Detection, quantification, malignancy prediction and growth forecasting of pulmonary nodules using deep learning in follow-up CT scans

Xavier Rafael Palou

DOCTORAL THESIS UPF / 2021

Thesis supervisors:
Prof. Miguel Ángel González Ballester,
Prof. Gemma Piella Fenoy,
PhD. Vicent Ribas Ripoll

Universitat
Pompeu Fabra
Barcelona

Acknowledgements

First and foremost, I want to express my gratitude to my supervisors, for their patience and guidance along this path. Without your support, this would not have been possible, and without your insights, I could not have reached this far. Especially, I want to express my thankfulness to Professor Miguel Ángel González Ballester, I have been really fortunate to have you as tutor and supervisor, your understanding, positivity and encouragement during all meetings and scientific discussions is something that I save for the future.

I would like to thank my colleagues at the University, especially PhD. Mario Ceresa, and at the data analytics in medicine group from Eurecat for their support, recommendations and feedbacks. Also, I would like to thank the clinical collaborators at the Vall d'Hebron Hospital, in particular to M.D. Esther Pallisa, M.D. Oscar Persiva and M.D. Anton Aubanell. Their patience in explaining the most basic concepts of radiology to me, and the time they took from their busy schedules for the data collection and annotation process, has been crucial for this thesis. Also, a special gratitude to PhD. Felip Miralles and Eurecat as an institution, for their patience with me and above all, for allowing me to carry out this work.

Finally, I want to thank my family and parents, they have always been there showing me their support and unconditional love. A very special thanks is to my wife, Janka. During this period, we have experienced good but also hard times. Thanks for your understanding, patience and love. I am deeply indebted to you and our daughter. You really are my main strength and motivation.

Abstract

Nowadays, lung cancer assessment is a complex and tedious task mainly performed by radiological visual inspection of suspicious pulmonary nodules, using computed tomography (CT) scan images taken to patients over time.

Several computational tools relying on conventional artificial intelligence and computer vision algorithms have been proposed for supporting lung cancer detection and classification. These solutions mostly rely on the analysis of individual lung CT images of patients and on the use of hand-crafted image descriptors. Unfortunately, this makes them unable to cope with the complexity and variability of the problem. Recently, the advent of deep learning has led to a major breakthrough in the medical image domain, outperforming conventional approaches. Despite recent promising achievements in nodule detection, segmentation, and lung cancer classification, radiologists are still reluctant to use these solutions in their day-to-day clinical practice. One of the main reasons is that current solutions do not provide support to automatic analysis of the temporal evolution of lung tumours. The difficulty to collect and annotate longitudinal lung CT cases to train models may partially explain the lack of deep learning studies that address this issue.

In this dissertation, we investigate how to automatically provide lung cancer assessment through deep learning algorithms and computer vision pipelines, especially taking into consideration the temporal evolution of the pulmonary nodules. To this end, our first goal consisted on obtaining accurate methods for lung cancer assessment (diagnostic ground truth) based on individual lung CT images. Since these types of labels are expensive and difficult to collect (e.g. usually after biopsy), we proposed to train different deep learning models, based on 3D convolutional neural networks (CNN), to predict nodule malignancy based on radiologist visual inspection annotations (which are reasonable to obtain). These classifiers were built based on ground truth consisting of the nodule malignancy, the position and the size of the nodules to classify. Next, we evaluated different ways of synthesizing the knowledge embedded by the nodule malignancy neural network, into an end-to-end pipeline aimed to detect pulmonary nodules and predict lung cancer at the patient level, given a lung CT image. The positive re-

sults confirmed the convenience of using CNNs for modelling nodule malignancy, according to radiologists, for the automatic prediction of lung cancer.

Next, we focused on the analysis of lung CT image series. Thus, we first faced the problem of automatically re-identifying pulmonary nodules from different lung CT scans of the same patient. To do this, we present a novel method based on a Siamese neural network (SNN) to rank similarity between nodules, overpassing the need for image registration. This change of paradigm avoided introducing potentially erroneous image deformations and provided computationally faster results. Different configurations of the SNN were examined, including the application of transfer learning, using different loss functions, and the combination of several feature maps of different network levels. This method obtained state-of-the-art performances for nodule matching both in an isolated manner and embedded in an end-to-end nodule growth detection pipeline.

Afterwards, we moved to the core problem of supporting radiologists in the longitudinal management of lung cancer. For this purpose, we created a novel end-to-end deep learning pipeline, composed of four stages that completely automatize from the detection of nodules to the classification of cancer, through the detection of growth in the nodules. In addition, the pipeline integrated a novel approach for nodule growth detection, which relies on a recent hierarchical probabilistic segmentation network adapted to report uncertainty estimates. Also, a second novel method was introduced for lung cancer nodule classification, integrating into a two stream 3D-CNN the estimated nodule malignancy probabilities derived from a pre-trained nodule malignancy network. The pipeline was evaluated in a longitudinal cohort and the reported outcomes (i.e. nodule detection, re-identification, growth quantification, and malignancy prediction) were comparable with state-of-the-art work, focused on solving one or a few of the functionalities of our pipeline.

Thereafter, we also investigated how to help clinicians to prescribe more accurate tumour treatments and surgical planning. Thus, we created a novel method to forecast nodule growth given a single image of the nodule. Particularly, the method relied on a hierarchical, probabilistic and generative deep neural network able to produce multiple consistent future segmentations of the nodule at a given time. To do this, the network learned to model the multimodal posterior distribution of future lung tumour segmentations by using variational inference and injecting the posterior latent features. Eventually, by applying Monte-Carlo sampling on the outputs of the trained network, we estimated the expected tumour growth mean and the uncertainty associated with the prediction.

Although further evaluation in a larger cohort would be highly recommended, the proposed methods reported accurate results to adequately support the radiological workflow of pulmonary nodule follow-up. Beyond this specific application, the outlined innovations, such as the methods for integrating CNNs into com-

puter vision pipelines, the re-identification of suspicious regions over time based on SNNs, without the need to warp the inherent image structure, or the proposed deep generative and probabilistic network to model tumour growth considering ambiguous images and label uncertainty, they could be easily applicable to other types of cancer (e.g. pancreas), clinical diseases (e.g. Covid-19) or medical applications (e.g. therapy follow-up).

Resum

Avui en dia, l'avaluació del càncer de pulmó és una tasca complexa i tediosa, principalment realitzada per inspecció visual radiològica de nòduls pulmonars sospitosos, mitjançant imatges de tomografia computada (TC) preses als pacients al llarg del temps.

Actualment, existeixen diverses eines computacionals basades en intel·ligència artificial i algorismes de visió per computador per donar suport a la detecció i classificació del càncer de pulmó. Aquestes solucions es basen majoritàriament en l'anàlisi d'imatges individuals de TC pulmonar dels pacients i en l'ús de descriptors d'imatges fets a mà. Malauradament, això les fa incapaces d'afrontar completament la complexitat i la variabilitat del problema. Recentment, l'aparició de l'aprenentatge profund ha permès un gran avenç en el camp de la imatge mèdica. Malgrat els prometedors assoliments en detecció de nòduls, segmentació i classificació del càncer de pulmó, els radiòlegs encara són reticents a utilitzar aquestes solucions en el seu dia a dia. Un dels principals motius és que les solucions actuals no proporcionen suport automàtic per analitzar l'evolució temporal dels tumors pulmonars. La dificultat de recopilar i anotar cohorts longitudinals de TC pulmonar poden explicar la manca de treballs d'aprenentatge profund que aborden aquest problema.

En aquesta tesi investiguem com abordar el suport automàtic a l'avaluació del càncer de pulmó, construint algoritmes d'aprenentatge profund i pipelines de visió per ordinador que, especialment, tenen en compte l'evolució temporal dels nòduls pulmonars. Així doncs, el nostre primer objectiu va consistir a obtenir mètodes precisos per a l'avaluació del càncer de pulmó basats en imatges de CT pulmonar individuals. Atès que aquests tipus d'etiquetes són costoses i difícils d'obtenir (per exemple, després d'una biòpsia), vam dissenyar diferents xarxes neuronals profundes, basades en xarxes de convolució 3D (CNN), per predir la malignitat dels nòduls basada en la inspecció visual dels radiòlegs (més senzilles de recol.lectar). A continuació, vàrem avaluar diferents maneres de sintetitzar aquest coneixement representat en la xarxa neuronal de malignitat, en una pipeline destinada a proporcionar predicció del càncer de pulmó a nivell de pacient, donada una imatge de TC pulmonar. Els resultats positius van confirmar la conveniència

d'utilitzar CNN per modelar la malignitat dels nòduls, segons els radiòlegs, per a la predicció automàtica del càncer de pulmó.

Seguidament, vam dirigir la nostra investigació cap a l'anàlisi de sèries d'imatges de TC pulmonar. Per tant, ens vam enfrontar primer a la reidentificació automàtica de nòduls pulmonars de diferents tomografies pulmonars. Per fer-ho, vam proposar utilitzar xarxes neuronals siameses (SNN) per classificar la similitud entre nòduls, superant la necessitat de registre d'imatges. Aquest canvi de paradigma va evitar possibles pertorbacions de la imatge i va proporcionar resultats computacionalment més ràpids. Es van examinar diferents configuracions del SNN convencional, que van des de l'aplicació de l'aprenentatge de transferència, utilitzant diferents funcions de pèrdua, fins a la combinació de diversos mapes de característiques de diferents nivells de xarxa. Aquest mètode va obtenir resultats d'estat de la tècnica per reidentificar nòduls de manera aïllada, i de forma integrada en una pipeline per a la quantificació de creixement de nòduls.

A més, vam abordar el problema de donar suport als radiòlegs en la gestió longitudinal del càncer de pulmó. Amb aquesta finalitat, vam proposar una nova pipeline d'aprenentatge profund, composta de quatre etapes que s'automatitzen completament i que van des de la detecció de nòduls fins a la classificació del càncer, passant per la detecció del creixement dels nòduls. A més, la pipeline va integrar un nou enfocament per a la detecció del creixement dels nòduls, que es basava en una recent xarxa de segmentació probabilística jeràrquica adaptada per informar estimacions d'incertesa. A més, es va introduir un segon mètode per a la classificació dels nòduls del càncer de pulmó, que integrava en una xarxa 3D-CNN de dos fluxos les probabilitats estimades de malignitat dels nòduls derivades de la xarxa pre-entrenada de malignitat dels nòduls. La pipeline es va avaluar en una cohort longitudinal i va informar rendiments comparables a l'estat de la tècnica utilitzats individualment o en pipelines però amb menys components que la proposada.

Finalment, també vam investigar com ajudar els metges a prescriure de forma més acurada tractaments tumorals i planificacions quirúrgiques més precises. Amb aquesta finalitat, hem realitzat un nou mètode per predir el creixement dels nòduls donada una única imatge del nòdul. Particularment, el mètode es basa en una xarxa neuronal profunda jeràrquica, probabilística i generativa capaç de produir múltiples segmentacions de nòduls futurs consistents del nòdul en un moment determinat. Per fer-ho, la xarxa apren a modelar la distribució posterior multimodal de futures segmentacions de tumors pulmonars mitjançant la utilització d'inferència variacional i la injecció de les característiques latents posteriors. Finalment, aplicant el mostreig de Monte-Carlo a les sortides de la xarxa, podem estimar la mitjana de creixement del tumor i la incertesa associada a la predicció.

Tot i que es recomanable una avaluació posterior en una cohort més gran, els mètodes proposats en aquest treball han informat resultats prou precisos per donar suport adequadament al flux de treball radiològic del seguiment dels nòduls pulmonars. Més enllà d'aquesta aplicació específica, les innovacions presentades com, per exemple, els mètodes per integrar les xarxes CNN a pipelines de visió per ordinador, la reidentificació de regions sospitoses al llarg del temps basades en SNN, sense la necessitat de deformar l'estructura de la imatge inherent o la xarxa probabilística per modelar el creixement del tumor tenint en compte imatges ambigües i la incertesa en les prediccions, podrien ser fàcilment aplicables a altres tipus de càncer (per exemple, pàncrees), malalties clíniques (per exemple, Covid-19) o aplicacions mèdiques (per exemple, seguiment de la teràpia).

Preface

The present thesis was carried out within the framework of an Industrial Doctorate (2017-DI087), under the supervision of Prof. Miguel Ángel González Ballester, ICREA Research Professor at the Department of Information and Communication Technologies of Universitat Pompeu Fabra (UPF) in Barcelona, Prof. Gemma Piella Fenoy, Professor at the Department of Information and Communication Technologies of Universitat Pompeu Fabra (UPF) in Barcelona, and Dr. Vicent Ribas Ripoll, head of the data analytics in medicine research line of the digital health unit at Eurecat.

The clinical research requirements, clinical background as well as the CT scans necessary to build most of the models in this thesis were obtained thanks to the close collaboration with the radiologists of the Vall d'Hebron Research Institute in Barcelona.

The Industrial Doctorates Plan (http://doctoratsindustrials.gencat.cat/) is promoted by the Generalitat de Catalunya and led by the "Agencia de Gestió d'Ajuts Universitaris i de Recerca" (AGAUR), to help companies to improve their competitiveness and productivity through the promotion of a better use of knowledge, technology and skills within universities, colleges and research organizations.

Eurecat

Eurecat (www.eurecat.org) is the result of the merging process of the most important Technology Centres in Catalonia, a process which started in 2015 and is still ongoing. It counts already with the sum of capacities of eight Centres and its multiplying effects.

Eurecat is currently the leading Technology Centre in Catalonia, and the second-largest private research organization in Southern Europe. Eurecat manages a turnover of 50M€ and 650 professionals, is involved in more than 200 R&D projects and has a customer portfolio of over 1600 business clients. Eurecat is currently participating in more than 60 EU funded collaborative projects, mainly in the Horizon 2020 Programme. In addition to this wide experience at European level, Eurecat is also a strong player in the various R&D programmes sponsored by the Spanish and Catalan administrations. Technology transfer is also an essential activity at Eurecat, with 36 international patents and 7 technology-based companies (eight in Spain and one in Latin America) started-up from the centre.

Vall d'Hebron Research Institute

Vall d'Hebron Research Institute (VHIR) is a public sector institution that promotes and develops the biomedical research, innovation and teaching at Vall d'Hebron University Hospital (HUVH), which is the largest hospital of the Catalan Institute of Health (ICS).

Since its creation in 1994, VHIR has more than 1100 people conducting research and around 750 helping to transfer it to society in the form of projects, technology transfer and innovation, communication or fundraising. In April 2015, VHIR was recognized with the 'HR Excellence in Research' accreditation from the European Commission (also known as "HRS4R").

Contents

List of figures **xx**

List of tables **xxiii**

1 INTRODUCTION **1**

1.1 Clinical context 1
1.2 Methodological context 2
1.3 Objectives of the thesis 6
1.4 Outline of the thesis 7

2 CLINICAL BACKGROUND **9**

2.1 Lung cancer 9
2.2 Pulmonary nodules 11
2.3 Computed Tomography 14
2.4 Managing pulmonary nodules 15
2.5 Challenges and limitations 16

3 METHODOLOGICAL BACKGROUND **17**

3.1 Deep learning for image analysis 17
  3.1.1 Image classification 19
  3.1.2 Object detection 21
  3.1.3 Image segmentation 23
  3.1.4 Image generation 24
  3.1.5 Temporal image analysis 25
3.2 Automatic lung cancer assessment 27
  3.2.1 Common lung cancer datasets 28
  3.2.2 Lung preprocessing 29
  3.2.3 Nodule detection 30
  3.2.4 Nodule segmentation 32
  3.2.5 Malignancy classification 33
  3.2.6 Temporal nodule analysis 35

3.3 Challenges and limitations 40

4 INTEGRATION OF PULMONARY NODULE MALIGNANCY IN A LUNG CANCER CLASSIFICATION PIPELINE 43
4.1 Introduction 43
4.2 Related Works 45
4.3 Materials and Methods 46
  4.3.1 Data 46
  4.3.2 Method 46
4.4 Results 53
  4.4.1 Malignancy classification 53
  4.4.2 Lung cancer 54
4.5 Discussion 56
4.6 Conclusions 59

5 RE-IDENTIFICATION AND GROWTH DETECTION OF PULMONARY NODULES WITHOUT IMAGE REGISTRATION USING 3D SIAMESE NEURAL NETWORKS 61
5.1 Introduction 61
5.2 Related work 62
  5.2.1 Automated nodule re-identification 62
  5.2.2 Siamese Neural Networks 63
5.3 Method 64
  5.3.1 Nodule re-identification 64
  5.3.2 Nodule growth detection pipeline 68
5.4 Experiments and results 70
  5.4.1 Evaluation datasets 70
  5.4.2 Nodule re-identification 71
  5.4.3 Nodule growth detection pipeline 72
  5.4.4 Automatic lung CTs registration 75
5.5 Discussion 77
5.6 Conclusions 84

6 END-TO-END AUTOMATIC PIPELINE FOR PULMONARY NODULE FOLLOW-UP ASSESSMENT 85
6.1 Introduction 85
6.2 Background 86
  6.2.1 Nodule detection 87
  6.2.2 Nodule quantification 87
  6.2.3 Lung cancer prediction 88
6.3 Method 89

6.4 Results 95
6.5 Discussion 98
6.6 Conclusions 101

7 AN UNCERTAINTY-AWARE HIERARCHICAL PROBABILISTIC NETWORK FOR EARLY PREDICTION, QUANTIFICATION AND SEGMENTATION OF PULMONARY TUMOUR GROWTH 103
7.1 Introduction 103
7.2 Related work 105
  7.2.1 Deep learning deterministic approaches 105
  7.2.2 Deep generative networks 107
  7.2.3 Uncertainty in deep learning 108
7.3 Method 109
  7.3.1 Hierarchical probabilistic U-Net 109
  7.3.2 U-HPNet 111
  7.3.3 Comparison with related works 114
  7.3.4 Tumour growth assessment 116
7.4 Experiments and Results 118
  7.4.1 VH-Lung 118
  7.4.2 Qualitative results 120
  7.4.3 Quantitative results 123
  7.4.4 Ablation studies 126
  7.4.5 Comparison with other networks 128
7.5 Discussion 129
7.6 Conclusion 133

8 CONCLUSIONS AND FUTURE WORK 135
8.1 Overview 135
8.2 Outlook and future work 141
8.3 Publication list 146

Appendices 148

A SUPPLEMENTARY MATERIAL I 149
A.1 Significance of radiologists' annotated features 149
A.2 The lung cancer pipeline 150

B SUPPLEMENTARY MATERIAL II 155
B.1 Nodule classification 155
B.2 Nodule detector architecture 156
B.3 Nodule re-identification 157

B.4 Lung nodule re-identification performance 159
B.5 Nodule growth pipeline performance 159

List of Figures

2.1 The left lung and right lung. The lobes of the lungs can be seen, and the central root of the lung is also present. Credits: Henry Vandyke Carter. Public domain. 10

2.2 Microscopic view of a non-small cell lung carcinoma (left) and a small-cell lung carcinoma (right). Credits: KGH and Librepath, CC 3.0. 10

2.3 On the left, a solid carcinoma tumour; on the right, a part-solid adenocarcinoma. 11

2.4 On the left, well-delineated solid lung nodule with smooth border. In the centre of the image, a lobulated nodule. On the right, a spiculated nodule. Image taken from [274]. 12

2.5 Diameter size (D1 and D2), growth (difference in diameter), time difference and malignancy for 5 different nodules. The top row is time-point 1, and the second row is time-point 2. 13

2.6 Image of a modern CT scanner (left) and an axial slice of a CT scan with a lung nodule (right). Credits: daveynin, CC 2.0. 14

3.1 Vanilla architecture of a CNN. 19

3.2 Overview of an Inception network block in GoogLeNet (Image taken from [290]). 20

3.3 Residual block by ResNet networks (Image taken from [114]). . . 20

3.4 Overview of the Faster R-CNN object detection network (Image taken from [243]). 21

3.5 Example of a SSD network architecture for object detection (Image taken from [186]). 22

3.6 Example of a U-Net network architecture. Credit Mehrdad Yazdani CC 4.0. 23

3.7 Example of an LSTM network architecture. Image taken from [218]. 27

4.1 PCA analysis and boxplot of radiologists annotated features per malignancy class. 48

4.2 Architecture of the two CNN networks designed for the malignancy classifiers. 48
4.3 Pipeline proposed for lung cancer classification. 52
4.4 Data distribution for lung cancer classification. 55
4.5 Performance comparison of the lung cancer pipelines. 56

5.1 Siamese network proposed for lung nodule re-identification. The network is composed of a feature extraction and a basic head network to perform the prediction. 65
5.2 Alternative head networks to configure different siamese networks. 65
5.3 Nodule growth detection pipeline. 69
5.4 FROC-curve of the malignant nodule detection algorithm for training and test partition. 73
5.5 Candidates predicted (yellow marks) at a maximum distance from the ground truth centroid (red circle). 74
5.6 Confusion matrix for nodule growth prediction. 75
5.7 Bland$-$Altman plot for agreement between ground truth and predicted nodule growth. 76
5.8 Comparison between real and predicted cases. Upper panel: diameter differences for all test set. Lower panel: axial slices at two time points of different nodules. 76

6.1 Pipeline architecture for the temporal analysis of lung nodules. . . 89
6.2 Architecture proposed for the nodule detection network of the pipeline. 90
6.3 Architecture of the 3D-SNN for the nodule re-identification component of the pipeline. 91
6.4 Hierarchical probabilistic U-Net network architecture overview. T_i is the nodule image i, S_i is the ground truth segmentation of the nodule T_i, and S'_i the predicted segmentation of the nodule T_i. 92
6.5 Nodule malignancy classification architecture of the TS-3DCNN-MAL network. 94
6.6 Comparative of radiologist growth measurements with results from nodule detector and nodule segmentation. 97
6.7 ROC curves of the nodule malignancy classification models. . . . 98
6.8 Grad-CAM features from the TS-3DCNN-TL-MAL network for 2 malign (upper row) and benign (bottom row) nodules. 99

7.1 General overview of the HPU network architecture. On the left of the picture we can observe the prior network and on the right the posterior. Both networks have different probabilistic latent blocks interleaved along the decoder component. 110

7.2 General overview of the proposed U-HPNet network architecture. This network is also composed by a prior network (on the left, with further details) and a posterior (on the right). Attached at the end of the posterior we observe the post-process module aimed at reporting the estimated future growth prediction, size and appearance with the associated uncertainty. 111

7.3 On the left, we show a more detailed view of the different components of a decoder layer of the U-HPNet. On the right, we see the elements that compose an attention block. 112

7.4 Four alternative network architectures proposed for lung nodule growth estimation. At the top we have the U-ResNet and the generative Bayesian dropout. In the centre, we show the probabilistic U-Net. Below we find the Pix2Pix cGAN network proposed. . . . 115

7.5 Description of the training and test set cohorts in terms of tumour growth size (in mm), time between studies (days) and the annotators (RX-0,1,2). 119

7.6 Comparison of ground truth annotations and predictions from the U-HPNet for the tumour case C94. 120

7.7 Comparison of ground truth annotations and predictions from the U-HPNet for the tumour case B01. 121

7.8 Comparison of ground truth annotations and predictions from the U-HPNet for the tumours C17, B64, B19 and C16. In this example, the network incorrectly predicts growth for the three first cases while for the last one it guesses the prediction, although the predicted growth size is far from the radiologist measurement. . . 122

7.9 Tumour growth prediction accuracy of the U-HPNet stratified by time to predict, nodule growth size and ground truth (i.e. Closest, RX0, RX1 and RX2). 124

8.1 Findings section window of the front end interface with the results of a CT study. 144

8.2 Screen with the results of the re-identification analysis. In this particular example, we observe the same nodule found in both CT studies, its growth and its estimated malignancy. 145

A.1 FROC curve achieved in testing for the 3D ResNet network. . . . 152

B.1 Architecture for the nodule detector. 156
B.2 FROC curve of the lung nodule detector computed for the test set. 157

List of Tables

2.1 Mapping of HU values with substance type. 15

4.1 Dataset used for building and testing the malignancy classifier. . . 47
4.2 Results of nodule malignancy classification on test set (at nodule level). 53
4.3 Cancer prediction at patient level from nodule malignancy. 54
4.4 Cross-validation and test (ND: nodule level, PT: patient level) results for the lung cancer pipelines. 56

5.1 List of the different siamese network configurations. The index column contains the acronyms of the networks, resulting from joining the first letter of the options placed in the next 4 columns. . 66
5.2 Layers selected from the pre-trained part of the SNNs. 67
5.3 Performance results (accuracy (acc), precision (prec) and recall (rec)) obtained on training (tr), validation (val) and test for the different SNN configurations. The meaning of the configured methods is detailed in Table5.1. 72
5.4 Number of CTs (in T2) containing N candidates located within a chosen euclidean distance from the actual nodule centroid. The accuracy score represents the number of CTs at N=1 respect to the total of CTs. 74
5.5 Results of the different nodule re-identification pipelines. The meaning of the configured methods is detailed in Table5.1. 74
5.6 Results after applying automatic registration using rigid and non-rigid approaches. 77

6.1 Performance of the re-identification component of the pipeline. . . 96
6.2 Nodule growth performance comparison between 3D-FasterCNN and HPU networks. 96
6.3 Performance of the different nodule malignancy classifiers of component of the pipeline. 97

7.1 Nodule growth prediction performances obtained by the U-HPNet using expected means. 124
7.2 Nodule growth balanced accuracy performances for the U-HPNet obtained within 2 standard deviations from the mean. 125
7.3 Nodule growth size performances of the U-HPNet using the estimated mean (MAE, MSE) and the interval composed by the mean and 2 standard deviations. A down arrow next to a metric means that the metric is more accurate the smaller the value. 125
7.4 Nodule segmentation performances of the U-HPNet for each of the ground truths. 126
7.5 GED nodule segmentation performance of the U-HPNet. Each score reports 1-IoU metric. 126
7.6 Different network setups of the U-HPNet configured in the ablation study. 127
7.7 Performance comparison between the different U-HPNet setups. . 127
7.8 Generative ability comparison between the different U-HPNet setups. 127
7.9 Performance comparison with alternative networks for tumour growth using the expected mean. 128
7.10 Generative performance of alternative networks for tumour growth. 129

A.1 Statistical significance of the radiologists annotated features. P-values are obtained through Wald test of significance. 149
A.2 Configurations of the Difference of Gaussian method for lung nodules detection. 150
A.3 Results from three different configurations of the Difference of Gaussian method for lung nodules detection. The total number of nodules in the test set was 105. 151
A.4 Confusion matrix results for the 3D ResNet network. 151
A.5 Classification results for the 3D ResNet network. 152
A.6 Pipeline parameters tested using grid-search and 5-fold CV. 153
A.7 Parameters for training the dense network. 153

B.1 Confusion matrix results for the 3D ResNet network. 155
B.2 Classification results for the 3D ResNet network. 155
B.3 Performances of the lung nodules detector at different FP in average per scan. 157
B.4 Extended experiment results of different SNN configurations. The meaning of the acronyms are detailed in the Table 1 of the manuscript. 158

B.5 Performances for the whole training and test datasets stratified by lung nodule change size and matching/non-matching nodule pairs. Each cell contains accuracy and total number of nodule pairs. . . . 159

B.6 Performances for the whole training and test datasets stratified by lung nodule change size. In each cell we provide the accuracy for nodule detection (ND), nodule re-identification (RI) and F1-score for nodule growth detection (GD). In parenthesis we show the total number of nodule pairs involved per pipeline process and change size. 159

Chapter 1

INTRODUCTION

1.1 Clinical context

According to World Health Organization[1] Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020[2]. Lung cancer, in particular, is the second most common in terms of new cases of cancer, with 2.21 million cases (11.4%), and the most aggressive in number of deaths (both sexes, all ages) with 1.8 million cases (18%) in 2020.

Early detection of lung cancer significantly improves the chances of patient survival. Depending on the lung cancer type, patients with an advanced stage of the disease have a 1-year survival rate of only 15-19%, compared with 81–85% for patients treated with the disease identified at early stages [35]. Unfortunately, in most cases, patients are unaware that they have a pulmonary nodule until physical symptoms appear, which most often occur in advanced stages of the disease. For this reason, early-stage detection and classification of benign and malignant pulmonary nodules plays an important role in clinical diagnosis.

Today, the gold standard for lung cancer detection consists in routinely taking a computed tomography (CT) scan, and detecting nodules (i.e. small and approximately spherical masses) in it [298]. Once pulmonary nodules have been identified, radiologists normally perform size and growth rate quantification studies to assess their malignancy. However, lung CT images often have a low signal-to-noise ratio, causing misclassifications of regions with weak or irregular contours. Also, lung cancer diagnosis through CT is often subjective and highly affected by observer's experience, fatigue and emotional state [227], which can lead to inconsistent results from the same radiologist at different times or from different radiologists examining the same CT image.

---

[1] https://www.who.int/news-room/fact-sheets/detail/cancer

[2] Global Cancer Observatory: Cancer Today. https://gco.iarc.fr/today, accessed April 2021

To support radiologists in the management of the lung cancer disease, several guidelines like LungRADs [12] and Fleischner [194] have been proposed. These guidelines are a compilation of well documented cases and a set of rule-based recommendations from the clinical experience designed to help clinicians to discern among pulmonary nodules, normal tissues and artefacts, as well as to determine the inherent malignancy of the nodules. However, they are constrained to a limited number of visual parameters (e.g. size, morphology, texture and location of the nodules) and to a fixed range of values. In addition, the recommended clinical actions are often too general and vague, making them insufficiently suitable for specific patients, or for supporting rare or borderline cases.

Unfortunately, the radiological imaging units of health institutions are often overloaded due to limited resources. This makes it unfeasible to have all the medical care required at any given time for any patient. Therefore, efficient and accurate computational support could help to address and unblock this complicated healthcare scenario.

To overcome current clinical limitations, this thesis proposes the development of accurate predictive methods to analyse lung CT scan images (potentially including follow-up scans) to automatically provide the most relevant information to the radiologists such as location, diameter, growth and malignancy of the nodules, as well as to predict their potential evolution.

1.2 Methodological context

Traditionally, automated support for lung cancer assessment has been addressed using conventional image analysis techniques applied to individual lung CT images. Specifically, some of these techniques have been used for the detection of pulmonary nodules (e.g. using Laplacian of Gaussian (LoG) filters [86], histogram of oriented gradients [59] or wavelet feature descriptors [221]), for the nodule segmentation (e.g. using region growing-based approaches [303]) or for the nodule malignancy classification (e.g. building conventional machine learning algorithms such as random forest [120], gradient boosting machines [82] or support vector machines [55]). Unfortunately, the resulting proposed solutions [80, 303], which mostly rely on pre-conceived notions of the suspicious regions through hand-crafted image descriptors, turned out to be not effective enough to fully capture the complexity and variability of pulmonary nodules [124].

Recently, the advent of deep learning [167], and in particular thanks to convolutional neural networks (CNN) [168], has allowed a breakthrough in the medical image analysis domain [31, 78, 100]. Specifically, for the lung cancer assessment problem, several studies using CNNs have shown outstanding performances, surpassing those of conventional approaches, for nodule detection [262], segmentation [200] and malignancy classification [51]. Some of the reasons that explain these achievements lie in the ability of CNNs to automatically extract high level feature imaging representations, in which its shared-weights architecture enables it to capture basic image properties, such as high correlation among local values and translation invariance [54].

Hence, in a relatively short time, CNNs have been widely investigated for analysing lung CT images. In particular, for the nodule detection problem, recent solutions [212, 334] propose the use of object detection networks such as Faster R-CNN [244]. This method together with SSD [186] and YOLO [241] are well-known architectures originally proposed for object detection on natural images. Specifically, Faster R-CNN although slightly slower than the others, are particularly suitable for nodule detection because of its flexibility on defining the initial anchor boxes. For the problem of nodule segmentation, U-Net based architectures [248] have been extensively studied, using either 2D/3D input patch imaging, and several variants can be found in the literature [133, 333]. This network usually provides high segmentation performances thanks to a convolutional encoder and decoder backbone tied at different levels by short-cuts, which allow by-passing high level features of the encoder to the decoder, in order to enhance the image reconstruction. For lung nodule malignancy classification (either using radiological observation or diagnostically confirmed labels), tailored 2D multi-view/multi-scale (commonly known as 2'5D) and 3D CNNs are frequently used for this problem (e.g. [8, 184, 266]), most of them relying on standard CNN architectures for classification [213, 214], such as VGG-16 [272], ResNet[115] or DenseNet [128].

Despite the high performances generally reported by these studies, few works have addressed the lung cancer assessment in an end-to-end manner [178]. One of the main reasons is the lack of available datasets with confirmed lung cancer diagnosis. To overcome this initial limitation, in **chapter 4** we developed tailored 3D CNNs aimed at predicting nodule malignancy (i.e. subjective measure provided by radiologists) for which data are more abundant and labels are cheaper to obtain. The proposed CNNs showed radiologist nodule malignancy classification performances. Subsequently, we designed a transfer learning framework to integrate nodule malignancy CNNs into an end-to-end lung cancer pipeline devoted to detect nodules and predict lung cancer at the patient level, given as input a raw lung CT image study. Interestingly, this approach allowed increasing dramatically the cancer classification performance of the pipeline.

Since current radiological practice for lung cancer assessment is based on the visual inspection of different lung CT studies performed at different time-points on the patients, solutions for analysing single time-point CT images remain insufficient, providing only partial support to the entire radiological workflow. Thus, a critical challenge is to provide algorithms capable of analysing and extracting relevant spatial and temporal patterns, to support clinicians in the lung nodule follow-up. To the best of our knowledge, studies using deep learning to address this problem are scarce, in part due to the lack of open annotated longitudinal data. Although we can find large longitudinal cohorts for lung cancer assessment, such as NLST [297], these are not publicly available and obtained through population-based screenings, which usually present findings with different characteristics from incidental cohorts and, therefore, are advised by different clinical recommendation guidelines [12]. In this dissertation, we were able to collect (in collaboration with radiologists of the Vall d'Hebron hospital at Barcelona) a new cohort composed of more than 150 labelled incidental cases to enable the temporal lung nodule analysis.

To provide automatic support to clinicians in the analysis of the temporal evolution of pulmonary nodules, we had to face several challenges. In first term, nodules have to be automatically re-identified from different lung CT images. Current solutions rely on image registration techniques [39, 207]. These are usually time-consuming and potentially need to modify the structure of the image to be able to align both lung CTs. Therefore, in **chapter 5**, we addressed these limitations by contributing with a novel and agile approach based on a 3D Siamese neural network (SNN) [152] avoiding the need of having the lung CT images previously registered. SNNs are a type of CNNs suitable for predicting similarity between images thanks to its original architecture in which the features from two input images are extracted using two sibling networks (normally sharing architecture and weights), and compared by a distance layer at the top of the network. They have been extensively used in computer vision matching problems such as tracking objects in videos [296]. However, to the best of our knowledge, SNNs have not been applied before to re-identify nodules in a series of lung CT scans. Another important task to be faced for the temporal lung nodule analysis is quantifying nodule growth. To address this issue, also in **chapter 5** we built an automatic pipeline able to, given two different CT studies of the same patient, detect, re-identify and quantify the nodule growth by subtracting the (major) diameter from the matching nodules, provided by a lung nodule detection network (i.e. 3D Faster R-CNN). However, due to the inherent ambiguity of the images (often contours of the nodules are not clearly delimited), it is desirable and safer reporting network uncertainty estimates when quantifying the size of the nodules. One way to learn model uncertainty is moving from one-input one-output to one-input multiple-output networks. This change of paradigm has already been tackled in deep neural networks through different approaches. One of the simplest approximations consists in ensembling multiple networks in order to provide multiple opinions [164]. Another approach consists in enabling dropout [279] at inference time in order to provide independent pixel-wise probabilities [142]. In **chapter 6**, we proposed a novel way to address nodule growth quantification extending a recent deep probabilistic and generative U-Net network [154], suitable for modelling ambiguous images, to provide nodule diameter and an uncertainty on this estimated measure. Another relevant task for temporal lung nodule analysis is to predict the lung cancer probability of the nodules. Unfortunately, few deep learning-based works have tackled lung cancer prediction using the temporal evolution of the nodules. Some solutions rely on CNNs combined with long short-term memory networks (LSTM) [121], which are a special type of recurrent neural networks (RNN) suitable for capturing long-term dependencies [71], or applying multi-stream CNN architectures [14] (approaches typically used for activity recognition [139]). In **chapter 6**, we extended the pipeline proposed in **chapter 5** with a new method able to provide the lung cancer probability of a lung nodule. In particular, we proposed a new 3D two-stream CNN. Our results confirmed the suitability of this approach by achieving high performances and surpassing 3D CNN approaches trained on single time-point images.

Finally, in **chapter 7**, we focused on the problem of predicting the evolution of the lung nodules, given a single image of the nodule and the time at which to provide the estimation. Cancer progression analysis has been traditionally addressed through complex and sophisticated mathematical models [254], such as those based on the reaction-diffusion equation [287, 292]. However, the number of parameters of such models is often limited (e.g. 5 in [318]), which might not be sufficient to capture the inherent complexities of the growing patterns of the tumours. Recently, deep learning has been used to predict future tumour growth, surpassing performances reported by traditional approaches [328]. Some of these solutions usually rely on architectures composed by CNNs in combination with LSTMs or using generative networks such as those based on adversarial learning [96] and variational auto-encoders [148] to estimate future images of the tumour [76, 231]. In the lung cancer domain, to the best of our knowledge, few works [177, 311] using deep learning have been proposed. Despite the reasonable performances reported by existing solutions, those have not adequately taken into account the inter-observer variability and uncertainty in the lung nodule estimations. Thus, we addressed the lung nodule forecasting problem, aiming to cover existing limitations. To do this, we adapted a recent hierarchical probabilistic framework [154] to model the posterior multimodal lung nodule growth probability distribution. Our approach was able to provide growth prediction, quantification and segmentation of the future nodule, together with a measure of uncertainty for the estimation. This network demonstrated better performances than a similar deterministic approach (i.e. U-ResNet [331]) and other alternative deep generative networks, such as a probabilistic U-Net [153], a conditional generative adversarial network [134] and a Bayesian dropout network [142].

1.3 Objectives of the thesis

The main goal of this thesis was to research and develop new methods based on deep learning using images from CT scans of the patients (single time-point and/or follow-up sets of images), to improve the performance of crucial parts of the radiological workflow regarding lung cancer management. Therefore, this thesis has been structured around the following specific objectives:

- Design and evaluate an automatic method to find the possible locations of pulmonary nodules in a lung CT. This functionality would help clinicians to rapidly detect all nodules present in a CT scan and annotate their position.

- Design and evaluate a method to automatically predict lung nodule malignancy using radiologists labels acquired from single CT image observation. This functionality would help the characterization of the malignancy patterns existing in the lung nodule image, as well as evaluating the ability of the automatic methods in comparison with the experts.

- Develop an end-to-end pipeline to predict lung cancer using the malignancy patterns automatically extracted from the nodules. This functionality would help clinicians to provide early indicators of cancer from a lung CT scan.

- Address the issue of the automatic spatial mapping of nodules between CT studies. The position of the nodule may be shifted between scans due to respiration or the positioning of the patient during the scanning process. Therefore, supporting clinicians in this problem would help them to save time.

- Design and evaluate a method to automatically predict cancer from series of lung nodule images using diagnosed cases. This functionality would support clinicians in the identification of lung cancer from the temporal evolution of the nodules, as well as assessing the contribution of the temporal feature in the prediction ability of the model.

- Develop an end-to-end pipeline to predict lung cancer using the temporal evolution of nodules in subsequent scans. This functionality would help them to provide more accurate lung cancer assessment as well as better therapeutic planning.

- Develop a method to predict, quantify and visualize lung nodule growth. Since growth is considered one of the most important factors in tumour malignancy, providing this method would help clinicians to anticipate the development of the disease and take more accurate and personalized treatments.

1.4 Outline of the thesis

The core contents of this thesis are presented in the following chapters.

Chapter 2 describes the clinical background. First, we define lung cancer and pulmonary nodules, and then the main diagnostic methods and management options are presented.

Chapter 3 details the methodological background of this thesis. We review the main deep learning methods and tools required for pulmonary nodule detection, segmentation and malignancy prediction.

Chapter 4 proposes different methods to integrate a 3D CNN for malignancy prediction into a lung cancer classification pipeline. This work has led to the following publications:

1. *Rafael-Palou Xavier, Bonavita Ilaria, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. Improving Lung Cancer Prediction with a Deep Learning Nodule Malignancy Classifier. International Journal of Computer Assisted Radiology and Surgery. Vol. 14 (Suppl. 1): 70-71, 2019.*

2. *Rafael-Palou Xavier, Bonavita Ilaria, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline. Computer Methods and Programs in Biomedicine. Vol. 185 (105172), pp. 1-9, 2020.*

Chapter 5 proposes a method to re-identify lung nodules located in different CT scan images. This work has led to the following publications:

1. *Rafael-Palou Xavier, Aubanell Anton, Bonavita Ilaria, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. 3D Siamese Neural Networks for Matching Pulmonary Nodules in Series of CT Scans. International Journal of Computer Assisted Radiology and Surgery. Vol. 15 (Suppl. 1), 2020.*

2. *Rafael-Palou Xavier, Aubanell Anton, Bonavita Ilaria, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. Re-Identification and Growth Detection of Pulmonary Nodules without Image Registration Using 3D Siamese Neural Networks. Medical Image Analysis. Vol. 67 (101823), pp. 1-12, 2021.*

Chapter 6 proposes a pipeline to detect, quantify and predict malignancy of pulmonary nodules using a pair of CT scan images. This work has led to the following publications:

1. *Rafael-Palou Xavier, Aubanell Anton, Bonavita Ilaria, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. Pulmonary Nodule Malignancy Classification Using its Temporal Evolution with Two-Stream 3D Convolutional Neural Networks. International Conference on Medical Imaging with Deep Learning (MIDL), 2020.*

2. *Rafael-Palou Xavier, Aubanell Anton, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. Detection, growth quantification and malignancy prediction of pulmonary nodules using deep convolutional networks in follow-up CT scans. To appear in Artificial Intelligence in Cancer Diagnosis, Volume 1: Lung and Kidney Cancer.*

Chapter 7 proposes a method to forecast lung nodule growth and its associated uncertainty given a single CT scan image. This work has led to the following publication(s):

1. *Rafael-Palou Xavier, Aubanell Anton, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. An Uncertainty-aware Hierarchical Probabilistic Network for Early Prediction, Quantification and Segmentation of Pulmonary Tumour Growth. Under review in Medical Image Analysis.*

Finally, **Chapter 8** concludes the thesis and discusses the outlook and directions for future work.

Chapter 2

CLINICAL BACKGROUND

This chapter describes the clinical background of this thesis. First, lung cancer is defined. Then, pulmonary nodules and principal malignancy factors are detailed. Lastly, the main methods and tools for pulmonary nodule management are introduced.

2.1 Lung cancer

Lung cancer is characterized by abnormal cells with an uncontrolled growth ability to potentially spread into nearby tissues or parts of the body [317]. These abnormal cells interfere with the normal function of the lung (Figure-2.1) in providing oxygen to the blood. The malignant behaviour of these cells is due to severe damage (or mutations) in the structure of their DNA sequence. Researchers have found that it takes a series of mutations to create a lung cancer cell [223]. Therefore, before becoming fully cancerous, cells can be precancerous, in that they have some mutations but still function normally as lung cells. However, after several cell divisions (in which the malignant genes are replicated), the eventual lung cell becomes more mutated and may not be as effective in carrying out its original function. In later stages of the disease, some cells may travel away from the original tumour and start growing in other parts of the body. This process is called metastasis.

DNA mutations in lung cells can be caused by the normal ageing process or through environmental factors. In lung cancer the main exogenous genotoxic agent is tobacco (e.g. acrolein, formaldehyde, acrylonitrile, 1,3-butadiene, acetaldehyde, ethylene oxide and isoprene) [58]. Specifically, long-term cigarette smoking is implicated in 85% of lung cancer cases [9], and therefore, is considered the major factor of risk. About 10–15% of cases occur in people who have never smoked [301], often caused by a combination of genetic factors and expo-

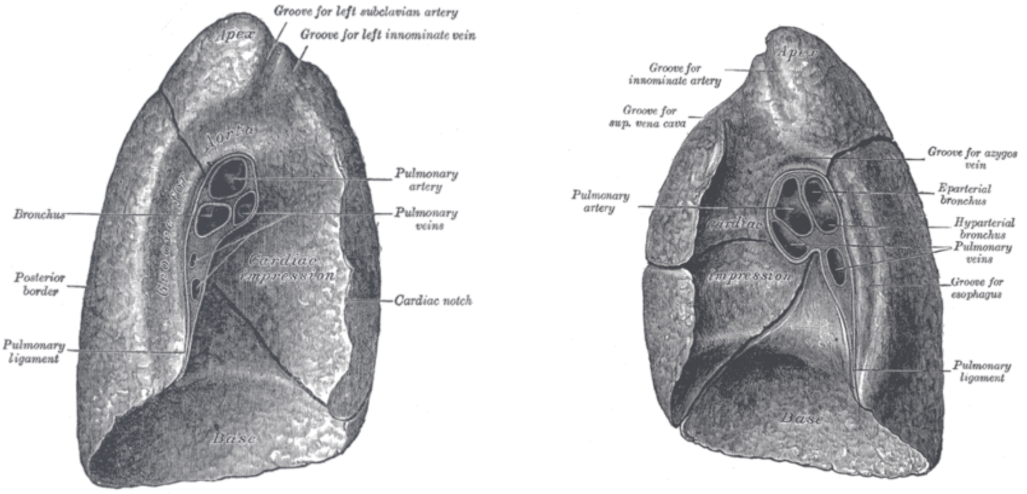


Figure 2.1: The left lung and right lung. The lobes of the lungs can be seen, and the central root of the lung is also present. Credits: Henry Vandyke Carter. Public domain.

sure to radon gas, asbestos, second-hand smoke, or other forms of air pollution [219, 220].

When cancer arises directly at any location of the structure of the lung (Figure-2.1), it is known as primary lung cancer [316], whereas it is secondary if it is originated in other part of the body and reaches the lung through metastasis. According to its histology (Figure-2.2), lung cancers are classified mainly in non-small cell lung carcinoma (NSCLC) and small cell lung carcinoma (SCLC). They differ clinically in terms of presentation, treatment, and prognosis [163]. Nearly 85% of lung cancers are NSLC, whereas the rest are SCLC [64]. NSCLC are further sub-divided into: adenocarcinoma, squamous-cell carcinoma and large-cell carcinoma [70]. Among all lung cancer types, adenocarcinoma is the most common with 40% of entire lung cancer incidence, followed by squamous cell carcinoma with 25-30% [6, 7]. Large cell carcinoma comprises 5–10% of total lung cancer [85].

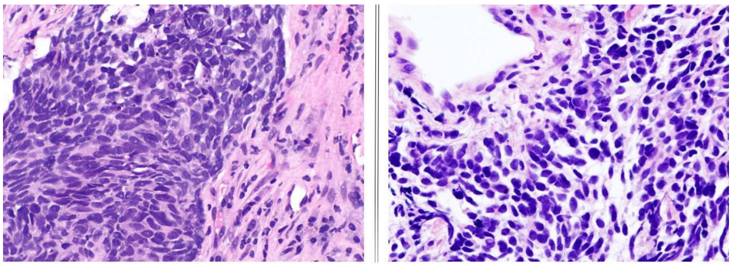


Figure 2.2: Microscopic view of a non-small cell lung carcinoma (left) and a small-cell lung carcinoma (right). Credits: KGH and Librepath, CC 3.0.

Treatment and prognosis for lung cancer depends on its stage. Stages I and II refer to cancers localized in the lungs, whilst latter stages (III or IV) refer to cancers that have spread to other organs. Early-stage lung cancer is non-specific and often asymptomatic[1], which explains why most cases are diagnosed at stage III or IV (representing 61% of all newly diagnosed lung cancers), and only 21% are diagnosed at stage I [201]. Therefore, early detection of lung cancer is crucial since it significantly improves the chances of survival.

2.2 Pulmonary nodules

One key characteristic of lung cancer is the presence of malignant pulmonary nodules. By definition, a lung nodule is a small round or oval-shaped growth in the lung, which may be well or poorly delineated, measuring less than three centimetres in diameter. A nodule smaller than 3 mm should be referred to as micronodule (difficult to be detected) [109]. If the diameter is larger than 3 cm, it is called a pulmonary mass and is more likely to represent a cancerous nodule.

There are two main types of pulmonary nodules: malignant (cancerous) and benign (non-cancerous). A wide variety of causes may originate the appearance of nodules in the lung, such as infections (from mycobacterium like tuberculosis, or fungal such as aspergillosis), non-infectious disorders (such as sarcoidosis) and abnormal growths or neoplasms. In the latter case, they still may be benign (e.g. fibroma, hamartoma) or malignant cancerous nodules (e.g. adenocarcinomas).

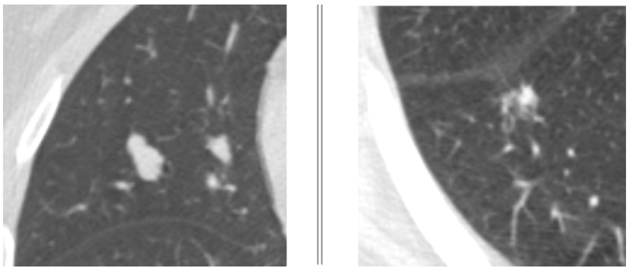


Figure 2.3: On the left, a solid carcinoma tumour; on the right, a part-solid adenocarcinoma.

Traditionally, to estimate the malignancy probability of a pulmonary nodule, radiologists consider different factors, some based on the clinical history (such as age, smoking history, and the presence of another malignancy 5 years prior to the time of evaluation) of the patient and others relying on the subjective radiographic appearance of the nodules [57, 102]. The typical imaging factors are:

---

[1] https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/signs-symptoms.html

- **Morphology**: Nodules can be categorized as solid, part-solid, and pure ground-glass (GGN). Criteria for making these distinctions have not been completely agreed upon and remain controversial. A solid nodule is a nodule that completely obscures the entire lung parenchyma within it. Part-solid nodules are those having sections that are solid, and ground-glass nodules are those with no solid parts with focal nodular areas of increased lung attenuation through which lung parenchymal structures can be observed. Part-solid and ground-glass nodules have a higher likelihood of being malignant when compared with solid nodules [91].

- **Edge characteristics**: This feature is typically specific of solid nodules. Nodule edges are classified as smooth, irregular, lobulated or spiculated. Marginal spiculation has been known for many years to be associated with malignancy, and more recent studies have confirmed spiculation as a risk factor for cancer [199]. Examples of nodules with different edge characteristics are shown in Figure-2.4.

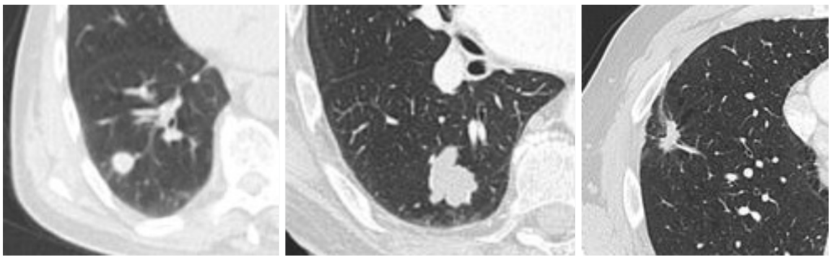


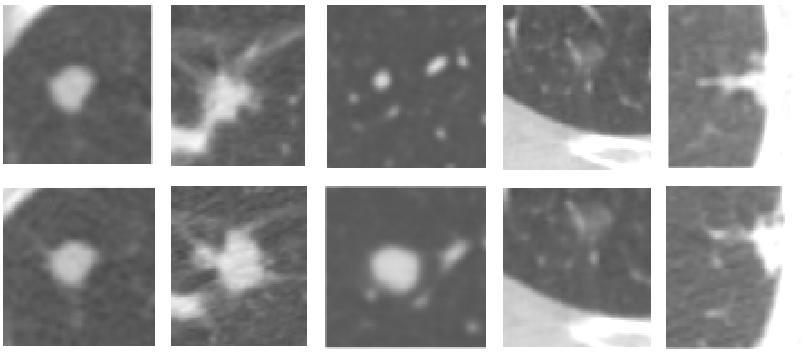
Figure 2.4: On the left, well-delineated solid lung nodule with smooth border. In the centre of the image, a lobulated nodule. On the right, a spiculated nodule. Image taken from [274].

- **Location**: Lung nodules can be located at any place of the lung tissue, although malignant ones are manifested with more frequency in the upper lobes, with a predilection for the right lung [288]. Adenocarcinomas and metastases tend to be in the periphery, while squamous cancers are more often found near the hila. Small solid nodules in a perifissural or subpleural location often represent intrapulmonary lymph nodes.

- **Size**: This feature is one of the most important indicators of malignancy, together with nodule growth. The size of the nodule is commonly described with the major diameter, and it has a significant relation with lung cancer probability [199]. Thus, very small nodules (<5 mm) [310] have low chances to be malignant, whereas larger nodules are more likely to be cancerous [310]. Other indicators, such as the volume of the nodule, have been recently introduced in clinical guidelines [111, 195], from the recent evidence found in the Dutch-Belgian Lung Cancer Screening trial (NELSON)

[123] regarding the major inter-observer agreement with volumetric measures. Radiological guidelines establish different cut-off thresholds in the size and morphology of the nodules for their management.

- **Growth**: Nodule growth, determined by imaging surveillance, allows assessing nodule malignancy [97]. Due to inter-observer variability, nodule growth is identified based on a minimum amount of diameter increase between two medical studies (e.g. more than 2 mm as suggested by Fleischner [195] and British Thoracic [42] Societies), independently of the morphology of the nodule [151, 197] (e.g. solid or part-solid). In general, a very rapid growth or stability in pulmonary nodules suggest a benign aetiology. Hence, [102] determined a malignancy likelihood close to 0 if growth was noted in fewer than 7 days, or stability over a 2-year period. In these terms, a wide range of rates has been reported in literature depending on methods used to measure, histological subtypes and/or radiological appearance of the nodules [42]. For instance, solid cancers generally double in volume (approximately 26% increase in diameter [53]) over between 100 and 400 days, while subsolid cancers (generally representing adenocarcinomas) frequently double in volume over 3 to 5 years. However, the proportion of the nodule that contained the solid component in subsolid nodules has been found to be an important factor of risk [286]. Figure-2.5 shows several nodules with different sizes, growths and malignancy associated.



D1:	5.74 mm	8.29 mm	6.95 mm	10.9 mm	5.66 mm
D2:	5.69 mm	14.4 mm	12.7 mm	11.4 mm	13.3 mm
Growth:	-0.05 mm	6,11 mm	5,75 mm	0.5 mm	7,64 mm
TDiff:	2 years	12 months	22 months	3 years	5 months
Malign:	No	Yes	Yes	Yes	Yes

Figure 2.5: Diameter size (D1 and D2), growth (difference in diameter), time difference and malignancy for 5 different nodules. The top row is time-point 1, and the second row is time-point 2.

Other related risk factors have also been studied, such as the existence of calcification in benign solid nodules, distance to cavity walls or multiplicity of nodules in the same patient.

2.3 Computed Tomography

Pulmonary nodules are detected in patients using common chest image studies such as X-rays and computed tomography (CT) (Figure-2.6). During a CT acquisition, a thin axial section of a patient is imaged by taking large series of two-dimensional X-ray projection images of this section from different directions. Using computer processing, many continuous axial slices can be obtained and then stacked to form a three-dimensional image of the body.

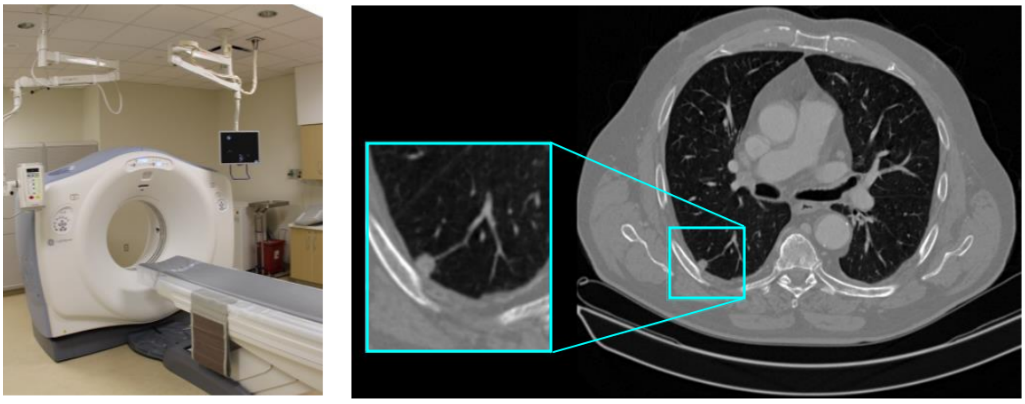


Figure 2.6: Image of a modern CT scanner (left) and an axial slice of a CT scan with a lung nodule (right). Credits: daveynin, CC 2.0.

In a CT scan of the lung, all tissues are visualized according to their absorption of X-rays. The level of absorption in CT scans is measured in Hounsfield Units (HU), which is a standard quantitative scale for describing radio-density in which every tissue has its own HU range (Table-2.1). Therefore, CT scans are calibrated to accurately measure this range of values.

CT scans can be made at different dose levels (i.e. amount of ionizing X-radiation). Typically, less noise is present when using higher amount of radiation. However, current CT scans manage to produce effective images, even with low doses of around 1.5 mSv (low-dose CT) [137]. These doses are still much higher than those produced by a single chest X-ray, which is estimated to have an effective dose of 0.1 mSv.

Thoracic CT is nowadays a common volumetric imaging tool for lung cancer diagnosis [209]. Thanks to it, radiologists can visualize the rich structures that compose the lungs, such as predominantly lung parenchyma, vessels and airways.

Substance	HU
Air	-1000
Lung	-500
Fat	0
Water	15
CSF	30
Kidney	+30 to +45
Blood	+10 to +40
Muscle	+37 to +45
Grey matter	+20 to +45
White matter	+40 to +30
Liver	+40 to +60
Soft tissue, contrast	+100 to +300
Bone	+700 (cancellous bone) to +3000 (cortical bone)

Table 2.1: Mapping of HU values with substance type.

Their utility for lung cancer diagnosis was scientifically proved by the results of the National lung cancer screening trial (NLST) in which a clear survival benefit (reduced mortality rates by 20%) for low-dose CT in current and former smokers was reported over patients diagnosed using radiographic studies [298].

2.4 Managing pulmonary nodules

In general, to assess nodule malignancy and to prescribe the most appropriate management, radiologists consider the clinical history of the patient, current imaging features, and previous imaging studies [195].

There are different options for managing lung nodules, such as not taking any further action, CT surveillance in intervals determined by nodule size and clinical risk, further imaging investigation with a PET/CT scan, further invasive investigation with non-surgical biopsies (e.g. CT-guided fine-needle biopsies), and/or concurrent definitive histological diagnosis and treatment through surgical excision (normally lobectomy or exceptionally sublobar excisions) [190].

Imaging is a key piece of information to assist in this decision-making. To support radiologists in this crucial and complex task, several clinical guidelines have been defined. Fleischner [195] provides recommendations for the follow-up and management of indeterminate pulmonary nodules detected incidentally on CT scans. This guideline does not apply to lung cancer screening, patients younger than 35 years, or patients with a history of primary cancer or immunosuppression.

Similarly, Lung-RADS guidelines [12], defined by the American College of Radiology, proposed a classification to aid with findings in low-dose CT screening exams for lung cancer. The goal of Lung-RADS is to standardize the follow-up and management decisions, but for the subset of patients intended for low-dose screening studies.

2.5 Challenges and limitations

Low-dose CT screening has been demonstrated to be an effective method for the early detection of lung cancer [298]. However, this method by itself is far from being perfect. Radiologists are forced to process large volumes of CT slices, usually with a low signal-to-noise ratio, which causes erroneous classifications of weak irregular limits or normal tissues. In addition to this, lung cancer diagnosis through CT is often subjective and highly affected by factors such as the fatigue and emotions of the observer [227], leading to inconsistent results from the same radiologist at different times or from different radiologists examining the same CT image.

Clinical guidelines have been designed to support clinicians in determining lung nodule malignancy and selecting the best management options [12, 195]. However, most of these recommendations are rather weak, relying on low-quality evidence, and followed by a minority of clinicians (approximately 40%) [294]. Therefore, the management of most patients presenting incidental lung nodules seems to largely rely on subjective clinical judgment.

The evidence suggests that more accurate, robust and reliable assessments are required. Therefore, there is a need to advance with the research, development and application of more effective and efficient methods and technologies that allow the detection, follow-up and diagnosis of lung cancer to assist in clinical decision-making and reduce the burden in radiological health units.

Chapter 3

METHODOLOGICAL BACKGROUND

This chapter reviews the methodological background of this thesis. First, we provide the fundamentals of deep learning by introducing the essential methods and the most successful network architectures for image analysis. Second, we introduce the common processes involved in lung cancer assessment using conventional and deep learning image analysis. Finally, we detail principal challenges and limitations encountered.

3.1 Deep learning for image analysis

Deep learning is a data-driven approach that allows to automatically discovering and learning multiple-levels of representations directly from the training data. To do this, deep neural networks are organized as the composition of simple but non-linear layers (or modules), where each layer transforms the input representation (starting from raw data) into a higher-level, slightly more abstract representation. More formally, the outputs of a layer l (the activation maps $A^{<l>}$) are obtained through the linear combination of the inputs (outputs of layer l-1, i.e. $A^{<l-1>}$), and then a non-linear activation function $g^{<l>}$:

$$A^{<l>} = g^{<l>}(W^{<l>}A^{<l-1>} + b^{<l>}),$$

where $W^{<l>}$ and $b^{<l>}$ are the weights and the biases of the layer l respectively.

With the composition of enough layers, complicated functions can be learned [167] to recognize complex objects in the images. To learn or training such networks, back propagation mechanisms are used, such as stochastic gradient descent optimization algorithms [250], to minimize the weights ($W = W_{\forall l}^{<l>}$ and $b = b_{\forall l}^{<l>}$) of the different layers based on the difference (or loss) between the predicted $\hat{y}$ outputs and the target y values. This process is known as empirical risk minimization and the cost function J is usually defined as the average of the errors committed on each training instance:

$$\mathrm{J}(W, b) = \frac{1}{m} \sum_{x_i} \mathcal{L}(\hat{y}_i, y_i),$$

where $\mathcal{L}$ is the loss function, which for a multi-class problem is usually expressed as the cross-entropy loss:

$$\mathcal{L}(\hat{y}_i, y_i) = -log(p(y_i|x_i)),$$

where $p(y_i|x_i)$ represents the probability of predicting the true class y_i of instance i computed by the last layer of the network, $A^{<L>}$.

A basic form of deep neural networks are the fully connected (FC) networks. An FC network consists of a stack of FC layers. An FC layer is a function from $\mathbb{R}^m$ to $\mathbb{R}^n$, and each neuron in one layer is connected to all neurons in the next layer. Unfortunately, the full connectivity of FC networks make them computationally expensive, and prone to overfitting for image data.

Much more specialized, and efficient, than an FC networks are the convolutional neural networks (CNN) [168]. CNNs are designed to process data in the form of multiple arrays (i.e. 1D signals, 2D images or 3D video/volumetric images). They can extract, without human intervention, accurate feature image representations thanks to their shared-weights architecture [167]. Also, this type of architectures allows capturing other basic image properties, such as high correlation among local values and translation invariance [54]. The main type of layer in CNNs is the convolutional layer. These layers are organized in feature unit maps, in which each unit is connected to local patches in the feature maps of the previous layer through a set of weights called filters. Particularly, to obtain the feature maps of a convolutional layer, the filters slide over (according to some initial stride and padding values) the input feature map applying, on each step, a convolution operation (e.g. sum of element-wise multiplications). The convolution operation is typically denoted with an asterisk:

$$s = (x * w),$$

where w are the weights (or kernel), x is the input and s the resulting vector (or feature map). Other important type of layers in CNNs are the Pooling layers. These layers aim to merge similar features, reducing the dimension and creating invariance to small shifts and distortions.

The advantages of CNN in terms of performance and efficiency for image analysis have made them one of the preferred solutions used today to solve most image analysis problems [144]. In the following subsections, we tackle some of the most well-known network architectures and proposed solutions, making special emphasis on those based on CNNs.

3.1.1 Image classification

One of the earliest CNNs for image classification was LeNet-5 [168]. Applied originally on the hand-written digit recognition problem (MNIST dataset [LeCun and Cortes]), this network is structured basically by a stack of convolutional and pooling layers with an FC layer at the head of the network, as schematized in Figure-3.1.

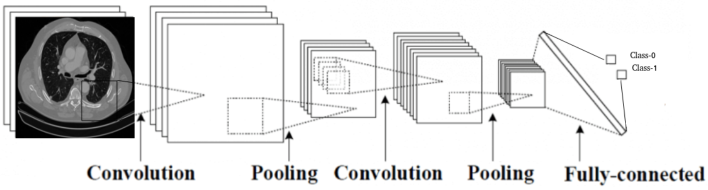


Figure 3.1: Vanilla architecture of a CNN.

Despite the success of LeNet-5, the first CNN that really supposed a major breakthrough in the computer vision field was AlexNet [158]. This network won the 2012-ILSVRC [63], a yearly international competition where researchers submit the results of their classification algorithms trained on ImageNet, a large scale annotated dataset of images. AlexNet unseat previous methods (i.e. based on traditional machine learning algorithms) reducing the classification error by more than 10%. This CNN incorporated three novel ideas in its architecture and in the way how it was trained. First, it used Rectified Linear Units (ReLU) instead of previously used activation units (e.g. hyperbolic tangent, sigmoid) to solve the vanishing gradient problem during training. Second, it used heavy data augmentation techniques by applying label-preserving transformations (e.g. mirroring, rotation, cropping) to make the training data more varied. Third, it used dropout layers to turn off neurons with a predetermined probability to reduce overfitting [279]. Another relevant deep network was VGG [272], which obtained 2nd place in 2014-ILSVRC. This network proposed using multiple small filters without pooling stacked together. This allowed increasing the representational depth of the networks while limiting the number of parameters. Also in VGG, 1x1 convolutions were used in between the convolutional layers to regulate complexity and to learn a linear combination of the resultant feature maps. Moreover, in this network max-pooling was placed after the convolutional layer, while padding was performed to maintain the spatial resolution. GoogLeNet [290] was another relevant architecture. This network defined inception blocks (composed by parallel convolution layers with different sizes) to leverage feature detection at different scales and dimensionality reduction (Figure-3.2). Other prominent contributions were presented in ResNet network [114]. This network posits the use of batch normalization[132] to reduce internal co-variance shift during training, the replacement of fully connected layers by convolutions thanks to one by one convolution [179], and the creation of residual blocks (Figure-3.3). These blocks allowed adding further (convolutional) layers in the network without performance damage, as by default the identity function was learnt. In case the filters could learn any new information, it is subtracted or added to the base representation.

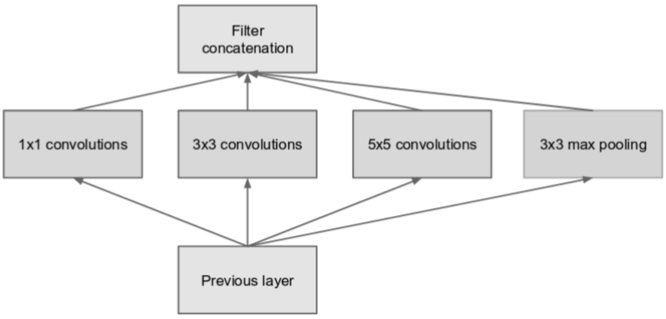


Figure 3.2: Overview of an Inception network block in GoogLeNet (Image taken from [290]).

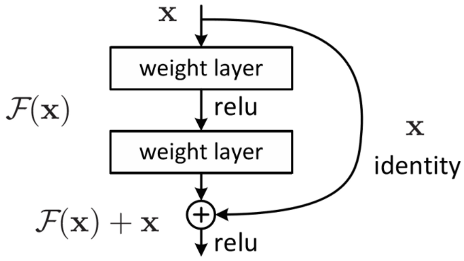


Figure 3.3: Residual block by ResNet networks (Image taken from [114]).

Beyond the aforementioned networks, new methods and architectures continue to be developed [144], although in recent years the performance increase seems to have reached a standstill, based on the scores obtained in public benchmark rankings (currently Top-1 accuracy[1] is already above 86% without extra training data). Nonetheless, recent architectures can be highlighted such as the inception-ResNet network [289], which combines the power of residual learning and inception blocks, the DenseNet [128] network, which defines direct con-

---

[1] https://paperswithcode.com/sota/image-classification-on-imagenet

nections between any two layers with the same feature-map size, or the SENet [125], which developed the SE (Squeeze-Excitation) block to adaptively recalibrate channel-wise feature responses by explicitly modelling interdependencies between channels. Other recent network is NasNet [338] created thanks to an automatic method for exploring new convolutional blocks (or cells) [118]. Also, it uses a new regularization technique called ScheduledDropPath to drop each path in the cell with a probability linearly increased over the course of the training. As a continuation of NasNet, we can also highlight the EfficientNet [293], which is a new family of networks obtained thanks to the recent progress in automatic network architectures search [118], together with a novel scaling method which aims to uniformly scale each dimension (width, depth and resolution) with a fixed set of scaling coefficients.

3.1.2 Object detection

Object detection is also a prolific and active research field in which deep learning has contributed largely during last years, providing top scores on well-known public benchmarks [2, 63]. In object detection, the model is tasked with localizing the objects present in an image, and at the same time, classifying them into different categories.

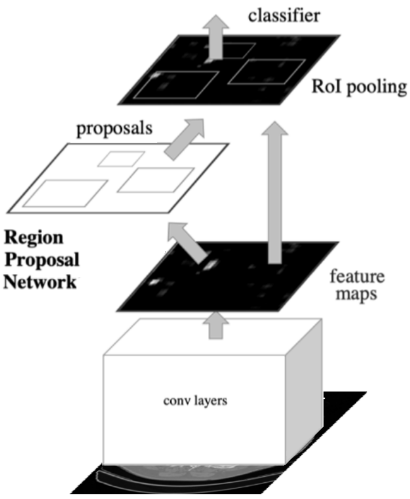


Figure 3.4: Overview of the Faster R-CNN object detection network (Image taken from [243]).

We could broadly differentiate between two types of deep network architectures for object detection. The first of them is a two-step detector based on the use of region proposals or rectangular boxes. One of the most well-known architectures is the Faster-RCNN [243] (which is an evolution of Fast R-CNN [89] and

R-CNN [90] networks). This type of networks (Figure-3.4) are composed by three sub-networks: a feature extraction CNN that reads the initial image and outputs a set of features maps; a region proposal network (RPN) that generates automatically several proposals (using different anchor boxes or predefined fixed windows) in a sliding window fashion on the feature maps; and finally a feature extraction (RoI Pooling layer) with two parallel FC layers to obtain the predicted bounding box coordinates and probability scores for each object class in each bounding box.

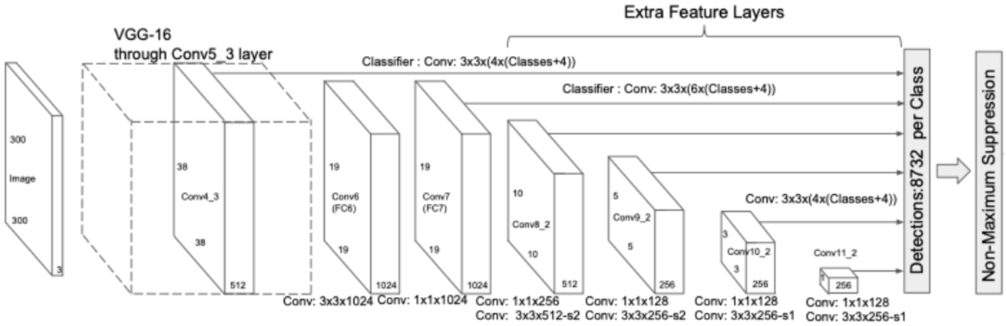


Figure 3.5: Example of a SSD network architecture for object detection (Image taken from [186]).

The second type of networks for object detection are one-stage detectors, in which region proposal and region classification are tackled together in the same CNN. They propose partitioning the input image into a grid of cells and then assigning the centre of the regions to one of the cells, allowing to identify objects with a single convolutional run. Examples of this type of networks are ”You Only Look Once” (YOLO) [240], or ”Single Shot multi box Detector” (SSD) [186]. This last architecture (Figure-3.5) incorporated three remarkable ideas to overcome YOLO (first version) performance limitation: first, it applies different anchor boxes (taken from Faster R-CNN) per each grid cell; second, it uses hard negative mining to prioritize complex cases in the computation of the loss function; third, it aggregates multi-scale features to pick up fine-grained local features while preserving coarse global features.

Beyond these two detectors, we highlight another recent network, RetinaNet [181]. This network basically incorporates two new concepts over previous works. First, the backbone of this network combines a ResNet [115] for deep feature extraction, a feature pyramid network (FPN) [180] to efficiently detect objects at multiple scales, and two task-specific subnetworks for classification and bounding box regression. Second, it defines the focal loss, a new loss function specifically designed to reduce the impact of the large number of ”easy” cases, or proposals without any object, allowing to focus in ”hard” cases during training. Also, we should mention that YOLO network has been evolved into its version 3 [241], where an upgraded architecture is used adding most of the previously mentioned features (i.e. FPN, focal loss, anchors) to offer better performances and same fast detection speed.

3.1.3 Image segmentation

Semantic image segmentation or pixel-wise classification is an essential topic in computer vision. It typically involves clustering together or isolating parts of an image that belong to the same object [187].

Several deep learning methods have been created to address the image segmentation problem [203]. One of the earlier deep segmentation architectures was the fully convolutional network (FCN) [188]. In this approach, all fully-connected layers were replaced by convolutional layers to manage arbitrary sizes of input images and generate a segmentation map of the same size. This approach presented different limitations, such as object localization problems. New techniques, like Conditional Random Fields [46], were added on top of it to improve initial performance limitations.

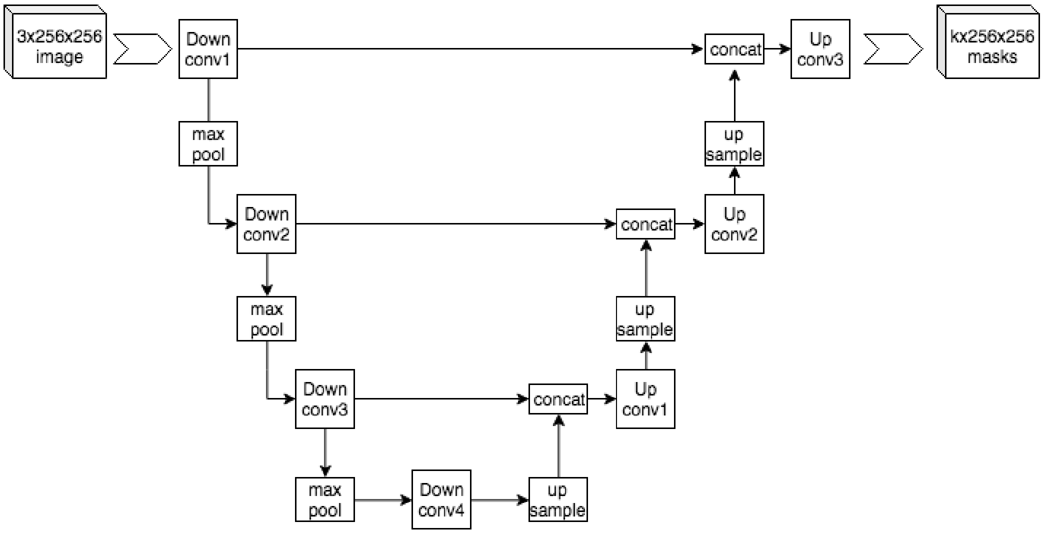


Figure 3.6: Example of a U-Net network architecture. Credit Mehrdad Yazdani CC 4.0.

As an alternative to the previous segmentation approach, an encoder-decoder backbone was proposed in [215]. In this network, the encoder part takes the input image and pass it through a set of convolutional layers (usually following the VGG architecture) to obtain a smaller feature vector, and the decoder part uses deconvolutions and upsampling layers to convert the feature vector into a map of pixel-wise class probabilities. Some networks, such as SegNet [25] and its probabilistic version BayesSegNet [142], evolved this architecture introducing the concept of passing information from the encoders (i.e. the max pooling indices) into the corresponding upsample layers of the decoders. However, the most well-known network using this approach is the U-Net [248]. This network uses a convolutional encoder and decoder backbone tied at different levels by shortcuts, which allow by-passing high level features of the encoder to the decoder, in order to enhance the image reconstruction task, diminished by the flow of data through the convolutional and pooling layers of the architecture (necessary at the same time, to improve the generalizability of the network). Several extensions of this architecture can be found in the literature, such as its 3D formulation [50] or the incorporation of ResNet-like blocks and a Dice-based loss [202]. Also, recent works have proposed integrating in the U-Net architecture recent mechanisms originally created for other data type problems (i.e. natural language processing), such as attention gates [217], to automatically learn to focus on target structures of varying shapes and sizes, or recurrent layers [10] to accumulate features and ensure better feature representation for segmentation tasks.

Another different alternative for image segmentation is represented by the Mask-RCNN [113]. This network is based on the Faster-RCNN architecture in which a new branch is added at the head for predicting class-specific object masks, in parallel with the existing bounding box regressor and the object classifier branches.

3.1.4 Image generation

Image generation with deep learning is one of the most challenging but more actively research areas in computer vision.

One of the most popular deep learning models to generate new images are generative adversarial networks (GAN) [96]. This framework consists of two networks, the generator and the discriminator, that compete with each other in a zero-sum game where the generator aims to increase the error rate of the discriminator network. Thus, the generator learns to map points from a latent space, usually sampled from a multivariate standard normal distribution, into observations that look as if they were sampled from the original dataset. The discriminator tries to predict whether an observation comes from the original dataset.

Another well-known approach for addressing image generation is deep auto-encoders (AE). This framework uses an encoder, which embeds the input into a representation vector, and a decoder, which projects the vector back to the original manifold. The representation vector is a compression of the original image into a lower dimensional, latent space. The idea is that, by choosing any point in a latent space, a novel image is generated by passing this point through a decoder (as it learned to convert points, or representations, in a latent space into viable images). Therefore, the learning process of this network consists on minimizing the reconstruction error, which is the error between the original image and the reconstruction from its representation. Since auto-encoders do not force continuity in space, images are poorly generated at sampling time.

One successful extension from auto-encoders are variational auto-encoders (VAE) [148, 246]. In particular, the encoder retrieves two vectors, the mean and log-variance vectors, which together define a multivariate distribution in the latent space. When a random point is sampled from this distribution, the decoder produces a similar image, guaranteeing the continuity in the latent space. The way to achieve this, is by making the output distribution of the encoder as close as possible to a standard multivariate normal distribution using the Kullback-Leibler divergence (KL) loss. Thus, the total loss function of the VAE is composed by the sum of the KL-divergence loss and the reconstruction loss. A variant of VAEs was created to generate multiple outputs from a single input. Precisely, conditional variational auto-encoders (CVAE) [275] were proposed to model the distribution of a high dimensional space as a generative model conditioned on the input. Therefore, the prior on the latent variable is conditioned by the input.

3.1.5 Temporal image analysis

Many computer vision problems, such as activity recognition, change detection, object tracking, require the analysis of temporal sequences of images. The emergence of deep neural networks have overcome results from state-of-art conventional methods as shown in public large scale datasets (e.g. UCF101 [278]) developed for this type of tasks. In these terms, one of the common deep architectures used are the Siamese neural networks (SNN) [38]. SNNs are designed as two sibling networks, connected by a distance layer at the top, trained to predict matching or mismatching between two input images. To achieve this, SNNs are usually trained using the contrastive loss more suitable for learning to differentiate a pair of instances. This loss function comes described as follows:

$$\mathcal{L}_{contrastive} = yD_w^2 + (1-y)(max(0, m - D_w))^2$$

where y is the binary label, m is a margin at which dissimilar paired inputs will not contribute further to the loss and D_w is a distance function (e.g. L2) between the two embedding vectors resulting from the sibling networks (i.e. $f(A)$ and $f(B)$). The original Siamese architecture, first introduced for the problem of signature verification, was extended by [152] using convolutional layers and adjusting the optimization metric with a weighted L1 distance between the twin feature vectors of both networks. SNNs have been extensively used in computer vision matching problems such as tracking objects in videos [296], matching pedestrians across multiple camera views [307], and matching corresponding patches in satellite images [131].

Other typical CNN architectures for image sequence analysis was defined in [271], in which two separate CNNs for recognition of spatial and temporal features were combined by late fusion. The spatial stream performed action recognition from still video frames, whereas the temporal stream was trained to recognize action from motion in the form of dense optical flow. Decoupling the spatial and temporal nets allowed using a pre-trained spatial net on the ImageNet [63]. Further extensions [43, 302] of this approach propose the use of 3D convolution filters by either using a single or two seamless 3D CNN networks. Alternatively, in [79] proposed a model based on frequency domain representation for predicting object movement in the video. One of the recent methods in modelling temporal data is temporal convolution networks (TCN) [166]. The critical advantage of TCN is the representation gained by applying the hierarchy of dilated causal convolution layers on the temporal domain, which successfully capture long-range dependencies.

Beyond the use of CNN architectures for the analysis of image time series, we can also find the use of recurrent neural networks (RNN) [252]. They are leading methods applied to longitudinal data, such as natural language [315]. RNN introduces the concept of state or memory of the network. The state of a network is updated with each image of the sequence during the training stage, and it is used to generate the output of the RNN. The main component of an RNN is the cell, which is applied to each image of the sequence. There may be multiple cells in an RNN. A cell receives as input both the current image of the sequence and the previous state of the network (initially, a zero matrix or null state, $h^{<0>}$) and it retrieves the following state (h^1). An RNN cell combines the current state and the image to generate a new state. This happens as following:

$$h^{<t>} = g(W_{rec}h^{<t-1>} + W_{input}x^{<t>} + b)$$

where b is the bias term, W_{rec} the recurrent weight matrix, W_{input} the input weight, x^t the input image, h^{t-1} the current state, h^t the new state and g an element-wise non-linearity (e.g. hyperbolic tangent). The final hidden state is eventually used in combination with a weight matrix V to compute the final prediction ($\hat{y}$).

$$\hat{y}^{<t>} = g(Vh^{<t>})$$

A significant limitation of RNN models is known as the ”vanishing gradient” problem, i.e. the impossibility to back propagate the loss value through a long-range temporal interval. To overcome this limitation, the Long Short-Term Memory (LSTM) networks [121] were designed for the next time-step status prediction in a temporal sequence capable of learning long-term dependencies. In particular, LSTM networks incorporate, within a layer or cell (A), the concept of gates. Up to 4 different gates were originally proposed in an LSTM: The input gate $\mathrm{i_t}$ designed to control the information to be stored, the forget gate $\mathrm{f_t}$ created for controlling the information to be forgotten, the cell state gate $\mathrm{c_t}$ to control what new information is going to be stored and the output gate $\mathrm{o_t}$ to decide what information is going to output the network at step $\mathrm{T_t}$.

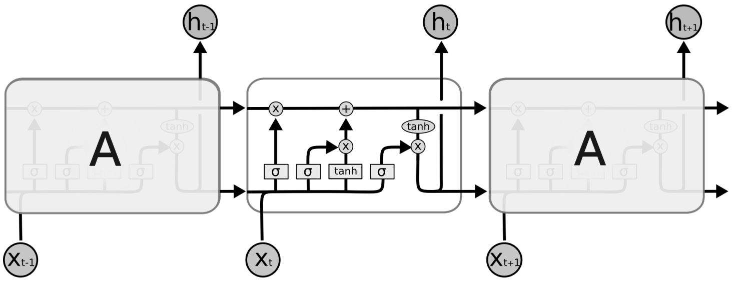


Figure 3.7: Example of an LSTM network architecture. Image taken from [218].

RNNs are usually combined with convolutional layers to learn compositional representations in space and time. Thus, at first stage CNN layers extracts features from the raw data and generates high-level representations, then at second stage, recurrent layers uses the features yielded by the CNN layers to learn time dependencies ([71]). More recent works have proposed combining both type of layers (i.e. CNN and RNN) into a new type, the ConvLSTM [267], specially suitable for learning spatio-temporal features. These layers are recurrent components that compound convolutions to determine the future state of the cell based on its local neighbours instead of the entire input.

3.2 Automatic lung cancer assessment

Lung cancer assessment using CT scan images is a complex and tedious work with large inter-observer variability. Clinical studies quantified manual lung nodule detection sensitivities close to 80% with an average of 1 false-positives per study, and inter-observer agreements below 34% [20, 119]. These performances make lung cancer as the third most frequently missed diagnosis based on expert readers' visual assessment, as corroborated in [255].

To support radiologists in this task, conventional image analysis and machine learning algorithms have been extensively studied for automatic pulmonary nodule assessment [80, 208, 303, 304]. However, the resulting predictive models, built from hand-crafted features on top of the images, turned out to be not effective enough to fully capture the complexity and variability of pulmonary nodules [124].

Deep learning emerged as a step forward over conventional methods thanks to, among others, its ability to automatically extract intricate feature representations directly from the data [224]. However, the application of this technology in the lung cancer domain was circumstantial until the recent release of the LIDC-IDRI dataset [16], the largest public annotated cohort of CT scan images, and the creation of two open medical image challenges, the LUng Nodule Analysis [4, 262] and the Data Science Bowl competition [3]. The outstanding scores achieved by methods relying on CNNs, drowning out those using conventional techniques, ended up convincing the scientific community about the advantages of using deep learning for lung cancer assessment, flooding rapidly the medical image analysis research literature [183, 264].

Next, we provide further details regarding the main image analysis datasets and tasks for automatic lung cancer assessment.

3.2.1 Common lung cancer datasets

The most commonly used lung CT datasets for research purposes are described below:

- **LIDC/IDRI**. With a total of 1018 CT scans, the LIDC [18] is the largest publicly available reference database for lung nodules. Each CT scan is associated with a file containing annotations from four experienced thoracic radiologists. The annotations are the result of a two-phase reading process in which the radiologists were asked to mark suspicious lesions and to provide additional characterization of lesions of diameter larger or equal to 3 mm which were marked as a nodule [19]. Additionally, the four radiologists annotated a malignancy rating ranging from 1 (highly unlikely for cancer) to 5 (highly suspicious for cancer) on the nodules $>=$ 3 mm [16].

- **LUNA16**. An updated version of the LIDC dataset was provided in the LUng Nodule Analaysis 2016 challenge [262], which includes only scans with at least one lesion of size $>=$ 3 mm marked as a nodule by at least three of the four radiologists. The LUNA16 dataset consists of 888 CT scans comprising a total of 1186 nodules. Annotations with coordinates of each nodule in the three spatial axes inferred from the original LIDC annotations are also provided.

- **TCIA**. For 157 cases the LIDC dataset provides diagnostic data at patient level obtained from biopsy, surgical resection, progression or reviewing of the radiological images showing stable nodules after two years [52].

- **DSB17**. From mid-January till early April 2017 the data mining platform Kaggle launched a global challenge (Data Science Bowl [3]) to build accurate methods able to determine probability of a case to be diagnosed with lung cancer. For this, a labelled dataset was made available with 2001 patients (1397, 198, 506 cases in its training, validation and test set respectively). The DSB dataset only includes per-subject binary labels indicating whether a subject was diagnosed with lung cancer in the year after the scanning. Note that this dataset does not provide information about nodules in the CT scans. For each patient, the CT scan data consists of a variable number of images (typically around 100-400 axial slices) of 512 x 52 pixels. The slices are provided in DICOM format. Around 70% of the provided labels in this dataset are negative cases.

- **NLST**. Launched in 2002, the initial findings were released in November 2010. The National Lung Screening Trial dataset [297] enrolled 53,454 current or former heavy smokers with ages between 55 and 74. Participants were required to have a smoking history of at least 30 pack-years and were either current or former smokers without signs, symptoms, or history of lung cancer. NLST was conducted by the American College of Radiology Imaging Network, a medical imaging research network focused on the conduct of multicentre imaging clinical trials, and the Lung Screening Study group, which was initially established by NCI to examine the feasibility of NLST. The total amount of data available (under contract agreement) is formed by 15,000 participants distributed in 622 participants with screen-detected lung cancer, 419 participants with non-screen-detected lung cancer (false negatives or post-screening cancers), 8,205 participants with at least 1 nodule detected on any screens, 5,754 participants with no lung cancer and no nodules.

3.2.2 Lung preprocessing

Prior to any analysis of lung CT images, it is necessary to carry out a series of image processing techniques to be able to successfully perform subsequent analyses on them. One of the first processes consists on converting the pixel values of lung CT images into Hounsfield Units (HU) [324]. After that, pixel image intensities (in HU) are typically masked or clipped, in order to be consistent with that of lung tissues. Additionally, pixel values are normalized, usually adjusting their values between 0 and 1 [159]. Moreover, due to the dataset contains CT scan images acquired at different resolutions, a re-sampling mechanism is performed on each CT to a fixed resolution (e.g. isotropic at 1x1x1 mm) in order to reduce the variance given by the different pixel size/coarseness (e.g. the distance between slices) of the scans.

More elaborated techniques are also usually conducted to attenuate the effect of the multiple structures that exist in the lungs and highlight regions of interest to avoid confusion with parenchyma, such as linear interpolation [60], median filter [314], morphological top-hat transformation [325] or Gaussian filters [175].

The lung parenchyma segmentation is another common preprocessing task in which the pulmonary tissue is selected from the CT slices so that the subsequent detection stages can operate in optimal conditions [62]. In particular, different strategies have been proposed such as knowledge-based techniques [225], threshold methods [17], region growing [33], mathematical morphology [40], active contour/shape models [45] or cluster analysis methods [101].

3.2.3 Nodule detection

This task aims to examine the entire lung CT volume, searching for small suspicious regions or nodules (usually between 3 mm to 30 mm) [270].

I) Conventional approaches

To support clinicians, several conventional image analysis techniques (either in 2D and 3D data) have been developed, as shown in [80, 303]. One of the widely used techniques is blob detectors, such as the Laplacian of Gaussian filter (LoG), which detect edges or regions of rapid intensity change by approximating the second derivative measurement on the image [86]. Another interesting technique is descriptors of histogram of oriented gradients (HOG) [59] in which the distribution (i.e. histograms) of directions of gradients are used as features. Gradients (x and y derivatives) of an image are used because they are larger around edges and corners than in flat regions. Alternatively, local binary patterns [263] and wavelet feature descriptors [221] are also two common techniques for extracting relevant features from CT scans. Other broader techniques have been used for extracting candidates such as curvature computation, voxel clustering, intensity thresholding or morphological operations [135, 208]. Although these techniques offer adequate sensitivity scores, they produce too many false positives. For reducing the false positive rate, low-level descriptors have been carefully and heuristically defined. The literature is plenty of studies that uses multitude of features to characterize the regions [135, 208, 260]. Some features aim to quantify the morphology of the regions such as size, curvedness, length of the axis. Others aim to measure the texture of the region e.g. mean grey intensity, entropy or uniformity, which give further insight into the distribution of tissue attenuation information lost when averaging intensity over a large area. On top of these features, classical machine learning algorithms have been built to discriminate among candidates or real nodules (e.g. Logistic Regression [56], Support Vector Machines [55], Random Forest [120]).

II) Two-stage detectors

Recently, deep learning has extensively tackled the problem of lung nodule detection by offering a multitude of diverse solutions [104, 247, 299]. Some of them address this task in a two-stage process consisting of, first, achieving high sensitivities (finding the large number of potential nodule candidates) and then reducing its false positive rate. For the first task, earlier works explored alternative deep networks such as autoencoders [162], or deep belief networks [127]. However, this task is currently tackled using CNN as they have been shown easier to be trained. For instance in [305], the authors proposed using transfer learning from an earlier 2D object detection CNN, named OverFeat [259], trained for object detection in natural images. Thus, from the CT scan they extract 2D sagittal, coronal and axial patches for each nodule candidate. In a posterior work [68], a 2D Faster R-CNN [244] modified with a deconvolutional layer for candidate detection on axial slices was used. In a more recent study [323], a 3D version of the U-Net architecture, named V-Net [202], was adapted for the detection of nodule candidates.

Regarding the false positive stage, some works use 2D CNNs on patch images of the nodule candidates. For instance, in [261], a 2D multiple view approach was proposed for analysing in parallel different image planes using shallow 2D CNNs with a late fusing feature mechanism. However, 3D-CNN approaches are currently used as they work more efficiently because of the 3D nature of the lung nodules [73, 323].

III) One-stage detectors

Alternatively, other studies address the lung nodule detection in an end-to-end fashion. One of the solutions [92] proposes a 3D CNN to detect lung nodules in sub-volumes of CT images. In [334], authors used 3D RPN based on a 3D Faster R-CNN with a U-net-like encoder-decoder structure for nodule detection to capture spatial image representations with high discrimination capabilities. Another approach [212] also used 3D Faster R-CNN-like scheme for directly identifying lung nodules, but using a U-Net-like architecture improved with the advantages of ResNet, DenseNet, and Dual path networks [334]. Alternatively to RPN, in [159], inspired by the class probability map of YOLO network [240], the search space was divided into a uniform grid to perform detection in each grid cell. For this, they used a 3D ResNet network [115]. In another work [An2], which ranked first at the nodule detection LUNA16 Challenge [262], authors adopted a 3D feature pyramid network (FPN) [180].

On a recent survey [6] made over 56 different studies with 74 separate patient cohorts, they reported a pooled specificity of 0.89 (95% CI 0.87-0.92) and a sensitivity of 0.86 (95% CI 0.83-0.89) and an AUC of 0.93 (95% CI 0.92-0.94) for diagnosing lung nodules on CT scans.

3.2.4 Nodule segmentation

Having the masks of the suspicious regions or nodules of the lung, allows applying further post-processes such as quantifying their diameter, area or volume, which in fact are fundamental information for estimating their malignancy. However, automatically segmenting or providing a pixel wise classification of the nodules is a challenging task due to the existing ambiguity (e.g. poor image resolution) in the lung CT images.

I) Conventional approaches

As in lung segmentation, conventional image analysis methods have been applied to segment the nodule from the lung parenchyma [21]. To do this, traditionally two categories of methods were used [176]: Region-based and edge-based. The former uses similarity or homogeneity between pixels. To this category belong methods such as thresholding, which was applied on the volumetric lung region in [15], or region growing, which was used in [303] where a seed point was given, and it had to be decided whether surrounding pixels could belong to the growing region. The latter category relies on the detection of contours in the image, assuming that different objects are separated by these type of structures. Some popular edge detection algorithms are based on differential operator, including Sobel and Laplacian [283]. More advanced methods rely on morphological criteria, and they were used in [160] for segmenting complex nodules with attached structures of similar attenuation values (e.g. vessels, airways, and pleura). Alternatively, temporal image subtraction [5, 13] was proposed as a technique to increase radiologist sensitivity in detecting nodule changes.

II) U-Net based networks

Deep learning has also successfully addressed nodule segmentation, outperforming limitations of prior conventional approaches. Most of these works are built upon the U-Net [248] segmentation network. For instance in [333], a 2D U-Net was modified with re-designed skip pathways aiming at reducing the semantic gap between the feature maps of the encoder and decoder sub-networks, demonstrating an average intersection over union (IoU) gain of 3.9 points over U-Net. A Probabilistic version of the 2D U-Net was proposed in [29, 153, 154] to generate multiple coherent segmentations for pulmonary nodules given different possible ground truths. Also in [331] a patch-based 3D U-Net was used for nodule segmentation from raw CT scans. The network was trained in an adversarial way, using as generator the 3D U-Net and as discriminator a 3D Inception with residual convolutional blocks. The Inception network incorporated additional context surrounding the nodules, allowing a larger receptive field for better classification. In [133], an automatic configurable U-Net named nn-Unet was defined in order to adapt the architecture and tuning parameters to the particular type of images to be segmented. This approach reached the top ranking score in a recent medical image segmentation challenge consisting of performing 10 different semantic segmentation tasks [273].

III) Alternative segmentation networks

In [312], authors proposed a network with 2 branches which combined multi-view 3D features and 2D local texture features simultaneously. Also, this approach provides multi-scale feature extraction and a novel central pooling to select features according to their spatial relevance to the target voxel. This method reported a Dice score of 0.82 for LIDC dataset, outperforming previous conventional segmentation methods [200]. Also in [185] a 2D Mask R-CNN model trained on the COCO data set was fine-tuned to segment pulmonary nodules. The model was tested on the LIDC-IDRI dataset. In [156] a 3D version was presented reporting detection and segmentation at same time with competitive results for lung nodule detection (0.936 sensitivity at 7 FP) and segmentation (70% of Dice) on LUNA16 data set.

3.2.5 Malignancy classification

Given the location of a lung nodule on a CT image of a patient, another important task is to automatically determine its possible malignancy or likelihood of being lung cancer.

I) Conventional approaches

Until recently, nodule classification and characterization in CT relied mostly on conventional machine learning algorithms. One of the first studies to estimate malignancy was [198] which created a predictive model from a relatively small balanced dataset of 31 malignant and benignant nodules. To build this type of models, typically hundreds of 2D and 3D features were extracted for each nodule (such as size, density, shape or texture). This task was so common for any medical imaging problem that it became an important research field by itself, known as radiomics [170]. One of the major challenges of this field is how to integrate radiomic feature descriptors with clinical, pathological, and genomic data to decode the different types of tissue biology [157]. Once the nodule features are computed, then different types of machine learning classifiers (such as logistic regression or support vector machines) are trained to ultimately provide the malignancy class and posterior probabilities estimates. A review of different works using this type of approach for chest and colon with CT scans can be found at [284]. Additionally, in the medical literature, we can find different statistical tools for inferring the probability of malignancy for a nodule. To do this, typically a logistic regression is fitted on a small data set, using sensible imaging features (such as nodule size, type, location, spiculation) and clinical information from the patient. One of the best known tools is the PanCan model [199]

II) Networks using radiological labels

Outstanding performances have been reported for lung nodule malignancy classification using deep neural networks trained on physicians annotations [313]. A frequent referenced dataset for this task is the LIDC/IDRI dataset, which contains a large set of CT scans ($>$1000) with nodule malignancy annotations (ranging from 1 to 5) performed by up to 4 radiologists. Several approaches can be found in the literature for nodule malignancy classification. For instance, in [266] authors handle pulmonary nodule classification by utilizing a multi-scale 2D CNN in which three images at different scales (resampled them to a fixed size) are input into the network. In [184, 213], a multi-view 2D CNN (to mimic 3D image volumes) was built using several 2D planes of the nodules. To avoid lack of information from 2D approaches, in [66], 3D CNNs were built and compared against 2D and 2D multi-view networks for nodule malignancy classification. The best reported method was a 3D DenseNet network [128] using a multi-output strategy, consisting of merging last layer features with earlier layer outputs. An alternative solution was proposed in [44], where a 3D CNN was built and merged together with radiological quantitative features to obtain higher performances.

III) Networks using confirmed labels

Alternatively, other deep learning works focused on predicting malignancy using diagnosed labels (or confirmed nodule malignancy). These labels are usually more difficult to collect since they have to be validated (usually under biopsy or after a close follow-up along more than two years) limiting the size of the cohort. Different studies have addressed this approximation, for instance, in [214], authors built a 2D multi-view CNN with a VGG-16 [272] backbone (pre-trained from ImageNet) to distinguish between benign and cancer (primary or metastasic) nodules. The inputs of this network were stacked on 3 images of same size corresponding to 3 orthogonal planes (e.g. axial, sagittal, coronal) from the centre of the nodule. Best model reported an accuracy of 68%. In [8] a shallow 3D CNN was built using data from DSB17 challenge. Results from testing this network in an independent subset (400 cases with almost 70% non-negative cases) were 83% of AUC and 86% of accuracy. In [326], a 3D RPN with a U-Net like backbone was presented to identify nodules with different sizes, in which a second branch was attached (containing two FC layers) to provide nodule malignancy classification. Interestingly, this network was trained in a two-stage process, first, it learned to detect the nodules, and then to predict the nodule malignancy reusing the weights learned from the first stage. Results from this network reported an 85% of AUC in TCIA (subset of LIDC with confirmed cases) and in a test set of 50 cases (25% malign) of the DSB17 an accuracy of 92%. Similarly, in [178], a modified 3D RPN with a U-Net backbone was shown to detect nodules. The feature map of the top-5 candidates were fed into a leaky noisy-or model [228], a causal probabilistic model to infer cancer probability from multiple nodule candidates assuming a leakage probability even when none of the nodules of a patient were predicted as malignant. This is specially appropriated for the DSB17 dataset since the labels are at the patient level (benign or cancer), rather than at the nodule level. This work won the DSB17 challenge, reporting a cancer performance of 0.87% of AUC.

3.2.6 Temporal nodule analysis

In the current clinical practice, a closer follow-up on the temporal evolution of the tumours is required to determine its cancer probability. To do this, radiologists mainly need to detect, match and quantify, each of the suspicious nodules from the different CT scan images taken from the patients (at different time-points) before to assess their malignancy. Nowadays, this work is basically manual and relies on the visual understanding and knowledge expertise of physicians. To reduce the amount of work, stress, cost and errors derived from this process, researchers on computer vision and artificial intelligence have tried to build different pipelines (or sequence of processes) to provide automatic support to clinicians. Next, we detail the main tasks addressed by these pipelines and significant work carried out for each of them.

I) Lung CT image alignment

For the correct automatic analysis of the pulmonary nodules from a series of CT scans of the lung, it is previously necessary that these images are adequately aligned spatially. In computer vision, this problem is known as image registration [39]. To perform registration, two images are involved, the moving image $I_M(x)$, which is deformed to fit the other image, the fixed image $I_F(x)$. Thus, registration is the problem of finding a displacement $u(x)$ that makes $I_M(x + u(x))$

spatially aligned to $I_F(x)$. An equivalent formulation is to say that registration is the problem of finding:

$$T(x) = x + u(x)$$

that makes $I_M(T(x))$ spatially aligned to $I_F(x)$. Good reviews on the subject are given in [196, 204].

In the context of temporal lung nodule analysis, the prior and follow-up lung CT exams have to be spatially aligned to facilitate, for instance, the correct matching of pulmonary nodules. Several factors may compromise the effectiveness of the lung CT registration process, such as the variability in the images size and resolution, and the variability in the position and breath cycle of the patient when performing the scanning [206]. Also, to obtain a good registration, the selection of the right transformation method and an appropriated evaluation metric are imperative for this task.

There are numerous conventional methods for registering medical images in the literature [276]. In particular, in [207], we can find a comprehensive evaluation and comparison of more than 20 algorithms on 30 thoracic CT pairs from the EMPIRE10, a pulmonary image registration challenge. From this work, top-5 algorithms used different non-rigid transformations. Also, in [295] we can find the evaluation of a commercial system for lung nodule matching applying a registration mechanism. The performance obtained was 92.7% of accuracy on three serial CT scans from 40 subjects with 143 nodules from the NLST. In another study [155], the automatic lung nodule matching ability was evaluated using another commercially available system using a conventional registration method. The performances obtained were between 79% and 92% of accuracy using annotations from 4 experts in 57 patients.

Despite the relevant results reported by some of these conventional methods, further demand for faster registration methods motivated the development of deep learning approaches for medical image registration [112]. For instance, in [47] a stacked denoising autoencoder was used to learn a similarity metric for assessing the quality of the rigid alignment of CT and MR images. This metric outperformed conventional ones, such as local cross correlation. In [77], a 3D CNN was designed to perform the deformable registration of inhale–exhale 3D lung CT image volumes. In [280], another CNN was used to both linearly and locally register inhale–exhale pairs of lung volumes. Both, the affine transformation and the deformation were jointly estimated, and the loss function used was composed of MSE and a regularization term. This method outperformed several state-of-the-art methods that do not use ground truth data, including Demons [189] and SyN [22].

Although satisfactory advances have been produced in lung CT alignment, concerns regarding time latency (e.g. 5 minutes according to [251] per case), and distortions introduced in the intrinsic structure of the lung images, still hinders their wide acceptance in the clinical practice [309].

II) Nodule malignancy classification

Despite the medical importance of monitoring the evolution of pulmonary nodules for determining its malignancy likelihood, few works have really taken into account the temporal dimension to provide a malignancy estimate. In addition, most of these works rely on a subset of the NLST (accessible under prior committee agreement), which is probably the largest longitudinal lung cancer dataset. However, cases from this cohort are constrained to certain parameters (e.g. yearly CT scans on a subset of the population), which limits the complexity of the data to analyse.

In this regard, a recent deep learning work [87] proposed a modified LSTM network for nodule malignancy classification using lung nodule follow-up images. In particular, they aggregate in the forget and input gates a method to weight the importance of the temporal distance between scan images. This method was trained using a subset of the NLST (with 1794 cases) and obtained performances around 82% of AUC. In another study [14], an end-to-end deep learning based pipeline was built for lung cancer prediction using two CT images per patient (current and previous year). This approach proposed a pipeline composed by three 3D CNN networks, one for analysing the lung CT image, another for analysing nodule patches, and a final one, to provide cancer risk prediction using outcomes from previous two components. The method reported high AUC score of 94.4% using a subset of NLST (for 6,716 cases, 86 cancer-positives), although predictions were restricted to 1-year cancer risk. More recently, in [308] an attention-based 2D CNN network was built using pre-trained weights and a multi-time-point classification in a Siamese structure. Attention was on slice-wise for reducing network parameters to learn an appropriate nodule malignancy classifier. The Siamese-style architecture was proposed by allowing various number of inputs to be processed concurrently while also reducing the number of overall weight parameters since they are shared across twin branches. Best results of this approach reported an AUC of 88% in a test set of 170 nodules with 2 time-points. Also in [129], a deep learning approach was described for predicting lung cancer risk at 3 years and lung cancer-specific mortality. This study, although not being focused on automatic image analysis, uses a multilayer perceptron to ensemble nodule and non-nodule features associated to lung abnormalities.

III) Nodule growth forecasting

Determining the growth of the nodule is central for a proper malignancy estimation and treatment prescription [199]. However, lung tumours are highly heterogeneous (e.g. in size, texture and morphology) and their growth assessment is subject to inter and intra-observer variability (up to 3 mm in diameter on spiculated nodules [106]), making it complex to derive general patterns of tumour growth.

Traditionally, the tumour growth prediction problem has been addressed through complex and sophisticated mathematical models [254], such as those based on the reaction-diffusion equation [287, 292] also known as Fisher-Kolmogorov model. These methods provide informative and interpretable results. However, the number of model parameters is often limited (e.g. 5 in [318]), which might not be sufficient to model the inherent complexities of the growing patterns of the tumours.

Deep learning has recently addressed the tumour growth estimation to overcome the limitations of conventional approaches. For instance, in [328], authors proposed the use of two (invasion/expansion) stream CNNs for pancreatic cancer. The network reads 2D patch images of the tumour and predicts future tumour segmentation as well as tumour volume growth rate. Interestingly, the method allowed integration with clinical data to enable personalization. Best method performances achieved 86% of Dice score and 8.1% relative volume difference (RVD). Those overcame state-of-the-art of conventional mathematical models [319] for that disease type. However, the size of the test set was too small (10 cases) to extract robust conclusions. Also, to make inference this network required multimodal images (i.e. dual phase contrast-enhanced CT and FDG-PET), as well as three time points spanning between three and four years, which represented strong pre-conditions for the usability of the model. Alternatively, a recent work used the ability of RNN for exploiting the temporal patterns on tumour growth prediction. For instance, in [329], a 3D convolutional LSTM network [267] was proposed for predicting pancreatic tumour growth. Interestingly, in this study, features from the clinical history of the patient were integrated in the network with the intention to find extra non-linear relationships between spatial and temporal features. This approach used a limited dataset (33 cases) and required having series (≥ 2) of previous images of the lesion, which for early tumour growth estimation is not the best scenario due to the aggressiveness of the disease. In [173], a 2D deep convolutional GAN was presented for discriminating between true tumour progression and pseudo-progression of glioblastoma multiforme. The results confirmed its suitability for prediction and feature extraction, although only one image per tumor was used in the study. In [76], a stacked 3D GAN was built for growth prediction of gliomas using temporal evolution of the tumour. Although high performances were reported (88% Dice score), the database was composed by only 18 subjects, in which all tumours always grew. In [235], different GAN networks were built to predict the evolution of white matter hyperintensities. They also demonstrated the potential of using GANs in a semi-supervised scheme, improving results of a deterministic U-ResNet [330]. Despite the satisfactory performances obtained with GANs, this type of networks suffers from mode collapse [94], that is, they hardly generate correct representations of the probability output distribution, so they may not be adequate to model uncertainty. Alternatively, in [140], a deep auto-encoder attached to an FC network was shown for colorectal tumour growth detection. Despite providing results close to the RECIST methodology[2] and radiomic measures, the use of the auto-encoder was for mere feature reduction. In [28], authors applied a VAE for progression of Alzheimer's disease from structural MRI images. Their experiments demonstrated that VAE outperforms conventional CNNs on doubtful cases, as it acts as a soft classifier learning a Gaussian distribution. Also, for patient risk analysis they observed that VAE produced less false positive cases, sampling from the latent space, than deterministic CNNs. However, CNNs provided better overall performances. In another study [239], a deep auto-encoder, conditioned on fixed characteristics such as age and diagnosis, was defined to generate sequences of 3D MRI for Alzheimer's disease progression. Despite results outperformed previous 2D versions, some artefacts and false structures were noted on the generated images. Moreover, additional terms were required to ensure loss stability, latent space continuity, reducing memory constraints and restoring 3D outputs. Given the ambiguity present in medical images, in [231] a deep probabilistic generative model (sPUNet) [29, 153] was used to model glioma growth for radiotherapy treatment planning. The model, based on a combination of a U-Net [248] and a CVAE [275], was able to generate multiple future tumour segmentation modes on a given input. Although they demonstrated the potential of providing multiple views over a single solution, they did not report nodule growth performances.

Regarding forecasting lung tumour growth, in [311] a network was proposed to combine convolutional layers and gated recurrent units with an attention mechanism [193]. The goal was to predict spatial and temporal trajectories over a course of radiotherapy using a longitudinal MRI dataset. Although the purpose of this study is similar to ours (i.e. future lung tumour growth estimation), the complexity of the problem differs in that the images analysed were MRI (instead of CT), the period of the predictions were weeks (instead of months/years), and the number of input images (i.e. 2-3) to the network was larger than in our case. In another recent work [177], a method was proposed to generate a future image of the nodule. To do this, a temporal module encoded the distance at which to make the prediction, and two 3D U-Net [202, 248] networks extracted the warped and texture image features of the lung nodule. The network was trained with more than 300 pairs (prior and current studies) of 3D nodule centred patches. Experiments reported a high balanced accuracy score of 86% for nodule progression,

---

[2] https://recist.eortc.org/

although a relative Dice score of 65% for future nodule segmentation.

The gap in the model's ability to provide future segmentations of the tumours, the use of a tailored criterion to determine nodule growth instead of conventional metrics (e.g. the longest diameter or double time volume) or not taking into account inter-observer variability, shows the need to continue with the investigation of more reliable and effective solutions.

3.3 Challenges and limitations

One of the main concerns when adopting data-driven approaches for medical imaging problems is precisely the quality of data from which to learn from. In automatic lung cancer assessment using CT scans, the acquisition protocol (e.g. pixel spacing, slice thickness, volume size, contrast agents) is a principal limiting factor on the quality of the generated data. CT scans usually work at the minimum radiation possible in order to reduce its interaction with the disease, obtaining images with limited resolution. Building highly predictive models on top of these ambiguous images (e.g. low contrast and SNR) is hard. Models struggle to extract useful image representations that let them distinguish among different labels (e.g. normal and malign tumour). At the same time, image ambiguity limits the quality of the annotations. The large variability in shape and texture of nodules (specially on smaller ones) without a clear image resolution lead to inter-observer variability, producing weak labels. As we have shown, several deep learning contributions to lung cancer assessment have focused on making the networks better on extracting predictive features from these images (e.g. multi-scale, multi-view, attention gates). However, due to the relevance of having accurate networks, further work is still required in these terms.

At the same time, as we have seen, the amount of available data for training deep learning models for lung cancer assessment is scarce. Restrictive data sharing policies, or the lack of resources in clinical institutions, makes it complex to release publicly annotated data for the research community. Labelling medical data is time-consuming, annotators are usually overloaded and do not dispose of the pace and time desired to perform meticulous annotations (e.g. semantic segmentation), limiting the amount of labelled data and introducing variability in the data. This poses a challenge for deep learning algorithms, since they require large amounts of data for training the large number of parameters of such deep networks. In this chapter we have seen several techniques to overcome this limitation, although still further research is still required. One clear case, worth to be investigated could be the possibility to transfer knowledge from models learnt with weak radiological labels but with more voluminous datasets (since they are annotated just by visual inspection), to models with few confirmed malignancy cases (more difficult to collect), although with stronger beliefs.

As already seen, providing automatic support for lung cancer assessment is a challenging task due to it is highly conditioned by the different sources of variability (e.g. data acquisition, tumour heterogeneity, quality of annotations, agreement between readers). Although several works have shown high accurate deep learning solutions, they mostly come as black boxes, i.e. they only provide a single outcome directly to the clinicians. This really limits the reliability and trustfulness of these models for such critical tasks. In these terms, we believe it is required further work on this topic to enhance the interpretability of the model outcomes. Related to this, an interesting research direction is quantifying the uncertainty of the model estimates. This could permit providing to the clinicians how much certain is the model on their predictions (e.g. nodule detection, segmentation, growth, malignancy). We also believe this problem still requires further research, as demonstrated by the scarce number of works which have addressed this important feature for clinicians.

Another factor observed in the literature review, is the large effort done by computer vision and artificial intelligence researchers on lung cancer assessment focusing on single time-point CT images. Although single time-point pipelines have the potential to automatize the early lung cancer prediction, they do not have information regarding the temporal evolution of the nodule, even if as stated by international clinical guidelines, lung nodule malignancy is highly determined by how it evolves (i.e. its growth rate). Providing automatic support on the lung nodule follow-up could have great implications on the current clinical practice. Actually, for each suspicious detected nodule, clinicians have to perform a closely follow-up. This means having to perform more costly studies, subsequently requiring further resources to compare and analyse each of these studies. This pose a clear bottle-neck for the already overloaded radiological units of health institutions. Unfortunately, few works have really addressed this topic. The main reason that explains this fact is the lack of available longitudinal annotated data. Therefore, most of the studies rely on small in-house datasets which limits the reliability and the performance of the methods. Further research is required for the longitudinal analysis of lung nodule malignancy, but also more efforts are required for collecting large and heterogeneous longitudinal datasets.

Beyond providing support to automatize lung nodule follow-up, little research has been addressed towards lung tumour growth forecasting. This functionality could help clinicians to determine following treatments and clinical interventions. Unfortunately, several challenges have to be faced, such as the lack of longitudinal annotated data (as already mentioned), the inter-observer variability on the growth annotations and the stochastic growth factor involved in the nodule malignancy (e.g. sub-solid nodules grow at different rates than solids, size is also another determining growth factor). Despite these complications, recent advances in recurrent and deep generative models have demonstrated to be promising lines of research on this topic.

In the following sections we aim to provide further details regarding some of these principal limitations, how we have faced them and which have been our contributions to most of these problems.

Chapter 4

INTEGRATION OF PULMONARY NODULE MALIGNANCY IN A LUNG CANCER CLASSIFICATION PIPELINE

4.1 Introduction

Lung cancer is the uncontrolled growth of abnormal cells in one or both lungs. These abnormal cells can form tumours and interfere with the normal functioning of the lung.

Although the 5-year survival for lung cancer has improved over the last fifty years, it is still one of the most common cancers, accounting for over 225,000 cases, 150,000 deaths, and \$12 billion in health care costs yearly in the U.S. [48]. It is also one of the deadliest cancers; only 17% of people in the U.S. diagnosed with lung cancer survive five years after the diagnosis, and the survival rate is even lower in developing countries.

Early detection of lung cancer significantly improves the chances of patient survival. However, in most cases, a patient is unaware that she/he has a pulmonary nodule until a chest X-ray or a low-dose computed tomography (CT) scan of the lungs is performed, typically after physical symptoms appear, which occur most often in advanced stages of the disease. For this reason, early stage detection of benign and malignant pulmonary nodules plays an important role in clinical di-

---

The work described in this chapter is based on the following journal publication: Rafael-Palou X, Bonavita I, Ceresa M, Piella G, Ribas V, González Ballester MA. Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline. Computer Methods and Programs in Biomedicine. Vol. 185 (105172), pp. 1-9, 2020.

agnosis. Today, the gold standard for lung cancer detection consists in routinely taking a CT scan, and detecting nodules (i.e. small and approximately spherical masses) in it. Once lung nodules are detected, radiologists perform size measurements to assess their malignancy. To support them in this task, several guidelines like LungRADs [12] and Fleischner [194] have been proposed. These guidelines are a compilation of well documented cases and a set of rule-based recommendations from the clinical experience designed to help clinicians to discern among pulmonary nodules, normal tissues and artefacts, as well as to determine the inherent malignancy of the nodules. However, they are constrained to a limited number of visual parameters (e.g. size, morphology, texture and location of the nodules) and to a fixed range of values.

Low-dose CT is an effective method for radiologists to early identify lung cancer [298], although it presents several limitations. First, radiologists need to process large volumes of CT slices, usually with a low signal-to-noise ratio, which causes erroneous classifications of regions with weak or irregular contours. In addition, lung cancer diagnosis through CT is often subjective and highly affected by observer's experience, fatigue and emotional state [227], which can lead to inconsistent results from the same radiologist at different times or from different radiologists examining the same CT image.

Emulating the decision process of radiologists to determine malignancy of a nodule would be an extremely useful tool to help physicians plan future interventions for patients. Several approaches can be found in the literature relying on artificial intelligence and computer vision techniques. Conventional solutions (e.g. [93, 141]) propose engineering handcrafted features extracted directly from the CT image to build standard machine learning classifiers. This approach achieves satisfactory results when nodule candidates are well-defined, but shows some shortcomings when the nodules present complex and different sizes, shapes and context. An alternative recent solution to this problem is the use of convolutional neural networks (e.g. [44, 266, 321]), which are able to learn automatically inherent representations directly from the raw images.

In this chapter, we aim to evaluate the relevance of nodule malignancy to automatically predict lung cancer. To this end, we propose to use 3D convolutional neural networks to build accurate nodule malignancy classifiers trained on a public dataset of CT images with the annotations made by radiologists on the pulmonary nodules.

The main contribution of this chapter with respect to previous works is twofold. First, we provide a framework to allow integrating nodule malignancy classifiers, built at nodule level, into a pipeline that does not take into account malignancy information, but predicts lung cancer at the patient level. To this aim, three different features obtained from the malignancy classifiers were concatenated to a baseline lung cancer classifier: predicted classes, probabilities and features extracted from the last layer of the network. Secondly, we quantified the contribution of the nodule malignancy classifiers for lung cancer prediction. For this objective we evaluated the three different types of integration, and we compared their performances with that of a baseline lung cancer pipeline. This baseline was implemented using well established methods so that we could reliably quantify the impact of the nodule malignancy integration within the cancer detection pipeline.

The chapter is organized as follows: in the next section we review the existing related work on nodule malignancy and cancer classification. Then, we present the methods and materials used. Finally, we provide the results and a thorough discussion about the main outcomes presented in the chapter.

4.2 Related Works

In the past years, numerous works have addressed the problem of classifying the malignancy of pulmonary nodules in CT scans; some of these works use as features the radiologists' annotations of the nodules and perform classification for example with rule-based [141] and statistical learning [108] methods or by building a machine learning classifier [93] or an ensemble of classifiers [336, 337].
In other works, in addition or as alternative to radiologist annotations, shape-based, margin-based, and texture-based features [67] or 3D features of the nodules [242] are computed directly from the image with classical image analysis techniques.
In more recent years, it has been shown that deep learning techniques can outperform standard techniques in discriminating benign from malignant nodules (e.g. [266, 277, 322]). In [127], a deep belief network is used to extract from nodules features that are fed to a convolutional neural network aimed at classifying nodule malignancy. In [162], deep features are extracted from an autoencoder. In [44], high malignancy classification accuracy is achieved by using a convolutional neural network and radiological quantitative features.

Despite the abundance of papers focusing on classification of nodule malignancy and on nodule detection [72, 145, 334] in CT scans, little effort has been put in providing a systematic analysis of the effects of combining both, to assess whether predicting malignancy at the nodule level is beneficial for cancer prediction at patient level. To the best of our knowledge, [265] is the closest work to ours that tackles this issue. However, the focus of [265] is limited to the transferability of deep features of nodules to the cancer prediction task and the input data are exactly located nodules. Our aim is, instead, to provide and evaluate different types of nodule malignancy integration within an end-to-end cancer detection pipeline that takes as input raw CT scans.

4.3 Materials and Methods

4.3.1 Data

LIDC and LUNA16 datasets

The LIDC [18] is the largest publicly available reference database for lung nodules. It contains a total of 1018 CT scans, each of which is associated with a file containing annotations from four experienced thoracic radiologists. The annotations are the result of a two-phase reading process in which the radiologists were asked to mark suspicious lesions and to provide additional characterization of lesions of diameter larger or equal to 3 mm which were marked as a nodule [19].

In this work, we use an updated version of the LIDC dataset provided in the LUNA16 challenge [262], which includes only scans with at least one lesion of size $>=$ 3 mm marked as a nodule by at least three of the four radiologists. The LUNA16 dataset consists of 888 CT scans comprising a total of 1186 nodules. Annotations with coordinates of each nodule in the three spatial axes inferred from the original LIDC annotations are also provided.

We obtained the malignancy outcome of our classifiers from the annotation files in the LIDC database as they provide, among other characteristics, the subjective assessment of each radiologist of the likelihood of malignancy of the nodule. The admitted malignancy scores are discrete values ranging from 1 (highly unlikely for cancer) to 5 (highly suspicious for cancer). Since for each nodule included in LUNA16 we have the assessment of three or four radiologists, in order to obtain a unique label we averaged their scores.

TCIA Diagnosis Data

For 130 cases the LIDC dataset provides diagnostic data at patient level obtained from biopsy, surgical resection, progression or reviewing of the radiological images showing stable nodules after two years [52].

We retained this small dataset from the data used for building the malignancy classifiers, and we used it for training and testing the baseline and integrated lung cancer classifiers.

4.3.2 Method

Malignancy Classifiers

We describe here the approach used to build the nodule malignancy classifiers that will be integrated in our cancer prediction pipeline.

The input data to train the malignancy classifier consists of 3D cubes measuring

(32, 32, 32) mm centred in the centroid of the nodule computed from the coordinates in the LUNA16 annotation file. Note that each CT scan (i.e. subject) can contain more than one nodule; hence, to avoid any data leakage we assigned all the nodules belonging to a subject to only one of the training, validation or test sets. Additionally, we performed clipping (using a filter of [-1000, 400] HU) and normalization of the cubes.

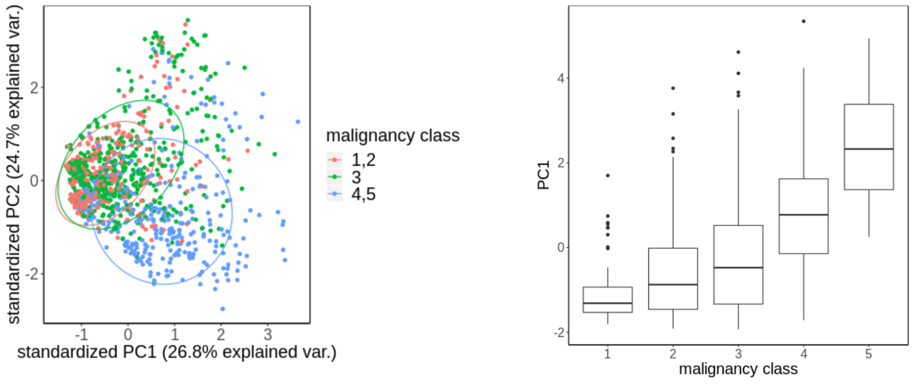
The malignancy score of each nodule was obtained from the original XML annotations using a parser provided by the second place winner of the DSB Kaggle competition [105] and averaging the radiologist scores as described in 4.3.1.

Given the binary nature of the final cancer prediction we want to achieve, we decided to remove nodules of ambiguous or intermediate malignancy from our experimental dataset. A Principal Component Analysis performed on some of the most relevant features annotated by radiologists showed that nodules of malignancy 1, 2 and 3 have similar feature distributions differently from those of malignancy 4 and 5 (Figure 4.1 b). Additionally, nodules of class 3 present higher variance, and form a less well-defined cluster in the principal components space (Figure 4.1 a). Therefore, we decided to remove them from our analysis. We hence opted for training and validating our classifiers on: *Dataset_145*, in which we selected only nodules labelled as 1, 4 or 5 and *Dataset_1&245*, in which we selected nodules of malignancy 4 and 5, and we merged them into one single category (renamed 1&2) nodules labelled as 1 and 2. Both datasets were split into stratified training (60%) and validation (40%) sets. As stated above, the test set for both the malignancy classifiers and the cancer pipeline consists of the TCIA data. However, only CTs containing at least one nodule with label 1, 4 or 5 (*Test_145*) or 1,2,4 or 5 (*Test_1&245*) were selected. Sample sets size and labels distribution are presented in Table 4.1.

Dataset	N subjects	N nodules			
		1(&2)	4	5	total
Dataset_145	247	72	213	48	333
Dataset_1&245	351	287	213	48	548
Test_145	65	15	65	9	89
Test_1&245	82	59	65	9	133

Table 4.1: Dataset used for building and testing the malignancy classifier.

The 3D nodule-cubes and corresponding malignancy labels are fed to a machine learning multi-class classifier. The model is based on a convolutional neural network (CNN). We designed two tailored networks for this purpose: a 3D CNN with 3 convolutional blocks, each followed by a 3D Max-Pooling layer and a dense block with a final dropout layer (*CNN_without_BN*), and a similar 3D



(a) (b)

Figure 4.1: PCA analysis and boxplot of radiologists annotated features per malignancy class.

CNN (*CNN_with_BN*) in which we applied batch normalization in each block, and we made more extensive use of dropout by adding it after each block. Note that for this last network we increased the kernel size of the last convolutional block and the number of units in the dense block.

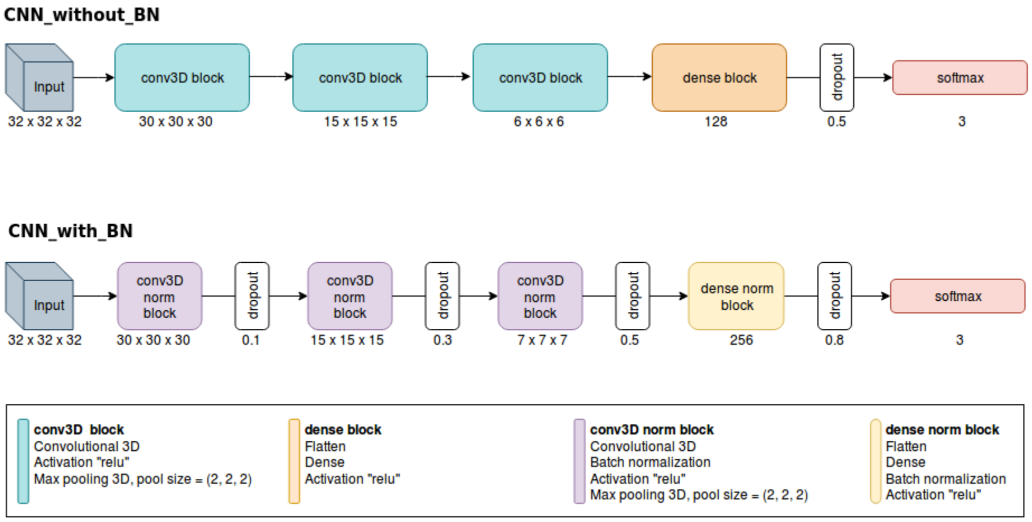


Figure 4.2: Architecture of the two CNN networks designed for the malignancy classifiers.

We trained the model weights with a batch training approach for 150 epochs and adopting early stopping with Adam optimizer [147] for regularization. We set the learning to 0.001, and we chose categorical cross-entropy as loss function.

Moreover, given the small size of our dataset, we used data augmentation on the training set (90 degrees of rotation, 0.02 of shear, zoom range of 0.1, shift of 0.05 and horizontal and vertical flip). Different training and validation batch sizes together with other input parameter combinations have been tested. A detailed description of the network architectures is presented in Figure 4.2.

The combination of the two network architectures and of the two datasets led to the creation of four malignancy classifiers:

- *CNN_without_BN_145* and *CNN_without_BN_1&245*
- *CNN_with_BN_145* and *CNN_with_BN_1&245*

Performances of these classifiers are presented in the results section whereas their integration and evaluation in the cancer prediction pipeline is described below.

The lung cancer pipeline

With the intention of setting a baseline method, we developed a two-stage lung cancer pipeline that did not take into account any information regarding nodule malignancy. We refer to this pipeline as our baseline method.

1. *Nodule detection*

   To build the automatic nodule detection stage, we used the LUNA16 dataset (reserving 10% for testing purposes) since it contains, for each CT, the location and diameter of the nodules. The first process performed was re-sampling each CT to an isotropic resolution (1, 1, 1) mm in order to reduce the variance given by the different pixel size/coarseness (e.g. the distance between slices) of the scans.

   Secondly, we performed a segmentation of the lungs from the re-sampled CTs, with the intention of reducing the analysis to the area of interest. For this task, we relied on a method proposed by the most cited kernel of the Data Science Bowl Kaggle competition [339]. This method consists in applying a threshold (i.e. -320 HU) to separate the air from the tissues. Then, it uses connected components to separate the lung air from outside, and finally it applies a morphological dilation to fill the existing gaps in the lung tissue.

   To detect nodule candidates in a CT, we used a 3D blob detector based on the Difference of Gaussian method [303]. This technique tries to detect nodules by retrieving those parts of the image that differ in properties, such as brightness or grey-level, compared to surrounding regions. One advantage of this method is its intuitive parameterization. In particular, we needed to tune 5 parameters: the minimum and maximum diameter of the region to look for (i.e. the minimum and maximum Gaussian standard deviations), the steps (i.e. the number of standard deviations to try between the defined ranges), a similarity threshold and the overlap score used for pruning closely located regions of interest. The configuration selected for this method was 5 mm and 60 mm as minimum and maximum nodule diameters, 5 steps, a threshold of 0.15 and 0.9 of overlapping. More details on the evaluation results of this method are available in the supplementary material (A.2).

As this candidate detection method tends to be optimistic (i.e. to accept several candidates similar in shape and texture to nodules), we implemented a classifier aimed at reducing the rate of false positive candidates. We chose to solve this task with a 3D CNN and, after empirical tests with different network architectures, we opted for the ResNet-50 [115]. To train this network, we used the same training set used for building the nodule detection method, along with a list of candidate nodule locations, provided by the LUNA16 challenge. The inputs to the network were volumes of (32, 32, 32) mm extracted from the nodule candidate positions. We used 0.0001 as initial learning rate, Adam optimization and binary cross-entropy for the loss function. Additionally, to improve the generalization ability of the network, in the training phase, we used data augmentation of the positive class by a factor of 1:240. In particular, we applied a 90 degree of rotation, 0.2 of shear, zoom range of 0.1, up and downs shifts of 0.5 and horizontal and vertical flips. The network reached its best performance in the training phase after 6 epochs with a batch size of 32. Further details regarding the evaluation of this method are also available in the supplementary material (A.2).

2. *Cancer classification*

The following stage of the pipeline consisted in building a lung cancer classifier, fed with the detected nodules, in order to predict cancer probability for each patient. For this purpose, we used the TCIA dataset that only provides patient labels (cancer or non-cancer). Given the lack of nodule labels, one of the main difficulties we had to face in building the classifier was to establish a nodules-patient labels relationship. We created a ground truth for the detected nodules from the ground truth of the patients by labelling all the nodules detected in a CT as 0/1 depending on the presence (1) or absence (0) of cancer in the patient scan. For example, if three nodules were detected by the pipeline in a CT scan of a patient with cancer, all the nodules were labelled as cancerous. Thus, we constructed a lung cancer classifier that predicts the probability of cancer of every nodule in a CT.

Then, since we were interested to report cancer predictions at the patient level, we reported as cancer probability of the patient the predicted cancer probability of his/her most cancerous nodule (i.e. the highest among the predicted cancer probabilities of all his/her nodules).

Additionally, in the classification, we included the main features provided by the 3D blob detector. In total, we selected three main features (radius, power and relative_z_position) referring respectively to size, signal intensity and location of the nodules. Texture related information was partially provided by the power feature (obtained from the 3D Difference of Gaussian method), which contains information about the spatial arrangement of grey intensities of the selected region of an image. Although further image descriptors could be envisaged, we limited our choice to those three not only to highlight the contribution of the nodule malignancy knowledge but also to approximate as closely as possible the features recommended in the current radiologist guidelines to focus on when screening nodules in a CT scan.

Several classification algorithms were used to train the classifiers, each accounting for a different classification strategy (i.e. linear, non-linear, distance-based, and tree-based). Moreover, different hyper-parameters were defined for each algorithm (Table A.6 of the supplementary material). In order to determine the best classification model, we used a grid-search and 5-fold cross-validation, a technique suitable for our sample size range [81].

Nodule Malignancy Integration

In order to assess the effects of the automatic nodule malignancy classification (section 4.3.2) for lung cancer prediction, we proposed three different methods to integrate the nodule malignancy models in the lung cancer pipeline: integration of predicted classes, integration of probabilities or integration of the models themselves (Figure 4.3).

1. Class integration

   The class integration method consisted in creating a new categorical feature containing the label predicted by the nodule malignancy classifiers. Thus, this feature was 0, 1 or 2 depending on whether the malignancy classifier predicted malignancy level of 1 (or 1&2 for *CNN_without_BN_1&245* and *CNN_with_BN_1&245* classifiers), 4, 5 respectively. To build the lung cancer classifier, we then concatenated this feature to the three basic features defined in 4.3.2 (cancer classification), namely, radius, power and z-position.

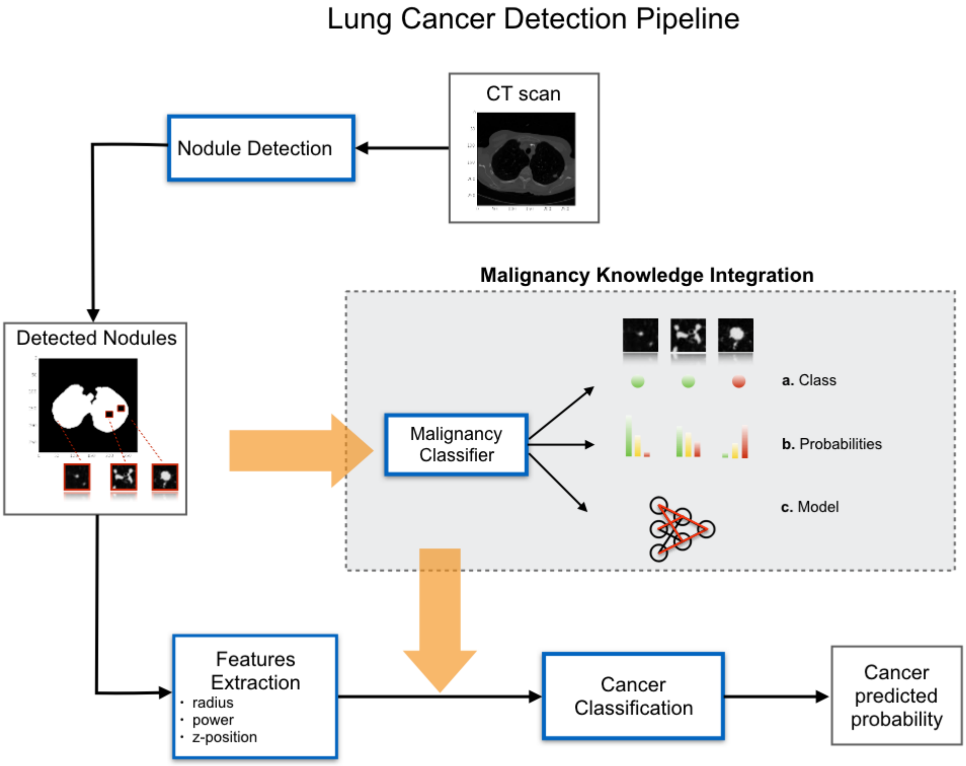


Figure 4.3: Pipeline proposed for lung cancer classification.

2. Probability integration

   The second integration method consisted in creating three new features, each containing the predicted probability of the nodule to be of malignancy level 1 (or 1&2 for *CNN_without_BN_1&245* and *CNN_with_BN_1&245* classifiers), level 4 or level 5. To build the lung cancer classifier, we then concatenated these three features to the three basic features.

3. Model integration

   The third integration method aimed to directly use the nodule malignancy models for the task of lung cancer prediction. Several techniques can be envisaged for this type of integration. We propose using transfer learning [95] since both problems have the same type of input data (CT scans) and a similar objective (identifying malignancy). To perform transfer learning, all the weights of the layers of the 3D malignancy networks were frozen, the last softmax layer was removed and replaced by a dense network (several configuration parameters of this network are presented in the supplementary material, Table A.7) and a final sigmoid layer. The first layer of the dense

Classifier	Precision	Recall	F1-score	F1-macro	Support
CNN_without_BN_145	0.83	0.81	0.82	0.68	89
CNN_with_BN_145	0.80	0.73	0.76	0.63	89
CNN_without_BN_1&245	0.83	0.83	0.83	0.67	133
CNN_with_BN_1&245	0.82	0.80	0.81	0.66	133

Table 4.2: Results of nodule malignancy classification on test set (at nodule level).

network was combined with the three basic features defined for the lung cancer classifier of the pipeline. The last layer of the final network outputs a value between 0 and 1 that represents the probability of lung cancer.

For tuning and evaluating the classifiers, independently of the integration method used, we applied a grid-search and 5-fold cross-validation as we did for building the cancer classifier of the pipeline.

4.4 Results

4.4.1 Malignancy classification

Although nodule classification is not the focus of our work, it is important to determine that these classifiers are able to extract useful information from the CTs before integrating them into the cancer pipeline. In Table 4.2 we summarize the weighted average performance metrics and the macro averaged F1-score on the test set of the four classifiers. The models *CNN_without_BN_145* and *CNN_with_BN_145* achieved the best performances with batch size of 32 in training and validation, while for *CNN_without_BN_1&245* and *CNN_without_BN_1&245* batch size of 32 and 16 respectively in training and validation were selected. In all the experiments we augmented each nodule in the training set by a factor between 10 and 25, augmenting more nodules of malignancy 5 given their lower representation in the dataset.

Overall, the more shallow architectures slightly outperformed the deeper ones; nevertheless, all the classifiers achieved a weighted F1-score above 0.75 with the best one (*CNN_without_BN_1&245*) achieving 0.83. These results indicate that the nodule deep features extracted by the CNN are good predictors of nodule malignancy.

Consistency between nodule-level malignancy predictions and patient-level diagnostic ground truth

To validate our hypothesis that the integration of a nodule malignancy classifier in a cancer detection pipeline can improve the predictions, we evaluated the consistency between the diagnosed cancer status of a patient and the predicted malignancy of his/her nodules. To do so, we inferred the cancer label of each patient from the malignancy labels of his/her nodules: if the CT scan of the patient contains at least one nodule with predicted malignancy 4 or 5, then the patient is positive to cancer, otherwise (i.e. all the nodules in the CT are benign) the patient is negative to cancer. Given this rule, we obtained cancer predictions at patient level in the cases where the predictions of nodule malignancy come from: 1) the radiologists, 2) the four malignancy classifiers. Performance metrics of these rule-based predictions are evaluated in the *Test_145* and *Test_1&245* sets (as they are the only provided with truth cancer labels) and are reported in Table 4.3. It is worth noting that both radiologist and CNN classifiers achieved comparably high, although not perfect, predictions (in Test_145 the best F1-score was 0.92 achieved by radiologists and *CNN_with_BN_145* while in Test_1&245 the best F1-score was 0.85 achieved by *CNN_without_BN_1&45* followed by 0.84 obtained from the radiologists prediction).

Dataset	Prediction source	Precision	Recall	F1-score	Support
Test_145	radiologist	0.89	0.94	0.92	65
Test_145	CNN_without_BN_145	0.86	0.96	0.91	65
Test_145	CNN_with_BN_145	0.88	0.96	0.92	65
Test_1&245	radiologist	0.89	0.79	0.84	82
Test_1&245	CNN_without_BN_1&245	0.87	0.84	0.85	82
Test_1&245	CNN_with_BN_1&245	0.83	0.78	0.80	82

Table 4.3: Cancer prediction at patient level from nodule malignancy.

4.4.2 Lung cancer

The pipeline described in section 4.3.2 was applied on the diagnosed TCIA dataset. From the 130 cases, we obtained that 100 (76.9%) were predicted with potential lung nodules, 11 cases (8.4%) were correctly predicted without any cancerous nodule and 19 cases (14.1%) were false negatives as they had some missing cancerous nodules.

On the 100 CT cases with detected nodules (227 nodules), we ran the cancer classification stage of the pipeline. The data was imbalanced with a non-cancer/cancer class ratio of 1:3.61. This ratio was respected during the random partitioning of the data in training and test datasets. In total, for training we had 75 cases (21 non-cancer, 54 cancer) with 220 nodules (48 non-cancer, 172 cancer). In contrast, for testing we had 25 cases (6 non-cancer, 19 cancer) with 57 nodules (12 non-cancer, 45 cancer). Figure 4.4 shows the distribution of nodules by patient and the box-plot of nodules for cancer and non-cancer CTs.

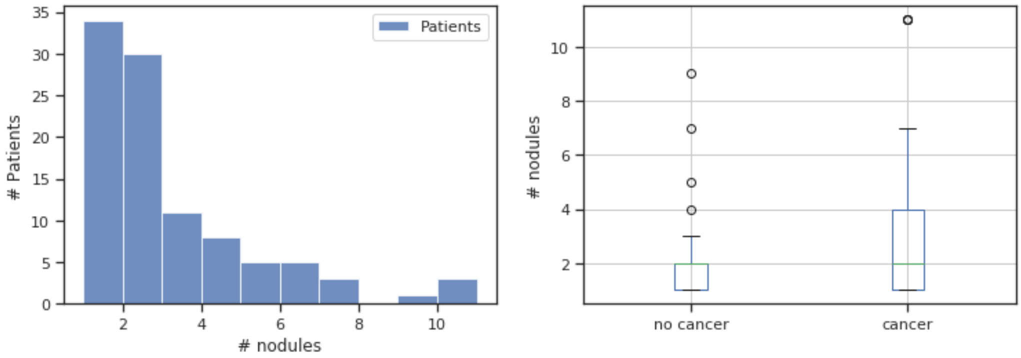


Figure 4.4: Data distribution for lung cancer classification.

The results of evaluating the different malignancy integration pipelines for lung cancer prediction are summarized in Table 4.4. This table shows the weighted precision, recall and F1-scores for cross-validation, test at the nodule level and test at the patient level. The baseline method achieved 0.65 +/- 0.021 of weighted F1 in cross-validation, whereas 0.55 in test at the nodule level and 0.593 in test at the patient level. The pipeline with malignancy probabilities integration method achieved the best results with 0.709 of weighted F1 in test at the nodule level and 0.74 of F1-weighted score in test at the patient level.

Figure 4.5 shows a bar-plot with the accuracy and the weighted F1-scores achieved by the different integration pipelines. The dashed lines represent the baseline classification performances. On the right, we show a precision-recall curve of the different lung cancer pipelines. This type of curves are especially appropriate when the classes are imbalanced as it shows the trade-off between precision and recall for different thresholds [253]. Basically, this type of curves is obtained applying several thresholds (th) on the predicted probabilities of the positive class (pr) of the dataset. The thresholds, which are typically the different predicted probabilities of the positive class (including 0.0 and 1.0), allow to define the positives cases ($pr > th$) and the negative cases ($pr < th$). Therefore, to obtain the list of pairs of precision and recall that conform the curve, we iteratively calculate these metrics for each of the different thresholds defined. To generate these curves, we have used the implementation provided by the scikit-learn library [229].

	Metric	Baseline Pipeline	Malignancy Integrated Pipelines		
			Class	Probabilty	Model
CV	prec	0.627+/-0.03	0.737+/-0.01	0.766+/-0.02	0.715+/-0.06
	rec	0.711+/-0.05	0.587+/-0.03	0.732+/-0.03	0.712+/-0.05
	F1	0.650+/-0.02	0.623+/-0.02	0.743+/-0.02	0.712+/-0.05
Test (ND)	prec	0.615	0.685	0.692	0.703
	rec	0.509	0.491	0.737	0.684
	F1	0.550	0.536	0.709	0.693
Test (PT)	prec	0.553	0.660	0.842	0.704
	rec	0.640	0.640	0.800	0.720
	F1	0.593	0.640	0.740	0.711

Table 4.4: Cross-validation and test (ND: nodule level, PT: patient level) results for the lung cancer pipelines.

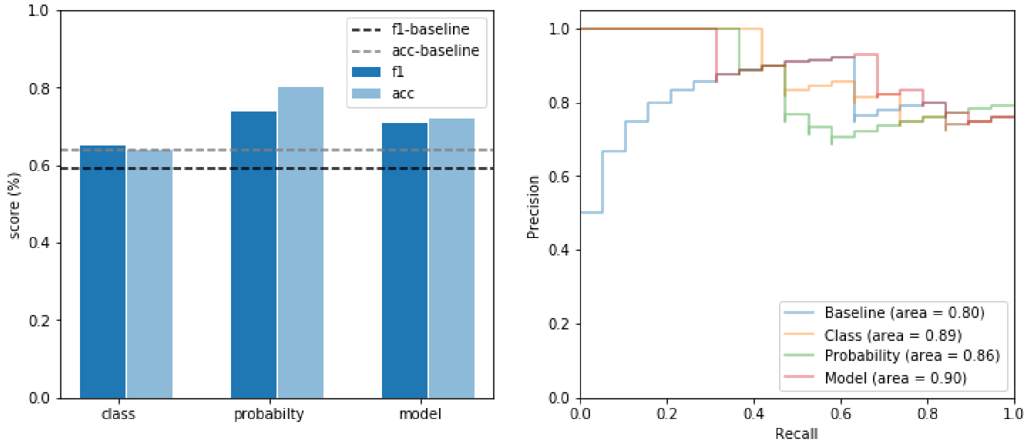


Figure 4.5: Performance comparison of the lung cancer pipelines.

4.5 Discussion

One of the most critical tasks that radiologists have to perform when examining lung CTs is to identify nodules from normal lung tissue. Highly malignant nodules are usually candidates of being lung cancer. Therefore, radiologists should precisely quantify the malignancy of the pulmonary nodules before planning expensive and sometimes traumatic clinical interventions.

Measuring nodule malignancy is a complex and tiresome process with significant levels of intra- and inter-observer variability. Several tools relying on image processing and conventional machine learning techniques or, more recently, convolutional neural networks have been proposed to support radiologists in this task. However, to the best of our knowledge, very few of them (e.g. [265]), independently of the technique selected, use nodule malignancy for the classification of lung cancer. With the intention of providing a realistic evaluation of the importance of nodule malignancy for the automatic lung cancer classification, in this work we have provided a framework with different methods to integrate nodule malignancy in a cancer detection pipeline.

With this aim, we created several nodule malignancy classifiers using 3D convolutional neural networks. To build these classifiers, beforehand, we knew the level of malignancy, the position and the size of the nodules to classify. The best nodule malignancy classifier (CNN_without_BN_1&245) achieved a relevant performance 0.83 of weighted F1-score when classifying the malignancy of the nodules in an independent test set.

The expected usefulness of these classifiers to the task of lung cancer prediction was validated by deriving a cancer classification from the nodule malignancy prediction on the TCIA diagnosed dataset. The best malignancy classifier (CNN_with_BN_145) achieved a performance of 0.92 of a weighted F1 score, comparable to the performance using the malignancy annotations given by the radiologists. However, it is worth noting that the evaluation was performed knowing a priory the location of the nodules and that the nodules annotated with a label 3 were removed due to their ambiguous malignancy.

To have a more realistic evaluation, we first created a baseline pipeline comprising a nodule detection and a cancer classification that uses a very simple set of descriptors (such as the radius, signal intensity and location of the candidates). We limited the number of features to this basic set to reasonably emulate the features recommended in the current radiologist guidelines. More radiomic features could have been added to further increase the performance. This would however complicate the assessment of the contribution of malignancy in cancer prediction, which is the main focus of this study. It remains for future work the extraction of more advanced features that could be useful, along with malignancy, to improve cancer prediction.

Eventually, to assess the effects of automatic nodule malignancy classification for lung cancer prediction, we provided three different ways to integrate the nodule malignancy classifiers into a lung cancer pipeline. The first approach aimed to use only the predicted classes as a new feature to add into the basic set of features of the baseline pipeline. The second approach consisted in creating three new features, representing the nodule malignancy probability distribution, and adding them to the features of the baseline. Finally, the last integration method consisted in using directly the malignancy model for lung cancer classification. In particular, we used a transfer learning technique which consisted on freezing the weights of the malignancy classifiers, removing the last layer and replacing it by new dense layers.

In total, three new pipelines were created by applying the different integration techniques within the baseline pipeline. The three pipelines and the baseline were trained using the TCIA dataset and evaluated using a grid-search with a 5-fold cross-validation.

Results show that the best pipeline with integrated nodule malignancy outperforms up to a 15.9% and 14.7% of weighted F1 score in comparison with the baseline at the nodule and patient level. The best pipeline was using the malignancy probabilities, and it achieved a difference of 2.9% of weighted F1 score at the patient level with respect to the second-best integration pipeline, the malignancy model integration. This result may appear surprising since the model integration adds to the classifier more features and hence more information. However, this extra information comes at the cost of an increased dimensionality of the problem, suggesting that this transfer learning approach may be better suited when a larger dataset would be available. Alternatively, a further fine-tuning (e.g. unfreezing or removing more layers) of the transfer learning proposed can be envisaged. Nevertheless, the model integration pipeline significantly outperformed the baseline by 11.8% of weighted F1 score at the patient level. In contrast, malignancy class integration did not significantly improve the lung cancer classification performance of the baseline. The poorer performance of the class compared to the other integration methods was expected, since the information was compressed into a single categorical feature not able to capture the complexity of the problem.

The findings of our study suggest that systematically integrating the assessment of nodule malignancy in an automated cancer detection system may improve significantly the ability of the system to identify cancer in lung scans. Emulating the malignancy assessment with powerful techniques such as deep learning, able to extract complex information directly from raw data, can relieve the difficulties and costs of a manual assessment. However, we believe that the lack of larger datasets with manual malignancy annotations and diagnostic cancer labels constitutes the main limitation of our study. If datasets of this kind become available in the future, our pipeline will highly benefit from the additional amount of information, which will likely result in more accurate predictions. Better predictions will eventually: reduce the need for time-consuming manual annotations and feature engineering approaches, provide a reliable support to radiologists and automatize to a greater extent cancer detection pipelines adopted in clinical applications.

Our work is, to the best of our knowledge, the first attempt to build this nodule-malignancy/patient-cancer integrated framework. Despite the encouraging results, several improvements can be envisaged to extend this approach. For instance, more advanced nodule detection methods [145, 320, 334] could be implemented for increasing the overall performance. Also, implementing an ensemble of all the malignancy classifiers instead of using them individually could enhance the classification performance. Furthermore, nodule malignancy could be also used for filtering nodule candidates detected by the cancer pipeline. Thus, rather than using all the detected nodules, we could use only the most malignant ones as input for the lung cancer classifier. Finally, another approach for building the lung cancer classifiers at the patient level would be to summarize all the nodules of the patient in a single row by computing several aggregated functions (e.g. max, min, mean) of the features (radius, power, z-position and CNN-malignancy features) obtained per each nodule. This would eliminate the need to infer the labels of each nodule of the patient, although at the cost of increasing the overall number of features.

4.6 Conclusions

In this chapter we have proven that it is feasible to build highly accurate malignancy classifiers relying on deep learning techniques to predict nodule malignancy. We have validated that they are also good predictors of lung cancer at the patient level when having the location of nodules beforehand. In order to provide a more realistic evaluation of nodule malignancy for lung cancer classification, we finally proposed a novel framework to quantify and assess nodule malignancy for lung cancer given only CTs and labels at the patient level. The experimental findings of this study suggest that systematically integrating the assessment of nodule malignancy in an automated cancer detection system improves up to 14.7% of F1-score the ability of the system to identify cancer in lung scans. The encouraging results presented are, to the best of our knowledge, the first attempt to build this nodule-malignancy/patient-cancer integrated framework to quantify nodule malignancy for future research in lung cancer classification.

Chapter 5

RE-IDENTIFICATION AND GROWTH DETECTION OF PULMONARY NODULES WITHOUT IMAGE REGISTRATION USING 3D SIAMESE NEURAL NETWORKS

5.1 Introduction

Few CAD systems [14] have been proposed for the automatic support of lung cancer follow-up. Major developments in the field are mainly limited by the lack of open datasets with annotated series of CTs. To analyse series of CT scans, prior and follow-up lung exams have to be initially registered to facilitate, for instance, the correct re-identification of pulmonary nodules. Several factors compromise the effectiveness of the registration process, such as the variability in the image size and resolution originated by the use of different CT scans, and the variability in the position and breath cycle of the patients when performing the scanning.

Although current medical image registration methods [276], especially non-linear [251], report accurate CT alignments, they are still slow and introduce some distortions in the intrinsic structure of the lung, hindering their wide clinical ac-

---

The work described in this chapter is based on the following journal publication: Rafael-Palou X, Aubanell A, Bonavita I, Ceresa M, Piella G, Ribas V, González Ballester MA. Re-Identification and growth detection of pulmonary nodules without image registration using 3D siamese neural networks. Medical Image Analysis. Vol. 67 (101823), pp. 1-12, 2021.

ceptance [309]. In addition, other complexities must be addressed, regardless of the quality of the image registration, to enable a proper nodule re-identification, such as the existence of several nodules close to each other, and/or the alteration in texture, size, and even location of the nodules due to disease progression. Therefore, more research is still needed to reliably include the nodule re-identification in different CT scans, in automated tools to support physicians in the analysis of longitudinal studies of lung cancer.

This work aims to take a step in this direction, and proposes a novel approach for the re-identification of pulmonary nodules. In particular, we propose a 3D siamese neural network [152] to predict the most likely matching nodules from a series of lung CT scans of the same patient. This approach does not require prior registration of the CT scans, avoiding some of the shortcomings that it entails. In addition, to demonstrate the value of this approach, we integrate it into an automated pipeline aimed to detect the growth of pulmonary nodules over time.

The contributions of this paper with respect to previous works is two-fold. First, we investigate and provide several models for re-identifying lung nodules in CT scans series, relying directly on 3D volumetric data, transfer learning, and siamese neural networks. In this sense, to the best of our knowledge, this would be the first time that the problem of pulmonary nodule re-identification is addressed through deep learning techniques. Secondly, we build and evaluate an automatic pipeline that integrates the proposed models to predict nodule growth from longitudinal CTs.

5.2 Related work

5.2.1 Automated nodule re-identification

Lung nodule re-identification (i.e. matching) between current and former CT examinations is necessary for assessing nodule growth or shrinkage. While the majority of lung cancer CAD systems found in the literature focus on the nodule detection task [191], relatively few automated nodule matching systems have been proposed (partly because of the limited availability of follow-up datasets).

An early CAD system for nodule re-identification in series of lung CT scans was proposed in [150]. They reported high performances (86% nodules re-identified) using 8 patients (310 nodules), although some parts of the system required manual intervention (lung apex identification) and no train/test split was reported. In [171] a commercial CAD system was evaluated for nodule re-identification for 30 patients (210 nodules) with lung metastasis, reaching a matching rate of 67%. In a cohort of 54 pairs of low-dose multi detector CT screening, a CAD system successfully matched 91.3% of nodules $\geq$4mm [30]. In another commercial CAD evaluation study [295], a matching rate of 92.7% was achieved in three serial CT scans from 40 subjects with 143 nodules from the NLST[1]. Another CAD system evaluation [155] for automated lung nodule matching using annotations from 4 experts in 57 patients reported between 79% and 92% of accuracy scores. Deep learning-based CAD systems for analysis of longitudinal lung cancer studies are practically non-existent in the literature. An exception is in [14], where a CAD system for end-to-end lung cancer screening is proposed. However, nodule matching was not directly tackled in the study.

All these CAD systems rely on registration of the lungs in the different CT examinations. Performing an accurate registration of lung images is particularly challenging [206] due to the high deformability of the lung tissue and the volume changes during the breathing cycle. Previous studies [122, 281] evaluated methods for rigid and non-rigid registration for matching lung nodules on sequential chest CT scans. [207] presented the results of the EMPIRE10 pulmonary image registration challenge, which comprised a comprehensive evaluation and comparison of more than 20 algorithms on 30 thoracic CT pairs. Top-5 algorithms were using different non-rigid transformations. Although non-rigid registration is usually more accurate than rigid registration, rigid registration is substantially more computationally efficient, potentially making it more useful in a busy clinical setting in which real-time processing is necessary. A more recent and complete review of registration methods for medical image series analysis can be found in [276]. The choice of the right registration method and of the correct evaluation metric to assess its performance are of crucial importance, as they can affect the results of the analysis.

5.2.2 Siamese Neural Networks

The problem of nodule re-identification can be closely related to the one of recognizing the same object in different images. This type of problems has been successfully addressed by Siamese neural networks [38] (SNNs). They are designed as two sibling networks, connected by a distance layer at the top, trained to predict matching or mismatching between two input images. The original architecture, first introduced for the problem of signature verification, was later extended by [152] using convolutional layers and adjusting the optimization metric with a weighted L1 distance between the twin feature vectors of both networks.

SNNs have been extensively used in computer vision matching problems such as tracking objects in videos [296], matching pedestrians across multiple camera views [307], and matching corresponding patches in satellite images [131].

In the medical image domain, SNNs have been used primarily to extract a la-

---

[1] https://www.cancer.gov/types/lung/research/nlst

tent representation for content-based image retrieval. For instance, [49] proposed a SNN, pre-trained on the ImageNet dataset and using a contrastive loss function [103] to retrieve similar images to the query, using a publicly available dataset of diabetic retinopathy fundus images. Another example is the work by [41], which applied SNNs to retrieve similar images from several medical image databases of lung, pancreas, and brain. As far as we know, SNNs have not yet been applied to re-identify nodules in a series of lung CT scans.

5.3 Method

5.3.1 Nodule re-identification

To solve the problem of nodule re-identification in a pair of CTs of the same patient taken at different time points, we propose building a SNN [152]. An appealing characteristic of SNNs is that they rely on a distance metric computed on features extracted automatically by a deep learning network. This should allow greatly accelerating and simplifying the nodule re-identification process, avoiding introducing a registration technique as a source of variability and error in the analysis.

Siamese neural networks are composed of a feature extraction component in which two subnetworks (with shared architecture and weights) process a pair of images at a time to produce two embedding feature vectors directly from the images. A second component (i.e. the head of the network) aims to classify whether the two embedding feature arrays are similar or not. To assess this, the features are passed to a pairwise distance layer that computes a similarity score.

In a previous study [37], we trained a deep convolution neural network (CNN) for nodule classification, able to effectively reduce the number of false positives in the nodule detection problem. In the present work, we have adjusted that network, improving its final performance. In particular, we propose a 3D CNN based on a ResNet-34 architecture that expects nodule patches of 32x32x32. As described in the original paper, the patches are pre-processed crops done around the centre of the annotated nodules of the lung CT. The nodule classification network was trained from scratch using a large amount of nodule candidates ($>$ 750K) from the LUNA-16 challenge dataset [262]. Further details on its architecture and performance are shown in the supplementary material (B.1).

In the current study, we removed the fully connected layers of the nodule classification network to use it as the backbone of the sibling networks of the feature extraction component of the SNNs. Figure-5.1 shows the SNN architecture for the nodule re-identification problem. In this figure, we can observe the two components. First, the feature extraction component, which pre-processes the input nodule patches (i.e. taken at different time points, T1 and T2) and uses the sibling network to extract the corresponding feature maps. Second, the classification component composed of the head of the network that predicts if both feature maps are similar or not. These feature maps (solid arrows in Figure-5.1) come from different levels of the pre-trained sibling networks. Further details about the feature maps and the network heads are described in Subsection 5.3.1.

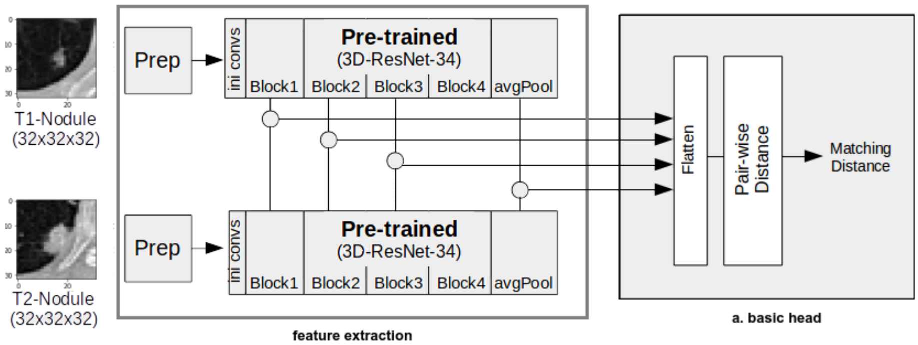


Figure 5.1: Siamese network proposed for lung nodule re-identification. The network is composed of a feature extraction and a basic head network to perform the prediction.

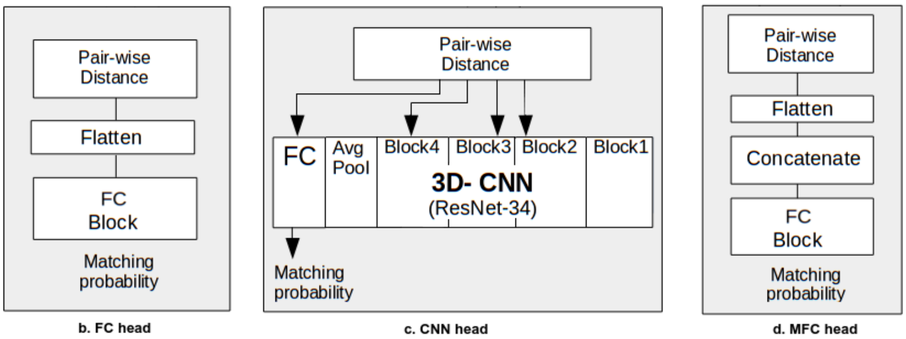


Figure 5.2: Alternative head networks to configure different siamese networks.

Different SNNs configurations were proposed (Table-5.1) to gain further insights into the best parameterizations. To allow a fair comparison of the configurations, we trained the SNNs with the same parameter values. Concisely, the number of epochs was set to 150, the learning rate to 1e-4, the batch size to 8, dropout to 0.3, the early stopping at 10 epochs without any significant improvement, and Adam [147] was used for optimization. Finally, random rotation, flip, and zoom were applied for data augmentation.

Below we describe in more detail the main configurations and parameters used in the experiments.

	Pre-trained	Feature maps	Head	Loss
FIBC	Frozen	Individual	Basic	Contrastive
UIBC	Unfrozen	Individual	Basic	Contrastive
FIFB	Frozen	Individual	FC	BCE
UIFB	Unfrozen	Individual	FC	BCE
FICB	Frozen	Individual	CNN	BCE
UICB	Unfrozen	Individual	CNN	BCE
FCMB	Frozen	Combined	MFC	BCE
UCMB	Unfrozen	Combined	MFC	BCE

Table 5.1: List of the different siamese network configurations. The index column contains the acronyms of the networks, resulting from joining the first letter of the options placed in the next 4 columns.

Pre-trained network weights

Two configuration values were proposed for this setting: frozen and unfrozen. Usually, the weights of the pre-trained networks in a SNN remain frozen. In this study the pre-trained network had a related but slightly different learning goal than the target (siamese) network. Thus, we allowed also the option of unfreezing the weights of the pre-trained network and updating them during the back-propagation steps of the siamese network training process. To un/freeze the networks, we dis/abled the option to update all the weights and biases of the pre-trained layers during training.

Feature maps

We propose two options: using the feature maps individually and combining the feature maps together. Feature maps extracted from the first layers of a CNN refer to low-level and less domain-specific representations (e.g. lines, circles, spikes), whereas features extracted from deeper layers are generally more high level and domain-related representation (e.g. morphology, texture). To analyse the potential of both general and more specific nodule features, we used features from different depths of the network (i.e., from the last layer of each of the 4 convolution blocks that holds the pre-trained Resnet-34 network). The resulting feature maps were obtained after a forward-passing through the network for each of the nodule images of the whole dataset. Table-5.2 shows the layer name, the number of filters per layer, the output dimension of each filter, and the total number of parameters for each of the selected feature maps.

We designed experiments to evaluate each of the possible feature maps, i.e. 4 individual features maps - one per layer - and 11 feature maps resulting from combinations: (6 over 2) + (4 over 3) + (1 over 4).

Layer	Filters	Dimension	Total params
layer1	64	[16,8,8]	65536
layer2	128	[8,4,4]	16384
layer3	256	[4,2,2]	4096
avgpool	1	[1,1,512]	512

Table 5.2: Layers selected from the pre-trained part of the SNNs.

Siamese heads

We proposed four different head networks, one meant to follow a more conventional siamese architecture and the others with more exploratory purposes, more precisely:

1. A basic head network (Figure-5.1) composed of a flatten (to homogenize all features to one dimension) and a pairwise distance (i.e. L1) layer, just after the feature extraction part of the network.

2. A fully connected (FC) head network (Figure-5.2b) composed of a pairwise distance, a flatten, and an FC block layer. The FC block comprises a FC layer (with 64 units), a batch norm, a ReLU, a dropout layer and a final FC layer (with one unit). This classifier head aims at finding non-linear patterns among the merged features (from both sibling networks).

3. A CNN head network (Figure-5.2c) composed of a pairwise distance layer and a clean (without pre-trained weights) ResNet-34 CNN. Several arrows connect the pairwise distance layer with this clean ResNet-34. There are as many arrows as pre-trained layers used to extract the features. The arrows redirect the features to a specific part of the clean ResNet-34. The redirection had to make compatible the dimensions of the output from the previous layer with the layers of the input. For instance, features extracted from last layer of block1 were linked to the initial layer of the block2, features from layer2 were linked to the initial layer of the block3 and so on. This head network aimed at exploring non-linear patterns between features but without losing the space dimension (i.e. no flattening of the features was done between the pairwise layer and the clean ResNet-34).

4. A multi-features combined (MFC) head network (Figure-5.2d) composed of a pairwise distance layer, a flatten layer, a concatenation layer (to merge all features), and a FC (already described above). This head network aimed at exploring combination of features from different parts of the network.

It is important to note that in the basic head network, the pairwise distance layer not only computes the batch-wise L1-distance between each component of the previously flattened input vectors, but also it sums the components up to eventually generate an output of size bs × shape (where bs is the batch size). This is done to accommodate the expected inputs of the contrastive loss function with which the basic head network is configured. For the rest of the head networks, the pairwise distance layer does not perform any reducing sum operation, leaving its input and output with the same size bs × 1 × z × y × x, and therefore, allowing its output to be exploited more deeply with additional layers (for example, convolutional or fully connected).

Loss functions

We explored two options: a contrastive loss and a binary cross entropy (BCE) loss function. Traditionally, SNNs are trained using a contrastive loss [103] function. This function encodes both similarity and dissimilarity (between the feature maps) independently in a loss function. It ensures that semantically similar pairs are embedded close together while forcing the dissimilar pairs to be apart from each other. Another option to train these networks is through a prediction error-based approach. For our case we adopted the binary cross entropy loss. This implied to apply a sigmoid function on the outputs to transform them into probability values (between 0 and 1).

5.3.2 Nodule growth detection pipeline

A valuable application of nodule re-identification is to predict nodule growth between current and follow-up CT scans of a patient. This is a crucial, complex, and time-consuming task for lung cancer assessment since nodule growth has a clear predictive importance for benignity and malignancy [102]. Thus, further efforts are required to support clinicians to increase the precision and effectiveness of such endeavour.

To this end, we propose an end-to-end pipeline (Figure-5.3) comprised of two different components: 1) a nodule detector that, given a pair of CTs of the same patient but taken at different time points, generates a list of nodule candidates per each CT, and 2) a nodule matching component (embedding the siamese networks) that, given the list of nodule candidates of the CTs, matches the nodules and computes the difference in diameter between them.

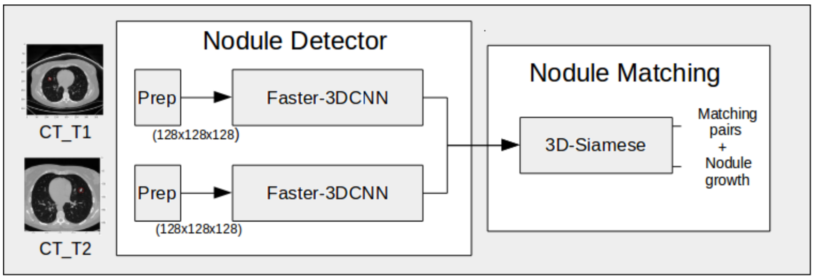


Figure 5.3: Nodule growth detection pipeline.

Nodule detector

To build the nodule detector, we followed the work of [178], with which they won the Data Science Bowl lung cancer challenge[2]. The authors proposed a 3D Faster R-CNN [244] scheme for nodule detection. The backbone of the network was similar to the U-net [248] architecture, in which the information flows not only in a classical bottom-up way but also between the encoder and decoder parts of the network thanks to some symmetric links (or short-cuts) that bound both parts of the network. The output of this network were probability feature maps, useful for the lung cancer classification problem.

To the original network, we proposed attaching a double CNN head as in [244]. One head was used for regression and the other for classification. The regression branch infers the centre (x,y,z locations) and the diameter of the nodule, while the classification branch predicts the probability of being a nodule.

The input lung CT was pre-processed before entering the nodule detection network. The image was resampled to an isotropic resolution ($1 \times 1 \times 1$ mm), pixel intensities clipped between [-1000, 600] HU and normalized between 0 and 1. The full lung image, without any previous lung segmentation, was then split in overlapping patches (due to memory constraints) of $128 \times 128 \times 128$ with an overlap of 32 pixels per dimension. Since the location of the patch may influence the decision of whether it is a nodule and whether it is malignant, we also introduce the location information in the network as in [178]. Thus, each patch was fed to the network together with its corresponding location crop of size $32 \times 32 \times 32 \times 3$, which contains the location of the patch image with respect to the whole lung image. The final network architecture used for nodule detection as well as the performance obtained in LUNA-16 [262] dataset can be found in the supplementary material (B.2).

---

[2] https://www.kaggle.com/c/data-science-bowl-2017

Nodule matching

This component performs the re-identification of the nodules among all CT pairs. To do this, for each pair of CTs, we took each candidate found at T1, and we paired with each of the candidates found at T2. The pairs were pre-processed following the specifications described in Section 5.3.1, and then they were fed to the SNN. The network, trained off-line, provided a matching probability for each pair of candidates. The pairs with the highest probability were selected as the matching ones.

To assess the performance of this process, we computed for each pair of CTs, whether the candidate at T2 predicted with the highest probability by the SNN, matched with the annotated nodule at T2. Additionally, we computed the time required for finding the matching nodules. We repeated this process for each of the SNN configurations.

Once having predicted all matching nodules for each pair of CTs, the pipeline returns the nodule growth along with the location and diameter of the matching nodules. The nodule growth is calculated directly by the difference between the predicted nodule diameters at T1 and T2 for each pair of lung CTs.

To evaluate the nodule growth detection, we selected all the correctly matched CT pairs and compared whether the nodule growth difference was of the same sign in both ground truth and predicted. True positive (TP) and false negative (FN) cases were those that had (in both ground truth and predicted) positive and negative growth differences, respectively. A false positive (FP) case was considered when the predicted growth difference was positive and the ground truth one was negative; and a false negative (FN) was considered in the opposite case.

5.4 Experiments and results

5.4.1 Evaluation datasets

LUNA-16

In this work we used an updated version of the LIDC dataset [18] provided in the LUNA-16 challenge [262], which includes only scans with at least one lesion of size $\geq$ 3 mm marked as a nodule by at least three of the four radiologists. The LUNA-16 dataset consists of 888 CT scans comprising a total of 1186 nodules. Annotations with coordinates of each nodule in the three spatial axes inferred from the original LIDC annotations are also provided.

VH-Lung

This dataset was designed specifically to identify and follow up suspicious lung nodules in time. Ethics approval was obtained from the Medication Research Ethics Committee of Vall d'Hebrón University Hospital (Barcelona) with reference number PR(AG)111/2019 presented on 01/03/2019.

Inclusion criteria were patients without a previous neoplasia, with a confirmed diagnosis, and with visible nodules ($\geq$ 5 mm) in at least two consecutive CT scans. The interval between current and previous CT examinations ranged from 32 to 2464 days. These nodules were located in the three spatial axes by two different specialists at each time point and quantified by another experienced radiologist. The size mean of annotated nodules was 11.08 $\pm$ 5.35 at T1 and 13.49 $\pm$ 5.18 at T2, and the growth size mean is 2.41 $\pm$ 4.38 mm.

The chest helical CT studies were performed using different scanners: Phillips (Brilliance 16/64, iCT 256), Siemens (SOMATOM Perspective/ Definition) and General Electrics (LightSpeed16). Acquisition and reconstruction protocols were set according to subject biometrics and clinical inquiry: 100–120 kV, 33-196 mAs and exposure time 439-1170 ms. Each image had 512 $\times$ 512 pixels with 16-bit grey resolution, spacing between slices 0.75-1.5 mm and slice thickness 1-5 mm.

In total, the dataset contains 151 patients with two thoracic CT scans. For each patient, the clinicians annotated only one relevant nodule in both CT scans. We randomly divided the dataset into two subsets, one for training (113 patients) with 70 cancers and 43 benign cases, and the other for testing (38 patients) with 25 cancers and 13 benign cases.

5.4.2 Nodule re-identification

In this paper we propose the use of SNNs for nodule re-identification. In order to train the SNNs, we first identified positive cases, i.e. pairs of the same nodule from the same patient taken at different time points (T1 and T2), as well as negative cases made up of pairs of mismatched nodules. In the VH-Lung dataset we had already annotated (N=151) positive cases. To create the negative cases, we used the nodule locations of the VH-Lung dataset at T1 together with a random nodule location of the annotated nodule locations at T2 (avoiding selecting the correct nodule location). In total, we build a balanced dataset (N=302) composed of 226 CT pairs in the training set and 76 CT pairs in the test set, thus respecting the initial training/test (75% / 25%) partition of the VH-Lung dataset.

We optimized the different SNNs (Table-5.1) with the training data using a stratified 10-fold cross-validation, and we tested them with the testing set. Results of the best SNNs configurations are shown in Table-5.3. Additional SNNs configuration results can be found in Table-B.4 (supplementary material).

Config.	Layer	tr_acc	val_acc
FIBC	layer2	0.790 ± 0.01	0.775 ± 0.05
UIBC	layer3	0.891 ± 0.01	0.864 ± 0.04
FIFB	layer1	0.939 ± 0.02	0.899 ± 0.03
UIFB	layer2	0.918 ± 0.03	0.890 ± 0.03
FICB	layer1	0.867 ± 0.03	0.857 ± 0.06
UICB	layer1	0.868 ± 0.06	0.888 ± 0.04
FCMB	layer1, layer2	0.938 ± 0.03	0.882 ± 0.03
UCMB	layer1, layer2, avgpool	0.954 ± 0.02	0.897 ± 0.04

Config.	Layer	test_acc	test_prec	test_rec
FIBC	layer2	0.709 ± 0.01	0.806 ± 0.01	0.550 ± 0.00
UIBC	layer3	0.798 ± 0.01	0.765 ± 0.02	0.863 ± 0.03
FIFB	layer1	0.921 ± 0.03	0.905 ± 0.05	0.944 ± 0.03
UIFB	layer2	0.896 ± 0.02	0.871 ± 0.05	0.934 ± 0.01
FICB	layer1	0.831 ± 0.04	0.824 ± 0.07	0.860 ± 0.06
UICB	layer1	0.859 ± 0.07	0.842 ± 0.09	0.900 ± 0.04
FCMB	layer1, layer2	0.918 ± 0.01	0.907 ± 0.02	0.934 ± 0.03
UCMB	layer1, layer2, avgpool	0.925 ± 0.02	0.904 ± 0.04	0.952 ± 0.03

Table 5.3: Performance results (accuracy (acc), precision (prec) and recall (rec)) obtained on training (tr), validation (val) and test for the different SNN configurations. The meaning of the configured methods is detailed in Table5.1.

In addition, we investigated the nodule re-identification performance in terms of nodule growth. In total, we found 14 cases (CT pairs) with an increase in nodule diameter $>$ 9 mm (aprox. Mean + 1.5*std), and 4 cases with a decrease in nodule diameter $>$ 4 mm (aprox. Mean – 1.5*std). We labelled these cases as large growth changes (Other similar studies [155] defined large nodules as $>$ 10 mm). We also found 50 cases with a nodule change ± 1 mm, labelling them as small growth changes, and the remaining 87 cases were labelled as medium growth changes. The results for our best method (FIFB) can be found in the Table-B.5 of the supplementary material.

5.4.3 Nodule growth detection pipeline

For the evaluation of the initial stage of the pipeline described in Section 5.3.2, we first computed the performance of the pipeline to detect the annotated nodules (one per CT). To do this, we proposed different thresholds (1, 4, 8, 16, 32, and 64) or number of nodule candidates, and we computed per each CT whether the annotated nodule was in each subset of predicted nodule candidates (ranked by probability). To have a better estimation of the nodule detection performance, we repeated this process on 10 random train-test partitions (respecting the proposed size of the initial partitions of the dataset) of the VH-Lung. Results are plotted in Figure-5.4. This FROC curve [262], shows the sensitivity, in average, of finding the (only) annotated nodule per scan at different nodule candidate rates. As we can observe, in training the detector reaches a sensitivity of 0.951 with 32 nodule candidates (missing 10.5 ± 1.02 annotated nodules in 226 different CTs), and in test set a sensitivity of 0.973 with the same threshold (missing 2.5 ± 1.02 nodules in 76 CTs).

We therefore configured the nodule detection component of the pipeline with a threshold of 32 candidates per CT, since it empirically showed a good balance between sensitivity (real nodules detected) and precision (number of nodule candidates not really targeted by the clinicians) both in training and test.

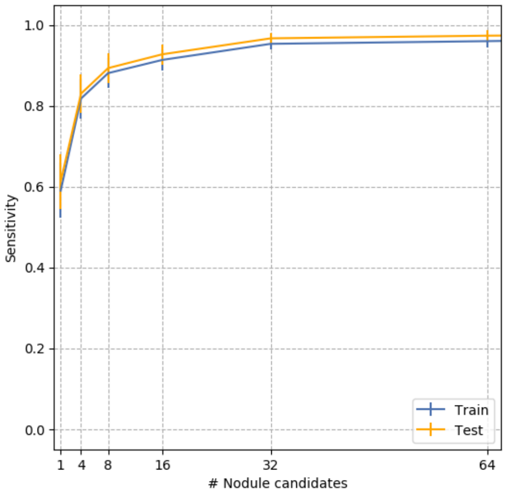


Figure 5.4: FROC-curve of the malignant nodule detection algorithm for training and test partition.

To gain insight into the complexity of the re-identification problem, we computed how many candidates were located within a chosen Euclidean distance from the nodule ground truth position (Figure-5.5). We defined 5 different distance thresholds: radius squared Euclidean distance (as used in the LUNA-16 challenge to accept a nodule detection as correct) and 4 fixed Euclidean distances (30, 20, 15, and 10 mm). For every distance, we computed the number of CTs in which 0, 1, 2, 5 or more than 10 candidates fell within the distance. Moreover, we computed an accuracy of detection for every distance choice by dividing the number of CTs for which only one candidate is within the distance by the total number of CTs. Results are shown in Table-5.4.

Next, we evaluated the performance of the best SNN (Table-5.3) for nodule re-identification using the location of the nodule candidates provided by the nodule

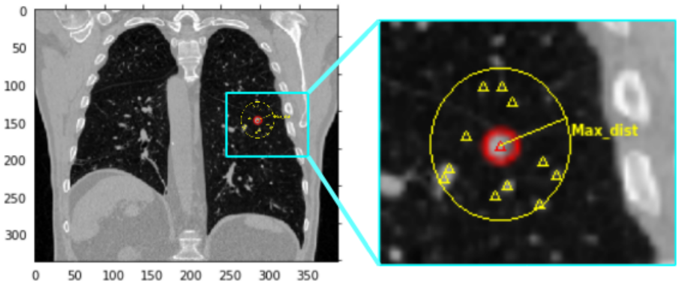


Figure 5.5: Candidates predicted (yellow marks) at a maximum distance from the ground truth centroid (red circle).

Distances	N=0	N=1	N=2	N=5	N>=10	Accuracy
radius[2]	0	18	6	2	3	0.500
30 mm	1	22	7	1	0	0.611
20 mm	1	26	6	0	0	0.722
15 mm	1	32	3	0	0	0.888
10 mm	1	34	1	0	0	0.944
5 mm	3	33	0	0	0	0.916
3 mm	5	31	0	0	0	0.861
1.5 mm	18	18	0	0	0	0.500

Table 5.4: Number of CTs (in T2) containing N candidates located within a chosen euclidean distance from the actual nodule centroid. The accuracy score represents the number of CTs at N=1 respect to the total of CTs.

detector. The best results were achieved by the FIFB network with only 4 CT-pairs incorrectly matched and an accuracy of 0.888. All results are presented in Table-5.5.

Configuration	Correct	Incorrect	Accuracy	Time(s)
FIBC	25	11	0.694	18.71
UIBC	27	9	0.750	36.01
FIFB	32	4	0.888	9.36
UIFB	30	6	0.834	12.73
FICB	30	6	0.834	20.12
UICB	28	8	0.777	20.16
FCMB	31	5	0.861	12.41
UCMB	31	5	0.861	19.10

Table 5.5: Results of the different nodule re-identification pipelines. The meaning of the configured methods is detailed in Table5.1.

As in the standalone evaluation of our method, we also conducted some experiments with the best pipeline (FIFB) to investigate nodule re-identification performance in terms of nodule growth. Results are shown in Table-B.6 of the supplementary material.

Then, we evaluated the performance of the best pipeline (i.e. the pipeline configured with the FIFB network) for the nodule growth detection task. A correct prediction was achieved when the difference on diameters between predicted and ground truth nodules had both the same sign. In this way, having 32 correctly identified cases (out of 36), we obtained a 0.92 of recall, a 0.88 of precision and a 0.90 of F1-score. The confusion matrix is shown in Figure-5.6.

Additionally, we assessed the precision in the measurement of the nodule growth prediction. Agreement between the predicted and ground-truth nodule growth vectors was assessed with a Bland-Altman [11, 136] plot (Figure-5.7). The mean difference between the two measurements was 0.17 mm with a 95% confidence interval (from -3.35 to 3.70 mm). Predicted and ground-truth nodule growth vectors were not found statistically different on the basis of a 1-sample t-test (p-value = 0.99). Also, we computed the mean absolute error of the predicted nodule growths (1.38 ± 1.17 mm), their mean squared error (3.26 ± 5.30 mm) and its coefficient of determination (r^2=0.71). Finally, Figure-5.8 shows the predicted and real difference of diameters for all CT pairs of the test dataset. To support the interpretation of this figure, we have included the axial slice with major diameter taken at time points T1 and T2 of an illustrative subset of nodules.

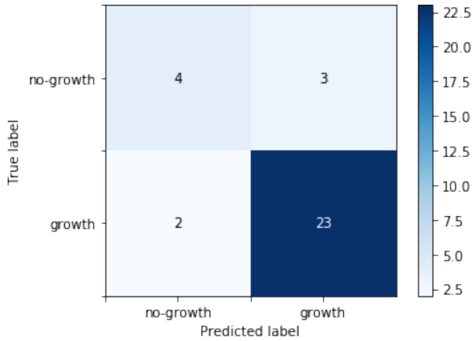


Figure 5.6: Confusion matrix for nodule growth prediction.

5.4.4 Automatic lung CTs registration

We also computed lung nodule re-identification using conventional image registration methods. To do this, we aligned the CT pairs of the VH-Lung dataset, and we computed how far apart were the nodule centroids, annotated by the radiologists, at T2 with the warped locations obtained after applying the transformation-fitted

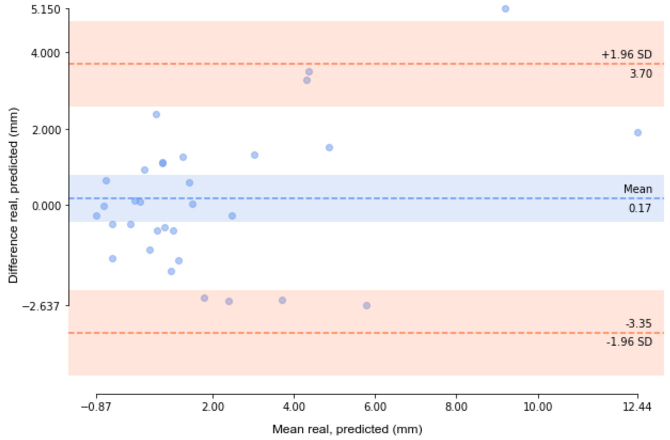
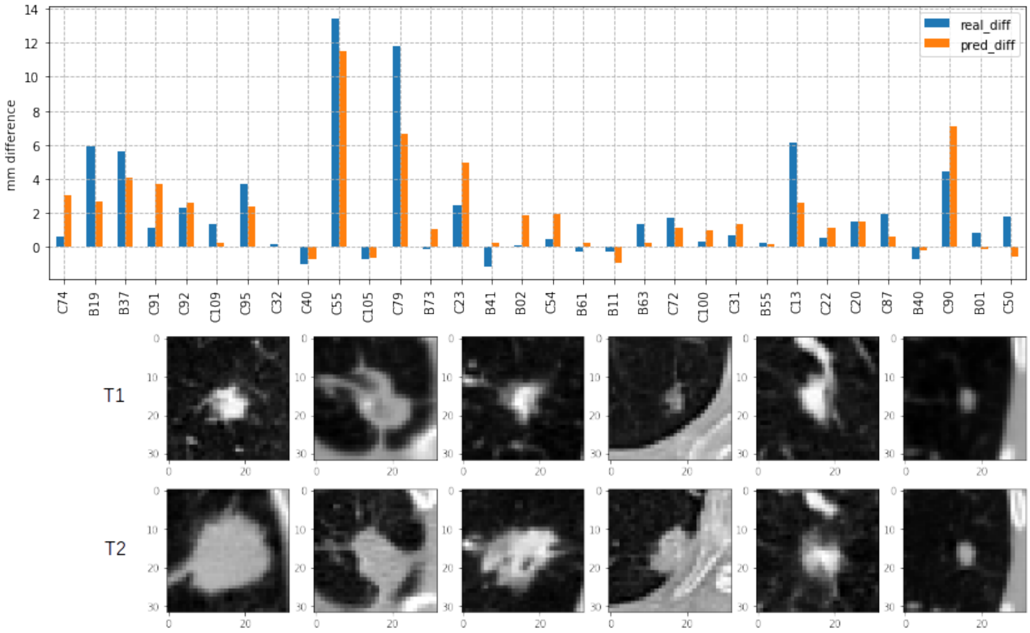


Figure 5.7: Bland−Altman plot for agreement between ground truth and predicted nodule growth.



ID:	C55	C40	C79	B19	C50	B01
Dx:	Malignant	Malignant	Malignant	Benign	Malignant	Benign
Orig_size_T1:	12.5	18.77	7.97	7.5	16.1	7.16
Orig_size_T2:	25.9	17.76	19.75	13.5	17.9	7.95
Date difference:	7m	1y-3m	1y-1m	2y-9m	1y-11m	2y-5m
Label:	Growth	No Growth	Growth	Growth	Growth	Growth
Prediction:	Correct	Correct	Correct	Correct	Error	Error

Figure 5.8: Comparison between real and predicted cases. Upper panel: diameter differences for all test set. Lower panel: axial slices at two time points of different nodules.

function on the nodule centroids at T1. To do this, we used two well-established methods for image alignment, one for rigid and the other for non-rigid registration. Rather than exploring and fine-tuning new registration setups, we leveraged the Elastix [149] database[3] of published registration configurations. This is a publicly-available repository of configurations aimed at promoting research reproducibility. Therefore, for the rigid approach we selected a recent configuration already applied for CT images on [7], and for the non-rigid approach we used an affine registration [234] previously applied for lung CTs.

Table-5.6 shows the nodule re-identification performances obtained for the two registration methods on the train, test and the whole dataset. Correct cases were those in which the Euclidean distances between the location of the centroids at T2 and the warped locations of the centroids at T1 were less than the nodule' radius squared (same threshold as proposed in LUNA-16 challenge). Accuracy was obtained after summing all correct alignments divided by the total of CT pairs in the dataset. We also computed mean absolute errors (MAE) between the ground truth and the warped centroids and the average time required for performing the alignments.

	Rigid			Non-Rigid		
	Accuracy	MAE (mm)	Time (s)	Accuracy	MAE (mm)	Time
Train (113 CT pairs)	0.672	30.8±44.2	52.6±10.0	0.761	23.8±39.7	82.2±12.5
Test (38 CT pairs)	0.684	29.6±38.7	52.9±7.7	0.605	30.2±44.3	82.8±9.5
All (151 pairs)	0.675	29.5±43.0	52.7±9.4	0.721	25.4±41.0	82.3±11.8

Table 5.6: Results after applying automatic registration using rigid and non-rigid approaches.

5.5 Discussion

In this chapter, we provide a novel way to address the nodule re-identification problem. In particular, we propose a deep SNN that can directly re-identify nodules located in a series of pairs of CT scans without the need for any image registration.

The SNN allows matching pulmonary nodules in different CTs in a single stage by outputting a similarity score (i.e. the probability of being the same nodule). In contrast, standard techniques require at least two stages: first registering the image and then identifying matching nodules with some distance function. Moreover, with the proposed solution, no additional deformations/perturbations of the lung scan are performed, so that nodule measurements can be done directly from the image itself. Another advantage is that the re-identification process is fast

---

[3] http://elastix.bigr.nl/wiki

since all weights of the network have already been calculated during the training phase.

We designed and tested several SNN architectures in order to fully understand the complexities of the problem and find the best network configuration. To this end, we collected a longitudinal cohort of two CT scans per patient taken at different time-points. In each of the CT scans of the patients, the most suspicious nodule was annotated according to two different radiologists. Despite the richness of the cohort in terms of heterogeneity in the parameters that affect the image acquisition (e.g. scanners, protocols and setups), in the selected nodules (e.g. size, growth, malignancy), and in the temporal differences between CT studies, the total number of cases to test our approach was limited (38 patients, 25% of the total). Thus, the test set may not be representative enough of the whole nodule spectrum. To mitigate this issue, despite having presented two different evaluation scenarios, more and diverse number of pulmonary nodules (with different morphologies, locations, sizes, growth rates, or degrees of malignancy) are recommended to collect for a more exhaustive validation of the present work.

As previously mentioned, we have provided two different evaluation scenarios with the intention of showing reliability and usefulness of our approach. In the first evaluation scenario, we trained the models with previous localized fixed image patches from 226 CTs pairs (doubling the original training partition with random negative cases) and we evaluated them using 10-fold cross validation as well as with image patches from 76 CT pairs from the independent test partition (doubling the original test partition with random negative cases). Results (Table-5.3) showed that, in general (7 out of 8 experiments), the networks obtained high accuracy scores, above 85% in validation and 80% in test. Indeed, several of the SNN configurations (e.g. FIFB, UCMB) achieved accuracy scores in test above 92%. Also, as shown in Table–5.3 there is no relevant performance gap between training, validation and test sets, which suggests that there is no overfitting.

Regarding the ability to re-identify matching nodules according to their growth (Table-B.5 supplementary material), the best SNN (FIFB) obtained a high accuracy score both in training (99.1%) and in test (97.3%), and no significant differences in performance were found despite their nodule growth rates. However, the performance for identifying non-matching nodules was lower than that of the matching cases. In particular, the performance in training was 96.4% and in test 86.8%. This slight drop in test performance was due to errors for predicting non-matching nodules with moderate (2 out of 6 errors) to large change in size (3 out of 6 errors) between time-points. Beyond growth factor, other visual aspects, such as the density and size of the nodules at T0, were not relevant as they were equally distributed among the 6 mismatched pulmonary nodules in the test set. However, 4 of them were found in the left lung and 2 of them in the superior lobe. Also, 3 of these nodules were attached to blood vessels, 2 were close to or attached to the lung wall, and 1 of them was difficult to distinguish from the surrounding lung tissue at T0, whereas at T1 it was clearly visible. The usual appearance of the edges of these nodules was irregular (4 out of 6).

One of the main factors contributing to the good performance is the use of transfer learning, namely initializing the backbone of the different SNNs with the weights of a previously trained 3D network. This can be noted by the fact that the simplest network configuration (FIBC), which it mainly performs a direct forward-pass mechanism of the input through the network, initialized with the weights of the transferred network, reaches, in our opinion, a considerable performance of 77.5% in validation and 71% in tests.

Regarding the loss functions configured in the different experiments, the methods using the BCE loss (which are based on probabilities) slightly outperformed the ones using the contrastive loss (which is based on distances). This can be seen in the difference in accuracy (3.5% in validation and 12% in test) obtained by the best network configured with probability-based loss function (FIFB) compared to the best network configured with loss function based on distance (UIBC).

Another finding was that unfreezing the weights of the pre-trained networks usually allowed for better performances. This is particularly evident in the UIBC case, which exceeded of almost 10% in validation and testing the corresponding frozen configuration (FIBC). Somehow, this finding was expected as weights were transferred from networks trained in a different, although closely related, domain.

With respect to the features used by the networks, we can observe (Table-5.3) that, in almost all the methods, the best performance was achieved by using features extracted by layer1 and/or layer2, while only for two methods it was achieved using features from layer3 and avgpool (i.e. the global average pooling). This may suggest that features encoding simple patterns (from earlier layers) are preferred for this problem, whereas layers that contains more specific features (from the last layers) are less useful. It is also worth noticing that networks combining features from different layers did not clearly outperform networks using features from a single layer. This is the case of UCMB in which the reported validation performances are just a bit lower (0.2%) than the performances reported by the FIFB configuration, although in the test, UCMB outperformed by 0.4% the performance of FIFB.

Concerning the type of heads with which the networks were configured, the best option was using fully connected layers (FC head). Surprisingly, networks with extra convolution layers before the fully connected layers (CNN head) achieved worse performances (1% and 6% less in validation and test, respectively) than networks with FC heads. This might suggest that adding extra convolution layers to find patterns between locally connected features increases the complexity of the model, leading to more weights to adjust but with the same amount of training data.

In the second evaluation scenario, more ambitious and practical, we integrated the SSNs into automatic pipelines intended first for the detection and re-identification of nodules, and then for the prediction of nodule growth given series of CTs of the same patient. This evaluation was done for both training (113 CT pairs) and testing (38 CT pairs) random partitions of the VH-Lung dataset.

The nodule detector component of the pipeline was configured to provide only the top-32 scored nodule candidates per CT. This threshold was empirically set based on the good balance between precision and recall in terms of nodule detection obtained in both training and test partitions of the VH-Lung dataset. In test, this component reported a nodule detection sensitivity of 97% in 32 nodule candidates (FP) per CT in average. This performance is far from 81.7% sensitivity in 0.125 FP per CT scan in [130] and from the results we obtained when training the nodule detector standalone (0.84 sensitivity with 1 FP, in average) in the LUNA-16 dataset. However, the comparison is not fair since the nodule detector was not trained to find the most questionable nodule per patient according to radiologist but for detecting any nodule in the lungs, that is why more nodule candidates were needed to find the annotated nodules in the VH-Lung dataset.

Regarding the nodule re-identification step of the pipeline, the performances obtained by the different SSNs networks (Table-5.5) were lower than when evaluating the models standalone. This was expected since, as opposed to in training, where the patched images were cropped around the ground truth centroid of the nodules, in the pipeline the patches were cropped around the position predicted by the nodule detector, making its correct matching more difficult if the centroid position was not as precise. However, 5 out of 8 networks reached a nodule matching accuracy score above 80%, and the best network (FIFB) reached an accuracy of 88.8%.

In Table-B.6 (supplementary material), we reported the performance of the different sub-processes of the best pipeline (FIFB) according to the growth of the nodules. Looking at the results, we can highlight that nodule detection and re-identification steps had high performances both in training (>92%, >85%) and testing (>94%, >88%). However, the training performance for growth detection in small nodules dropped down to 47%. This was not the case for moderate and large nodule changes in neither training nor testing. Different interrelated factors may explain this limitation. One reasonable factor could be the different data proportions between training and test set for this type of nodules. A second factor could be the errors in the ground truth annotations. Another factor could be the limitations from the nodule detector when out-coming the diameter for these nodules. More experiments and tests are required to improve this particular case.

Independently of the growth of the nodules, some common visual appearances were found along with the nodules incorrectly re-identified by the pipeline. In particular, from the 2 non-detected nodules at T0, we would highlight that both were solid and difficult to distinguish from the lung parenchyma ($<$ 9 mm of diameter). From the 4 non-re-identified pair of nodules, 3 of them were malignant and greater than 10 mm at T0. Also, they were located on the right lung and close to or attached to the wall of the lung with irregular edges. Among the 5 pairs of nodules with incorrect growth classification, all of them were solid, 4 of them were malignant and 3 had sub-centimeter diameters at T0. Moreover, 3 of them were in the lower right lobe of the lungs, whereas the others were in the upper left lobe. Furthermore, 3 of them were close to the lung wall, 2 had an attached vessel whereas another was close to the mediastinum. Regarding the characteristics of its edges, 2 were irregular and the other 3 smooth.

In terms of computational time, our approach achieved satisfactory performances being able to re-identify the nodules of the complete test set in times ranging from half a minute (in the worst case, UIBC) to less than 10 seconds (for the best configuration, FIFB), as can be seen in Table-5.5. This is a particularly appealing feature of our method, since even the most recent techniques for registration of lung CT images, necessary by any standard pipeline for nodule re-identification, require significantly more time, for instance 5 minutes according to [251] or approximately 1 minute by [335] per case. These processing times fluctuate substantially depending on the technique and the quality of the image registration.

To have a better intuition of the performances obtained using the proposed pipelines for the automatic nodule re-identification problem, we compared them with two conventional methods for lung image registration (Table-5.6). Both registration mechanisms were slower and did not outperform the performances reported by any of the configured pipelines. The accuracy differences using the worst (FIBC) and best (FIFB) pipelines compared with the rigid alignment were between 1% and 20.4%, and with the non-rigid alignment between 8.9% and 28.3%. Despite these differences in performance, more advanced registration techniques and further fine-tunning of its parameters would lead to greater re-identification performances. For example, in [98], the authors compared rigid and non-rigid registration methods for matching 60 diverse nodules in 60 lung CT pairs obtaining average registration errors (Euclidean distances between baseline and follow-up after alignment) between 9.5 and 10 mm. Also, in [138], the authors using a rigid registration along with a rib based adjustment mechanism reported registration errors of 17 ± 7 mm for 69 lung nodules in 50 subjects with series of two CTs.

Compared to the latest CAD systems providing nodule re-identification [155, 295], our method reports similar performances (92% accuracy) when evaluated standalone, but slightly below when integrated in pipelines. A number of factors may explain this difference. First, our approach is fully automated, whereas in those systems the position of the reference nodule, to match with, was given by the radiologists. Second, in those systems the data they used for evaluation was from lung cancer screening population, which makes the underlying lung tissue structure more consistent when compared to patients with lung metastases or from incidental cases like ours. Third, in our study, the total number of patients was more than double the number of patients used in these studies (40 and 53), which makes re-identification more difficult since the similarity of the lung structures between nodules is less plausible. In another related study [138] for lung nodule re-identification, they reported rates from 29% to 100% in 69 nodules from 50 different patients. However, in their experimental dataset, no severe lesions were reported (e.g. 14 nodules had no changes in diameter between corresponding nodules), and their method was evaluated using the entire cohort, making it difficult to know their ability to generalize to new cases.

Although the focus of the paper is the nodule re-identification, we also quantified and assessed nodule growth. To do this, we selected the best network for nodule re-identification (FIFB) and integrated it in the nodule-growth pipeline. In total, nodule growth was correctly detected in 27 cases and erroneously in 5 cases. However, only 2 of these errors were false negatives (that is, the pipeline failed to predict growth); one of them was on a benign nodule (B01) with growth difference of less than 1 mm, whereas the other was on a malignant nodule (C50) with growth difference of 1.8 mm. As shown in Figure-5.7, there is an agreement when comparing predicted and real nodule growths as most of the measures fall between the two standard deviations of the mean, there is a non-significant difference between them (p=0.99), and they show a good correlation score (r^2=0.71). Despite this positive results, the values obtained for the 95% limits of agreement ($>$ 3 mm) are still high. This was somehow expected as quantifying lung nodules is complex and subject to multiple variability factors [174] (e.g. slice thickness, reconstruction kernel algorithms, attachment of vessels, patient inspiration depth). An example of this was shown in a previous study [61], in which up to six different open software packages measured the volumetry of solid lung nodules, and reported large nodule inter-variabilities (from 16.4% to 22.3%) on repeated CTs of the same patient in a cohort of 20 patients.

In our case, as we can see in the BA plot (Figure-5.7), the cases that experiment higher disagreements are those nodules with larger mean nodule growth (i.e. observations located in the right part). A reason that could explain it is that the nodule detector (which reports the nodule diameter) was trained in a database (LUNA-16) with a smaller nodule size distribution (8.30 ± 4.75 mm) than the one used for the evaluation of the pipeline (VH-Lung dataset with 12.45 ± 4.32 mm). Alternatives to address this issue could range from gathering more annotated data, increasing the distribution of large nodules by applying further data augmentation, implementing more sophisticated mechanisms (e.g. attention networks [256]) in the nodule detector, or instead of using the predicted diameter and centroid of the nodule detector, implementing deep nodule segmentation networks.

From a clinical point of view, the majority of the nodule differences were correctly classified (growth, no-growth) as shown in Figure-5.6. Indeed, we reported a mean absolute error of 1.38 ± 1.17 mm in diameter with respect to the ground truth, which is slightly less than the 1.73 and 2.2 mm of the variability error reported in different retrospective analysis [146, 245] measuring changes in solid and subsolid nodules ($<$2 cm) using only their diameter.

This study, however, is subject to several limitations. First, the limited number of cases to build our models. In the medical domain, longitudinal data is scarce, and much more complex to collect and manage than single time-point studies. Specially for lung cancer assessment, gathering large quantities of samples is even more difficult for different reasons. First, the disease in the early stages is asymptomatic and very aggressive, so when patients are explored, their pulmonary nodules often have clear signs of malignity, and radiologists do not require further studies for its diagnosis. Second, data is usually incomplete or missing, which suppose a real challenge in evolutionary studies. Although there are different initiatives that aim to screen large populations at risk (e.g. NLST), the access to these assets is not publicly open. Thus, having an insufficiently large dataset can negatively impact the performance of deep learning-based models. This is even more concerning for re-identification of lung nodules, since for each patient, twice as many images and annotations are needed. Another main limitation of the study is that the only expert annotation provided for nodule quantification was the major axial diameter. Although the diameter is the most common radiological measure used in practice for nodule growth assessment, using 3D measurements could lead to a more accurate quantification. In addition, if we had had nodule measurements from more experts, we could have better explained the clinical variability, reporting more accurately the performance of our pipeline with respect to nodule growth prediction. Another limitation of our method could be on re-identifying structures with strong size variations. Some actions may be done to amend this aspect. First, retraining the model with larger input patch sizes. Second, making further data augmentation especially on image pairs with large size variation or collecting more cases of this typology. However, according to radiologists' recommendations, clinical guidelines [12], and literature [165], the challenge is to provide automatic support for growth detection at small/medium nodule change sizes, since larger nodules are easier to identify and substantial differences in growth ratio indicate a clear symptom of either malignancy [270] or benignity [102]. Finally, in this work, we focus on training and evaluating several SNNs to explore different configurations. Finer tuning of hyperparameters (e.g. the learning rates, batch sizes or dropout values) may lead to improved results.

Nevertheless, the automated re-identification of regions of interest in medical images over time, without the need to warp the inherent image structure, could be an appealing application beyond lung cancer assessment such as therapy follow-up as well as for different diseases located at different organs (e.g. prostate, breast cancer) in the body.

Several future works have been described in the paper, and some others are envisaged to extend the research presented in this paper. For example, it would be interesting to longitudinally evaluate the pipeline for more than one nodule per patient, or exploring the nodule spatial localization for the re-identification problem. Also, applying different feature fusion techniques, introducing different manners to weigh the feature maps, applying new techniques to reduce the dimensionality of the problem, as well as the use of segmentation could be some other research lines that would be worth exploring beyond this paper.

5.6 Conclusions

In this paper, we address the problem of automatic re-identification of pulmonary nodules in lung cancer follow-up studies, using siamese neural networks (SNNs) to rank similarity between nodules, which overpasses the need of image registration. This change of paradigm avoids possible image disturbances and provides computationally faster results. Different configurations of the conventional SNN were examined, ranging from the application of transfer learning, using different loss functions, to the combination of several feature maps of different network levels. The best results during the off-line training of the SNNs reached accuracies (0.89 in cross-validation and 0.92 in test) similar to those reported by state-of-the-art registration mechanisms. Finally, we embedded the best SNN into a two-stage nodule growth detection pipeline. Nodule re-identification results reported by the pipeline in an independent test set were fast ($<$10 seconds, matching 38 pairs of CTs) and precise (0.88 accuracy score). Nodule growth predictions were also accurate (0.92 sensitivity score), and both the predicted, and the ground truth measurements were not significantly different (p=0.99).

Chapter 6

END-TO-END AUTOMATIC PIPELINE FOR PULMONARY NODULE FOLLOW-UP ASSESSMENT

6.1 Introduction

The use of computed tomography (CT) scan images has increased dramatically over the last decades, becoming a crucial tool for the diagnosis and follow-up of malignant lung tumours [192, 230]. Radiologists are able to detect, measure and monitor the evolution of abnormal tissues in their lungs by visually inspecting CT scans of the patient's chest. However, tumours, specially at early stages, are complex to detect and diagnose due to large heterogeneity in their morphology, size, texture, localization and growth rates [27]. Moreover, spatial resolution in computerized axial tomography images is often limited by the acquisition protocol [24]. This leads to some ambiguities and conflicts for radiologists when having to determine the next study, whether to discharge the patient from the follow-up, or whether to resolve a clinical intervention for the patient [107]. Therefore, the experience and expertise of physicians is fundamental for the early diagnosis and

---

The work described in this chapter is based on the following publications: Rafael-Palou X, Aubanell A, Ceresa M, Piella G, Ribas V, González Ballester MA. Pulmonary Nodule Malignancy Classification Using its Temporal Evolution with Two-Stream 3D Convolutional Neural Networks. Medical Image with Deep Learning (MIDL), 2020.
Rafael-Palou X, Aubanell A, Ceresa M, Piella G, Ribas V, González Ballester MA. Detection, growth quantification and malignancy prediction of pulmonary nodules using deep convolutional networks in follow-up CT scans. To appear in Artificial Intelligence in Cancer Diagnosis: Lung and Kidney Cancer.

prognosis of lung cancer. Unfortunately, the aggressive nature of this disease, its important incidence in the adult population, and the constant need for specialized professionals, make it necessary to have accurate and efficient tools to reduce the workload of clinicians as well as to help them in making critical decisions.

The idea of providing automatic support for the detection and diagnosis of lung cancer is not new, and large efforts have been made with conventional machine learning and artificial intelligence techniques [54, 141, 285]. Recently, the advent of deep neural networks [167] has allowed a major breakthrough in the medical image domain [31, 78, 100]. Specifically for lung cancer, outstanding performances have been achieved in a very short period of time, outperforming conventional approaches for nodule detection [262], pixel segmentation [200], or lung cancer classification [51]. Despite this, most of the research focuses primarily on a single CT scan. This fact conditions the potential of these contributions, since they do not consider the temporal evolution of the tumour, which, indeed, is one of the most important clinical factors influencing prognosis [165].

In this chapter, we take a step forward in supporting the radiological workflow, by proposing an automatic tool that takes into account the evolution of the pulmonary nodules in the predictive modelling task. To do this, we defined a data-driven approach with a flexible and configurable four-stage pipeline, which 1) automatically detects nodules, 2) re-identifies them from different CT scans of a given patient, 3) quantifies their growth, and 4) predicts their malignancy. To configure each of the pipeline components, we have integrated existing solutions [37, 236, 237] and proposed new ones based on deep convolutional neural networks. Hence, in the remainder of this chapter we describe the background, present the pipeline and its different components, and show the results of its evaluation in a longitudinal cohort of more than 30 patients. We conclude this work by discussing the present solution and establishing future works for the automatic temporal lung nodule assessment.

6.2 Background

In this section, we review some of the most relevant and recent works proposed for supporting radiologists in the lung cancer assessment. From the different tasks encompassed by radiologists in the management of this disease, we focus on the most essential ones, such as nodule detection, nodule quantification, and lung cancer prediction.

6.2.1 Nodule detection

This task consists on screening the entire lung CT volume, searching for small suspicious regions or nodules (usually between 3 mm to 30 mm) [270]. Nowadays, this problem, as in most of the computer vision research areas, is addressed by convolutional neural networks (CNN) [54] able to extract, without human intervention, accurate feature image representations thanks to their shared-weights architecture and translation invariance characteristics. A common approach for automatic nodule detection consists on dividing the problem in two steps [68, 321]: candidate detection and false positive reduction. In the first stage, 2D region proposal networks, such as faster region-based networks (Faster-RCNN) [244], are used to extract suspicious regions of interest from the whole CT scan. In the second stage, these regions are classified as normal tissues or nodules using 3D CNN networks, in which the input is 3D image patches around the centre of the nodules. Other recent approaches directly address this problem in a single step [159, 178, 334]. They re-adapt region proposal networks with 3D deeper architectures (such as ResNet [116] or DenseNet [128]) to directly predict 3D bounding boxes surrounding the nodules.

6.2.2 Nodule quantification

Another important task for lung cancer assessment is determining the size of the nodule. Currently, radiologists calculate the size of the nodule by visual inspection on the CT scan, locating and measuring the largest diameter (in mm) [195]. Usually this measure is extrapolated to 3D dimensions, by means of mathematical operations [257], to approximate the volume of the tumour. Although this process is simple and fast, it entails significant intra and inter-observer variability in the size of the nodule, which can go up to 3 mm in diameter [106]. Since this variability may negatively impact the disease management, several deep learning solutions have addressed nodule size measurement to support clinicians. Some works [178, 334] propose learning the diameter of the tumour by extending the nodule detection network (either in 2D or 3D) with a new output in the network. Other solutions build semantic segmentation networks to automatically determine the pixels of the nodules, from which the diameter or volume can later be extracted. One of the most common successful architectures for segmentation is the U-Net [248]. This type of networks uses a convolutional encoder and decoder backbone, tied at different levels by short-cuts, which allow by-passing high level features of the encoder to the decoder, in order to enhance the image reconstruction task. Several extensions of this architecture can be found, such as its 3D formulation [50] or the incorporation of ResNet-like blocks and a Dice-based loss layer, more suitable for segmentation tasks [202]. A more recent approach, nnU-

Net [133] has been successfully applied to a multitude of medical segmentation problems (including pulmonary nodule segmentation). One of the benefits of this approach is the automatic fine-tuning of several configuration parameters to the particular type of images to be segmented.

Despite the high performances reported by U-Net-like networks, they address the segmentation problem from a deterministic point of view. However, due to the inherent ambiguity of the problem (often contours of the nodules are not clearly delimited), it is desirable reporting network uncertainty estimates when predicting the size of the nodules. One way to learn model uncertainty is moving from one-input one-output to one-input multiple-output networks. This change of paradigm has already been tackled in deep neural networks through different approaches. One of the simplest approximations consists in ensembling multiple networks in order to provide multiple opinions [164]. Another approach consists in enabling dropout [279] at inference time in order to provide independent pixel-wise probabilities [142]. Another approximation is by deep generative networks, such as generative adversarial networks [96]. This type of networks try to learn, in an unsupervised manner, a direct mapping from a random noise to an output image. To do this, a generator network creates new valid images (from the random noise) with the intention to fool a discriminator network that evaluates whether an image is valid or fake. An extension of this type of networks are conditional GANs [134], in which the goal is to learn structured outputs conditioned on an input image. To do this, the discriminator receives as input the target image to which conditioning the generator. Similar to cGANs, we can find the conditional variational autoencoders (CVAE) [275]. This type of networks propose learning a multi-dimensional latent space that encodes all possible output images. During training, the latent space distribution defined by the encoder is approximated to a normal distribution to ensure continuity and avoid ‘mode collapse’ commonly seen in GAN approaches [94]. Also, the random vector sampled from the latent space, together with the target image are passed to the decoder (only during training) in order to generate a new plausible image. A recent work, hierarchical probabilistic U-Net (HPU) [154], has been proposed to cover the gap between the generative ability of producing new structured images of the CVAE, with the accuracy of segmenting images of the U-Net. To do this, during training, a posterior U-Net like network, conditioned on the radiologist ground truth nodule, is added to transfer the latent features to a prior U-Net like network by injection, at different levels of the decoder part of this network.

6.2.3 Lung cancer prediction

To support radiologists in the lung cancer prediction, several works have been proposed relying on 2D and 3D inputs, using different deep learning architectures (e.g. CNN, RNN) [44, 66, 282], but mostly relying on single CT scan images (commonly derived from the LIDC dataset [19]). Therefore, very few deep learning works have addressed the temporal evolution of pulmonary nodules to support the clinical decision-making. In [14], an end-to-end deep learning based pipeline was presented for lung cancer prediction using two CT studies per patient (current and previous year). This approach proposes three 3D CNN networks, one for analysing the lung CT image, another for analysing nodule patches, and a final one, to provide cancer risk prediction using outcomes from previous two components. In [129], a deep learning approach is proposed for predicting lung cancer risk at 3 years and lung cancer-specific mortality. This study, although not being focused on automatic image analysis, uses a multilayer perceptron to ensemble nodule and non-nodule features associated to lung abnormalities.

6.3 Method

Our pipeline takes as input two images from the same patient at different time-points, identifies the lung nodules, and estimates their malignancy and growth. The pipeline consists of 4 main components (see Figure-6.1): 1) nodule detection, which is done independently on each image; 2) nodule re-identification, which finds the correspondence between nodules across time points; 3) nodule malignancy classification; and 4) nodule growth quantification.

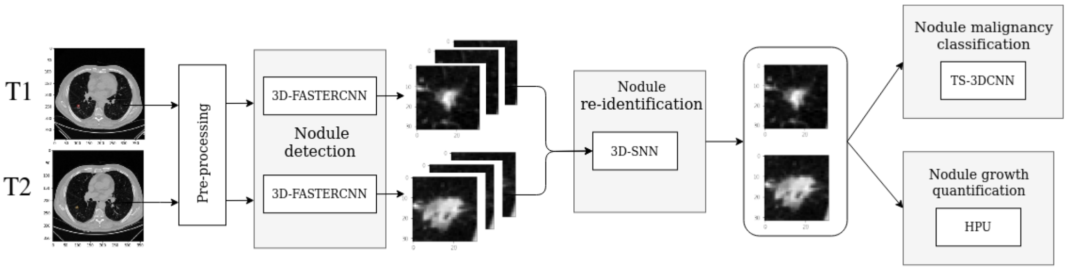


Figure 6.1: Pipeline architecture for the temporal analysis of lung nodules.

Pre-processing

Lung CT images are usually originated by different scanners and at different image resolutions. Therefore, the pipeline makes an initial pre-processing step with the intention to standardize the input images. Precisely, first, the images (T1 and T2) are resampled to an isotropic resolution of 1x1x1 mm3. Second, the image pixel intensities are clipped between [-1000, 600] Hounsfield Units to filter out non-tissue related regions. Finally, the pixels are normalized between 0 and 1.

Nodule detection

Both pre-processed lung CT scan images, without any previous lung segmentation, are separately analysed to find possible nodules. Thus, each lung image is split in overlapping patches (due to memory constraints) of [128x128x128] with an overlap of 32 pixels per dimension. Since the location of the patch may influence the decision of whether it is a nodule, we also compute the location information of the patch with respect to the whole lung image as in [178], and we send it all together to the nodule detection network.

The nodule detection network (Figure-6.2) was developed and tested in our previous work [237], and consists of a 3D Faster-RCNN [244] using as a backbone a U-Net like framework. The input of the network is a lung patch image of [128x128x128]. The location information of this patch is concatenated in the decoder part of the network. The output of the network is the location of the nodules (x, y, z coordinates), the diameter, and a probability of being a nodule. Once all patches of the lung CT are analysed by the network, those are resembled (due to overlapping areas between patches) and clear out the repeated findings.

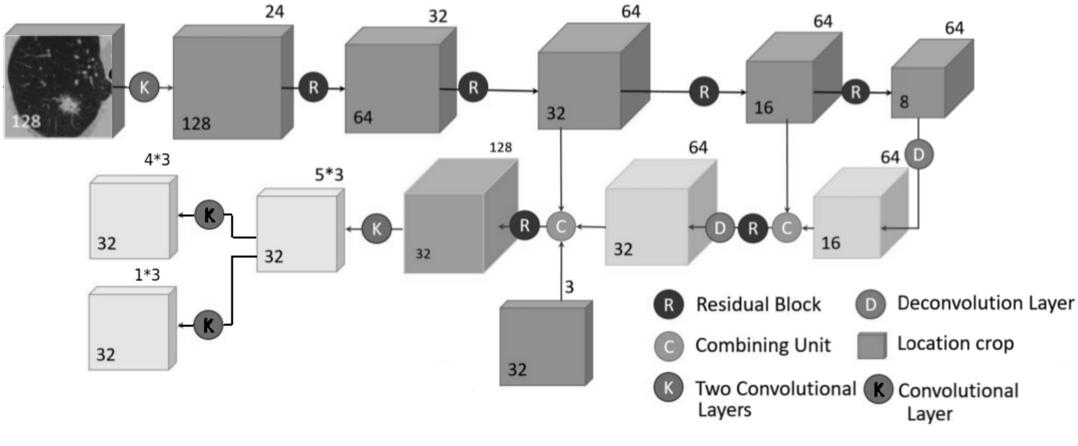


Figure 6.2: Architecture proposed for the nodule detection network of the pipeline.

For further information regarding the configuration parameters for training the 3D Faster-RCNN, we refer the reader to the supplementary material (B.2).

Nodule re-identification

Once the nodules from two different CT scans of the same patient are detected, a second component automatically matches or re-identifies these nodules using a 3D Siamese neural network (3D-SNN) as presented in our previous work [237]. A SNN [152] is made up of two components: feature extraction and classification. In the first, two subnetworks (with shared architecture and weights) process a pair of images at a time to produce two embedding feature vectors directly from the images. In the second, a head network determines whether the two embedding feature arrays are similar (i.e., correspond to the same nodule).

From the different re-identification network setups of our previous work [237], we used the one that obtained the best results (i.e. FIFB). This setup (Figure-6.3) consisted on freezing the sibling networks of the feature extraction component with the weights of a pre-trained network, initially built for nodule identification [37]. From the different convolution blocks of the pre-trained network, we used the output from the first block as input for the head component of the 3D-SNN, since they reported the best performances. The classification head component was configured with a L1-pairwise distance, a flattening layer, and a fully connected (FC) block, comprising an FC layer (with 64 units), a batch norm, a ReLU, a dropout layer and a final FC layer (with one unit).

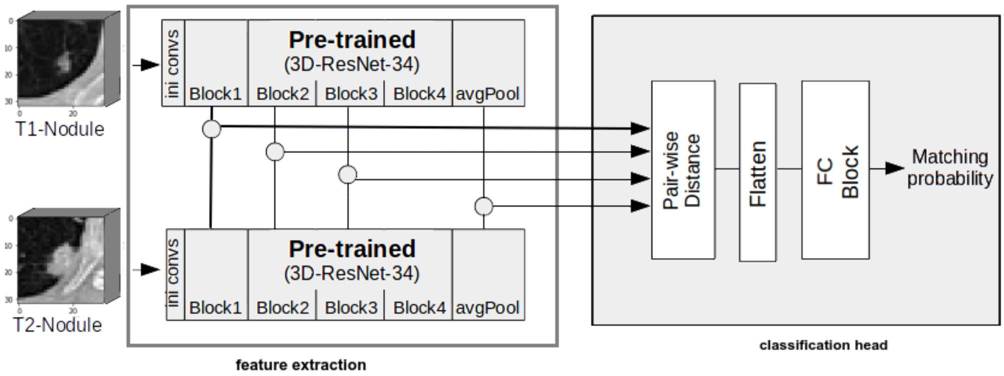


Figure 6.3: Architecture of the 3D-SNN for the nodule re-identification component of the pipeline.

For further information regarding the configuration parameters used for training the 3D-SNNs, we refer the reader to the subsection 5.3.1 of this thesis.

Nodule growth quantification

The set of re-identified nodules are analysed by the nodule growth quantification pipeline component. A couple of methods are proposed for this component. The first one, already used in [237], consists on computing the diameter difference of the paired nodules from the re-identification component. The diameters of both nodules are taken from the nodule detection component of the pipeline.

The second approach differs from the previous one in that we use a probabilistic generative network (HPU) [154] to provide not only nodule growth, but also the uncertainty associated with such prediction. This network (Figure-6.4) is composed by two sub-networks, the prior, which models the prior distribution of possible segmentation maps for a given input image T_i, and the posterior, which models the joint probability distribution of the input image T_i and its annotated segmentation S_i.

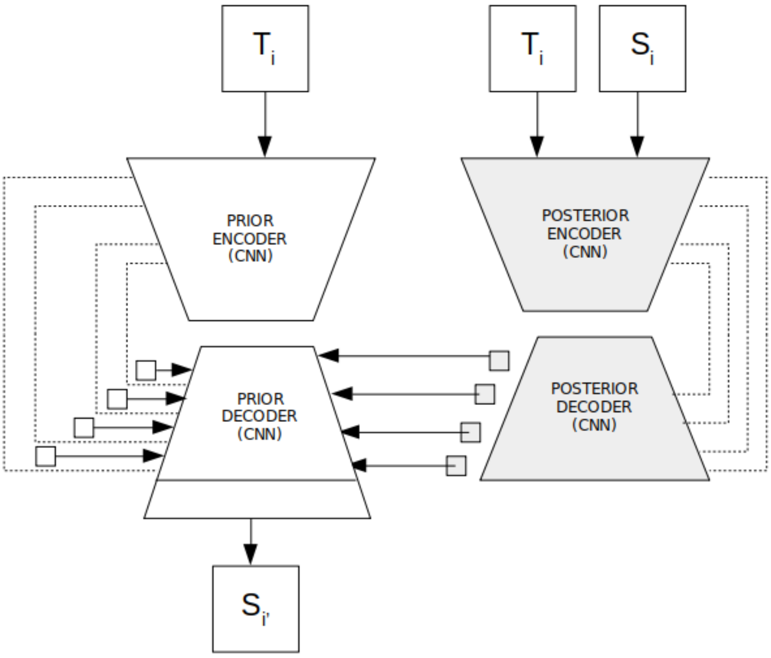


Figure 6.4: Hierarchical probabilistic U-Net network architecture overview. T_i is the nodule image i, S_i is the ground truth segmentation of the nodule T_i, and S_i' the predicted segmentation of the nodule T_i.

During the training, both networks learn, in parallel, to adapt their latent distributions to be able to generate consistent segmentations. To do this, the different latent features, defined at different levels of the decoder of the posterior network, are injected to the corresponding following layer from the decoder of the prior network. In this way, gradients can flow through both networks using any stochastic gradient descent-based optimization algorithm. At inference time, a random sample z_i from the distribution defined by the different latent features, located at different levels of the decoder of the prior network, are injected into the following layer of this same network to output a new segmentation S_i'.

Following the same notations as in the original paper [154], the loss function used for training this network, also named as ELBO, is composed by the sum of the cross-entropy loss (below formulated as P_c) between the segmentation ground truth Y and the predicted segmentation S, given an input X and a sample z, and the distance D_{KL} (Kullback-Leibler divergence) between the prior and the posterior distributions. Since it is a non-deterministic network, the authors [154] proposed to evaluate this network with the generalized energy distance (GED^2), a metric to account for quality of the segmentation and variability in generating segmentations, according to the variability in the ground truths:

$$D^2_{GED}(P_{gt}, P_{out}) = 2\mathbb{E}\left[d(S,Y)\right] - \mathbb{E}\left[d(S,S')\right] - \mathbb{E}\left[d(Y,Y')\right],$$

where d is a distance measure (in our case, $1 - IoU$), S and S' are independent segmentations from the predicted distribution P_{out}, Y and Y' are independent segmentations from the ground truth distribution P_{gt}.

In the original paper, the HPU network was already trained for segmenting lung nodules using data from the LIDC dataset [19], which contains nodule segmentation annotations from up to four radiologists. However, when no nodule was marked by a radiologist, an empty segmentation image was used. This made the network to model as well the probability that there is no nodule. For our particular settings, this was undesired, as the nodule detector already filters out non-nodule cases. Thus, we retrained the HPU network using the same specifications and architecture as in the original paper, but omitting empty segmentation cases. Then, we use the HPU (prior) network to estimate the nodule growth and a measure of dispersion (standard deviation). To do this, we run N times (N=1000) the HPU network for the nodule at T1, obtaining N segmentations. From each of these nodule segmentations, we extract the major diameter, obtaining a random vector of N diameters. We repeat this same process, but for the nodule at T2. Then, we obtain the mean diameter growth as the mean difference of both random diameter vectors, and the standard deviation, as the squared root of the sum of the variances of the difference of both random diameter vectors.

Nodule malignancy classification

The re-identified nodules are also analysed by the nodule malignancy pipeline component. To provide nodule malignancy we propose three different approaches, one using nodule malignancies annotated by radiologists (i.e. not confirmed cases of cancer) from a single time-point image, and the other two using nodule malignancies confirmed by diagnosis (either from biopsy or without significant growth increase during at least 2 years) with two patches of the same nodule taken at two different time-points.

The first approach consists on re-using the best network to quantify nodule malignancy from [37]. This network (3D-CNN-MAL) receives as input a single volumetric nodule of 32x32x32. The network has a tailored architecture composed of 4 blocks of 3D CNNs, interleaved with dropout and a final dense layer with a softmax layer at the end. The outputs of this network are 3 probabilities corresponding to 3 categories of nodule malignancy (benign, suspicious and malignant). Further details regarding the configuration parameters for training this network can be found at section 4.3.2 of this thesis.

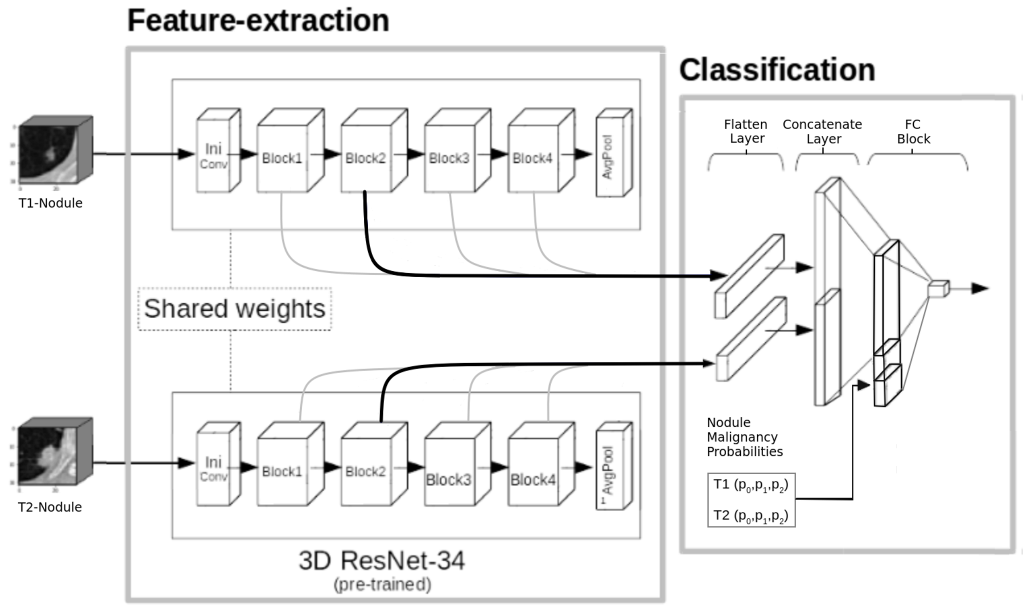


Figure 6.5: Nodule malignancy classification architecture of the TS-3DCNN-MAL network.

The second approach uses our method presented in [236]. In particular, this approach analyses, at a time, two volumetric input patch images of 32x32x32 centred around the nodule. The two patches correspond to the two CT scans made on the same patient, but at different time-points. The network is a two-stream 3D CNN, in which two feature extraction sub-networks, with the same architecture and weights, analyse in parallel the nodule patches, while the classification network part provides a cancer probability risk. Given the limited amount of longitudinal data, the siblings of the TS-3DCNN were transferred from a pre-trained 3D ResNet-34 network, used for identifying pulmonary nodules. We used the features from the last layer of the second block of the 3D ResNet-34, as the ones that reported better performances in our previous work [237]. The classification head component of the TS-3DCNN was configured with a flattening, a concatenation, and an FC block layer comprising an FC layer (with 64 units), a batch norm, a ReLU, a dropout and a final FC layer (with one unit). Figure-5 shows the architecture of the TS-3DCNN network. The third approach builds upon the integration of the other two previous approaches. Given a new nodule to classify, we predict the malignancy using 3D-CNN-MAL network and integrate this information in the TS-3DCNN network. Precisely, 6 extra features (corresponding to the 3 outcomes of the 3D-CNN-MAL for each time-point nodule) were concatenated with the features of the last fully connected layer of the TS-3DCNN.

To allow a fair comparison between these two last approaches, we defined the same initial training settings. Thus, binary cross-entropy was set as the loss function, the number of epochs was set to 150, the learning rate to 1e-4, the batch size to 32, dropout to 0.3, the early stopping strategy to 10 epochs without improvement of the validation loss, and Adam was used for optimization. Moreover, random rotation and flip were applied for data augmentation.

6.4 Results

In this section, we report the performance of the pipeline obtained on the VHLung test set, broken down into each of its components.

Nodule detection

First, we evaluate the ability of the pipeline to detect the annotated nodules (one per each CT) among all nodules predicted by the 3D-FasterCNN network. Therefore, we measured the performance of the network to find the annotated nodules in the least number of predicted nodule candidates. Results show that, taking a reasonable threshold of top-32 predicted nodule candidates per CT, the pipeline obtained a sensitivity score of 0.973 on the 76 CT scans of the test set (taken individually), missing only 2 of them.

Nodule re-identification

To evaluate the second stage of the pipeline, we used the 36 pairs of CT scans where the pipeline found the radiologist's annotated nodules. For each pair of CTs, we input 64 (32 per each time-point) nodule candidates into the 3D-SNN network to obtain those that correspond. In total, the 3D-SNN network reported only 4 CT-pairs incorrectly matched, with an accuracy of 0.888. Table-1 provides a summary of the results for the re-identification step, stratifying them by the initial size of the nodules.

	Small	Medium	Large	Total
Accuracy	1.0	0.84	0.75	0.88
#Nodule-pairs	13	19	4	36

Table 6.1: Performance of the re-identification component of the pipeline.

Nodule growth quantification

Two different approaches for nodule growth quantification were evaluated on the 32 matching nodules obtained from the nodule re-identification component. The first approach used the predicted nodule diameter measurements from the 3D-FasterCNN network, while the second used the predicted nodule diameter measurements from the HPU network. For a proper usage of the HPU in the context of the nodule growth quantification, we retrain this network according to [154] with same data from LIDC dataset but omitting 'empty' cases where radiologists did not mark any nodule in the axial slices. The model reported a GED^2 of 0.38 and a reconstruction Dice of 0.91.

Table-6.2 shows the mean absolute error, mean squared error and r-coefficient of correlation with respect to the ground truth. Results show the mean and 2 standard error associated with 95% of confidence, obtained with 1000 bootstraps with replacement of the test set. Figure-6.6 shows estimated nodule growth sizes from both networks per each nodule of the test set.

	MAE	MSE	R^2
3D-FasterCNN	1.400 ± 0.422	3.333 ± 1.927	0.637 ± 0.344
HPU	1.348 ± 0.370	2.889 ± 1.561	0.667 ± 0.418

Table 6.2: Nodule growth performance comparison between 3D-FasterCNN and HPU networks.

These results show that the HPU network (segmentation based approach) provides closer estimates to radiologists annotations than the 3D-FasterCNN network.

Nodule malignancy classification

Three different methods (3DCNN-MAL, TS-3DCNN and TS-3DCNN-MAL) were evaluated on the resulting 32 matching lung nodules (65% of them cancerous) from the re-identification step. For the evaluation of these methods, we directly used the 3DCNN-MAL classifier on the evaluation cases, while for the other two classifiers, to avoid data leakage for this evaluation, we retrained them before being evaluated, using the training partition of the VHLung with a 10-fold cross-

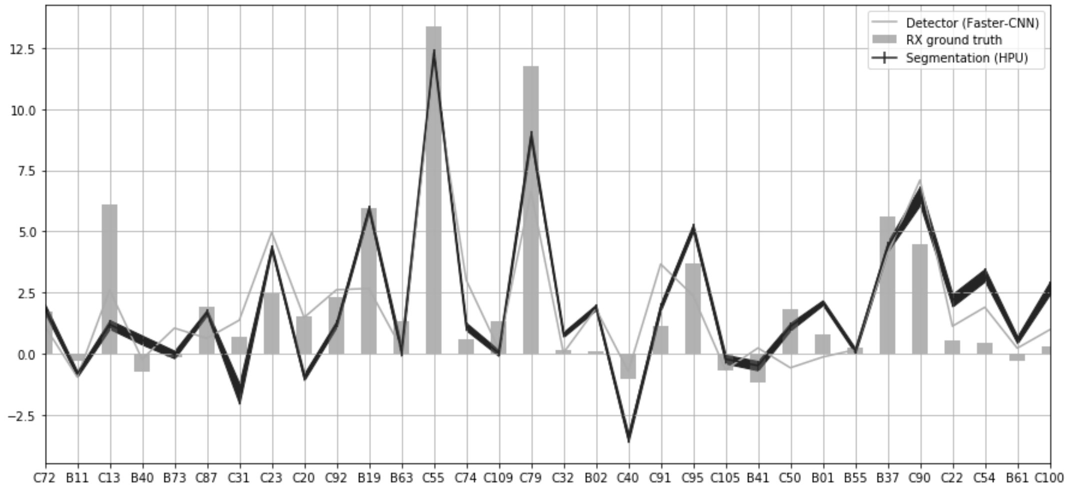


Figure 6.6: Comparative of radiologist growth measurements with results from nodule detector and nodule segmentation.

validation. Particularly for the 3DCNN-MAL classifier, as it outputs 3 probabilities, we assumed cancer prediction when this classifier reported as maximum probability either the category suspicious or malignancy. Table-6.3 shows the resulting classification performances for these models on the 32 matching nodules. We computed, precision (PREC), recall (REC) and specificity (SPEC), as well as balanced accuracy (BA). Due to the data was unbalanced towards the cancer case, the BA was used as the reference metric. Results show for each of the metrics the mean and 2 standard error associated with 95% of confidence, obtained with 1000 bootstraps with replacement of the matching nodules.

	Bacc	Prec	Rec	Spec
3DCNN-MAL	0.776+/-0.153	0.808+/-0.152	1.0+/-0.0	0.552+/-0.307
TS-3DCNN	0.810+/-0.203	0.899+/-0.167	0.791+/-0.309	0.829+/-0.296
TS-3DCNN-MAL	0.825+/-0.201	0.910+/-0.179	0.821+/-0.330	0.830+/-0.348

Table 6.3: Performance of the different nodule malignancy classifiers of component of the pipeline.

The 3DCNN-MAL classifier obtained a balanced accuracy score of 0.77, while the TS-3DCNN achieved a 0.81. However, the TS-3DCNN-MAL, which integrated the outcomes of the 3DCNN-MAL model, improved the balanced accuracy score of the TS-3DCNN model, a 1.5%. For further comparison of these models, we show Figure-6.7 with the ROC-curves of the two best models.

For a further intuition and visual interpretability of the areas of the images that the TS-3DCNN-MAL network took most seriously in deciding which class to assign to the image, we extracted the Grad-CAMs features for this classifier [258].

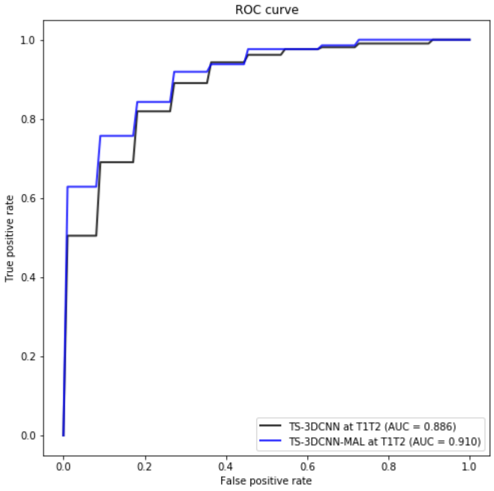


Figure 6.7: ROC curves of the nodule malignancy classification models.

In particular, on the last layer of the first block of the TS-3DCNN-MAL network we obtained the gradients and the feature activations, and they were multiplied after being pooled on the channel dimension. Figure-6.8 shows the results of this visualization technique on the three-planes of different lung nodules (either benign as malign).

6.5 Discussion

The still incomplete knowledge of malignant patterns in the course of multifactorial diseases such as lung cancer makes it necessary to support physicians with automatic, fast and reliable predictive tools to reduce the workload in radiological services. Unfortunately, the vast majority of research is still focused on specific tasks with data from single time-points [36, 44, 66, 159, 178, 282, 334], which limits their potential impact and usability in real clinical settings.

In this chapter, we presented a computer vision pipeline aimed at automatizing the main tasks involved in the lung cancer follow-up. To do this, we relied on a deep learning solution aiming at modelling the temporal evolution of this disease to assess nodule growth and malignancy. The pipeline was formed by 4 different components: nodules detection, nodule re-identification, nodule growth and nodule malignancy classification. The evaluation of the pipeline was done on an independent test set composed by 38 CT pairs. To train the models, we used two different training datasets, LIDC [19] for nodule detection and nodule growth estimation, and VHLung for nodule re-identification and cancer classification.

The evaluation of the first component of the pipeline, aimed at detecting the

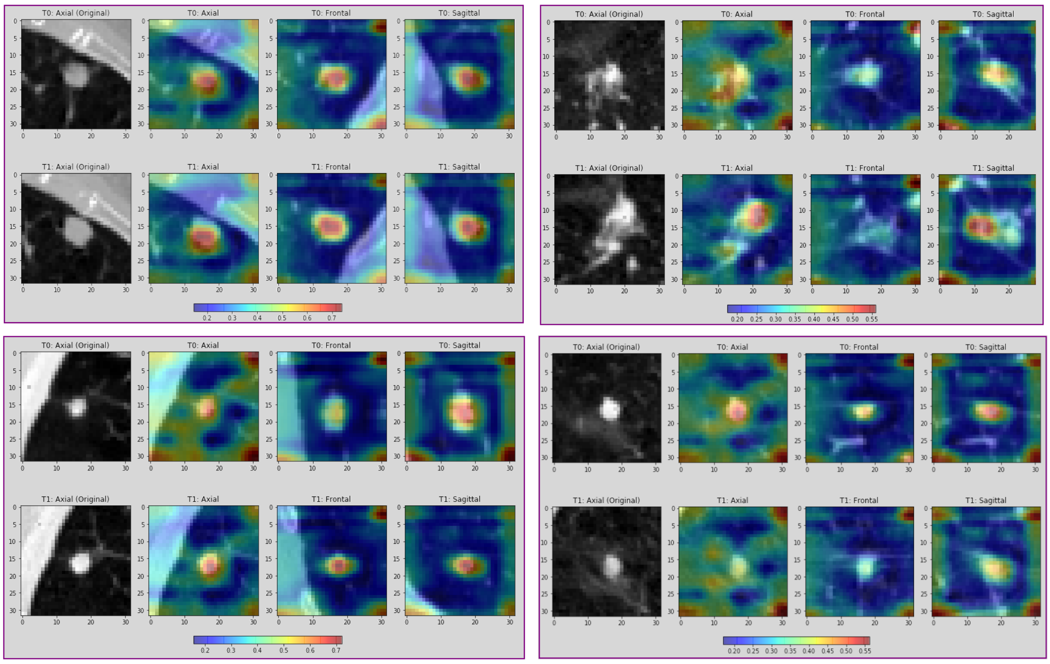


Figure 6.8: Grad-CAM features from the TS-3DCNN-TL-MAL network for 2 malign (upper row) and benign (bottom row) nodules.

suspicious nodules marked by the radiologists. To this end, we re-used a deep convolutional network based on the 3D Faster-CNN scheme already published in a previous work [237]. This network reported on the available data for the evaluation of the pipeline (test set of the VHLung dataset) a sensitivity score of 97.6% for 32 nodule candidates. When evaluating this model in a larger test set (e.g. LIDC test set), the performance decreased to 84% sensitivity at 1 false positive. This value is slightly below compared with top performances (81.7% at 0.125 FP) in LUNA-161 benchmarks. However, it is not clear if those performances are realistic or just an overfit on the provided dataset.

For the nodule re-identification component of the pipeline, we re-used the best model reported in [237]. One of the main benefits of using this approach for matching nodules is that no previous registration of the lung CTs was required (which usually it is slow and introduce artefacts in the original images). The results for nodule re-identification reported also high performances (88.8% of accuracy). However, we should note that they are still a bit below the performances reported, by the same method, when performing in an isolated way, the matching task (92% of accuracy).

Two different approaches were proposed and compared for the nodule growth component. The first one relied directly on the predicted nodule diameters reported by the model of the nodule detection component of the pipeline. Therefore, the growth was computed from the subtraction of both measurements. The second approach for nodule growth estimation relied on a hierarchical probabilistic U-Net (HPU) [154]. This method, based on the generation of several feasible segmentations of the nodule, allowed us to provide an estimation of the growth of the nodule together with an uncertainty of the reliability of the model on this measure. This method obtained the best result, specifically, a mean absolute error (MAE) of 1.34 mm (with a standard error of +/-0.37 mm at 95% of confidence). This approach slightly outperformed by 0.05 mm of MAE the previous approach based on the nodule detection network. Somehow this result was expected since the nodule detection network was trained without any information regarding the contour of the nodules. Nonetheless, both approaches reported errors that were below 2 mm, the threshold determined by radiological guidelines from which to consider nodule growth [195]. Beyond these results, the ability of the HPU-based approach to provide a measure of uncertainty could help clinicians to make better decisions, since it provides how confident the model is about its predictions.

Regarding lung cancer classification, our best model (TS-3DCNN-MAL) obtained 82.5% of balanced accuracy score. This performance is competitive with those reported from recent cancer classification systems. For instance, in [44] they achieved 86% and 87% of precision and recall, while we obtained 91% and 81.9%. In [14], they reported an AUC of score 92.6% while our model obtained 91.1%.

Despite the notable performances of the pipeline, our work still presents several limitations. First, the great heterogeneity and complexity of the problem makes the amount of data used for the evaluation of the pipeline too small. Hence, greater emphasis is needed on collecting new data for a more comprehensive evaluation. Second, although in the clinical practice the nodule growth is measured with the size of the diameter, we believe building a growth detection method relying on 3D volumetric measures should capture more accurately the patterns of nodule growth. Third, more efforts could be done on visualization and interpretability techniques to allow a better understanding of how the models of the pipeline behave, and thus an easier implantation of this tool in clinical domains.

Finally, several future works can be envisaged. The integration of non-image data, such as the clinical history of the patient, could be an added-value on the pipeline for modelling the whole context of the disease. Also, further efforts in fine-tuning the current networks or adopting recent advances in computer vision [32, 226] could lead to an overall improvement of the performances reported.

6.6 Conclusions

In this chapter, we address the problem of supporting radiologists in the longitudinal management of lung cancer. Therefore, we proposed a deep learning pipeline, composed of four stages that completely automatized from the detection of nodules to the classification of cancer, through the detection of growth in the nodules. In addition, the pipeline integrated a novel approach for nodule growth detection, which relied on a recent hierarchical probabilistic U-Net adapted to report uncertainty estimates. Also, a second novel method was introduced for lung cancer nodule classification, integrating into a two stream 3D-CNN network the estimated nodule malignancy probabilities derived from a pre-trained nodule malignancy network. The pipeline was evaluated in a longitudinal cohort and reported comparable performances to the state-of-the-art.

Chapter 7

AN UNCERTAINTY-AWARE HIERARCHICAL PROBABILISTIC NETWORK FOR EARLY PREDICTION, QUANTIFICATION AND SEGMENTATION OF PULMONARY TUMOUR GROWTH

7.1 Introduction

Pulmonary nodule malignancy is usually assessed based on relatively few parameters such as longest axial diameter, tumour growth and time between observations[1]. Depending on these values and the recommendations made by international radiological guidelines [195], experts make conjectures and draw their conclusions. From the different malignancy parameters, pulmonary tumour growth is one of the most important indicators when assessing lung cancer by computed

---

The work described in this chapter is based on the following publication: Rafael-Palou X, Aubanell A, Ceresa M, Piella G, Ribas V, González Ballester MA. An Uncertainty-aware Hierarchical Probabilistic Network for Early Prediction, Quantification and Segmentation of Pulmonary Tumour Growth. Under review in Medical Image Analysis

[1] https://my.clevelandclinic.org/health/diseases/14799-pulmonary-nodules

tomography (CT) [306]. In particular, clinicians commonly assess tumour growth by imaging surveillance, measuring the nodule diameter along different CT studies taken at different time-points [26].

Anticipating the tumour growth rate would help clinicians to prescribe more accurate tumour treatments and surgical planning. However, lung tumours are highly heterogeneous (e.g. in size, texture and morphology) and their assessment is subject to inter and intra-observer variability (up to 3 mm in diameter on spiculated nodules [106]), making it complex to derive general patterns of tumour growth.

Due to the importance of supporting clinicians in this task, several efforts have been done from the computer vision and artificial intelligence community. Traditionally, the tumour growth prediction problem has been addressed through complex and sophisticated mathematical models [254], such as those based on the reaction-diffusion equation [287, 292] also known as Fisher–Kolmogorov model. These methods provide informative results and explainability. However, the number of model parameters is often limited (e.g. 5 in [318]), which might not be sufficient to model the inherent complexities of the growing patterns of the tumours.

Recently, deep learning and in particular deep convolutional neural networks (CNN) have shown a great ability to automatically extract high-level representations from image data [211]. This has enabled performance improvements over conventional approaches in various medical imaging problems, such as nodule detection [262], segmentation [200], re-identification [237] and malignancy classification [51].

Tumour growth estimation has also been addressed with deep learning for brain, pancreatic and/or colorectal cancer, using data from longitudinal CT/PET or magnetic resonance imaging (MRI) [140, 328, 329]. Proposed deep architectures usually rely on CNNs and recurrent neural networks (RNN) [121], for extracting spatial and temporal tumour growth patterns and correlations. Recently, generative networks such as those based on adversarial learning [96] and variational auto-encoders [148] have also been proposed to enhance grow prediction and clinical interpretability by estimating future images of the tumour [76, 231].

Few works have tackled lung tumour growth estimation [177, 311]. In [177], they proposed two 3D CNNs to extract warping and texture patterns to predict malignancy risk and future aspects of the tumour. In contrast, in [311], they proposed a CNN combined with an RNN extended with an attention mechanism [193] to find temporal patterns to provide trajectories of lung tumour evolution using MRI images. We provide further details of these recent works in section 7.2. These works, however, address the problem of growth prediction in a deterministic way, providing a single prediction without considering uncertainties. Therefore, the models do not usually take into account neither the variability in the annotations of the experts, nor the risk of failure. This could partially explain why, in clinical settings, the credibility of these models is questioned and their adoption limited.

Along with the recent interest on tumour growth prediction and uncertainty with deep learning, this work aims to take a step forward in these promising research directions. In particular, we propose a probabilistic-generative model able to predict, given a single time-point image of the lung nodule, multiple consistent structured output representations. To do this, the network learns to model the multimodal posterior distribution of future lung tumour segmentations by using variational inference and injecting the posterior latent features. Eventually, by applying Monte-Carlo sampling on the outputs of the trained network, we estimate the expected tumour growth mean and the uncertainty associated with the prediction.

The contribution of this work is three-fold. First, to the best of our knowledge, this is the first time pulmonary nodule growth is estimated using deep learning and nodule diameter annotations from multiple experts. Second, this is the first time that model uncertainty is reported using a deep learning approach to predict lung nodule growth. Third, a new deep learning solution is presented, building on an existing hierarchical generative and probabilistic segmentation framework, for lung nodule growth prediction, quantification and visualization.

The rest of the chapter is organized as follows. Section 7.2 describes the most recent works on tumour growth estimation. Section 7.3 details the proposed method for modelling lung tumour growth and its related uncertainty. Sections 7.4 and 7.5 report and discuss the experimental results of applying our approach, and other competing solutions, on a longitudinal cohort. Finally, the conclusions are summarized in Section 7.6.

7.2 Related work

7.2.1 Deep learning deterministic approaches

Deep learning, and in particular CNNs, seems a perfect match for leveraging tumour growth for its intrinsic capability of automatically extracting deep representations and correlations between multiple images [34].

One of the earliest deep learning studies addressing tumour growth estimation was for pancreatic cancer [328]. The authors proposed the use of two (invasion/expansion) stream CNNs, relying on 2D patch images of the tumour, for predicting future tumour segmentations as well as tumour volume growth rates. Interestingly, the method allowed integration with clinical data to enable personalization. Best method performances achieved 86% of Dice score and 8.1% relative volume difference (RVD). Those overcame state-of-the-art of conventional mathematical models [319] for that disease type. However, the size of the test set was too small (10 cases) to extract robust conclusions. Also, to make inference this network required multimodal images (i.e. dual phase contrast-enhanced CT and FDG-PET), as well as three time points spanning between three and four years, which represented strong pre-conditions for the usability of the model.

Aiming to go beyond black-box predictions for lung tumour malignancy [51, 129], recent work [177] proposed a method to generate a future image of the nodule. To do this, a temporal module encoded the distance at which to make the prediction, and two 3D U-Net [202, 248] networks extracted the warped and texture image features of the lung nodule. The network was trained with more than 300 pairs (prior and current studies) of 3D nodule centred patches. Experiments reported a high balanced accuracy score of 86% for nodule progression, although a relative Dice score of 65% for future nodule segmentation. The gap in the model's ability to provide future segmentations of the tumours, the use of a tailored criterion to determine nodule growth instead of conventional metrics (e.g. the longest diameter or double time volume) or not taking into account inter-observer variability, shows the need to continue with the investigation of more reliable and effective solutions.

An alternative approach, especially suitable for temporal series, are the RNNs, in particular the Long Short-Term Memory (LSTM) networks [121]. They were designed for the next time-step status prediction in a temporal sequence capable of learning long-term dependencies. Some recent works have used this type of architectures for tumour growth prediction. For instance, in [329], a 3D convolutional LSTM network [267] was proposed for predicting pancreatic tumour growth. Interestingly, in this study, features from the clinical history of the patient were integrated in the network with the intention to find extra non-linear relationships between spatial and temporal features. This approach used a limited dataset (33 cases) and required having series (≥ 2) of previous images of the lesion, which for early tumour growth estimation is not the best scenario due to the aggressiveness of the disease. Regarding lung tumour growth, in [311] a network was proposed to combine convolutional layers and gated recurrent units with an attention mechanism [193]. The goal was to predict spatial and temporal trajectories over a course of radiotherapy using a longitudinal MRI dataset. Although the purpose of this study is similar to ours (i.e. future lung tumour growth estimation), the complexity of the problem differs in that the images analysed were MRI (instead of CT), the period of the predictions were weeks (instead of months/years), and the number of input images (i.e. 2-3) to the network was larger than in our case.

7.2.2 Deep generative networks

Another way to tackle tumour growth prediction is by using deep generative models. One of the most popular is generative adversarial networks (GAN) [96]. This framework consists of two networks, the generator and the discriminator, that compete with each other in a zero-sum game where the generator aims to increase the error rate of the discriminator network. Thus, the generator learns to map points from a latent space, usually sampled from a multivariate standard normal distribution, into observations that look as if they were sampled from the original dataset. The discriminator tries to predict whether an observation comes from the original dataset.

GANs have been recently applied to predict future tumour/disease growth over time. For instance, in [173] they proposed a 2D deep convolutional GAN for discriminating between true tumour progression and pseudo-progression of glioblastoma multiforme. The results confirmed its suitability for prediction and feature extraction, although only one image per tumour was used in the study. In [76] they built a stacked 3D GAN for growth prediction of gliomas using temporal evolution of the tumour. Although high performances were reported (88% Dice score), the database was composed by only 18 subjects, in which all tumours always grew. In [235], they compared different GAN networks to predict the evolution of white matter hyperintensities. They also demonstrated the potential of using GANs in a semi-supervised scheme, improving results of a deterministic U-ResNet [330]. Despite the satisfactory performances obtained with GANs, this type of network suffers from mode collapse[94], that is, they hardly generate correct representations of the probability output distribution, so they may not be adequate to model uncertainty.

Another well-known approach for addressing image generation is deep auto-encoders (AE). This framework uses an encoder which embeds the input into a representation vector, and a decoder, which projects the vector back to the original manifold. The representation vector is a compression of the original image into a lower dimensional, latent space. The idea is that, by choosing any point in a latent space, a novel image is generated by passing this point through a decoder (as it learned to convert points, or representations, in a latent space into viable images). Therefore, the learning process of this network consists on minimizing the reconstruction error, which is the error between the original image and the reconstruction from its representation. Since auto-encoders do not force continuity in space, images are poorly generated at sampling time.

One successful extension from auto-encoders are variational auto-encoders (VAE) [148, 246]. In particular, the encoder retrieves two vectors, the mean and log-variance vectors, which together define a multivariate distribution in the latent space. When a random point is sampled from this distribution, the decoder produces a similar image, guaranteeing the continuity in the latent space. The way to achieve this, is by making the output distribution of the encoder as close as possible to a standard multivariate normal distribution using the Kullback-Leibler divergence (KL) loss. Thus, the total loss function of the VAE is composed by the sum of the KL-divergence loss and the reconstruction loss. A variant of VAEs was created to generate multiple outputs from a single input. Precisely, conditional variational auto-encoders (CVAE) [275] were proposed to model the distribution of a high dimensional space as a generative model conditioned on the input. Therefore, the prior on the latent variable is conditioned by the input.

Few works have applied auto-encoders and their variants for tumour/disease growth prediction. In [140] they proposed using a deep auto-encoder attached to a fully connected network architecture for colorectal tumour growth detection. Despite providing results close to the RECIST methodology[2] and radiomic measures, the use of the auto-encoder was for mere feature reduction. In [28], the authors applied a VAE for progression of Alzheimer disease from structural MRI images. Their experiments demonstrated that VAE outperforms conventional CNNs on doubtful cases as it acts as a soft classifier learning a Gaussian distribution. Also, for patient risk analysis they observed that VAE produced less false positive cases, sampling from the latent space, than deterministic CNNs. However, CNNs provided better overall performances. In another study [239], they conditioned a deep auto-encoder on fixed characteristics like age and diagnosis, to generate sequences of 3D MRI for Alzheimer's disease progression. Despite results outperformed previous 2D versions, some artefacts and false structures were noted on the generated images. Moreover, additional terms were required to ensure loss stability, latent space continuity, reducing memory constraints and restoring 3D outputs.

7.2.3 Uncertainty in deep learning

Contradictorily, given the multifactorial and complex nature of the problem, uncertainty in the prediction of tumour growth was not addressed in any of the aforementioned studies. However, uncertainty information about the output of a network could make them safer and more reliable since it would allow indicating potential mis-segmented or low confident regions, or guiding user interactions for refinement of the results. Two common approaches have been proposed for modelling uncertainty in deep learning, Monte Carlo dropout networks (MCDNs) [83] and Bayesian neural networks (BNNs) [268]. MCDNs use dropout layers as a Bayesian inference approximation in deep Gaussian processes, and although their implementation is easy, criticism has emerged recently regarding the type of

---

[2] https://recist.eortc.org/

uncertainty that is captured [222]. BNNs use variational inference to learn the posterior distribution of the weights given a dataset. These weights are implicitly described as (multivariate) probability distributions. This has several consequences. First, it makes the neural network non-deterministic; for every forward pass, we must sample from each weight distribution to obtain a point estimate. Repeated applications of this sampling technique, through Monte Carlo sampling, will result in different predictions which can then be analysed for uncertainty. Second, it changes the backpropagation algorithm, since we cannot flow back the gradients through a sampling operation.

Uncertainty estimation in deep neural networks has been widely investigated for medical image tasks. For instance, in segmentation of multiple sclerosis lesions, some works [210, 249] showed that by filtering out predictions with high uncertainty, the models improved lesion detection accuracy. For brain tumour segmentation, other work [75] demonstrated that MCDNs can be calibrated to provide meaningful error bars overestimates of tumour volumes. Moreover, the uncertainty metric based on MCDNs also showed promising results in disease grading of retinal fundal images [23, 172]. In [182] a Bayesian method predicted patient-specific tumour cell densities with credible intervals from high resolution MRI and PET imaging modalities.

Unfortunately, few works have modelled uncertainty for tumour growth estimation. In [231] a deep probabilistic generative model (sPUNet) [29, 153] was used to model glioma growth for radiotherapy treatment planning. The model, based on a combination of a U-Net [248] and a CVAE [275], was able to generate multiple future tumour segmentation modes on a given input. Although they demonstrated the potential of providing multiple views over a single solution, they did not report nodule growth performances.

7.3 Method

We present a novel approach to estimate the future growth of pulmonary nodules along with its uncertainty. Our approach exploits the generative and probabilistic nature of a recent framework, the hierarchical probabilistic U-Net [154] (HPU), to estimate the output probability distribution of lung nodule growth, conditioned on an initial image of the nodule. Before delving into the details, in the following sub-section we describe the basics of the underlying framework.

7.3.1 Hierarchical probabilistic U-Net

A segmentation framework that provides multiple segmentation instances for ambiguous images was proposed in [154]. This network, schematized in Figure-7.1,

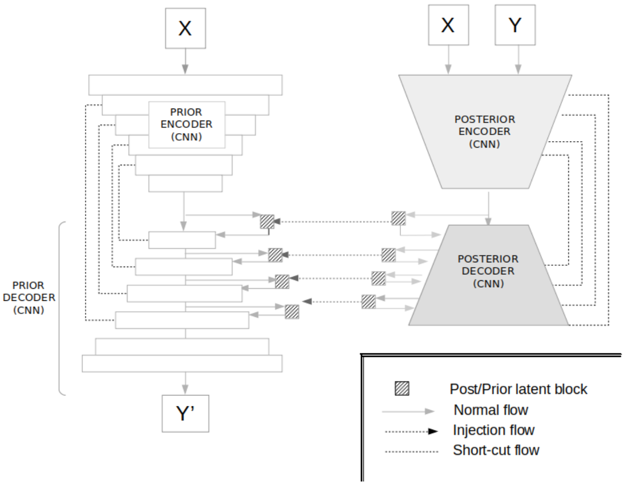


Figure 7.1: General overview of the HPU network architecture. On the left of the picture we can observe the prior network and on the right the posterior. Both networks have different probabilistic latent blocks interleaved along the decoder component.

is composed of two inter-related sub-networks, the posterior and the prior. Both follow a CVAE scheme with a couple of changes. First, the encoder-decoder structure is implemented by a 2D U-Net [248] extended with residual blocks [74, 117] (U-ResNet) and filters adjusted to the input size. Second, instead of a single probabilistic latent block (see Figure-7.3) at the end of the encoder, several probabilistic latent blocks are interleaved at different levels of the hierarchy of the decoder, to provide fine-grained segmentation samples closer to the ground truth probabilistic distribution.

The inference process of this network consists on forward-passing an input image, X, through the prior network. Specifically, along the decoder part of the network, feature activation maps are concatenated with vectors, $\mathrm{z_i}$ ($i \leq L$, being L the number of latent hierarchies), obtained from sampling different latent distributions interleaved in the decoder. As a result, we obtain a predicted segmentation, Y'.

The training process of this network aims to pull to each other the prior distribution p, encoded by the prior network, and the posterior distribution q, defined by the posterior network, while minimizing the loss of the reconstructed images. This is the same as maximizing the evidence lower bound (ELBO) in variational inference. Therefore, the KL divergence loss (D_{KL}) between the posterior and the prior distributions is added to the reconstruction objective ($\mathcal{L}_{\mathrm{rec}}$) obtained through the log likelihood (represented by the pixel-wise categorical distribution P_{c}) between the reconstructed image Y', and the ground truth segmentation Y. Additionally a weighting factor β, is multiplied to the $\mathrm{D_{KL}}$ term to balance the overall loss

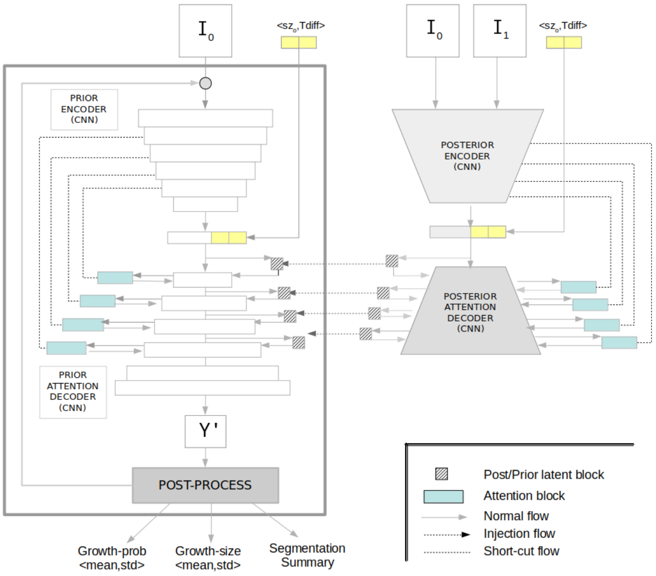


Figure 7.2: General overview of the proposed U-HPNet network architecture. This network is also composed by a prior network (on the left, with further details) and a posterior (on the right). Attached at the end of the posterior we observe the post-process module aimed at reporting the estimated future growth prediction, size and appearance with the associated uncertainty.

function:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{\,\text{z}\sim\text{Q}}[-logP_{\text{c}}(Y|Y')] + \beta\sum_{i=0}^{L}\mathbb{E}_{\,\text{z}_{<\text{i}}\sim\text{Q}}D_{\text{KL}}(q_{\text{i}}(z_{\text{i}}|z_{<\text{i}},X,Y)||p_{\text{i}}(z_{\text{i}}|z_{<\text{i}},X))$$

where $\mathbb{E}_{\,\text{z}\sim\text{Q}}$ is the expectation operator, and z a vector sampled from the posterior distribution Q.

7.3.2 U-HPNet

Based on the HPU framework, we propose a network (U-HPNet) able to generate plausible future nodule segmentations conditioned on the nodule image, its diameter at time T_0, and the temporal distance at which to make the prediction. To do this, the U-HPNet uses variational inference to approximate the estimated output distribution to the ground truth, in our case, provided by different graders. Figure-7.2 shows the overall architecture of the proposed network.

Architecture

Both sub-networks of the U-HPNet (prior and posterior) receive as input an axial nodule image I_0 at time T_0, while the posterior receives also the axial nodule image

I_1 at T_1. The images are centred patches of 32x32 pixels rather than 128x128 as in the original network. We down-scaled the input size of the network to focus in the relevant parts of the image (i.e. contour and close surrounding of the nodule), and to reduce the number of parameters of the network, especially convenient for small datasets [233]. Smaller patches were discarded due to the size of the nodules, and larger patches (e.g. 64x64) experimentally did not report any performance gain.

We conditioned the latent space, learnt by the network, with a couple of extra features: the time difference (Tdiff) at which to predict nodule growth, and the diameter size (sz_0) of the nodule image at time T_0. Tdiff is an ordinal value representing the main time-elapses defined by radiological guidelines (i.e. 6, 12, 24 or more months) [195]. Sz_0 is a numerical value provided (in our case) by radiologists to better estimate the tumour growth. In particular, with this feature, we aimed to facilitate the network to learn the intrinsic patterns followed by the experts when measuring tumours from the images. Both features (Tdiff, sz_0) were normalized between 0 and 1, and concatenated with the encoder output.

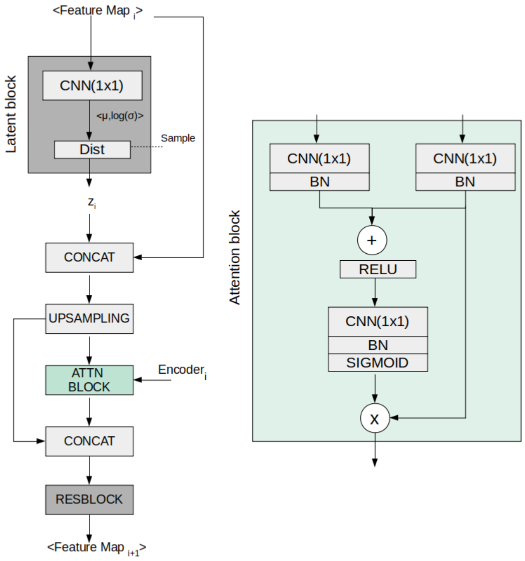


Figure 7.3: On the left, we show a more detailed view of the different components of a decoder layer of the U-HPNet. On the right, we see the elements that compose an attention block.

Regarding the network architecture, both sub-networks use the same 2D U-ResNet as in the original HPU, but adapted to the proposed input size (32x32). Also, up to 4 prior/posterior latent blocks are interleaved in the decoder of the sub-networks, generating latent feature vectors (z) of 1, 4, 16, 64 dimensions respectively.

Additionally, we integrated a soft attention mechanism in the decoder part of the sub-networks with the intention of detecting small and minor changes in the structure of the nodule images. To do this, we followed a recent work [217] in which a grid-attention mechanism was integrated in a U-ResNet. The attention mechanism aims at progressively suppressing feature responses in irrelevant background regions. To do this, attention gates are integrated before the concatenation operation to merge only relevant activations. Figure-7.3 provides further details regarding the components of the attention mechanism and how it was integrated in the decoder of the sub-networks.

Loss function

On the conventional ELBO loss function used in the original HPU paper, we incorporated a couple of modifications in the reconstruction loss ($\mathcal{L}_{\text{rec}}$) term. In particular, we used the L1 distance between the predicted $D1'$ and the ground truth $D1$ tumour diameters, and the intersection over union (IoU) between the predicted Y' and ground truth Y tumour segmentation. Also, a weighting (γ) factor was used on the combined loss to balance the ranges of both terms.

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{IoU}}(Y, Y') + \gamma \mathcal{L}_{\text{L1}}(D1, D1')$$

We used the L1 loss to prioritize the diameter fidelity, and consequently improve network performance. Also, we used IoU loss as a good approximation function when learning on imbalanced data conditions [216], which in our case was caused by having a much smaller number of pixels belonging to the tumour than to the background. In our experiments, we found better performances setting γ to one.

Post-processing

The generative ability of the proposed network offers the possibility to produce future nodule segmentations, sampling from the latent space and injecting the resulting vectors in the network, for a given input. This may be useful from a medical exploratory point of view, but for practical reasons a more useful outcome should be presented to the clinicians. To this end, we formulated a generic and embeddable post-processing module that converts multiple predicted segmentations into a lung nodule growth prediction, size and segmentation visualization with the uncertainty associated to each of them. Precisely, the post-processing module applies Monte-Carlo sampling by running the network K times (K=1000) with the same input image. In particular, for each iteration, a sample from all the hierarchical latent blocks of the prior network is injected in the corresponding location of the decoder part of the (prior) network, to produce a new segmentation. As a result, we obtained K random nodule segmentations. For each predicted segmentation, we extracted its longest diameter $D1'$, using conventional image processing libraries. With the vector of K nodule diameters, we computed the vector of predicted nodule growths, Δ, by subtracting the input nodule diameter size $D0$ (the aforementioned sz_0) to the predicted diameters $D1'$. From the resulting vector Δ of predicted nodule growths, we computed its mean and standard deviation as measures of nodule growth size and its associated uncertainty.

In addition, we computed the probability that the nodule growth is at least of 2 mm (threshold recommended in clinical guidelines for tumour growth [195]). For this, for each of the K nodule growths, we used the logistic function $f(\Delta_{\mathrm{i}}) = 1/(1+e^{-\Delta_{\mathrm{i}}+2})$. From the resulting K-length vector of probabilities, we considered the mean and the standard deviation as the estimated nodule growth probability and its associated uncertainty.

Finally, the post-processing module also outputs two images, both corresponding to the predicted future tumour appearance (at T_1). In particular, and inspired by [142], one of the images is the per-pixel mean of all K predicted segmentations and the other the per-pixel standard deviation.

7.3.3 Comparison with related works

Since we did not find any other deep generative network to provide lung tumour growth predictions and their associated uncertainty, we adapted 4 different state-of-the-art deep architectures to compare the performance of our method (see Figure-7.4), one deterministic network and three generative.

To allow a fair comparison, all these networks had the same U-ResNet backbone proposed for the U-HPNet, with the same number of layers and filters. Also, these models were configured with same data augmentation, optimization algorithm, batch size and learning rate than the U-HPNet network. Moreover, these networks had the same input (I_0, sz_0 and Tdiff) and output as the U-HPNet (i.e. an estimated future segmentation of the nodule). For the non-deterministic models, the output was post-processed to evaluate tumour growth prediction, diameter growth and the segmentation performance.

As for the deterministic (or baseline) approach, we used a single U-ResNet like network, Figure-7.4a. This network was trained using a conventional loss function, formed by a pixel-wise binary cross entropy, without any additional configuration.

The first generative selected method consisted on a Bayesian dropout network (BAYES_TD) inspired by the Bayesian SegNet network proposed in [142]. This approach provides a probabilistic pixel-wise semantic segmentation by enabling dropout at inference time. Therefore, this approach aims to find the posterior distribution over the convolutional weights, W, given the observed image I_0 and labels Y, i.e. $p(W|I_0, Y)$. According to the authors, the best configuration was obtained using dropout in the central part of the network. Thus, we followed the same suggestion and we setup dropout (p=0.5) layers in the 3 last encoder and

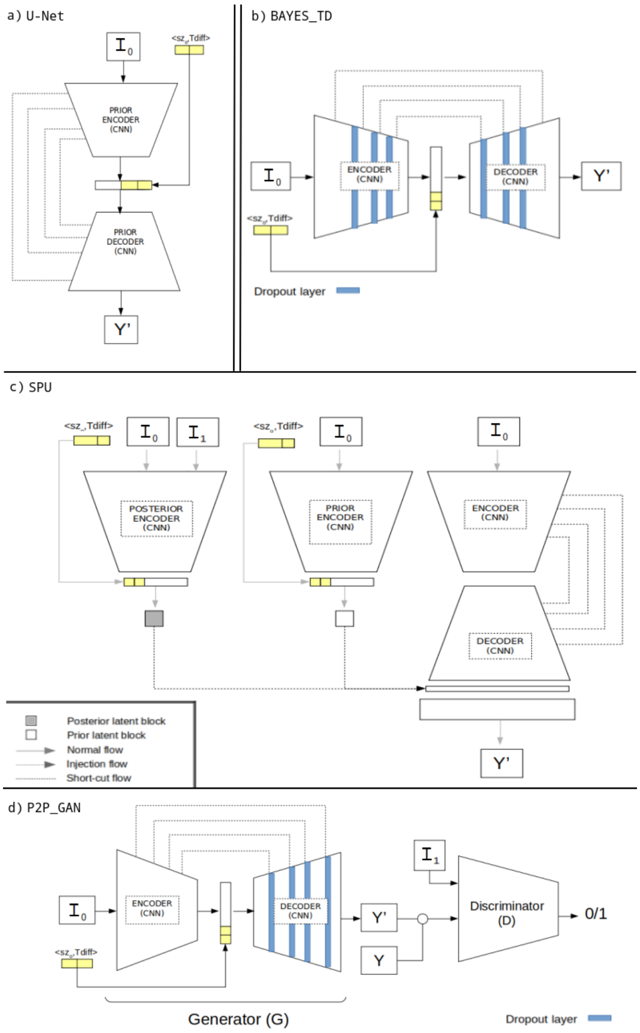


Figure 7.4: Four alternative network architectures proposed for lung nodule growth estimation. At the top we have the U-ResNet and the generative Bayesian dropout. In the centre, we show the probabilistic U-Net. Below we find the Pix2Pix cGAN network proposed.

3 initial decoder blocks of the U-ResNet, Figure-7.4b. This network was trained using pixel-wise binary cross entropy. Hence, we used dropout at inference time as a way to get samples from the posterior distribution.

The second proposed generative network was the former version of the HPU, the standard probabilistic U-Net (SPU) [153]. This approach goes beyond the notion of reporting a per-pixel probability map, by capturing the co-variances between pixels and providing consistent structured outputs. To do this, two networks: the prior (having as input a nodule I_0) and the posterior (which also receives the nodule I_1), learn to map the input into a low dimensional latent space which encodes the distribution of all possible segmentation variants for the given input. In particular, we configured a latent vector of 6 features (or dimensions) as in the original paper, Figure-7.4c. This network was trained to maximize the ELBO function composed by the pixel-wise binary cross entropy between the predicted and the ground truth segmentation, and the KL-divergence between the posterior (which can see the future image of the nodule) and the prior distributions. By sampling on the latent features of the prior network, this method allows generating multiple segmentations at inference time.

The last generative approach consisted on a conditional GAN named Pix2Pix [134]. The framework allows learning, in a model-free fashion, a mapping between two images. In our case, the two images were a tumour image I_0 and a segmentation image Y at T_1. The proposed network (P2P_GAN) is composed by two networks; a generator formed by U-ResNet configured with dropout (p=0.5) along the decoder (no specific locations were indicated by the authors), and a discriminator composed by the encoder part of a U-ResNet, Figure-7.4d. These two networks learn to generate images that are as similar as real ones, as well as to discriminate between images that are increasingly similar between real and fake ones. This network was trained as suggested by the authors, using the $\mathcal{L}_{\mathrm{cGAN}}$ loss:

$$\mathcal{L}_{\mathrm{cGAN}}(G,D) = \mathbb{E}_{\mathrm{I_1,Y}}[logD(I_1,Y)] + \mathbb{E}_{\mathrm{I_0,I_1,z}}[log(1-D(I_1,G(I_0,z)))],$$

which represents the sum between the discriminator D loss (i.e. binary cross entropy) of a nodule I_1 and the segmentation ground truth Y, and one minus the discriminator loss of a nodule I_1 and the segmentation Y' produced by the generator at T_1, i.e. $Y' = G(I_0, z)$. Additionally, a second term was added into this loss to figure out the fidelity of the generated samples with the ground truth. Thus, the L1 distance was computed between the generated sample Y' and the truth Y. The final loss G^* is as follows:

$$G^* = arg\,\underset{G}{min}\,\underset{D}{max}\,\mathcal{L}_{\mathrm{cGAN}}(G,D) + \lambda\mathcal{L}_{\mathrm{L1}}(G).$$

This approach allowed generating multiple samples by adding noise in the form of dropout, applied during training and testing time.

7.3.4 Tumour growth assessment

We assess tumour growth in terms of growth prediction, size and segmentation mask. For this, we considered not only the generative nature of our approach, but also the fact of having per each nodule the opinion of up to three different radiologists. In the following, we provide further details regarding the tumour growth assessment.

Metrics

We proposed to evaluate how well the distributions produced by the generative model and the given ground-truth distributions agree. To this purpose, we considered two evaluation scenarios:

1) Using the expected value of the distributions, we computed conventional metrics such as precision (Prec), recall (Rec), specificity (Spec) and balanced accuracy (Bacc) for growth prediction, mean absolute error (MAE) and mean squared error (MSE) for nodule growth, and Dice for segmentation fidelity.

2) Using confidence intervals, we defined the following metrics:

- For nodule growth prediction:

  We proposed the metric Bacc_2std. This computes the balanced accuracy between the radiologist tumour growth predictions (i.e. 1 if the tumour growth size was above 2 mm) and the predicted tumour growths at 2 standard deviations away from the estimated growth size means.

  To do this, we re-defined a true positive case as when the ground truth and the lower value of the predicted interval were above 2 mm. A false negative was when the ground truth was above 2 mm, but the lower value of the interval was not. A true negative was when the ground truth and the upper value of the interval were less or equal to 2 mm. False positive was when the ground truth was less or equal to 2 mm, but the upper value of the interval was not.

- For nodule growth size:

  We proposed the ratio P(RX∈2std). This reports the proportion of tumours (over all tumours), whose growth size is within the interval (2 standard deviations away from the estimated tumour growth mean).

  To do this, we compared, for each tumour, if the distance between the tumour growth size (ground truth) and the predicted growth size distribution was below the distance between the estimated mean with 2 standard deviations and the predicted growth size distribution. To compute this distance we used the Mahalanobis distance D_{MH}, which is the distance of a test point x, from the centre of mass m, divided by the width of the ellipsoid defined by the covariance matrix C in the direction of the test point.

  $$D_{MH}^2 = (x - m)^{\mathrm{T}} C^{\text{-}1} (x - m).$$

- For nodule segmentation:

  We used the estimation of the Generalized Energy Distance (GED) [291] metric. This metric reports the segmentation performance in terms of the variability in the ground truth as in the generated samples of the network.

$$D^2_{GED} = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} d(Y_i', Y_j) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d(Y_i', Y_j') - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} d(Y_i, Y_j).$$

where m and n are the number of generated and ground truths segmentations, Y_i' and Y_j are a predicted and ground truth tumour segmentation and d is the distance obtained using the 1-IoU metric. The resulting GED distance will be better the closer to 0.

Ground truths

To evaluate the models we used the annotations provided by the 3 different radiologists (RX0, RX1 and RX2). However, given that the annotations of the different radiologists may diverge (although all of them may still be correct), we derived two more ground truths, precisely, the mean of the radiologists annotations (RX_mean), and the radiologist annotations that stands closest to our predictions (RX_closest). Although the former is direct to obtain, the second has some particularities:

- For the deterministic model, we computed the closest radiologist tumour growth size and tumour growth prediction selecting the radiologist annotation with the minimum growth size difference with respect to the prediction. We computed the closest radiologist segmentation, selecting the segmentation with the highest Dice with respect to our prediction.

- For the generative models, we computed the closest radiologist tumour growth size and tumour growth prediction selecting the radiologist annotation with the minimum Mahalanobis distance between the radiologist growth size and the estimated output distribution. We computed the closest radiologist segmentation, selecting the radiologist segmentation with the highest average Dice score obtained from the generated samples of the network.

7.4 Experiments and Results

7.4.1 VH-Lung

In this study, we used a longitudinal lung CT dataset [237] for the follow-up analysis of incidental pulmonary nodules. In total, the cohort contains 161 patients (10 more cases compared to the previous version) with two thoracic CT scans per patient. The most relevant pulmonary nodule in each patient was located in each study by two different specialists. We address the reader to the source article for further details regarding the ethics, inclusion criteria, and acquisition protocol of the dataset.

A new feature included in this updated version of the cohort is that up to 3 different clinicians ($RX0$, $RX1$ and $RX2$) reported the diameter size of the nodules ($D0$, $D1$) at the two different time-point (T_0, T_1) studies. From here, we computed the tumour growth by subtracting the diameters ($D1 - D0$). The tumour growth mean in the dataset was 2.52±3.85 mm for RX0, 2.76±3.63 mm for RX1 and 2.68±4.01 mm for RX2. The inter-observer mean absolute difference was 1.55 mm, whereas the inter-observer mean standard deviation was 0.97 mm (both metrics were computed pair-wise) [232]. The time interval between current and previous CT studies ranged from 32 to 2464 days.

Tumour segmentations were obtained in a semi-automatic way, being visually verified and curated with the annotations provided by each of the radiologists (that is, location of the centroid, diameter and growth of the tumours) with a residual margin of 0.25 mm.

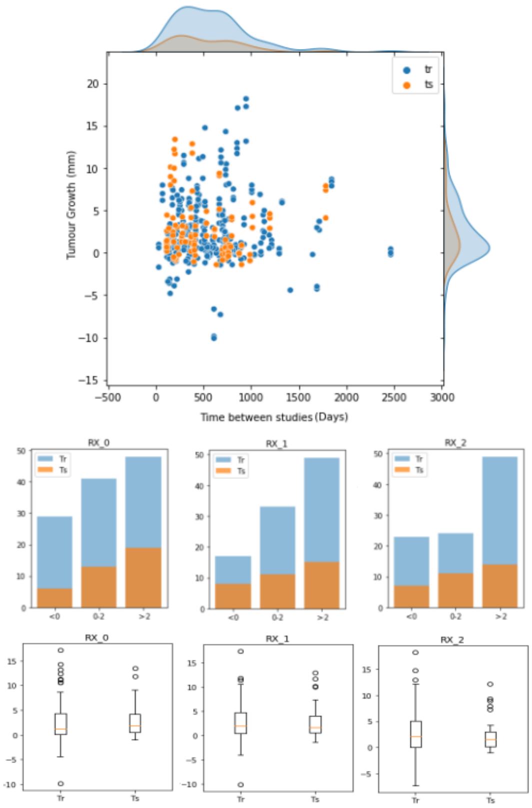


Figure 7.5: Description of the training and test set cohorts in terms of tumour growth size (in mm), time between studies (days) and the annotators (RX-0,1,2).

To train and evaluate the proposed methods, the whole data was randomly divided into training (70%) and test (30%) sets. In this process we assured that all entrances of the same nodule were in the same set in order to avoid data leakage between partitions. Therefore, the training set was composed of 313 (122 unique) nodule growth annotations from up to three different radiologists (118 for RX0, 99 for RX1 and 96 for RX2), whereas the test set was formed by 104 (39 unique) nodule growth annotations (38 for RX0, 34 for RX1 and 32 for RX2). Hence, for each data entrance (i.e. nodule growth annotation) of these partitions we had 2 nodule images (at T_0 and T_1), 2 nodule segmentations (at T_0 and T_1), a growth label (indicating whether it grew (1) or not (0)) and a growth size (in mm) corresponding to a particular radiologist. Further details regarding training and test set partitions can be seen in Figure-7.5.

7.4.2 Qualitative results

We present some qualitative results (Figures-7.6,7.7,7.8) obtained from the U-HPNet using lung nodules from the test set. We recall that these results were obtained from a post-processing (see section 7.3.2) aimed at obtaining the estimated growth probability, size and visualization together with its associated uncertainty.

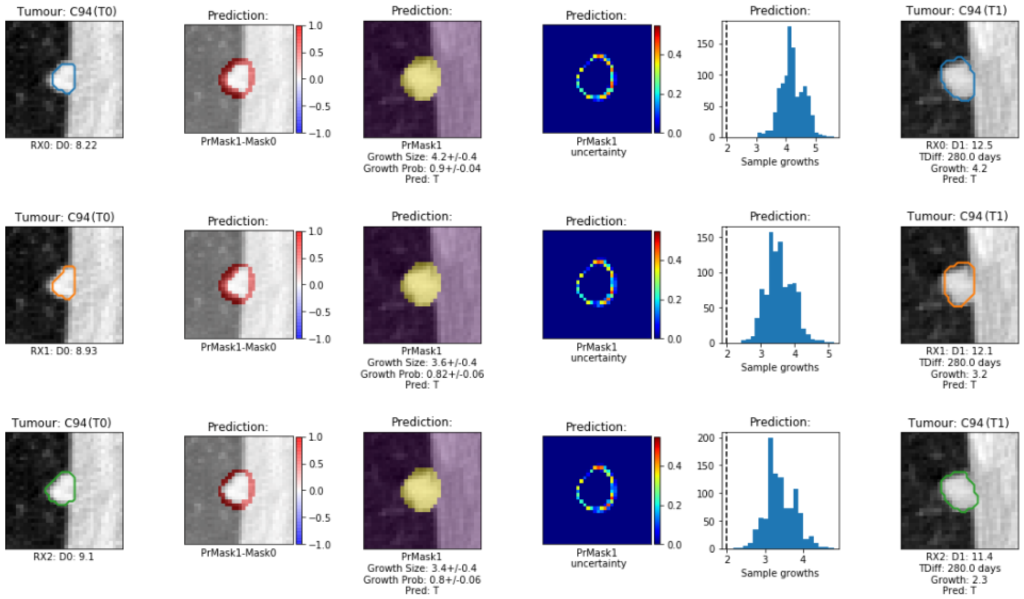


Figure 7.6: Comparison of ground truth annotations and predictions from the U-HPNet for the tumour case C94.

The figures are composed of six columns, the first one shows the nodule at T_0 overlapped with the segmentation of a radiologist. The second column shows, overlapped with the nodule image at T_0, the difference between the ground truth segmentation at T_0 and the estimated mean segmentation at T_1. The third column provides, overlapped with the nodule image at T_1, the estimated tumour mean

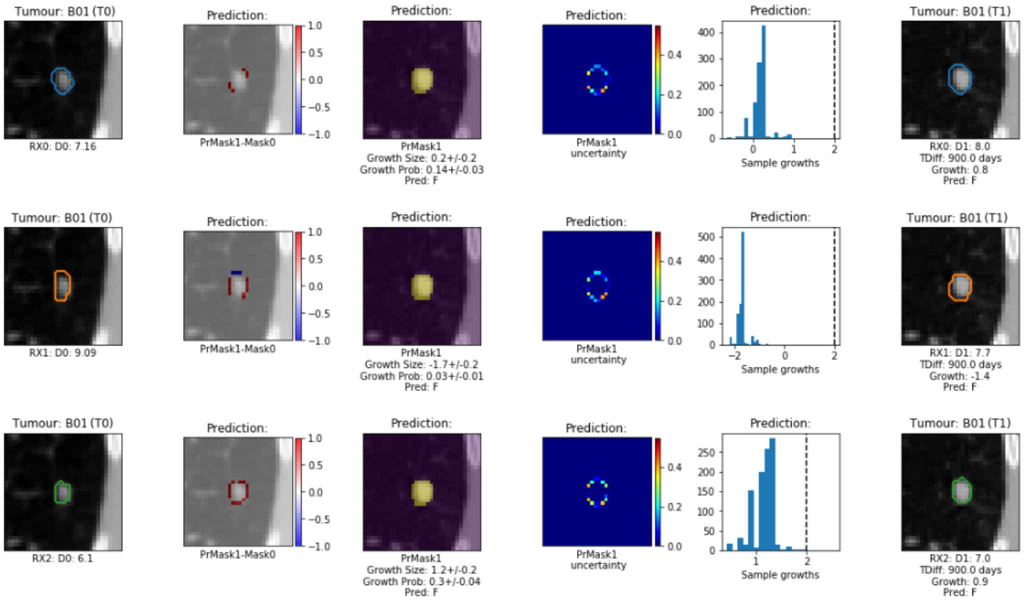


Figure 7.7: Comparison of ground truth annotations and predictions from the U-HPNet for the tumour case B01.

segmentation. The fourth column visualizes the estimated uncertainty probability map with the per pixel-standard deviation. The fifth column shows the histogram of the (K=1000) estimated tumour growths (i.e. predicted diameter at T_1 minus radiologist diameter at T_0). The last column shows the nodule image at T_1 overlapped with the segmentation of the radiologist at T_1.

The first tumour case (Figure-7.6) was cancerous with 280 days between image studies. The tumour, attached to the chest wall, was annotated by three radiologists. Our method correctly predicted the existence of growth ($>$ 2 mm) for each of the three radiologists, reporting high growth probabilities or estimated means (0.9, 0.82 and 0.8) and low uncertainties (0.04, 0.06 and 0.04) or standard deviations. The predicted tumour growth sizes were especially close (4.2, 3.6 and 3.4 mm) to those reported by the first two radiologists (4.2, 3.2 and 2.3 mm).

The second tumour case (Figure-7.7) was benign, with almost 3 years between studies (900 days). All three radiologists did not detect any relevant tumour growth (diameter difference $\leq$ 2 mm) for this case. Our method correctly predicted the existence of no growth for each of the three radiologists, reporting low tumour growth probabilities (0.14, 0.03 and 0.3), especially for the first two radiologists, with low uncertainties (0.03, 0.01 and 0.05). The predicted tumour growth sizes were approximately less than 0.6 mm (0.2, -1.7 and 1.2 mm) apart from those reported by the radiologists (0.8, -1.4 and 0.9 mm). Moreover, the network agrees with the second radiologist to correctly predict tumour recession

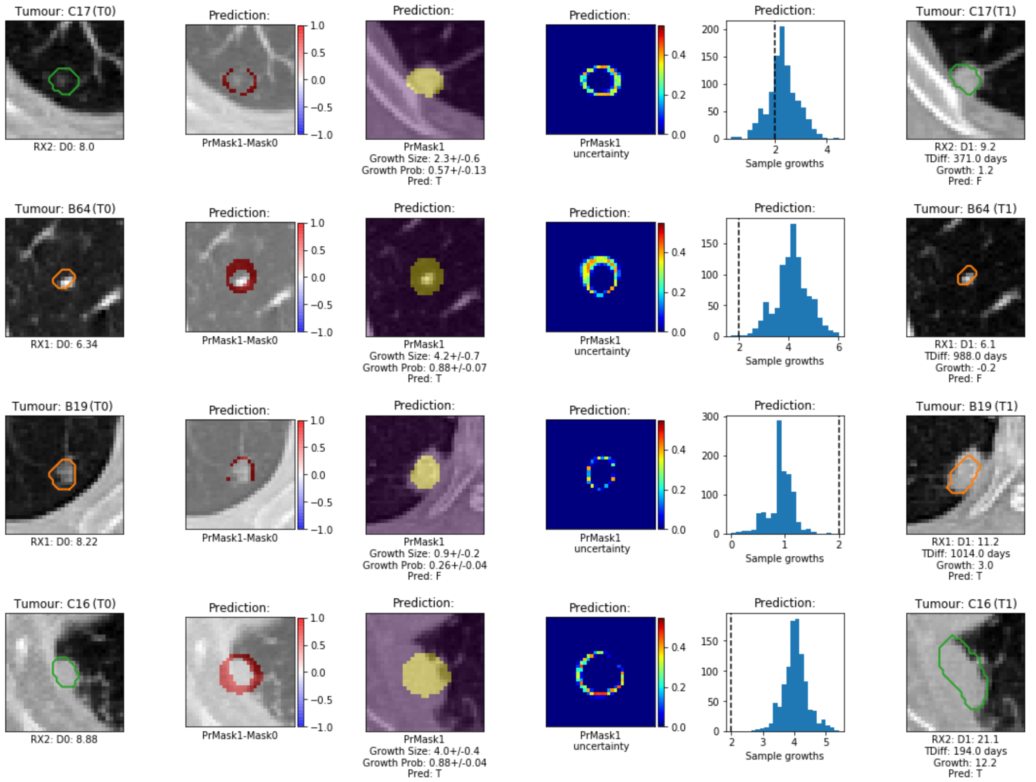


Figure 7.8: Comparison of ground truth annotations and predictions from the U-HPNet for the tumours C17, B64, B19 and C16. In this example, the network incorrectly predicts growth for the three first cases while for the last one it guesses the prediction, although the predicted growth size is far from the radiologist measurement.

and also correctly guess to provide slightly higher probability, and greater tumour growth size for the third radiologist than for the other two.

Figure-7.8 shows 4 tumour examples in which our model struggles to find correct predictions. In particular, for the tumour case C17, the model predicted a tumour growth size of 2.3±0.6 mm whereas the radiologist reported less than 2 mm (i.e. 1.2 mm). Nevertheless, the network provides a tumour growth probability close to 0.5 with an uncertainty of 0.14, which implies that the network is not highly confident on the nodule predictions. Looking at the ground truth provided for this case (first column of the figure), this mistake could be due to a probable overestimation of the diameter size of the nodule at T_0. For the second tumour case, B64, the model predicted a growth size of 4.2±0.7 mm whereas the radiologist detected tumour recession (i.e. -0.2 mm). In this case, the network incorrectly provides high tumour growth probability and low uncertainty. However, if we look at the estimated per-pixel uncertainty image, the model correctly outputs a relevant quantity of uncertainty surrounding the nodule. In the third tumour case, B19, the model predicted a tumour growth size of 0.9±0.2 mm whereas the radiologist found a tumour growth of 3.0 mm. This could be caused by a relevant change in the context of the nodule, i.e. the nodule at T_1 was attached to the lung wall whereas at T_0 it was aerial. In the last case, C16, the model correctly predicted tumour growth. However, the radiologist indicated a high tumour increase of 12.2 mm, whereas the network only detected 4.0±0.4 mm. If we observe the estimated segmentation mean image of the nodule, we see that the network missed detecting the longitudinal growth direction of the tumour due to not enough representation of this kind of tumour growth behaviour in the training set.

7.4.3 Quantitative results

Next, we detail the test performances of the proposed networks in terms of estimated nodule growth size, prediction and segmentation fidelity (for further details on the used metrics, see section 7.3.4). To obtain these performances, the methods were optimized with the training data using a 5-fold cross-validation, and tested with the testing set. The setup of the learning hyperparameters was the same for all methods, thus we used 1e-4 of learning rate, 8 of batch size, 200 epochs and Adam [147] as optimization algorithm.

The performances provided in the tables of the following sections show for each of the metrics, the mean and the standard deviation obtained from a bootstrapping process (with N=1000 iterations), in which for every iteration we performed a resample with replacement from the test set (N=104).

Nodule growth prediction

The performances of the U-HPNet regarding lung nodule growth prediction are shown in Table-7.1. These results were obtained using the annotations from three different radiologists (RX0, RX1 and RX2), their mean and the radiologists' annotations closest to our predictions (closest), with the intention to provide a more complete analysis of the performance of our method and to detect disparities between radiologists annotations (see section 7.3.4 for further details).

Despite the heterogeneity in the morphology, density and location of the nodules, the fact of using a single image (I_0) of the tumour and the variety on the time at which to make the predictions, our method was able to satisfactorily report positive performances, such as 0.74 of balanced accuracy, 0.71 of recall and 0.76 of specificity.

Further details regarding the tumour growth prediction of the U-HPNet are provided in Figure-7.9. This figure shows the prediction accuracy stratified by the time to predict, the real growth of the nodules and the ground truth type. From this figure, we can observe that, the best performances were usually obtained when

RX	Bacc	Prec	Rec	Spec
RX0	0.49±0.07	0.46±0.11	0.42±0.11	0.56±0.11
RX1	0.68±0.08	0.63±0.13	0.64±0.13	0.72±0.10
RX2	0.67±0.09	0.58±0.15	0.59±0.15	0.75±0.10
Mean	0.55±0.04	0.50±0.07	0.47±0.07	0.63±0.06
Closest	**0.74±0.07**	**0.65±0.12**	**0.71±0.12**	**0.76±0.08**

Table 7.1: Nodule growth prediction performances obtained by the U-HPNet using expected means.

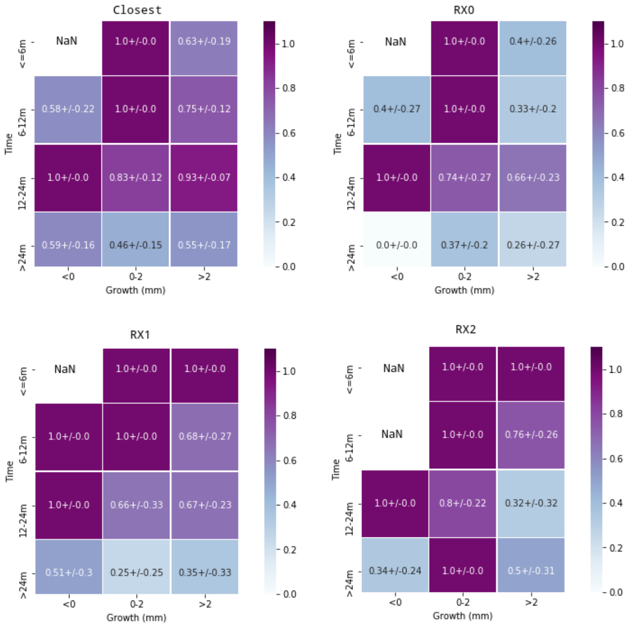


Figure 7.9: Tumour growth prediction accuracy of the U-HPNet stratified by time to predict, nodule growth size and ground truth (i.e. Closest, RX0, RX1 and RX2).

predicting in the range of 12 to 24 months, whereas the worst performances were obtained when predicting above 24 months. Also, the best performances were usually obtained when the nodules had a growth between 0 and 2 mm.

Additionally, we provide Table-7.2 with the balanced accuracy obtained at 2 standard deviations away from the expected means.

From this table, we observe that the performances are clearly below the 0.74 of balanced accuracy obtained using single point estimates (i.e. the expected mean). The reason is that Bacc_2std reports the balanced accuracy at 2 standard deviations away from the expected mean. Therefore, at this extreme, our approach is still able to correctly predict tumour growth in 57% of the test cases.

RX	Bacc_2std
RX0	0.37±0.08
RX1	0.53±0.09
RX2	0.57±0.10
Mean	0.45±0.05
Closest	**0.57±0.08**

Table 7.2: Nodule growth balanced accuracy performances for the U-HPNet obtained within 2 standard deviations from the mean.

Nodule growth size

We also computed the nodule growth size performance of the U-HPNet. Table-7.3 provides the mean absolute error (MAE) and mean squared error (MSE), obtained from the comparison of the estimated tumour growth size mean of the network for the different ground truths. Also, this table reports the metrics regarding the probability of finding the radiologist measurements within 2 standard deviation P(RX∈2std) from the estimated tumour growth size mean.

RX	MAE ↓	MSE ↓	P(RX∈2std)
RX0	2.99±0.42	16.60±4.94	0.20±0.06
RX1	2.52±0.45	13.11±4.40	0.21±0.07
RX2	2.62±0.41	11.86±2.98	0.28±0.08
Mean	2.83±0.23	14.40±2.39	0.17±0.04
Closest	**1.74±0.34**	**7.55±2.87**	**0.44±0.08**

Table 7.3: Nodule growth size performances of the U-HPNet using the estimated mean (MAE, MSE) and the interval composed by the mean and 2 standard deviations. A down arrow next to a metric means that the metric is more accurate the smaller the value.

As we observe from Table-7.3 the best performances on tumour growth size reported a MAE of 1.74 mm close to the 1.55 mm of inter-observer mean absolute difference. Moreover, this method reported that in 44% of the cases, the exact tumour growths annotated by the radiologists were found at 2 standard deviations from the mean. Further details are exposed in the discussion section.

Nodule segmentation

Subsequently, we report the performance of the U-HPNet for predicting accurate future nodule segmentations. To do this, we computed for each nodule of the test set, the average Dice score obtained from each generated tumour segmentation with respect to the proposed radiologist segmentation. Table-7.4 summarizes the resulting Dice performances for each of the radiologists.

RX0	RX1	RX2	Mean	Closest
0.74±0.02	0.77±0.02	0.75±0.02	0.76±0.02	**0.78±0.02**

Table 7.4: Nodule segmentation performances of the U-HPNet for each of the ground truths.

The best Dice score was 78%, achieved for the closest radiologists ground truths. Complementary, we also computed the GED metric to report the ability of the network to generate accurate and diverse future tumour segmentations (Table-7.5). From this result, we remark that a high segmentation agreement (i.e. 0.14 of 1-IoU) was found between ground truths (YY). This may explain why a small variability (i.e. 0.04 of 1-IoU) was also found between predicted segmentations (Y'Y').

GED	2*(Y'Y)	Y'Y'	YY
0.29±0.04	0.48±0.04	0.04±0.01	0.14±0.01

Table 7.5: GED nodule segmentation performance of the U-HPNet. Each score reports 1-IoU metric.

As a general observation about the results shown previously (Tables-7.1,7.2,7.3 and 7.4), we should mention that the best performances were always achieved with the radiologists' annotations closest to our predictions. This ground truth is as important as any of the others since, as mentioned in section 7.3.4, the annotations of each of the radiologists were assumed to be equally valid. Therefore, we already expected that our method would work somewhat better with this criterion since, by definition, for each tumour growth prediction we compared it with the radiologist's annotation closest to that prediction. Regarding the rest of the ground truths, the results obtained with the RX1 and RX2 annotations were found near to the best performances, while the results obtained with RX0 annotations and the mean of the radiologists' annotations were the lowest.

7.4.4 Ablation studies

An ablation study was made to isolate the effects of the different components of the U-HPNet using the radiologists' annotations closest to our predictions. Table-7.6 shows the different network setups evaluated and their acronym for better identification.

Table-7.7 shows the performances obtained for tumour growth prediction, size and segmentation for each of the network setups using their estimated means.

Acronym	$Loss_{rec}$	Attention	D0
BD0	BCE	✗	✓
ID0	IoU	✗	✓
IDD0	IoU+L1	✗	✓
IDAOD0	IoU+L1	✓	✗
U-HPNet	IoU+L1	✓	✓

Table 7.6: Different network setups of the U-HPNet configured in the ablation study.

	Prediction (Bacc)	Size ↓ (MAE)	Segmentation (Dice)
BD0	0.66±0.09	1.93±0.35	0.74±0.02
ID0	0.72±0.08	1.80±0.35	0.79±0.02
IDD0	**0.75±0.08**	1.84±0.37	0.80±0.02
IDAOD0	0.74±0.08	1.79±0.38	**0.81±0.02**
U-HPNet	0.74±0.08	**1.74±0.34**	0.78±0.02

Table 7.7: Performance comparison between the different U-HPNet setups.

An interesting observation to note from these results is that the configurations using IoU clearly outperformed the setup using BCE (BD0). Particularly, a rise of 0.09 in Bacc was achieved with the IDD0, an improvement of 0.19 mm in MAE was obtained with the U-HPNet and an increase of 0.07 in Dice score was reached with the IDAOD0. Also, the networks using attention (i.e. U-HPNet and IDAOD0) obtained the best performances in terms of MAE, in particular the U-HPNet obtained the lowest value with 1.73 mm. Regarding the Dice score, all networks using IoU loss obtained performances above 0.78, although the IDAOD0 with 0.81 was the one with the highest performance.

	Prediction (Bacc_2std)	Size (P(RX∈2std))	Segmentation ↓ (GED)
BD0	0.08±0.04	**0.87±0.05**	**0.24±0.02**
ID0	0.59±0.08	0.38±0.08	0.31±0.02
IDD0	0.64±0.08	0.33±0.08	0.31±0.03
IDAOD0	**0.67±0.08**	0.36±0.08	0.30±0.04
U-HPNet	0.57±0.08	0.44±0.08	0.29±0.04

Table 7.8: Generative ability comparison between the different U-HPNet setups.

Table-7.8 shows the performances of the generative ability of the different network configurations. The best option regarding prediction performance was IDAOD0 with 0.67 of Bacc_2std. The best network for size and segmentation was BD0, although it reported an unacceptable prediction performance of 0.08 in Bacc_2std due to a high variability in the generated samples. If we do not consider this option, the best option was the U-HPNet either in P(RX∈2std) and GED.

7.4.5 Comparison with other networks

We evaluated 4 different alternative deep networks for nodule growth estimation using the radiologists' annotations closest to our predictions, to enable their comparison with the proposed method. In particular, we evaluated 1 deterministic (U-Net) and 3 generative architectures (GAN-P2P, BAYES_TDO, SPU). Table-7.9 shows the performances obtained for these models regarding nodule growth prediction (Bacc), size (MAE) and segmentation quality (Dice) using the predicted value for the deterministic approach and, using the expected mean of the output distribution for the generative approaches.

	Prediction (Bacc)	Size ↓ (MAE)	Segmentation (Dice)
U-Net	0.64±0.09	2.94±0.43	0.77±0.02
BAYES_TD	0.67±0.08	2.29±0.45	**0.78±0.02**
SPU	0.73±0.08	2.14±0.46	0.77±0.01
P2P_GAN	0.69±0.07	2.62±0.43	0.71±0.02
U-HPNet	**0.74±0.08**	**1.74±0.34**	**0.78±0.02**

Table 7.9: Performance comparison with alternative networks for tumour growth using the expected mean.

From the four alternative methods, the SPU obtained the best Bacc score with 0.73 and MAE with 2.14 mm. In contrast, the BAYES_TD method obtained the best Dice score with 0.78. If we compare these results with the U-HPNet none of them could outperform their results neither in terms of prediction, size nor segmentation.

In Table-7.10, we summarize the performances regarding the generative ability to report accurate results. In particular, we provide nodule growth prediction using BA_2std metric, nodule size using P(RX∈2std) and estimated nodule segmentation using GED.

The best Bacc_2std score was 0.46 for the BAYES_TD, the best P(RX∈ $2std$) was for the SPU with 0.68 and the best GED with 0.25 mm for the GAN-P2P. If we compare these results with the U-HPNet, we observe that other methods showed better segmentation and size generative ability to capture the ground truth, however this made them to be less accurate with the lowest prediction performances.

	Prediction (Bacc_2std)	Size ↓ (P(RX∈2std))	Segmentation (GED)
BAYES_TD	0.46±0.08	0.49±0.08	0.27±0.03
SPU	0.28±0.06	**0.68±0.08**	**0.23±0.02**
GAN-P2P	0.26±0.07	0.67±0.08	0.25±0.04
U-HPNet	**0.57±0.08**	0.44±0.08	0.29±0.04

Table 7.10: Generative performance of alternative networks for tumour growth.

7.5 Discussion

With the aim of supporting radiologists in the early detection of lung cancer, we proposed a new predictive method capable of estimating tumour growth at a given time. In line with current clinical practice, our method predicts tumour progression when there is substantial growth (i.e., more than 2 mm) in the longest diameter of the pulmonary nodule [195]. Although this criterion is commonly used for its simplicity and applicability, it entails significant inter-observer [106] variability that may impact on the reliability of the predictive models. Along with the inter-observer variability, other inter-related factors may also have a direct impact on the trustworthiness of the estimator, such as the ambiguity, partiality or scarcity of the data to model. Therefore, in medical settings it is important that predictive models also provide a measure of uncertainty, which is especially of interest when complex or doubtful cases have to be assessed.

In this work we have taken this aspect into account, and we have built a predictive model capable of also estimating the associated uncertainty when predicting tumour growth. To do this, we collected a longitudinal dataset with more than 160 selected pulmonary tumours with two CT images per case (taken at different time-points), labelled by up to three different radiologists. To model these data, we opted for a generative deep learning approach as opposed to the deterministic approaches used to date [177, 311] for lung tumour growth prediction. The suitability of the generative approach was already proved in [231], where they modelled glioma tumour growth using an early probabilistic and generative framework [153] to estimate the tumour growth output distribution. Nonetheless, tumour growth prediction was not quantified, model uncertainty was not reported, and multiple observer variability was not addressed.

To address the aforementioned aspects, we relied on a more recent hierarchical generative and probabilistic framework [154] to estimate the output distribution of the future lung tumour appearance (at T_1) conditioned on the previous image of the nodule (at T_0). Our method (U-HPNet) extended this framework with the following modifications. First, we used smaller image patches (32x32) to focus on the tumour and its immediate surrounding tissues, and to reduce the number of parameters to be adjusted by the network. Second, we added two new features to the network to extract additional patterns from the tumour images: the time to predict, and the diameter of the nodule (at T_0). Third, we integrated an attention mechanism [217] in the decoder part of the network to boost its performance. Fourth, we proposed a new reconstruction loss function composed of the IoU and the L1 distance to provide more accurate segmentation and diameter estimations. Finally, we created a new post-processing module that applies Monte-Carlo sampling to estimate the mean and standard deviation of the tumour growth prediction, diameter growth and segmentation of a given nodule at a specific time.

We evaluated the U-HPNet using the annotations provided by 3 different radiologists, but also with their average and the radiologists' annotations closest to our estimates, to provide a more complete assessment of our approach and to detect possible divergences between the experts.

Regarding the evaluation of our approach using the expected values, the best results were obtained using the radiologists' annotations closest to our predictions. This ground truth criterion always reports real radiological annotations (specifically, the closest ones to our predictions), therefore since we take all radiologists' opinions equally, in a sense, this criterion is equally comparable to any of the three radiological criteria available in the study. In particular, we achieved 74% of tumour growth balanced accuracy (Bacc), 1.73 mm of diameter mean absolute error (MAE) and 78% of Dice score (Tables-7.1, 7.3, 7.4). Near to these results, we found the performances obtained with RX1 and RX2 annotations. Specifically, for RX1 we achieved 0.68 of Bacc, 2.52 of MAE and 0.77 of Dice score (Tables-7.1, 7.3, 7.4). Lower performances were found using the RX0 annotations and the mean of all radiologists, especially on tumours with a growth size greater than 2 mm and predictions over 24 months (Figure-7.9).

Compared to similar recent work in the literature [177], they reported higher balanced accuracy scores (86%) but much lower segmentation Dice scores (64%) than us. Results however are not fully comparable since both networks used different in-house cohorts, with different tumour case complexities, and both defined tumour progression differently, theirs relied on a tailored volumetric threshold and ours on the diameter growth convention established in radiological guidelines ([195]).

We also evaluated the ability of the network to produce consistent samples matching with the ground truths. To this end, we proposed different metrics (see section 7.3.4), i.e. the balanced accuracy for tumour growth prediction in an interval of 2 standard deviations (Bcc_2std), the probability of matching with the tumour growth size in an interval of 2 standard deviations ($P(RX \in 2std)$) and the generalized energy distance for tumour segmentation (GED). Our method achieved the best performances with the closest radiologists criterion, in particular 57% of Bacc_2std, and 44% of P(RX∈2std) (Tables-7.2,7.3). These values reflect that our approach still has room for improvement to make the estimated tumour growth sizes more accurate (i.e. bringing the tumour growth size mean closer to the radiologists ground truths). However, we should stress that these performances (as seen in Figure-7.9) were affected especially by complex cases with higher uncertainty (i.e. with a temporal prediction distance above 24 months). Different solutions could be applied to improve these performances, such as acquiring more tumour cases (e.g. especially on those cases where the method was not as accurate), using more aggressive data augmentation techniques (e.g. generating synthetic tumours); or using volume images, rather than single slices, to extract better predictive features. Breaking down the GED performance (Table-7.5), we observed that the network obtained 23% of segmentation variability between predicted and ground truths (Y'Y), being not far from 14% of inter-observer variability (YY). Also, the network showed a relatively small variability of 4% between the generated sample segmentations (Y'Y'). This may indicate that the network, during training, preferred to concentrate the predictions around the mean rather than predict highly disperse values in order to optimize performance.

For a better understanding of the effects of the main components of the network, we provided an ablation study with different network configurations using the closest radiologists criterion (Table-7.6). From this analysis, we obtained that the largest improvement was achieved replacing binary cross entropy (BCE) by IoU in the reconstruction loss. This can be observed by comparing BD0 and ID0 networks. Specifically, the Bacc increased approximately 7%, MAE decreased almost 0.2 mm, and the Dice improved to nearly 5%. Moreover, the Bacc_2std raised to almost 60% (Table-7.7). Different reasons may explain the suitability of using IoU for this problem. First, this loss is robust to data unbalance. Second, IoU had values with similar magnitude to the KL-divergence distance, allowing a better optimization of the network than using BCE. Despite the benefits of using IoU, we realized that the P(RX∈2std) and GED decreased significantly due to higher variability around the estimated mean. A second network configuration (IDD0) allowed improving previous performance limitations. In particular, this network incorporated the L1 distance between the predicted and ground truth diameters in the reconstruction loss together with IoU. Results showed that the IDD0 network increased its growth prediction performance (3% in Bacc and 5% in Bacc_2std) and segmentation ability (1% in Dice and GED), despite slight decrease of performance in diameter growth prediction (0.05 mm in MAE and 4% in P(RX∈2std)). Adding attention (current U-HPNet network) in the decoder part of the sub-networks, outperformed the IDD0, precisely, reducing 0.1 mm in MAE and increasing 10% in P(RX∈2std). However, it implied a certain increase also in the estimated diameter growth variability, reducing 1% of its Bacc and 7% of Bacc_2std. A final comparison was performed between IDAOD0 and U-HPNet to obtain the importance of adding nodule diameter (at T_0) in the input of the U-HPNet. According to the results, using this feature we reduced 0.6 mm of MAE, increased 8% the P(RX∈2std)), and consequently improved the Bacc almost 2%. This reflects that using D0, the network was able to predict more accurately the diameter growth of the nodule. However, this feature increased the estimated diameter growth variability, resulting in 9% decrease of Bacc_2std and 2% of Dice.

Due to the lack of similar studies for early lung tumour growth prediction, we built different alternative networks to allow their comparison. In particular, we proposed a deterministic (U-Net), and 3 different generative networks: Bayesian dropout (BAYES_TD), probabilistic U-Net (SPU), Pix2Pix GAN (P2P_GAN). The comparison was performed using the closest radiologists criterion. Results from Tables-7.9,7.10 showed that, using the estimated sample means, the generative approaches outperformed the performances reported by the deterministic network (U-Net). This result consolidates the suitability of the generative approach for this type of problem. Also, among all generative methods, the U-HPNet obtained the best performance metrics using the estimated means (i.e. in tumour growth prediction, size and segmentation). Regarding the metrics measuring the generative ability of the networks, the U-HPNet obtained the best Bacc_2std although the poorest P(RX∈2std). In contrast, the SPU reported a large sample variability in tumour growth size as shown by the highest performance in P(RX∈2std) and GED, but one of the lowest performances in Bacc_2std. The GAN-P2P, similarly to the SPU, obtained high P(RX∈2std) and GED, but poor Bacc_2std due to high variability in the sample distribution of tumour growth size. The BAYES_TD in contrast obtained lower variability in tumour growth size, achieving better P(RX∈2std) and GED than the U-HPNet but lower Bacc_2std. Interestingly, BAYES_TD and U-HPNet networks reported rather similar generative performances, despite employing two different ways to generate samples (by weight randomization and by randomly selecting a vector in the latent space). Thus, combining both approaches could help to disentangle different types of uncertainties (as in [126] for tumour segmentation) and disclose a potential increase in the performance of the network.

Our method still has a number of limitations. First, the number of cases analysed in this study was low, which clearly impacted on the reported performances of our approach due to its data eager nature. However, this data was clinically validated, and selected by different radiologists according to their relevance and interest. Second, segmentations were generated semi-automatically according to the original diameter, growth and centroid annotations with a final visual expert validation. However, we believe using manual expert segmentations could make our method more precise, especially in the contour of the tumours. Third, our method relied on a single axial slice of the tumour to predict tumour growth. However, tumour growth is a tri-dimensional biological process, hence using volumetric images may allow capturing further relevant features and patterns to explain better the tumour progression. Nevertheless, using 2D information made our solution more compact, with fewer parameters to fit, faster to train and more suitable for smaller datasets.

Beyond this work, further efforts in fine-tuning the proposed approaches are required such as exploring different number of layers, latent hierarchies, loss weights factors and other optimization parameters. Also, future extensions are suggested along this chapter, such as exploring a 3D version of the network, deepen in the uncertainty ability of the network, evaluate its integration with Bayesian dropout or adversarial learning, and incorporate the newest advances in deep learning to extract better spatial and temporal features from the tumours.

7.6 Conclusion

In this chapter, we addressed early lung tumour growth prediction as a multimodal output problem, as opposed to existing solutions that provide deterministic outputs. Several reasons motivated our decision, such as the complexity of the problem, the inter-observer variability, or the importance of estimating uncertainty in medical settings. To this end, we adapted an existing deep hierarchical generative and probabilistic framework to encode the initial image of the nodules in a continuous multidimensional latent space, to sample from it, and to generate multiple consistent future tumour segmentations conditioned on the given nodule.

Our network (U-HPNet) extended the original framework with the intention to predict and quantify tumour growth, as well as to visualize the future semantic appearance of the tumour. Therefore, we added new context features (i.e. the time to predict, and the initial nodule diameter measured by the specialist), we used a new reconstruction loss (combining IoU and diameter distance), and we integrated an attention mechanism in the decoder parts of the network. Finally, we attached a new post-processing module on the network to perform Monte Carlo sampling, and retrieve the estimated tumour growth probability, size and segmentation, along with their associated uncertainty.

The network was trained and evaluated on a longitudinal cohort with more than 160 cases. Best performances reported a tumour growth balanced accuracy of 74%, a tumour growth size MAE of 1.77 mm and a tumour segmentation Dice score of 78%. Finally, we compared the performance of our method with 4 different networks based in a U-Net, probabilistic U-Net, Bayesian dropout and Pix2Pix GAN. The U-HPNet outperformed the proposed alternatives for tumour growth prediction, size and segmentation.

Chapter 8

CONCLUSIONS AND FUTURE WORK

8.1 Overview

The main goal of this thesis was to provide useful and effective image analysis tools, relying on recent advances of deep learning, to support the current clinical workflow involved in the management of lung cancer disease. The motivation lies in the complexity and cost of accurately detecting and diagnosing pulmonary nodules in early stages through visual inspection of computed tomography (CT) lung images, as well as the lack of tools to support physicians in the follow-up of suspicious lung tumours. In the following paragraphs, a summary of the main contributions introduced in each chapter of the thesis is provided, highlighting the strengths and limitations that need to be further addressed.

In **Chapter 4**, we addressed the problem of accurately predicting lung nodule malignancy given beforehand the location of the nodules, as well as the challenge of automatically providing lung cancer prediction at the patient level in an end-to-end manner given a raw lung CT image study. Traditional solutions for automatic lung nodule malignancy classification [93, 336, 337] have been approached through machine learning algorithms based on nodule image descriptors [67, 242]. However, these approaches usually show generalization problems, specially on doubtful and infrequent cases. In this work, inspired by outstanding performances on nodule malignancy classification using 2D and 2'5D CNNs [72, 145, 334], we investigated the use of 3D CNNs for this problem, achieving radiologist performances and confirming the suitability of this approach. Beyond learning accurate CNNs, only in [265] tackled the concept of transferring deep nodule features for lung cancer prediction. Consistent to this idea, we developed an integration framework to enable transfer learning from nodule malignancy models, for which data are more abundant and labels cheaper to obtain, to predict lung cancer (diagnostically confirmed) at the patient level, usually more costly and difficult to get. To validate our approach, we built a basic two-stage lung cancer pipeline able to detect lung nodules and provide lung cancer prediction at the patient level. Several works have addressed the lung nodule detection problem using adjustable and agile conventional image analysis techniques [80, 303], such as wavelet feature descriptors [221]. Recently, accurate but data demanding deep object detection networks, such as Faster R-CNN [244], have been adapted for tumour identification, surpassing performances of prior approaches. In this work, we proposed a two-steps solution combining the flexibility of conventional image analysis techniques (i.e. 3D Laplacian of Gaussian [86] filters) and the classification performance of a CNN (i.e. a 3D ResNet-50 [114]) for detecting lung nodules. For the second stage of the pipeline, conventional machine learning algorithms were trained using few radiological image descriptors (i.e. size, texture and morphology) of the nodules, extracted during the detection stage, to predict lung cancer classification at the patient level. The successful results of integrating the CNN for nodule malignancy classification into the pipeline confirmed the convenience of the approach. Nevertheless, some limitations were identified in this work. First, the pipeline was configured with a basic two-stage nodule detector, using a more modern deep object detector [244], as we already did in **Chapter 5**, we could obtain better performances. Second, the nodule malignancy CNN had a shallow architecture, thus more recent and deeper architectures could obtain better results. Third, we used a naive but intuitive algorithm to provide lung cancer prediction for cases with multiple nodules; more elaborated strategies could be envisioned (e.g. local causal probability models [228]) to improve the predictions of the pipeline. Fourth, using further radiomic descriptors as well as clinical history data of the patient could potentially enhance the reported lung cancer classification performances of the pipeline. Despite these limitations, the work presented in this chapter was, to the best of our knowledge, the first attempt to build a nodule-malignancy/patient-cancer integrated framework.

In **Chapter 5**, we made a step forward in our research by incorporating the temporal dimension into our goal of providing automatic lung cancer assessment. Despite the medical importance of monitoring the evolution of pulmonary nodules for determining their malignancy likelihood, few works [14, 87, 308] have really taken into account the temporal dimension to provide a malignancy estimate. In addition, most of these works rely on a subset of the NLST (accessible under prior committee agreement), which is probably the largest longitudinal lung cancer dataset. However, cases from this cohort are not publicly available and are constrained to certain parameters (e.g. yearly CT scans on a subset of the population), which limits the complexity of the data to analyse. In this work, given the unavailability of open longitudinal lung cancer datasets, we collected a rich and heterogeneous longitudinal dataset composed of more than 150 confirmed cases with two CT scans of the same patient taken at two different time-points, from the Vall d'Hebron Hospital. Once the data was appropriately stored and anonymized, we focused on a way to automatically re-identify nodules located at different lung CT studies of the same patient. Typically, this problem is addressed through image registration processes [39], consisting on aligning the prior and current lung CT scan images of the patients. However, several factors compromise the effectiveness of the registration process, such as the variability in the image size and resolution originated by the use of different CT scans, and the variability in the position and breath cycle of the patients when performing the scanning. Current medical image registration methods [276], especially non-linear [251], report accurate CT alignments. However, they are still slow and potentially introduce some distortions in the intrinsic structure of the lung, hindering their wide clinical acceptance [309]. To overcome these limitations, we proposed a fast and accurate solution that does not require having lung CT images previously registered, avoiding some of the shortcomings that it entails. In particular, our solution consisted on adopting a 3D Siamese neural network (SNN) [152]. SNNs have been extensively used in computer vision matching problems, such as tracking objects in videos [296], or in the medical image domain, to extract a latent representation for content-based image retrieval [49]. However, to the best of our knowledge, SNNs had not been applied before to re-identify nodules in a series of lung CT scans. Thus, for this problem, we configured several SNNs to find the most suitable one for our problem. To do this, we made emphasis on the use of transfer learning, namely configuring the backbone of the different SNNs with the architecture and weights of the 3D ResNet built for nodule classification, a problem for which we had a larger set of data compared to that of the temporal evolution of pulmonary nodules. Despite achieving state-of-the-art performances for nodule re-identification, either in a standalone mode or integrated within a developed pipeline aimed at automatically providing nodule growth quantification, some limitations were discovered in our approach. The SNN struggled to identify nodules that had undergone a large change in their size. Although this was infrequent in our cohort, different simple solutions were envisaged to overcome this limitation, such as stronger focus on data augmentation, adding the nodule location in the network or using larger patch images. A couple of limitations were encountered in the data collection process, and thus also affecting our subsequent studies. First, data annotations for nodule quantification were based on the major axial diameter. Although the diameter is the most common radiological measure used in practice for nodule growth assessment, using 3D measurements (e.g. volume) could lead to a more accurate quantification. Second, the cohort size was small, which, as seen in the literature, is a common occurrence. Gathering large longitudinal lung cancer cohorts based on incidental cases is difficult due to the asymptomatic nature of the disease, and the overloaded radiological units of the health institutions. Population screening based cohorts have usually more voluminous sets of cases, which makes them more suitable for deep learning, however their access is restricted, their management is assessed by different clinical guidelines [12], and the case heterogeneity is reduced (e.g. larger number of controls, prone to contain common malignancies, fixed timings between studies). Nonetheless, the automated re-identification of regions of interest in medical images over time, without the need to modify the image structure, could be an appealing application beyond lung cancer assessment such as therapy follow-up as well as for different diseases located in other organs (e.g. prostate, breast cancer) in the body.

In **Chapter 6**, we presented a novel pipeline, relying on deep learning, for supporting the automatic analysis of the lung nodule follow-up. In particular, we extended the pipeline proposed in **Chapter 5** to provide in an end-to-end fashion detection, re-identification, growth quantification and cancer classification of the nodules given a pair of CT image studies (e.g. prior and current) of the patient. Few works have tried to automatize the radiological workflow, taking into account the temporal evolution of nodules [14]. However, to the best of our knowledge, this is the first attempt to cover nodule growth quantification and cancer prediction in an end-to-end approach. The methodological novelty part of this pipeline, compared with the one presented in **Chapter 5**, lies in the nodule growth quantification and cancer classification components. Specifically, for the nodule growth component, instead of relying on the outcomes of the nodule detection network (i.e. 3D Faster R-CNN, built in Chapter 5), we proposed a nodule segmentation network to provide more accurate measurements. Also, since lung CT images are ambiguous, mostly caused by the image acquisition protocol, we extended this network to provide a measure of the uncertainty of the model on their estimations. Providing uncertainty in the predictions of a model is important in the medical domain. This allows clinicians to make crucial subsequent decisions more safely as well as increase their reliability on the predictive models. One of the alternatives to provide uncertainty for image segmentation in deep learning is ensembling multiple networks in order to provide multiple opinions [164]. However, diversity is not assured and limited by the subset of models. Other common approach is using dropout [279] at inference time in order to provide independent pixel-wise probabilities [142]. However, providing only pixel-wise probabilities ignores co-variances between pixels, which may drive to inconsistent results. In this work, we relied on a recent 2D hierarchical probabilistic U-Net (HPU) [154] able to generate multiple consistent nodule segmentations. This network combines the generative and probabilistic capability of conditional variation autoencoders [275] with the segmentation ability of U-Nets [248]. The HPU improves a previous work [153] with a hierarchy latent space formulation that enables modelling ambiguities at different scales. Our approach extended the HPU network with a Monte-Carlo sampling post-processing to estimate the lung nodule diameter mean and its uncertainty. This approach slightly outperformed our previous solution based on a deterministic nodule detection network for measuring nodule growth. Somehow, this result was expected since the nodule detection network was trained without any information regarding the contour of the nodules. Nonetheless, both approaches reported errors that were below to 2 mm, the threshold determined by radiological guidelines from which to consider nodule growth [195]. Further extensions of this method are envisioned for future work, such as implementing a 3D version of the network to obtain better spatial image representations of the nodule growth, or using 3D metrics (such as volume) to provide more precise and stable nodule growth measures. To enable the pipeline to predict the lung cancer probability of a nodule given its temporal evolution, we found few deep learning related works, mostly relying on data from lung cancer screening programs. For instance, in [87] they aggregated in the forget and input gates of an LSTM for lung nodule cancer prediction, a method to weight the importance of the temporal distance between scan images. In [14], they proposed an end-to-end deep learning-based pipeline for which, the cancer risk model, was composed by a two stream 3D CNN networks to analyse patches from the prior and current lung CT images. More recently, in [308] an attention-based 2D CNN network was built using pre-trained weights and a multi-time-point classification in a Siamese structure. Our approach was similar to [14], proposing a 3D two-stream CNN (TS-3DCNN-MAL) able to predict lung cancer nodule probability, given two images of the same nodule taken at different time-points. However, given the limited training data for building from scratch this network, the sibling architecture of this network was re-used (including the weights) from the 3D ResNet built for nodule classification (**Chapter 4**). In addition, into the head component of the TS-3DCNN-MAL network, we concatenated the malignancy probability predictions reported by a 3D CNN built for nodule malignancy prediction (presented in **Chapter 4**). Results from the evaluation of the TS-3DCNN-MAL network surpassed the results of the networks using a single time-point image, and they reported state-of-the-art performances similar to approaches trained with more voluminous datasets. Despite the high performances achieved, we strongly recommend evaluating this network and the whole pipeline on a larger and multicentric dataset. In addition, further efforts could be done on visualization and interpretability aspects to allow a better understanding of how the models of the pipeline behave, and thus an easier implantation of this tool in clinical domains.

In **Chapter 7**, we faced the challenge of forecasting lung tumour growth. Traditionally, tumour progression has been addressed through complex and sophisticated mathematical models [254], such as those based on the reaction-diffusion equation [287, 292]. However, these models are limited in number of parameters (e.g. 5 in [318]), which might not be sufficient to capture the growing patterns of the tumours. Recently, deep learning has been used to predict future tumour growth, surpassing performances reported by traditional approaches [328]. This kind of solutions usually relies on architectures combining CNNs and LSTMs [71] or using generative networks, such as those based on adversarial learning [96] and variational auto-encoders [148] to estimate the future image of the tumour [76, 231]. In the lung cancer domain, to the best of our knowledge, few deep learning-based works [177, 311] have been proposed. Moreover, despite the multifactorial and complex nature of the problem, uncertainty in the lung tumour growth was not addressed in any of the aforementioned studies. In **Chapter 7**, we adapted the HPU [154], to provide a novel tumour growth forecasting network (e.g. adding context features, incorporating attention gates in the decoder, and using a new reconstruction loss function). Hence, the proposed deep neural network allows producing, given a single image of the nodule, a growth prediction, a size estimation and a future semantic segmentation of the nodule at a given time. In addition to this, we added into the network a new component, based on Monte-Carlo sampling, to compute the uncertainty associated to each of the predictions. We performed an ablation study to quantify the gains of each of the integrated features of the model. Also, we provided a comparative study implementing alternative deterministic (i.e. U-ResNet [331]) and generative state-of-art networks adapted to this specific problem, such as a probabilistic U-Net [153], a conditional generative adversarial network [134] and a Bayesian dropout network [142]. The generative aspect of the network led us to provide two kinds of evaluations, one regarding the performance of the estimated predictions and another regarding the ability of the network to produce consistent samples matching with the ground truths. Although our approach reported better performances with respect to the rest of tested alternatives, some limitations were found in this solution. First, the ground truth segmentations were generated semi-automatically according to the original diameter, growth and centroid annotations with a final visual expert validation. However, we believe that using manual expert segmentations could make our method more precise, especially in the contour of the tumours. Second, the method relied on a single axial slice of the tumour to predict tumour growth. Nonetheless, tumour growth is a tridimensional biological process, hence using volumetric images may have allowed capturing further relevant features and patterns to better explain the tumour progression. Third, given the limited size of the cohort, further evaluation of the approach in a larger dataset could help to verify the robustness and correctness of the solution. Despite all these limitations, the presented solution provides some unique characteristics that makes it useful for the clinical practice, such as the combination of numerical and visualization results, along with a measure of network reliability in these predictions. A further aspect worth investigating could be the disentanglement of the latent features of the network. This could provide advanced clinical functionalities, such as giving additional control of the factors/conditions to explore the future growth (or appearance) of the nodules.

Beyond this specific application, the outlined innovations, such as the methods for integrating CNNs into computer vision pipelines, the re-identification of suspicious regions over time based on SNNs, without the need to warp the inherent image structure, or the proposed deep generative and probabilistic network to model tumour growth considering ambiguous images and label uncertainty, they could be easily applicable to other types of cancer (e.g. pancreas), clinical diseases (e.g. Covid-19) or medical applications (e.g. therapy follow-up).

8.2 Outlook and future work

In the previous section, we summarized the presented work towards temporal lung nodule assessment, highlighting its strengths and limitations as well as pointing out specific future work. However, there is much work ahead to make the proposed methods widely used in the daily clinical practice. In the following section, we detail each of the foreseen tasks.

Integration with other data types

The predictive models presented in this work (e.g. nodule detection, re-identification, malignancy prediction and growth forecasting) could be extended/combined with other kinds of data, such as the clinical history of the patient (e.g. smoking, medical history, family cancer and alcohol antecedents), genomics data (e.g. genetic variants associated with lung cancer) [300] and/or other imaging studies (e.g. X-rays, PET, MRI) [161]. Interactions between these data, may uncover novel patterns and thus increase the predictive ability of the resulting models [332]. A straightforward way to integrate these data could be done by concatenation at specific levels of the networks. However, further research is required for fusing and filtering these data on the models.

Emphasis on fine-tuning the models

For time reasons, we did not explore all the parameter possibilities when training the proposed deep neural networks in our experiments. Therefore, it would be highly advantageous to perform further fine-tuning of the proposed networks using different parameter configurations (e.g. input size, number of layers, kernel size, dropout percentage) and different meta-learning parameter settings (e.g. learning rate, batch size or epochs) from those used during the training of the networks.

Model interpretability

To improve the reliability of the provided tools, more efforts should be made towards being able to explain the reasons behind the reported predictions. Currently, there already exist several mechanisms that make the networks more interpretable, such as those based on back-propagation and those based on input perturbation. In the former ones, the signal from the output can be propagated back to the input layer, and several approaches have been provided to capture which features have more importance. Usually, these techniques output a heat-map overlapped with the original image. This allows checking which parts of the input were more involved in the model outcomes [205]. In the perturbation-based approaches, a portion of the input is changed and the effect of this change on the model output is observed [99]. These methods have high computational complexity, but they are easy to visualize. We have already presented some interpretation work for lung nodule malignancy classification by implementing the grad-CAM technique [258]. However, we believe further efforts are required for providing, for instance, a holistic interpretation of all decision/predictions made by the pipeline in an easy and accessible way to the clinicians.

Deepening on uncertainty estimation

Modelling uncertainty is an important feature in medical scenarios since it informs clinicians about the trustworthiness of a model's outputs, making them safer and more reliable [143]. In **Chapter 6 and 7**, thanks to the generative and probabilistic ability of the proposed models, we provided a measure of how uncertain the models were on their predictions (e.g. tumour growth quantification and forecasting). To do this, we relied on a Monte-Carlo sampling post-processing, consisting on generating an enough number of future nodule reconstructions from the proposed networks to estimate the mean and standard deviation. Further work could be devoted to distinguish the different types of uncertainty [65], typically epistemic (i.e. knowledge uncertainty) and aleatoric (i.e. data uncertainty), to enhance interpretability of the results and/or to potentially reduce model (epistemic) uncertainty [126].

Given the importance of modelling uncertainty for detecting low confident predictions, further work should be addressed on the rest of proposed networks (e.g. nodule detection) to enhance their potential acceptability in the clinical practice. For instance, in [210] they applied a threshold on the uncertainty measures computed for each of the detected candidates for Multiple Sclerosis lesion detection. Beyond this task, other applications using uncertainty can be foreseen for lung cancer assessment, such as rejecting cases with higher predictive uncertainty [84], detecting outliers depending on the estimated uncertainty, extracting the most robust feature embeddings from the network, or in general applying uncertainty for learning better models (e.g. filtering small part of training samples with the highest predictive uncertainty [88]).

Increasing the amount of data

Having more annotated data would be highly recommendable and advantageous for improving current work. These data could be partially used to fine-tune, retrain and evaluate the presented methods. Therefore, having extra amount of data would help to enhance performance and robustness of the methods, to improve the reliability of the models for the clinical practice, as well as to be able to extract further conclusions from the predictive methods.

Beyond allocating more resources to extend the number of collected cases of the VH cohort, it is also equally important to ensure that the variability of the problem is well represented (e.g. using different acquisition protocols, different nodule sizes, morphologies and textures) in the cohort to build robust and reliable predictors. Additionally, it would be highly recommendable to extend the cohort with further time-points for each of the cases. This would permit learning better the temporal evolution of the nodules and capturing better representative patterns and predictive features.

Because of the cost of collecting and annotating new data, further emphasis could be put on alternative technical solutions to overcome this limitation. First, we could investigate more recent transfer learning techniques from existing models (e.g. from different medical disciplines or domains) to our problem. Second, we could use more advanced techniques for data augmentation (e.g. mix-up [327] to randomly interpolate both inputs and labels) in order to improve the generalization ability of the models. Third, we could spend further efforts on recent methods able to generate realistic synthetic data (e.g. based on GANs [96]). Lastly, another technique worth exploring consists on having the data itself to provide their supervision [69] (e.g. withhold some information about the data, and task the network with predicting it). This is known as self-supervised learning, and it is a type of unsupervised learning, which is recently gaining a lot of attention from the computer vision community.

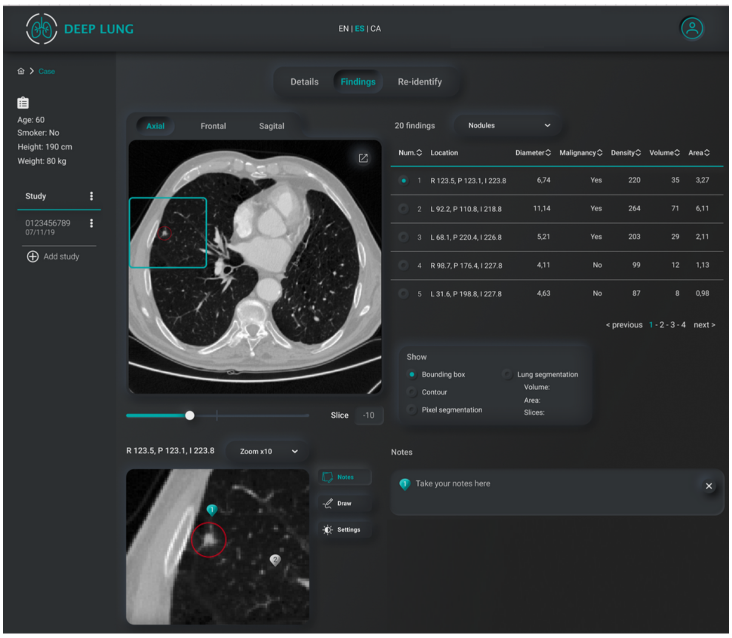
Pilot test

Although we have exhaustively evaluated each of the different proposed methods both individually and integrated in a pipeline, we have not tested them in a clinical setting. Thus, it is part of future work to validate the effectiveness of the pipeline in a realistic scenario. To do this, a pilot test should be conducted on a specifically designed cohort of patients and for a limited period of time. For instance, two groups of patients could be randomly selected, and only one of them would be further managed through the support of the provided pipeline.

Graphical user interface

Related with the previous task, with the support of the digital health and software development units of Eurecat, we are currently working on a graphical user interface to wrap-up the functionalities proposed by the temporal lung cancer assessment pipeline [238].

This web interface has the goal to demonstrate the viability and facilitate the uptake of the current solution. To this end, we aimed to provide an intuitive and informative front-end that should allow clinicians to be more effective and efficient in their day-to-day routines. Precisely, the front-end is organized to summarize in few screens all crucial information and main functionalities resulting from the automatic temporal lung nodule analysis.



Age: 60
Smoker: No
Height: 190 cm
Weight: 80 kg

0123456789
07/11/19

Slice -10

R 123.5, P 123.1, I 223.8

20 findings

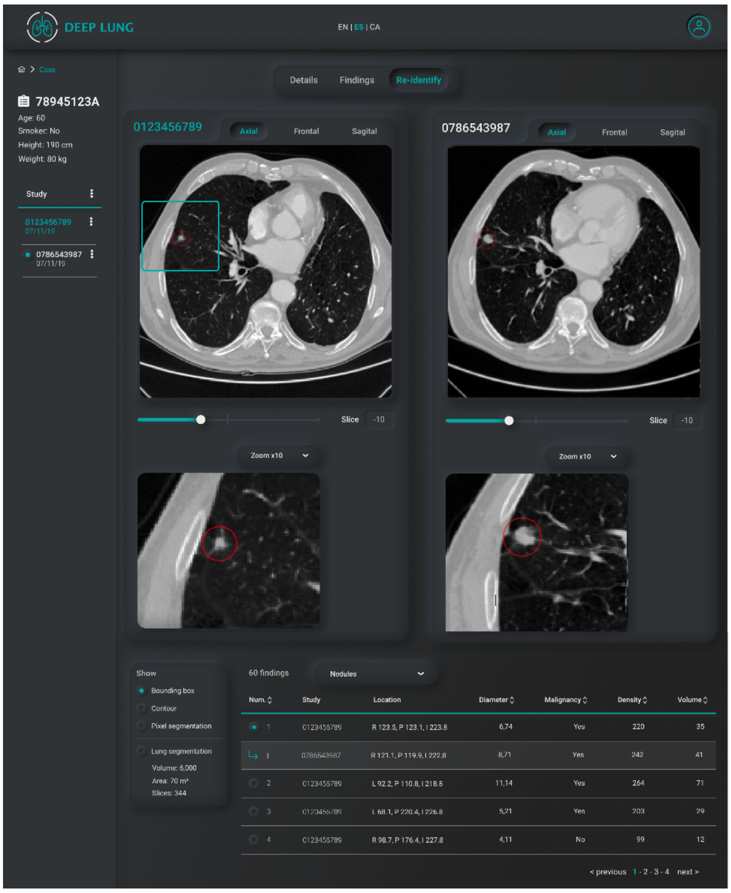
Num.	Location	Diameter	Malignancy	Density	Volume	Area
● 1	R 123.5, P 123.1, I 223.8	6,74	Yes	220	35	3,27
○ 2	L 92.2, P 110.8, I 218.8	11,14	Yes	264	71	6,11
○ 3	L 68.1, P 220.4, I 226.8	5,21	Yes	203	29	2,11
○ 4	R 98.7, P 176.4, I 227.8	4,11	No	99	12	1,13
○ 5	L 31.6, P 198.8, I 227.8	4,63	No	87	8	0,98

● Bounding box
○ Contour
○ Pixel segmentation
○ Lung segmentation
Volume:
Area:
Slices:

Take your notes here

Figure 8.1: Findings section window of the front end interface with the results of a CT study.

Figure-8.1 shows the window to present the findings detected on a single CT scan image. In the main panel of this screen, we find, on the left side, a slice navigator with the axial slice of the lung in which a nodule was found. On the right side, we find a clickable table with the nodules detected by the pipeline. The table also shows information regarding the location, diameter, ”subjective” malignancy, and other related features computed from the pipeline results (e.g. nodule area in the slice, volume, and density). Below the slice navigator, we observe a zoom component, which shows an augmented view of the area of interest in which the nodule was found. This zoom component is updated every time the user moves the cursor above the main slice navigator. The nodules are automatically surrounded by a red circle. Other extra functionalities are available for the radiologists such as showing the pixel segmentation of the nodule, adding comments directly on the slice image or a full-screen view.



78945123A
Age: 60
Smoker: No
Height: 190 cm
Weight: 80 kg

Study
0123456789
07/11/19
0786543987
07/11/19

● Bounding box
○ Contour
○ Pixel segmentation
○ Lung segmentation
Volume: 6,000
Area: 70 m²
Slices: 344

60 findings

Num.	Study	Location	Diameter	Malignancy	Density	Volume
● 1	0123456789	R 123.5, P 123.1, I 223.8	6,74	Yes	220	35
↳ 1	0786543987	R 121.1, P 119.9, I 222.8	8,71	Yes	242	41
○ 2	0123456789	L 92.2, P 110.8, I 218.8	11,14	Yes	264	71
○ 3	0123456789	L 68.1, P 220.4, I 226.8	5,21	Yes	203	29
○ 4	0123456789	R 98.7, P 176.4, I 227.8	4,11	No	99	12

Figure 8.2: Screen with the results of the re-identification analysis. In this particular example, we observe the same nodule found in both CT studies, its growth and its estimated malignancy.

Figure-8.2 shows the screen with the results of the temporal evolution of a selected nodule. The re-identified nodules can be examined directly using the two slice navigators located on the centre of the screen. A table with the detected and re-identified nodules from both studies is also shown on the bottom part of the screen. Furthermore, the screen also shows the exploration capabilities of the tool by showing a zoom component, updated when the user moves the cursor over the slice navigator.

8.3 Publication list

Journal papers

- **Rafael-Palou Xavier**, Bonavita Ilaria, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline. Computer Methods and Programs in Biomedicine. Vol. 185 (105172), pp. 1-9, 2020.

- **Rafael-Palou Xavier**, Aubanell Anton, Bonavita Ilaria, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. Re-Identification and Growth Detection of Pulmonary Nodules without Image Registration Using 3D Siamese Neural Networks. Medical Image Analysis. Vol. 67 (101823), pp. 1-12, 2021.

- **Rafael-Palou Xavier**, Aubanell Anton, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. An Uncertainty-aware Hierarchical Probabilistic Network for Early Prediction, Quantification and Segmentation of Pulmonary Tumour Growth. *Under review in Medical Image Analysis*.

Book chapters

- **Rafael-Palou Xavier**, Aubanell Anton, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. Detection, growth quantification and malignancy prediction of pulmonary nodules using deep convolutional networks in follow-up CT scans. *To appear in Artificial Intelligence in Cancer Diagnosis, Volume 1: Lung and Kidney Cancer*.

Conference papers

- **Rafael-Palou Xavier**, Bonavita Ilaria, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. Improving Lung Cancer Prediction with a Deep Learning Nodule Malignancy Classifier. International Journal of Computer Assisted Radiology and Surgery. Vol. 14 (Suppl. 1): 70-71, 2019.

- **Rafael-Palou Xavier**, Aubanell Anton, Bonavita Ilaria, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. 3D Siamese Neural Networks for Matching Pulmonary Nodules in Series of CT Scans. International Journal of Computer Assisted Radiology and Surgery. Vol. 15 (Suppl. 1), 2020.

- **Rafael-Palou Xavier**, Aubanell Anton, Bonavita Ilaria, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A. Pulmonary Nodule Malignancy Classification Using its Temporal Evolution with Two-Stream 3D Convolutional Neural Networks. International Conference on Medical Imaging with Deep Learning (MIDL), 2020.

Invited talks

- **Rafael-Palou Xavier**, Pezzano Giuseppe, Bonavita Ilaria, Subías Paula, Aubanell Anton, Pallissa Esther, Persiva Oscar, Ruiz Laura, Ribas Vicent. Deeplung: Lung Cancer Detection with Deep Learning. Big Data Congress, 2019.

- Aubanell Anton, **Rafael-Palou Xavier**, Persiva Oscar, Varona Diego, Sánchez Ángel, Ribas Vicent. Deeplung: detección precoz del cáncer de pulmón a través de técnicas de inteligencia artificial. VIII Congres Nacional de Radiolegs de Catalunya, 2019.

- Aubanell Anton, **Rafael-Palou Xavier**, Persiva Oscar, Varona Diego, Sánchez Ángel, Ribas Vicent. DeepLung: inteligencia artificial para detectar nódulos pulmonares malignos. 35 Congreso Nacional de Radiología, ISBN - 978-84-09-30018-1, 2020.

- **Rafael-Palou Xavier**, Aubanell Anton, Pezzano Giuseppe, Domínguez Rubén, Rodríguez Ángel, Capdevial Sandra, Subías Paula, Ruíz Laura, Ceresa Mario, Piella Gemma, Ribas Vicent, González Ballester Miguel A, Miralles Felip. DeepLung-CT: A system for automatic temporal analysis of pulmonary nodules for cancer assessment through CT scans. Presentation of proof of concepts at Centre of Innovation for Data tech and Artificial Intelligence, 2021.

Other journal publications

- **Rafael-Palou, Xavier**, Turino Cecilia, Steblin Alexander, Sánchez-de-la-Torre Manuel, Barbé Ferran, and Vargiu Eloisa. ”Comparative analysis of predictive methods for early assessment of compliance with continuous positive airway pressure therapy”. BMC Medical Informatics and Decision Making 18, no. 1 (2018): 1-14.

- Valdés, María Gabriela, Galván-Femenía Iván, Ribas Ripoll Vicent, Duran Xavier, Yokota Jun, Gavaldà Ricard, **Rafael-Palou Xavier**, and de Cid

Rafael. ”Pipeline design to identify key features and classify the chemotherapy response on lung cancer patients using large-scale genetic data”. BMC Systems Biology 12, no. 5 (2018): 55-74.

Appendix A

SUPPLEMENTARY MATERIAL I

A.1 Significance of radiologists' annotated features

We discussed in the paper our choice to restrict the radiomics features included in the pipeline to a minimal set of classical nodule characteristics. For the sake of completeness, we report here the statistical significance of the complete set of features detected by radiologists through visual inspection (reported in the LIDC annotation file). To find the p-values we performed a generalized linear model with the annotated features as predictors and the five malignancy categories as outcome.

Nodule Feature	P-value
diameter (mm)	<2e-16
subtlety	0.002223
internal structure	0.646398
calcification	<2e-16
sphericity	0.000756
margin	0.005173
lobulation	4.42e-08
spiculation	8.16e-07
texture	0.165262

Table A.1: Statistical significance of the radiologists annotated features. P-values are obtained through Wald test of significance.

A.2 The lung cancer pipeline

Here we present the results of the first stage of the lung cancer pipeline. Those were obtained using an independent testset (10% of the data) of the LUNA16 dataset.

Nodules detection

Tables A.2 and A.3 show the description and results of three different configurations tested for the nodule detection part of the lung cancer pipeline.

	Option 1	Option 2	Option 3
Minimum radius	10	5	5
Maximum radius	40	30	60
Steps	10	10	5
Threshold	0.2	0.15	0.15
Overlap	0.9	0.7	0.9

Table A.2: Configurations of the Difference of Gaussian method for lung nodules detection.

	DoG configurations		
	Option 1	Option 2	Option 3
Total detected nodules	21	29	73
Total detected candidates	1142	7130	76631
Min,Max,Mean,Std radius of detected nodules (real)	3.51 12.14 8.03 2.33	2.8 12.14 6.82 2.74	1.7 12.14 4.69 2.65
Min,Max,Mean,Std radius of detected nodules (predicted)	5.0 12.6 8.6 2.31	3.05 12.29 7.33 2.74	2.5 8.66 4.41 2.35
Min,Max,Mean,Std intensity of detected nodules (pred)	0.21 0.45 0.31 0.07	0.16 0.57 0.32 0.13	0.15 1.31 0.46 0.3
Total missing nodules	84	76	32
Min,Max,Mean,Std radius of missing nodules (real)	1.64 8.36 3.2 1.16	1.64 8.36 3.15 1.25	1.64 6.28 2.97 1.12

Table A.3: Results from three different configurations of the Difference of Gaussian method for lung nodules detection. The total number of nodules in the test set was 105.

False Positive Reduction

Tables A.4, A.5 and Figure A.1 present the results achieved by the 3D ResNet deep convolutional network used for the false positive reduction task.

	Predicted	
Real	False (0)	True (1)
Candidate (0)	75726	54
Nodule (1)	58	86

Table A.4: Confusion matrix results for the 3D ResNet network.

	Precision	Recall	F1-score	Support
Candidate (0)	1.00	1.00	1.00	75780
Nodule (1)	0.61	0.60	0.61	144

Table A.5: Classification results for the 3D ResNet network.

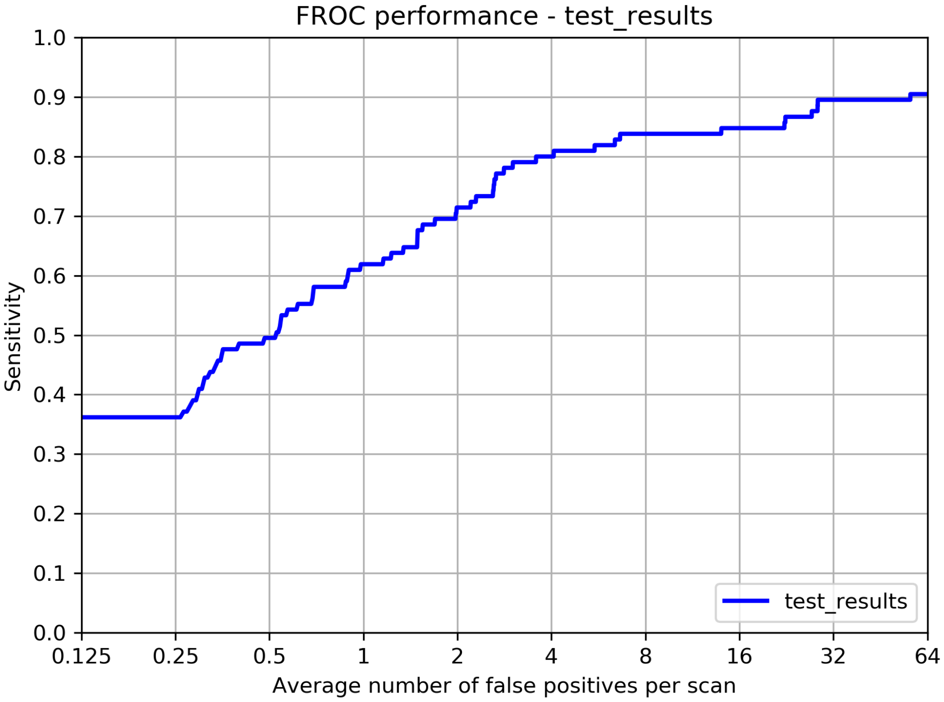


Figure A.1: FROC curve achieved in testing for the 3D ResNet network.

Cancer classification

In this section we describe the pipeline parameters (Table A.6) used for training the machine learning classifiers as well as the parameters used for the dense fully connected network (Table A.7) for lung cancer prediction.

Table A.6: Pipeline parameters tested using grid-search and 5-fold CV.

Algorithm	Options

k-NN	n_neighbors = [1,3,5,7,9,11]
	weights = ['uniform', 'distance']

LR	C = [0.001,0.01,0.1,0.5,1,3]
	class_weight = ['balanced']
	penalty = ['l1', 'l2']

RF	n_estimators = [100,150,200,250,500,750]
	criterion = ['entropy','gini']
	max_depth = ['None',2,4,6]
	class_weight = ['balanced']

SVM	C = [0.001,0.01,0.1,0.5,1,3]
	gamma = [0.005,0.01, 0.05,0.1,1,3]
	kernel = ['radial','poly']
	degree = [3,5,7,9]
	class_weight = ['balanced']

Table A.7: Parameters for training the dense network.

Method	Options
Hidden-Layers	(size/4),(size/3), (size/2),(size)
Alpha	1e-5,1e-3,1e-2,1,3,10
Activation	'relu', 'sigmoid'
Solver	'lbfgs'
Max_iter	200
Tol	1e-4

(*) The value of 'size' is the output of the N-1 layer of the nodule malignancy model together with the 3 features of the lung cancer baseline pipeline.

Appendix B

SUPPLEMENTARY MATERIAL II

B.1 Nodule classification

The model implemented for nodule classification is a 3D ResNet-34, borrowed from [110]. We used this architecture rather than the ResNet-50 described in our previous work [37] because we achieved better sensitivity and FROC scores. To train the network, we used the Adam optimization algorithm, a batch size of 128, a weighted binary cross entropy loss function (to attenuate the heavy data imbalance) and 3D data augmentation (flip, rotation, lighting, and zoom transforms).

Results

Tables (B.1, B.2) describe the evaluation results of the network in the test set of the LUNA-16. This partition represented 10% of the total amount of data, and it was composed of 75780 candidates (labeled as 0) and 144 nodules (labeled as 1).

	Predicted	
Real	0	1
Candidate (0)	75677	103
Nodule (1)	15	129

Table B.1: Confusion matrix results for the 3D ResNet network.

	Precision	Recall	F1-score	Support
Candidate (0)	1.00	1.00	1.00	75780
Nodule (1)	0.56	0.90	0.69	144

Table B.2: Classification results for the 3D ResNet network.

B.2 Nodule detector architecture

The nodule detection network was implemented using the available code of two recent and successful works [178] and [334]. The network reports five regression values that correspond to the centre of the candidate (x, y, z), its diameter d and the probability of being a nodule. For this latter value, the network uses a sigmoid activation function and no activation function is used for the others.

The network was trained using a batch size of 8, Adam as optimization algorithm, and a learning rate of 0.1 with a decay of 0.001 every 100 epochs, with a total of 450 epochs. Moreover, we used hard negative mining [269] with a factor of 20 times the batch size, as well as random rotation, flip, and zoom for 3D data augmentation. The final network architecture is shown in Figure-B.1.

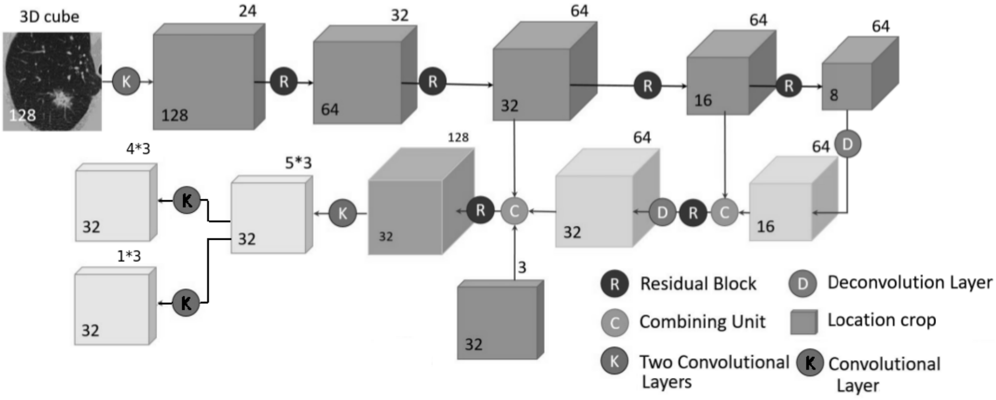


Figure B.1: Architecture for the nodule detector.

We followed [334] for designing the loss function of this network. Therefore, we set up 3 anchor boxes of 5,10 and 20 mm based on the nodules' distribution. For each anchor we defined 5 parts in the loss function, a classification loss L_{cls} for whether the current box is a nodule or not, regression loss L_{reg} for nodule coordinates x, y, z and nodule size d. Whether an anchor overlapped a ground truth bounding box with the intersection over union (IoU) higher than 0.5, we considered it as a positive anchor (p^* = 1). On the other hand, if an anchor has IoU with all ground truth boxes less than 0.02, we considered it as a negative anchor (p^* = 0).

The multi-task loss function for an anchor i was defined as:

$$L(p_i, t_i) = \alpha L_{cls}(p_i, p_i^*) + p_i^* L_{reg}(t_i, t_i^*)$$

The α is set as 0.5. For L_{cls} , we used a binary cross entropy loss function. For L_{reg} , we used smooth l_1 regression loss function [89].

Results

We evaluated the performance of this network with the test set partition of the LUNA-16 dataset. This partition was composed of 88 CT scans (out of 888 in total) with 105 annotated nodules. For each nodule annotation, we had its location (x, y, z) and its diameter. The resulting nodule detection performances are shown in Table-B.3. The first column of the table has the different 0 positive (FP) ratios (averaged per scan), and the rest of columns show the sensitivity obtained (mean, upper, and lower bounds). Upper and lower bounds were obtained after a 1000 bootstrapping. Figure-B.2 shows the FROC curve reported by this method.

FPRate	Mean	Lower	Upper
0.125	0.5799	0.45217	0.7176
0.25	0.6926	0.56976	0.7978
0.50	0.7961	0.71544	0.8666
1.0	0.8421	0.76800	0.9036
2.0	0.8755	0.81132	0.9306
4.0	0.9290	0.87962	0.9743
8.0	0.9419	0.89423	0.9809

Table B.3: Performances of the lung nodules detector at different FP in average per scan.

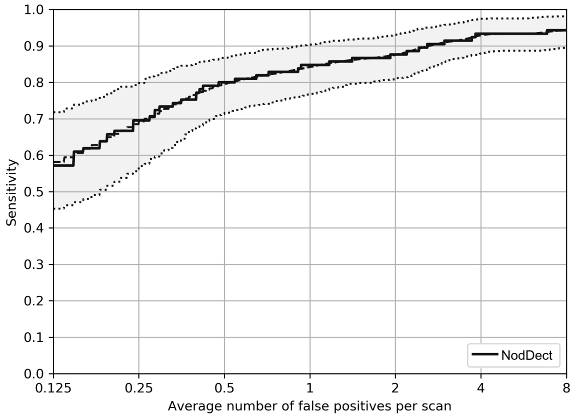


Figure B.2: FROC curve of the lung nodule detector computed for the test set.

B.3 Nodule re-identification

To gain further intuitions about the performances obtained between different SNNs, we show the results of additional experiments using different configurations (described in Section 3.1 of the manuscript):

Config	Layer	tr_acc	val_acc	test_acc	test_prec	test_rec
FIBC	['layer1']	0.817 ± 0.010	0.771 ± 0.005	0.792 ± 0.001	0.803 ± 0.004	0.781 ± 0.003
FIBC	['layer3']	0.723 ± 0.008	0.695 ± 0.002	0.710 ± 0.001	0.860 ± 0.006	0.513 ± 0.006
FIBC	['avgPool']	0.601 ± 0.009	0.588 ± 0.050	0.619 ± 0.021	0.735 ± 0.075	0.434 ± 0.196
UIBC	['layer2']	0.861 ± 0.015	0.853 ± 0.052	0.792 ± 0.016	0.754 ± 0.020	0.868 ± 0.044
UIBC	['avgPool']	0.845 ± 0.039	0.822 ± 0.039	0.767 ± 0.045	0.770 ± 0.034	0.763 ± 0.112
UIBC	['layer1']	0.841 ± 0.024	0.784 ± 0.065	0.777 ± 0.021	0.760 ± 0.037	0.818 ± 0.062
FIFB	['layer2']	0.880 ± 0.042	0.856 ± 0.041	0.864 ± 0.024	0.877 ± 0.040	0.853 ± 0.067
FIFB	['layer3']	0.850 ± 0.027	0.829 ± 0.047	0.830 ± 0.034	0.854 ± 0.030	0.800 ± 0.083
FIFB	['avgpool']	0.573 ± 0.023	0.637 ± 0.054	0.664 ± 0.033	0.641 ± 0.029	0.774 ± 0.174
UIFB	['layer3']	0.894 ± 0.027	0.883 ± 0.049	0.891 ± 0.024	0.866 ± 0.046	0.929 ± 0.039
UIFB	['layer1']	0.937 ± 0.038	0.879 ± 0.061	0.914 ± 0.029	0.883 ± 0.034	0.958 ± 0.054
UIFB	['avgpool']	0.761 ± 0.064	0.811 ± 0.088	0.789 ± 0.062	0.729 ± 0.059	0.929 ± 0.050
FICB	['layer2']	0.828 ± 0.067	0.830 ± 0.060	0.828 ± 0.034	0.871 ± 0.060	0.779 ± 0.083
FICB	['layer3']	0.799 ± 0.024	0.819 ± 0.071	0.780 ± 0.037	0.765 ± 0.053	0.818 ± 0.058
FICB	['avgpool']	0.519 ± 0.035	0.548 ± 0.033	0.580 ± 0.077	0.465 ± 0.328	0.295 ± 0.289
UICB	['layer3']	0.844 ± 0.054	0.888 ± 0.047	0.845 ± 0.038	0.797 ± 0.043	0.929 ± 0.026
UICB	['layer2']	0.888 ± 0.031	0.880 ± 0.050	0.862 ± 0.043	0.854 ± 0.084	0.892 ± 0.061
UICB	['avgpool']	0.620 ± 0.066	0.639 ± 0.076	0.642 ± 0.059	0.639 ± 0.221	0.492 ± 0.188
FCMB	['layer1', 'layer2', 'layer3', 'avgpool']	0.930 ± 0.035	0.881 ± 0.060	0.905 ± 0.027	0.900 ± 0.035	0.913 ± 0.041
FCMB	['layer1', 'layer2', 'layer3']	0.935 ± 0.025	0.870 ± 0.044	0.900 ± 0.024	0.878 ± 0.052	0.934 ± 0.029
FCMB	['layer1', 'layer2', 'avgpool']	0.932 ± 0.040	0.865 ± 0.045	0.912 ± 0.029	0.898 ± 0.039	0.932 ± 0.038
FCMB	['layer2', 'layer3']	0.882 ± 0.037	0.845 ± 0.041	0.850 ± 0.031	0.887 ± 0.032	0.805 ± 0.080
UCMB	['layer1', 'layer2', 'layer3']	0.946 ± 0.047	0.893 ± 0.046	0.904 ± 0.031	0.877 ± 0.048	0.945 ± 0.027
UCMB	['layer2', 'layer3']	0.920 ± 0.031	0.889 ± 0.046	0.908 ± 0.020	0.904 ± 0.036	0.916 ± 0.051
UCMB	['layer1', 'layer2']	0.952 ± 0.023	0.882 ± 0.039	0.942 ± 0.020	0.930 ± 0.037	0.958 ± 0.027
UCMB	['layer1', 'layer2', 'layer3', 'avgpool']	0.944 ± 0.024	0.875 ± 0.057	0.917 ± 0.019	0.896 ± 0.041	0.947 ± 0.037

Table B.4: Extended experiment results of different SNN configurations. The meaning of the acronyms are detailed in the Table 1 of the manuscript.

B.4 Lung nodule re-identification performance

	Train				**Test**			
	Small	**Medium**	**Large**	**Total**	**Small**	**Medium**	**Large**	**Total**
Matching	0.972 (37)	1.0 (66)	1.0 (10)	0.991 (113)	1.0 (13)	0.952 (21)	1.0 (4)	0.973 (38)
Non matching	0.933 (15)	0.962 (53)	0.977 (45)	0.964 (113)	1.0 (5)	0.9 (20)	0.769 (13)	0.868 (38)
Total	0.952 (52)	0.981 (119)	0.985 (55)	0.977 (226)	1.0 (18)	0.926 (41)	0.884 (17)	0.921 (76)

Table B.5: Performances for the whole training and test datasets stratified by lung nodule change size and matching/non-matching nodule pairs. Each cell contains accuracy and total number of nodule pairs.

B.5 Nodule growth pipeline performance

	Train				**Test**			
	Small	**Medium**	**Large**	**Total**	**Small**	**Medium**	**Large**	**Total**
ND	0.945 (37)	0.954 (66)	0.7 (10)	**0.929** **(113)**	1.0 (13)	0.90 (21)	1.0 (4)	**0.947** **(38)**
RI	0.914 (35)	0.841 (63)	0.71 (7)	**0.857** **(105)**	1.0 (13)	0.84 (19)	0.75 (4)	**0.888** **(36)**
GD	0.47 (32)	0.837 (53)	0.857 (5)	**0.74** **(90)**	1.0 (12)	0.929 (16)	0.842 (4)	**0.90** **(32)**

Table B.6: Performances for the whole training and test datasets stratified by lung nodule change size. In each cell we provide the accuracy for nodule detection (ND), nodule re-identification (RI) and F1-score for nodule growth detection (GD). In parenthesis we show the total number of nodule pairs involved per pipeline process and change size.

Bibliography

[An2] 3D CNN for lung nodule detection and false positive reduction. `https://grand-challenge-public.s3.amazonaws.com/f/challenge/71/8ac994bc-9951-420d-a7e5-21050c5b4132/20180102_081812_PAtech_NDET.pdf`. Accessed: 2021-06-03.

[2] Cifar-100. `https://www.cs.toronto.edu/~kriz/cifar.html`. Accessed: 2018-06-14.

[3] Kaggle, data science bowl 2017. `https://www.kaggle.com/c/data-science-bowl-2017/`. Accessed: 2018-06-13.

[4] Lung nodule analysis 2016. `https://luna16.grand-challenge.org/`. Accessed: 2018-06-13.

[5] Abe, H., Ishida, T., Shiraishi, J., Li, F., Katsuragawa, S., Sone, S., MacMahon, H., and Doi, K. (2004). Effect of temporal subtraction images on radiologists' detection of lung cancer on CT: Results of the observer performance study with use of film computed tomography images. *Academic Radiology*, 11(12):1337–1343.

[6] Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S., Karthikesalingam, A., King, D., Ashrafian, H., and Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digital Medicine*, 4(1):1–23.

[7] Al-Dhamari, I., Bauer, S., Paulus, D., Lissek, F., and Jacob, R. (2017). Acir: automatic cochlea image registration. In *Medical Imaging 2017: Image Processing*, volume 10133, page 1013310. International Society for Optics and Photonics.

[8] Alakwaa, W., Nassef, M., and Badr, A. (2017). Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). *Lung Cancer*, 8(8):409.

[9] Alberg, A., Brock, M., and Samet, J. (2016). Chapter 52: Epidemiology of lung cancer. *Murray & Nadel's Textbook of Respiratory Medicine, 6th edn. Saunders Elsevier*.

[10] Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. (2018). Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*.

[11] Altman, D. G. and Bland, J. M. (1983). Measurement in medicine: the analysis of method comparison studies. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(3):307–317.

[12] American College of Radiology (2014). Lung CT screening reporting and data system (lung-RADS). *Reston, VA: American College of Radiology*.

[13] Aoki, T., Murakami, S., Kim, H., Fujii, M., Takahashi, H., Oki, H., Hayashida, Y., Katsuragawa, S., Shiraishi, J., and Korogi, Y. (2014). Temporal subtraction method for lung nodule detection on successive thoracic CT soft-copy images. *Radiology*, 271(1):255–261.

[14] Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6):954.

[15] Armato, S. G., Giger, M. L., Moran, C. J., Blackburn, J. T., Doi, K., and MacMahon, H. (1999). Computerized detection of pulmonary nodules on CT scans. *Radiographics*, 19(5):1303–1311.

[16] Armato, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., et al. (2011). 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.

[17] Armato III, S. and MacMahon, H. (2003). Automated lung segmentation and computer-aided diagnosis for thoracic CT scans. In *International Congress Series*, volume 1256, pages 977–982. Elsevier.

[18] Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., and Reeves, Anthony P, . C. L. P. (2015). Data from LIDC-IDRI. the Cancer Imaging Archive `http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX`.

[19] Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., et al. (2011). 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.

[20] Armato III, S. G., McNitt-Gray, M. F., Reeves, A. P., Meyer, C. R., McLennan, G., Aberle, D. R., Kazerooni, E. A., MacMahon, H., van Beek, E. J., Yankelevitz, D., et al. (2007). The lung image database consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans. *Academic Radiology*, 14(11):1409–1421.

[21] Ather, S., Kadir, T., and Gleeson, F. (2020). Artificial intelligence and radiomics in pulmonary nodule management: current status and future applications. *Clinical Radiology*, 75(1):13–19.

[22] Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41.

[23] Ayhan, M. S. and Berens, P. (2018). Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*.

[24] Aziz, Z., Padley, S., and Hansell, D. (2004). CT techniques for imaging the lung: recommendations for multislice and single slice computed tomography. *European Journal of Radiology*, 52(2):119–136.

[25] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495.

[26] Bankier, A. A., MacMahon, H., Goo, J. M., Rubin, G. D., Schaefer-Prokop, C. M., and Naidich, D. P. (2017). Recommendations for measuring pulmonary nodules at CT: a statement from the Fleischner society. *Radiology*, 285(2):584–600.

[27] Bashir, U., Siddique, M. M., Mclean, E., Goh, V., and Cook, G. J. (2016). Imaging heterogeneity in lung cancer: techniques, applications, and challenges. *American Journal of Roentgenology*, 207(3):534–543.

[28] Basu, S., Wagstyl, K., Zandifar, A., Collins, L., Romero, A., and Precup, D. (2019). Early prediction of alzheimer's disease progression using variational autoencoders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 205–213. Springer.

[29] Baumgartner, C. F., Tezcan, K. C., Chaitanya, K., Hötker, A. M., Muehlematter, U. J., Schawkat, K., Becker, A. S., Donati, O., and Konukoglu, E. (2019). Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer.

[30] Beigelman-Aubry, C., Raffy, P., Yang, W., Castellino, R. A., and Grenier, P. A. (2007). Computer-aided detection of solid lung nodules on follow-up MDCT screening: evaluation of detection, tracking, and reading time. *American Journal of Roentgenology*, 189(4):948–955.

[31] Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210.

[32] Bello, I., Zoph, B., Vaswani, A., Shlens, J., and Le, Q. V. (2019). Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295.

[33] Bellotti, R., De Carlo, F., Gargano, G., Tangaro, S., Cascio, D., Catanzariti, E., Cerello, P., Cheran, S. C., Delogu, P., De Mitri, I., et al. (2007). A cad system for nodule detection in low-dose lung CTs based on region growing and a new active contour model. *Medical Physics*, 34(12):4901–4910.

[34] Bengio, Y., Goodfellow, I., and Courville, A. (2017). *Deep learning*, volume 1. MIT press Massachusetts, USA:.

[35] Blandin Knight, S., Crosbie, P. A., Balata, H., Chudziak, J., Hussell, T., and Dive, C. (2017). Progress and prospects of early detection in lung cancer. *Open biology*, 7(9):170070.

[36] Bonavita, I., Rafael-Palou, X., Ceresa, M., Piella, G., Ribas, V., and Ballester, M. A. G. (2020). Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline. *Computer Methods and Programs in Biomedicine*, 185:105172.

[37] Bonavita, I., Rafael-Palou, X., Ceresa, M., Piella, G., Ribas, V., and González Ballester, M. A. (2019). Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline. *Computer Methods and Programs in Biomedicine*, 185(105172):1–9.

[38] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a ”siamese” time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744.

[39] Brown, L. G. (1992). A survey of image registration techniques. *ACM Computing Surveys (CSUR)*, 24(4):325–376.

[40] Brown, M. S., Goldin, J. G., Suh, R. D., McNitt-Gray, M. F., Sayre, J. W., and Aberle, D. R. (2003). Lung micronodules: automated method for detection at thin-section CT—initial experience. *Radiology*, 226(1):256–262.

[41] Cai, Y., Li, Y., Qiu, C., Ma, J., and Gao, X. (2019). Medical image retrieval based on convolutional neural network and supervised hashing. *IEEE Access*, 7:51877–51885.

[42] Callister, M., Baldwin, D., Akram, A., Barnard, S., Cane, P., Draffan, J., Franks, K., Gleeson, F., Graham, R., Malhotra, P., et al. (2015). British thoracic society guidelines for the investigation and management of pulmonary nodules: accredited by nice. *Thorax*, 70(Suppl 2):ii1–ii54.

[43] Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

[44] Causey, J. L., Zhang, J., Ma, S., Jiang, B., Qualls, J. A., Politte, D. G., Prior, F., Zhang, S., and Huang, X. (2018). Highly accurate model for prediction of lung nodule malignancy with CT scans. *Scientific Reports*, 8(1):9286.

[45] Chalana, V., Sannella, M., and Haynor, D. R. (2000). General-purpose software tool for serial segmentation of stacked images. In *Medical Imaging 2000: Image Processing*, volume 3979, pages 192–204. International Society for Optics and Photonics.

[46] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

[47] Cheng, X., Zhang, L., and Zheng, Y. (2018). Deep similarity learning for multimodal medical images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):248–252.

[48] Choi, W.-J. and Choi, T.-S. (2013). Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach. *Entropy*, 15(2):507–523.

[49] Chung, Y.-A. and Weng, W.-H. (2017). Learning deep representations of medical images using siamese CNNs with application to content-based image retrieval. *Advances in Neural Information Processing Systems. Workshop on Machine Learning for Health (ML4H)*.

[50] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer.

[51] Ciompi, F., Chung, K., Van Riel, S. J., Setio, A. A. A., Gerke, P. K., Jacobs, C., Scholten, E. T., Schaefer-Prokop, C., Wille, M. M., Marchiano, A., et al. (2017). Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Scientific Reports*, 7:46479.

[52] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al. (2013). The cancer imaging archive (TCIA): maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057.

[53] Collins, V. P. (1956). Observation on growth rates of human tumors. *Am. J. Roentgenol*, 76:988–1000.

[54] Coroller, T. P., Grossmann, P., Hou, Y., Velazquez, E. R., Leijenaar, R. T., Hermann, G., Lambin, P., Haibe-Kains, B., Mak, R. H., and Aerts, H. J. (2015). CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology*, 114(3):345–350.

[55] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[56] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242.

[57] Cummings, S. R., Lillington, G. A., and Richard, R. J. (1986). Estimating the probability of malignancy in solitary pulmonary nodules: a bayesian approach. *American Review of Respiratory Disease*, 134(3):449–452.

[58] Cunningham, F., Fiebelkorn, S., Johnson, M., and Meredith, C. (2011). A novel application of the margin of exposure approach: segregation of tobacco smoke toxicants. *Food and Chemical Toxicology*, 49(11):2921–2933.

[59] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

[60] De Boor, C. and De Boor, C. (1978). *A practical guide to splines*, volume 27. springer-verlag New York.

[61] de Hoop, B., Gietema, H., van Ginneken, B., Zanen, P., Groenewegen, G., and Prokop, M. (2009). A comparison of six software packages for evaluation of solid lung nodules using semi-automated volumetry: what is the minimum increase in size to detect growth in repeated CT examinations. *European Radiology*, 19(4):800–808.

[62] De Nunzio, G., Tommasi, E., Agrusti, A., Cataldo, R., De Mitri, I., Favetta, M., Maglio, S., Massafra, A., Quarta, M., Torsello, M., et al. (2011). Automatic lung segmentation in CT images with accurate handling of the hilar region. *Journal of Digital Imaging*, 24(1):11–27.

[63] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

[64] Denisenko, T. V., Budkevich, I. N., and Zhivotovsky, B. (2018). Cell death-based treatment of lung adenocarcinoma. *Cell Death & Disease*, 9(2):1–14.

[65] Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

[66] Dey, R., Lu, Z., and Hong, Y. (2018). Diagnostic classification of lung nodules using 3D neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 774–778. IEEE.

[67] Dhara, A. K., Mukhopadhyay, S., Dutta, A., Garg, M., and Khandelwal, N. (2016). A combination of shape and texture features for classification of pulmonary nodules in lung CT images. *Journal of Digital Imaging*, 29(4):466–475.

[68] Ding, J., Li, A., Hu, Z., and Wang, L. (2017). Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 559–567. Springer.

[69] Doersch, C., Gupta, A., and Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430.

[70] Doll, R. and Hill, A. B. (1950). Smoking and carcinoma of the lung. *British Medical Journal*, 2(4682):739.

[71] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634.

[72] Dou, Q., Chen, H., Jin, Y., Lin, H., Qin, J., and Heng, P.-A. (2017). Automated pulmonary nodule detection via 3D convnets with online sample filtering and hybrid-loss residual learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 630–638. Springer.

[73] Dou, Q., Chen, H., Yu, L., Qin, J., and Heng, P.-A. (2016). Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection. *IEEE Transactions on Biomedical Engineering*, 64(7):1558–1567.

[74] Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., and Pal, C. (2016). The importance of skip connections in biomedical image segmentation. In *Deep learning and data labeling for medical applications*, pages 179–187. Springer.

[75] Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., and Cardoso, M. J. (2018). Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. *arXiv preprint arXiv:1806.08640*.

[76] Elazab, A., Wang, C., Gardezi, S. J. S., Bai, H., Hu, Q., Wang, T., Chang, C., and Lei, B. (2020). GP-GAN: Brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal MR Images. *Neural Networks*.

[77] Eppenhof, K. A. and Pluim, J. P. (2018). Pulmonary CT registration through supervised learning with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 38(5):1097–1105.

[78] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

[79] Farazi, H. and Behnke, S. (2019). Frequency domain transformer networks for video prediction. *arXiv preprint arXiv:1903.00271*.

[80] Firmino, M., Morais, A. H., Mendoça, R. M., Dantas, M. R., Hekis, H. R., and Valentim, R. (2014). Computer-aided detection system for lung cancer in computed tomography scans: review and future prospects. *Biomedical Engineering Online*, 13(1):41.

[81] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer Series in Statistics New York, NY, USA.

[82] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[83] Gal, Y. and Ghahramani, Z. (2016a). Bayesian convolutional neural networks with Bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*.

[84] Gal, Y. and Ghahramani, Z. (2016b). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

[85] Gandini, S., Botteri, E., Iodice, S., Boniol, M., Lowenfels, A. B., Maisonneuve, P., and Boyle, P. (2008). Tobacco smoking and cancer: a meta-analysis. *International Journal of Cancer*, 122(1):155–164.

[86] Ganeshan, B., Miles, K. A., Young, R. C., and Chatwin, C. R. (2009). Texture analysis in non-contrast enhanced CT: impact of malignancy on texture in apparently disease-free areas of the liver. *European Journal of Radiology*, 70(1):101–110.

[87] Gao, R., Tang, Y., Xu, K., Huo, Y., Bao, S., Antic, S. L., Epstein, E. S., Deppen, S., Paulson, A. B., Sandler, K. L., et al. (2020). Time-distanced gates in long short-term memory networks. *Medical Image Analysis*, 65:101785.

[88] Ghesu, F. C., Georgescu, B., Mansoor, A., Yoo, Y., Gibson, E., Vishwanath, R., Balachandran, A., Balter, J. M., Cao, Y., Singh, R., et al. (2021). Quantifying and leveraging predictive uncertainty for medical image assessment. *Medical Image Analysis*, 68:101855.

[89] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[90] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587.

[91] Godoy, M. C. and Naidich, D. P. (2009). Subsolid pulmonary nodules and the spectrum of peripheral adenocarcinomas of the lung: recommended interim guidelines for assessment and management. *Radiology*, 253(3):606–622.

[92] Golan, R., Jacob, C., and Denzinger, J. (2016). Lung nodule detection in CT images using deep convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 243–250. IEEE.

[93] Gonçalves, L., Novo, J., Cunha, A., and Campilho, A. (2018). Learning lung nodule malignancy likelihood from radiologist annotations or diagnosis data. *Journal of Medical and Biological Engineering*, 38:1–19.

[94] Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

[95] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*, volume 1. Cambridge: MIT press.

[96] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

[97] Gould, M. K., Donington, J., Lynch, W. R., Mazzone, P. J., Midthun, D. E., Naidich, D. P., and Wiener, R. S. (2013). Evaluation of individuals with pulmonary nodules: When is it lung cancer?: Diagnosis and management of lung cancer: American college of chest physicians evidence-based clinical practice guidelines. *Chest*, 143(5):e93S–e120S.

[98] Gu, S., Wilson, D., Tan, J., and Pu, J. (2011). Pulmonary nodule registration: Rigid or nonrigid? *Medical Physics*, 38(7):4406–4414.

[99] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42.

[100] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410.

[101] Gurcan, M. N., Sahiner, B., Petrick, N., Chan, H.-P., Kazerooni, E. A., Cascade, P. N., and Hadjiiski, L. (2002). Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system. *Medical Physics*, 29(11):2552–2558.

[102] Gurney, J. (1993). Determining the likelihood of malignancy in solitary pulmonary nodules with bayesian analysis. part i. theory. *Radiology*, 186(2):405–413.

[103] Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

[104] Halder, A., Dey, D., and Sadhu, A. K. (2020). Lung nodule detection from feature engineering to deep learning in thoracic CT images: a comprehensive review. *Journal of Digital Imaging*, 33(3):655–677.

[105] Hammack, D. (2017). Forecasting lung cancer diagnoses with deep learning. Technical report.

[106] Han, D., Heuvelmans, M. A., Vliegenthart, R., Rook, M., Dorrius, M. D., De Jonge, G. J., Walter, J. E., van Ooijen, P. M., De Koning, H. J., and Oudkerk, M. (2018). Influence of lung nodule margin on volume-and diameter-based reader variability in CT lung cancer screening. *The British Journal of Radiology*, 91(1090):20170405.

[107] Han, P. K., Klein, W. M., and Arora, N. K. (2011). Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making*, 31(6):828–838.

[108] Hancock, M. C. and Magnan, J. F. (2016). 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 4:044504.

[109] Hansell, D. M., Bankier, A. A., MacMahon, H., McLoud, T. C., Muller, N. L., and Remy, J. (2008). Fleischner society: glossary of terms for thoracic imaging. *Radiology*, 246(3):697–722.

[110] Hara, K., Kataoka, H., and Satoh, Y. (2018). Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555.

[111] Harris, M., Clark, J., Coote, N., Fletcher, P., Harnden, A., McKean, M., and Thomson, A. (2011). British thoracic society guidelines for the management of community acquired pneumonia in children: update 2011. *Thorax*, 66(Suppl 2):ii1–ii23.

[112] Haskins, G., Kruger, U., and Yan, P. (2020). Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1):1–18.

[113] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE.

[114] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034.

[115] He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

[116] He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

[117] He, K., Zhang, X., Ren, S., and Sun, J. (2016c). Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer.

[118] He, X., Zhao, K., and Chu, X. (2021). Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622.

[119] Helm, E. J., Silva, C. T., Roberts, H. C., Manson, D., Seed, M. T., Amaral, J. G., and Babyn, P. S. (2009). Computer-aided detection for the identification of pulmonary nodules in pediatric oncology patients: initial experience. *Pediatric Radiology*, 39(7):685–693.

[120] Ho, T. K. (1995). Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE.

[121] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[122] Hong, H., Lee, J., and Yim, Y. (2008). Automatic lung nodule matching on sequential CT images. *Computers in Biology and Medicine*, 38(5):623 – 634.

[123] Horeweg, N., van Rosmalen, J., Heuvelmans, M. A., van der Aalst, C. M., Vliegenthart, R., Scholten, E. T., ten Haaf, K., Nackaerts, K., Lammers, J.-W. J., Weenink, C., et al. (2014). Lung cancer probability in patients with CT-detected pulmonary nodules: a prespecified analysis of data from the nelson trial of low-dose CT screening. *The Lancet Oncology*, 15(12):1332–1341.

[124] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510.

[125] Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141.

[126] Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., and Welling, M. (2019). Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–145. Springer.

[127] Hua, K.-L., Hsu, C.-H., Hidayati, S. C., Cheng, W.-H., and Chen, Y.-J. (2015). Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and therapy*, 8.

[128] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708.

[129] Huang, P., Lin, C. T., Li, Y., Tammemagi, M. C., Brock, M. V., Atkar-Khattra, S., Xu, Y., Hu, P., Mayo, J. R., Schmidt, H., et al. (2019a). Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. *The Lancet Digital Health*, 1(7):e353–e362.

[130] Huang, W., Xue, Y., and Wu, Y. (2019b). A cad system for pulmonary nodule prediction based on deep three-dimensional convolutional neural networks and ensemble learning. *PLOS ONE*, 14(7):1–17.

[131] Hughes, L. H., Schmitt, M., Mou, L., Wang, Y., and Zhu, X. X. (2018). Identifying corresponding patches in sar and optical images with a pseudo-siamese CNN. *IEEE Geoscience and Remote Sensing Letters*, 15(5):784–788.

[132] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

[133] Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al. (2018). nnU-Net: Self-adapting framework for U-Net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*.

[134] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134.

[135] Jacobs, C., van Rikxoort, E. M., Twellmann, T., Scholten, E. T., de Jong, P. A., Kuhnigk, J.-M., Oudkerk, M., de Koning, H. J., Prokop, M., Schaefer-Prokop, C., et al. (2014). Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. *Medical Image Analysis*, 18(2):374–384.

[136] jaketmp (2018). jaketmp/pycompare: Looks both ways.

[137] Jang, J., Jung, S. E., Jeong, W. K., Lim, Y. S., Choi, J.-I., Park, M. Y., Kim, Y., Lee, S.-K., Chung, J.-J., Eo, H., et al. (2016). Radiation doses of various CT protocols: a multicenter longitudinal observation study. *Journal of Korean medical science*, 31(Suppl 1):S24.

[138] Jo, H. H., Hong, H., and Goo], J. M. (2014). Pulmonary nodule registration in serial CT scans using global rib matching and nodule template matching. *Computers in Biology and Medicine*, 45:87 – 97.

[139] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

[140] Katzmann, A., Muehlberg, A., Sühling, M., Noerenberg, D., Holch, J. W., Heinemann, V., and Groß, H.-M. (2018). Predicting lesion growth and patient survival in colorectal cancer patients using deep neural networks. In *Medical Imaging with Deep Learning*.

[141] Kaya, A. and Can, A. B. (2015). A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics. *Journal of Biomedical Informatics*, 56:69–79.

[142] Kendall, A., Badrinarayanan, V., and Cipolla, R. (2015). Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*.

[143] Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*.

[144] Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516.

[145] Khosravan, N. and Bagci, U. (2018). S4nd: Single-shot single-scale lung nodule detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 794–802. Springer.

[146] Kim, H., Park, C. M., Song, Y. S., Sunwoo, L., Choi, Y. R., Im Kim, J., Kim, J. H., Bae, J. S., Lee, J. H., and Goo, J. M. (2016). Measurement variability of persistent pulmonary subsolid nodules on same-day repeat CT: what is the threshold to determine true nodule growth during follow-up? *PLoS One*, 11(2):e0148853.

[147] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

[148] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

[149] Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluim, J. P. (2009). Elastix: a toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205.

[150] Ko, J. and Betke, M. (2001). Chest CT: automated nodule detection and assessment of change over time–preliminary experience. *Radiology*, 218(1):267—273.

[151] Kobayashi, Y., Sakao, Y., Deshpande, G. A., Fukui, T., Mizuno, T., Kuroda, H., Sakakura, N., Usami, N., Yatabe, Y., and Mitsudomi, T. (2014). The association between baseline clinical–radiological characteristics and growth of pulmonary nodules with ground-glass opacity. *Lung Cancer*, 83(1):61–66.

[152] Koch, G. (2015). Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning. Workshop on Deep Learning, vol. 2.*, volume 2.

[153] Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., Eslami, S. A., Rezende, D. J., and Ronneberger, O. (2018). A probabilistic U-Net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975.

[154] Kohl, S. A., Romera-Paredes, B., Maier-Hein, K. H., Rezende, D. J., Eslami, S., Kohli, P., Zisserman, A., and Ronneberger, O. (2019). A hierarchical probabilistic U-Net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*.

[155] Koo, C. W., Anand, V., Girvin, F., Wickstrom, M. L., Fantauzzi, J. P., Bogoni, L., Babb, J. S., and Ko, J. P. (2012). Improved efficiency of CT interpretation using an automated lung nodule matching program. *American Journal of Roentgenology*, 199(1):91–95.

[156] Kopelowitz, E. and Engelhard, G. (2019). Lung nodules detection and segmentation using 3D mask-rcnn. *arXiv preprint arXiv:1907.07676*.

[157] Kothari, S., Phan, J. H., Stokes, T. H., and Wang, M. D. (2013). Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association*, 20(6):1099–1108.

[158] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

[159] Kuan, K., Ravaut, M., Manek, G., Chen, H., Lin, J., Nazir, B., Chen, C., Howe, T. C., Zeng, Z., and Chandrasekhar, V. (2017). Deep learning for lung cancer detection: tackling the kaggle data science bowl 2017 challenge. *arXiv preprint arXiv:1705.09435*.

[160] Kuhnigk, J.-M., Dicken, V., Bornemann, L., Bakai, A., Wormanns, D., Krass, S., and Peitgen, H.-O. (2006). Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans. *IEEE Transactions on Medical Imaging*, 25(4):417–434.

[161] Kumar, A., Fulham, M., Feng, D., and Kim, J. (2019). Co-learning feature fusion maps from pet-ct images of lung cancer. *IEEE Transactions on Medical Imaging*, 39(1):204–217.

[162] Kumar, D., Wong, A., and Clausi, D. A. (2015). Lung nodule classification using deep features in CT images. In *2015 12th Conference on Computer and Robot Vision*, pages 133–138. IEEE.

[163] Kumar, V., Abbas, A. K., and Aster, J. C. (2017). *Robbins basic pathology e-book*. Elsevier Health Sciences.

[164] Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2016). Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*.

[165] Larici, A. R., Farchione, A., Franchi, P., Ciliberto, M., Cicchetti, G., Calandriello, L., del Ciello, A., and Bonomo, L. (2017). Lung nodules: size still matters. *European Respiratory Review*, 26(146):170025.

[166] Lea, C., Flynn, M. D., Vidal, R., Reiter, A., and Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165.

[167] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

[168] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

[LeCun and Cortes] LeCun, Y. and Cortes, C.

[170] Lee, G., Lee, H. Y., Park, H., Schiebler, M. L., van Beek, E. J., Ohno, Y., Seo, J. B., and Leung, A. (2017). Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: state of the art. *European Journal of Radiology*, 86:297–307.

[171] Lee, K. W., Kim, M., Gierada, D. S., and Bae, K. T. (2007). Performance of a computer-aided program for automated matching of metastatic pulmonary nodules detected on follow-up chest CT. *American Journal of Roentgenology*, 189(5):1077–1081.

[172] Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1):1–14.

[173] Li, M., Tang, H., Chan, M. D., Zhou, X., and Qian, X. (2020a). DC-AL-GAN: pseudoprogression and true tumor progression of glioblastoma multiform image classification based on DCGAN and AlexNet. *Medical Physics*, 47(3):1139–1150.

[174] Li, Q., Gavrielides, M. A., Sahiner, B., Myers, K. J., Zeng, R., and Petrick, N. (2015). Statistical analysis of lung nodule volume measurements with CT in a large-scale phantom study. *Medical Physics*, 42(7):3932–3947.

[175] Li, Q., Sone, S., et al. (2003). Selective enhancement filters for nodules, vessels, and airway walls in two-and three-dimensional CT scans. *Medical Physics*, 30(8):2040–2051.

[176] Li, Y., Lu, H., Yang, S., Serikawa, S., and Kitazono, Y. (2013). A review of image segmentation methods. In *The 1st IEEE/IIAE International Conference on Intelligent Systems and Image Processing 2013 (ICISIP2013)*.

[177] Li, Y., Yang, J., Xu, Y., Xu, J., Ye, X., Tao, G., Xie, X., and Liu, G. (2020b). Learning tumor growth via follow-up volume prediction for lung nodules. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 508–517. Springer.

[178] Liao, F., Liang, M., Li, Z., Hu, X., and Song, S. (2019). Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-or network. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3484–3495.

[179] Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

[180] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125.

[181] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

[182] Lipková, J., Angelikopoulos, P., Wu, S., Alberts, E., Wiestler, B., Diehl, C., Preibisch, C., Pyka, T., Combs, S. E., Hadjidoukas, P., et al. (2019). Personalized radiotherapy design for glioblastoma: integrating mathematical tumor models, multimodal scans, and Bayesian inference. *IEEE Transactions on Medical Imaging*, 38(8):1875–1884.

[183] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

[184] Liu, K. and Kang, G. (2017). Multiview convolutional neural networks for lung nodule classification. *International Journal of Imaging Systems and Technology*, 27(1):12–22.

[185] Liu, M., Dong, J., Dong, X., Yu, H., and Qi, L. (2018). Segmentation of lung nodule in CT images based on mask r-cnn. In *2018 9th International Conference on Awareness Science and Technology (iCAST)*, pages 1–6. IEEE.

[186] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer.

[187] Liu, X., Deng, Z., and Yang, Y. (2019). Recent progress in semantic image segmentation. *Artificial Intelligence Review*, 52(2):1089–1106.

[188] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.

[189] Lorenzi, M., Ayache, N., Frisoni, G. B., Pennec, X., (ADNI, A. D. N. I., et al. (2013). Lcc-demons: a robust and accurate symmetric diffeomorphic registration algorithm. *NeuroImage*, 81:470–483.

[190] Loverdos, K., Fotiadis, A., Kontogianni, C., Iliopoulou, M., and Gaga, M. (2019). Lung nodules: A comprehensive review on current approach and management. *Annals of thoracic medicine*, 14(4):226.

[191] Loyman, M. and Greenspan, H. (2019). Lung nodule retrieval using semantic similarity estimates. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, page 109503P. International Society for Optics and Photonics.

[192] Lung Screening Trial Research Team, N. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409.

[193] Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

[194] MacMahon, H., Austin, J. H., Gamsu, G., Herold, C. J., Jett, J. R., Naidich, D. P., Patz Jr, E. F., and Swensen, S. J. (2005). Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner society. *Radiology*, 237(2):395–400.

[195] MacMahon, H., Naidich, D. P., Goo, J. M., Lee, K. S., Leung, A. N., Mayo, J. R., Mehta, A. C., Ohno, Y., Powell, C. A., Prokop, M., et al. (2017). Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner society 2017. *Radiology*, 284(1):228–243.

[196] Maintz, J. A. and Viergever, M. A. (1998). A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36.

[197] Matsuguma, H., Mori, K., Nakahara, R., Suzuki, H., Kasai, T., Kamiyama, Y., Igarashi, S., Kodama, T., and Yokoi, K. (2013). Characteristics of subsolid pulmonary nodules showing growth during follow-up with CT scanning. *Chest*, 143(2):436–443.

[198] McNitt-Gray, M. F., Hart, E. M., Wyckoff, N., Sayre, J. W., Goldin, J. G., and Aberle, D. R. (1999). A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results. *Medical Physics*, 26(6):880–888.

[199] McWilliams, A., Tammemagi, M. C., Mayo, J. R., Roberts, H., Liu, G., Soghrati, K., Yasufuku, K., Martel, S., Laberge, F., Gingras, M., et al. (2013). Probability of cancer in pulmonary nodules detected on first screening CT. *New England Journal of Medicine*, 369(10):910–919.

[200] Messay, T., Hardie, R. C., and Tuinstra, T. R. (2015). Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the lung image database consortium and image database resource initiative dataset. *Medical Image Analysis*, 22(1):48–62.

[201] Miller, K. D., Nogueira, L., Mariotto, A. B., Rowland, J. H., Yabroff, K. R., Alfano, C. M., Jemal, A., Kramer, J. L., and Siegel, R. L. (2019). Cancer treatment and survivorship statistics, 2019. *CA: a Cancer Journal for Clinicians*, 69(5):363–385.

[202] Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE.

[203] Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., and Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[204] Modersitzki, J. (2004). *Numerical methods for image registration*. Oxford University Press on Demand.

[205] Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

[206] Murphy, K., Van Ginneken, B., Reinhardt, J., Kabus, S., Ding, K., Deng, X., and Pluim, J. (2010). Evaluation of methods for pulmonary image registration: The empire10 study. *Grand Challenges in Medical Image Analysis*, 2010:11–22.

[207] Murphy, K., van Ginneken, B., Reinhardt, J. M., Kabus, S., Ding, K., Deng, X., Cao, K., Du, K., Christensen, G. E., Garcia, V., Vercauteren, T., Ayache, N., Commowick, O., Malandain, G., Glocker, B., Paragios, N., Navab, N., Gorbunova, V., Sporring, J., de Bruijne, M., Han, X., Heinrich, M. P., Schnabel, J. A., Jenkinson, M., Lorenz, C., Modat, M., McClelland, J. R., Ourselin, S., Muenzing, S. E. A., Viergever, M. A., De Nigris, D., Collins, D. L., Arbel, T., Peroni, M., Li, R., Sharp, G. C., Schmidt-Richberg, A., Ehrhardt, J., Werner, R., Smeets, D., Loeckx, D., Song, G., Tustison, N., Avants, B., Gee, J. C., Staring, M., Klein, S., Stoel, B. C., Urschler, M., Werlberger, M., Vandemeulebroucke, J., Rit, S., Sarrut, D., and Pluim, J. P. W. (2011). Evaluation of registration methods on thoracic CT: The empire10 challenge. *IEEE Transactions on Medical Imaging*, 30(11):1901–1920.

[208] Murphy, K., van Ginneken, B., Schilham, A. M., De Hoop, B., Gietema, H., and Prokop, M. (2009). A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. *Medical Image Analysis*, 13(5):757–770.

[209] Naidich, D. P., Marshall, C. H., Gribbin, C., Arams, R. S., and McCauley, D. I. (1990). Low-dose CT of the lungs: preliminary observations. *Radiology*, 175(3):729–731.

[210] Nair, T., Precup, D., Arnold, D. L., and Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59.

[211] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1.

[212] Nasrullah, N., Sang, J., Alam, M. S., Mateen, M., Cai, B., and Hu, H. (2019). Automated lung nodule detection and classification using deep learning combined with multiple strategies. *Sensors*, 19(17):3722.

[213] Nibali, A., He, Z., and Wollersheim, D. (2017). Pulmonary nodule classification with deep residual networks. *International Journal of Computer Assisted Radiology and Surgery*, 12(10):1799–1808.

[214] Nishio, M., Sugiyama, O., Yakami, M., Ueno, S., Kubo, T., Kuroda, T., and Togashi, K. (2018). Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning. *PloS one*, 13(7):e0200721.

[215] Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528.

[216] Oksuz, K., Cam, B. C., Kalkan, S., and Akbas, E. (2020). Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[217] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

[218] Olah, C. (2015). Understanding LSTM networks.

[219] on the Evaluation of Carcinogenic Risks to Humans, I. W. G., Organization, W. H., and for Research on Cancer, I. A. (2004). *Tobacco smoke and involuntary smoking*. Number 83. Iarc.

[220] O'Reilly, K. M., Mclaughlin, A. M., Beckett, W. S., and Sime, P. J. (2007). Asbestos-related lung disease. *American Family Physician*, 75(5):683–688.

[221] Orozco, H. M., Villegas, O. O. V., Sánchez, V. G. C., Domínguez, H. d. J. O., and Alfaro, M. d. J. N. (2015). Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. *Biomedical Engineering Online*, 14(1):9.

[222] Osband, I. (2016). Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. In *NIPS Workshop on Bayesian Deep Learning*, volume 192.

[223] O'Connor, C. M., Adams, J. U., and Fairman, J. (2010). Essentials of cell biology. *Cambridge, MA: NPG Education*, 1:54.

[224] O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D., and Walsh, J. (2019). Deep learning vs. traditional computer vision. In *Science and Information Conference*, pages 128–144. Springer.

[225] Park, M., Wilson, L. S., and Jin, J. S. (2000). Automatic extraction of lung boundaries by a knowledge-based method. In *Selected papers from the Pan-Sydney workshop on Visualisation-Volume 2*, pages 11–16. Australian Computer Society, Inc.

[226] Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. (2018). Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR.

[227] Patz, E. F., Pinsky, P., Gatsonis, C., Sicks, J. D., Kramer, B. S., Tammemägi, M. C., Chiles, C., Black, W. C., and Aberle, D. R. (2014). Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Internal Medicine*, 174(2):269–274.

[228] Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

[229] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[230] Pelc, N. J. (2014). Recent and future directions in CT imaging. *Annals of Biomedical Engineering*, 42(2):260–268.

[231] Petersen, J., Jäger, P. F., Isensee, F., Kohl, S. A., Neuberger, U., Wick, W., Debus, J., Heiland, S., Bendszus, M., Kickingereder, P., et al. (2019). Deep probabilistic modeling of glioma growth. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 806–814. Springer.

[232] Popovic, Z. B. and Thomas, J. D. (2017). Assessing observer variability: a user's guide. *Cardiovascular Diagnosis and Therapy*, 7(3):317.

[233] Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., and Nielsen, M. (2013). Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 246–253. Springer.

[234] Qiao, Y., van Lew, B., Lelieveldt, B. P., and Staring, M. (2015). Fast automatic step size estimation for gradient descent optimization of image registration. *IEEE Transactions on Medical Imaging*, 35(2):391–403.

[235] Rachmadi, M. F., Valdés-Hernández, M. d. C., Makin, S., Wardlaw, J., and Komura, T. (2020). Automatic spatial estimation of white matter hyperintensities evolution in brain MRI using disease evolution predictor deep neural networks. *Medical Image Analysis*.

[236] Rafael-Palou, X., Aubanell, A., Bonavita, I., Ceresa, M., Piella, G., Ribas, V., and Ballester, M. Á. G. (2020a). Pulmonary nodule malignancy classification using its temporal evolution with two-stream 3D convolutional neural networks. In *Medical Imaging with Deep Learning*.

[237] Rafael-Palou, X., Aubanell, A., Bonavita, I., Ceresa, M., Piella, G., Ribas, V., and Ballester, M. A. G. (2020b). Re-identification and growth detection of pulmonary nodules without image registration using 3D siamese neural networks. *Medical Image Analysis*, 67.

[238] Rafael-Palou, X., Aubanell, A., Ceresa, M., Ribas, V., Piella, G., and Ballester, M. A. G. (2021). Detection, growth quantification and malignancy prediction of pulmonary nodules using deep convolutional networks in follow-up CT scans. *arXiv preprint arXiv:2103.14537*.

[239] Ravi, D., Blumberg, S. B., Mengoudi, K., Xu, M., Alexander, D. C., and Oxtoby, N. P. (2019). Degenerative adversarial neuroimage nets for 4D simulations: Application in longitudinal MRI. *arXiv preprint arXiv:1912.01526*.

[240] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788.

[241] Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

[242] Reeves, A. P., Xie, Y., and Jirapatnakul, A. (2016). Automated pulmonary nodule CT image characterization in lung cancer screening. *International Journal of Computer Assisted Radiology and Surgery*, 11(1):73–88.

[243] Ren, S., He, K., Girshick, R., and Sun, J. (2015a). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

[244] Ren, S., He, K., Girshick, R., and Sun, J. (2015b). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

[245] Revel, M.-P., Bissery, A., Bienvenu, M., Aycard, L., Lefort, C., and Frija, G. (2004). Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? *Radiology*, 231(2):453–458.

[246] Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

[247] Riquelme, D. and Akhloufi, M. A. (2020). Deep learning for lung cancer nodules detection and classification in CT scans. *AI*, 1(1):28–67.

[248] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer.

[249] Roy, A. G., Conjeti, S., Navab, N., and Wachinger, C. (2018). Inherent brain segmentation quality control from fully convnet Monte Carlo sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 664–672. Springer.

[250] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

[251] Rühaak, J., Polzin, T., Heldmann, S., Simpson, I. J., Handels, H., Modersitzki, J., and Heinrich, M. P. (2017). Estimation of large motion in lung CT by integrating regularized keypoint correspondences into dense deformable registration. *IEEE Transactions on Medical Imaging*, 36(8):1746–1757.

[252] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

[253] Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10(3):e0118432.

[254] Sarapata, E. A. and de Pillis, L. (2014). A comparison and catalog of intrinsic tumor growth models. *Bulletin of Mathematical Biology*, 76(8):2010–2024.

[255] Schiff, G. D., Hasan, O., Kim, S., Abrams, R., Cosby, K., Lambert, B. L., Elstein, A. S., Hasler, S., Kabongo, M. L., Krosnjar, N., et al. (2009). Diagnostic error in medicine: analysis of 583 physician-reported errors. *Archives of Internal Medicine*, 169(20):1881–1887.

[256] Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., and Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53:197 – 207.

[257] Schwartz, M. (1961). A biomathematical approach to clinical tumor growth. *Cancer*, 14(6):1272–1294.

[258] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

[259] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.

[260] Setio, A. A., Jacobs, C., Gelderblom, J., and Ginneken, B. (2015). Automatic detection of large pulmonary solid nodules in thoracic CT images. *Medical Physics*, 42(10):5642–5653.

[261] Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., Wille, M. M. W., Naqibullah, M., Sánchez, C. I., and Van Ginneken, B. (2016). Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169.

[262] Setio, A. A. A., Traverso, A., De Bel, T., Berens, M. S., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M. E., Geurts, B., et al. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Medical Image Analysis*, 42:1–13.

[263] Shan, C. (2012). Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters*, 33(4):431–437.

[264] Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248.

[265] Shen, W., Zhou, M., Yang, F., Dong, D., Yang, C., Zang, Y., and Tian, J. (2016). Learning from experts: developing transferable deep features for patient-level lung cancer prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 124–131. Springer.

[266] Shen, W., Zhou, M., Yang, F., Yang, C., and Tian, J. (2015). Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging*, pages 588–599. Springer.

[267] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *arXiv preprint arXiv:1506.04214*.

[268] Shridhar, K., Laumann, F., and Liwicki, M. (2019). Uncertainty estimations by softplus normalization in Bayesian convolutional neural networks with variational inference. *arXiv preprint arXiv:1806.05978*.

[269] Shrivastava, A., Gupta, A., and Girshick, R. (2016). Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769.

[270] Siegelman, S. S., Khouri, N. F., Leo, F., Fishman, E. K., Braverman, R., and Zerhouni, E. (1986). Solitary pulmonary nodules: CT assessment. *Radiology*, 160(2):307–312.

[271] Simonyan, K. and Zisserman, A. (2014a). Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*.

[272] Simonyan, K. and Zisserman, A. (2014b). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[273] Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., et al. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*.

[274] Snoeckx, A., Reyntiens, P., Desbuquoit, D., Spinhoven, M. J., Van Schil, P. E., van Meerbeeck, J. P., and Parizel, P. M. (2018). Evaluation of the solitary pulmonary nodule: size matters, but do not ignore the power of morphology. *Insights into Imaging*, 9(1):73–86.

[275] Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491.

[276] Song, G., Han, J., Zhao, Y., Wang, Z., and Du, H. (2017a). A review on medical image registration as an optimization problem. *Current Medical Imaging Reviews*, 13(3):274–283.

[277] Song, Q., Zhao, L., Luo, X., and Dou, X. (2017b). Using deep learning for classification of lung nodules on computed tomography images. *Journal of Healthcare Engineering*, 2017:8314740.

[278] Soomro, K., Zamir, A. R., and Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

[279] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of Machine Learning Research*, 15(1):1929–1958.

[280] Stergios, C., Mihir, S., Maria, V., Guillaume, C., Marie-Pierre, R., Stavroula, M., and Nikos, P. (2018). Linear and deformable image registration with 3D convolutional neural networks. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 13–22. Springer.

[281] Sun, S., Rubin, G. D., Paik, D., Steiner, R. M., Zhuge, F., and Napel, S. (2007). Registration of lung nodules using a semi-rigid model: Method and preliminary results. *Medical Physics*, 34(2):613–626.

[282] Sun, W., Zheng, B., and Qian, W. (2016). Computer aided lung cancer diagnosis with deep learning algorithms. In *Medical imaging 2016: computer-aided diagnosis*, volume 9785, page 97850Z. International Society for Optics and Photonics.

[283] Sun, X. and Wang, X. (2011). Study of edge detection algorithms for lung CT image on the basis of matlab. In *2011 Chinese Control and Decision Conference (CCDC)*, pages 810–813. IEEE.

[284] Suzuki, K. (2013a). Machine learning in computer-aided diagnosis of the thorax and colon in CT: a survey. *IEICE Transactions on Information and Systems*, 96(4):772–783.

[285] Suzuki, K. (2013b). Machine learning in computer-aided diagnosis of the thorax and colon in CT: a survey. *IEICE Transactions on Information and Systems*, 96(4):772–783.

[286] Suzuki, K., Koike, T., Asakawa, T., Kusumoto, M., Asamura, H., Nagai, K., Tada, H., Mitsudomi, T., Tsuboi, M., Shibata, T., et al. (2011). A prospective radiological study of thin-section computed tomography to predict pathological noninvasiveness in peripheral clinical ia lung cancer (japan clinical oncology group 0201). *Journal of Thoracic Oncology*, 6(4):751–756.

[287] Swanson, K. R., Alvord, E. C., and Murray, J. (2002). Quantifying efficacy of chemotherapy of brain tumors with homogeneous and heterogeneous drug delivery. *Acta Biotheoretica*, 50(4):223–237.

[288] Swensen, S. J., Silverstein, M. D., Ilstrup, D. M., Schleck, C. D., and Edell, E. S. (1997). The probability of malignancy in solitary pulmonary nodules: application to small radiologically indeterminate nodules. *Archives of internal medicine*, 157(8):849–855.

[289] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

[290] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

[291] Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.

[292] Talkington, A. and Durrett, R. (2015). Estimating tumor growth rates in vivo. *Bulletin of Mathematical Biology*, 77(10):1934–1954.

[293] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

[294] Tanner, N. T., Porter, A., Gould, M. K., Li, X.-J., Vachani, A., and Silvestri, G. A. (2017). Physician assessment of pretest probability of malignancy and adherence with guidelines for pulmonary nodule evaluation. *Chest*, 152(2):263–270.

[295] Tao, C., Gierada, D. S., Zhu, F., Pilgram, T. K., Wang, J. H., and Bae, K. T. (2009). Automated matching of pulmonary nodules: evaluation in serial screening chest CT. *American Journal of Roentgenology*, 192(3):624–628.

[296] Tao, R., Gavves, E., and Smeulders, A. W. (2016). Siamese instance search for tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1420–1429.

[297] Team, N. L. S. T. R. (2011a). The national lung screening trial: overview and study design. *Radiology*, 258(1):243–253.

[298] Team, N. L. S. T. R. (2011b). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409.

[299] Thakur, S. K., Singh, D. P., and Choudhary, J. (2020). Lung cancer identification: a review on detection and classification. *Cancer and Metastasis Reviews*, pages 1–10.

[300] Thawani, R., McLane, M., Beig, N., Ghose, S., Prasanna, P., Velcheti, V., and Madabhushi, A. (2018). Radiomics and radiogenomics in lung cancer: a review for the clinician. *Lung cancer*, 115:34–41.

[301] Thun, M. J., Hannan, L. M., Adams-Campbell, L. L., Boffetta, P., Buring, J. E., Feskanich, D., Flanders, W. D., Jee, S. H., Katanoda, K., Kolonel, L. N., et al. (2008). Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. *PLoS Med*, 5(9):e185.

[302] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497.

[303] Valente, I. R. S., Cortez, P. C., Neto, E. C., Soares, J. M., de Albuquerque, V. H. C., and Tavares, J. M. R. (2016). Automatic 3D pulmonary nodule detection in CT images: a survey. *Computer Methods and Programs in Biomedicine*, 124:91–107.

[304] van Ginneken, B. (2017). Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. *Radiological Physics and Technology*, 10(1):23–32.

[305] Van Ginneken, B., Setio, A. A., Jacobs, C., and Ciompi, F. (2015). Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International symposium on biomedical imaging (ISBI)*, pages 286–289. IEEE.

[306] van Klaveren, R. J., Oudkerk, M., Prokop, M., Scholten, E. T., Nackaerts, K., Vernhout, R., van Iersel, C. A., van den Bergh, K. A., van't Westeinde, S., van der Aalst, C., et al. (2009). Management of lung nodules detected by volume CT scanning. *New England Journal of Medicine*, 361(23):2221–2229.

[307] Varior, R. R., Haloi, M., and Wang, G. (2016). Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer.

[308] Veasey, B. P., Broadhead, J., Dahle, M., Seow, A., and Amini, A. A. (2020). Lung nodule malignancy prediction from longitudinal CT scans with siamese convolutional attention networks. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:257–264.

[309] Viergever, M. A., Maintz, J. A., Klein, S., Murphy, K., Staring, M., and Pluim, J. P. (2016). A survey of medical image registration – under review. *Medical Image Analysis*, 33:140 – 144. 20th anniversary of the Medical Image Analysis journal (MedIA).

[310] Wahidi, M. M., Govert, J. A., Goudar, R. K., Gould, M. K., and McCrory, D. C. (2007). Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer?: Accp evidence-based clinical practice guidelines. *Chest*, 132(3):94S–107S.

[311] Wang, C., Rimner, A., Hu, Y.-C., Tyagi, N., Jiang, J., Yorke, E., Riyahi, S., Mageras, G., Deasy, J. O., and Zhang, P. (2019a). Toward predicting the evolution of lung tumors during radiotherapy observed on a longitudinal mr imaging study via a deep learning algorithm. *Medical Physics*, 46(10):4699–4707.

[312] Wang, S., Zhou, M., Liu, Z., Liu, Z., Gu, D., Zang, Y., Dong, D., Gevaert, O., and Tian, J. (2017). Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical Image Analysis*, 40:172–183.

[313] Wang, X., Mao, K., Wang, L., Yang, P., Lu, D., and He, P. (2019b). An appraisal of lung nodules automatic classification algorithms for CT images. *Sensors*, 19(1):194.

[314] Wang, Z. and Zhang, D. (1999). Progressive switching median filter for the removal of impulse noise from highly corrupted images. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 46(1):78–80.

[315] Wen, T.-H., Gasic, M., Mrksic, N., Su, P.-H., Vandyke, D., and Young, S. (2015). Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

[316] Wild, C. P., Stewart, B. W., and Wild, C. (2014). *World cancer report 2014*. World Health Organization Geneva, Switzerland.

[317] Williams, C. J. H. (1992). *Lung cancer: the facts*. Oxford University Press, USA.

[318] Wong, K. C., Summers, R. M., Kebebew, E., and Yao, J. (2015). Tumor growth prediction with reaction-diffusion and hyperelastic biomechanical model by physiological data fusion. *Medical Image Analysis*, 25(1):72–85.

[319] Wong, K. C., Summers, R. M., Kebebew, E., and Yao, J. (2016). Pancreatic tumor growth prediction with elastic-growth decomposition, image-derived motion, and FDM-FEM coupling. *IEEE Transactions on Medical Imaging*, 36(1):111–123.

[320] Xie, H., Fang, S., Zha, Z.-J., Yang, Y., Li, Y., and Zhang, Y. (2019a). Convolutional attention networks for scene text recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1s):3.

[321] Xie, H., Yang, D., Sun, N., Chen, Z., and Zhang, Y. (2019b). Automated pulmonary nodule detection in CT images using deep convolutional neural networks. *Pattern Recognition*, 85:109–119.

[322] Xie, Y., Zhang, J., Xia, Y., Fulham, M., and Zhang, Y. (2018). Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. *Information Fusion*, 42:102–110.

[323] Xu, Y.-M., Zhang, T., Xu, H., Qi, L., Zhang, W., Zhang, Y.-D., Gao, D.-S., Yuan, M., and Yu, T.-F. (2020). Deep learning in CT images: automated pulmonary nodule detection for subsequent management using convolutional neural network. *Cancer Management and Research*, 12:2979.

[324] Zeb, I., Li, D., Nasir, K., Katz, R., Larijani, V. N., and Budoff, M. J. (2012). Computed tomography scans in the evaluation of fatty liver disease in a population based study: the multi-ethnic study of atherosclerosis. *Academic Radiology*, 19(7):811–818.

[325] Zeng, M., Li, J., and Peng, Z. (2006). The design of top-hat morphological filter and application to infrared target detection. *Infrared Physics & Technology*, 48(1):67–76.

[326] Zhang, C., Sun, X., Dang, K., Li, K., Guo, X.-w., Chang, J., Yu, Z.-q., Huang, F.-y., Wu, Y.-s., Liang, Z., et al. (2019a). Toward an expert level of lung cancer detection and classification using a deep convolutional neural network. *The Oncologist*, 24(9):1159.

[327] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017a). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

[328] Zhang, L., Lu, L., Summers, R. M., Kebebew, E., and Yao, J. (2017b). Convolutional invasion and expansion networks for tumor growth prediction. *IEEE Transactions on Medical Imaging*, 37(2):638–648.

[329] Zhang, L., Lu, L., Wang, X., Zhu, R. M., Bagheri, M., Summers, R. M., and Yao, J. (2019b). Spatio-temporal convolutional LSTMs for tumor growth prediction by learning 4D longitudinal patient data. *IEEE Transactions on Medical Imaging*, 39(4):1114–1126.

[330] Zhang, Z., Liu, Q., and Wang, Y. (2018). Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753.

[331] Zhao, C., Han, J., Jia, Y., and Gou, F. (2018). Lung nodule detection via 3D U-Net and contextual convolutional neural network. In *2018 International Conference on Networking and Network Applications (NaNA)*, pages 356–361. IEEE.

[332] Zhou, M., Leung, A., Echegaray, S., Gentles, A., Shrager, J. B., Jensen, K. C., Berry, G. J., Plevritis, S. K., Rubin, D. L., Napel, S., et al. (2018a). Non–small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. *Radiology*, 286(1):307–315.

[333] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018b). Unet++: A nested U-Net architecture for medical image segmentation. In *Deep learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer.

[334] Zhu, W., Liu, C., Fan, W., and Xie, X. (2018). Deeplung: Deep 3D dual path nets for automated pulmonary nodule detection and classification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 673–681. IEEE.

[335] Zikri, Y. K. B., Helguera, M., Cahill, N. D., Shrier, D., and Linte, C. A. (2019). Toward an affine feature-based registration method for ground glass

lung nodule tracking. In *ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing*, pages 247–256. Springer.

[336] Zinovev, D., Furst, J. D., and Raicu, D. S. (2011). Building an ensemble of probabilistic classifiers for lung nodule interpretation. *2011 10th International Conference on Machine Learning and Applications and Workshops*, 2:155–161.

[337] Zinovev, D., Raicu, D. S., Furst, J. D., and Armato, S. G. (2009). Predicting radiological panel opinions using a panel of machine learning classifiers. *Algorithms*, 2:1473–1502.

[338] Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2017). Learning transferable architectures for scalable image recognition. *arXiv preprint arXiv:1707.07012*.

[339] Zuidhof, G. (2017). Full preprocessing tutorial. `https://www.kaggle.com/gzuidhof/full-preprocessing-tutorial`. Accessed: 2018-06-11.