FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Decision Support Methods for Lung Cancer Follow-Up in Computed Tomography

Carlos Alexandre Nunes Ferreira



Doctoral Programme in Electrical and Computer Engineering

Supervisor: Aurélio Campilho, Emeritus Professor, FEUP

Co-supervisor: Miguel Coimbra, Full Professor, FCUP

Co-supervisor: Isabel Ramos, Full Professor, FMUP

Co-supervisor: Pedro Sousa, Radiologist, ULSGE

March 14, 2025

© Carlos Alexandre Nunes Ferreira, 2024

Decision Support Methods for Lung Cancer Follow-Up in Computed Tomography

Carlos Alexandre Nunes Ferreira

Doctoral Programme in Electrical and Computer Engineering

March 14, 2025

Resumo

O cancro do pulmão é a principal causa de mortalidade oncológica a nível mundial, sendo frequentemente detetado em estádios avançados, onde o tratamento é menos eficaz. Devido à forte associação entre o cancro do pulmão e o tabagismo, têm-se adotado estratégias preventivas centradas em políticas de controlo do tabaco para reduzir as taxas de incidência. Assim, quando um novo caso é diagnosticado em contexto hospitalar, o seguimento do cancro pulmonar necessita de ser altamente preciso para garantir que a progressão não seja ignorada ou permaneça sem tratamento adequado e atempado. Isto deve-se ao facto de a remoção cirúrgica de tecido, a quimioterapia e a radioterapia nem sempre serem as opções imediatas de tratamento. Devido a isto, um seguimento adequado dos pacientes permite aos clínicos monitorizar de perto a evolução dos nódulos e tomar decisões informadas sobre quando são necessárias intervenções mais invasivas.

Apesar da importância fundamental do seguimento na gestão do cancro do pulmão, existem poucos métodos automáticos concebidos especificamente para apoiar esta tarefa. As estratégias atuais de seguimento são baseadas em tarefas automáticas, como a deteção, segmentação e caracterização de nódulos pulmonares, para facilitar a tomada de decisão clínica. As propostas existentes para sistemas de apoio à decisão no seguimento do cancro pulmonar focam-se, assim, na otimização destas tarefas isoladas, ao invés de abordar o seguimento de uma perspetiva holística. Esta abordagem fragmentada pode resultar em decisões menos precisas, uma vez que os erros acumulados das tarefas automáticas podem ter implicações clínicas significativas. Nesta tese, propõe-se o desenvolvimento de sistemas de apoio à decisão para o seguimento do cancro do pulmão. Concretamente, a investigação está estruturada em torno de três objetivos principais: *Avaliar* o estado atual da doença, *Aantecipar* a sua progressão futura e *Reavaliar* a condição do paciente através da identificação eficaz de nódulos pulmonares previamente relatados em relatórios médicos textuais.

Na fase de *Avaliação*, foram desenvolvidas duas abordagens distintas para recomendações de seguimento de acordo com as diretrizes Lung-RADS: um método de ponta-a-ponta, que integra a incerteza de anotação durante o treino para sugerir diretamente o intervalo de seguimento, e um método modular baseado nas características dos nódulos. Constatou-se que estas abordagens podem ser usadas para responder a diferentes necessidades clínicas: o método de ponta-a-ponta é eficaz para fins de estratificação, apresentando uma melhor capacidade de generalização, enquanto o método baseado em características permite um controlo mais detalhado sobre as decisões de cada módulo, especialmente relevante em casos complexos onde a interpretabilidade é essencial. A modelação da incerteza no método de ponta-a-ponta apresentou um desempenho superior aos restantes modelos, especialmente em cenários de grande discordância entre as opiniões dos radiologistas, revelando-se portanto importante considerar a variabilidade das anotações durante o treino de modelos.

Na fase de *Antecipação*, foi desenvolvido um modelo para prever o crescimento dos nódulos pulmonares após um e dois anos, usando tanto dados de imagem como o historial clínico do paciente. Os resultados de previsão foram particularmente relevantes na estimativa da progressão de nódulos malignos, podendo servir como um sistema útil de alerta para os clínicos, em substituição

das estimativas atuais de 348 dias para que um nódulo maligno duplique o seu volume. O modelo demonstrou uma melhor capacidade para prever o tamanho dos nódulos de forma simultânea para vários anos, em vez de usar modelos independentes para cada ano. Relativamente aos dados de entrada, o historial clínico do paciente mostrou um superior desempenho em comparação com os dados de imagem isolados, uma vez que fornece uma visão mais abrangente da condição do paciente e da sua influência no crescimento dos nódulos.

Na fase de *Reavaliação*, introduz-se um novo método para correlacionar relatórios médicos em texto com anotações de nódulos em dados de imagem. Isto é conseguido através do desenvolvimento de uma base de dados anotada e validada com pares de referência texto-imagem, e de um sistema automático para extrair e associar estes pares. Demonstrou-se que este método pode ser útil para melhorar a avaliação longitudinal dos nódulos pulmonares, reduzindo a variabilidade nas avaliações clínicas. Além disso, mostrou-se particularmente valioso quando os clínicos não guardam as anotações dos nódulos nas imagens nos softwares ou quando estas estão ausentes; ao ter dados de texto e imagem emparelhados, o sistema pode restaurar a posição espacial dos nódulos a partir de relatórios médicos textuais passados, garantindo a precisão do seguimento dos nódulos mencionados.

Globalmente, os métodos propostos nesta tese proporcionam avanços significativos no seguimento do cancro do pulmão ao integrar modelos automáticos de apoio à decisão. Estas abordagens facilitam a personalização dos cuidados dos pacientes, ajustando os intervalos de seguimento de forma flexível e melhorando a precisão das previsões de crescimento dos nódulos. Além disso, a capacidade de correlacionar eficazmente dados textuais e de imagem reduz a incerteza nas decisões clínicas e melhora a precisão das avaliações longitudinais. A incorporação destes desenvolvimentos em softwares de análise de imagem médica permite abordagens mais orientadas a dados ou centradas no paciente, conduzindo a melhores resultados clínicos e à otimização da alocação de recursos na gestão do cancro do pulmão.

Abstract

Lung cancer is the leading cause of cancer-related mortality worldwide, often detected at advanced stages where treatment is less effective. Due to the high association of lung cancer with smoking, preventive strategies have focused on tobacco control policies to reduce incidence rates. Nevertheless, once a case falls within the medical scope, a highly accurate follow-up is essential to ensure that no progression is overlooked or left untreated in a timely manner. Surgical resection, chemotherapy, and radiotherapy are not always the first-line treatment options, and a robust follow-up strategy allows clinicians to monitor nodule evolution closely and make informed decisions about when more invasive actions are warranted.

Despite the critical role of follow-up in managing lung cancer, there is a notable lack of automated methods specifically designed to address this need. Existing strategies for follow-up often build upon individual tasks such as nodule detection, segmentation, and characterization, which are essential components of clinical decision-making. Various proposals have been made for decision support systems in follow-up care, yet they generally rely on optimizing these isolated tasks rather than integrating them into a holistic follow-up system. This fragmented approach can lead to accumulated errors that, when combined, may have significant clinical implications. Therefore, this thesis proposes frameworks to enhance decision support for lung cancer follow-up. More specifically, this research is structured around three primary objectives: evaluating the current state of the disease, forecasting disease progression, and reassessing the patient's condition by easily identifying pulmonary nodules that were previously reported in radiology reports.

For the *Evaluate* phase, two distinct approaches were developed for follow-up recommendations according to the Lung-RADS guidelines: an end-to-end method, embedding annotation uncertainty during training, to directly suggest follow-up timing; and a feature-based method that predicts nodule characteristics to then determine the follow-up intervals. It was found that these approaches can be used to address different clinical needs. The end-to-end method is effective for stratification purposes with better generalization capabilities, while the feature-based method provides detailed insights into nodule characteristics, allowing for more granular control over patient management decisions, particularly in complex cases where interpretability is crucial. Modeling uncertainty in the end-to-end method exibited improved model performance, particularly in cases of annotation discordance or absence, highlighting the importance of accounting for variability in data labeling.

In the *Forecast* phase, a model is developed to predict lung nodule growth one and two years ahead, using both imaging data and patient history as inputs. The findings emphasize the model's robustness in accurately forecasting the progression of malignant nodules, highlighting its potential to serve as an early warning system for clinicians, as an alternative to the current estimates of 348 days for a malignant nodule to double in volume. The model demonstrated a better ability to predict nodule size over multiple years simultaneously, rather than predicting for each year individually. Moreover, using patient history as input revealed superior performance when compared to using imaging data alone, as it provided a more comprehensive view of the patient's condition

and its influence on nodule growth dynamics.

The *Reassess* phase introduces a novel framework for correlating medical text reports with nodule annotations in image data. This is achieved by developing a dataset annotated with validated text-image reference pairs and an automated system to extract and link these pairs. Such a framework is found useful in enhancing the longitudinal assessment of nodules, reducing variability in clinical evaluations. Additionally, it is demonstrated to be valuable when clinicians do not annotate nodules in the images or when annotations are missing entirely; by having paired text-image data, the system can restore the positional context of nodules from historical reports, ensuring continuity of follow-up accuracy.

Overall, the methods proposed in this thesis provide substantial advancements with the potential to incorporate and support clinical practice by integrating machine learning models into lung cancer follow-up protocols. These approaches facilitate the personalization of patient care by tailoring follow-up intervals and improving the accuracy of nodule growth predictions. Furthermore, the ability to effectively correlate textual and imaging data not only reduces uncertainty in clinical decisions but also enhances the precision of longitudinal assessments. By incorporating these innovations, healthcare providers can implement more data-driven, patient-specific interventions that are likely to improve outcomes and optimize resource allocation in lung cancer management.

Acknowledgments

This PhD journey has certainly been a rollercoaster ride, marked by its share of ups and downs. Pursuing this research during the pandemic was a challenging task, with Lamego becoming my safe haven. Balancing this research with the responsibilities of TEC4Health was equally challenging. Completing this work without succumbing to burnout, questioning my sanity, or sacrificing family and friendships has been a monumental task.

For all these reasons, I extend my profound professional gratitude to Professor Aurélio Campilho, for his rigorous oversight that have shaped me into a more meticulous researcher; to Professor Miguel Coimbra, for the many opportunities he has provided throughout my career; and to Dr. Isabel Ramos and Dr. Pedro Sousa, as well as the teams at ULS São João and ULS Gaia/Espinho, for their invaluable clinical support. No researcher stands alone, and I am deeply thankful to all the colleagues at INESC TEC, C-BER, and the Bioimaging Lab for their camaraderie during these long years of research. A special note of appreciation goes to João Pedrosa, Tânia Melo, Joana Rocha, and Sofia Cardoso Pereira for being by my side throughout this journey. I am also grateful to my co-authors for their contributions to the scientific publications, particularly to Radboudumc, which hosted me for two months during an Erasmus period.

To my family, thank you. To my parents, who, with the little they had, always tried to give me so much, and who have been my stepping stones, enabling me to reach new heights. To my brother, for always being there, making me laugh, and inspiring me to teach. To my grandparents, for instilling in me the values of kindness and respect. To Maria João, you have been my heart and soul — thank you for showing up every day to support and be supported. To Miki and Camila, my final days of this PhD would have been far more difficult without you.

To my friends, Carlos Guimarães, José Alves, Cristina Martins, and João Vale—those forged in the fires of martensite, whose bonds will never break. To those in Lamego, for always being my safe harbor. And to those at IEEE, for being not only my volunteer activity but also the catalyst for my professional growth.

"Põe quanto és no mínimo que fazes".
Carlos Alexandre Ferreira.

This research was funded by the Foundation for Science and Technology (FCT) grant contract 10.54499/SFRH/BD/146437/2019 (https://doi.org/10.54499/SFRH/BD/146437/2019), supported by Norte Portugal Regional Operational Program (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Social Fund (ESF).











Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Objectives and Contributions . . . 3
- 1.3 Thesis Outline . . . 7

2 Clinical Background in Medical Imaging for Lung Cancer **8**
- 2.1 Lung Cancer . . . 8
  - 2.1.1 Anatomical Context . . . 8
  - 2.1.2 Patient Pathway . . . 9
- 2.2 Computed Tomography Diagnosis . . . 10
  - 2.2.1 Computed Tomography . . . 10
  - 2.2.2 Lung Nodules . . . 12
  - 2.2.3 Annotation Process of CT Scans . . . 14
  - 2.2.4 Follow-up Guidelines . . . 15
- 2.3 Discussion . . . 17

3 Review on Automatic Methods for Lung Cancer Follow-up **18**
- 3.1 Introduction . . . 18
  - 3.1.1 Overview . . . 18
  - 3.1.2 Performance Evaluation . . . 21
- 3.2 Lung Nodule Detection . . . 23
- 3.3 Lung Nodule Segmentation . . . 28
- 3.4 Lung Nodule Characterization . . . 31
- 3.5 Discussion . . . 34

4 Automatic Follow-up Recommendation using Lung-RADS Guidelines **36**
- 4.1 Introduction . . . 37
- 4.2 Related Work . . . 39
- 4.3 Material . . . 40
- 4.4 Methods . . . 42
  - 4.4.1 End-to-End Classification . . . 42
  - 4.4.2 Feature-Based Classification . . . 45
  - 4.4.3 Model Evaluation . . . 46
- 4.5 Results . . . 48
  - 4.5.1 End-to-End Classification . . . 48
  - 4.5.2 Feature-Based Classification . . . 50
  - 4.5.3 Discussion . . . 51
- 4.6 Concluding Remarks . . . 57

5 Forecast of Lung Nodule Diameter **58**
5.1 Introduction . . . 58
5.2 Related Work . . . 61
5.3 Material and Methods . . . 63
5.3.1 Dataset . . . 63
5.3.2 Experiments . . . 66
5.3.3 Model Training . . . 68
5.3.4 Model Evaluation . . . 69
5.4 Results . . . 70
5.4.1 Growth Regression . . . 70
5.4.2 Analysis of the Classes with Malignant or Growing Nodules . . . 73
5.4.3 Bland-Altman Analysis . . . 75
5.4.4 Feature Importance Analysis . . . 77
5.5 Discussion . . . 78
5.5.1 Growth Regression . . . 78
5.5.2 Analysis of the Classes with Malignant or Growing Nodules . . . 80
5.5.3 Bland-Altman Analysis . . . 81
5.5.4 Feature Importance Analysis . . . 81
5.5.5 Comparability . . . 82
5.5.6 Limitations . . . 83
5.6 Concluding Remarks . . . 83

6 Lung Nodule Image Annotation from Medical Reports **85**
6.1 Introduction . . . 86
6.2 Related Work . . . 87
6.3 Materials . . . 90
6.3.1 Medical Reports Annotation . . . 91
6.3.2 Nodule Matching . . . 93
6.3.3 Clinical Assessment . . . 96
6.3.4 Data Insights . . . 96
6.4 Methods . . . 100
6.4.1 Named Entity Recognition . . . 101
6.4.2 Relation Extraction . . . 104
6.4.3 Text Simplification . . . 107
6.4.4 Anatomic Segmentation and Nodule Detection . . . 108
6.4.5 Matching Decision . . . 110
6.5 Results . . . 113
6.5.1 Natural Language Processing . . . 113
6.5.2 Computer Vision . . . 115
6.5.3 Discussion . . . 117
6.6 Concluding Remarks . . . 119

7 Conclusions **121**
7.1 Future Trends . . . 123

References **125**

A List of Publications **142**

List of Figures

1.1 Overview of the main contributions in this thesis. 6

2.1 Schematic representation of the lungs anatomy. 9
2.2 Different phases and data used in the lung cancer journey. 11
2.3 HU scale and intensity ranges of anatomic structures. 11
2.4 A representation of a transverse chest LDCT. 12
2.5 Nodule examples. 14

3.1 Number of scientific publications on lung nodule detection, segmentation, characterization, and follow-up since 2004 up to 2023. 20
3.2 Standard lung nodule detection pipeline. 23
3.3 Diagram of the Faster R-CNN and YOLO networks. 27
3.4 Manual and automatic nodule segmentation examples. 30

4.1 Schematic representation of the functionality of the different methods used in this study. 38
4.2 Examples of nodule middle axial slices of different Lung-RADS classes. 41
4.3 Diagram showing the relationship between the number of annotators and the level of agreement regarding Lung-RADS classes in LIDC-IDRI. 42
4.4 WRN Arquitecture. 43
4.5 DenseNet Architecture. 43
4.6 iW-Net. 46
4.7 Examples of correct results from the automatic methods in this chapter. 49
4.8 Cumulative accuracy in the different experiments as a function of the uncertainty weight. 49
4.9 Examples of incorrect results from the automatic methods in this chapter. 52
4.10 Confusion matrices between GT, end-to-end class prediction, and feature-based class prediction for Lung-RADS. 54
4.11 Lung nodule examples, showing LR class results and with Grad-CAM highlights. 56

5.1 A malignant nodule and its growth over two years. 59
5.2 General pipeline for lung nodule follow-up. 60
5.3 Overview of the three experimental forecast set-ups 61
5.4 Overview of the three workflows used in the forecast experiments. 67
5.5 ResNet50 Architecture. 67
5.6 Example of a nodule forecast over two years. 71
5.7 Examples of the ROI of the nodules corresponding to some of the best one year forecast results. 71
5.8 Bland-Altman plots of the deviation between the obtained results and GT. 76

5.9 Examples of the ROI of the nodules corresponding to some of the worst one year forecast results. . . . 79

6.1 Summary of the CLIP approach. . . . 89
6.2 Schematic overview of the assay phases in this study. . . . 91
6.3 Summarized flowchart encompassing the criteria used to match the nodules. . . . 95
6.4 Example of some matched nodules. . . . 98
6.5 Example of some mismatched nodules. . . . 99
6.6 Submodules of the automatic method for converting medical reports into nodule annotations in CT scans. . . . 102
6.7 NER Architecture 103
6.8 Schematic representation of some of the relations that can be established in an RE example sentence. . . . 105
6.9 RE Architecture. . . . 106
6.10 Schematic representation of the automatic lung nodule detection system's architecture, composed of an initial candidate detection followed by a false-positive reduction step. . . . 109
6.11 Examples of correctly matched nodules. . . . 116
6.12 Examples of nodule matches with inaccuracies. . . . 118

List of Tables

2.1 Nodule characteristics organized in 1-6 classes. . . . 13
2.2 Comparison of lung cancer follow-up guidelines . . . 16

3.1 Search terms used in each category. . . . 19
3.2 Performance of selected automatic lung nodule detection algorithms. . . . 25
3.3 Performance of selected automatic lung nodule segmentation algorithms. . . . 29
3.4 Automatic lung nodule characterization algorithms. . . . 32

4.1 Correspondence of Lung-RADS classes with nodule characteristics and estimated population prevalence. . . . 38
4.2 Relation between LUNG-RADS classes with the quantity of nodules in LIDC-IDRI. 41
4.3 Illustrative examples of the computation of the sample weights. . . . 44
4.4 Performance of end-to-end methods implemented to classify Lung-RADS Classes. 50
4.5 Performance of tasks and the feature-based method for classifying Lung-RADS classes. . . . 51
4.6 Comparison of end-to-end and feature-based approaches . . . 57

5.1 Criteria for nodule growth classification. . . . 59
5.2 Number of nodules in ranges of diameter and growth rate. . . . 65
5.3 Number of nodules used from the NLST database based on their characteristics. . 65
5.4 Tabular patient data. . . . 66
5.5 MAD (mm) and standard deviation (mm) for different experiments on the test set. 72
5.6 MAD (mm) and standard deviation (mm) for different experiments on malignant nodules. . . . 73
5.7 Experiments in which the predicted VDT of malignant nodules closely approximates the GT. . . . 74
5.8 MAD (mm) and standard deviation (mm) for different experiments on nodules classified as grown according to three guidelines. . . . 75
5.9 Mean and standard deviation over the total range of variation for all features. . . 77
5.10 Comparison of results with the state of the art. . . . 82

6.1 Number of nodules with matching and mismatching based on diameter. . . . 97
6.2 Number of nodules with matching and mismatching based on nodule characteristics and location. . . . 98
6.3 Fleischner classification table used in this study. . . . 100
6.4 Confusion matrix comparing the Fleischner guidelines' follow-up recommendation considering the pool of nodules from manual image annotations and those in the medical reports. . . . 100
6.5 Performance of Named Entity Recognition . . . 114

6.6 Performance in the identification of lung nodules. 115
6.7 Confusion matrices comparing the Fleischner guidelines' follow-up recommendations between the GT from medical reports and scenarios where nodules are identified, considering both the perfect and real scenarios within the NLP module. 117

Acronyms and Abbreviations

AI Artificial Intelligence
AL Agreement Level
AUC Area Under the Receiver Operating Characteristic Curve

BERT Bidirectional Encoder Representations from Transformers
BIO Beginning, Inside, and Outside
BMI Body Mass Index
BTS British Thoracic Society

CAD Computer-Aided Diagnosis
CLIP Contrastive Language-Image Pre-Training
CNN Convolutional Neural Network
CRF Conditional Random Fields
CT Computed Tomography
CV Computer Vision

DL Deep Learning

FFNN FeedForward Neural Network
FN False Negative
FP False Positive

GAN Generative Adversarial Network
GDPR General Data Protection Regulation
GGO Ground-Glass Opacity
GPT Generative Pre-trained Transformer
GT Ground Truth

HU Hounsfield Unit

ICVF IntraCellular Volume Fraction
ID Identification
IoU Intersection over Union

KNN K-Nearest Neighbors

LDCT Low Dose Computed Tomography
LIDC-IDRI Lung Image Database Consortium and Image Database Resource Initiative
LLL Left Lower Lobe
LLM Large Language Model
LNDb Lung Nodule Database
LSTM Long Short-Term Memory
LUL Left Upper Lobe
LUNA-16 LUng Nodule Analysis 2016

MAD Mean Absolute Difference
MIP Maximum Intensity Projection
ML Middle Lobe
MLP MultiLayer Perceptron
MSE Mean Squared Error

NER Named Entity Recognition
NLP Natural Language Processing
NLST National Lung Screening Trial
NN Neural Network

PET Positron Emission Tomography
PET/CT Positron Emission Tomography – Computed Tomography
POS Part-of-Speech
PY Pack-Year

R-CNN Region-based Convolutional Neural Network
RE Relation Extraction
ReLU Rectified Linear Unit
RepVL-PAN Vision-Language Path Aggregation Network
RLL Right Lower Lobe
RMSE Root Mean Squared Error
RNN Recurrent Neural Network
ROI Region of Interest
RPN Region Proposal Network
RUL Right Upper Lobe
RVD Relative Volume Difference

SAM Segment Anything Model
STM Spatial-Temporal Mixer
SVM Support Vector Machine

TN True Negative
TP True Positive
TS Text Simplification

USA United States of America

VDT Volume Doubling Time

WHO World Health Organization
WRN Wide Residual Network

YOLO You Only Look Once

Chapter 1

Introduction

This chapter serves as the introductory section of the thesis. Section 1.1 outlines the context of lung cancer in clinical practice, highlighting the role of imaging modalities in the diagnosis and management of pulmonary nodules. Section 1.2 defines the main objectives of the study, emphasizing the key contributions of the research. Section 1.3 presents the structure of the thesis, providing an overview of each chapter and serving as a guide for the reader.

1.1 Motivation

Lung cancer is the most lethal cancer globally, responsible for 1.8 million deaths in 2020 as reported by the World Health Organization (WHO) (Sung et al., 2021). Although it is the second most common cancer, following breast cancer in terms of incidence, lung cancer typically remains undetected until it reaches an advanced stage. This late detection significantly reduces the efficacy of treatment options and adversely affects patient survival prospects. The 5-year survival rate for lung cancer is just 17% (Wong et al., 2017).

The prevention of lung cancer is of paramount importance due to its strong association with smoking habits. According to Zhou (2019), more than 85% of lung cancer deaths can be attributed to tobacco smoking. Efforts to reduce or eliminate smoking can therefore significantly decrease the annual incidence of this disease. However, it is essential to recognize that smoking is not the sole factor contributing to lung cancer. Other significant factors include exposure to haze and carcinogens, family history, and inherited cancer genes. While tobacco control policies are critical, comprehensive prevention strategies must also address these additional risk factors to effectively combat lung cancer on a broader scale.

Given the asymptomatic nature of lung cancer in its early stages, early detection is particularly challenging. The disease can develop for an extended period without causing noticeable symptoms. Additionally, when symptoms do appear, they often mimic those of benign conditions, such as a persistent smoker's cough, mouth bleeding, chest pain, dyspnea, or fatigue. Screening programs using Computed Tomography (CT) are crucial for early detection, yet they are not

widely implemented. In many countries lacking such programs, a substantial portion of the at-risk population, primarily smokers (17% in Portugal (Guerreiro et al., 2023)), remains unscreened.

The difficulties for the implementation of CT screening programs are significant and include logistical challenges and the substantial resources required for effective deployment. Furthermore, the potential risks associated with repeated exposure to ionizing radiation, though mitigated by advancements in low-dose CT technology, remain a consideration for long-term screening strategies. Addressing these challenges requires innovative approaches, such as leveraging mobile CT units to reach underserved populations, improving patient education on the benefits of screening, and optimizing scheduling systems to reduce bottlenecks. It is essential that these efforts ensure not only broader access but also the effectiveness and safety of the screening process, so that the benefits of early detection are clearly realized.

Once detected, lung nodules identified in CT scans — the primary indicator of lung cancer — require careful evaluation to determine the appropriate course of action. Immediate treatment, such as surgical removal of lung tissue or radio/chemotherapy, can have life-long impacts on a patient's respiratory function, making non-invasive management strategies highly desirable. Alternatively, ongoing surveillance of detected nodules allows physicians to monitor their progression before deciding on any invasive interventions. If a nodule does not exhibit significant growth or changes over time, clinical follow-up can effectively manage the patient's risk, minimizing unnecessary interventions while maintaining vigilance against the potential development of lung cancer.

In this context, follow-up refers to the systematic and regular monitoring of patients with detected lung nodules, aimed at tracking changes over time and guiding clinical decisions. Approximately 3.21 million individuals worldwide were undergoing follow-up for lung cancer, according to 2019 statistics (IHME, Global Burden of Disease Study, 2019). This number places significant pressure on healthcare systems. Guidelines like Lung-RADS (American College of Radiology, 2019) are used to standardize the follow-up process, providing structured recommendations based on nodule size and characteristics. A CT scan is categorized according to predefined criteria, helping medical doctors determine the timing and nature of subsequent actions. By offering a harmonized framework, Lung-RADS enhances the consistency and objectivity in follow-up management, ensuring that clinical decisions are based on uniform criteria rather than individual interpretation. This standardization supports better alignment of follow-up schedules, reducing variability in clinical practice and facilitating appropriate responses to observed changes in nodules.

As the demand for lung cancer follow-ups continues to grow, there is an increasing need to optimize patient monitoring to maintain quality care and efficient use of healthcare. While prevention and early-stage screening are critical for reducing lung cancer incidence and mortality by approximately 30% (Paci et al., 2017), the burden of follow-up care remains substantial, particularly given the number of patients requiring ongoing surveillance. The integration of Artificial Intelligence (AI) in decision support systems offers a promising avenue to mitigate these challenges. AI can enhance the analysis of the large volume of data from CT scans, aiding in recommending accurate and timely follow-up actions. By improving the efficiency of lung cancer follow-up protocols, AI allows healthcare professionals to allocate their attention to more complex cases. Ultimately, AI

also contributes to the trend of non-invasive monitoring, enabling a quick and effective balance in patient reassessment as needed.

1.2 Objectives and Contributions

Deciding to undergo follow-up for lung cancer implies that healthcare professionals **evaluate** the current state of the disease to determine the frequency of monitoring, allowing for timely interventions. It is a proactive measure to **forecast** potential complications or advancements in the disease. Regular follow-up appointments provide an opportunity for patient **reassess** in terms of progress, so that any future changes will be promptly addressed, and strategies to manage the health condition effectively are defined.

The primary goal of this thesis is precisely to advance the state of the art in decision support methods for lung cancer follow-up, particularly focused on the three main stages highlighted in the follow-up process: evaluate, forecast, and reassess. More specifically, the objectives of this research are:

- **Evaluate**: Assess the severity of lung cancer in a CT scan. By analyzing the characteristics of detected lung nodules, automatically provide the timing for the next patient reassessment or identify any additional exams based on a follow-up recommendation guideline;
- **Forecast**: Anticipate potential disease progression, particularly the growth of nodules which is related to the malignancy of cancer, using historical patient data and imaging patterns that may indicate the future course of the cancer;
- **Reassess**: Establish a framework that allows for the identification of nodules in CT scans that were previously mentioned in medical text, enabling medical doctors to easily recognize what was reported in the past and compare it to the current status of findings in the new CT scan.

These objectives motivated and guided the six main contributions presented in this thesis. Among them, the identification and analysis of the state of the art helped to define the directions of this thesis, serving as a support contribution (referred to as **S.**). The support contribution is as follows:

S. Review of the state of the art in computer-aided lung cancer methods in CT: This contribution resulted in a co-authored book chapter at IOP Science (Pedrosa et al., 2022). This is a literature review contribution and provides an overview of the methods proposed for each screening stage of nodules: detection, segmentation, and characterization. It also compares methods for patient diagnosis and management, such as predicting biopsy results from CT image data. Additionally, the chapter discusses the datasets currently available in the literature and future directions for the scientific community. Analyzing these steps is important since the follow-up stage is preceded by screening steps to evaluate the nodules and CT

scans. Moreover, these AI-based methods are crucial when comparisons between the same nodule at different time points are necessary for an accurate understanding of disease progression.

Moving on to the main contributions, this thesis makes several advancements in AI, with a strong focus on Computer Vision (CV) for CT in the processing of thoracic scans and slices centered on lung nodules. It also involves machine learning regression methods using patient input data and Natural Language Processing (NLP) for medical reports. Grouped by the addressed objectives, the following five main contributions are a result of the research presented in this document:

I. Evaluate: One of the lung cancer follow-up guidelines was chosen to automatically recommend the timing of new assessments based on CT slices centered on pulmonary nodules. Two approaches were followed:

(a) End-to-end approach: Provide a follow-up recommendation directly from the nodule cuboid, without outputting any further characterization. Different models are proposed, applying weights during training to account for inter-observer disagreement in the Ground Truth (GT) labels. These weights are assigned to each sample based on the level of agreement among observers: samples with higher agreement have a greater influence on the training loss, while those with higher disagreement contribute less to the updates of the model parameters. This contribution was published in the proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) under the title "Classification of Lung Nodules in CT Volumes Using the Lung-RADS Guidelines with Uncertainty Parameterization" (Ferreira et al., 2020).

(b) Feature-based approach: All steps up to the follow-up recommendation are automatically predicted, namely the characterization of calcification, texture, and size of nodules, including the delineation of the solid part in part-solid nodules when needed. This allowed to define the time-to-act established in the guidelines. This contribution was published in the proceedings of the 2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON) under the title "A Comparative Study of Feature-Based and End-to-End Approaches for Lung Nodule Classification in CT Volumes to Lung-RADS Follow-up Recommendation" (Ferreira et al., 2024a).

II. Forecast the size of lung nodules into one and two years using patient data and the imaging cuboid centered on the lung nodule. The advantages and disadvantages of using a dedicated model for each year *vs.* an adaptive model capable of forecasting both years simultaneously is studied. Different strategies were applied for forecasting malignant nodules and nodules that exhibited growth. This contribution was published in the journal Biomedical Signal Processing and Control under the title "Towards automatic forecasting of lung nodule diameter with tabular data and CT imaging" (Ferreira et al., 2024c).

III. Reassess: Detection of the position of the lung nodules to allow the patient reassessment in next examinations. This research focused on two key areas:

(a) Building a dataset: It was needed to annotate medical reports and objectively validate the corresponding nodules in CT images where these findings were located. Using asynchronous manual annotations of the nodules as a basis for another study, it was possible to compare the number of findings reported textually with the manual annotations. This contribution was published in the journal Scientific Data under the title "LNDb v4: pulmonary nodule annotation from medical reports" (Ferreira et al., 2024b).

(b) Building a text to image tool: With the validated GT, an automatic tool was developed to extract relevant and structured information from medical reports to serve as a basis for an image module that automatically restored nodule locations. This is useful for reassessment if the spatial positions of the nodules are not saved in information systems and for annotating large quantities of images from texts. This research resulted in a patent submission (Ferreira, C. A., Ferreira, A., Coimbra, M., Guimarães, N., Jorge, A., Campilho, A., 2024, "RePick Annotator") due to its significant application value in being added to medical image visualization and annotation systems.

Altogether, the contributions detailed in this thesis have the potential to drive advancements in decision support systems for lung cancer follow-up. By addressing each objective - evaluation, forecasting, and reassessment — through novel machine learning frameworks, the research introduces a cohesive strategy that not only enhances medical efficiency but also adapts dynamically to the complexities of patient care. These AI-driven solutions allow to automate follow-up scheduling, forecast nodule growth with accuracy, and recover image annotations for consistent longitudinal comparisons. Furthermore, by integrating multimodal data sources such as imaging and clinical history, the proposed methods enhance the contextual understanding of disease progression, facilitating more precise, personalized interventions. These innovations streamline healthcare processes and shift the focus from reactive treatments to proactive, data-informed clinical decisions.

The five main contributions are presented in Figure 1.1. Specifically, for the evaluation objective, the two proposed approaches are depicted. In approach **I. (a)**, the follow-up timing is predicted directly, while in approach **I. (b)**, features are predicted prior to determining the follow-up timing. In the forecast objective, contribution **II.**, both image data and patient data are used as the basis for predicting the future diameter of a nodule. Finally, in the reassessment objective, a database (**III. (a)**) and an automatic tool (**III. (b)**) were developed to manually or automatically extract relevant entities, respectively, from medical reports regarding lung nodules. These extracted entities serve to restore image annotations of the nodules.

A complete list of the publications that resulted from the research performed during this doctoral program can be found in Appendix A.

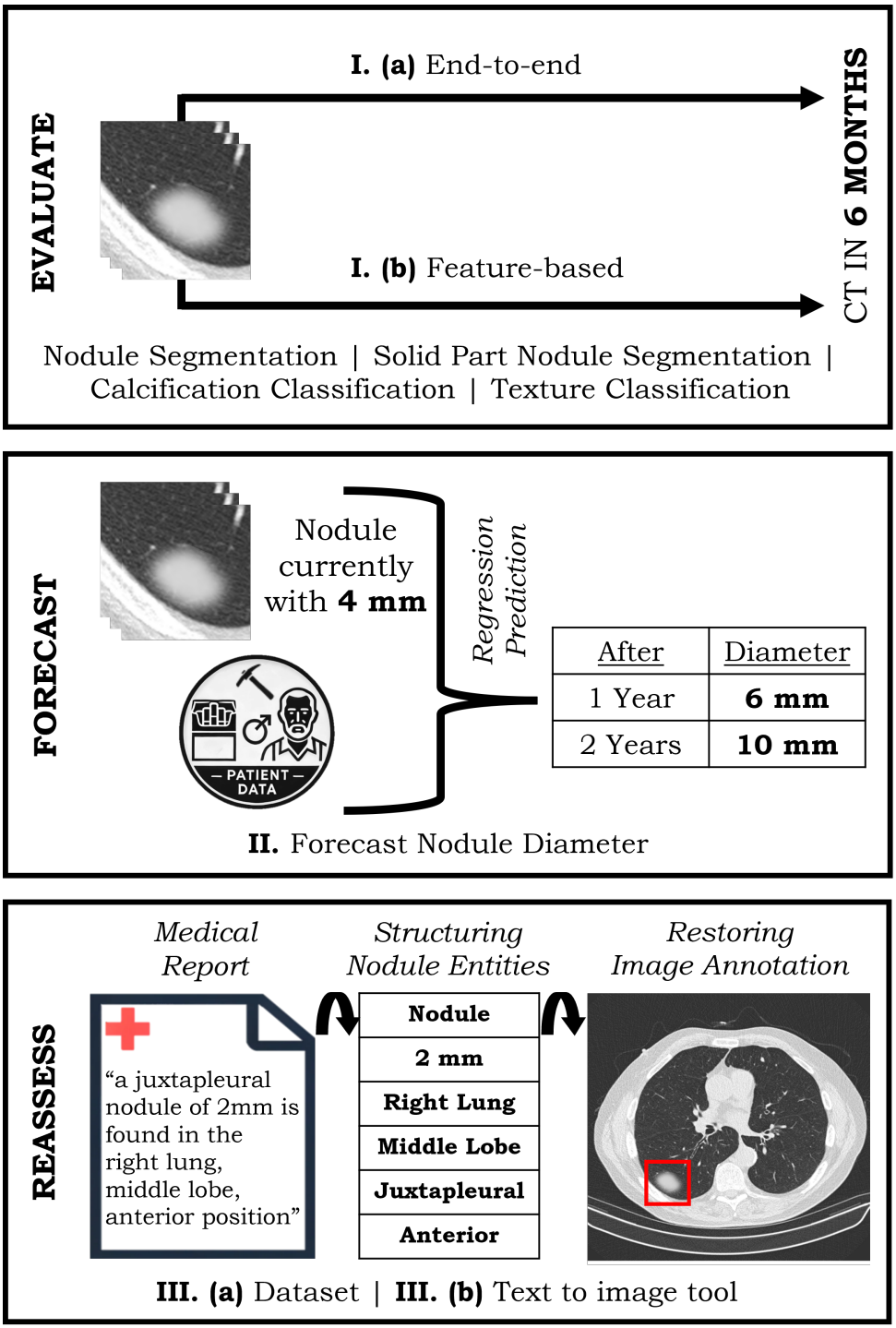


Figure 1.1: Overview of the main contributions in this thesis.

1.3 Thesis Outline

This thesis comprises six chapters in addition to this introductory chapter, where the motivation and objectives of this study are presented.

Chapter 2 describes the anatomy of the lung and the medical management of lung cancer patients. Additionally, this chapter describes the diagnostic imaging modality for lung cancer – CT, the imaging features of pulmonary nodules, the annotation process, and the follow-up guidelines.

Chapter 3 provides an overview of the general methodologies employed in image analysis for lung cancer screening, specifically focusing on nodule detection, nodule segmentation and nodule characterization. This chapter contains part of the content referred to as contribution **S.**.

Chapter 4 addresses the evaluation objective by determining the appropriate monitoring timing for patients according to Lung-RADS guidelines. Two approaches, end-to-end and feature-based, are considered to evaluate the results and understand the advantages and disadvantages. This chapter reflects contributions **I. (a) and I. (b)**.

Chapter 5 focuses on forecasting the diameter of lung nodules in one and two years time frames. This prediction uses variations of input data from both patient data and CT image cuboids around the lung nodules. The analysis is focused on malignant and growing nodules. The chapter also includes details on how the dataset was generated and discusses the new perspectives brought by this research. The chapter is derived from contribution **II.**.

Chapter 6 focuses on reassessment by bridging medical text reports with CT image annotations. It details the creation of a dataset by annotating medical reports and validating their correspondence in CT images. The methodology covers the use of NLP for extracting relevant information from reports and CV techniques for identifying and matching nodules in CT scans. Results, implications, and limitations are discussed. This chapter is based on contributions **III. (a)** and **III. (b)**.

Finally, the concluding chapter, summarizes the key findings of the thesis. It reflects on the implications of the developed AI models and methodologies for clinical practice, outlines potential directions for future work, and emphasizes the need for ongoing innovation to address the challenges identified in this thesis.

Chapter 2

Clinical Background in Medical Imaging for Lung Cancer

This chapter serves as the clinical foundation for this thesis, enabling a deeper understanding of the methodologies and analyses in subsequent chapters. Section 2.1 provides a detailed overview of the anatomical context of the lungs, listing the primary structures and their roles, and describes the typical patient pathway for lung cancer, from risk factors and symptoms to diagnostic and treatment procedures. Section 2.2 addresses the role of Computed Tomography (CT) in diagnosing lung cancer, explaining the imaging features of pulmonary nodules, the diagnostic annotation process, and the guidelines for follow-up. Section 2.3, the discussion, provides a summary of the chapter and outlines the main challenges for the subsequent chapters.

2.1 Lung Cancer

2.1.1 Anatomical Context

The lungs are located within the thorax and are responsible for oxygen renewal in the bloodstream while removing the carbon dioxide and other waste gases. The right and left lungs have three and two lobes, respectively. The air reaches the lungs via the trachea, which branches into the main bronchi. Each of these bronchi subdivide in progressively smaller structures, thus allowing gas exchange within the pulmonary capillaries. The lungs and thoracic wall are covered by a thin layer of tissue called pleura. The pleura, main blood vessels, fissures, and lobes are important anatomical landmarks that aid in the characterization and follow-up of lung nodules.

The heart (and other mediastinum organs) is in the central region of the thoracic cavity, separating the lungs. Structures like the rib cage help to protect these organs from external impacts. Both diaphragm and intercostal muscles participate in the breathing process (Gray and Lewis, 1918). Figure 2.1 depicts the main lung anatomical structures.

A lung nodule appears when abnormal cells grow out of control in a region of the lung. Usually, this process is triggered by a mutation in the genetic code of the cell, which can be

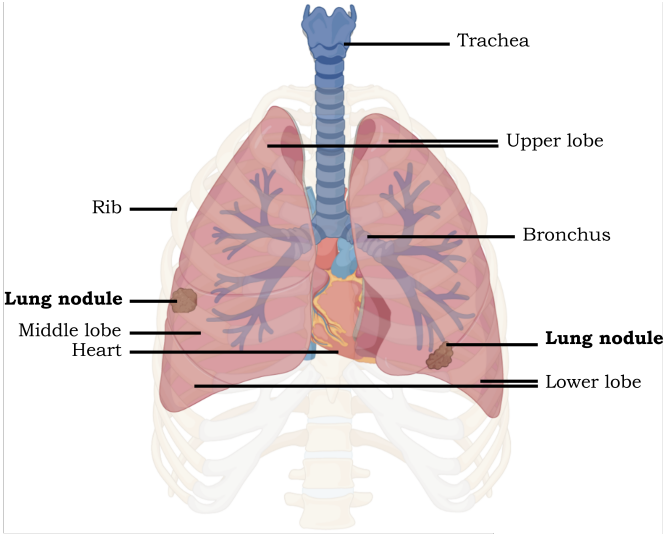


Figure 2.1: Schematic representation of the lungs anatomy.

congenital (inherited gene changes), or appear later in life (acquired gene changes). The cancer cells can spread or metastasize to other organs, compromising additional vital functions.

2.1.2 Patient Pathway

Lung cancer, being a silent disease in the early stages, is not easy to detect. Warning signs may come from routine exams, the genetic profile (family history and existing mutations) or a lifestyle linked to risk behaviors. The risks are age, dust works, heavy smoking and sedentary habits. By combining these risks with other signs and symptoms, the medical doctor may recommend further exams to infer more about the likelihood of having lung cancer.

While X-ray may sometimes be used for initial check-ups, its sensitivity and specificity are significantly lower compared to advanced imaging techniques. The frequency of missed lung carcinoma on X-ray can vary from 12% to 90% (Tack and Howarth, 2019), particularly because nodules smaller than 10 mm may be obscured by other anatomical structures of the lung. CT scans provide a much more detailed view of the lungs, enabling better detection and visualization of pulmonary nodules. Particularly, Low Dose Computed Tomography (LDCT) has emerged as a preferred method for lung cancer screening, offering the advantage of reducing radiation exposure while maintaining high diagnostic accuracy. Other non-imaging exams may also be prescribed, such as sputum cytology, where examining the sputum under a microscope can sometimes reveal the presence of lung cancer cells.

In the United States of America (USA), one of the few countries with a screening program, there is no need to wait for warning signs or symptoms to appear. The American Cancer Society conducts annual lung cancer screenings with LDCT for individuals aged between 55 and 74 years who are in good health, currently smoke or have quit within the past 15 years, and have a smoking history equivalent to a pack a day for 30 years (Wender et al., 2013). This initiative has demonstrated a decrease in lung cancer mortality by approximately 15–20% (van der Aalst et al., 2016;

Pinsky et al., 2013). Other countries remain undecided, awaiting the results of ongoing trials. European countries, such as the Netherlands, Germany, Belgium, France, Denmark, and Italy, have initiated their respective trials (Sanchez-Salcedo et al., 2015). In Portugal, although not yet implemented, lung cancer screening is already in the project phase in some private institutions (Galha, 2020).

If after the screening/early detection step, the presence of malignant cancerous nodules is likely, then a biopsy/collection of cells in the region (to have a histopathological profile and/or protein markers), blood, or sputum exams can be made. With all the collected information, the medical doctor can decide which procedure to take. If the type of lung cancer mutation is possible to be identified in the aforementioned clinical exams, then personalized medicine is an option, choosing the most appropriate chemo and immuno treatments. However, radiation therapy, surgery to remove the malignant tissue, or interventional therapy, to remove fluid or clear obstructions for better patient breathing, remain common solutions in clinical practice.

Whether there is a therapeutic decision or not (if it is a benign case), there should be a periodic follow-up to check for resurgence, resection, or progression of the lung nodules. Target exams and the LDCT continue to be traditional methods. Positron Emission Tomography – Computed Tomography (PET/CT) is usually only an option if the diagnosis is already known, typically by biopsy, to obtain lung cancer staging. Tissues with a higher metabolic rate (higher energy usage), such as lung cancer tumors, absorb more of the tracer substance than other tissues. These areas stand out on the PET/CT scan.

All in all, in the lung cancer journey, the patient will go through different medical specialties, namely family medicine (routine / first warning signs), passing through pneumology when lung disease is suspected, involving radiology in the annotation of the prescribed imaging exams and thoracic surgeons, medical oncologists, radiation oncologists, pathologists if lung cancer is confirmed (Latimer and Mott, 2015). Figure 2.2 shows the phases normally required in the management of lung cancer.

2.2 Computed Tomography Diagnosis

2.2.1 Computed Tomography

A CT scan extends X-ray imaging into a three-dimensional context by capturing multiple two-dimensional projections from different angles. It uses computer-processed X-ray measurements to generate tomographic images (virtual “slices”), allowing visualization inside the body without surgery. An X-ray beam passes through the body, with different tissues absorbing energy at varying rates. The transmitted X-rays are detected on the opposite side and converted into electrical charges through scintillation-based photodiodes or direct semiconductor absorption. These charges are digitized and processed using reconstruction algorithms to generate a detailed image of internal structures. Dense parts of the body, such as bone and nodules, are coded as clear white areas on the image. Softer parts, such as the heart and lungs, are coded as darker areas. The

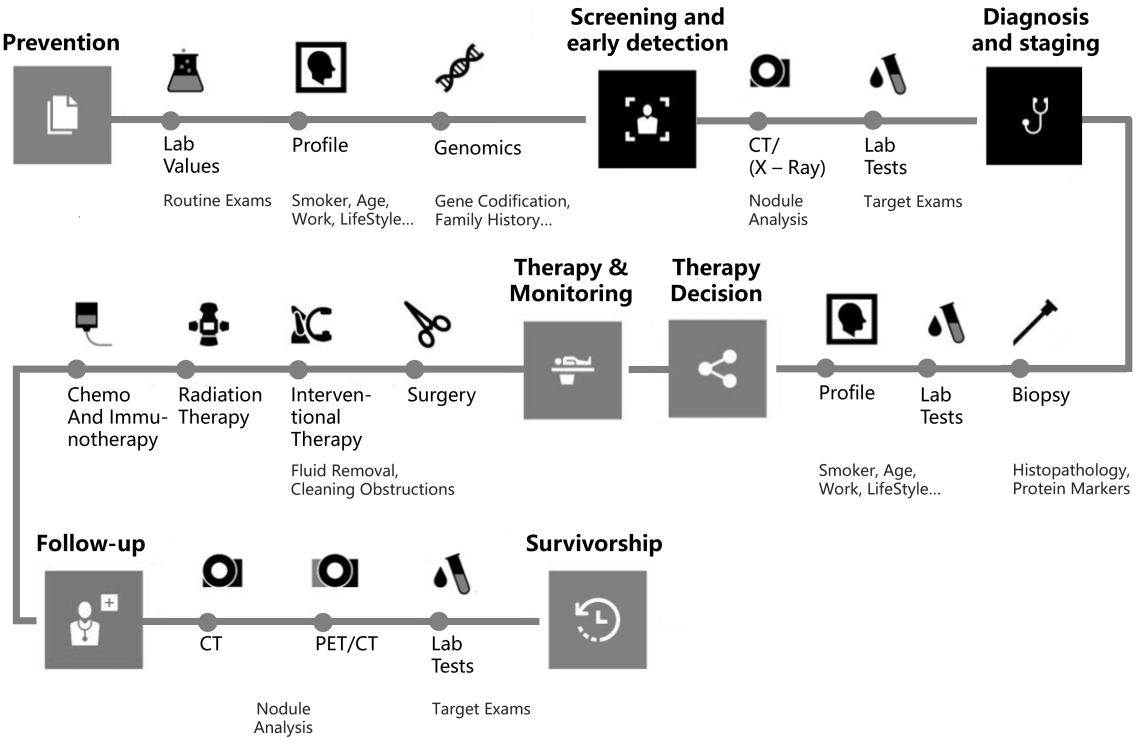
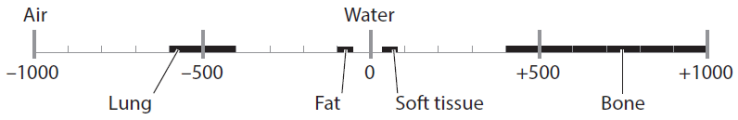


Figure 2.2: Different phases and data used in the lung cancer journey. The different phases are represented by a square, with the data associated with each phase appearing next to the circles along the continuous line extending from the square.

densities are expressed using the Hounsfield Unit (HU), a standardized measure based on a scale centered on the value of water (HU= 0). This scale enables fast visualization of tissues of different densities during scan assessment. The resulting scale is shown in Figure 2.3. In the end, each voxel of the scan has a corresponding attenuation value (or tissue density value) proportional to the amount of energy that was dispersed during the radiation travel.

CT started to be used in some biological observations instead of the standard X-ray procedure due to its ability to provide cross-sectional images with greater detail, quality, and a 3D perspec-



Bone	+400 → +1000
Soft tissue	+40 → +80
Water	0
Fat	−60 → −100
Lung	−400 → −600
Air	−1000

Figure 2.3: HU scale and intensity ranges of anatomic structures. Image from Holmes and Misra (2017).

tive. However, CT scans generally involve higher radiation doses, and it is traditionally recommended not to undergo more than 25 conventional CT procedures in a lifetime due to cumulative radiation exposure risks (Huppmann et al., 2010). With advancements in imaging technology, LDCT has become widely adopted, particularly in lung cancer screening, as it significantly reduces the radiation dose to approximately 20-30% of that from a standard CT while still providing adequate image quality to identify pulmonary nodules. This evolution in CT technology allows for safer, more frequent scans, making LDCT a preferred option in modern clinical practice. Other types of CT, such as contrast-enhanced CT, can also be used, particularly for cancer staging, where a contrast agent is injected into the patient's bloodstream to enhance the visualization of anatomical tissues (Rawson and Pelletier, 2013).

CTs provide a 3D perspective, but in practice, the review is done slice by slice, focusing on the three main anatomic planes (sagittal, coronal, and transverse) for a comprehensive analysis. Figure 2.4 shows an example of an LDCT scan, as well as relevant anatomical structures in the most visualized plane, the transverse. Depending on the acquisition resolution, some structures may be affected by image reconstruction issues. For instance, large values of slice thickness introduce the known partial-volume effect, where the increase in the voxel volume results in higher density averaging and thus blurriness (Goldman, 2008; Soret et al., 2007). In order to avoid this effect, most scans have a slice thickness around 1.5 mm or less.

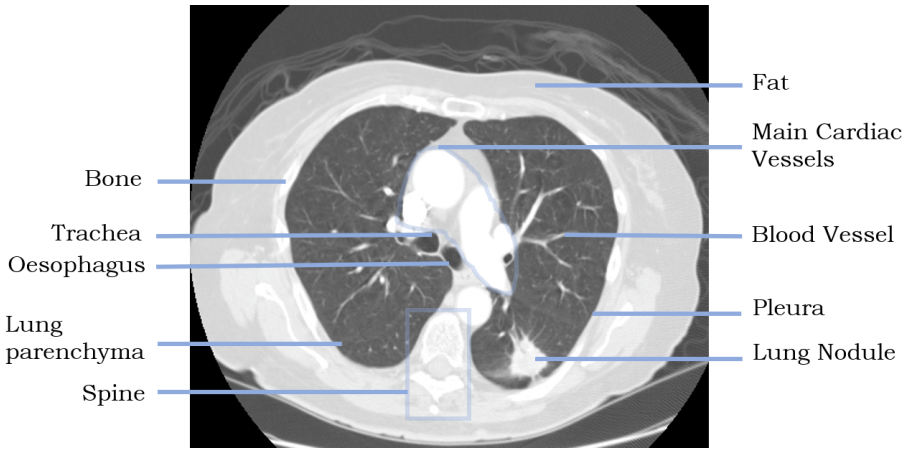


Figure 2.4: A representation of a transverse chest LDCT.

2.2.2 Lung Nodules

CT show existing nodules in the pulmonary region, with the HU intensity of nodules typically higher than those in the parenchyma and lower than those in bones. The 3D view provided by CT enables nodule detection, segmentation, characterization, and even follow-up recommendations. Nodules are usually characterized by their shape parameters (lobulation, margin, sphericity, and spiculation), intensity (texture, subtlety, and calcification), location, and internal structure (whether composed of soft tissue, fluid, fat, or air). Table 2.1 presents the characteristics and corresponding 6 classes mentioned throughout this section.

Table 2.1: Nodule characteristics organized in 1-6 classes.

Nodule Feature	**1**	**2**	**3**	**4**	**5**	**6**
Internal Structure	Soft tissue	Fluid	Fat	Air	-	-
Texture	Non-solid	...	Part solid	...	Solid	-
Subtlety	Extremely subtle	Very subtle	Subtle	Relatively obvious	Obvious	-
Calcification	Popcorn	Laminated	Solid	Non-central	Central	Absent
Sphericity	Linear	...	Ovoid	...	Round	-
Margin	Poorly defined	...	...	...	Sharp	-
Lobulation	No lobulation	...	...	...	Marked	-
Spiculation	No spiculation	...	...	...	Marked	-
Malignancy	Highly unlikely	Moderately unlikely	Indeterminate	Moderately suspicious	Highly suspicious	-

Despite the possible characterizations referred to in the literature, usually size, location, texture, and calcification are mostly taken into account in the radiologists' routine annotations of the CT scans, as they are the most important ones in the recommending follow-up. Regarding size, the threshold for considering a large or small nodule is an equivalent diameter of 5 mm, *i.e.* the diameter of the sphere having the same nodule volume. Location-wise, nodules can be divided as peri-fissural, juxta-vascular, juxta-pleural, or isolated (Van Ginneken et al., 2010). This means that nodules are located attached to fissures, blood vessels, pleura, or none of them, respectively. Texture-wise, nodules can be classified as solid, part-solid, and Ground-Glass Opacity (GGO) (also known as non-solid). Solid nodules have well-defined borders and high uniform contrast regarding the lung parenchyma. Non-solid nodules have a diffused distribution over the parenchyma, commonly appearing as cloudy structures. Hence, they present a low margin contrast with the lung parenchyma and their shape is difficult to define. Finally, part-solid nodules combine characteristics from both solid and non-solid abnormalities, with a solid core and diffused boundaries. Calcification is commonly noted as a reliable indicator of a benign nodule (Erasmus et al., 2000), and further classification into specific calcification patterns (complete, central, popcorn, non-central, or other) are also common. Examples of these nodule types are shown in Figure 2.5.

The malignancy of a nodule cannot be conclusively determined solely from CT visual features, as it is often subjective and relies on multiple low-level features such as calcifications, nodule texture, and spiculations. The assessment remains speculative, heavily based on the radiologist's experience and empirical knowledge. Although automated approaches using Artificial Intelligence (AI) have been developed to predict malignancy confirmed by biopsy, these methods still fall short of the expected performance and trustworthiness (Silva et al., 2020). A biopsy is still required for definitive verification and thus remains standard practice.

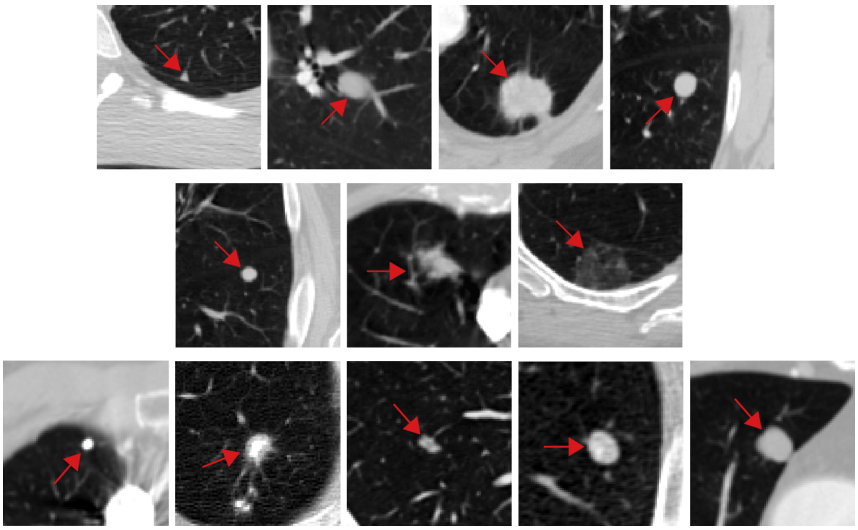


Figure 2.5: Nodule examples. Central axial slice of each nodule is shown, where red arrows indicates the nodule. Classification based on location (1st row, from left to right): peri-fissural, juxta-vascular, juxta-pleural and isolated nodule. Classification based on texture (2nd row, from left to right): solid, sub-solid and non-solid nodules. Classification based on calcification (3rd row, from left to right): complete, central, popcorn, non-central and absent calcification.

2.2.3 Annotation Process of CT Scans

The pneumologist is the physician who makes the diagnosis of lung diseases in patients with symptoms or warning signs. For this, the radiologist compiles and communicates all the information about abnormalities that can be observed in imaging exams, particularly in LDCT. Even if lung cancer is suspected, the imaging assessment does not exclusively focus on nodules, allowing the pneumologist to gather more comprehensive information, as the diagnosis could involve other conditions or reveal additional lung diseases.

The search for nodules is conducted using either a scanning or drilling search strategy. In scanning, a radiologist examines an entire slice for nodules before moving to the next, often employing techniques such as Maximum Intensity Projection (MIP) to assess depth information. Alternatively, in drilling, the radiologist focuses on a single quadrant of the volume at a time, scrolling through all slices of the scan to account for 3D information (Drew et al., 2013; Diaz et al., 2015).

Annotating CTs scans is a demanding task for radiologists due to the need to analyze numerous slices with abnormalities that vary in shape, size, and characteristics. Moreover, small nodules can easily be confused with other lesions, such as scars, which rarely have diagnostic relevance. When attached to other structures, such as vessels or the pleura, nodules can be obscured and/or have altered shapes. Even after correct nodule identification, their characterization is highly subjective, heavily dependent on experience, and can vary widely between centers, making it prone to error.

Studies have shown that radiologists' sensitivity in detecting these abnormalities ranges from 30% to 97%, with false positive (FP) rates of 0.6 to 2.1 per patient (Rubin, 2015). To enhance diagnostic accuracy, obtaining a second opinion from another radiologist is strongly recommended. Despite the significant economic and human resource implications, previous studies have revealed

that a second reader can aid in the detection of up to 19% more nodules (Wormanns et al., 2005; Wang et al., 2012).

Typically, CT findings are documented in text reports and imaging annotations within information systems. Text reports, while detailed, often suffer from inconsistency in structure and completeness, particularly in describing the size and characterization of abnormalities. Imaging annotations, although precise, may not be compatible with all information systems and can be challenging to interpret quickly.

Nowadays, automatic decision support systems, including Computer-Aided Diagnosis (CAD), can act as an additional tool in clinical practice by enhancing the documentation and diagnostic process in lung cancer care. These systems can swiftly and accurately analyze CT scans, identifying subtle patterns that may be missed by human observers. CAD systems, in particular, provides quantitative assessments and highlighting suspicious areas, thereby assisting radiologists in their evaluations. When integrated seamlessly with existing health information systems, they improve the accessibility and standardization of both textual and visual data. Furthermore, they aid radiologists in prioritizing cases and making more informed decisions by standardizing the presentation of findings, which enables clearer communication between oncologists, radiologists, and surgeons. As a strong procedural ally, these systems enhance diagnostic accuracy and alleviate some of the workload, streamlining workflows and ultimately supporting decision-making, contributing to more efficient lung cancer management.

2.2.4 Follow-up Guidelines

The follow-up guidelines for lung cancer management are crucial for determining the appropriate clinical pathway for patients. These guidelines are designed to support clinical decisions by providing structured approaches to evaluation, guiding follow-up and patient management. Three well-defined guidelines are commonly referenced: the Lung-RADS criterion (American College of Radiology, 2019), the Fleischner guidelines (MacMahon et al., 2017), and the British Thoracic Society (BTS) guidelines (Callister et al., 2015).

Table 2.2 summarizes the main differences of guidelines across several key aspects: geographical application, function, features, applicability, and follow-up recommendations. The function of these guidelines can be in the context of screening, which involves detecting nodules for the first time, or incidental findings, which are found during imaging exams conducted for other reasons. In terms of features, in addition to patient-specific factors, the guidelines emphasize key characteristics of nodules, such as size, texture, and location, which are typically derived from detection, segmentation, and characterization processes employed by radiologists or automated systems to inform clinical decision-making. The follow-up recommendations differ slightly. While Lung-RADS and Fleischner guidelines offer similar follow-up options, the BTS guidelines uniquely incorporate longitudinal assessments for patient management by considering nodule growth over time.

Table 2.2: Comparison of lung cancer follow-up guidelines

Guideline	**Geographical Application**	**Function**	**Features**	**Applicability**	**Follow-up Recommendations**
Lung-RADS	America	Screening	– Calcification – Texture – Nodule diameter – Solid diameter of sub-solid nodules	Per nodule basis, highest ranking nodule determines follow-up	– No follow-up or CT at 12 months – CT at 6 months – CT at 3 months – PET/CT, CT with contrast, or biopsy
Fleischner	Europe	Incidental findings	– Number of nodules – Sizes (diameter or volume) – Texture – Patient risk indicators (age, smoking, margins, location)	Whole patient evaluation	– No follow-up or CT at 12 months – CT at 6 months – CT at 3 months – PET/CT, CT with contrast, or biopsy
BTS	United Kingdom	Screening and incidental findings	– Nodule size – Growth – Texture – Spiculation – Location – Patient demographics and history	Based on largest nodule	– Discharge – CT at 3 months – CT at 1 year – CT at 2 years – Specific imaging and therapeutic interventions

The seamless integration of these guidelines into clinical workflows is a crucial aspect of their utility. This integration might involve incorporating the guidelines directly into imaging analysis and annotation software used by medical doctors, allowing for easy access and application of guideline criteria during patient assessments. For instance, a medical doctor could simply click to apply guideline recommendations or input specific parameters directly into the software, streamlining the diagnostic process. Machine learning methods can support this integration by automating or aiding in the application of guidelines. However, the practical integration of such technologies into existing clinical systems is still evolving and presents challenges. While some progress has been made, the complexity and variability of software platforms, as well as the need for continuous updates and validation, can make full integration difficult. Ensuring that these tools are user-friendly and seamlessly embedded into the clinician's routine is essential for maximizing their utility and impact.

2.3 Discussion

This chapter lays the groundwork for understanding the complexity of detection and characterization of lung nodules in clinical practice, highlighting the variability and difficulties radiologists face when interpreting LDCT scans. While LDCT plays a key role in detecting nodules, the reliance on radiologist expertise introduces significant variability in evaluations. Structured guidelines aim to bring consistency to follow-up decisions, ensuring that nodules are monitored effectively, especially when immediate treatment is not required or recommended.

Automated approaches offer promising support to radiologists in refining the follow-up process, reducing inconsistencies and enhancing the precision of monitoring strategies. These developments are intended to help radiologists identify the optimal timing for interventions, thereby improving patient care and enabling timely treatment when necessary. This transition towards more reliable and integrated follow-up methods will be further explored in the next chapters.

Chapter 3

Review on Automatic Methods for Lung Cancer Follow-up

This chapter provides a comprehensive review of the automatic methods related to the steps for lung cancer follow-up. This constitutes contribution **S.** of this thesis, as outlined in Chapter 1. Following a reflection on the current state of proposed methods and the definition of metrics predominantly used in these approaches in section 3.1, the main publications concerning nodule detection, segmentation, and characterization are detailed in sections 3.2, 3.3, and 3.4, respectively. The key conclusions drawn from this review are presented in section 3.5.

The contents of this chapter, especially from section 3.2 onwards, are based on a book chapter[1]. All the authors contributed equally.

3.1 Introduction

3.1.1 Overview

The adoption of Artificial Intelligence (AI) in radiology has seen a significant increase in recent years, enhancing the accuracy and efficiency of diagnostic tasks. AI systems, particularly in the context of lung cancer, provide substantial benefits in follow-up tasks. Computer-Aided Diagnosis (CAD) systems act as a second opinion, focusing on detection, segmentation, and characterization of lesions in a manner that is objective and data-driven. These systems primarily evaluate visual data, which can help to reduce some forms of subjective interpretation inherent to manual assessments. However, it is crucial to integrate visual analysis with patient-specific contextual information, such as medical history and clinical data, to ensure comprehensive and personalized follow-up assessments. The deployment of AI in radiology is not intended to replace the interpretative skills of radiologists but to complement them by providing objective analyses that can support and verify human assessments.

[1] João Pedrosa, Guilherme Aresta, and **Carlos Ferreira**. *Computer-aided lung cancer screening in computed tomography: state-of the-art and future perspectives*, chapter 4. IOP Publishing, January 2022. doi: 10.1088/978-0-7503-3355-9ch4.

The literature has been abundant with proposals for automatic methods specifically in the tasks crucial for patient follow-up: detection, segmentation, and characterization of lung nodules. Figure 3.1 illustrates the trend of literature proposals for these three tasks together with follow-up publications over the past 20 years. Each bar highlights the number of publications related to Deep Learning (DL) in dark blue, followed by other methodological publications in light red, so that the sum of the two colored bars represents the total number of publications per year for each task.

This literature analysis was conducted using the Scopus repository, focusing on four primary categories, as summarized in Table 3.1. For each category, specific queries were formulated by searching for the exact terms in the title, abstract, and keywords. The terms applied across all searches included: `"CT"`, `"computed tomography"`, `"computerized tomography"`, `"tomography"`, `"tomographic"` combined with `"CAD"`, `"computer-aided"`, `"computer-aid"`, `"computer aided"`, `"computer aid"`, `"computer assisted"`, `"automatic"`, `"automated"`, and `"automation"`.

Table 3.1: Search terms used in each category.

Category	**Search Terms**
Detection	`"lung nodule detection"`, `"detection of lung nodules"`, `"lung nodules detection"`, `"detecting lung nodules"`, `"pulmonary nodule detection"`, `"detection of pulmonary nodules"`, `"pulmonary nodules detection"`, `"detecting pulmonary nodules"`
Segmentation	`"lung nodule segmentation"`, `"segmentation of lung nodules"`, `"lung nodules segmentation"`, `"segmenting lung nodules"`, `"pulmonary nodule segmentation"`, `"segmentation of pulmonary nodules"`, `"pulmonary nodules segmentation"`, `"segmenting pulmonary nodules"`
Characterization	`"lung nodule classification"`, `"classification of lung nodules"`, `"lung nodules classification"`, `"classifying lung nodules"`, `"pulmonary nodule classification"`, `"classification of pulmonary nodules"`, `"pulmonary nodules classification"`, `"classifying pulmonary nodules"`, `"lung nodule characterization"`, `"characterization of lung nodules"`, `"lung nodules characterization"`, `"characterizing lung nodules"`, `"pulmonary nodule characterization"`, `"characterization of pulmonary nodules"`, `"pulmonary nodules characterization"`, `"characterizing pulmonary nodules"`
Follow-up	`"lung nodule follow-up"`, `"follow-up of lung nodules"`, `"lung nodules follow-up"`, `"following up lung nodules"`, `"pulmonary nodule follow-up"`, `"follow-up of pulmonary nodules"`, `"pulmonary nodules follow-up"`, `"following up pulmonary nodules"`, `"lung cancer follow-up"`, `"follow-up of lung cancer"`, `"following up lung cancer"`

For deep learning-specific studies, the search was refined using the terms `"deep learning"`, `"deep neural network"`, `"convolutional neural network"`, `"CNN"`, `"convnet"`, and `"convnets"`.

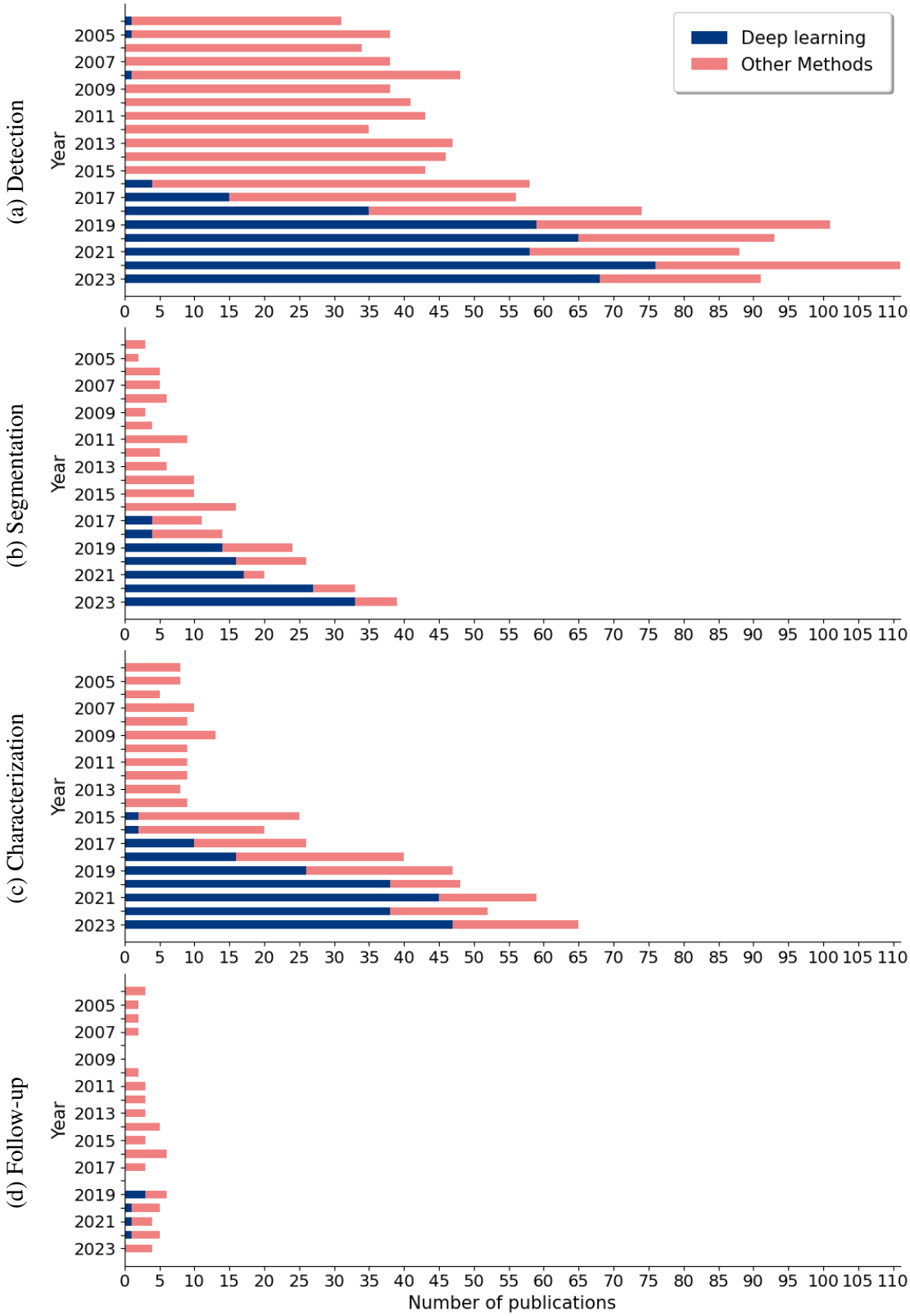


Figure 3.1: Number of scientific publications on lung nodule detection, segmentation, characterization, and follow-up since 2004 up to 2023. Information extracted from Scopus.

Observing Figures 3.1a, 3.1b, and 3.1c, there is a notable increase in contributions following the DL boom. DL is a branch of AI that focuses on training large Neural Networks (NNs) to learn and make decisions from vast amounts of data. This approach has transformed the landscape of computational methods. Although the potential of DL proposals was already evident by 2012 (Krizhevsky et al., 2012), it was from 2015 onwards that literature contributions became more abundant. This surge was driven by significant hardware improvements and the introduction of influential architectures such as the U-Net (Ronneberger et al., 2015).

In recent years, a new DL approach has emerged with the introduction of transformers, highlighted by the publication "Attention is All You Need" (Vaswani et al., 2017). Transformers leverage attention mechanisms to process sequential data more efficiently, addressing limitations of previous models in handling long-range dependencies. This approach has revolutionized Natural Language Processing (NLP), giving rise to advanced models like ChatGPT. While this represents a significant leap forward, it is not yet predominant enough to be highlighted in the bar charts. Nevertheless, the ongoing advancements in transformers and their applications suggest that the early stages of another major shift in AI methodologies may be occurring.

Since follow-up is intimately dependent on the detection, segmentation, and characterization of nodules, the published research focused on follow-up are not predominant in the literature, as seen in Figure 3.1d. In fact, the mentioned three tasks (detection, segmentation and characterization) are the tools that provide enough information for patient management using the follow-up guidelines. Therefore, this may be the reason why publications focused solely on patient follow-up are not found. For this reason the next subsections will focus solely on methods for the detection, segmentation, and characterization of nodules.

Before moving on to the next sections, it is important to note that the rapid development of proposals in the literature is closely linked to the availability of large, diverse, and publicly accessible datasets. These datasets need to be annotated by clinicians, detailing the location, contour, and characteristics of nodules. The increased availability of large datasets has been facilitated by the push towards digitalization, the establishment of standards, and heightened awareness of how to share data publicly while adhering to General Data Protection Regulation (GDPR) regulations. The most commonly referenced datasets in the literature include Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) (Armato III et al., 2011), National Lung Screening Trial (NLST) (Team, 2011), and Lung Nodule Database (LNDb) (Pedrosa et al., 2019). Detailed discussions on the nuances and specifics of these datasets will be provided in the following chapters. LUng Nodule Analysis 2016 (LUNA-16) (Setio et al., 2017) is a subset of LIDC-IDRI, specifically for Computed Tomography (CT) scans with a slice thickness of 3 mm or less.

3.1.2 Performance Evaluation

Evaluating the performance of AI algorithms is essential for validating their potential for follow-up recommendations. Performance metrics provide a quantitative basis for assessing how well these algorithms perform crucial tasks like detection, segmentation, and characterization of lung

nodules. This evaluation not only substantiates the benefits of CAD systems as a second opinion but also ensures their reliable integration into clinical practice, supporting radiologists' interpretations and enhancing the precision of follow-up assessments.

The performance of an algorithm is typically evaluated by comparing its predicted scores, which indicate the likelihood of a sample belonging to a specific class, with the actual classifications, referred to as the Ground Truth (GT). In a binary classification setting, the score is converted into a decision criterion (*e.g.*, by binarization via a threshold), which allows each inferred sample to be labeled as either a True Positive (TP), False Positive (FP), True Negative (TN), or False Negative (FN). Here, TP and TN represent the number of samples correctly classified as belonging to the positive or negative class, respectively. Conversely, FP and FN denote the number of samples incorrectly classified as belonging to the positive or negative class, respectively. These quantities are then used to compute various performance metrics as shown below:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{3.1}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3.2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{3.3}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{3.4}$$

$$\text{Jaccard} = \frac{\text{TP}}{\text{TP}+\text{FN}+\text{FP}} \tag{3.5}$$

$$\text{Dice} = \frac{2 \times \text{TP}}{2 \times \text{TP}+\text{FP}+\text{FN}} \tag{3.6}$$

Accuracy measures the proportion of correctly classified samples among all the samples, providing an overall view of the model's performance. However, it can be misleading in cases of class imbalance, where one class may dominate the dataset. Sensitivity (or recall) measures the proportion of correctly identified positive cases, while precision indicates the proportion of true positive predictions among all positive predictions. The F1 Score offers a more balanced measure, combining both sensitivity and precision into a single metric that is particularly useful when the focus is on correctly identifying the positive class, especially in datasets where FN and FP have significant consequences. The Jaccard (or Intersection over Union (IoU)) and Dice are two of the most commonly used metrics for evaluating the quality of segmentations, such as those of lung nodules, where a specific region or class needs to be segmented from the pulmonary background. Although both Jaccard and Dice measure the overlap between the predicted and true sets, Dice tends to be more lenient, giving slightly more weight to TP.

By varying the decision criteria applied to the prediction score, different values of TP, FP, TN, and FN can be obtained. This allows the system's performance to be evaluated across different trade-offs of sensitivity and specificity. In scenarios where the concept of TN does not apply, such as in lung nodule detection, performance can be evaluated by assessing the detection sensitivity as a function of the number of FP detections.

When dealing with continuous value predictions, it is essential to use metrics such as Mean Absolute Difference (MAD) and Root Mean Squared Error (RMSE) to assess the performance of

regression models by measuring prediction errors.

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.7}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.8}$$

where, y_i represents the true values, $\hat{y}_i$ denotes the predicted values, and n is the total number of observations. MAD calculates the average magnitude of errors, treating all deviations equally. In contrast, RMSE squares the errors, thus emphasizing larger discrepancies and making it more sensitive to outliers.

3.2 Lung Nodule Detection

Automatic lung nodule detection involves identifying locations on the CT scan where a nodule likely exists. The standard CAD pipeline for nodule detection in CT images typically consists of two main stages: 1) initial candidate detection and 2) FP reduction, as shown in Figure 3.2. This process can also be integrated with lung segmentation, as it helps to delineate the search region more precisely.

The objective of the initial candidate detection step is to maximize the number of detected TP nodules. Due to the heterogeneity of lung nodules, this detection step is highly complex. To achieve high detection sensitivity, many FPs are also identified. These false detections motivate a second step known as FP reduction. In this stage, features are extracted from the candidates and used to reduce the number of FPs through classification. While the goal is to preserve the initial detection sensitivity, achieving an optimal balance between sensitivity and reducing FPs remains challenging in CAD detection systems.

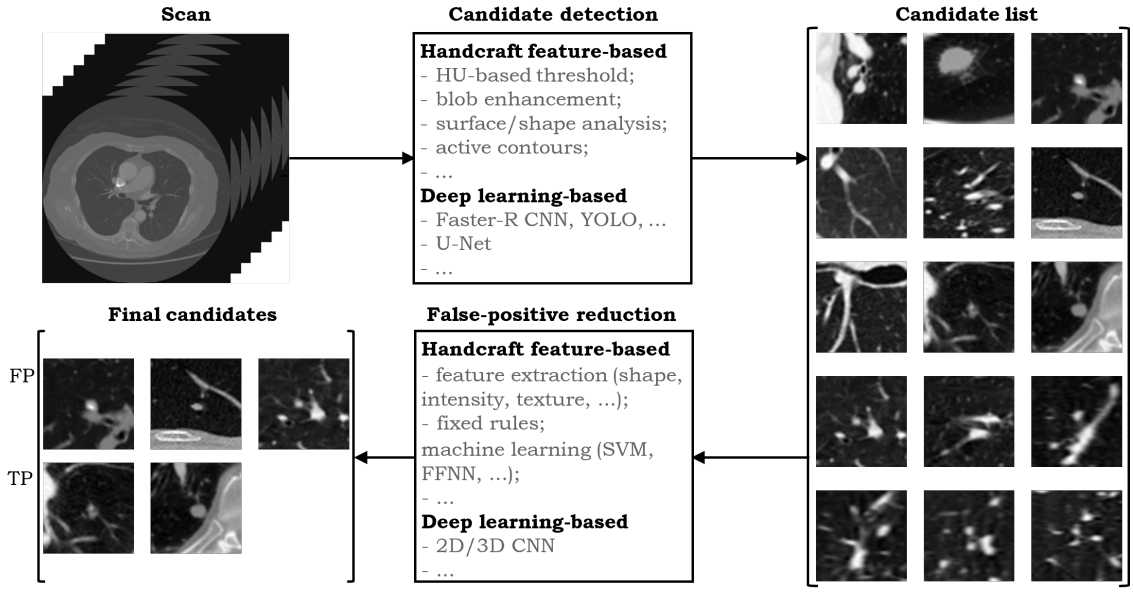


Figure 3.2: Standard lung nodule detection pipeline.

A summary of the performance of the methods detailed in this section is available in Table 3.2. The table shows whether the nodule detection method is 2D, 2.5D, or 3D, with specific mention if a DL technique was applied. Most studies report the performance in terms of sensitivity, either at a given (or several) number of FPs per CT scan or at different Agreement Levels (ALs) (the number of radiologists that reported a given nodule upon reading the CT), making direct comparisons between methods a challenging process. However, an increasing trend in reported detection performance (both in terms of higher sensitivity and fewer FPs per scan) can be observed, especially since the advent of DL techniques. Performance is reported in terms of the evaluation perspective, considering the quantity of data and the variability in the datasets used.

In the past, lung nodule detection methods predominantly relied on intensity and sphericity rules, despite the variability in nodule characteristics. Initial approaches used fixed thresholds, such as -700 Hounsfield Unit (HU) (Enquobahrie et al., 2007; Lopez Torres et al., 2015), to distinguish lung nodule tissue from other structures. These thresholds could be refined using techniques like mixture of Gaussians (da Silva Sousa et al., 2010; Santos et al., 2014) or histogram analysis (Choi and Choi, 2014) to account for the diverse textures of nodules. Some methods also explored the natural sphericity of nodules (Dehmeshki et al., 2007), employing Hessian-based approaches to enhance spherical structures while suppressing vessel-like ones (Li et al., 2008), using metrics like shape-index and curvedness to refine candidate detection (Retico et al., 2009). These early detection systems often produces rough segmentations, necessitating additional steps to improve candidate segmentation, such as morphological operations, active contours Wang et al. (2015); Cascio et al. (2012); Wang et al. (2015) and intensity region growing (Brown et al., 2014; de Carvalho Filho et al., 2014), to better isolate nodules from surrounding structures like vessels and the pleura.

Earlier systems often struggled with high FP rates, primarily due to the complexity of the chest CT scans and the similarity of nodules to structures such as scars and lymph nodes. There was no clear FP reduction phase; rather, some fixed rules were applied, which could easily lead to overfitting in smaller datasets. As nodule detection evolved, methods progressed to more sophisticated approaches incorporating statistical models and machine learning. The major breakthrough came with the introduction of pure classification phases, where models like Support Vector Machines (SVMs) (Hearst et al., 1998) and FeedForward Neural Networks (FFNNs) were used. These models leverages a wide array of features, including texture and contextual information extracted from gradient-based techniques (Tan et al., 2011; Teramoto and Fujita, 2013; Setio et al., 2015) allowing them to learn pixel-level nuances that purely handcrafted methods could not capture. The performance of these systems heavily depends on the number and type of features used, leading to extensive research on creating, selecting, and refining feature sets. Consequently, some contributions have focused on automating the optimization process, further enhancing the capability of models to accurately distinguish nodules.

Table 3.2: Performance of selected automatic lung nodule detection algorithms.

Author	Method	Dataset	Sensitivity/FPs (No. of scans)	Comments
Dehmeshki et al. (2007)	3D	Private	0.90/15 (70)	
Enquobahrie et al. (2007)	3D	Private	0.94/7.1 (250); 0.91/6.9 (250)	Isolated nodules
Li et al. (2008)	3D	Private	0.75/1.6 (117)	
Retico et al. (2009)	3D	Private	0.44/1.0 (42)	Juxta-pleural nodules
da Silva Sousa et al. (2010)	3D	Private	0.85/0.4 (33)	
Tan et al. (2011)	3D	LIDC-IDRI	0.88/4.0 (125)	
Cascio et al. (2012)	3D	LIDC-IDRI	0.88/2.5 (84)	
Teramoto and Fujita (2013)	3D	LIDC-IDRI	0.80/4.2 (84)	
Brown et al. (2014)	3D	LIDC-IDRI	0.79/1.0 (108)	AL $\geq$ 3
de Carvalho Filho et al. (2014)	2D	LIDC-IDRI	0.86/1.8 (28)	Isolated nodules
Choi and Choi (2014)	3D	LIDC-IDRI	0.98/6.8 (84)	
Santos et al. (2014)	3D	LIDC-IDRI	0.91/1.2 (28)	
Lopez Torres et al. (2015)	3D	LIDC-IDRI	0.70/8.0 (899); 0.80/8.0 (899)	
Setio et al. (2015)	3D	LIDC-IDRI	0.94/1.0 (888)	Solid nodules; Nodule radius > 5mm
Wang et al. (2015)	3D	LIDC-IDRI	0.88/4.0 (108)	
Setio et al. (2016)	2.5D-DL	LUNA-16	0.85/1.0 (888)	
Dou et al. (2016)	3D-DL	LUNA-16	0.85/1.0 (888)	
Ding et al. (2017)	2.5D-DL	LUNA-16	0.92/1.0 (888)	
Aresta et al. (2018)	2.5D-DL	LUNA-16	0.93/0.3 (888)	
Cheng et al. (2019)	2D-DL	Private	0.74/2.0 (1000)	
Zheng et al. (2020)	2.5D-DL	LUNA-16	0.93/1.0 (888)	
Kaluva et al. (2020)	3D-DL	LNDb	0.71 (58)	AL $\geq$ 2
Katz et al. (2020)	3D-DL	LNDb	0.58/1.0 (58)	AL $\geq$ 2
Mei et al. (2021)	2D-DL	LIDC-IDRI	0.95 (1012)	
Liu (2022)	2.5D-DL	LUNA-16	0.93 (888)	
Mkindu et al. (2023)	3D-DL	LUNA-16	0.94/1.0 (888); 0.97/4.0 (888)	

Label: NA - Not Available.

Recent advancements in DL have significantly improved lung nodule detection by reducing the necessity for manual feature design. These methods leverage Convolutional Neural Networks (CNNs), a type of NN particularly effective for image processing tasks. CNNs use layers of convolutional filters to automatically learn and extract relevant features from raw input images, which enhances overall performance. DL has been successfully applied for both initial candidate detection and FP reduction in lung nodule detection.

Training DL networks effectively requires substantial amounts of data and computational power, especially for 3D data such as CT scans. Consequently, many approaches use 2D representations of scans for initial nodule detection. This architecture allows the network to generate a 3D probability map through a sliding window approach. In some cases, a 2.5D approach is used, which involves analyzing multiple adjacent 2D slices to provide some 3D contextual information without the full complexity of 3D data.

Although primarily designed for segmentation purposes, the U-Net architecture (Ronneberger et al., 2015), a type of CNN, has demonstrated success in detecting nodule candidates using axial slices (Cheng et al., 2019). U-Net consists of an encoder that captures context through successive convolutional layers and a decoder that enables detailed feature extraction via upsampling and convolutional layers. Through rough predictions, thresholding, and morphological erosion, potential nodules are detected. However, it is clearly less efficient compared to other strategies focused solely on detection.

Other authors have applied dedicated object detection networks, such as the Faster Region-based Convolutional Neural Network (R-CNN) (Ren et al., 2017), achieving high performance on candidate detection by assessing axial slices in a depth-wise sliding fashion (Ding et al., 2017; Kaluva et al., 2020). Initially, a NN performs near pixel-wise feature extraction, generating feature maps. These feature maps serve as input for the Region Proposal Network (RPN), which proposes regions likely to contain objects. A second network then analyzes these proposed regions, evaluating different rectangular areas to accommodate various object sizes and scales. Regions with higher classification scores and greater similarity to the ground truth are retained for final object detection. To adapt the Faster R-CNN for lung nodule detection, an upsample layer was added to modify the feature map output, making it more suitable for the small size of the nodules relative to the scan. Additionally, a composite image of three adjacent axial slices is used as input to provide 3D context. The resulting candidate list is then reduced through a FP reduction step, adhering to the standard detection pipeline (Zheng et al., 2020).

The You Only Look Once (YOLO) object detection framework has become the most adopted approach for nodule detection (Aresta et al., 2018), particularly because it avoids the need for FP reduction. The YOLO framework revolutionized object detection by framing it as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities. Initially, YOLO divides the image into a grid, with each grid cell predicting a fixed number of bounding boxes and confidence scores for the presence of objects. Unlike Faster R-CNN, which has an intermediate step for generating region proposals and lacks an objectness score, YOLO includes an objectness score for each proposal. This results in more proposals from YOLO, but

with varied objectness scores, while Faster R-CNN performs a higher level of filtering through its RPN step. The key advantage of YOLO is its speed, as it processes images in real-time due to a single-stage architecture, contrasting with the multi-stage approach of models like Faster R-CNN. A representative illustration of the stages and outputs of Faster R-CNN and YOLO can be found in Figure 3.3.

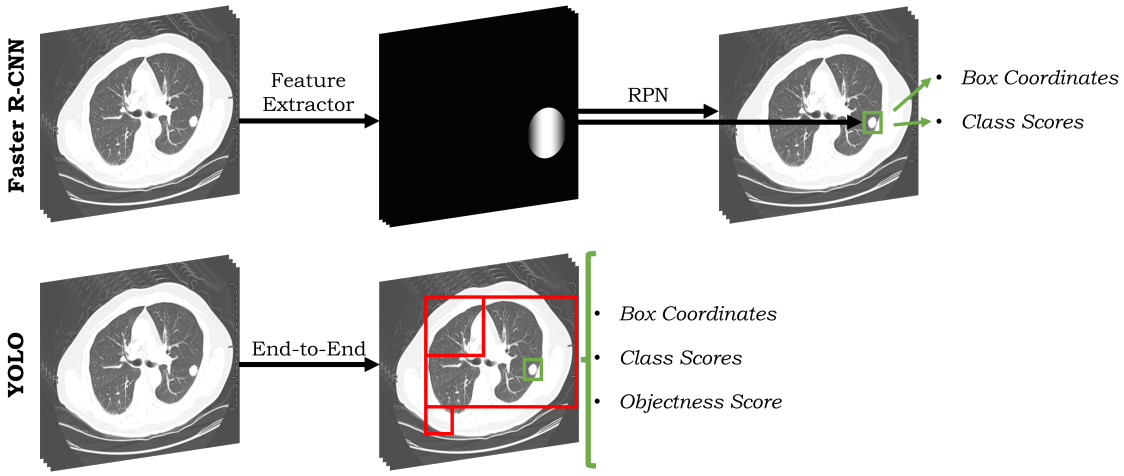


Figure 3.3: Diagram of the Faster R-CNN and YOLO networks.

Over the years, YOLO has undergone several iterations, each improving upon its predecessor. YOLOv2 introduced batch normalization and high-resolution classifiers. YOLOv3 further improved detection by adopting a feature pyramid network, allowing for better detection of small objects by incorporating multi-scale predictions. YOLOv4 (Mei et al., 2021) and YOLOv5 (Liu, 2022) focuses on optimizing speed and accuracy, incorporating advanced techniques such as mosaic data augmentation, techniques for better information flow between layers, and methods to combine features from different scales. These enhancements have made YOLO robust for real-time applications, with a significant reduction in the computational burden while maintaining high accuracy. For nodule detection, YOLO's ability to predict both nodules and nodule confounders using a hard sample mining scheme has proven effective. Instead of adjacent slices, stacks of axial images are used to distinguish nodules from blood vessels, providing a reliable method for early and accurate nodule detection. New versions of YOLO continue to be proposed. Due to its widespread use and robust detection results, YOLO approaches remain the preferred choice, making them my preferred solution.

CNNs have also shown significant efficacy in reducing FP in nodule detection, diminishing the need for extensive domain-specific knowledge (Katz et al., 2020). While it is possible to use features from pre-trained models, the best results are often achieved by training a network specifically tailored to the task from scratch (Setio et al., 2016). Networks that integrate 3D context by analyzing multiple 2D slices of cubic patches centered on candidate nodules have demonstrated superior performance. This method ensures comprehensive visibility of nodules within a designated patch size, with each slice being processed by specialized CNN blocks. To prevent overfitting, techniques such as dropout and data augmentation are employed, with hyperparameters

being meticulously tuned to account for the limitations of smaller datasets. Additionally, employing multi-scale analysis, where 3D patches are examined at different scales by adjusting the size and stride of pooling layers, has proven to be highly effective (Dou et al., 2016). The outputs from these networks are then aggregated to produce a final candidate classification, distinguishing between nodule and non-nodule cases, significantly improving detection performance.

Although the integration of transformers in lung nodule detection is still emerging, their potential is significant. Transformers, as demonstrated by Mkindu et al. (2023), are applied to nodule detection using attention mechanisms to selectively focus on the most relevant parts of the input data, thus enhancing feature extraction and handling long-range dependencies more effectively than traditional models. The shifted window-based vision transformer encoder in this architecture divides the input data into smaller, manageable sections, which are processed independently and then combined. This approach allows for more efficient and accurate processing of lung nodule features. By leveraging these attention mechanisms, transformers can better identify subtle patterns and variations in medical images, leading to more precise nodule detection and marking a promising advancement in the field.

3.3 Lung Nodule Segmentation

Segmentation involves identifying the voxels within a Region of Interest (ROI) that belong to a lung nodule (Figure 3.4), typically following the detection stage. Complete segmentation offers a more precise representation of a nodule's size and shape. When automated, segmentation can enhance diagnostic insights without imposing significant time constraints, which has motivated the development of various methodologies discussed in this section. However, in clinical practice, full 3D nodule segmentation is uncommon due to its time-consuming nature. Radiologists often rely on measuring the largest axial diameter to estimate nodule size, but this approach can lead to inaccuracies due to geometric 3D nodule complexity. As a result, follow-up criteria are designed to accommodate either 3D volume or axial diameter measurements.

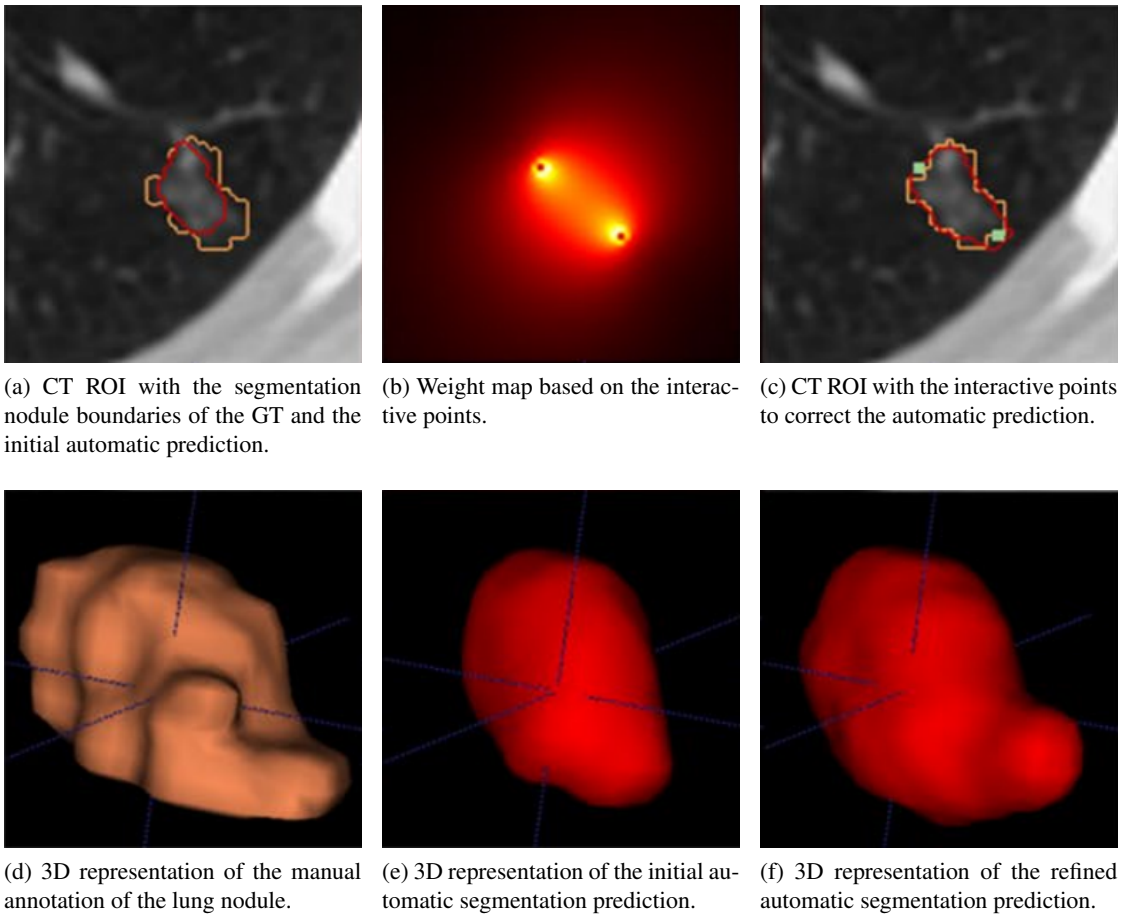
Table 3.3 lists some of the relevant segmentation, characterizing the approach (2D, 2.5D or 3D, with or without DL), identifying the dataset used, and showing the obtained IoU and Dice measures. It is important to recognize that the definition of the segmentation GT varies among studies, which often limits comparisons and lacks clarity on whether metrics are assessed in 2D or 3D. Although manual segmentation typically occurs on axial planes, the scientific community often prefers 3D segmentation for its comprehensive view. However, 2D and 2.5D methods, which use multiple CT slices to provide 3D context, have been proposed to reduce the complexity and computational demands of full 3D segmentation.

Non-DL approaches to lung nodule segmentation leverage the natural characteristics of solid nodules, which typically exhibit high contrast with lung parenchyma and a spherical shape. Traditional image processing methods such as thresholding and morphological operations (Mukhopadhyay, 2016) have been widely applied. Other techniques include region-growing (Kalpathy-Cramer et al., 2016; Kubota et al., 2011; Lassen et al., 2015; Moltz et al., 2009; Zhao et al., 2015),

Table 3.3: Performance of selected automatic lung nodule segmentation algorithms.

Author	Method	Dataset	IoU/Dice (No. of nodules)	Comments
Ye et al. (2010)	3D	Private	NA/0.85 (181)	
Tan et al. (2013)	3D	Private; LIDC-IDRI	0.65/NA (23)	
Messay et al. (2015)	3D	LIDC-IDRI	0.74/NA (66)	AL ≥ 3
Zhao et al. (2015)	2D	Private	0.89/NA (281)	
Gonçalves et al. (2016)	3D	LIDC-IDRI	0.65/NA (569)	
Feng et al. (2017)	2D-DL	LIDC-IDRI	NA/0.76 (1390)	
Wang et al. (2017)	2.5D-DL	LIDC-IDRI	0.71/0.82 (493)	
Aresta et al. (2018)	2D-DL	LIDC-IDRI	0.63/0.79 (NA)	
Jung et al. (2018)	3D	LIDC-IDRI	0.64/0.78 (18)	
Tong et al. (2018)	2D-DL	LIDC-IDRI	NA/0.74 (1245)	
Wang et al. (2018)	3D-DL	LIDC-IDRI	NA/0.95 (NA)	
Wu et al. (2018)	3D-DL	LIDC-IDRI	NA/0.74 (1404)	
Shakibapour et al. (2019)	2D	LIDC-IDRI	0.70/NA (705)	AL ≥ 3
Aresta et al. (2019)	3D-DL	LIDC-IDRI	0.48/NA (2284)	
Huang et al. (2019b)	2D-DL	LIDC-IDRI	NA/0.79 (1186)	
Atwal and Phoulady (2020)	3D-DL	LNDb	0.40/NA (127)	
Cao et al. (2020)	2.5D-DL	LIDC-IDRI	NA/0.83 (986)	
Dong et al. (2020)	2.5D-DL	LIDC-IDRI	NA/0.93 (874)	
Galdran and Bouchachia (2020)	3D-DL	LNDb	0.57/NA (127)	
Rassadin (2020)	3D-DL	LNDb	0.51/NA (127)	
Rocha et al. (2020)	2D; 2D-DL	LIDC-IDRI	NA/0.67 (2653); NA/0.83 (531)	
Shi et al. (2020)	3D-DL	LIDC-IDRI	NA/0.85 (1154)	
Tian et al. (2020)	3D-DL	LNDb	0.57/NA (127)	
Usman et al. (2020)	2.5D-DL	LIDC-IDRI	NA/0.88 (893)	
Wu et al. (2020)	3D-DL	LIDC-IDRI	NA/0.80 (200)	
Hou et al. (2023)	2D-DL	Private	0.65/0.78 (NA)	

Label: NA - Not Available.



(a) CT ROI with the segmentation nodule boundaries of the GT and the initial automatic prediction.

(b) Weight map based on the interactive points.

(c) CT ROI with the interactive points to correct the automatic prediction.

(d) 3D representation of the manual annotation of the lung nodule.

(e) 3D representation of the initial automatic segmentation prediction.

(f) 3D representation of the refined automatic segmentation prediction.

Figure 3.4: Manual and automatic nodule segmentation examples, with ■ representing the manual annotation, and ■ representing the automatic prediction. Adapted from Aresta et al. (2019).

deformable models, level-set methods (Awad et al., 2012; Farag et al., 2013; Jung et al., 2018), and graph-cut algorithms (Ye et al., 2010). Additionally, voxel-wise segmentation has been explored through intensity (Tan et al., 2013) and shape-related feature extraction, such as those from Hessian matrices (Gonçalves et al., 2016), and training classifiers like SVM or NNs (Messay et al., 2015). The handcrafted nature of these techniques makes them challenging to extend to Ground-Glass Opacity (GGO) and part-solid nodules, which have diffuse textures, irregular shapes, and low contrast with the surrounding tissue.

In recent years, DL techniques have been at the forefront of advancements in lung nodule segmentation. For instance, Wang et al. (2017) and Huang et al. (2019b) employed 2D CNN architectures to achieve pixel-by-pixel predictions for nodule segmentation. In Wang's approach, variations in input image scales were introduced to handle nodules of different sizes. Conversely, Huang's method concatenated outputs from convolutional layers throughout the network in the final layer to enhance segmentation accuracy. Notably, the U-Net architecture, available in both 2D and 3D forms (Çiçek et al., 2016), remains a popular choice due to its ability to capture both local and global contexts, with various adaptations implemented by numerous researchers (Aresta

et al., 2018; Tong et al., 2018; Rocha et al., 2020; Cao et al., 2020; Usman et al., 2020; Dong et al., 2020; Wu et al., 2020; Kaluva et al., 2020; Galdran and Bouchachia, 2020; Atwal and Phoulady, 2020; Hou et al., 2023). Researchers have also explored multi-task strategies by using features from the segmentation network's bottleneck to predict both nodule characteristics and the segmentation output (Wu et al., 2018; Tian et al., 2020; Rassadin, 2020).

Given the significant demand for annotated data to train these models, there is a growing interest in weakly supervised or unsupervised methods (Shakibapour et al., 2019), as well as the use of synthetic data. For example, Feng et al. (2017) used slice-level labels to indicate nodule presence rather than voxel-wise annotations. Approaches like active self-paced learning, which iteratively label unannotated data for subsequent training cycles, have also been considered (Wang et al., 2018). Furthermore, Shi et al. (2020) proposed generating synthetic data using a Generative Adversarial Network (GAN) to create augmented samples from real semantic labels with random styles, which are subsequently used for training a 3D U-Net.

Despite the high accuracy of DL methods, they often face criticism in medical applications due to lack of interpretability and deterministic outputs, which fail to account for inter-observer variability and limit the ability of medical professionals to interact with and adjust system decisions. To address these challenges, Aresta et al. (2019) introduced the iW-Net architecture, allowing experts to make minimal-interaction corrections to automatic nodule segmentations. This model consists of two 3D encoder-decoder networks resembling U-Nets. The initial block generates a segmentation proposal for user evaluation, where corrections are made by marking two points along the largest diameter of the nodule to create a weight map. This map, combined with the initial segmentation, is employed in the second block to produce a refined segmentation. In Figure 3.4, the correction points, resulting weight map, initial segmentation, and refined segmentation are shown to illustrate the output of the iW-Net. Simulations showed significant performance improvements, especially for GGO and small nodules.

3.4 Lung Nodule Characterization

Lung nodule characterization refers to the detailed description of nodule characteristics beyond those obtained in the detection and segmentation steps, *i.e.*, nodule location and size. As described in subsection 2.2.2, a wide range of characteristics for lung nodules is typically used. The development of approaches for the prediction of these characteristics either because they serve as a basis for deciding the timing of follow-up or because they provide insights into the malignancy of the nodules. Table 3.4 summarizes the nodule characterization methods discussed in this subsection, mentioning the form of discretization used in each.

Most approaches rely on the extraction of a ROI around the nodule, and 2D (Ferreira et al., 2018; Kawathekar and Areeckal, 2022), 2.5D (Tu et al., 2017; Galdran and Bouchachia, 2020; Marques et al., 2021), and 3D approaches have been proposed (Zinovev et al., 2011; Hussein et al., 2017; Dai et al., 2018; Wu et al., 2018; Shen et al., 2019; Kaluva et al., 2020; Tian et al., 2020), but

Table 3.4: Automatic lung nodule characterization algorithms.

Author	Method	Dataset (No. of nodules)	Nodule Features								
			Cal	ISt	Mar	Lob	Sph	Spi	Sub	Tex	Others
Zinovev et al. (2011)	2D	LIDC-IDRI (914)	2C	-	2C	2C	2C	2C	2C	2C	Mal
Niehaus et al. (2015)	2D	LIDC-IDRI (2660)	-	-	-	-	-	2C	-	-	-
Chen et al. (2016)	2D-DL	LIDC-IDRI (2400)	6R	4R	5R	5R	5R	5R	5R	5R	Mal
Hussein et al. (2017)	3D-DL	LIDC-IDRI (1144)	2C	-	2C	2C	2C	2C	-	2C	Mal
Shen et al. (2017)	2D-DL	LIDC-IDRI (756)	-	-	2C	-	-	-	2C	-	Mal
Tu et al. (2017)	2D-DL; 2.5D-DL	LIDC-IDRI (570)	-	-	-	-	-	-	-	3C	-
Dai et al. (2018)	3D-DL	LIDC-IDRI (1011)	6C	4C	5C	5C	5C	5C	5C	5C	Mal
Ferreira et al. (2018)	2.5D-DL	LIDC-IDRI (907)	-	-	-	-	-	-	-	3C	-
Wu et al. (2018)	3D-DL	LIDC-IDRI (1404)	6C	4C	5C	5C	5C	5C	5C	5C	Mal; Seg
Shen et al. (2019)	3D-DL	LIDC-IDRI (NA)	2C	-	2C	-	2C	-	2C	2C	Mal
Wang et al. (2020)	2D-DL	LIDC-IDRI (2632)	-	-	5C	4C	4C	4C	5C	5C	Mal
Galdran and Bouchachia (2020)	2.5D-DL	LNDb (178)	-	-	-	-	-	-	-	3C	-
Kaluva et al. (2020)	3D-DL	LNDb (178)	-	-	-	-	-	-	-	3C	-
Liu et al. (2019)	2D-DL	LIDC-IDRI (1000)	6R	4R	5R	5R	5R	5R	5R	5R	Mal
Tian et al. (2020)	3D-DL	LNDb (178)	-	-	-	-	-	-	-	3C	Seg
Qiu et al. (2020)	2D-DL	LIDC-IDRI (2084)	-	-	-	-	-	2C	-	-	-
Marques et al. (2021)	2.5D-DL	LIDC-IDRI (2668); LNDb (77)	2C	-	2C	2C	2C	2C	2C	3C	-
Kawathekar and Areeckal (2022)	2D	LNDb (NA)	-	-	-	-	-	-	-	3C	-
Zhang et al. (2022)	2D	Private (1015)	-	-	-	-	-	2C	-	-	-
Momoki et al. (2024)	3D-DL	Private (66)	-	-	2C	2C	2C	2C	-	3C	-

Nodule Features: C for classification and R for regression. NA - Not Available. Cal: Calcification; ISt: Internal Structure; Mal: Malignancy; Mar: Margin; Lob: Lobulation; Seg: Segmentation; Sph: sphericity; Spi: Spiculation; Sub: Subtlety; Tex: Texture.

this seems to have little impact on performance. This may be related to the fact that radiologists mainly base their classification on the axial slice and occasionally resort to other planes.

Fully dedicated methods for nodule characterization are scarce, and most approaches are designed in multi-task scenarios. Particularly when combined with nodule segmentation (Wu et al., 2018; Tian et al., 2020), the idea is to use the downsampling part of the network, which inherently performs feature extraction, to also predict nodule characteristics.

Following the LIDC-IDRI annotation proposals, most nodule characteristics are annotated on a 1–5 scale (with calcification and internal structure annotated on 1–6 and 1–4 scales, respectively). Depending on the number of classes considered, both classification (Zinovev et al., 2011; Niehaus et al., 2015; Hussein et al., 2017; Shen et al., 2017, 2019; Marques et al., 2021; Kawathekar and Areeckal, 2022) and regression methods (Chen et al., 2016; Liu et al., 2019) can be found. Specific methods focused on a single characteristic are primarily found for texture (Tu et al., 2017; Ferreira et al., 2018; Galdran and Bouchachia, 2020; Kaluva et al., 2020; Tian et al., 2020; Kawathekar and Areeckal, 2022) and spiculation (Niehaus et al., 2015; Qiu et al., 2020). Internal structure and calcification are, by nature, not ordinal scales, and their prediction through regression is not recommended.

Non-DL approaches use handcrafted features, exploring the differences in intensity across different nodule classes. It is common the use of nodule intensity features (Zinovev et al., 2011; Kawathekar and Areeckal, 2022), textural (Zinovev et al., 2011; Kawathekar and Areeckal, 2022) and shape features (Niehaus et al., 2015). These features are used as input of traditional classifiers such as decision trees (Zinovev et al., 2011) and K-Nearest Neighbors (KNN) (Kawathekar and Areeckal, 2022) for texture characterization tasks. Nevertheless, recent approaches tend to favor DL architectures. Most works propose custom CNN architectures (Tian et al., 2020; Marques et al., 2021), though other classification networks such as DenseNet (Dai et al., 2018; Wang et al., 2020) or ResNet (Liu et al., 2019; Kaluva et al., 2020; Galdran and Bouchachia, 2020) are also employed, which have the advantage of pretrained weights being easily available compared to custom CNNs.

Given the limited data available, artificial generation of nodule samples based on a generative adversarial network has been proposed (Wang et al., 2019; Zhang et al., 2022; Momoki et al., 2024), where oversampling of minority classes was performed to balance the data during training.

Stand-alone feature characterization can be a suitable tool for clinical use, particularly for nodule features used in clinical guidelines such as nodule texture, spiculation, and calcification. While the first two have received some attention in the literature, calcification is notoriously absent. This might be due to the fact that binary classification of calcification can be easily obtained through thresholding (typically using a threshold of 200 HU). While there could be clinical value in different calcification patterns, these do not feature prominently in clinical guidelines.

3.5 Discussion

This chapter presents some of the relevant research on CAD systems for CT follow-up in lung cancer. Lung cancer remains the deadliest form of cancer, primarily due to its late diagnosis. Consequently, it is crucial to ensure that appropriate measures are taken during follow-up when findings are reported, during patient treatment or monitoring. Although much research has focused on individual tasks such as lung nodule detection, segmentation, and characterization, the ultimate goal of patient follow-up has received comparatively less attention. This may be because follow-up procedures often rely on pre-established frameworks. Nonetheless, it is important to analyze these proposals from a holistic perspective and evaluate their effectiveness. Such evaluation might become more feasible as the state of the art in nodule detection, segmentation, and characterization becomes saturated and as relevant datasets for this topic become available. Additionally, challenges targeting patient-level tasks, such as the LNDb, are becoming more prevalent, which could further facilitate this shift.

The radiologists errors in nodule detection underscore the potential benefits of current CAD systems. These systems can reduce the number of radiologists needed to confirm diagnoses by quickly assessing full CT scans with high sensitivity, allowing radiologists to focus on recognition tasks, such as distinguishing nodules from FPs. In terms of segmentation, CAD systems can provide initial segmentations allowing future refinements by the radiologists, if needed. For instance, a semi-automatic approach was early proposed by LIDC-IDRI (Armato III et al., 2011), and recent interactive 3D segmentation tools, such as those proposed by Aresta et al. (2019), can further enhance workflow and reduce the variability among radiologists. Furthermore, CAD systems could in a first phase characterize the nodules, followed by the review of the characterization by the radiologists, particularly for more complex tasks as the determination of the nodule type such as GGO, juxtapleural and spiculated nodules.

The human aspect of the process must be considered by developing systems that integrate radiologists into the loop, working synergistically with their capabilities and addressing their limitations. Radiologists' active involvement in this process is crucial for a two-way learning experience: firstly, to tailor algorithm to radiologists' needs, and secondly, to familiarize radiologists with the potential benefits of these solutions. Understanding the basics of machine learning and DL systems could be important for the successful radiologist adoption of automated systems Research into integrating these tools into radiologists' workflows is also promising. For example, combining eye-tracking tools with DL-based lung nodule detection approaches can enhance synergy between automated methods and radiologists, as demonstrated in Aresta et al. (2020) and Khosravan et al. (2019). Additionally, these approaches could serve as educational tools for medical students and radiology interns.

To ensure that CAD systems for follow-up are effective, research should align with the approved follow-up guidelines. This alignment encourages a shift toward closer integration with clinical practice, simplifying the adoption of CAD systems both practically and in terms of radiologist acceptance. Ensuring that system decisions are based on validated clinical evidence is

crucial. Moreover, incorporating patient data, such as age, smoking history, and medical history, into CAD systems for lung cancer follow-up could represent an additional step toward this goal. Guidelines like British Thoracic Society (BTS) have already started emphasizing the importance of such integration.

Moreover, for end-to-end patient-level decision systems, it is crucial that CAD systems can effectively communicate the rationale behind their decisions. Without this transparency, radiologists' trust in these algorithms may quickly erode, especially when the algorithm fails in seemingly straightforward cases, such as missing a large nodule. Such failures can significantly hinder the adoption of CAD systems, regardless of their overall performance. Therefore, researchers should prioritize incorporating explainability and uncertainty mechanisms into their algorithms, as these factors are likely essential for successful integration into clinical practice. Comprehensive validation of these algorithms is equally imperative, ensuring they fulfill their intended purpose and preventing catastrophic failures upon deployment. This validation must involve data from diverse sources, including multiple centers, equipment types, annotators, nodule characteristics, and populations, particularly since DL solutions are susceptible to dataset bias. Additionally, performance reporting on nodule detection, segmentation, and characterization should extend beyond a singular value, demonstrating robustness across various dimensions such as nodule size, texture, location, and other relevant factors.

Finally, the analysis of automatic methods for the stages of lung cancer follow-up outlined in this chapter, including detection, segmentation, and characterization of lung nodules, can inspire methodologies for patient follow-up. It is well known that characterization methods are particularly valuable when follow-up classification is intended. Segmentation techniques can provide critical insights into nodule size prediction, while detection methods are essential for the reassessment of nodules in CT scans. As previously discussed, there are clear limitations in annotation quality, representativeness of diverse data sources, and validation across different data types. Therefore, the following chapters are aimed at integrating and enhancing the methodological contributions presented here, while addressing the identified challenges. Altogether, these efforts strive to develop effective strategies to tackle the complexities of lung cancer follow-up.

Chapter 4

Automatic Follow-up Recommendation using Lung-RADS Guidelines

This chapter addresses two contributions of this thesis: (**I. (a)**) an end-to-end approach, and (**I. (b)**) a feature-based approach, both focusing on automatic methods for follow-up recommendations according to the Lung-RADS guidelines. The first contribution[2] takes as input Computed Tomography (CT) scan regions centered on nodules and automatically outputs follow-up recommendations. This method incorporates annotation uncertainty parameters for dealing with multiple experts divergent opinions. In the second[3], the feature-based approach was adopted to mimic the medical doctors when establishing the patient follow-up, allowing a comparisons between the two approaches. This comparison sheds light on the balance of advantages and disadvantages between end-to-end and feature-based methods, particularly regarding performance and interpretability.

The two contributions are based on two conference papers and involved collaboration with radiologists from São João Hospital, Porto. The author of this thesis is the first author of both studies and was the primary contributor to their conception, the development of the methods, and the assessment of the results.

These two contributions are integrated into this chapter, organized as follows: introduction (4.1), related work (4.2), materials (4.3), methods (4.4), results (4.5), and concluding remarks (4.6).

[2] **Carlos A Ferreira**, Guilherme Aresta, João Pedrosa, João Rebelo, Eduardo Negrão, António Cunha, Isabel Ramos, and Aurélio Campilho. Classification of lung nodules in CT volumes using the Lung-RADS™ guidelines with uncertainty parameterization. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 791–794. IEEE, 2020. doi: 10.1109/ISBI45749.2020.9098586.

[3] **Carlos A Ferreira**, Isabel Ramos, Miguel Coimbra, and Aurélio Campilho. A comparative study of feature-based and end-to-end approaches for lung nodule classification in CT volumes to Lung-RADS follow-up recommendation. In *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)*. IEEE, 2024. doi: 10.1109/MELECON56669.2024.10608773.

4.1 Introduction

Identifying pulmonary nodules is a crucial step for lung cancer diagnosis. In many cases, invasive interventions, such as biopsies or lung resections, may not be justified. Many nodules are benign, showing minimal growth and limited clinical significance. Removing a pulmonary nodule is not straightforward and can have significant consequences, including impacts on respiratory function and overall health. As a result, a systematic follow-up approach is often preferred to avoid unnecessary procedures and associated risks. The use of standardized guidelines such as Lung-RADS (American College of Radiology, 2019), Fleischner Society recommendations (MacMahon et al., 2017), and British Thoracic Society (BTS) (Callister et al., 2015) aids in categorizing nodules based on risk factors, guiding medical doctors in choosing the most appropriate management strategy.

For pulmonary nodules, guidelines like Lung-RADS provide a structured framework for assessing the likelihood of malignancy and determining the need for further investigation or treatment, helping medical doctors balance the risks and benefits of various management options. The Fleischner Society guidelines take a more holistic approach by evaluating the patient's overall condition, while the BTS guidelines emphasize longitudinal assessment points. This is why there is significant interest in guidelines like Lung-RADS, which focus specifically on the nodules themselves, with recommendations based on the most concerning nodule.

Lung-RADS considers factors such as location, calcification, texture, size, and the presence of solid cores in nodules to provide follow-up recommendations. These recommendations include: 1 and 2) no follow-up/CT at 12 months; 3) CT at 6 months; 4A) CT at 3 months; or 4B) CT with contrast or biopsy. These five recommendations may also be further stratified, serving as a referral pathway or indicating potential malignancy. This referral system is typically employed following an initial screening, allowing clinicians to decide whether the patient should proceed to follow-up imaging or treatment based on the detected nodule characteristics, or if no further action is required. Table 4.1 presents the correspondences between these Lung-RADS follow-up management recommendations, the characteristics of the nodules, and the population prevalence percentage estimated from U.S. clinical trials.

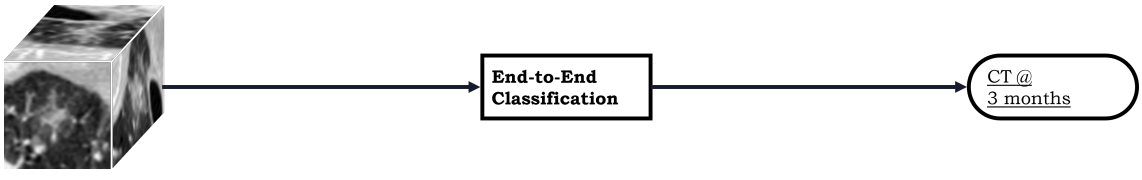
The integration of Artificial Intelligence (AI) as a procedural and decision-making support system can revolutionize lung cancer follow-up management. Lung-RADS is structured around a decision matrix that guides follow-up actions based on nodule characteristics, and AI can enhance this process by assisting at each stage or providing direct decision outputs. Feature-based methods apply AI to segment and characterize nodules, facilitating automated categorization for follow-up decisions, with each feature explicitly predicted by the models, namely nodule calcification, nodule texture, nodule diameter, and the solid diameter of part-solid nodules. In contrast, end-to-end methods employ AI to directly analyze nodule images and provide follow-up recommendations without breaking down the process into individual feature predictions, thereby focusing solely on the final decision output. These approaches streamline the follow-up process, bringing substantial efficiencies and showcasing the potential to accelerate patient care while significantly

Table 4.1: Correspondence of Lung-RADS classes with nodule characteristics and estimated population prevalence.

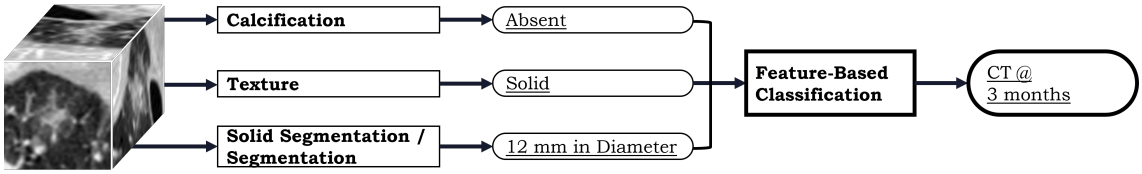
	Nodule Characteristics	*Prevalence*
Negative — 1	Calcified	39%
Benign — 2	2.1) Solid, Diameter $<$ 6 mm 2.2) Part-Solid, Diameter $<$ 6 mm 2.3) Ground-Glass Opacity (GGO), Diameter $<$ 30 mm	45%
Probably Benign — 3	3.1) Solid, 6 mm $\leq$ Diameter $<$ 8 mm 3.2) Part-Solid, Solid Diameter $<$ 6 mm 3.3) GGO, Diameter $\geq$ 30 mm	9%
Suspicious — 4A	4A.1) Solid, 8 mm $\leq$ Diameter $<$ 15 mm 4A.2) Part-Solid, 6 mm $\leq$ Solid Diameter $<$ 8 mm	4%
Very Suspicious — 4B	4B.1) Solid, Diameter $\geq$ 15 mm 4B.2) Part-Solid, Solid Diameter $\geq$ 8 mm	2%

reducing the radiologists' workload. Importantly, despite negative findings on lung cancer screening, nearly 18% of patients still undergo unnecessary follow-up procedures (Tailor et al., 2023), underscoring the urgent need for enhanced decision support. Recent studies demonstrate AI's ability to reduce CT image analysis time by approximately 30%, highlighting its critical role in optimizing patient care and minimizing inefficiencies in the follow-up process (Gang et al., 2021).

Building upon the identified opportunities for contribution, this thesis presents two Computer-Aided Diagnosis (CAD) proposals for classifying lung nodules into one of the five Lung-RADS recommendations. These proposals are depicted in Figure 4.1 and include both a feature-based and an end-to-end approach. Both methods require a CT cuboid centered around the nodule's center of mass as input.



(a) End-to-end approach.



(b) Feature-based approach.

Figure 4.1: Schematic representation of the functionality of the different methods used in this study.

In end-to-end classification, the class is determined solely based on image information. In contrast, for feature-based classification, each feature is individually assessed before integrating all features to derive the follow-up class, as outlined in Table 4.1. These features include the classification of calcification presence, texture (GGO, part-solid, and solid), and segmentations of both the nodule and, if applicable, the solid core of part-solid nodules.

4.2 Related Work

In the literature, approaches for follow-up predictions according to Lung-RADS often focus on nodule-wise analysis. However, full patient-level analyses are frequently not provided. This means that having a follow-up recommendation for an individual nodule is insufficient if it is not the most concerning nodule for the patient. It is important to determine the accuracy of classifying the most concerning nodules. This serves as a warning that a reported performance may obscure its behavior in a real clinical context.

The challenge of accurately categorizing nodules for follow-up is explored through a methodology proposed by Chang et al. (2022). Nodule segmentation is semi-automatic, with parameters such as location, margin, and texture determined by a Neural Network (NN) that processes features from radiomic filters applied to the CT image region containing the nodule. This approach enables the integration of follow-up recommendations, and the evaluation resulted in a 0.95 acceptance rate, as medical doctors reviewed the automatic predictions and confirmed their clinical relevance and adequacy to real cases. Lin et al. (2023) classified end-to-end Lung-RADS follow-up classes from CT images using an ensemble of radiomic features, including some derived from a deep learning extractor, and some from clinical data, achieving an accuracy of 0.80. In both studies, four Lung-RADS classes (2, 3, 4A, and 4B) were considered. Additionally, Nishio et al. (2018) trained a VGG-16 model pre-trained on ImageNet with transfer learning through fine-tuning on a dataset from Kyoto University Hospital, consisting of 1,240 patients, to classify nodules into three classes (2, 3 and 4A) and achieved an accuracy of 0.68. Finally, a previous study by the author of this thesis proposed an end-to-end classification of Lung-RADS follow-up recommendations into two classes for referral purposes (1 and 2 for the negative class; 3 and 4 for the positive class) (Ferreira et al., 2019) using a Wide Residual Network (WRN) (Zagoruyko and Komodakis, 2016), achieving an accuracy of 0.87. Among the referenced studies, only (Ferreira et al., 2019) uses the same dataset explored in this chapter, so caution is needed when comparing these results, as data conditions and class definitions vary.

It is clear that one of the main challenges with Lung-RADS is that it relies on data not currently available in public datasets, specifically the measurement of the solid diameter of part-solid nodules, which has likely hindered further work on this topic. Despite this limitation, Lung-RADS remains one of the most frequently referenced guidelines in the literature. To the best of the author's knowledge, automatic follow-up prediction algorithms for the British BTS guidelines have not yet been proposed. In contrast, for the Fleischner Society guidelines, the LNDb challenge was established to include nodule detection, segmentation, and texture characterization, with the aim of

performing patient-level follow-up recommendations according to Fleischner guidelines (Pedrosa et al., 2021). The author of this thesis was one of the organizers of this challenge. Most proposals emerging in this area follow the framework set by this challenge. Kaluva et al. (2020), in addition to the common framework, employed a rule-based approach to compensate for potential errors that could occur in different stages up to the follow-up recommendation according to Fleischner.

The presence of AI algorithms in clinical workflows for follow-up recommendations is not yet fully integrated. While AI systems that leverage Lung-RADS are capable of prioritizing high-risk nodules, their use in everyday clinical practice is still limited. Efficient access to these recommendations, ideally within existing information systems through straightforward interfaces, could enhance diagnostic processes and workflow efficiency. Both end-to-end and feature-based approaches offer potential for more personalized and timely patient management.

4.3 Material

This chapter uses the public Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset (Armato III et al., 2011), which includes 1018 CT thoracic scans from seven academic centers. The axial slices are 512×512 pixels with a resolution ranging from $[0.542,\ 0.750]$ mm/pixel. The number of slices varies from 100 to 600 with a thickness from 0.6 to 5 mm. To ensure uniformity, all scans were isotropically resampled to 0.7 mm/voxel with a range $[-1000,\ 400]$ Hounsfield Unit (HU).

Twelve radiologists participated in the study. Each case was characterized by four of these radiologists, who were asked to annotate abnormal structures of interest on a chest CT scan as follows: nodules with a diameter less than 3 mm, nodules with a diameter of at least 3 mm, and non-nodules with a diameter of at least 3 mm. Non-nodules are lesions, such as scars, that, although they have no direct cancerous manifestation, may be of interest for the diagnosis of lung cancer or other pathologies. The dataset provides annotations for nodules with a diameter of at least 3 mm through detection, segmentation, and characterization, while only centroids are provided for other structures. The number of annotations for nodules with a diameter of at least 3 mm was limited to six per CT per physician.

Initially, the radiologists made annotations individually, without knowing their colleagues' opinions. In a subsequent stage, they had the opportunity to review the annotations without the requirement of consensus, merely providing an additional perspective that might influence their own annotations. Each nodule with a diameter of at least 3 mm thus has up to four segmentations and scores for various characteristics, including nodule texture (ranging from 1 to 5) and calcification (ranging from 1 to 6), according to Table 2.1.

Each volume-of-interest is centered on the nodule's center of mass and spans $64 \times 64 \times 64$ voxels. The segmentation and characterizations are determined by averaging the annotations provided by the experts.

Based on the interpretation of the scores in the annotation of the LIDC-IDRI dataset, each nodule is assigned to a Lung-RADS class. For computational purposes, these classes are represented as $l \in \{\mathsf{LR} = 1,2,3,4,5\}$, where the class 4 corresponds to 4A and the class 5 corresponds to 4B. Then, the following considerations were made:

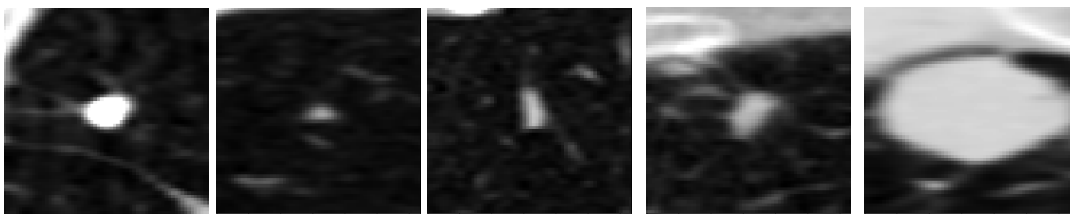
- calcified nodules exhibit scores within the range $[1,\ 5]$, with a score of 6 indicating an absence of calcification;
- textures within the range $]3.67,\ 5]$ were defined as solid nodules, those within $[1,\ 2.33[$ as GGOs, and part-solid otherwise;
- the diameter of the nodules is computed as the equivalent diameter of a sphere with the same volume as the Ground Truth (GT) masks, given the round shape of the segmentations.

Additionally, since solid core diameters of part-solid nodules were not provided, voxel-wise segmentations were annotated by consensus among 2 radiologists from Centro Hospitalar and Universitário do São João, Porto, Portugal.

The dataset has a total of 2,666 nodules, each annotated by at least one expert. Table 4.2 details the correspondences between LUNG-RADS classes and the percentage of nodules within the LIDC-IDRI dataset. Figure 4.2 shows the central axial view of a nodule from each one of the Lung-RADS classes. From the table and figure, it is possible to notice that: 1) the negative class is strongly dominated by calcified nodules with high HU; 2) there is a predominance of solid nodules in almost all classes; 3) larger nodule diameters correspond to greater malignancy; 4) there is a slight class imbalance.

Table 4.2: Relation between LUNG-RADS classes with the quantity of nodules in LIDC-IDRI.

					Total				Total				Total		
Lung-RADS Class	**1**	2.1	2.2	2.3	**2**	3.1	3.2	3.3	**3**	4A.1	4A.2	**4A**	4B.1	4B.2	**4B**
% in the Dataset	**14.2**	18.5	2.6	10.0	**31.1**	21.4	2.2	0	**23.6**	19.1	2.0	**21.0**	9.0	1.2	**10.2**



(1) Negative (2) Benign (3) Probably Benign (4A) Suspicious (4B) Very Suspicious

Figure 4.2: Examples of nodule middle axial slices of different Lung-RADS classes.

Lastly, Figure 4.3 presents a diagram to help illustrate some of the challenges associated with this dataset. The diagram shows the relationship between nodule agreement, which can involve up to four annotators depending on whether they agree that it is indeed a nodule, and the level of agreement, defined as the maximum Lung-RADS agreement, meaning the highest number of annotators assigning the same Lung-RADS class. Even when there is consensus that a finding is a

nodule, significant variations in the level of agreement can still be observed. This resulted in one, two, three, or four annotators per nodule in 29%, 19%, 18%, and 34% of cases, respectively.

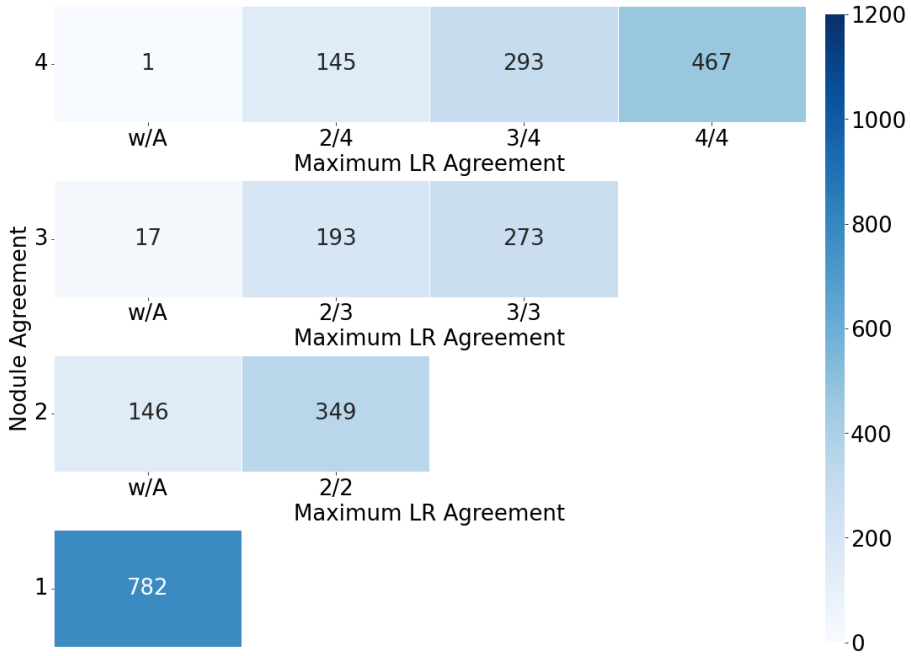


Figure 4.3: Diagram showing the relationship between the number of annotators (on the left) and the level of agreement regarding Lung-RADS classes in LIDC-IDRI (at the bottom), expressed as the fraction of maximum Lung-RADS agreement over the total number of annotators for each nodule. w/A denotes no agreement, and Lung-RADS is abbreviated as LR.

4.4 Methods

4.4.1 End-to-End Classification

The end-to-end method was designed from scratch, meaning that in addition to its conceptual development, the algorithm was also trained and tested adequately. For this purpose, two networks were used: a Wide Residual Network (WRN) (Zagoruyko and Komodakis, 2016) and a DenseNet (Huang et al., 2017). Both networks use "shortcut connections" to facilitate the back-propagation of errors, enhancing the classification performance when compared to more traditional Convolutional Neural Networks (CNNs).

In the WRN, the input of each layer block is added to the corresponding output, whereas in the DenseNet, each layer block concatenates inputs from all preceding layers, allowing their feature maps to be propagated through all subsequent layers. It is also noteworthy that, compared to the ResNet, the WRN is characterized by having a smaller depth and larger width. In this study, the WRN consists of 5 main blocks with 22 residual block repetitions. The DenseNet has 100 dense layers with a growth rate (number of filters added per dense block) of 4. These two architectures are represented in Figure 4.4 and Figure 4.5, respectively.

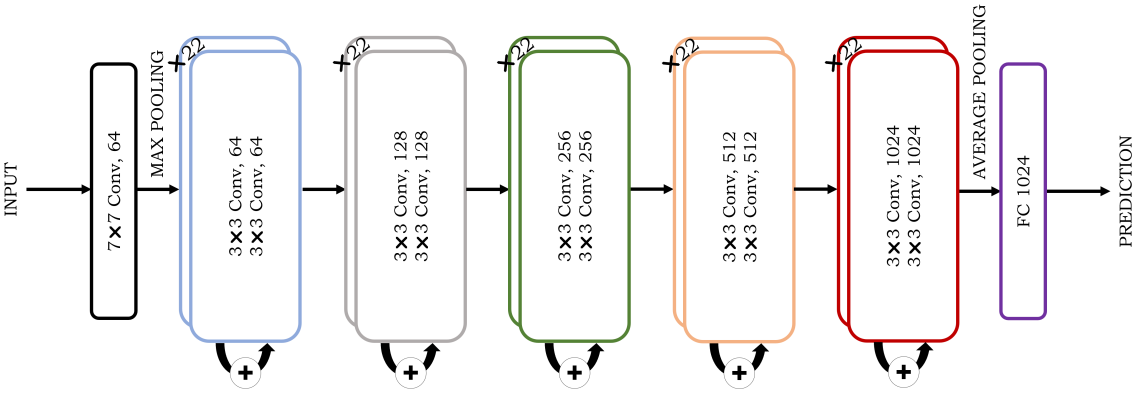


Figure 4.4: WRN Arquitecture. The operations are represented by rounded-corner rectangles. Convolutional layers (Conv) are composed of a sequence of operations: a convolution, a batch normalization, and a Rectified Linear Unit (ReLU). FC refers to a fully connected layer. The filter size is indicated before the layer type, and the number of filters or neurons is indicated after the comma.

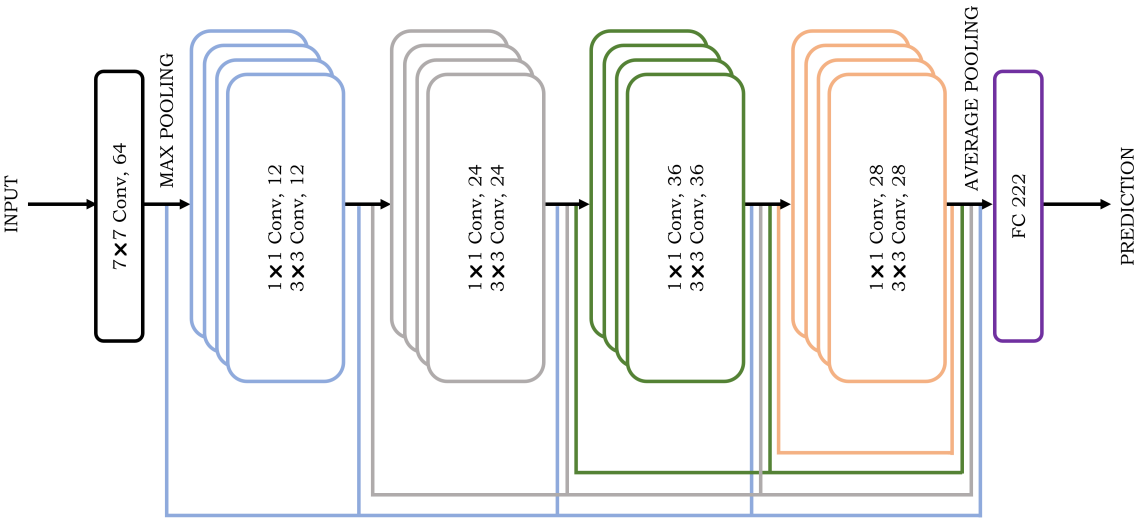


Figure 4.5: DenseNet Architecture. The operations are represented by rounded-corner rectangles. Convolutional layers (Conv) are composed of a sequence of operations: a convolution, a batch normalization, and a ReLU. FC refers to a fully connected layer. The filter size is indicated before the layer type, and the number of filters or neurons is indicated after the comma.

The models were trained using the three middle orthogonal planes from each nodule's volume of interest. Although this approach is referred to as 2.5D, it provides sufficient depth perception beyond a single 2D plane. Initial experiments revealed that this was enough to infer the characteristics of a nodule. Each of the planes is treated as an independent sample. Then, the final classification of a nodule is computed by averaging the prediction of its corresponding orthogonal planes.

The deep learning architectures were trained using early stopping of 25 epochs, batch size of 32 and data augmentation through vertical and horizontal flips and random rotations from 0° to 360°. The optimizer was the ADAM algorithm (Kingma and Ba, 2015) with learning rate of 10^{-4}.

As previously mentioned, the dataset contains significant annotation uncertainty. Each nodule's characterization carries an uncertainty due to annotation disagreements on whether an abnormality is a nodule and on the assignment of the Lung-RADS grade. Generically, samples with higher agreement should have a higher impact during training. For this purpose, each sample is assigned a weight α'_i, as defined in Equation (4.1):

$$\alpha'_i = \left(1 - \frac{\sigma_i}{\sigma_{M,i} + \varepsilon_i}\right) \frac{n_i}{4} \tag{4.1}$$

where n_i represents the number of annotators for nodule i, σ_i is the standard deviation of the corresponding set of grades $\mathbf{g}i$, $\sigma_{M,i}$ is the theoretical maximum value that σ_i can take based on the range of possible grades, and ε_i is a small positive value added to avoid division by zero. Some illustrative examples are presented in Table 4.3:

Table 4.3: Illustrative examples of the computation of the sample weights.

n_i	$\mathbf{g}_i$	σ_i	$\sigma_{M,i}$	α'_i
4	{5,5,5,5}	0	2	1
4	{1,3,3,5}	1.41	2	0.29
3	{3,3,3}	0	1.89	0.75
3	{2,3,4}	0.82	1.89	0.43
2	{4,4}	0	2	0.5
2	{1,5}	2	2	0
1	{2}	0	0	0.25

These weights are then normalized to ensure that the loss is on the same range when training with and without considering uncertainty:

$$\alpha_i = \alpha'_i \frac{N}{\sum_{k=1}^{N} \alpha'_k} \tag{4.2}$$

where N is the total number of samples. Finally, a second weight term is added to consider the class imbalance:

$$\beta_l = \frac{N}{|\mathsf{LR}| N_l} \tag{4.3}$$

where N_l is the number of samples belonging to grade l, and $|\mathsf{LR}|$ represents the total number of Lung-RADS grades. The loss $\mathscr{L}$ is the cross-entropy weighted by Equation (4.2) and (4.3):

$$\mathscr{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{l \in \mathsf{LR}} \alpha_i \beta_l y_{i,l} log(p_{i,l}) \tag{4.4}$$

where $y_{i,j}$ is the one-hot-encoded GT and $p_{i,g}$ is the inferred grade $l \in \{\mathsf{LR} = 1,2,3,4,5\}$, representing the probability computed by applying the softmax function to the outputs of the final layer of the network. The softmax function ensures that the outputs are normalized to form a probability distribution, with all values summing to 1. The softmax function is defined as follows:

$$f_{\text{softmax}}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{|\text{LR}|} e^{z_j}} \tag{4.5}$$

where z_i represents the i-th output of the network.

4.4.2 Feature-Based Classification

The Lung-RADS recommendation classes can be derived from various tasks, namely classifying calcification and texture, and segmenting the nodule and, if applicable, the solid core of part-solid nodules. For calcification classification, the approach of Marques et al. (2021) was followed, while texture characterization was based on Ferreira et al. (2018), and nodule segmentation on Aresta et al. (2019). Each of these tasks is detailed in the following sections:

Calcification Classification: The prediction of calcification is performed using the study by Marques et al. (2021), which proposes a method to automatically predict the characteristics and malignancy of lung nodules, by inputting three parallel axial CT slices around the nodule's center. Although the study was designed to be a multitask prediction approach for nodule characterization, one of the experiments involved a stand-alone comparison, solely focusing on obtaining a single characteristic, such as calcification, which is the approach used here. The model uses a CNN with three convolutional layers.

Since this is an imbalanced classification problem, class weights are used to aid in learning, and the final activation layer is softmax, with cross-entropy loss guiding the training. For the optimization of the models, the Nesterov Adam optimizer (Dozat, 2016) was used, with a learning rate of 0.002 and a batch size of 32. A patience parameter of 50 epochs was employed for the early stopping criterion. Augmentation was performed in the same manner as in the end-to-end method.

Texture Classification: The orthogonal medial slices of the nodule cuboid serve as input for texture prediction using a method proposed by the author of this thesis (Ferreira et al., 2018). This study is entirely focused on classifying the three texture classes (GGO, part-solid, and solid), with variations in model configurations and input data. The backbone of the model is a CNN with three convolutional layers.

To balance the classes, the underrepresented classes in the training set were replicated with small rotations (11.25°). The parameters of the network were optimized using the ADAM algorithm (Kingma and Ba, 2015), with a learning rate of 0.001. Additionally, a softmax final activation layer was used, and cross-entropy loss was applied.

Segmentation: The segmentation method used was the iW-Net proposed by Aresta et al. (2019), as shown in Figure 4.6. This method is based on two sequential 3D U-Nets (Ronneberger et al., 2015) consisting of 4 convolutional layers each. The first U-Net performs the automatic nodule segmentation, while the second one requires human interaction to draw points if deemed necessary

to refine the segmentation. Given the focus on fully automated prediction, only the first U-Net, which performs segmentation without the need for human intervention, was used.

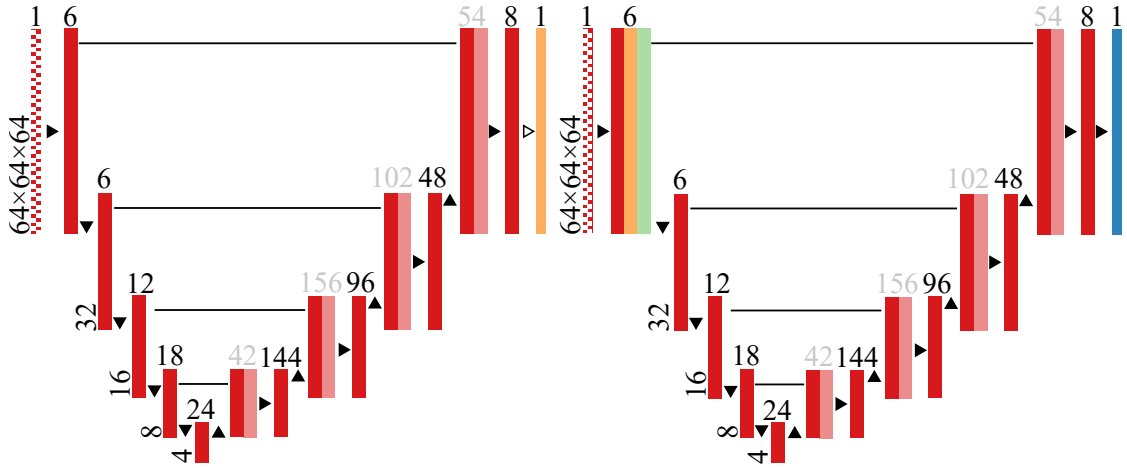


Figure 4.6: **iW-Net**: a network for guided segmentation of lung nodules, composed by a block responsible for predicting the initial segmentation and a second block for its correction. input image; intermediary feature maps; initial segmentation prediction; weight map computed from the user's input; corrected segmentation. $\blacktriangleright$ $3 \times 3 \times 3 \times N$ convolution, followed by batch normalization and ReLU activation (N is the number of feature maps, indicated on the top of each layer); $\blacktriangledown$ $3 \times 3 \times 3 \times N$ convolution with stride $2 \times 2 \times 2$, followed by batch normalization and ReLU activation; $\blacktriangle$ $2 \times 2 \times 2$ nearest neighbor up-sample; $\triangleright$ $3 \times 3 \times 3 \times N$ convolution with sigmoid activation. The feature map depth and width/height are shown above on the left side of each feature map, respectively. Image from Aresta et al. (2019).

The network was trained using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 and a batch size of 8 samples. An early stopping criterion was incorporated to halt training after 3 epochs without improvement. The loss function for the first U-Net architecture was defined as: $1 - \text{IoU}$, where IoU was previously defined in Equation (3.5). To enhance the diversity of the training data, the dataset was artificially augmented through random transformations, including rotations, translations, flips, and zooms.

When the nodule is predicted to be part-solid, the pixels in the segmentation prediction with a HU lower than -300 are removed to obtain the segmentation of the solid core. This threshold is often used as a reference for distinguishing between GGO and solid nodules, as in Scholten et al. (2015).

Both the segmentation of the nodule and the segmentation of the solid core of part-solid nodules are performed in 3D. To use the diameter in accordance with Table 4.1, the volume is considered as an ellipsoid, and the diameter is the equivalent diameter of a sphere.

4.4.3 Model Evaluation

For the evaluation of the various models, specific metrics were employed in the segmentation and classification tasks. For segmentation, the Dice was computed for the 3D cuboid encompassing the

lung nodule, and the Mean Absolute Difference (MAD) was calculated based on the differences in the equivalent diameters of the GT masks and the segmented predicted nodules. Equations (3.6) and (3.7) were previously presented for Dice and MAD, respectively. For classification, evaluating model performance requires a variety of metrics to assess different aspects of accuracy and class-specific performance. The accuracy has been introduced previously (Equation 3.1). In multi-class scenarios, the concept remains the same but is applied across all classes, considering the proportion of correctly classified instances for each class individually and averaging the results. To evaluate model performance, especially in imbalanced datasets (Sokolova and Lapalme, 2009), macro recall and macro precision are employed, adapting the concepts of recall (Equation 3.2) and precision (Equation 3.3) to multi-class scenarios. These metrics ensure that each class is considered equally, preventing larger classes from disproportionately influencing the evaluation. The relevant equations are:

$$\text{Accuracy} = \frac{\sum_{c=1}^{C} \text{TP}_c}{N} \tag{4.6}$$

$$\text{Macro Recall} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{4.7}$$

$$\text{Macro Precision} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \tag{4.8}$$

C represents the number of classes, N is the total number of observations, and N_c is the number of observations in class c. TP_c denotes the true positives for class c, while FN_c and FP_c are the false negatives and false positives for class c, respectively.

Furthermore, the dispersion of results across different classes is characterized by metrics that quantify the agreement between annotators and the automatic predictions. Two such metrics were used: Cohen's kappa ($\mathscr{K}$) (Spitzer et al., 1967) and the average agreement (Ag).

$$\mathscr{K} = \frac{P_o - P_e}{1 - P_e} \tag{4.9}$$

$$Ag = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{m=1}^{M} [y_{i,m} = \hat{y}_i] \tag{4.10}$$

P_o represents the observed proportion of agreement among annotators, while P_e denotes the expected proportion of agreement by chance, computed under the premise that annotator choices align with the observed frequencies of each class in the annotations. This simulates a scenario where assignments to classes are based on their observed distribution rather than random selection. For the Ag metric, M is the number of annotators, $y_{i,m}$ is the label provided by the mth annotator for the ith observation, and $\hat{y}_i$ is the automatic prediction for the ith observation. Ag measures the average agreement between the annotators' labels and the automatic predictions across all observations.

A quantitative analysis is primarily conducted at the nodule level. To assess the statistical significance of performance differences between two classification models for follow-up recommendation, the McNemar's test is employed. Additionally, a brief quantitative analysis of classification metrics is made to determine if the models are robust in identifying the most concerning nodules.

For qualitative evaluation, confusion matrices are used to identify where the main errors occur. The reasons for these errors are evaluated by analyzing task-specific errors in the feature-based case and using Grad-CAMs (Selvaraju et al., 2017) in the end-to-end case. These methods help to understand the decision-making process and the model's attention, respectively. Considerations regarding inference time and memory space are also taken into account.

All models were trained using 10-fold cross-validation. The evaluation process ensures that the fold distribution is preserved, meaning each instance is evaluated within its originally assigned test set, regardless of the different models presented in this chapter. The experiments were conducted using Keras-TensorFlow v2.1.6 on a desktop with an Intel(R) Core(TM) i7-5820K CPU @ 3.30 GHz, 32 GB RAM, and an 8 GB GPU.

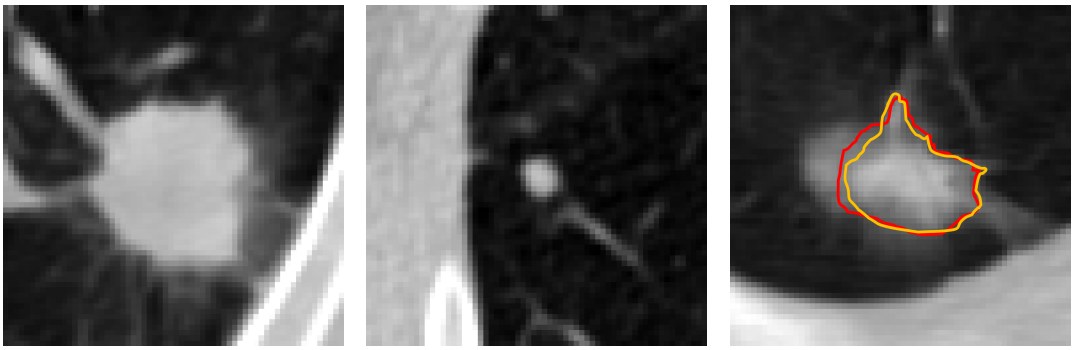
4.5 Results

4.5.1 End-to-End Classification

Figure 4.7 shows a series of results that illustrate the effectiveness of the methods described in this chapter. It presents various examples where the models performed well.

The end-to-end approach takes imaging data as input and directly returns the corresponding Lung-RADS grade, whose results are presented in this subsection. The Figure 4.7a depicts a correct classification of a large nodule for the end-to-end method as 4B, which corresponds to the highest level of follow-up surveillance. The method accurately predicted this grade, likely due to the significant portion of the Region of Interest (ROI) being occupied by the lung nodule. Subsequent Figures 4.7b-4.7f will be discussed later, where they correspond to feature-based methods addressed in the next subsection.

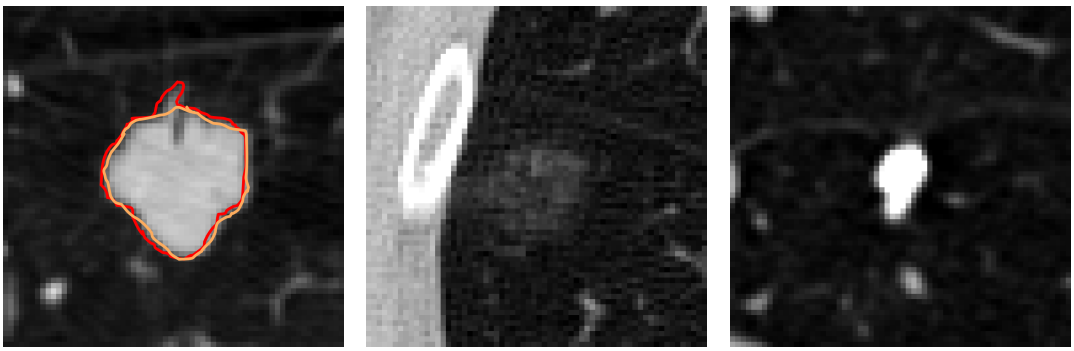
In this end-to-end method, annotation uncertainty was modeled to potentially enhance the model's performance. It is well-known that there is often a lack of consensus among radiologists regarding the classification of findings as nodules, and varying characterizations can lead to different Lung-RADS classifications. Figure 4.8 shows two plots of the cumulative accuracy *vs.* the annotators uncertainty weights for the WRN model and for the DenseNet model. Both plots exhibit a decrease in cumulative accuracy up to an uncertainty weight of 0.25, followed by an increase with minor fluctuations towards the end. This pattern suggests that applying uncertainty weights generally enhances performance. The improvements are more significant for reliably annotated nodules, indicating that uncertainty weighting helps prioritize these cases, ultimately resulting in a positive overall balance.



(a) [End-to-end approach] Lung-RADS class: 4B.

(b) [Feature-based approach] Lung-RADS class: 3.

(c) Core solid segmentation of a part-solid nodule.

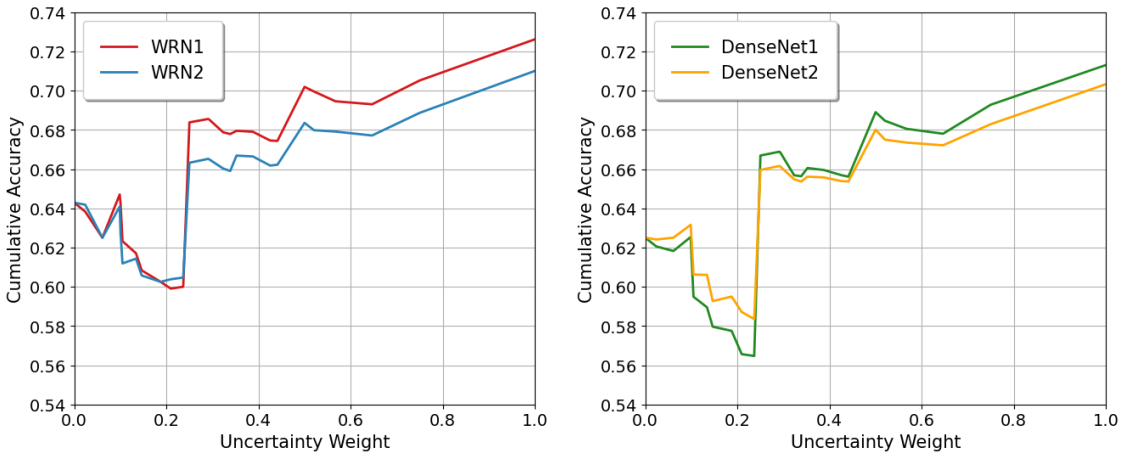


(d) Nodule segmentation.

(e) GGO nodule.

(f) Calcified nodule.

Figure 4.7: Examples of correct results from the automatic methods in this chapter, with ■ for manual annotation and ■ for automatic prediction in segmentation.



(a) **WRN1**: WRN with sample weights; **WRN2**: WRN without sample weights.

(b) **DenseNet1**: DenseNet with sample weights; **DenseNet2**: Dense Net without sample weights.

Figure 4.8: Cumulative accuracy in the different experiments as a function of the uncertainty weight.

Regarding performance, Table 4.4 presents the accuracy, macro recall, and macro precision to evaluate the model as a whole. Additionally, recall and precision are provided to assess performance for each of the Lung-RADS classes. The Lung-RADS class weights play an important role in enhancing the model's performance, even if the improvement is small in some cases. Secondly, although class imbalance control methods were used, performance across the different classes varies significantly. The best performances are observed in the "extreme" classes (1, 2, and 4B), while the worst are in the "central" classes (3 and 4A). Thirdly, WRN outperformed DenseNet, indicating that WRN's architecture is more effective for this classification task.

Table 4.4: Performance of end-to-end methods implemented to classify Lung-RADS Classes.

Network	**Weights**	**Recall**				
		1	2	3	4A	4B
WRN	✓	0.776 ± 0.081	0.764 ± 0.080	**0.645** ± 0.071	**0.693** ± 0.108	0.805 ± 0.082
	✗	**0.788** ± 0.069	0.702 ± 0.070	0.640 ± 0.104	0.690 ± 0.123	**0.836** ± 0.051
DenseNet	✓	0.785 ± 0.082	**0.831** ± 0.039	0.549 ± 0.121	0.643 ± 0.099	0.787 ± 0.084
	✗	0.774 ± 0.100	0.755 ± 0.065	0.619 ± 0.099	0.643 ± 0.065	0.820 ± 0.064

Network	**Weights**	**Precision**				
		1	2	3	4A	4B
WRN	✓	**0.910** ± 0.058	0.771 ± 0.061	**0.568** ± 0.076	**0.732** ± 0.074	**0.831** ± 0.091
	✗	0.870 ± 0.071	**0.789** ± 0.046	0.559 ± 0.094	0.700 ± 0.074	0.780 ± 0.120
DenseNet	✓	0.889 ± 0.054	0.724 ± 0.049	0.563 ± 0.074	0.724 ± 0.040	0.817 ± 0.097
	✗	0.859 ± 0.085	0.776 ± 0.059	0.561 ± 0.084	0.701 ± 0.061	0.723 ± 0.079

Network	**Weights**	**Macro Recall**	**Macro Precision**	**Accuracy**
WRN	✓	**0.737** ± 0.084	**0.762** ± 0.072	**0.727** ± 0.020
	✗	0.731 ± 0.083	0.740 ± 0.081	0.712 ± 0.027
DenseNet	✓	0.719 ± 0.085	0.743 ± 0.063	0.712 ± 0.015
	✗	0.722 ± 0.786	0.724 ± 0.074	0.706 ± 0.016

4.5.2 Feature-Based Classification

The feature-based method assigns the corresponding Lung-RADS grade (4.7b) by extracting the solid core of part-solid nodules (4.7c), the nodule's segmentation (4.7d), texture (4.7e), and calcification (4.7f) from imaging data. Figure 4.7 shows examples where the model performed well. These nodules are relatively isolated from surrounding anatomical structures, making them clear cases for radiologists, which likely led to proper feature extraction in the CNN and, consequently, accurate automatic predictions.

With respect to the feature-based method, the results not only include the overall performance of the method but also the individual tasks that contribute to determining the Lung-RADS class, given that this is a modular method. The results for characterizing the segmentation and characterization of calcification and texture are computed independently for the entire dataset, meaning

they do not come from the pipeline of the feature-based process when these methods were invoked. This approach provides insight into the performances a priori before proceeding to the fully composed method. Table 4.5 shows the results of these methods. The Dice for segmentation evaluation is competitive with the state of the art, demonstrating only moderate impacts on segmentation when considering the MAD. In terms of calcification, the classifications are nearly optimal as it is an easier challenge. Regarding texture, there is a clear deficit in accuracy for part-solid nodules.

Table 4.5: Performance of tasks and the feature-based method for classifying Lung-RADS classes.

	Macro Recall	**Macro Precision**	**Accuracy**
Lung-RADS	0.719 ± 0.026	0.734 ± 0.058	0.701 ± 0.026
Texture	0.684 ± 0.040	0.650 ± 0.054	0.862 ± 0.019
Calcification	0.866 ± 0.039	0.938 ± 0.031	0.954 ± 0.011

		Dice	**MAD**
Segmentation		0.688 ± 0.251	1.56 ± 0.14
Solid Segmentation (Part-Solid Nodules)		0.721 ± 0.112	2.22 ± 0.15

		1	2	3	4A	4B
Lung-RADS	**Recall**	0.742 ± 0.080	0.750 ± 0.033	0.464 ± 0.060	0.804 ± 0.056	0.834 ± 0.058
	Precision	0.917 ± 0.050	0.752 ± 0.036	0.556 ± 0.069	0.606 ± 0.044	0.840 ± 0.089

		GGO	Part-Solid	Solid
Texture	**Recall**	0.837 ± 0.075	0.301 ± 0.097	0.915 ± 0.018
	Precision	0.938 ± 0.010	0.301 ± 0.096	0.711 ± 0.057

A detailed analysis revealed that the errors in feature-based classification are induced by:

- 12.4% in the calcification prediction;
- 17.7% in the texture prediction;
- 47.5% in the mismatching of diameter ranges in GT;
- 2.8% in the mismatching of diameter ranges in solid cores of part-solid nodules;
- 19.6% where one of the characteristics is predicted differently, but the follow-up recommendation remains correct because, as seen previously, various combinations are possible.

4.5.3 Discussion

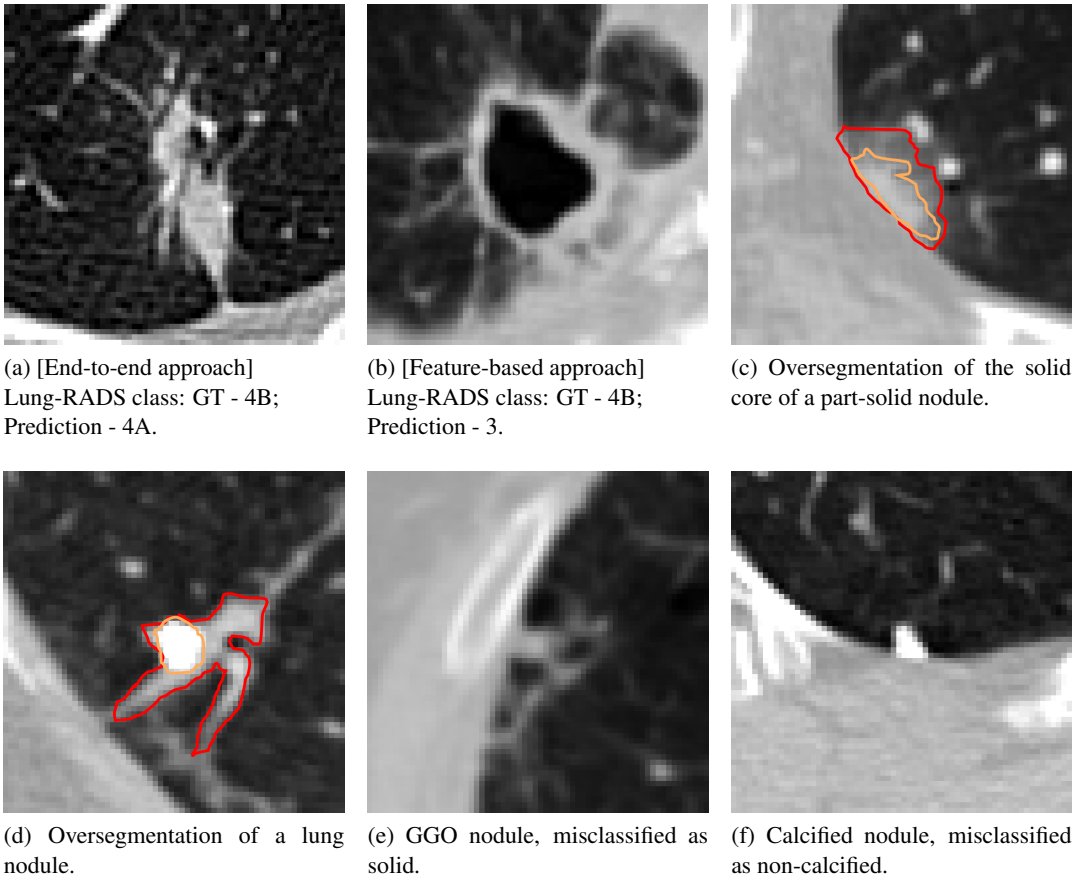
The considerations, design, and development of these feature-based and end-to-end methods differ significantly, resulting in neural architectures that do not follow a consistent pattern across approaches. Given their differences, it is naturally anticipated that their performance would also be affected, presenting a limitation. Nevertheless, the primary goal of this chapter was to compare two approaches which, despite having different specificities, provide insights into the practical applicability of such methods for recommending follow-up according to Lung-RADS.

The end-to-end method serves as an excellent framework for applying uncertainty weights, enabling direct classification and guiding the training. This method enables more comprehensive exploration of uncertainty weighting, which is less feasible with the feature-based approach due to its fragmented nature. These uncertainty weights generally allow for improved performance of

deep learning architectures, as observed by the fact that nodules with greater certainty have greater importance in training and vice versa, thus enhancing performance.

Continuing with the considerations regarding the end-to-end experiments, it is observed that WRN outperformed DenseNet. Typically, by taking the input from all previous convolutional layers, DenseNet tends to achieve better performance results and a smoother decision function with improved generalization performance. In this case, despite the differences in the architectures, WRN had more parameters to optimize, which may have led to deeper learning and better performance.

Figure 4.9 contains examples where the predictions are incorrect. Specifically, an example of incorrect prediction from the various end-to-end experiments is presented in Figure 4.9a. The prediction was a neighboring class. This nodule is located in a region of interest with many surrounding anatomical elements, making it difficult to perceive the nodule's boundaries. The challenge in extracting the relevant features that may have affected the prediction.



(a) [End-to-end approach] Lung-RADS class: GT - 4B; Prediction - 4A.

(b) [Feature-based approach] Lung-RADS class: GT - 4B; Prediction - 3.

(c) Oversegmentation of the solid core of a part-solid nodule.

(d) Oversegmentation of a lung nodule.

(e) GGO nodule, misclassified as solid.

(f) Calcified nodule, misclassified as non-calcified.

Figure 4.9: Examples of incorrect results from the automatic methods in this chapter, with ■ for manual annotation and ■ for automatic prediction in segmentation.

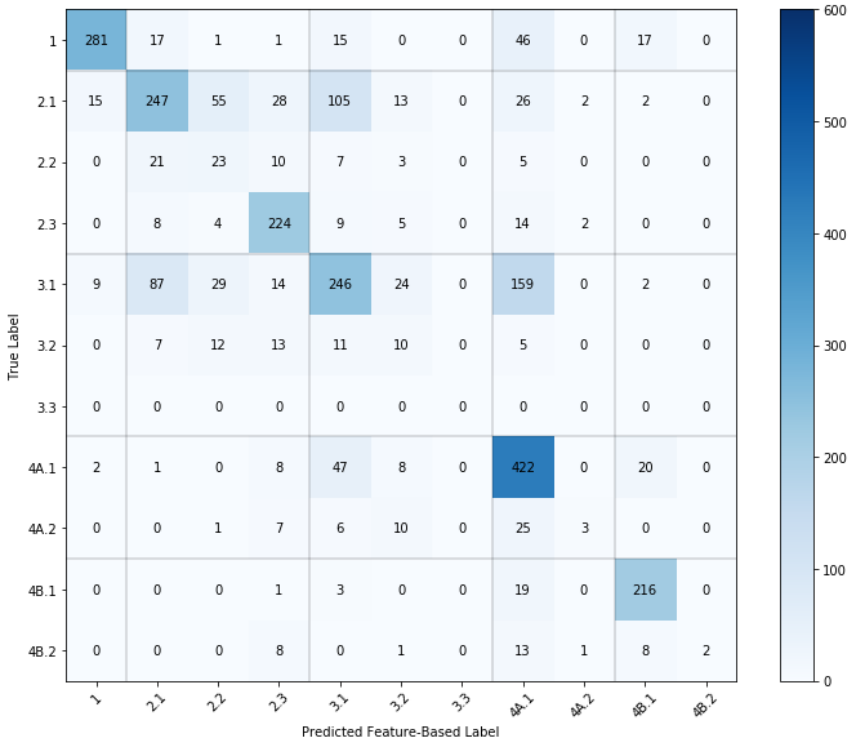
From this point on, for comparison with the feature-based model, the WRN model with sample weights is used, as it is determined to be the best in the end-to-end experiments.

Moving to the analysis of errors in the feature-based classification, an example of a misclassified nodule can be seen in 4.9b, where the cavitary nodule presents significant challenges due to its distinct characteristics compared to more common nodules. Only a few cases exhibit mismatches in diameter ranges within the solid core of part-solid nodules; one such example is shown in 4.9c, caused by oversegmentation. Segmentation, in general, shows a considerably high error rate, highlighting the need for further improvements in this task. Fixed thresholds, as commonly applied in guidelines, may not always provide optimal estimations. Nodules with diameters near threshold boundaries, even with similar predictions, may fall into different diameter ranges in Lung-RADS. The segmentation in 4.9d extends beyond the calcified nodule. Regarding texture, there is a notable deficit in accuracy for part-solid nodules, likely due to their close resemblance to GGO and solid nodules, leading to predictions being assigned to one of these two texture classes. In the example from 4.9e, a GGO nodule was misclassified as solid due to its proximity to the lung edges. Fortunately, calcification classification has shown high performance, likely due to the distinct HU ranges between calcified and non-calcified nodules. Calcified nodules typically exhibit HU above 200, while non-calcified nodules generally fall below 80 HU (Khan et al., 2010). An incorrect example is presented in 4.9f, once again likely affected by being connected to lung borders.

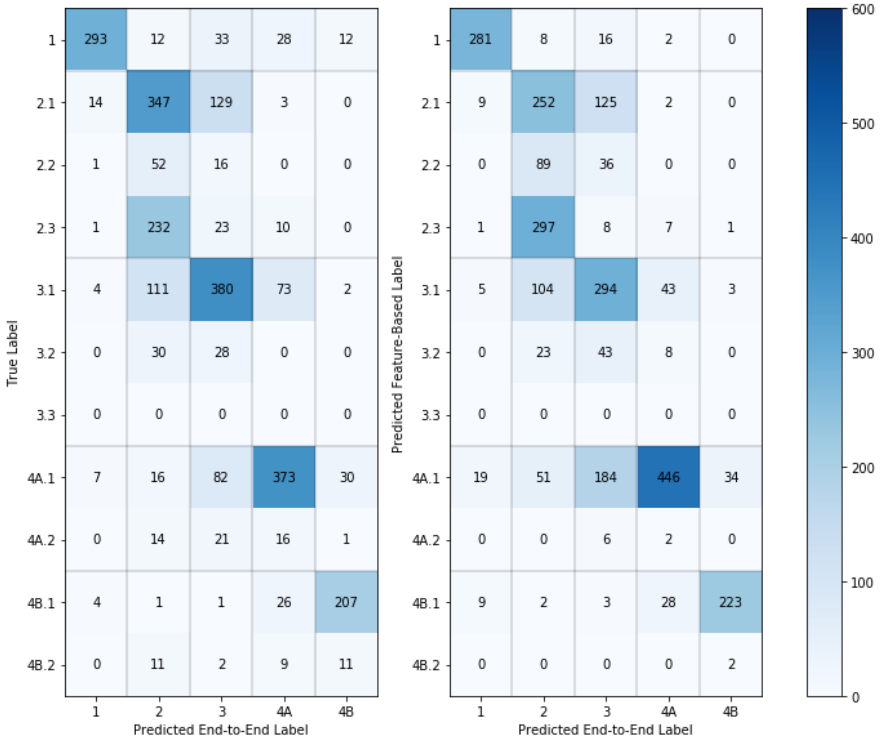
Quantitatively, the feature-based classification method and the end-to-end classification method (WRN with uncertainty weights) exhibit similar performance, with accuracy of 0.701 and 0.727, respectively. Despite this, a statistically significant difference between them was observed. This leads to the rejection of the null hypothesis (H_0) that there is no difference in the proportion of discordant predictions between the two models, with a p-value less than the alpha level of 0.05.

The *Ag* among radiologists' annotations is 0.670 for the feature-based method and 0.701 for the end-to-end method. Additionally, in Figure 4.10, confusion matrices are also depicted to provide insight into the different behaviour for the five classes, respectively for GT *vs.* Feature-based ($\mathscr{K}$: 0.612); GT *vs.* End-to-end ($\mathscr{K}$: 0.646); and Feature-based *vs.* End-to-end ($\mathscr{K}$: 0.642).

Figure 4.10a shows a tendency towards predicting class 3.1. Moreover, several cases are classified into the same Lung-RADS class despite potential errors occurring along the pipeline in the feature-based method. For instance, when a GGO or a solid nodule has a diameter of less than 6 mm, the Lung-RADS class remains the same in both cases, meaning that a texture classification error would not impact the decision. In Figure 4.10b, in the end-to-end method, errors primarily occur in classes 2 and 3. The performance is better for the "extreme" classes (1 and 4B) because they contain calcified and large nodules, respectively, while the more central classes include a mix of nodules with significantly different characteristics, making classification more difficult. When the prediction is not accurate, it usually falls within a nearby class, minimizing clinical impact. Finally, in Figure 4.10c, a comparison between the feature-based and end-to-end methods reveals an inconsistency in class predictions between the two methods.



(a) GT *vs.* Feature-based.



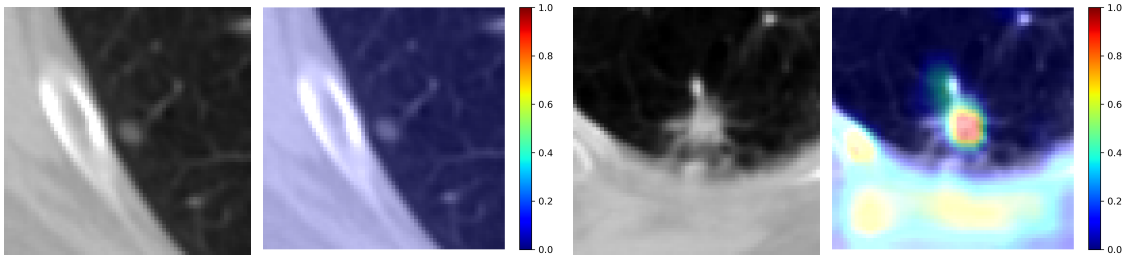
(b) GT *vs.* End-to-end. (c) Feature-based *vs.* End-to-end.

Figure 4.10: Confusion matrices between GT, end-to-end class prediction, and feature-based class prediction for Lung-RADS.

With these configurations, the end-to-end method had an inference time of 171 seconds, while the feature-based method took 117 seconds. Concerning computational space, the end-to-end method uses 5.4 GB, whereas the feature-based method, with all feature components summed up, used 71.4 MB. The difference in inference time demonstrates that the feature-based method is approximately 46.2% faster. Furthermore, the feature-based method occupies significantly less memory space, *i.e.*, 98.7% less than the end-to-end method. This disparity in space utilization is primarily attributed to the resource-intensive nature of the WRN employed in the end-to-end method.

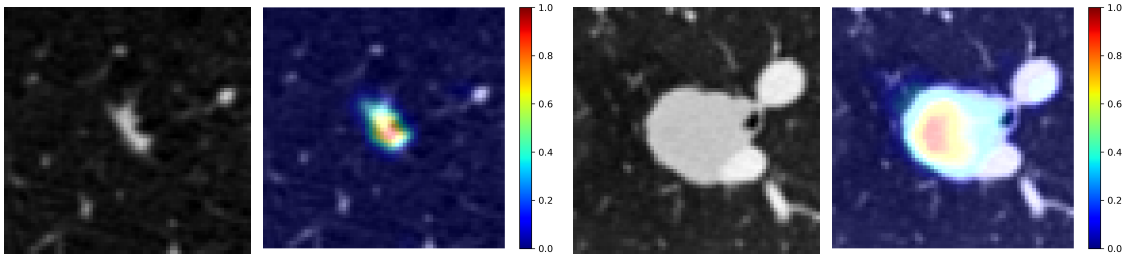
The most concerning nodule, defined as the one with the highest Lung-RADS class among the detected nodules, was identified with an accuracy of 0.747 using the end-to-end method, compared to 0.727 accuracy with the feature-based method. In addition to the previous results, this analysis checks whether the Lung-RADS class result for the most concerning nodule corresponds between the GT and the automatic methods. In both cases, there is no systematic error patterns, meaning that when the most concerning nodule is misclassified, it can lead to either a more conservative or a more lenient follow-up recommendation. Nevertheless, the end-to-end method continues to demonstrate better performance in identifying the most concerning nodule, with results closely matching those obtained when evaluating individual nodules in the previously reported performance analyses in this chapter.

To identify the main errors that can lead to incorrect end-to-end classification of follow-up, Grad-CAM (Selvaraju et al., 2017) is employed to highlight the most important regions of the image for prediction. Figure 4.11 presents examples of nodules where: 4.11a both methods fail; 4.11b only the end-to-end method fails; 4.11c only the feature-based method fails; and 4.11d both methods correctly classify the nodule. The class predicted by each method is given in the caption. Moreover, in the feature-based method, the reasons for error are specified, while in the end-to-end method, the prediction probability is included, along with the uncertainty weight. The correct prediction for both methods involves a large nodule. The feature-based method makes incorrect predictions, including both an overestimation of size (in 4.11a) and an incorrect texture determination, where a part-solid nodule is classified as solid (in 4.11c). In the case of the end-to-end method, since it needs to recognize the imaging features of nodules along with their differentiation, it struggles to classify and identify the relevant areas in 4.11a, resulting in an attention map with uniform tonality. In 4.11b, the juxtapleural location causes an erroneous prediction.



(a) **GT**: 3; **Feature-based**: 4A (Size error); **End-to-end**: 2 (0.604); α_i': 0.425.

(b) **GT**: 4A; **Feature-based**: 4A; **End-to-end**: 4B (0.544); α_i': 1.



(c) **GT**: 3; **Feature-based**: 4A (Texture error); **End-to-end**: 3 (0.633); α_i': 0.

(d) **GT**: 4B; **Feature-based**: 4B; **End-to-end**: 4B (0.997); α_i': 1.

Figure 4.11: Lung nodule examples, showing LR class results and with Grad-CAM highlights.

A comparative summary of the differences between the end-to-end method and the feature-based method is presented in Table 4.6. Overall, it becomes clear that the feature-based method is competitive with the end-to-end method in terms of performance. The feature-based method leverages medical knowledge for predicting the recommended follow-up class, but it may be affected by cumulative errors stemming from the various algorithms employed. It is important to note that the design of the feature-based method was not developed from scratch; rather, the results are derived from feature components that were originally designed to fulfill their primary functions, such as segmentation or classification, and not specifically for this purpose. On the other hand, the end-to-end method performs classification by learning the dataset's intricacies, but it blends nodules with vastly different characteristics within the dataset. This approach does not require exhaustive measurements and might be more suitable for stratification, where cases are often prioritized before being presented to a physician, making priority-based filtering particularly appropriate. Regarding explainability, a feature-based method, though potentially slightly less effective, provides clearer insights, which could offer compensatory benefits in clinical practice. If an error occurs in a feature component, it is easier for humans to intervene, leading to adjustments in the recommended follow-up.

Table 4.6: Comparison of end-to-end and feature-based approaches

	Feature-Based	**End-to-End**
Application	Clinical Practice	Stratification Purposes
Performance	Competitive	Statistically Different
Explainability	Clinical Methodology	Grad-CAM Based
Flexibility	Possible to change feature decisions	None
Inference Time	46% faster	Resource-intensive
Physical Memory	98.7% less memory space	

4.6 Concluding Remarks

This chapter addresses the objective of evaluating follow-up recommendations for lung cancer. Building on the Lung-RADS guidelines for follow-up recommendations and focusing on nodules that have already been detected, the goal is to automatically recommend the appropriate follow-up class to medical doctors. To this end, a comparison between a feature-based method and an end-to-end method is conducted. While there is competitiveness between the models, the primary takeaway is the potential avenues these methods could offer to medical doctors. The end-to-end method typically demonstrates superior performance, but it is less explainable and may have less obvious utility in clinical practice; stratification could be a possibility. Conversely, the feature-based method, which mimics clinical practice, is more tailored for adoption but falls slightly short in performance.

This chapter focuses on the Lung-RADS guidelines, but further investigation into the Fleischner and BTS guidelines would also be valuable. A significant challenge in this domain is the availability of data. The Fleischner and BTS guidelines require evaluating CT scans, which generally demands more patients for training. Since Lung-RADS bases its decisions on individual nodules, and each CT scan may contain several, the need for data is not as high as with the other guidelines. Despite that, Lung-RADS requires the segmentation of the solid core of part-solid nodules, which is uncommon in the literature. Collaboration with a clinical partner was fundamental in annotating this type of data to ensure the classification of Lung-RADS classes was possible.

Another noteworthy aspect of these methodological proposals is the incorporation of annotation uncertainty into the training process. It is common in the literature for datasets to lack multiple annotators per sample, making this dataset particularly valuable for this purpose. Although averaging score annotations and considering the majority score remain the most prevalent practices, there is growing interest in incorporating disagreement into the modeling process. The approach proposed here considers not only the variability in the annotations but also the number of annotators, which enhances the model's performance and provides valuable insights for addressing similar challenges.

Chapter 5

Forecast of Lung Nodule Diameter

The primary aim of this chapter was to forecast the progression of lung cancer by predicting the future diameter of lung nodules. This proposal uses various image data and discrete patient information to predict nodule diameter one and two years after the first exam. The most significant highlight is the prediction of malignant nodule diameters, advancing the current proposals in the literature. The chapter constitutes contribution **II.** of this thesis. Accordingly, it is divided into sections covering the introduction (5.1), related work (5.2), materials and methods (5.3), results (5.4), discussion (5.5), and concluding remarks (5.6).

The chapter is based on a publication[4] developed in collaboration with researchers at Radboudumc, Nijmegen, Netherlands, during a two-month internship by the author of this thesis under the Erasmus+ program. The thesis author is the first author of the study and was the primary contributor to the concept, the development of the methods, and the evaluation of the results.

5.1 Introduction

In lung cancer screening, radiologists examine the 3D images from chest low-dose Computed Tomography (CT) scans to detect lung nodules, followed by the application of management guidelines. These guidelines, including the Lung-RADS (American College of Radiology, 2019), the Fleischner (MacMahon et al., 2017), and the British Thoracic Society (BTS) (Callister et al., 2015), guide the radiologists on determining a patient's diagnosis and subsequent follow-up based on the findings from the CT scans.

When a single CT scan suggests a potentially malignant lung nodule, it often prompts immediate follow-up, which may include a biopsy to confirm malignancy or a Positron Emission Tomography – Computed Tomography (PET/CT) scan to stage the cancer (Latimer and Mott, 2015). The growth of the nodule over time, especially when observed at multiple intervals, is a

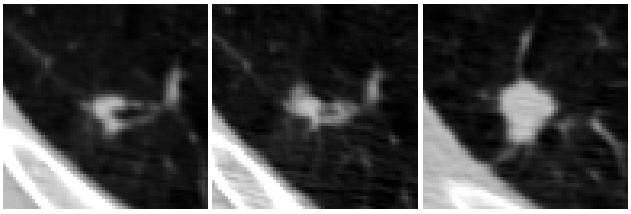
[4] **Carlos A Ferreira**, Kiran Vaidhya Venkadesh, Colin Jacobs, Miguel Coimbra, and Aurélio Campilho. Towards automatic forecasting of lung nodule diameter with tabular data and CT imaging. *Biomedical Signal Processing and Control*, 96:106625, 2024. doi: 10.1016/j.bspc.2024.106625.

critical factor in assessing malignancy risk. The guidelines classify nodule growth using distinct criteria, as shown in Table 5.1.

Table 5.1: Criteria for nodule growth classification.

Guideline	**Criteria for Growth**
Lung-RADS	1.5 mm increase in diameter (American College of Radiology, 2019)
Fleischner	2 mm increase in diameter (Nair et al., 2018)
BTS	25% increase in volume (Callister et al., 2015)

Figure 5.1 shows the growth of a malignant nodule over two years. Here, the nodule's equivalent diameter is estimated by modeling the nodule as a sphere, meaning that the nodule is represented a sphere with the same volume. Additionally, the Volume Doubling Time (VDT) is employed to quantify growth, representing the number of days required for the nodule volume to double. Malignant nodules are typically assumed to have a VDT of 348 days (Heuvelmans et al., 2017a).



	Y0	Y1	Y2
Diameter [mm]	7.7	9.0	14.3
Volume [mm^3]	239	381	1542
VDT (days)		542	136
Lung-RADS		**No Growth**	**Growth**
Fleischner		**No Growth**	**Growth**
BTS		**Growth**	**Growth**

Figure 5.1: A malignant nodule and its growth over two years: (from left to right) at year 0 (Y0), after 1 year (Y1) and after 2 years (Y2). The table presents the equivalent diameter, volume and VDT values and identification of the growing status according to the guidelines in the following two years.

To enhance the accuracy of malignancy assessment (Qiao et al., 2023), there is the need for advanced nodule size forecasting systems (Khanday and Sofi, 2021), given the known relationship between nodule growth and malignancy. Accurately classifying malignant nodules remains a significant challenge for medical doctors (Christakis, 2000), motivating the development of precise automatic tools. Forecasting nodule sizes involves anticipating future scenarios, a task that current clinical practice has yet to fully realize. This is where technology and machine learning methods can play a transformative role, offering less invasive alternatives to traditional biopsy methods.

Building upon the state-of-the-art methodologies, detailed in the next section, the proposals outlined in this chapter present an automatic method for forecasting lung nodule growth. This method is designed to be integrated into the lung nodule screening pipeline, following the phases of nodule detection, segmentation, and characterization. Nodule detection ensures the identification of the relevant nodules, segmentation provides the current nodule size for future diameter prediction, and characterization may inform the model about the nodule type, potentially refining the forecasting process. The proposed diameter forecasting method is applicable to any nodule,

with particular emphasis on growing and malignant nodules. Figure 5.2 show the general pipeline for forecasting future nodule diameter.

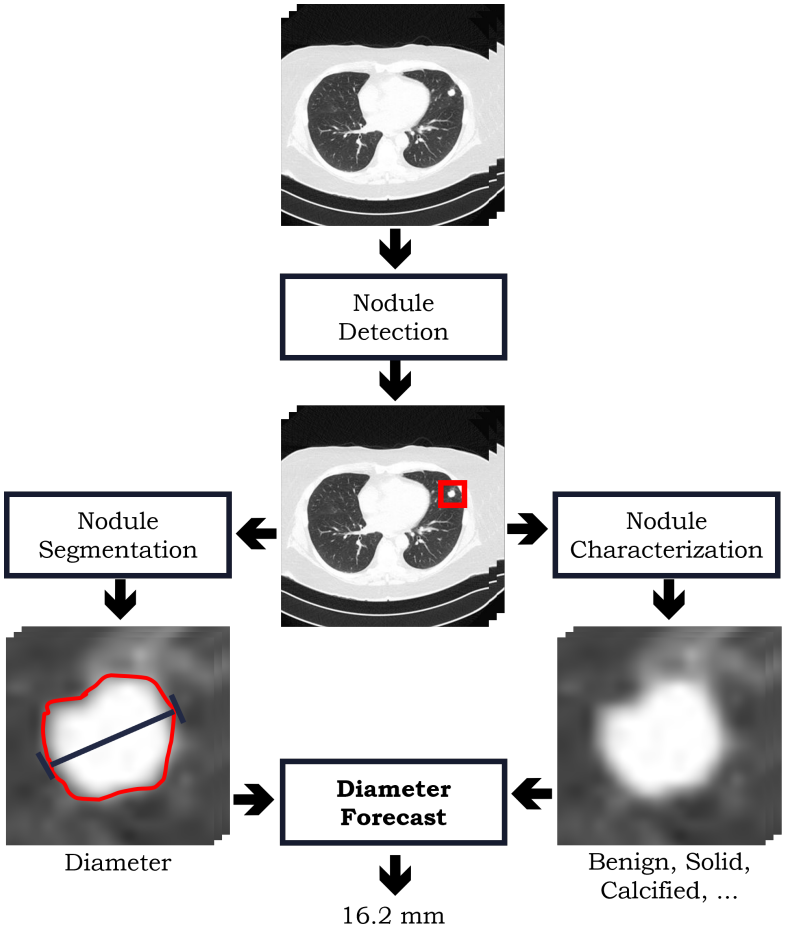


Figure 5.2: General pipeline for lung nodule follow-up.

Using Figure 5.1 as an example, for the input image at Y0, the diameter forecast at one year is given by the optimal regression result for this time frame, 9 mm in this case. Conversely, if the aim is to forecast the nodule size for two years, the optimal output would be 14.3 mm.

The proposed methods use different types of input data, including 2D images centered around the nodule (*I*), tabular data with patient information (*D*), and combinations of both (*I&D*). These methods provide forecasts for nodule diameter sizes at one or two-year intervals, with dedicated models for each year or adaptive selection for the desired forecast year. The backbone of the method is based on a Neural Network (NN), and when using image data, feature extraction is performed using a ResNet50. Figure 5.3 illustrates the method for two years ahead.

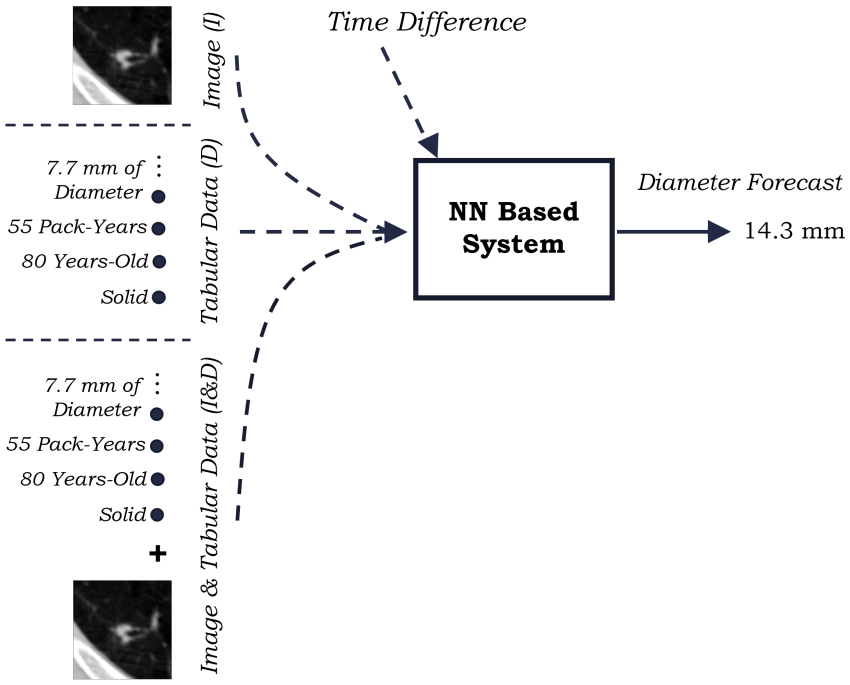


Figure 5.3: Overview of the three experimental set-up, for the following types of input data: 2D image data centered around the nodule (*I*), tabular data with patient information (*D*) and both types of data (*I*&*D*).

Thus, the main contributions of this research are:

i) A regression method for forecasting nodule growth, relying on a single time point. To the best of the author's knowledge, this is the first time a regression method has been proposed for this challenge.

ii) An in-depth analysis of the use of image and tabular data as inputs for forecasting nodule growth.

iii) A method that forecasts growth at more than one time point (in this case, one or two years after the time point in the input data).

iv) A feasibility study of these forecasting methods for malignant nodules and nodules classified as having grown.

5.2 Related Work

Recent literature has seen increased contributions to the automatic forecasting of lung nodule growth. This task has historically been challenging to tackle because of the need to acquire longitudinal data from the same patient over time, and conventional machine learning techniques have limited effectiveness due to the complexity and variability of medical images. However, recent advances in deep learning techniques have demonstrated significant potential in this domain. By using deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), researchers may automatically obtain visual features from medical images,

enabling more accurate and efficient algorithms. Lung nodule growth forecast methods could fall into three categories: classification, which determines whether growth occurs; numeric regression, which provides a growth value; and shape forecasts, which provides an estimate of the nodule shape on a future time.

Yang et al. (2022) addressed nodule growth by framing it as a binary classification task - stability or growth, based on Fleischner guidelines. They used five traditional machine learning models, namely logistic regression, Support Vector Machines (SVMs), random forest, MultiLayer Perceptron (MLP) and boosting methods, on a test dataset of 69 lung nodules, with input data comprising age and ten specific radiological features. The five models yielded Area Under the Receiver Operating Characteristic Curve (AUC) values of 82%, 81%, 77%, 81% and 71%, respectively, for the one year classification forecast. On the other hand, Fang et al. (2022) considered three classes - stability, growth and decrease, according to the Lung-RADS criteria. Their approach used two cuboids from the same nodule with a one year follow-up difference to forecast the class in the second year. They employed Siamese networks with backbones of a conventional CNN to which a Spatial-Temporal Mixer (STM) was added. By testing on 1,587 nodules from the National Lung Screening Trial (NLST) database, they achieved an AUC of 83.4%. Furthermore, they also experimented with the same network with two classes - stable and growth, according to BTS guidelines, obtaining an AUC of 85.8% for 627 nodules with subsolid characteristics (Liao et al., 2022).

In the context of forecasting the segmentation of nodules at a future time, Zhang et al. (2018) were the pioneers. They used dual-phase contrast-enhanced CT and PET/CT in a local patch-based classification task to forecast the shape of nodules at a third time. Their approach involved employing two-stream CNN architectures that were merged during learning by calculating pixel intensity ratios between the images of the two time points and comparing segmentation masks to represent tumor invasion and growth. A stream is a separate pathway comprising layers that process input data from a specific modality or source and contribute to the network's final output by extracting features from that specific data. Zhang et al. (2018) achieved a Relative Volume Difference (RVD) of 8.1% and a Dice of 85.9% in a test dataset of 10 patients. Rafael-Palou et al. (2022) proposed a similar network with two streams but based on a U-Net. The first stream used the 2D axial image of the initial time as input, while the second stream, which was activated only during the training processes, used both the patch at the initial time and the patch at the intended forecast time. They applied Monte Carlo sampling to estimate the mean and uncertainty of lung growth. Their method resulted in a Mean Absolute Difference (MAD) of 1.74 mm in growth size and a nodule segmentation Dice of 78% on an independent test dataset of 104 nodules.

To forecast both the segmentation and image at a future time point, Zhang et al. (2020) have proposed a method that employed a convolutional Long Short-Term Memory (LSTM) incorporating 3D inter-slice contexts. Their model used two time points and three channels, namely IntraCellular Volume Fraction (ICVF)-CT-Mask, to get results at a third time point. Their method resulted in an RVD of 11.2% and a Dice of 83.2% in segmentation results, which were evaluated using 11 patients. Another research group from Shanghai Jiao Tong University proposed a method based on two sequential networks, similar to 3D U-Nets (Li et al., 2020; Tao et al., 2022).

The first network computed the displacement field, followed by a texture network incorporating refinements to improve the visual appearance. The time interval, which varied for each sample, was concatenated after the encoding stage of the texture network. The evaluation of the mask segmentation proposal resulted in a Dice of 64.7% with 60 nodules. AUCs of 85.7% and 84.3% were obtained in classifying growth and non-growth nodules for Ground-Glass Opacity (GGO) and solid nodules, respectively. The high-risk class was considered with VDTs higher than 500 days for solid nodules and more than 1,500 days for subsolid nodules. There are also emerging approaches using transformers, such as Wang et al. (2022), who intended to forecast a third time point using the mask and image of the first two time points. Their method resulted in a Root Mean Squared Error (RMSE) and Dice of 11.32 mm and 89.31%, respectively, using 265 NLST nodules and a private dataset.

Finally, it is worth noting that Xiao et al. (2020) proposed a unique approach by developing a model that generates future images of lung nodules using the combined characteristics of variational autoencoder and RNNs. The model took images and tabular data, such as smoking, family cancer history, alcohol history and working history, as input and forecasted an image at a third time point. It is important to highlight that the output of this method is solely an image, and to evaluate it quantitatively, segmentation in the generated data needs to be performed. The authors used 80 CT scans from a private dataset and the NLST in their experiments, achieving a Dice of 82.5%.

Forecasting lung nodule diameter involves acquiring images at multiple temporal points to ensure accurate forecasting, underscoring the complexity of predicting nodule growth over time. Even existing methodological approaches need to be revisited, notably by introducing regression-based forecasting models that could be important for accurately determining the extent of nodule growth and for planning subsequent patient follow-up strategies. Despite several literature proposals, no method has been entirely based on a regression forecast. Most of the existing methods focus on the entire nodule pool rather than specifically on growing nodules, leading to an optimistic bias in the reported results due to the predominance of stable nodules. Additionally, using non-imaging patient data still needs to be further explored. Yang et al. (2022) and Xiao et al. (2020) are isolated contributions that have previously incorporated such data into their analyses. This underuse is an avenue that needs to be further explored by incorporating a broader spectrum of patient information in order to improve the forecasting accuracy resulting in a better patient care outcome.

5.3 Material and Methods

5.3.1 Dataset

The NLST database (Team, 2011), used in this study, contains data from individuals aged 55 to 74 years, who were current or former heavy smokers at the time of the study and had no prior history of lung cancer or related symptoms. The NLST adhered to stringent ethical guidelines, including

obtaining approval from an ethics committee and ensuring informed consent from participants. The inclusion criteria required a minimum smoking history of 30 Pack-Years (PYs), calculated by multiplying the average number of daily cigarette packs by the number of years of smoking. The database was collected between 2002 and 2004, with 26,722 participants from 33 medical centers across the United States of America who underwent annual low-dose CT screenings for three years. The availability of multiple scans for each patient at different time points enables researchers to conduct a comprehensive longitudinal analysis of patient evolution over time. Nevertheless, it is important to note that some patients could not complete the planned three rounds of screening due to variable participant compliance, dropout rates, and medical conditions, resulting in fewer time points for these individuals. Alongside the imagiologic data, the NLST study collected comprehensive medical, family, occupational, and lifestyle histories, providing a more thorough overview of the participants' health and facilitating a more in-depth analysis of the study findings.

To augment the utility of the NLST database, a team of healthcare professionals from Radboud University Medical Center provided annotations for the detected nodules (Venkadesh et al., 2021). These healthcare professionals conducted a single-reading annotation of the nodules, without seeking consensus or conducting additional readings, with specific attention to measuring their size and characteristics. Since radiologists' annotations of nodules tend to approximate a spherical shape, the diameter used throughout the entire study was based on the equivalent diameter of a sphere. Beyond this scope, the database is accessible to the scientific community upon formal request and includes only the slice-wise locations of the nodules.

To analyze the progression of the nodules over time, image registration of the CT scans from the NLST database was conducted to verify the correspondences of the same nodule annotated in different years. Nodules that had matches at least two time points were selected. Here, the notation 'Y' is used, followed by a numeric digit to designate the year of CT acquisition concerning the study's onset. T1 represents a one year difference, which can occur between Y0→Y1 or Y1→Y2. T2 denotes a two year difference, between Y0→Y2. Therefore, the following pairs of nodules at different time points are possible: Y0→Y1, Y1→Y2 and Y0→Y2.

Table 5.2 shows the distribution of the number of nodules based on partitions of the nodules' input diameter size and their growth. There are approximately 2.5 times more nodule annotations in the one year time frame compared to the two year time frame. Nevertheless, the distribution is quite similar for both cases, with a significant predominance of nodules between 4 and 8 mm and a diameter variation between -0.75 and 0.75 mm.

Table 5.3 presents the number of nodules used in this study and their characteristics corresponding to the 3,054 patients for whom there were multiple time points (57.07% are men and 42.93% are women). The table is organized by temporal difference (1 or 2 years) and distinguishes between cases where all screening time points are available and only two temporal points exist. The different types of nodules include solid, part-solid, GGO, perifissural and calcified. The malignancy of nodules is determined through biopsy tests. It is observed that the population under study is dominated by solid nodules, with malignant nodules accounting for only 8.17%. More-

Table 5.2: Number of nodules in ranges of diameter and growth rate.

		Growth [mm] @ 1 year										
		]-1.5[	[-1.5, -0.75[	[-0.75, 0[	[0, 0.75[	[0.75, 1.5[	[1.5, 2.25[	[2.25, 3[	[3, 3.75[	[3.75, 4.5[	[4.5[	Total
Diameter [mm]	] 4[	0	13	201	403	93	27	12	8	3	25	785
	[4, 8[	67	382	1586	1776	415	96	33	28	16	82	4481
	[8,12[	69	112	248	314	136	62	36	23	14	52	1066
	[12,16[	56	50	65	83	50	32	22	11	3	21	393
	[16,20[	29	10	19	16	12	16	6	8	1	11	128
	[20,24[	9	5	3	10	4	3	5	2	1	4	46
	[24,28[	5	2	2	1	2	2	2	1	1	3	21
	[28,32[	6	0	2	2	1	1	0	0	0	2	14
	[32,36[	2	0	0	0	2	1	0	0	0	2	7
	[36[	3	2	0	3	2	0	0	2	0	1	13
	Total	246	576	2126	2608	717	240	116	83	39	203	6954

		Growth [mm] @ 2 years										
		]-1.5[	[-1.5, -0.75[	[-0.75, 0[	[0, 0.75[	[0.75, 1.5[	[1.5, 2.25[	[2.25, 3[	[3, 3.75[	[3.75, 4.5[	[4.5[	Total
Diameter [mm]	] 4[	0	2	68	116	32	11	5	4	2	24	264
	[4, 8[	31	170	549	660	168	50	22	13	3	56	1722
	[8,12[	20	53	81	106	56	25	10	14	8	28	401
	[12,16[	19	16	22	27	18	13	4	4	3	12	138
	[16,20[	7	3	5	7	4	7	1	1	2	7	44
	[20,24[	4	3	2	0	2	1	2	1	0	2	17
	[24,28[	2	0	0	0	1	1	0	0	0	3	7
	[28,32[	3	0	1	1	0	0	0	0	0	2	7
	[32,36[	1	0	1	0	0	0	1	0	0	0	3
	[36[	1	0	0	0	1	0	0	1	0	1	4
	Total	88	247	729	917	282	108	45	38	18	135	2607

Label: The bluer colors indicate a higher % of representation (□ → ■).

Table 5.3: Number of nodules used from the NLST database based on their characteristics.

				Growth			Type				
Time Difference	Groups	Total	Malignancy	Lung-RADS	Fleischner	BTS	Solid	Part-Solid	GGO	Perifissural	Calcified
T1	Y0→Y1	989	138	170	140	334	664	85	202	38	0
	Y0→Y1*	2451	197	201	142	719	1619	199	457	175	0
	Y1→Y2	1063	33	81	65	325	693	58	215	96	1
	Y1→Y2*	2451	205	229	178	705	1636	222	424	167	2
	Total	6954	573	681	525	2083	4612	564	1298	476	3
T2	Y0→Y2	156	8	20	14	56	105	14	29	8	0
	Y0→Y2*	2451	197	324	257	831	1619	199	457	175	0
	Total	2607	205	344	271	887	1724	213	486	183	0
	Total	7110	581	1426	1246	2502	4717	578	1327	484	3

Label: * includes the number of nodules with three time points.

over, the BTS criteria are more permissive. In other words, the percentage of nodules considered to have grown is 35.19% in BTS, which is considerably higher than in the case of Fleischner and Lung-RADS, where the percentages are 17.52% and 20.05%, respectively.

The input images employed in this study have 64×64 pixels, with a pixel size of 0.7 mm×0.7 mm. These images correspond to the medial axial slice extracted from a cuboid centered around the annotated nodule's centroid. The Region of Interest (ROI) was normalized from

−1000 to 400 Hounsfield Unit (HU), a range commonly used in nodule classification tasks (Marques et al., 2021), and transformed to a normalized intensity range of [0, 1].

In this research, "tabular data" refers to the numerical patient-related feature data, detailed in Table 5.4. The decision to include these features is supported by the literature (Pedrosa et al., 2022). Each feature has been scaled to the range [0,1], except for the nodule type. The classes of the frequency of alcohol consumption variable were converted to a scale of $[0, 1]$ as they refer to quantities (1: never, 2: monthly or less often, 3: two to four times a month, 4: two to three times a week, and 5: four or more times a week). In the case of nodule type, a number is assigned to each class (1: Solid/Calcified, 2: Part-Solid, 3: GGO and 4: Perifissural). Since the class of calcified nodules was minimally represented, it was merged with the solid class, given their generally higher HU intensity.

Table 5.4: Tabular patient data, and corresponding minimum, maximum and average, when applicable.

Type	*Abbreviation*	*Description*	*Minimum*	*Maximum*	*Average*
Binary	SmokeWork	Exposure to smokers at work	0	1	0.87
Binary	WorkType	Work related to cancer risk	0	1	0.29
Binary	CancerFamily	Direct family member had lung cancer	0	1	0.22
Binary	SmokeHouse	Participant lived with smoker	0	1	0.89
Binary	LungDiseases	Diagnosed with lung disease before	0	1	0.66
Binary	OtherCancer	Diagnosed with other cancer before	0	1	0.04
Numeric	Age	Patient's age	55	74	62.33
Numeric	BMI	Body Mass Index	15	52	27.74
Numeric	QuitYears	Years since quit smoking	0	22	3.74
Numeric	PYs	Pack-Years	30	260	57.44
Numeric	Diameter	Input nodule diameter [mm]	3	44	6.89
Classes	Alcohol	Frequency of alcohol consumption	1	5	2.85
Classes		Nodule type	1	4	—

5.3.2 Experiments

This research forecasts the future diameter of a nodule by employing various experimental workflows, depending on the input data (see Figure 5.4) that are either 2D regions of the lung nodule in the CT image (*I*), or the tabular data (*D*) described in Table 5.4 or a combination of both (*I*&*D*). All workflows include an NN consisting of a single fully connected hidden layer with 16 neurons. Thus:

- *I*: The initial processing layer following the CT image input is a convolutional layer with three filters, enabling the transformation of grayscale, single-channel inputs, into a three-channel format. The three-channel allows the use of common pre-trained models, as the ResNet50 used in this research, due to its capabilities for feature extraction (Figure 5.5). These layers are trained to tune their parameters to this application. Subsequently, the multidimentional output of the

convolutional component of ResNet50 is flattened into a one-dimensional array. The result is a feature vector with 512 values, which is the input for the hidden layer NN_*I*.

• *D*: The input is the tabular data with 16 input neurons, directly connected to the hidden layer NN_*D*.

• *I&D*: First, each data type is processed separately being the image feature vector obtained through ResNet50 (He et al., 2016). The image feature vector has 512 values and the textual data has 16 input values. The hidden layer NN_*I&D* is set up separately, containing 16 neurons for each data type. The neurons from both data types are then concatenated.

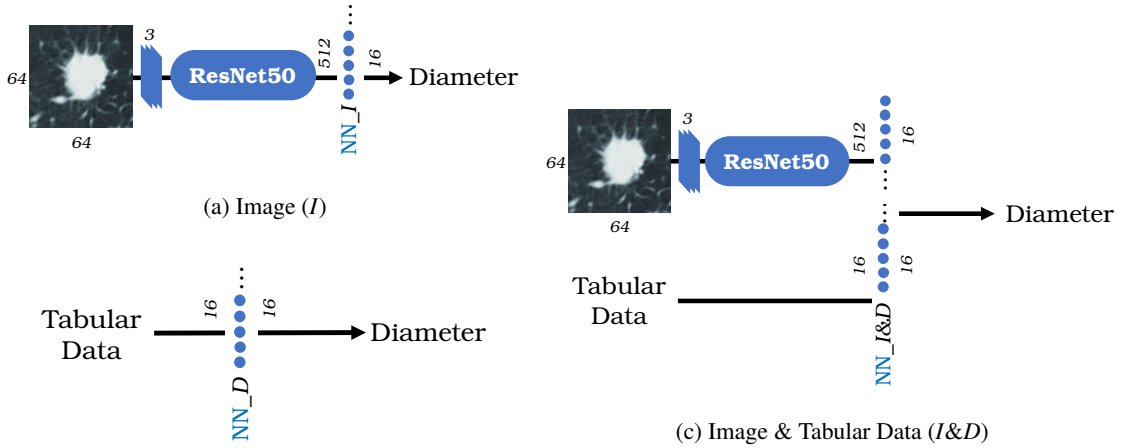


Figure 5.4: Overview of the three workflows used in the forecast experiments.

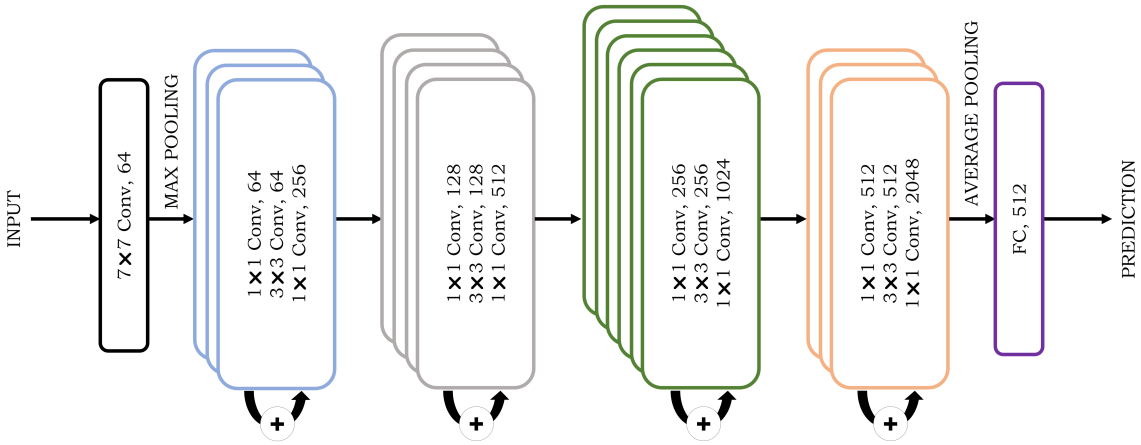


Figure 5.5: ResNet50 Architecture. The operations are represented by rounded-corner rectangles. Convolutional layers (Conv) are composed of a sequence of operations: a convolution, a batch normalization, and a Rectified Linear Unit (ReLU). FC refers to a fully connected layer. The filter size is indicated before the layer type, and the number of filters or neurons is indicated after the comma.

Finally, an activation layer with a linear function is applied prior to the last neuron, which is the one targeted for predicting the diameter of a nodule at a future time. This activation aligns with the need for continuous scale predictions in the context of nodule diameter regression forecasting.

Two options are considered: use separate models for predicting the nodule diameter for the two different year intervals, or using a single model capable of forecasting nodule diameter for both intervals. The objective was to explore the trade-off between resource optimization and performance. In the first case, these models are referred as "dedicated" because separate models were required for yearly predictions. In the second case, these models are referred as "adaptive", where a binary flag was incorporated into the model, to indicate which year is going to be forecast (0 for forecasting at one year and 1 for forecasting at two years).

To address the dataset's imbalance, where stable nodules are predominant, the classes are weighted differently to ensure that the training process accounts for the uneven distribution of nodule growth. In particular, the growth of malignant nodules is expected to be larger than the growth of other types of nodules. Two groups of nodules are considered: growing and non-growing classes, assessing each nodule's conformity to the growth criteria according to the various guidelines in Table 5.1. The class weights varies with the guideline and with the class, according to the expression in Equation (5.1).

$$W_{c,g} = \frac{N}{2N_{c,g}} \tag{5.1}$$

where $W_{c,g}$ and $N_{c,g}$ represent the class weight and the number of nodules, respectively, for class c according to guideline g, as in Table 5.1. Class c refers to whether the nodule is classified as growing or not growing. N denotes the total number of nodules in the dataset.

Overall, 48 results ($3\times2\times2\times4$) are obtained, according to the options below:

- **Input Data**: 2D ROI around the nodule (*I*), tabular data with patient information (*D*), and both (*I&D*);
- **Time Difference**: forecasting for time frames of 1 year and 2 years ahead of the input data;
- **Model Configuration**: a "dedicated" configuration, requiring a separate model for each of the desired time frames, and an "adaptive" configuration, where a single model serves to provide predictions for the target time frames by adjusting a flag;
- **Class weights**: without weights (w/o) and weighted for growth according to Lung-RADS, Fleischner, and BTS.

5.3.3 Model Training

Cross-validation was performed by dividing the data into 5 folds, with 3 being used for training, 1 for validation, and 1 as a fixed test fold, remaining unchanged across different training and validation sets permutations.

In the *I* model, pre-trained weights from ImageNet were employed. In the *I&D* model, the pre-trained weights from the *I* model were used for image feature extraction. This choice in the *I&D* model allows starting in a suboptimal space, and given the greater extent of the *I* part compared

to the D part, otherwise the propagation of learning would be concentrated only on the D part, making it impossible to train the feature extraction part of I.

Data augmentation was introduced to improve model generalization and robustness by dynamically applying random rotations in all directions and horizontal/vertical flips on the images during each epoch, also helping to mitigate the risk of overfitting. Consequently, each training sample received a randomly transformed version, with potential variations in the transformations applied in subsequent epochs. It is worth noting that the validation and the test sets do not use augmented data. In instances where tabular data was missing, a random value was introduced as a form of augmentation in each epoch, thus increasing the dataset's variability.

For training, the Mean Squared Error (MSE) loss function is used. The loss function, L, is defined by Equation (5.2). $W_{c,g} = 1$ if the experiment does not consider class weights; otherwise, the calculation is performed according to Equation (5.1) for each class and guideline considered in this study.

$$L = \frac{1}{N}\sum_{i=1}^{N} W_{c,g}(d_i - \hat{d_i})^2 \tag{5.2}$$

d_i is the Ground Truth (GT) label of sample i and $\hat{d_i}$ is the predicted regression value of sample i.

The learning process used the Nesterov Adam optimizer (Dozat, 2016), with a learning rate of 0.001 and a batch size of 16. In order to mitigate overfitting, an early stopping criterion was incorporated into the training phase, with a patience parameter set to 10 epochs. The maximum number of epochs was 60. For model training, it is used Python 3.5 and TensorFlow 2.3.0 on a system with an NVIDIA GTX1080, with 8 GB of GPU memory.

5.3.4 Model Evaluation

For evaluating the regression results, the MAD of the diameter prediction is computed, as defined by Equation (3.7). This metric measures the average magnitude of errors in the predictions, providing a straightforward interpretation by considering the absolute differences between predicted and actual values, which ensures that all errors are treated equally regardless of their direction. The evaluation encompassed the entire test set, and involved an analysis of different classes, explicitly focusing on malignancy and growth, as these are the primary targets of interest in a practical scenario.

A comparative analysis is conducted with various studies in the field, employing MAD and RMSE metrics. The RMSE was defined in Equation (3.8) and is used to quantify the model's predictive performance by emphasizing larger errors in the context of diameter prediction. Furthermore, the image-based and tabular data-driven methodologies are evaluated, without weighted training, in a data reduction scenario where half of the data from the training and validation sets are used, to understand the impact on the previous metrics not only to evaluate model performance but also to explore its ability to learn from limited data and identify any saturation effect.

As previously noted, a VDT of 348 days is the threshold for malignant nodules (Heuvelmans et al., 2017a). The VDT represents the time it takes for the nodule volume to double and is calculated using the diameters at two different time points, d_{t_0} and d_{t_n}, over the time interval T in days, assuming the nodule is approximately spherical. The formula for VDT is given by:

$$VDT = \frac{T \cdot \log(2)}{\log\left(\frac{d_{t_n}}{d_{t_0}}\right)^3} \tag{5.3}$$

An analysis was conducted to identify the experiments that, on average, resulted in a VDT with a smaller difference from the GT compared to the reference assumption of 348 days from the literature. Experiments meeting the conditions above allow progression towards a dynamic forecasting approach rather than assuming a fixed default value.

Modified Bland-Altman plots are employed to perceive a tendency in the prediction error, which can either be an overforecast or an underforecast. In these modified plots, the vertical axis represents the differences between the predicted values and the GT for nodule growth prediction, while the horizontal axis corresponds to the GT nodule diameters, instead of the traditional average of the two measurements on the horizontal axis. This change is made to better understand the relationship between the prediction error and the GT size. Both stable and growing nodules are analyzed. The Bland-Altman approach provides a visual means to evaluate the bias (or mean difference) and the limits of agreement, which are calculated as the mean difference ± 1.96 times the standard deviation (representing the 95% confidence interval). The Pearson correlation coefficient between the prediction error and the GT is also computed, given by:

$$r = \frac{\sum(d_i - \bar{d})(\Delta_i - \bar{\Delta})}{\sqrt{\sum(d_i - \bar{d})^2 \sum(\Delta_i - \bar{\Delta})^2}} \tag{5.4}$$

where d_i is the GT label of sample i, $\Delta_i = d_i - \hat{d}_i$ is the prediction error for sample i, and $\bar{d}$ and $\bar{\Delta}$ are their respective means. A positive correlation indicates that, as the nodule diameter increases, the model tends to overestimate the predicted value, while a negative correlation suggests that the model tends to underestimate the prediction.

Finally, the importance of each feature used in the regression is assessed. Each feature was independently randomized, and the subsequent impact on the predictions was observed in terms of absolute value, allowing for a better understanding of which features are relevant.

5.4 Results

5.4.1 Growth Regression

This subsection starts by showing some selected forecasting results. Both Figure 5.6 and Figure 5.7 illustrate the input-output relationship of the models, using the network trained with class weights based on Lung-RADS guidelines, which specifically receives image data (I).

In Figure 5.6, results for two separate forecast years are shown, providing an indication of the predictive model's performance. On the other hand, Figure 5.7 highlights some of the best one-year forecast results, where it is evident that all nodules are isolated, despite the relative proximity of other anatomical structures. The forecasted diameters are also presented in the caption of each subfigure. The different cases correspond to scenarios of substantial growth (5.7a), slight growth (5.7b), reduction (5.7c), and stability (5.7d).

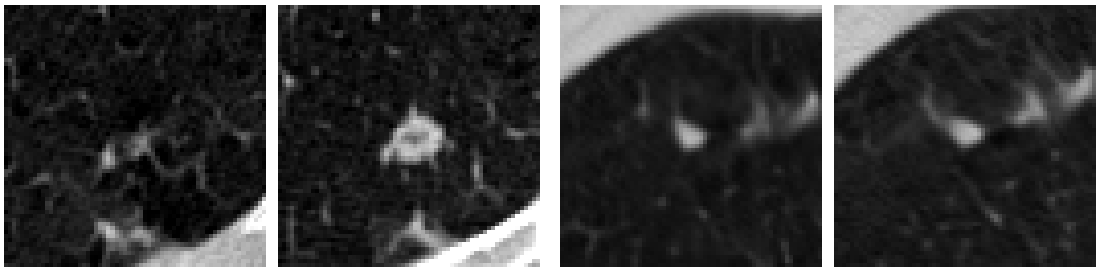


(Y0) **GT:** 9.20 mm.

(Y1) **GT:** 11.90 mm; **Prediction:** 12.71 mm.

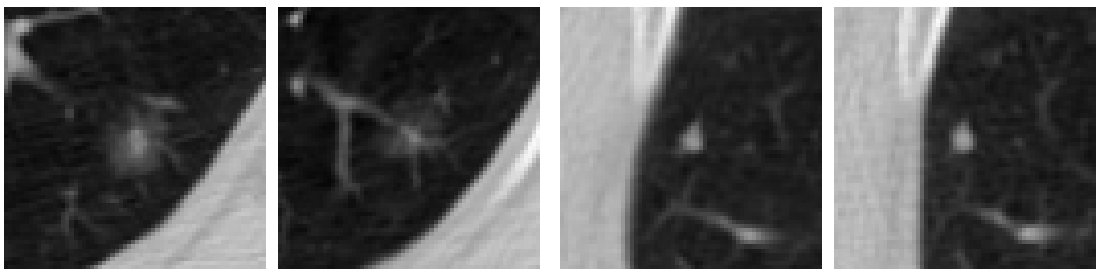
(Y2) **GT:** 14.70 mm; **Prediction:** 14.99 mm.

Figure 5.6: Example of a nodule forecast over two years.



(a) **Y0:** 5.30 mm; **GT:** 10.00 mm; **Prediction:** 10.01 mm; **Difference:** **0.01** mm.

(b) **Y0:** 5.50 mm; **GT:** 6.60 mm; **Prediction:** 6.58 mm; **Difference:** **−0.02** mm.



(c) **Y0:** 11.70 mm; **GT:** 10.10 mm; **Prediction:** 10.07 mm; **Difference:** **−0.03** mm.

(d) **Y0:** 4.50 mm; **GT:** 4.50 mm; **Prediction:** 4.47 mm; **Difference:** **−0.03** mm.

Figure 5.7: Examples of the ROIs of the nodules corresponding to some of the best one year forecast results achieved using the network trained with class weights based on Lung-RADS guidelines, which specifically receives image data (I).

The evaluation of the lung nodule growth forecasting experiments is presented, with Table 5.5 showing the MAD and corresponding standard deviations for the test set. Results are reported for the *I*, *D*, and *I&D* experiments, considering both adaptive and dedicated approaches at T1 and T2 time points. It should be noted that 48 results are achieved, with variations in input data, class weights (rows of Table 5.5), time differences, and model configurations (columns of Table 5.5). Regarding the last parameter, *i.e.*, the model configuration, in the dedicated approach for forecasting the growth of nodules, the results are obtained separately for one year and two years, while in the adaptive approach, the results can be obtained jointly for both temporal points.

Table 5.5: MAD (mm) and standard deviation (mm) for different experiments on the test set.

	Time Difference	T1		T2	
	Model Configuration	*Dedicated*	*Adaptive*	*Dedicated*	*Adaptive*
I	W/o *weights*	1.72 ± 0.20	1.49 ± 0.14	2.13 ± 0.23	1.59 ± 0.09
	Lung-RADS *weights*	2.48 ± 0.19	2.48 ± 0.13	2.42 ± 0.22	2.62 ± 0.08
	Fleischner *weights*	3.10 ± 0.18	2.70 ± 0.25	3.13 ± 0.66	2.96 ± 0.40
	BTS *weights*	1.84 ± 0.06	1.66 ± 0.08	2.27 ± 0.36	1.78 ± 0.13
D	W/o *weights*	0.99 ± 0.03	1.00 ± 0.03	1.31 ± 0.04	1.26 ± 0.04
	Lung-RADS *weights*	2.13 ± 0.17	1.99 ± 0.24	2.50 ± 0.05	2.64 ± 0.18
	Fleischner *weights*	2.51 ± 0.28	2.34 ± 0.21	2.90 ± 0.12	3.11 ± 0.18
	BTS *weights*	1.19 ± 0.07	1.19 ± 0.08	1.51 ± 0.07	1.56 ± 0.06
I&D	W/o *weights*	1.56 ± 0.07	1.38 ± 0.07	1.66 ± 0.06	1.60 ± 0.06
	Lung-RADS *weights*	2.62 ± 0.18	2.02 ± 0.18	2.17 ± 0.14	2.16 ± 0.13
	Fleischner *weights*	3.15 ± 0.23	2.56 ± 0.23	2.72 ± 0.39	2.69 ± 0.29
	BTS *weights*	1.84 ± 0.07	1.61 ± 0.08	1.63 ± 0.21	1.66 ± 0.10

Label: The table highlights the MAD top 3 and bottom 3 results by coloring them green and red, respectively. The color shades are proportional to their ranking position, with the best result being the most green and the worst being the most red. Any result outside of the top 3 or bottom 3 is colorless.

The MAD values in the Table 5.5 highlights the top three and bottom three results by color-coding, with green to representing the best results and red to representing the worst results. The color intensity is proportional, with the highest-ranking result shown in the deepest shade of green. The second-best result is displayed at 2/3 of the intensity of the top shade, and the third-best result appears at 1/3 of the deepest green shade. The fourth-best result remains uncolored. A similar principle is applied to the worst outcomes, marked in red.

Upon analyzing Table 5.5, it is evident that the most favorable results are obtained when using tabular data as input. Conversely, the results degrade when combining tabular data with the image component. The least favorable outcomes occur when using only the image as input.

Using class weights during training does not provide significant benefits, as it is clear for all the three guidelines, particularly for the Fleischner guidelines that shows the worst results.

In terms of the forecast year, two key observations are noteworthy. Firstly, better results are achieved when forecasting one year compared to a two year forecast. Secondly, the adaptive approach employed in this study does not lead to significant performance deterioration.

5.4.2 Analysis of the Classes with Malignant or Growing Nodules

This subsection aims to present the results of forecasting the diameter of nodules, corresponding to experiments with the malignant nodules confirmed by biopsy (Table 5.6) and with nodules classified as growing nodules according to the three guidelines (Table 5.8).

For the analysis of malignant nodules, Table 5.6 follows the same format of Table 5.5, using the same 48 experiments. The layout keeps the same structure, with input data and class weights in rows, forecast years in columns, and a distinction between dedicated and adaptive schemes. The color scheme remains consistent, with three color-coded values representing the best results in green and the three worst in red for all metrics.

Table 5.6: MAD (mm) and standard deviation (mm) for different experiments on malignant nodules.

	Time Difference	T1		T2	
	Model Configuration	*Dedicated*	*Adaptive*	*Dedicated*	*Adaptive*
I	W/o *weights*	4.15 ± 0.27	3.92 ± 0.08	6.13 ± 0.41	5.43 ± 0.10
	Lung-RADS *weights*	3.57 ± 0.13	3.47 ± 0.13	5.09 ± 0.51	4.32 ± 0.33
	Fleischner *weights*	4.22 ± 0.18	3.93 ± 0.10	5.32 ± 0.33	5.10 ± 0.20
	BTS *weights*	4.24 ± 0.05	3.68 ± 0.05	5.17 ± 0.33	5.34 ± 0.25
D	W/o *weights*	3.52 ± 0.07	3.57 ± 0.11	5.27 ± 0.12	5.18 ± 0.12
	Lung-RADS *weights*	2.82 ± 0.03	2.83 ± 0.03	4.09 ± 0.06	3.71 ± 0.10
	Fleischner *weights*	2.88 ± 0.05	2.83 ± 0.02	3.89 ± 0.10	3.52 ± 0.03
	BTS *weights*	3.12 ± 0.08	3.15 ± 0.08	4.89 ± 0.10	4.62 ± 0.07
I&D	W/o *weights*	4.16 ± 0.34	3.76 ± 0.24	5.14 ± 0.14	5.15 ± 0.11
	Lung-RADS *weights*	3.20 ± 0.09	3.28 ± 0.34	4.55 ± 0.46	4.05 ± 0.08
	Fleischner *weights*	4.24 ± 0.39	3.59 ± 0.07	4.21 ± 0.06	4.38 ± 0.28
	BTS *weights*	3.73 ± 0.11	3.19 ± 0.37	5.16 ± 0.19	4.71 ± 0.14

Label: The table highlights the MAD top 3 and bottom 3 results by coloring them green and red, respectively. The color shades are proportional to their ranking position, with the best result being the most green and the worst being the most red. Any result outside of the top 3 or bottom 3 is colorless.

As shown in Table 5.6, employing class weights during training is crucial for improving the regression prediction. Lung-RADS produces the most favorable results, followed by Fleischner and BTS. These findings contradict the results in Table 5.5, where using class weights did not provide any advantage.

Concerning input data and forecast year, the results in Table 5.6 align with those in Table 5.5. This consistency underscores the superiority of using tabular data as input over images and shows that the one year forecast is better than the two year forecast. Additionally, employing an adaptive approach for forecasting both years does not result in a performance degradation.

Table 5.7 displays experiments in which the average predicted VDT of malignant nodules is closer to the GT than the reference of 348 days (Heuvelmans et al., 2017a). The predicted results are arranged from top to bottom based on their proximity to the GT. These results align with the findings in Table 5.6, highlighting the importance of using class weights during training to forecast nodule diameters closer to GT values.

Table 5.7: Experiments in which the predicted VDT of malignant nodules closely approximates the GT.

Input Data	*Weights*	*Time Difference*	*Model Configuration*	VDT_{GT}	$VDT_{predicted}$
D	Fleischner	T1	*Adaptive*	404.68	409.61
I&D	Fleischner	T2	*Dedicated*	315.13	321.29
D	Lung-RADS	T1	*Dedicated*	404.68	397.29
I&D	BTS	T1	*Adaptive*	404.68	416.07
I&D	Lung-RADS	T2	*Adaptive*	315.13	327.85
D	Fleischner	T2	*Adaptive*	315.13	299.47
D	Lung-RADS	T2	*Adaptive*	315.13	346.86
D	Fleischner	T1	*Dedicated*	404.68	364.93

For the analysis of growing nodules, Table 5.8 is presented. Here, three blocks of rows represent the results (MAD) for nodules that have grown according to the three guidelines (Lung-RADS, Fleischner, and BTS). In terms of experiments, different types of input (image, tabular data, and both) are shown in rows, while the forecast year (either one or two years ahead, and in a dedicated or adaptive way) is shown in columns. Since the results of using all class weights for each group of nodules are not represented, only 24 results are available in Table 5.8 for each guideline.

When analyzing nodules that grow, if the predicted value is lower than the growth threshold for the guideline, the threshold value is used. This approach anticipates that the forecast of a nodule's diameter may be implemented after a binary classification algorithm for growth *vs.* stability classification. With the expectation of reliability in the classification algorithm, if the regression prediction falls below the growth threshold, it is adjusted to match the threshold. For Lung-RADS, Fleischner, and BTS, this corresponds to an increase of 1.5 mm in diameter, 2 mm in diameter, and 25% in volume, respectively.

Table 5.8 presents three major groups corresponding to nodules classified as having grown according to specific guidelines. The color scheme has been separated to emphasize each group's best and worst results. The top 3 and bottom 3 results have been color-coded, with shades of green and red indicating their respective rankings.

As depicted in Table 5.8, the results derived from the adaptive approach show substantial competitiveness compared to the dedicated approaches. Moreover, it is noticeable that one year forecasts consistently offer higher reliability compared to two year forecasts. However, concerning input data, the preferred approach involves the combination of tabular and image data. Class weights are recommended for these results in the analysis of growing nodules.

Table 5.8: MAD (mm) and standard deviation (mm) for different experiments on nodules classified as grown according to three guidelines (Lung-RADS, Fleischner and BTS).

	Time	T1				T2			
	Conf.	*Dedicated*		*Adaptive*		*Dedicated*		*Adaptive*	
	Weights	✗	✓	✗	✓	✗	✓	✗	✓
Lung-RADS	*I*	2.31 ± 0.03	2.23 ± 0.09	2.25 ± 0.05	2.35 ± 0.10	3.03 ± 0.17	2.82 ± 0.19	3.09 ± 0.09	2.74 ± 0.19
	D	2.55 ± 0.01	2.10 ± 0.01	2.55 ± 0.01	2.15 ± 0.05	3.42 ± 0.01	3.08 ± 0.05	3.42 ± 0.01	2.79 ± 0.08
	I&D	2.37 ± 0.11	2.02 ± 0.12	2.20 ± 0.12	2.23 ± 0.13	2.96 ± 0.08	3.00 ± 0.18	2.96 ± 0.10	2.70 ± 0.19
Fleischner	*I*	2.46 ± 0.03	2.52 ± 0.06	2.40 ± 0.04	2.55 ± 0.08	3.35 ± 0.20	3.31 ± 0.21	3.41 ± 0.08	3.83 ± 0.18
	D	2.69 ± 0.01	2.24 ± 0.05	2.69 ± 0.01	2.27 ± 0.06	3.72 ± 0.01	3.22 ± 0.12	3.72 ± 0.01	2.96 ± 0.06
	I&D	2.55 ± 0.13	2.85 ± 0.14	2.34 ± 0.07	2.37 ± 0.17	3.27 ± 0.08	3.00 ± 0.14	3.26 ± 0.09	2.99 ± 0.17
BTS	*I*	1.53 ± 0.12	1.51 ± 0.12	1.40 ± 0.14	1.46 ± 0.15	2.04 ± 0.12	1.96 ± 0.14	1.73 ± 0.06	1.83 ± 0.19
	D	1.33 ± 0.02	1.24 ± 0.03	1.33 ± 0.02	1.25 ± 0.02	1.76 ± 0.01	1.77 ± 0.03	1.77 ± 0.02	1.72 ± 0.02
	I&D	1.46 ± 0.08	1.53 ± 0.15	1.32 ± 0.05	1.33 ± 0.11	1.70 ± 0.03	1.72 ± 0.13	1.69 ± 0.09	1.66 ± 0.19

Label: The table highlights the MAD top 3 and bottom 3 results in each group of nodules considered as "grown" according to the guidelines (Lung-RADS, Fleischner and BTS). The color shades are proportional to their ranking position, with the best result being the most green and the worst being the most red. Any result outside of the top 3 or bottom 3 is colorless. ✗ indicates the absence of class weights, and ✓ indicates the presence of class weights, respectively, for nodules evaluated according to the Lung-RADS, Fleischner, and BTS growth guidelines.

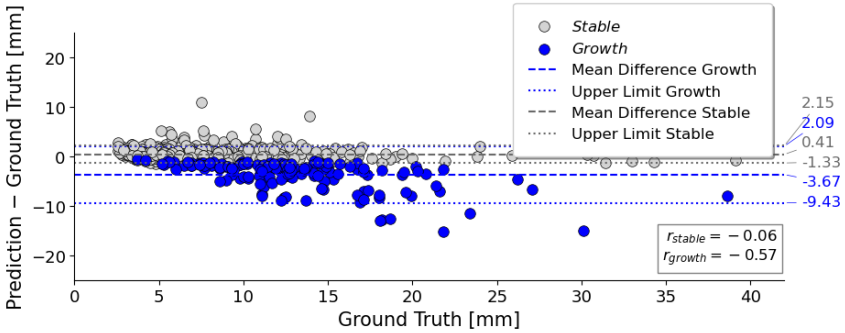
5.4.3 Bland-Altman Analysis

This subsection aims to analyse the level of agreement between the diameter prediction and the corresponding GT, by illustrating the distribution of prediction errors for all nodules diameters across various experiments. To achieve this, Figure 5.8 shows a modified Bland-Altman plot (*i.e.* the nodule diameter difference between the prediction and the GT *vs.* the GT nodule diameter). The experiments aim to forecast the diameter after one year.

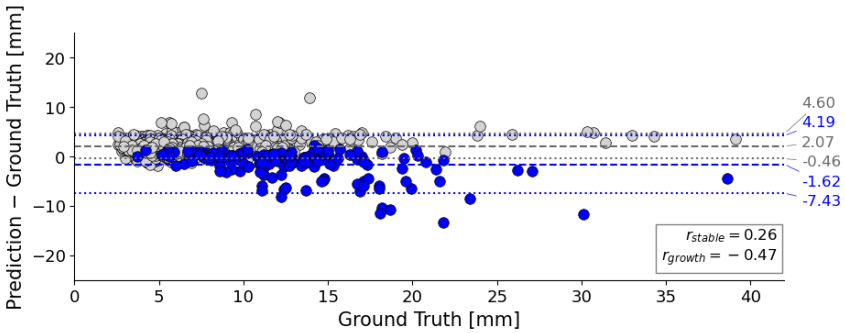
The four four Bland-Altman plots are obtained by varying the input data (tabular or image) and the use of class weights (without class weights or with the class weights for Lung RADS. Each sample is color-coded, with grey indicating stable nodules and blue for the growing nodules according to the Lung-RADS definition.

Three horizontal lines, with their respective values, are represented for each group of nodules (stable and growing), providing key statistical parameters: bias (or mean difference) and the confidence limits, which are calculated as the mean difference ± 1.96 times the standard deviation, referred to as the upper limits in the plot. Additionally, the Pearson correlation coefficients for the two group of nodules, r_{stable} and r_{growth}, are shown in the lower right corner of the plot.

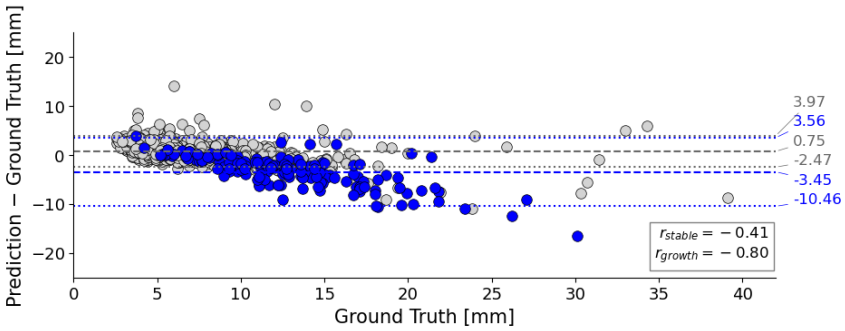
In Figure 5.8, the use of tabular data as input often leads to systematic errors, typically characterized by a constant average bias. For stable nodules, this bias may consistently be slightly positive by a few millimeters, whereas for growing nodules, it may be more negative. Also, prediction errors from image data inputs are more dispersed. Notably, both input types show a negative correlation between the magnitude of absolute error and nodule diameter — as the diameter increases, there is a greater underforecast of the diameter. This correlation is more pronounced in models that use image data.



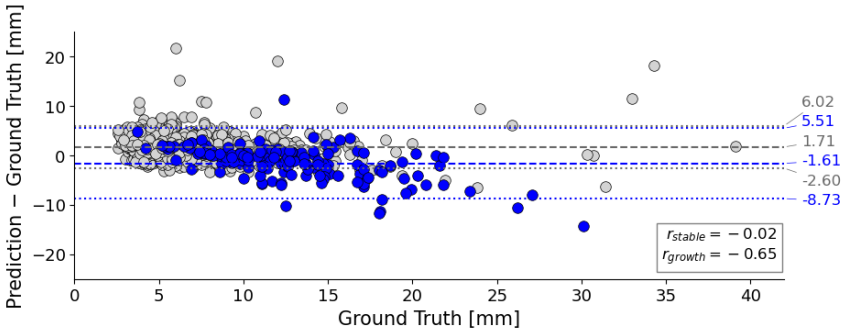
(a) **Input:** Tabular Data; **Class Weights:** None.



(b) **Input:** Tabular Data; **Class Weights:** Lung-RADS.



(c) **Input:** Image; **Class Weights:** None.



(d) **Input:** Image; **Class Weights:** Lung-RADS.

Figure 5.8: Bland-Altman plots of the deviation between the obtained results and GT, as a function of the nodule diameter [mm] and the growth criteria according to Lung-RADS. Pearson correlation coefficient *r* between the prediction error and GT is shown for both stable and growing nodules.

5.4.4 Feature Importance Analysis

This section aims to illustrate which features have the most significant impact on forecasting the diameter of a nodule. To achieve this, each feature was individually randomized to assess its influence on the model's performance. By substituting each feature with a random value, the output value is directly affected, resulting in variations between the outputs with and without randomization for each set of input data. These variations are quantified as the mean and standard deviation of the output deviations in the test set for each feature, as shown in Table 5.9. Four experimental setups are presented: predictions made over a one-year period, experiments using only tabular data, experiments combining image and tabular data, and the impact of using or not using Lung-RADS class weights in the training process.

All tabular features in Table 5.4 and the image component are used in Table 5.9. The features are ordered from top to bottom and then from left to right sequentially listing the features with the highest values, with and without randomization. The largest differences would represent features with more importance in the decision, namely the image data, nodule diameter [mm] at the initial time point, nodule type, Body Mass Index (BMI), PYs, and the number of years since quitting smoking.

Table 5.9 emphasizes the impact of the nodule diameter feature [mm] at the initial time point when using the model with tabular data as input. The use of class weights, namely Lung-RADS, reduces the influence of the feature mentioned above, leading to an increase in the standard deviation of all other features. Noteworthy is the influence of the image component when combined with tabular data, with the remaining features having almost negligible effect.

Table 5.9: Mean and standard deviation over the total range of variation for all features.

	Input	*D*		*I&D*	
	Lung-RADS *weights*	✗	✓	✗	✓
Feature	Image			7.89 ± 2.92	9.44 ± 4.24
	Diameter	33.19 ± 3.64	2.66 ± 3.05	0.15 ± 0.15	0.09 ± 0.67
	Solid	1.02 ± 0.34	1.53 ± 1.05	0.02 ± 0.17	0.12 ± 0.68
	Part-Solid	1.02 ± 0.11	0.63 ± 0.10	0.17 ± 0.14	0.25 ± 0.64
	GGO	0.98 ± 0.21	0.47 ± 0.12	0.11 ± 0.16	0.09 ± 0.67
	BMI	0.21 ± 0.10	1.08 ± 0.27	0.03 ± 0.17	0.07 ± 0.67
	Perifissural	0.79 ± 0.04	0.07 ± 0.18	0.16 ± 0.16	0.23 ± 0.64
	PYs	0.44 ± 0.07	0.34 ± 0.15	0.02 ± 0.17	0.39 ± 0.62
	QuitYears	0.11 ± 0.04	0.34 ± 0.14	0.08 ± 0.17	0.08 ± 0.67

	Input	*D*		*I&D*	
	Lung-RADS *weights*	✗	✓	✗	✓
Feature	SmokeHouse	0.06 ± 0	0.07 ± 0.10	0.08 ± 0.17	0.36 ± 0.65
	Age	0.13 ± 0.06	0.12 ± 0.07	0.08 ± 0.17	0.15 ± 0.67
	CancerFamily	0.08 ± 0.03	0.11 ± 0.10	0.11 ± 0.16	0.10 ± 0.67
	WorkType	0.03 ± 0	0.13 ± 0.05	0.07 ± 0.16	0.14 ± 0.67
	LungDiseases	0.03 ± 0.03	0.15 ± 0.06	0.12 ± 0.17	0.06 ± 0.67
	Alcohol	0.07 ± 0.04	0.12 ± 0.07	0.03 ± 0.17	0.11 ± 0.66
	OtherCancer	0.02 ± 0.04	0.13 ± 0.11	0.06 ± 0.17	0.08 ± 0.67
	SmokeWork	0.01 ± 0	0.13 ± 0.05	0.05 ± 0.16	0.08 ± 0.67

Label: ✗ indicates the absence and ✓ indicates the presence of weights based on Lung-RADS guidelines in the training of methods designed to better address growing nodules.

5.5 Discussion

5.5.1 Growth Regression

The results in subsection 5.4.1 offer insights into the influence of various factors on the predictions. Forecasting for one year results in better performance, likely due to the shorter time horizon being more predictable and less complex than forecasting for two years. Additionally, approximately 2.5 times more data is available. The similarity of results between the adaptive and dedicated approaches indicate that the model is robust in considering multi-year data, rendering it a viable option in real-world scenarios. Additionally, the adaptive approach offers computational efficiency, as it allows forecasting across multiple years without requiring separate models for each time period, thereby avoiding the need for additional computational weights and reducing both computation time and memory usage.

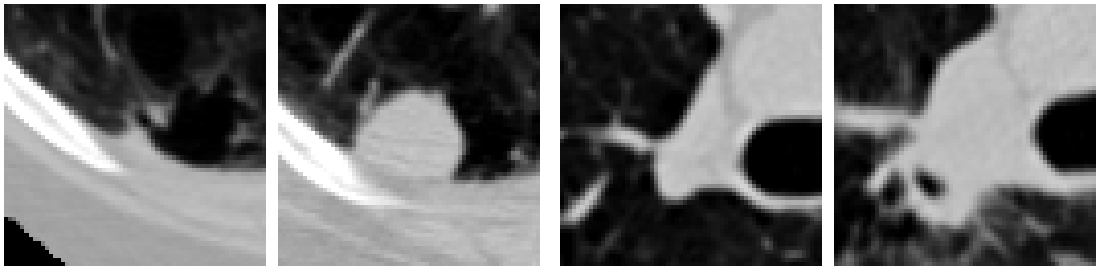
Regarding the input data, the best performance was achieved using tabular data, likely attributed to incorporating the diameter at the starting temporal point as a feature. Since this study's dataset is predominantly of small and stable nodules, which exhibit marginal growth variations, incorporating minor marginal differences in the future temporal point compared to the starting point enables predictions that are generally close to the desired result. However, it is important to note that this alignment between the input data and dataset characteristics is somewhat artificial and relies on the specific test dataset for producing reliable results. Conversely, relying solely on the image component resulted in a performance decline, which may suggest the need for more of the 2D axial medial axis image of the ROI surrounding the nodule for accurate prediction. The absence of information about the nodule's diameter size or segmentation in the input images limits the model's capacity to capture crucial features necessary for growth forecast. It should be noted that a wide range of nodules, from 3 mm to over 40 mm, needs to be dealt. Extracting meaningful features from the images requires the model to perceive the nodule's location accurately, distinguish it from surrounding anatomical elements, and leverage relevant clues to comprehend the growth trend and value. Combining image features with tabular data yields less accurate results than using tabular data alone, but this combination outperforms using image alone. It seems to balance both modalities' limitations and advantages, resulting in an intermediate performance.

Using class weights did not significantly improve the overall performance. Although the class weights were designed to benefit the classes of nodules that the model initially struggled to predict due to the dataset's nature, their introduction can potentially deteriorate the overall performance by improving predictions for underrepresented classes at the expense of overrepresented ones, leading to a trade-off between class-specific accuracy and overall model balance. Conversely, the performance tended to be worse when the weight was based on a guideline that categorized fewer nodules as growing. It is known that Fleischner categorizes fewer nodules as growing, followed by Lung-RADS and BTS, resulting in poorer results for the first and worst for the last guideline, respectively, aligning with the trend.

Figure 5.9 shows a selection of poor results to facilitate a critical analysis and explore potential explanations for these outcomes. Using solely image data as input, the results are derived from a

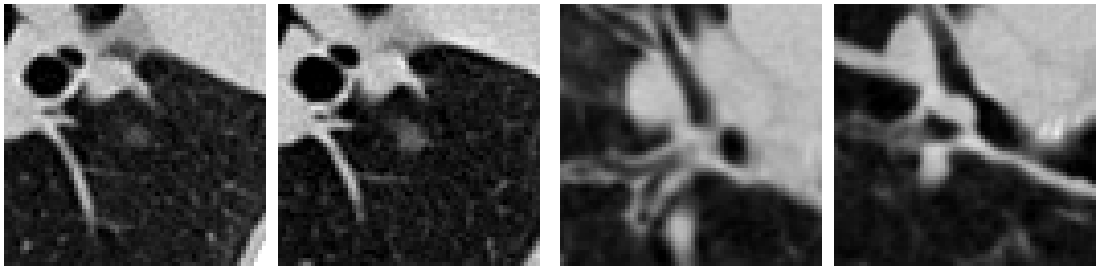
dedicated model specifically designed to forecast one year. Lung-RADS weights are used during the training process.

The forecast fails in different scenarios, exhibiting two distinct types of errors. On the one hand, there are cases where the prediction indicates stability when the GT corresponds to significant growth (5.9a and 5.9b). On the other hand, there are instances where the prediction suggests significant growth when the GT represents stability (5.9c and 5.9d).



(a) **Y0:** 4.80 mm; **GT:** 18.10 mm; **Prediction:** 6.75 mm; **Difference:** −**11.35** mm.

(b) **Y0:** 14.40 mm; **GT:** 30.10 mm; **Prediction:** 15.89 mm; **Difference:** −**14.21** mm.



(c) **Y0:** 5.40 mm; **GT:** 6.20 mm; **Prediction:** 21.45 mm; **Difference:** **15.25** mm.

(d) **Y0:** 4.80 mm; **GT:** 6.00 mm; **Prediction:** 27.61 mm; **Difference:** **21.61** mm.

Figure 5.9: Examples of the ROIs of the nodules corresponding to some of the worst one year forecast results achieved using the network trained with class weights based on Lung-RADS guidelines, which specifically receives image data (I).

The proximity of the nodules to other anatomical structures appears to significantly impact the forecast quality. The best results are observed for isolated nodules (Figure 5.7), while the worst results are obtained for nodules primarily near the lung walls (Figure 5.9). The limited information regarding size and segmentation in the input images further hampers the model's accuracy in identifying nodules. These findings underscore the inherent challenges in accurately perceiving the location and boundaries of nodules, which affects its ability to extract precise features for forecasting future diameter values.

From another perspective, there does not appear to be a discernible relationship between nodule diameter, texture, and performance. Small and large nodules and nodules with different textures are present in both the best and worst prediction results.

5.5.2 Analysis of the Classes with Malignant or Growing Nodules

Nodules with clinical significance should have diameter forecasts that are as accurate as possible to avoid underestimating potential treatment or follow-up needs. It has been observed that for malignant nodules and nodules exhibiting growth, applying class weights during training is beneficial for diameter forecasting. According to Table 5.6, using Lung-RADS weights yields the most favorable results for malignant nodules, likely due to the substantial overlap between these nodules and those classified as growing according to Lung-RADS. Table 5.8 demonstrates that the results are better for groups with more nodules classified as grown, favoring representativeness and possibly simplifying the prediction difficulty. Therefore, more favorable results are obtained for the BTS group, followed by Lung-RADS and Fleischner.

Regarding the forecast year, the results align with the previous findings. Specifically, the one year forecast surpasses the two year forecast and the adaptive approach often exhibits comparable or better performance compared to the dedicated approach. Training with a larger dataset may be preferable regardless of the potential difference in the target variable. This could explain why the results of the adaptive approach are highly competitive, despite accommodating different forecast years.

Minimal deviations from the earlier statements are observed when considering the variations in input data. Tabular data remains preferable in most cases to the use of images, with explanations similar to those in the preceding subsection. However, in the case of growing nodules, there are instances where combining the image component with the tabular data yields better results compared to using tabular data alone. This could be attributed to tabular data's tendency to underestimate the nodules' future diameter value, while the image component provides a valuable balance for the forecast.

Table 5.7 offers contextualization for the results. Presently, a VDT of 348 days (Heuvelmans et al., 2017a) is commonly considered for malignant nodules, but some of the proposed experiments have the potential to offer a closer approximation to actual future nodule size values. These findings underscore the viability of the recommendations presented here for practical real-world implementation.

Nevertheless, the analysis in this subsection may have bias. For instance, class weights are employed to counteract biases from imbalanced datasets, enhancing the model's ability to learn about clinically significant yet underrepresented nodule types, such as malignant or growing ones. However, using class weights and specific data representation strategies may lead to overfitting. Models might perform exceptionally well on the emphasized nodule types during training, compromising their ability to generalize across a broader range of nodule types. This could result in models overly tailored to certain classes, overlooking the complex and varied patterns of nodule growth. Therefore, while using class weights is a viable solution to dataset imbalance, it requires thorough consideration and validation. The experiments could have been tested on an external dataset, but this validation is always challenging due to the difficulty of getting longitudinal data.

5.5.3 Bland-Altman Analysis

Figure 5.8 highlights the challenge of accurately forecasting nodules that exhibit growth, especially when using tabular data as input. This may be attributed to the previously discussed issue of incorporating the diameter at the initial temporal point as a feature, which could result in minimal growth variations imposed on the predictions. However, it is important to note that using class weights helps to evenly distribute prediction errors without displaying a clear pattern, as they occur in stable and growing nodules.

The significant prediction errors observed using image data can be attributed to the inherent challenges in accurately extracting pertinent features and capturing the intricacies within the dataset. This challenge is particularly pronounced as nodules increase in diameter, as larger nodules exhibit heightened complexity and variability in appearance. Such correlation between nodule size and prediction accuracy underscores the need of integrating size-dependent factors into regression models for lung nodule growth, demanding for further exploration of innovative methodologies in this field.

5.5.4 Feature Importance Analysis

The study's findings indicate that specific features significantly impact the accuracy of nodule diameter forecast. Notably, the image data, initial nodule diameter, nodule type, BMI, PY, and years since quitting smoking were identified as the most influential factors. It is important to note that these features are consistently referenced in the literature, underlining their relevance in the regression task. Although other features were considered, their impact in this study was insubstantial.

When the models were trained using Lung-RADS class weights, the influence of the initial nodule diameter was attenuated. This observation suggests that the class weights play a role in reducing the dominance of the diameter feature in the regression process. Nevertheless, it is essential to acknowledge that using class weights also increases standard deviation across all features. This phenomenon implies that the models may become less resistant to noise or exhibit a broader range in their predictions, potentially introducing a degree of uncertainty.

Furthermore, the results underscore the impact of image data on prediction accuracy. An analysis of models that integrate image and tabular data reveals that other features derived from the tabular data have minimal influence on the prediction outcomes. This observation implies that combining tabular data with image data leads to an imbalance in feature representation. This arises because the number of neurons originating from the image component equals the sum of all other features represented in the tabular data. Consequently, the image data has a more pronounced influence on the regression method.

5.5.5 Comparability

This subsection compares the results presented throughout the chapter to the state of the art. Table 5.10 provides a reference for comparing nodule diameter forecasts at future time points across various methods, despite differences in time frames, datasets, learning challenges, and the number of nodules used for training and testing. Notably, Wang et al. (2022) was the only one who used the same database as in this study, NLST, but they used only 265 nodules, an incomparably smaller pool of nodules than the one presented in this chapter.

Table 5.10: Comparison of results with the state of the art.

	Proposed Approach				
	Full Dataset				
	D, W/o *weights*	*I*, W/o *weights*	*D*, Lung-RADS *weights*	*I&D*, W/o *weights*	*I&D*, Lung-RADS *weights*
MAD [mm]	0.99	1.72	2.13	1.56	2.62
RMSE [mm]	1.79	2.65	2.58	2.62	3.44

	State of the Art		Proposed Approach	
			1/2 Dataset	
	Rafael-Palou et al. (2022)	Wang et al. (2022)	*D*, W/o *weights*	*I*, W/o *weights*
MAD [mm]	1.74		0.99	2.14
RMSE [mm]		11.32	1.79	2.97

Compared to Rafael-Palou et al. (2022), the results are of the same order of magnitude. One of the main reasons for the better performance of the method presented here is likely the presence of a target for regression, whereas the other authors attempted to obtain future segmentation or images of nodules. This may have led to models that were not carefully designed to achieve reliable size-related results.

In the literature, the challenge was different, either by attempting to forecast the image or segmenting a nodule in a future temporal point, which may bring some level of explainability. However, this study conducted an exhaustive analysis to discern the pivotal features contributing to the decision-making process. Moreover, particular emphasis was placed on scrutinizing the results on malignant and growing nodules. These aspects hold significant importance in nodule growth-related matters and often remain underrepresented in the state of the art.

Beyond comparison with the state of the art, this chapter also details the impact on the performance of reducing the training dataset by half while maintaining the same test set, as illustrated in the last two columns of Table 5.10 across two distinct experiments. Contrary to findings from the literature, such as those by Tang et al. (2022), which suggest that image data (unstructured) contains richer information than tabular data (structured), the results showed no significant difference when using tabular data, whereas performance was worse using image input data when the training data was reduced. This discrepancy may stem from the unused complexity of image

data, in contrast to structured data, which, being pre-processed, is already closer to unlocking its potential utility. In this context, structured data is analogous to "engineered features", providing valuable context that image data lacks. On the other hand, "statistical features" are derived through Computer Vision (CV), suggesting that reliance on statistical features might reduce the required training data. It is expected that with dataset expansion, the efficacy of statistical features could converge with or even surpass that of engineered features.

5.5.6 Limitations

The results presented in this chapter demonstrate competitive performance with the state of the art. However, there are undeniable limitations. One of the most significant limitations is the potential impact of interobserver and intraobserver errors on the model's performance. The study by Revel et al. (2004) showed that in a cohort of 54 nodules under 20 mm, a minimum intra-reader measurement error of 1.32 mm and a minimum interobserver measurement error of 1.73 mm were observed. Notably, if the Lung-RADS criterion considers growth from 1.5 mm, smaller than the interobserver discrepancy, this discrepancy could lead to the misclassification of nodules as stable when they might grow, and vice versa (Heuvelmans et al., 2017b). In this sense, the Fleischner criteria may be more robust, given that the value of growth is greater than the errors mentioned here. It is worth noting that cumulative errors in this study, in measuring the diameter at the initial and forecasted times, may affect the interpretation of the results. Nevertheless, it is encouraging that the best results surpass the interobserver error by more than double.

Another limitation of the database is that the exact exam dates are unavailable. Only the year of examination is provided. This lack of precise temporal information may result in variations of a few months, which can impact the regression results.

It is important to recognize another constraint, namely, the consideration of the equivalent diameter and the use of the medial slice when using images as input. It remains to be seen to what extent the equivalent diameter is a suitable measurement for prediction and whether the chosen imaging plane is optimal for these purposes. Alternatively, exploring segmentation or volume-based approaches and adopting a 2.5D/3D approach might have yielded potential advantages.

Finally, this study did not specifically analyze the impact of sex or gender on the characteristics of nodules among the 3,054 patients (57.07% men and 42.93% women), which may limit the generalizability of the findings across different demographic groups.

5.6 Concluding Remarks

This chapter presents a novel approach for forecasting the growth of lung nodule diameters through a regression model that integrates diverse data sources, including both image and tabular data. This method appears to be the first regression-focused study in this area, diverging from previous works that, focus on forecasting future images or segmenting nodules at subsequent time points, aiming for more visual and interpretable objectives. The proposed model allows for flexible forecasts at multiple time points, providing precise estimations one or two years from the initial input data.

Particular emphasis was placed on scrutinizing the results for malignant and growing nodules, which are the ones that follow-up procedures must monitor or treat. This study has an exhaustive analysis to discern the pivotal features contributing to the decision-making process. The model addresses the guidelines of Lung-RADS, Fleischner, and BTS by incorporating class weights using growth criteria from these guidelines in training, thereby prioritizing and driving predictions toward higher growth values.

The recommended model configuration for implementation likely involves an adaptive model that forecasts multiple years with minimal computational expense. It incorporates both tabular and image data, along with class weights, to enhance forecasts for growing nodules. These experiments aligned closely with the study data, resulting in more accurate predictions than the consideration of 348 days of VDT for malignant nodules, a value commonly recommended in the literature (Heuvelmans et al., 2017a).

The developed predictive model is optimized for integration into a clinical workflow after detecting, segmenting, and classifying nodules. By providing accurate growth forecasts, it establishes a sustainable basis for rethinking current follow-up intervals. This approach allows medical doctors to anticipate changes in nodule growth effectively, essentially offering a predictive view into the future. Given lung cancer's low survival rates, the capacity for precise nodule growth prediction represents a significant breakthrough.

Chapter 6

Lung Nodule Image Annotation from Medical Reports

This chapter involved a collaboration with a Natural Language Processing (NLP) group at INESC TEC and radiologists from São João Hospital and Gaia/Espinho Hospital. In both cases, the thesis author was the main contributor and was responsible for the conceptualization of the ideas. The NLP experiments from this chapter benefited from the involvement of the aforementioned group.

Two contributions, corresponding to **III. (a)** and **III. (b)**, are seamlessly integrated into this chapter. The first contribution[5], a journal paper, outlines how pulmonary nodules descriptions in medical reports were processed and made available to the scientific community in a tabular format through the LNDb_v4 dataset. Additionally, it compares the manual annotations in the dataset with the textual descriptions of these nodules in the reports. This dataset version enabled the development of an automated method for the second contribution[6], presented in a submitted patent, for using medical texts to reconstruct the location of pulmonary nodules in Computed Tomography (CT) images. This contribution serves as a proof of concept and allows medical doctors to accurately identify the nodules mentioned in a prior CT scan during follow-up or facilitates a better comparison of the findings in a new CT scan.

The introduction (6.1) and related work (6.2) are the first sections of the chapter. The materials (6.3) section address the first contribution, with the description of the LNDb_v4. Then, the methods (6.4) and results (6.5) sections focus mainly on content derived from the second contribution, detailing how lung nodules in medical reports can be automatically mapped to annotations in CT. The concluding remarks (6.6) complete this chapter.

[5] **Carlos A Ferreira**, Célia Sousa, Inês Dias Marques, Pedro Sousa, Isabel Ramos, Miguel Coimbra, and Aurélio Campilho. LNDb v4: pulmonary nodule annotation from medical reports. *Scientific Data*, 11(1):512, 2024. doi: 10.1038/s41597-024-03345-6.

[6] **Carlos A Ferreira**, Alexandre Ferreira, Miguel Coimbra, Nuno Guimarães, Alípio Jorge, and Aurélio Campilho. RePick Annotator. *European Patent (Pending)*, 2024.

6.1 Introduction

Radiologists play a crucial role in lung cancer follow-up by meticulously examining and annotating CTs to identify pathological findings, such as lung nodules or masses. Nevertheless, the search for nodules is an arduous, error-prone, and time-consuming process due to the intricate three-dimensional nature of CTs. Furthermore, nodules exhibit distinct textures and shapes, adding complexity to the detection process. These different characteristics, along with the similarity of nodules with other anatomical structures, contribute to scanning errors ($\sim$30%), recognition inaccuracies ($\sim$25%), and decision-making difficulties ($\sim$45%) (Del Ciello et al., 2017). To address these issues and provide additional diagnostic support, Artificial Intelligence (AI) solutions are being considered for integration into clinical practice. However, the systematic development of such solutions encounters obstacles related to data quantity, representativeness, and the extensive annotation required for algorithm training.

Medical practice gives us an opportunity to gather more data to train AI algorithms, but observed findings are typically documented only in written medical reports, and not conveniently as marked spatial Regions of Interest (ROIs) associated with the images themselves. This is understandable since for medical doctors, free text is an easy way to record and access medically relevant information. Having access to the marked spatial ROIs would not only benefit the training of AI algorithms, but also other healthcare professionals who need to review these reports at a later stage, who may find this information incomplete or inconsistent (Qadan et al., 2019). If a radiology report omits a key finding or fails to effectively communicate the importance of a reported finding, subsequent clinical decisions regarding management or future follow-up may be affected (Bruno et al., 2015; Onder et al., 2021).

For future reports, the marking of spatial ROIs could be promoted using specialized software. This approach allows for more precise documentation of findings but comes with a higher temporal cost and requires learning new tools. However, even if these difficulties are mitigated, it does not address the issue of annotating past reports. An alternative method could involve extracting annotations from existing medical reports using text mining, followed by mapping this information onto the corresponding regions in CT images. Currently, fully labeled datasets for AI training are available, such as CheXpert (Irvin et al., 2019) and DeepLesion (Yan et al., 2018), as well as image datasets with manual annotations, like LIDC-IDRI (Armato III et al., 2011).

In 2019, the Lung Nodule Database (LNDb) was released (Pedrosa et al., 2020a), comprising of chest CT images with manual nodule segmentation and texture characterization. Additionally, medical reports for the same exams were collected as part of this study. The information related to the nodules in these medical reports has been gathered, structured, described and shared with the scientific community through the LNDb_v4 dataset (Pedrosa et al., 2023) presented in this chapter. The information from these reports was compared to the manually annotated findings to assess the matches and mismatches between the two data sources. To the best of the author's knowledge, this is the first time where manual and report-based annotations have been made available and their correspondence has been analyzed, particularly in the context of CT scans. The aim of sharing

this data is to encourage the community to consider the reuse of data such as medical reports, and to highlight the accuracy, consistency, and potential biases associated with each annotation source.

In addition to the description, validation, and comparison of data from medical reports, this chapter also presents a methodological contribution that allows reassessing findings in images based on medical reports, specifically nodules in CT scans. This approach supports the goal of revisiting and re-evaluating pulmonary nodules, significantly increasing the number of available image annotations, as long as corresponding pairs of medical reports and images are available.

In summary, prior to this thesis, the dataset known as LNDb was already available, consisting of CT images with lung nodules manually annotated by radiologists. Two main outputs were developed based on this dataset. First, a new version of the dataset, LNDb_v4, was created by cross-referencing the manually annotated nodules with descriptive information from the corresponding medical reports for all CT scans, resulting in a newly filtered set of image nodule annotations. Second, once the data was validated, an automatic method was developed to obtain, directly from the medical reports, the image annotations, streamlining nodule identification for radiologists during follow-up.

6.2 Related Work

In recent years, the development of automated systems for annotating pathological findings in medical images has harnessed significant attention, particularly through the use of information extracted from medical reports. This process typically involves analyzing textual reports written by radiologists and mapping the extracted information to corresponding ROIs within medical images, such as CT scans. The primary motivation behind these approaches is to enhance the efficiency and accuracy of medical image analysis, especially in scenarios where manual annotation of large datasets is both labor-intensive and prone to human error.

Several modular methodologies have been proposed to bridge the gap between textual information in medical reports and spatial annotations in images. Modular approaches often begin with text mining algorithms that extract detailed characteristics of the findings, such as size, shape, and location. These characteristics are then correlated with image data through computer vision techniques. Examples of such modular systems include BioBERT (Lee et al., 2020), ClinicalBERT (Huang et al., 2019a), and Generative Pre-trained Transformer (GPT)-3 (Brown et al., 2020), which can be tailored for biomedical text mining, as well as Convolutional Neural Network (CNN) and You Only Look Once (YOLO)-based models (Redmon et al., 2016) that process image data separately. By integrating these specialized algorithms at each step, such systems allow for the improvement of individual components without requiring an overhaul of the entire system. One advantage of these modular approaches is that they can be optimized incrementally and often require less data compared to end-to-end models, making them more accessible for specific tasks where data availability might be limited.

A crucial advancement in this area is the concept of zero-shot learning (Xian et al., 2018), where models are designed to perform tasks they were not explicitly trained on, enabling them to

generalize more effectively to new, unseen data. Zero-shot learning models can map textual information directly to image data in an end-to-end manner, bypassing intermediate steps (Palatucci et al., 2009). This approach reduces the need for task-specific datasets and allows models to handle a broader range of scenarios with minimal retraining, making them particularly useful in dynamic clinical environments. Additionally, zero-shot learning enhances the model's ability to adapt to novel cases, which is essential in medical contexts where unexpected conditions frequently arise.

A notable advancement in the field is the Segment Anything Model (SAM) (Kirillov et al., 2023). Designed for promptable segmentation, where prompts (user interactions that can range from simple points to free-form text) guide the segmentation process, SAM leverages transformer-based architectures to process both the prompts and image data. These transformers work through multi-head self-attention mechanisms, allowing the model to focus on different parts of the input data simultaneously, capturing various aspects of the image that might be relevant to the prompt. Mathematically, the prompt is encoded into a vector, which is then integrated into the transformer's attention mechanism. This vector influences how the attention heads allocate their focus across the image, helping to direct the model's attention to the areas that are most relevant to the prompt. This process is supported by a mask decoder, which refines the segmentation output based on the attention information, ensuring that SAM can accurately delineate ROIs, even in complex and previously unseen scenarios.

Another significant development is Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021), shown in Figure 6.1. CLIP works by encoding both text and images into high-dimensional vectors using a dual-encoder architecture, where one encoder processes text and the other processes images. The model then employs a contrastive learning approach, which aims to maximize the similarity between correct pairs of text and image vectors while minimizing the similarity between incorrect pairs. This contrastive method ensures that CLIP can effectively match textual descriptions with relevant image regions, making it particularly powerful for detailed and context-aware tasks, such as identifying specific pathologies in medical images. CLIP's ability to understand and align complex visual and textual information without the need for task-specific retraining sets it apart as a versatile tool in the integration of AI in medical imaging.

Building on these innovations, YOLO-World, introduced in early 2024, represents a major advancement in real-time open-vocabulary object detection (Cheng et al., 2024). This model extends traditional YOLO models by incorporating a vision-language modeling approach, allowing it to detect objects that were not explicitly labeled in the training data. YOLO-World employs a new architecture called the Re-parameterizable Vision-Language Path Aggregation Network (RepVL-PAN), which integrates transformer layers to enhance the interaction between visual features and linguistic inputs. The model processes prompts (user-provided textual cues) alongside image data, dynamically adjusting the detection process based on the textual input. Specifically, the prompt is encoded into a vector that influences the attention mechanism within the transformer layers, guiding the model to focus on relevant regions of the image that correspond to the prompt. Unlike SAM, which is designed primarily for segmentation tasks, YOLO-World is optimized for object detection, meaning it can quickly and efficiently locate and identify specific objects within an im-

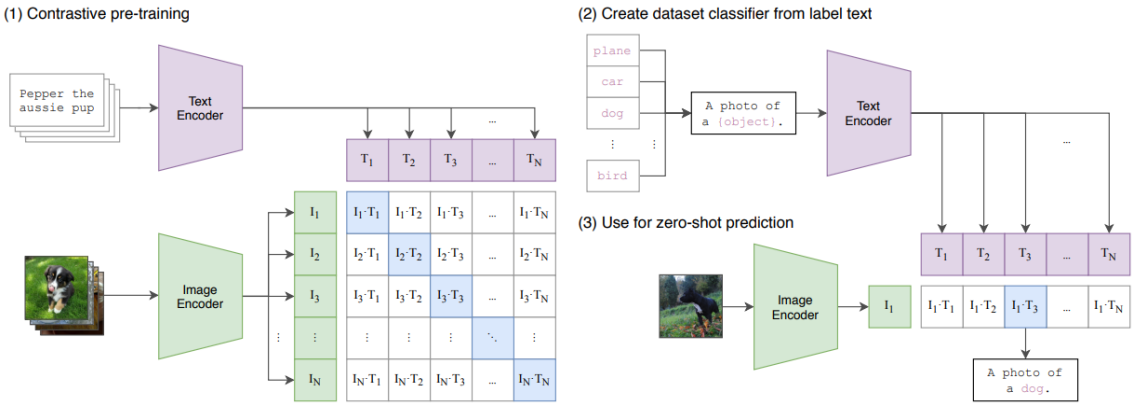


Figure 6.1: Summary of the CLIP approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image I_N and text T_N) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes. This image was extracted from Radford et al. (2021).

age based on the prompt. This approach combines the efficiency and speed of YOLO models with the flexibility of language-based inputs, making it highly adaptable to various medical imaging applications, especially those requiring rapid and precise detection of anomalies in complex images.

Interestingly, while these methodological advancements are focused on annotating images based on text, automatically generating textual reports from images has been more commonly explored in the literature (Rio-Torto et al., 2024; Kayser et al., 2022; Hou et al., 2021). This focus likely arises from the fact that generating text from images aligns more closely with existing clinical workflows, where summarizing findings in a report is a routine task. Moreover, the validation of such systems is often seen as more straightforward, as the generated reports can be directly compared to existing clinical reports, providing a clear benchmark for accuracy and reliability. However, while these systems can reduce the workload of medical doctors, there is often resistance to fully automated solutions due to concerns regarding accuracy and interpretability. Consequently, these systems are often viewed as decision-support tools rather than replacements for human judgment, complementing rather than replacing the role of physicians.

Despite these advancements, several challenges persist in the field. One of the primary issues is the variability in medical reports, which may differ significantly in structure, terminology, and level of detail across institutions or even among different radiologists. This variability complicates the task of creating generalized models that can consistently interpret and annotate findings. Additionally, there is a pressing need to refine these models to better handle the complexities of medical language, which often includes ambiguous terms, abbreviations, and context-dependent meanings. Addressing these challenges is essential for advancing the adoption of automated annotation systems in clinical practice and ensuring they can be effectively integrated into diverse healthcare environments.

6.3 Materials

The LNDb database, developed within the LNDetector project[¶], was used. The project aimed to create a comprehensive lung cancer screening system, incorporating modules for AI-based detection, segmentation, classification, and nodule follow-up to aid in pulmonary cancer management. Throughout the clinical study for data collection, hospital protocols remained unchanged. Radiologists and various physicians involved in the process performed their routine clinical practice. A CT scan was deemed eligible if the slice thickness was 1 mm.

All data acquisition procedures were conducted in accordance with the approval granted by the São João Hospital Ethics Committee. It should be noted that data collection was done prospectively, with a guarantee of anonymization. Among the 292 patients scanned between 2016 and 2018, 55.8% were male. The ages ranged from 19 to 98 years, with a median age of 66.

During the medical reporting process, radiologists routinely describe the thoracic findings without specifying spatial coordinates or considering potential inclusion in a clinical study. The exact number of distinct radiologists who contributed to the text reports across the dataset is unknown. In the LNDb database, five radiologists manually annotate nodular findings. Each CT scan received between one and three annotations from different radiologists. Guidelines also requires the annotation of non-nodules, which are findings closely resembling nodules but not actual nodules themselves.

Upon the release of the dataset, the LNDb Grand Challenge was launched[§] (Pedrosa et al., 2021), encouraging the community to make proposals for nodule detection, nodule segmentation, texture classification, and follow-up recommendation according to Fleischner guidelines (MacMahon et al., 2017). The imaging data and manual annotations is described in Pedrosa et al. (2019) and stored on the Zenodo platform (Pedrosa et al., 2020a). Data is divided into 5 folds, with one of them reserved for testing. Contributions for evaluating results on the test set are still being accepted.

This section outlines the process of describing lung nodules from medical reports in a tabular format and locating them spatially by cross-referencing with manually annotated image data (Pedrosa et al., 2023), resulting in LNDb_v4 (Ferreira et al., 2024b), a new version of the LNDb.

Due to the temporal gap between the textual reports and the manual annotations, they do not influence each other, and no definitive Ground Truth (GT) exists. In order to release the structured descriptions of lung nodules mentioned in the medical reports and evaluate their match with the previously released manual annotations, a data science methodology comprising two phases was established:

- **Medical Reports Annotation**: identifies the nodule descriptions in medical reports and extracts their features into a structured file;
- **Nodule Matching**: outlines the criteria used to determine the match.

[¶] https://www.inesctec.pt/en/projects/lndetector
[§] https://lndb.grand-challenge.org/

Figure 6.2 provides a schematic representation of the phases with illustrative images for each one, which is further described in the following subsections. Underlining is used in the report to highlight various attributes. The central image illustrates how this information translates into tabular terms, while the lung image shows the nodule chosen among the possibilities based on the written description.

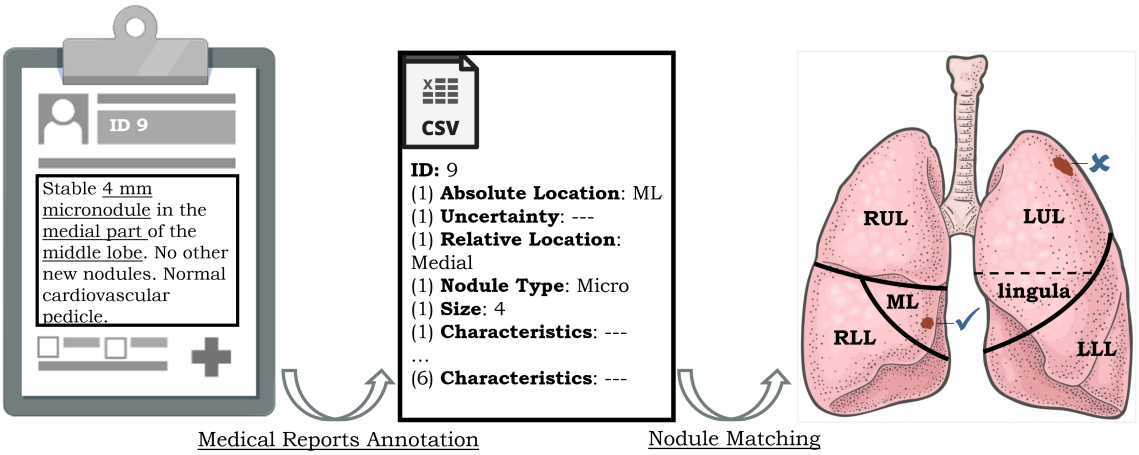


Figure 6.2: Schematic overview of the assay phases in this study: i) initially, all information related to the lung nodule is highlighted in the medical report; ii) the pieces of text are converted to their respective fields in the entities table; iii) the correspondence is verified between the two nodules that had been manually annotated, and only one of them corresponds to the Middle Lobe (ML) in terms of relative position and written characteristics, given by ✓, unlike the other one, given by ✗. The medical report was translated from Portuguese to English for better reader comprehension. CSV refers to a data file format (Comma-Separated Values), while RUL, LUL, RLL, and LLL indicate anatomical lung regions: Right Upper Lobe, Left Upper Lobe, Right Lower Lobe, and Left Lower Lobe, respectively.

6.3.1 Medical Reports Annotation

The medical reports were provided in Portuguese, the official language of the partner hospital, and were anonymized. The reports consist of a single text section that describes all observations related to atypical landmarks found in thoracic organs during the CT scans, including conditions such as cardiomegaly, emphysema, and effusions, among others. As such, the reports do not specifically focus on nodules.

Here, the term "entity" refers to text portions describing a nodule or a set of nodules with shared characteristics. The entities are characterized by the following attributes:

– **Absolute Location:** Right Upper Lobe (RUL), Middle Lobe (ML), Right Lower Lobe (RLL), Left Upper Lobe (LUL), Left Lower Lobe (LLL), upper lobes, lower lobes, right lung, left lung, lingula, and lingula/LLL.

The first five are the lobes, which are often represented with their acronyms. The lingula is a part of the LUL, and the remaining ones consist of a collection of lobes,

each designated accordingly. Figure 6.2 offers a schematic representation of the lung lobes.

– **Relative Location:** apical/superior, basal/inferior, anterior, posterior, medial, lateral, centrilobular, and justapleural/peripheral.

These references indicate relative positions. It is possible that more than one of these options may be mentioned for the same entity. The first three pairs refer to the axial, coronal, and sagittal planes, respectively. The last pair indicates the position within the lobe.

– **Lesion Type:** micronodule, nodule, mass and granuloma (*i.e.*, nodules with calcifications).

– **Size:** numerical value indicating the nodule diameter.

If the volume is provided, and considering the typical shape of the lesions, the diameter is considered as the equivalent diameter of a sphere. If multiple measurements are indicated for the size axes of the nodule, the average is taken.

– **Characteristics:** texture, calcification, internal structure, lobulation, malignancy, margin, sphericity, spiculation and subtlety.

The characteristics are extracted according to Table 2.1 (Opulencia et al., 2011). All characteristics are ranked continuously from 1 to 5, except for categorical variables like internal structure, which has 4 rankings, and calcification, which has 6 rankings.

– **Quantity:** numerical value indicating the number of nodules identified.

– **Negative:** “no”, “without” and negative verb forms.

This attribute indicates the absence of nodules and changes the interpretation of related descriptions.

– **Uncertainty:** “how many?” and “is it?”.

These situations are considered when there is uncertainty about the quantity of nodules mentioned in a sentence. The first case applies when there are nodular references in the plural form, while the second case applies when words indicating probability, such as “probably” or “maybe”, are used with the description of the nodules. It was necessary to interpret the text and both expressions of uncertainty may be found simultaneously for the same entity.

The annotation and conversion of the data into a tabular format were performed manually by the author of this thesis, who has over 7 years of experience in handling lung cancer CT data. A radiologist was consulted to validate any uncertainties that arose during the process.

An example of the manual process performed in this stage of medical report annotation is shown below. The medical reports are in Portuguese, but for clarity, the original medical report is presented first, followed by its English translation and the subsequent output of this stage:

Original Input: "`lobo médio persiste um nódulo de 7 mm. O estudo comparativo com os exames prévios mostra estabilidade do nódulo em comparação com o exame de 29-09-16 e discreto aumento das dimensões em comparação com o exame de 29-01-2016; o aumento é inferior a 10%. Sugere-se nova avaliação em cerca de 12 meses. Sem outras alterações relevantes.`"

Translated Version: "`middle lobe persists a 7 mm nodule. Comparative study with previous exams shows stability of the nodule compared to the exam of 29-09-16 and a slight increase in size compared to the exam of 29-01-2016; the increase is less than 10%. Further evaluation is suggested in about 12 months. No other relevant changes.`"

Output - in the format of the attribute type, which is underlined, followed by the extracted entity: – Size: "7mm"; Lesion Type: "nodule"; Absolute Location: "ML".

6.3.2 Nodule Matching

The nodule matching process enables the identification of matches between nodules described in the reports and those delineated in the image annotations, using specific criteria developed for this purpose. When a match is established, the spatial coordinates are obtained; otherwise, mismatching is identified regardless of whether it arises from the medical reports or the manual annotations. This process was verified automatically.

The matching process involves two phases: candidate search and matching decision. A flowchart in Figure 6.3 illustrates the process. During the candidate search, each medical report entity is matched with manually annotated nodules based on specific criteria. If any of the criteria is absent in the medical report during the candidate search phase, only the remaining criteria are checked. The criteria are as follows:

– **Absolute Location**: Converts lobe references from reports into spatial coordinates using an automated method by Hofmanninger et al. (2020), which segments lobes via a 2D U-Net model. This model achieved a Dice score of 0.97 on a slice-by-slice basis, assigning each nodule to one of the five lobes. All nodules in the LUL below the upper coordinate of the LLL are considered for the lingula reference.

– **Relative Location**: The lobe's centroid serves as a reference to verify the position of nodules described in reports along the three anatomical axes. Nodules labeled as peripheral are located within the outer quarter of each axis, near lung walls, while centrilobular nodules occupy the central half of the lobe.
– **Size**: The reported diameter was matched to the annotated equivalent diameter on the CT scan, allowing a $\pm$ 3 mm tolerance to account for inconsistencies in measurement methods, ensuring that only closely matching nodules are selected.
– **Characteristics**: Correspondence between the descriptive attributes in the medical report and the radiologist-assigned scores was assessed, allowing a $\pm$ 1 score tolerance for candidate inclusion.

In the matching decision phase, before evaluating each criterion, the number of candidates is verified. If a previously annotated nodule has already been matched with an entity from the same CT scan, it is removed from the pool of candidates. If no nodules remain, it is considered a mismatch unless the uncertainty criterion "is it?" is present. If there is more than one candidate, the next criterion is considered. If only one candidate remains, the match is considered found. The matching criteria were as follows:

– **Uniqueness**: Among the pool of annotated nodules, only those that appear less frequently as candidates for different entities are considered;
– **Agreement**: This criterion selects the findings that have a higher number of radiologists as annotators;
– **Size + Lesion Type**: If the report specifies a size, the annotated nodule with the closest size is chosen; otherwise, the largest nodule is selected. Based on data evidence, micronodules are considered up to 6 mm, as reports frequently classify them within this range.

If the uncertainty parameter "how many?" is present, the process stops at the first criterion. Otherwise, all criteria must be passed and fulfilled. Ties may occur in the first and second matching criteria in some cases, while the decision in the last criterion is made unequivocally based on the considerations. Considering uniqueness, agreement, and size in the matching decision is deemed reasonable, and it is believed that these considerations do not restrict the potential evidences from this study.

Regarding the negative attribute, it denotes the absence of a nodule or an opposing characteristic. The quantity attribute prompts the search for more than one nodule with the specified attributes described in the medical report.

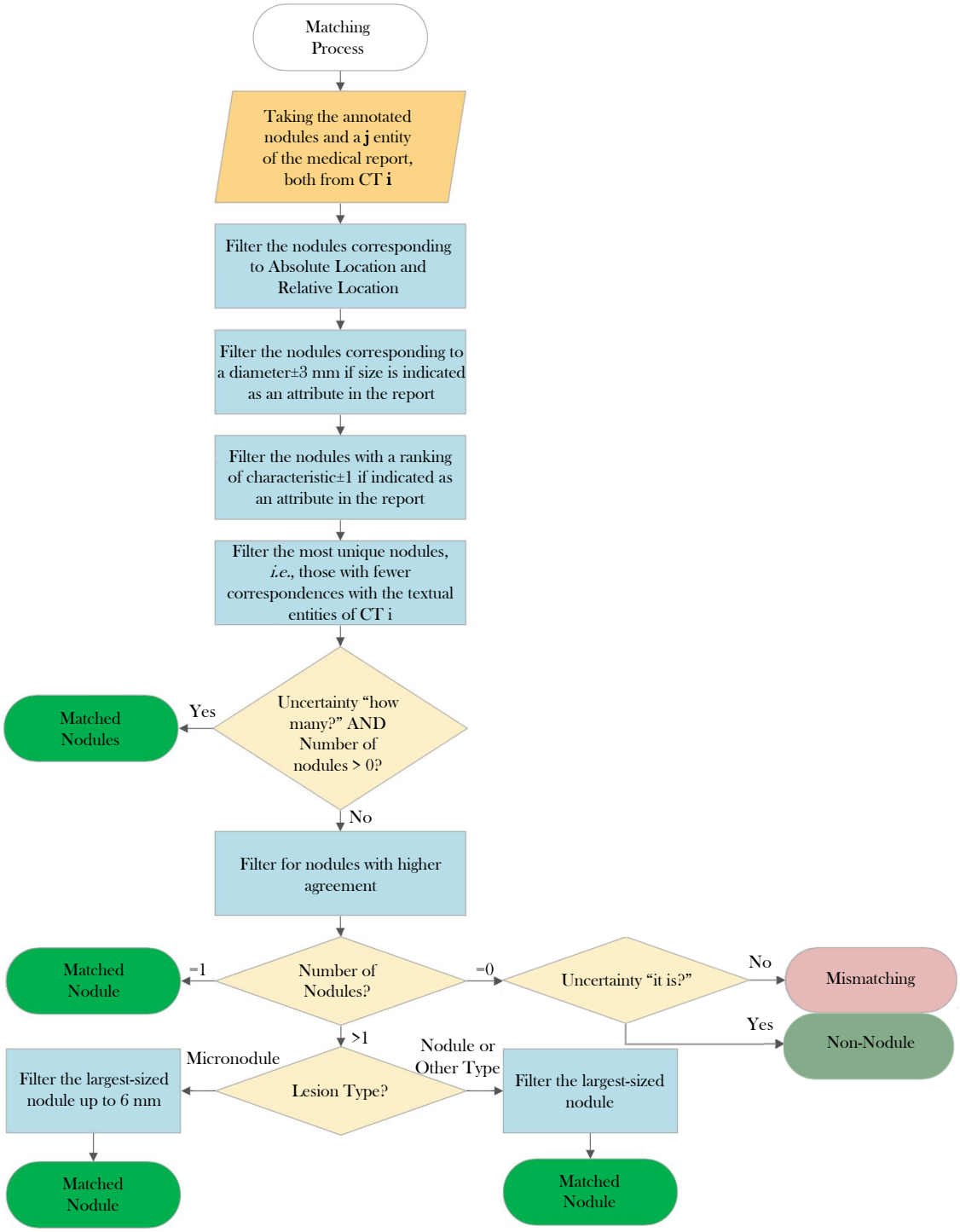


Figure 6.3: Summarized flowchart encompassing the criteria used to match the nodules.

6.3.3 Clinical Assessment

In this subsection, a radiologist was asked to evaluate the matching between the medical reports and the manual annotations in the test set. This clinical perspective confirms that a radiologist can identify the same nodules, as identified by the method described here, using only the textual information in the clinical reports.

For the clinical assessment task, a radiologist with 7 years of experience, who had not previously participated in the study, was invited to determine if he would establish the same correspondence between an entity in a medical report and a manually annotated nodule. This validation task consisted of the following steps:

– Initially, a file containing the sentences from the medical reports of each nodule and their respective centroids was provided to this radiologist.

– The radiologist was asked to classify each matching as *correct*, *imprecise*, or *incorrect*. *Correct* means that the nodule was accurately matched with its corresponding entity in the medical report. *Imprecise* indicates that the nodule was matched, but there were other nodules with similar characteristics that could also have been matched. *Incorrect* means that the nodule was wrongly matched with an entity in the medical report, showing no apparent correlation between the two.

– The radiologist viewed and classified all matching cases for each exam using a CT image visualization software (Pedrosa et al., 2020b).

In 73.9% of cases, the result was *correct*. In 7.5% of cases, the results were considered *imprecise* due to ambiguities in the calcification classifications of the nodules. For the cases with an *incorrect* decision, the radiologist in the study noted that:

- 5.0% corresponded to a scar in the manual annotation;
- 5.0% had a different location, within a fissure, which was not evaluated in this study;
- 3.8% were identified as a different type of nodule;
- 2.4% had a different size than reported;
- 2.4% where the correspondences between phrases and two nodules were swapped.

It is uncertain to what extent these results may have been influenced by the lack of systematicity in textual reporting or differing considerations among radiologists. The radiologist in this study identified some matches as scars, which were not expected in the annotation. Different sizes and types of lesions may also stem from differing considerations among the radiologists initially involved in textual reporting and manual annotation compared to the new radiologist in this context.

6.3.4 Data Insights

This subsection describes a experiment conducted to assess the validity of the GT data from medical reports generated. In the data insights, the data are analyzed to determine if the obtained matches align with the findings described in the literature.

Evidence from the literature suggests that medical reports are generally more pragmatic, whereas manual image annotation in dedicated studies tends to result in over-annotation, often including irrelevant findings (Heneghan et al., 2017; Ioannidis, 2016; Pham et al., 2014). Nevertheless, it is uncommon that these differences lead to significant discrepancies in diagnosis (Majkowska et al., 2020; Li et al., 2023). This subsection aims to verify whether this evidence from the literature aligns with the data science study described in this chapter.

It was found that in medical reports, matching is complete in 58.9% of them and only one entity remains unmatched in 30.7%. Complete means that all entities mentioned in the medical reports had matches. In manual image annotation, the correspondence with all nodules in the medical reports was achieved for 26.0% of cases, while in 25.3% there is no overlap at all between the entities of medical reports and manual annotations. This result is consistent with the literature (Heneghan et al., 2017; Ioannidis, 2016; Pham et al., 2014), meaning that manual image annotations tend to result in a larger number of annotations when compared to findings in medical reports.

The differences in the number of matched and mismatched nodules is listed in Table 6.1, presenting various diameter ranges and the number of nodules in each category. No significant differences in nodule types were evident, except for a discrepancy in the percentage of nodules smaller than 3 mm in manual annotations, likely not warranting textual relevance. As for the characteristics of nodules, comparison between the matched and mismatched nodules from manual annotations is provided in Table 6.2. Similarly, no significant differences between the two groups were noted. Minor variations were detected in malignancy, where the lower category suggests benign nodules. This implies that medical reports may prioritize clinically relevant information, possibly omitting less impactful findings in patient follow-up and management.

Table 6.1: Number of nodules with matching and mismatching between manual image annotations and medical reports based on diameter.

Diameter (mm)	?	≤3	]3,6]	]6,9]	]9,12]	]12,15]	]15,18]	]18,21]	]21,24]	]24,27]	]27,30]	>30	Total
Matches	0	66	250	78	28	12	5	3	4	2	0	2	**450**
Manual Annotation	0	128	273	71	16	7	0	0	0	1	0	0	**496**
Text Report	25	16	38	4	5	2	2	3	0	1	1	3	**100**

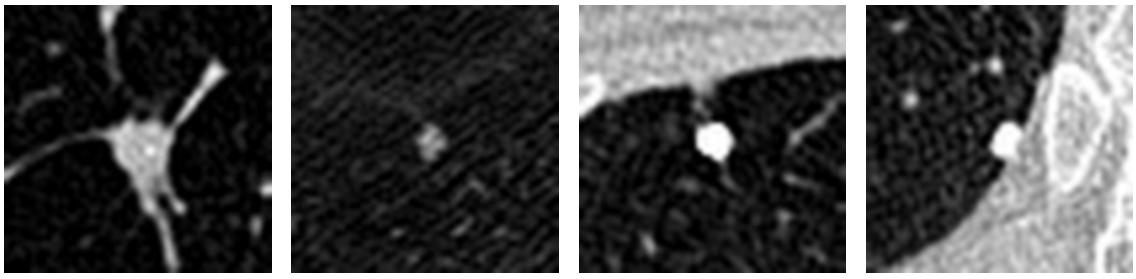
Label: The bluer colors indicate a higher % of representation (□ → ■). The % is calculated per row. The column ? quantifies the number of entities where the size is not indicated.

Figures 6.4 and 6.5 display image examples of the ROI around the centroid of respectively matched and mismatched nodules. These figures illustrate different types of nodules: isolated, Ground-Glass Opacitys (GGOs), juxtapleural, and calcified nodules. There appears to be a noticeable association between the mismatched nodules and the surrounding anatomical elements, as well as less typical nodule shapes, which could contribute to their omission in the medical reports.

Table 6.2: Number of nodules with matching and mismatching based on nodule characteristics and location.

Nodule Characteristics		*Ranking*					
		1	2	3	4	5	6
Calcification	*Matches*	0	0	29	26	38	357
	Mismatches	0	3	85	7	53	348
Internal Structure	*Matches*	443	3	3	1		
	Mismatches	489	1	0	6		
Lobulation	*Matches*	256	139	45	10	0	
	Mismatches	335	124	30	4	3	
Malignancy	*Matches*	91	154	133	59	13	
	Mismatches	188	157	115	31	5	
Margin	*Matches*	5	15	52	145	233	
	Mismatches	12	17	48	167	252	
Sphericity	*Matches*	5	70	86	131	158	
	Mismatches	16	100	101	118	161	
Spiculation	*Matches*	341	78	24	4	3	
	Mismatches	394	70	28	4	0	
Subtlety	*Matches*	12	40	90	118	190	
	Mismatches	33	79	117	118	149	
Texture	*Matches*	12	10	33	69	326	
	Mismatches	16	8	35	73	364	
		Lobes					
		LUL	LLL	RUL	ML	RLL	
Absolute Location	*Matches*	103	101	125	47	74	
	Mismatches	122	92	140	49	93	

Label: The bluer colors indicate a higher % of representation (□ → ■). The % is calculated per line. The ranking is the one followed in Table 2.1 for all characteristics.



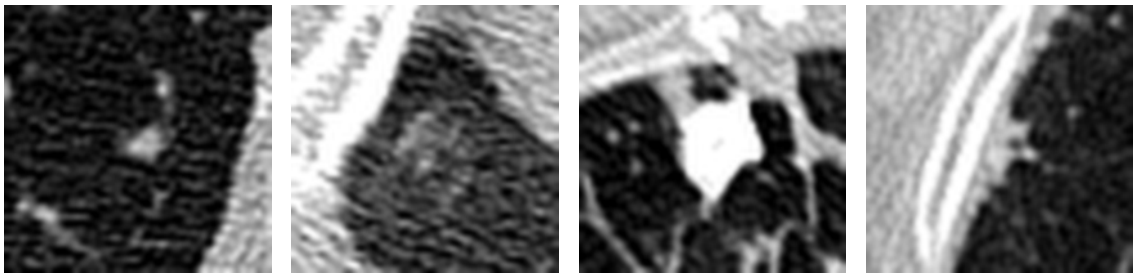
(a) **Agreement**: 3; **Diameter**: 7.52 mm; **Texture**: Solid; **Calcification**: No; **Location**: Juxtavascular.

(b) **Agreement**: 3; **Diameter**: 4.77 mm; **Texture**: GGO; **Calcification**: No; **Location**: Isolated.

(c) **Agreement**: 3; **Diameter**: 4.68 mm; **Texture**: Solid; **Calcification**: Yes; **Location**: Juxtapleural.

(d) **Agreement**: 3; **Diameter**: 4.97 mm; **Texture**: Solid; **Calcification**: Yes; **Location**: Juxtapleural.

Figure 6.4: Example of some matched nodules. Agreement ranges from 1 to 3, depending on the number of physicians who annotated each finding as a nodule.



(a) **Agreement**: 3; **Diameter**: 4.47 mm; **Texture**: GGO; **Calcification**: No; **Location**: Isolated.

(b) **Agreement**: 2; **Diameter**: 8.71 mm; **Texture**: GGO; **Calcification**: No; **Location**: Juxtapleural.

(c) **Agreement**: 3; **Diameter**: 10.30 mm; **Texture**: Solid; **Calcification**: Yes; **Location**: Juxtapleural.

(d) **Agreement**: 3; **Diameter**: 7.68 mm; **Texture**: Solid; **Calcification**: No; **Location**: Juxtapleural.

Figure 6.5: Example of some mismatched nodules. Agreement ranges from 1 to 3, depending on the number of physicians who annotated each finding as a nodule.

Finally, for the sets of nodules annotated and identified in the medical reports, the Fleischner score was used, categorizing each exam into four classes: 1) No routine follow-up required or optional CT at 12 months according to patient risk; 2) CT at 6–12 months required; 3) CT at 3–6 months required; and 4) CT, Positron Emission Tomography (PET), or tissue sampling at 3 months required. The Fleischner score is directly computed using a set of rules that consider the number of nodules (single or multiple), their volume (<100 mm^3, 100-250 mm^3, and $\geq$250 mm^3), and texture (solid, part-solid, and GGO), as shown in Table 6.3. A comparison between the manual image annotation and the medical reports can be found in Table 6.4, using confusion matrices. It is evident that the discrepancy mentioned earlier regarding mismatched nodules does not significantly affect the Fleischner results, as the outcomes are the same for the different sets in 77.1% of cases. This perspective highlights that, in the majority of cases, especially for Fleischner's follow-up recommendations, the key is to accurately reference the most clinically relevant nodules, and over-annotation and omission of findings may have a slight impact on patient management, as anticipated in the literature.

The release of pulmonary nodule data in tabular form offers extensive opportunities for further research in lung cancer. Rather than focusing solely on comparing detection performance, which may target nodules of limited clinical significance, it encourages a reassessment of the relevance of such comparisons and highlights the importance of evaluating clinical follow-up recommendations. Furthermore, it facilitates the comparison of detection algorithm performance between nodules annotated manually in images and those identified from medical reports.

Table 6.3: Fleischner classification table used in this study.

Single Nodule	Volume $<100mm^3$	$100–250mm^3$	$\geq 250mm^3$
GGO	0	1	1
Part-solid	0	2	2
Solid	0	1	3
Multiple Nodules	**Volume $<100mm^3$**	**$100–250mm^3$**	**$\geq 250mm^3$**
GGO/Part-solid	2	2	2
Solid	0	2	2
Mixed	Classify for GGO/part-solid and solid nodules independently and give the highest class		

Table 6.4: Confusion matrix comparing the Fleischner guidelines' follow-up recommendation considering the pool of nodules from manual image annotations and those in the medical reports.

		Medical Reports				
		1	2	3	4	*Total*
Manual Annotation	1	118	3	11	5	**137**
	2	0	5	0	0	**5**
	3	23	8	43	2	**76**
	4	12	1	2	59	**74**
	Total	**153**	**17**	**56**	**66**	**292**

Label: The bluer colors indicate a higher % of representation (□ → ■). The % is calculated per row and column.

This validation process was conducted using data stored in Pedrosa et al. (2023)[†], and the corresponding code is available on Github[‡].

6.4 Methods

Earlier in this chapter, LNDb_v4 is presented as a new version of LNDb, now incorporating information extracted from medical reports. The goal was to establish a basis for developing an automated method for getting image annotations from a medical report. A modular automatic approach was adopted (Figure 6.6). It has two major modules: **NLP** and **Computer Vision (CV)**.

The NLP module is designed to process clinical text, taking it as input and returning a structured format that presents simplified and interrelated entities referring specifically to lung nodules extracted from the text. It is subdivided into three distinct submodules: **Named Entity Recognition (NER)**, **Relation Extraction (RE)**, and **Text Simplification (TS)**.

[†] https://zenodo.org/records/8348419
[‡] https://github.com/carlosalexnflve/lndb_medicalreports2annotation.git

The goal of NER is to identify entity attributes, such as those already mentioned in subsection 6.3.1. Consequently, the attributes considered are: Absolute Location, Relative Location, Lesion Type, Size, Characteristics, Quantity, and Negative. The "Uncertainty" attribute was not considered due to the limited number of cases.

The RE submodule is responsible for grouping relationships between the entity attributes. For instance, RE can determine that a "nodule" is related to a specific "lung lobe", thereby structuring the information in a meaningful way.

Finally, the TS submodule focuses on reducing complex expressions to their simpler forms. For example, phrases such as "nodular finding", "slightly elongated contours", and "of considerable volume" are simplified to "nodule", "elongated", and "large". This simplification ensures that the extracted information is concise and standardized, facilitating further processing and analysis.

The CV module takes as input the structured file from the NLP module along with the CT scan, providing as output the image annotations of the nodules. Its function is to accurately locate and label nodules in the images based on the information extracted from the clinical text. It is divided into three submodules: **Anatomic Segmentation**, **Nodule Detection**, and **Matching Decision**.

The anatomic segmentation submodule focuses on segmenting anatomical regions, such as lung lobes, to define the search area for potential pathological lesions. The segmentation frames the regions of interest within specific anatomical positions (*e.g.*, anterior/posterior, lateral/medial, superior/inferior), thus narrowing the search scope.

The nodule detection submodule identifies regions within the segmented area that exhibit features consistent with nodules. This involves analyzing the segmented regions to pinpoint areas with a high likelihood of being nodules.

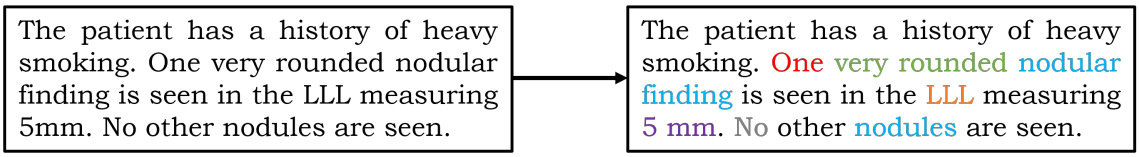
Finally, the matching decision submodule evaluates the detected candidates, comparing them against the characteristics extracted from medical texts. The candidate with the highest probability of fitting the required characteristics is selected accordingly, ensuring that the nodule mentioned in the text report is identified.

For the purposes outlined here, the dataset was divided into 5 folds, with one used for testing and the remaining four for training.

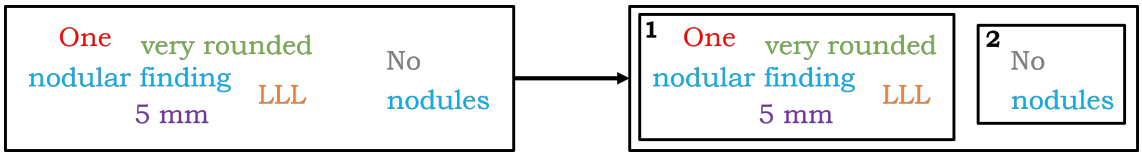
6.4.1 Named Entity Recognition

The subsection describes the development of an algorithm created by Alexandre Ferreira as part of his Master's dissertation, under the co-supervision of the author of this thesis, Nuno Guimarães (Assistant Researcher at INESC TEC), and Alípio Jorge (Full Professor at the Faculty of Sciences, University of Porto (FCUP)).

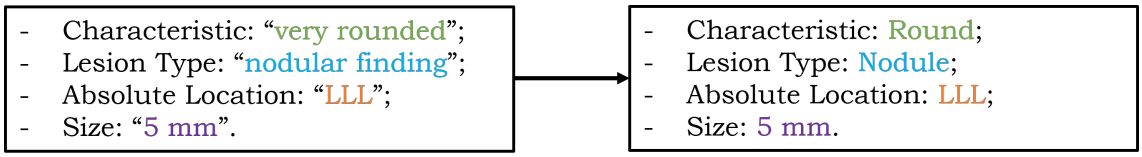
NER involves the identification of specific attributes related to pulmonary nodules in medical reports. The attributes identified through NER include Absolute Location, Relative Location, Lesion Type, Size, Characteristics, Quantity, and Negative.



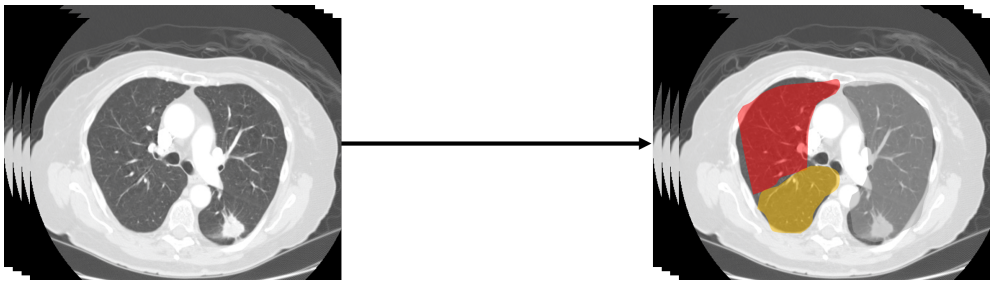
(a) **Named Entity Recognition**, where: red indicates Quantity; green indicates Characteristic; blue indicates Lesion Type; orange indicates Absolute Location; purple indicates Size; gray indicates Negative.



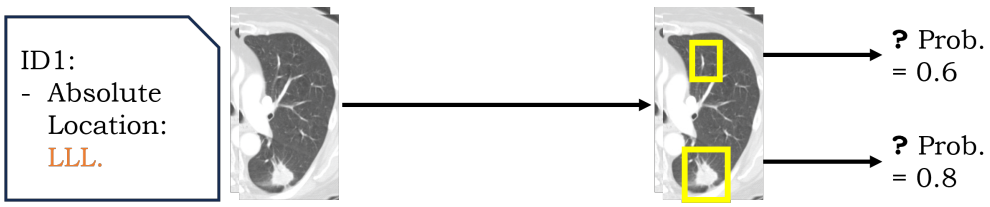
(b) **Relation Extraction**, where each numbered rectangle represents a different cluster of entities.



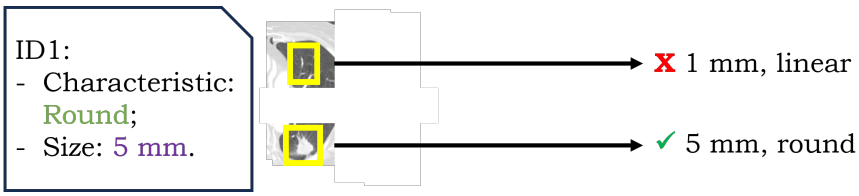
(c) **Text Simplification**.



(d) **Anatomic Segmentation**, where each colored region represents a lung lobe.



(e) **Nodule Detection**, where Prob. stands for Probability.



(f) **Matching Decision**.

Figure 6.6: Submodules of the automatic method for converting medical reports into nodule annotations in CT scans.
Note: English is used here for illustrative purposes only, as the original medical records are in Portuguese.

An example of the complete NER architecture is represented in Figure 6.7. The process of NER is essential for extracting structured information from unstructured medical text, enabling more efficient data analysis. To accurately perform NER, the text is first tokenized, breaking down the sentences into individual words. For instance, the sentence "The patient has a pulmonary nodule in the upper right lobe" would be tokenized into ["The", "patient", "has", "a", "pulmonary", "nodule", "in", "the", "upper", "right", "lobe"].

Following tokenization, vectorization is employed, where these tokens are transformed into numerical representations. Here, the BERTimbau Large model (Souza et al., 2020), which is pre-trained on a large journalistic corpus, is used. Bidirectional Encoder Representations from Transformers (BERT) generates context-aware embeddings for each token, ensuring that the meaning of each word is understood in its specific context. Importantly, during the training process, the BERT model itself is not iteratively updated or fine-tuned. Instead, BERT's pre-trained vectors are typically fixed, and the downstream layers (CharCNN, BiLong Short-Term Memory (LSTM), and Conditional Random Fields (CRF)) are trained on top of these embeddings. This means that BERT provides a contextual representation, which the following layers use without altering these base embeddings.

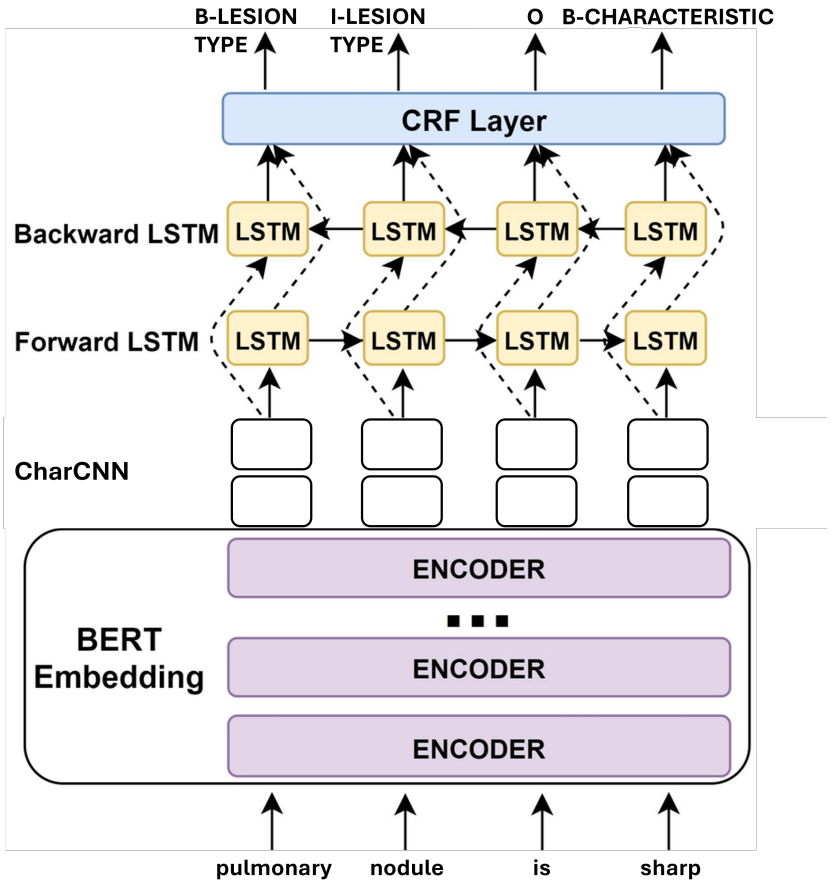


Figure 6.7: NER Architecture: BERT embeddings feed into CharCNN for subword patterns, then into BiLSTM for context, and finally, CRF for consistent tagging.

After vectorization, the next step involves classifying the tokens into their respective categories. This is achieved through a combination of CharCNN, BiLSTM, and CRF. While BERT operates at the word level, CharCNN is used to capture subword and character-level features. This works by breaking down each word into individual characters before processing. The CharCNN processes these characters to recognize patterns within the word, such as prefixes, suffixes, or morphological variants, which are especially important in the medical domain where specific terms might be spelled differently or have common roots. Despite BERT handling words as units, CharCNN helps the model capture finer-grained information that BERT might miss, enhancing the model's ability to recognize medical terminology accurately.

The outputs from the CharCNN are then passed into a BiLSTM network, which processes the sequence of tokens in both forward and backward directions to understand the relationships between tokens across the sentence, allowing the model to consider the entire context of the sentence when making classifications. Finally, the CRF layer is applied to ensure that the sequence of tags follows a logical and consistent pattern, improving the accuracy of the tagging process. CRF is particularly useful in sequence labeling tasks like NER because it considers the entire sequence of predictions when making a decision. For instance, in the sentence "The patient has no pulmonary nodules in the upper right lobe", the CRF ensures that if "pulmonary" is labeled as part of a lesion type, the subsequent word "nodule" is likely to continue this entity, rather than being labeled inconsistently. The CRF takes into account the context and the probability of certain tag sequences, ensuring that the tags assigned to each token make sense together, resulting in more accurate and coherent tagging.

An important aspect of this NER process is the use of Beginning, Inside, and Outside (BIO) tagging. This method labels each token to indicate whether it is the beginning (B), inside (I), or outside (O) of an entity. For example, in the phrase "pulmonary nodule in the upper right lobe", the tokens might be tagged as ["B-LESION TYPE", "I-LESION TYPE", "O", "O", "B-ABSOLUTE LOCATION", "I-ABSOLUTE LOCATION", "I-ABSOLUTE LOCATION"], helping the model accurately categorize entities composed of multiple words. The NER model was trained with specific hyperparameters, including a learning rate of 0.001, batch size of 8, and dropout rate of 0.4, to optimize its performance in identifying the medical attributes.

The recognition of attributes is assessed using precision, recall, and F1 metrics, as previously referenced (Equations 3.2, 3.3 and 3.4). The evaluation is conducted under both strict and relaxed criteria. Under the strict criterion, a correct identification is only recognized if all tokens belonging to an entity are accurately identified. In contrast, the relaxed criterion considers a correct identification if any token within the GT entity is correctly identified, counting it as a True Positive (TP).

6.4.2 Relation Extraction

The subsection describes the development of an algorithm created by Alexandre Ferreira as part of his Master's dissertation, under the co-supervision of the author of this thesis, Nuno Guimarães (Assistant Researcher at INESC TEC), and Alípio Jorge (Full Professor at FCUP).

RE focuses on identifying whether a relationship exists between entities within a given text. For instance, RE might determine whether a "lung nodule" and "upper right lung lobe" mentioned in different parts of a report are related. Another example can be found in Figure 6.8, while the complete RE architecture is represented in Figure 6.9.

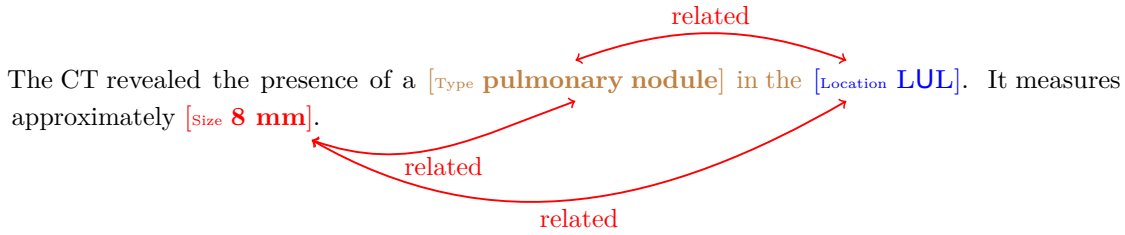


Figure 6.8: Schematic representation of some of the relations that can be established in an RE example sentence.

The input to the RE process includes several key components:

- **Identified Entities**: These are specific terms recognized in the text, such as "nodule" and "upper right lung lobe".
- **Textual Context**: The entire document is considered, as entities related to each other may appear in different sentences. This broader context is crucial for accurately capturing relationships.
- **Linguistic Features**: Part-of-Speech (POS) tags and dependency parses provide additional syntactic and grammatical information, helping the model understand the structure of the text and how entities are connected.

The model leverages a transformer-based model, specifically the BERT used in NER, to generate contextual embeddings of tokens. BERT processes the relevant text portions, creating embeddings that represent each token in a high-dimensional space, capturing its nuanced meaning in context. These embeddings are derived from the specific sentences or segments of the text provided — whether a single sentence when both entities are within the same sentence, or multiple sentences when the entities appear in different sentences. This allows BERT to understand how entities fit within their immediate context, which is crucial for accurate RE.

In the RE process, both identified entities and the surrounding textual context are transformed into embeddings, but this happens in slightly different ways:

- **Textual Context Embeddings**: The relevant sentences or text segments containing the entities are passed through BERT. BERT generates embeddings for every token in these sentences, capturing the contextual meaning based on the surrounding words.
- **Entity-Specific Embeddings**: The identified entities, such as "nodule" and "upper right lung lobe", are embedded within the same process as the context. These entities receive their own embeddings as part of the overall embedding of the sentence or sentences. However, because these entities are the focal points for RE, their embeddings might be emphasized or further processed in subsequent stages.

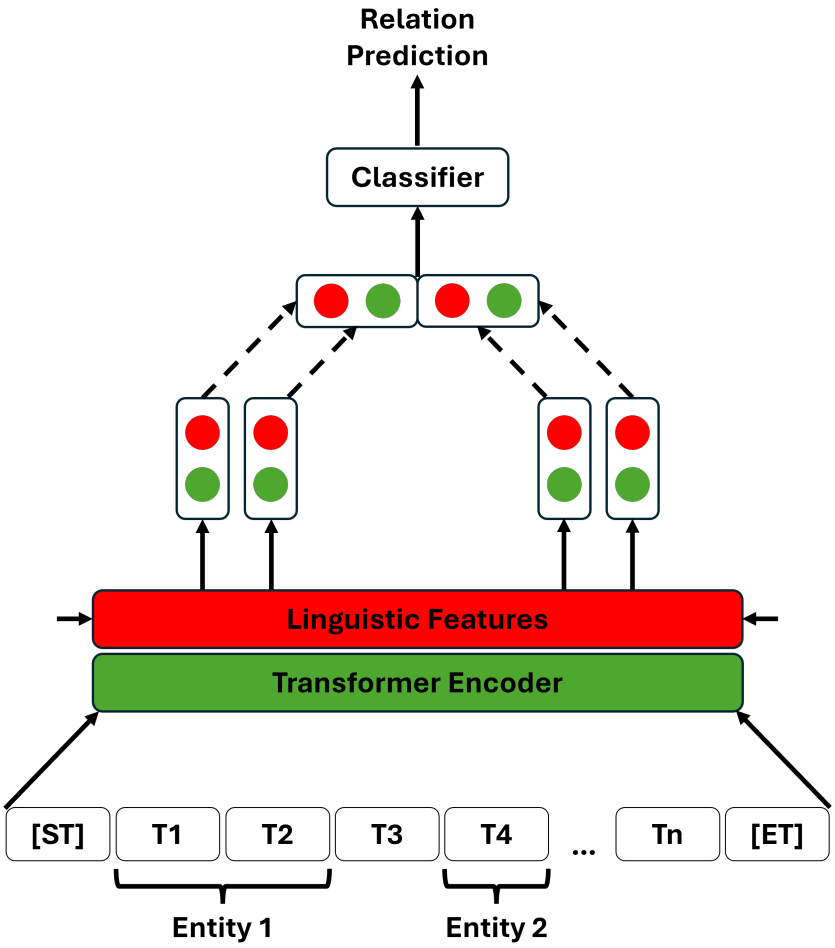


Figure 6.9: RE architecture with [ST] as the start of the text and [ET] as the end of the text. Let $T_1, T_2, T_3, \ldots, T_n$ represent the tokens in the text, where n is the total number of tokens, and T_n is the token at the last index.

These embeddings from BERT are then combined with linguistic features such as POS tags and dependency parses. The POS tags indicate the grammatical roles of words (*e.g.*, whether a word is a noun or verb), and dependency parses show the syntactical relationships between words (*e.g.*, how a subject relates to a verb). The integration of these features is achieved through concatenation. This means that the BERT embeddings are concatenated with the corresponding vectors representing POS tags and dependency parses, forming a rich, multi-dimensional input for the Neural Network (NN).

The POS tagger is trained on the Bosque dataset (Rademaker et al., 2017), part of the Universal Dependencies project, providing high-quality linguistic annotations for Portuguese. The dependency parser is similarly trained on annotated datasets to understand syntactic structures. These tools preprocess the text, producing the necessary linguistic features that are then fed into the RE model along with the BERT embeddings.

At the core of the RE model is a binary classification architecture designed to determine whether a specific relationship exists between entities. Neural networks are employed to perform this classification. The architecture involves feeding the concatenated embeddings (from BERT and linguistic features) into a neural network, which processes these inputs through multiple layers to output a prediction. This prediction indicates whether a relationship exists between the entities in question.

During the training phase, the model is exposed to pairs of entities along with their labeled relationships. For example, if a report states “The nodule is located in the upper right lung lobe”, the model learns from this example, where “nodule” and “upper right lung lobe” are labeled as related (or not). Entities with the same Identification (ID) are given a relation of “1” (indicating a relationship), while entities with different IDs are given a relation of “0" (indicating no relationship). The neural network adjusts its parameters through supervised learning to minimize prediction errors, ultimately learning to classify relationships accurately in new, unseen data. The evaluation of entity relationships is then performed using accuracy (Equation 3.1). For this model, a learning rate of 0.001 and a batch size of 8 were used.

The output of the RE model is a binary classification for each pair of entities:

– **Relationship Label**: A binary label indicates whether a relationship exists (“1”) or not (“0”).
– **Confidence Score**: Each prediction is accompanied by a confidence score that indicates the model’s certainty regarding the relationship, providing a measure of reliability for the extracted information.

After the model outputs its predictions, these results can be used to identify connected entities and build relationship trees. This involves associating related entities to clearly map out their connections, allowing for a structured understanding of relationships within the text.

6.4.3 Text Simplification

TS is aimed at transforming complex or descriptive expressions into simpler, more canonical forms. The primary objective of this step is to standardize the language, making it more predictable for subsequent phases, such as CV processing.

The simplification process begins with representing both the complex expressions and their desired simplified forms as vectors in a multi-dimensional space. These vectors are generated using BERT word embeddings. Unlike previous steps, the complex expressions in this stage represent only a single token, allowing them to maintain the same dimensional size as the simpler, more predictable forms. For instance, the terms “nodule”, “elongated”, and “large” could serve as the target simplified forms. Complex phrases such as “nodular finding”, “slightly elongated contours”, or “of considerable volume” are converted into vectors using the chosen embedding model.

Once these vectors are obtained, the next step involves calculating the similarity between the vector of a complex expression and the vectors of the target simplified forms. This is done using

the cosine similarity $S(A,B)$, defined by:

$$S(A,B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{6.1}$$

where $A \cdot B$ represents the dot product of the two vectors A and B, and $\|A\|$ and $\|B\|$ are the magnitudes of vectors A and B, respectively. A cosine similarity value close to 1 indicates that the vectors are nearly identical, while a value closer to 0 suggests that they are orthogonal or unrelated.

After calculating the cosine similarity, the expression is replaced with the target form that has the highest similarity score. In other words, the model identifies the most similar vector to the complex expression based on the expected terms and the category of the token identified in the NER phase. For example, since “nodular finding” is found to have a higher cosine similarity with “nodule” than with “mass” and “micronodule”, "nodule” is selected as the intended output.

6.4.4 Anatomic Segmentation and Nodule Detection

The anatomical segmentation serves as the initial step for the CV module. The NLP module automatically identifies Absolute Location and Relative Location, which guides the anatomical segmentation of lung lobes and the filtering of admissible points based on the indicated relative positions (apical/superior, basal/inferior, anterior, posterior, medial, lateral, centrilobular, and juxtapleural/peripheral). The automatic lobe segmentation uses the same method applied for the GT, specifically the approach by Hofmanninger et al. (2020), which employs a 2D U-Net trained on a slice-by-slice basis with a diverse set of datasets covering various diseases. This segmentation relies directly on the pre-trained algorithm without additional training, achieving a Dice of 0.97. For nodules indicated to be in the lingula, all nodules within the LUL and located below the upper boundary of the LLL are considered.

The segmented regions of the lung lobes are filtered according to Relative Location, by retaining points within the lobe that fulfill the conditions along the three anatomical axes. For nodules described as peripheral, their position is identified within the outer quarter of each anatomical axis within the lobe, adjacent to the lung walls. Conversely, nodules described as centrilobular are considered to be located in the central region of the lung, encompassing the lobe area up to half the size of the anatomical axes.

Following this, lung nodule detection is performed. Based on the identified spatial constraints, the goal is to obtain lung nodule candidates within defined regions that align with reported anatomical locations. The nodule detection used the approach developed by Aresta et al. (2020), with specific contributions from the author of this thesis in the data analysis implementation. The detection module comprises an initial candidate detection phase, followed by a False Positive (FP) reduction step, as seen in Chapter 3. Figure 6.10 illustrates this process. The detection algorithm is based on the YOLOv3 architecture (Redmon and Farhadi, 2018) and outputs bounding boxes of potential nodules on the scan.

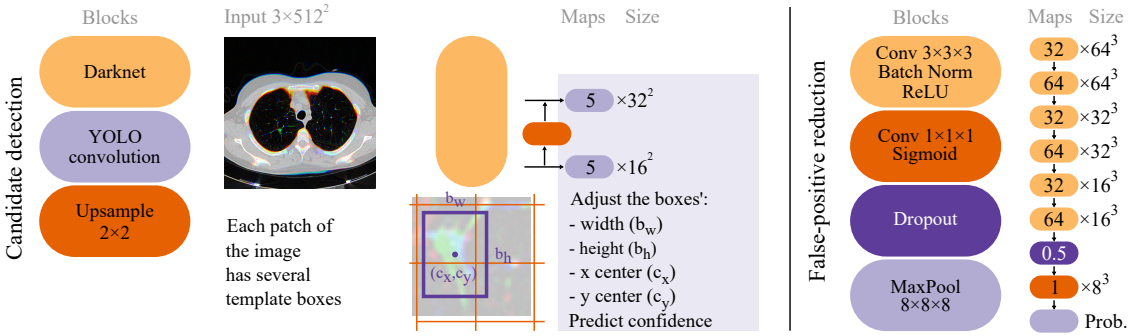


Figure 6.10: Schematic representation of the automatic lung nodule detection system's architecture, composed of an initial candidate detection followed by a false-positive reduction step. Image from Aresta et al. (2020).

YOLOv3 is an object detection architecture that predicts object bounding boxes and their labels in a single pass through the network. The feature extraction network generates a feature map F with dimensions $M \times M \times N$, where M represents the grid size and N the number of feature maps. Each grid cell $F_{i,j}$ corresponds to a specific region of the input image, effectively partitioning the image into a grid. Each cell is assigned a set of predefined anchor boxes, defined by their width (b_w), height (b_h), x-center (c_x), and y-center (c_y). For each anchor, YOLO predicts the adjustments to these parameters relative to the grid cell, along with a confidence score ($\hat{g}$) that indicates the likelihood of the box containing a nodule. The anchor boxes are optimized using k-means clustering based on the Intersection over Union (IoU) metric, simplifying the model and ensuring that the anchors align well with the object sizes in the training set. The YOLO network further refines the predefined template boxes across multiple scales, allowing the model to adapt to variations in nodule size and improve detection accuracy.

The network is trained on slices containing nodules by minimizing the detection loss:

$$L_{det} = \sum_{s=1}^{2} \lambda_s \left(\alpha_1 M_{centers} + \alpha_2 M_{dimensions} + \alpha_3 M_{confidence} \right) \tag{6.2}$$

where $M_{centers}$, $M_{dimensions}$, and $M_{confidence}$ are the mean square errors associated with the centroid, width/height, and confidence predictions, respectively. The weights α_i and λ_s are predefined to balance these components in the overall loss function.

The input to the model is a $512 \times 512 \times 3$ image composed of three neighboring axial CT slices, chosen to leverage 2D feature extractors and benefit from transfer learning with pre-trained networks on natural images. This approach reduces the computational complexity compared to 3D networks while preserving enough volumetric context to help differentiate nodules from other structures, such as blood vessels. The model is trained using the Adam optimizer (Kingma and Ba, 2015), with data augmentation techniques including random crops, rotations, translations, and slight color changes to maintain the 3D contextual information from the stacked slices. During inference, the model processes the CT volume slice by slice, generating bounding boxes and associated probabilities for nodule presence. To minimize FP, predictions located outside the lung

regions are excluded. For predictions with overlapping bounding boxes (where the inter-centroid distance is less than half the bounding box size), the centroids are averaged within a window of the bounding box dimensions, retaining the prediction with the highest confidence score for each merged region.

The initial candidate detection was trained on the LNDb dataset, while the FP reduction step used the LIDC-IDRI dataset to address overfitting to specific acquisition parameters or annotation styles, thereby enhancing the robustness of the entire detection pipeline. This dual-dataset strategy reduces the risk of overestimating performance and facilitates better generalization across varied clinical settings. By targeting 20 nodules per scan, the approach balances sensitivity with the FP rate, aiming for precise identification through anatomical filtering and nodule characterization. The binary classification of nodule *vs.* non-nodule in the false-positive reduction network was optimized using the binary cross-entropy loss.

6.4.5 Matching Decision

At this stage, a pool of candidate nodules has been identified within the CT images. Following the initial spatial identification, it is necessary to proceed with automated predictions regarding nodule size and characteristics, comparing these predictions with the information provided in the medical reports.

In cases where the size and characteristics are not explicitly mentioned, the default approach is to select the largest nodule. Conversely, if the report refers to a micronodule, it is considered to have a maximum size of up to 6 mm. This criterion aligns with the considerations established for the GT.

The segmentation method employed in this study was the iW-Net, as proposed by Aresta et al. (2019), illustrated in Figure 4.6 and further detailed in Chapter 4. This method is based on a 3D U-Net architecture (Ronneberger et al., 2015), consisting of four convolutional layers. The input comprises volumes of $64 \times 64 \times 64$ voxels with a spacing of 1 mm between voxels, and a binarization threshold of 0.5 was applied. The network was trained using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 and a batch size of 8 samples. An early stopping criterion was implemented, halting the training after 3 epochs without improvement. The loss function for the U-Net architecture was defined as $1 - \text{IoU}$. To increase the diversity of the training data, artificial augmentation techniques were applied, including random rotations, translations, flips, and zooms.

For predicting nodule characteristics, the methodology proposed by Marques et al. (2021) was employed, as also referenced in Chapter 4. This study introduces a multi-task framework for the automated prediction of lung nodule characteristics. Texture is classified into three distinct classes (solid, part-solid, and GGO), while other nodule characteristics — namely calcification, internal structure, lobulation, malignancy, margin, sphericity, spiculation, and subtlety — are evaluated as binary features, denoting either presence or absence. The network architecture consists of a CNN with three convolutional layers, which takes as input three parallel axial CT slices centered on the nodule, each with dimensions of 64×64 pixels and a resolution of 0.7 mm per pixel. The final

activation layer uses a softmax function, with cross-entropy loss guiding the training process. For binary classifications, a binarization threshold of 0.5 was applied, while for multiclass predictions, the class with the highest probability was selected. Model optimization was conducted using the Nesterov Adam optimizer (Dozat, 2016), with a learning rate of 0.002 and a batch size of 32. An early stopping criterion with a patience of 50 epochs was considered. Data augmentation included vertical and horizontal flips, as well as random rotations from 0° to 360°.

Throughout the matching process, the nodules are rescaled within the range $[-1000, 400]$ Hounsfield Unit (HU). After this rescaling, the automated segmentation and characterization steps are performed. Subsequently, it is necessary to compare the predicted segmentation with the GT reference to determine if a candidate nodule from the pool matches the reference. A margin of ± 3 mm is allowed relative to the dimensions specified in the medical report to account for measurement variability. This step involves checking if the predicted diameter D of the nodule, as given through the segmentation, lies within an acceptable range $[D_{\text{min}}, D_{\text{max}}]$, where $D_{\text{min}} = D_{\text{reported}} - 3$ mm and $D_{\text{max}} = D_{\text{reported}} + 3$ mm.

The diameter bounds, in millimeters, are then converted into a voxel-based representation. First, the nodule's dimensions are used to estimate its volume as an ellipsoid, which more accurately represents its actual size and shape. Using this estimated volume, the equivalent bounds in voxel units, denoted here as the lower and upper limits in voxel count, $\text{Voxel}_{\text{min}}$ and $\text{Voxel}_{\text{max}}$, are calculated. Specifically, $\text{Voxel}_{\text{min}}$ and $\text{Voxel}_{\text{max}}$ are determined by dividing the estimated volume of the nodule by the volume of a single voxel, taking into account the voxel spacing in each dimension of the CT scan. This conversion accounts for the precision of voxel-based segmentation, providing a voxel-based interval that aligns with the expected nodule size derived from the medical report.

The probability that the nodule size falls within the desired range is computed using a probability array P derived from segmentation predictions, representing the voxel-wise likelihood of the segmented region corresponding to the nodule. The probability values are flattened into a one-dimensional array, allowing the calculation of the probability that the number of voxels within the segmented region exceeds $\text{Voxel}_{\text{min}}$ and is less than $\text{Voxel}_{\text{max}}$.

To determine if the voxel count lies within the specified range, the probability of including exactly k voxels in the segmented region (where $k \in \mathbb{N}$) is computed as:

$$P(k \text{ voxels}) = \sum_{\text{combinations of } k} \left(\prod_{i \in S_k} P_i \right) \left(\prod_{j \notin S_k} (1 - P_j) \right) \tag{6.3}$$

where S_k represents the set of k selected voxels, P_i denotes the probability that voxel i is part of the segmented region, and $j \notin S_k$ represents voxels not included in this selection.

To assess whether the voxel count falls within the inclusive limits $\text{Voxel}_{\text{min}}$ and $\text{Voxel}_{\text{max}}$, the probabilities for all voxel counts within this range are summed:

$$P(\text{Voxel}_{\text{min}} \leq k \leq \text{Voxel}_{\text{max}}) = \sum_{k=\text{Voxel}_{\text{min}}}^{\text{Voxel}_{\text{max}}} P(k \text{ voxels}) \tag{6.4}$$

where each $P(k \text{ voxels})$ corresponds to the probability of obtaining exactly k voxels in the segmented region, as defined in Equation 6.3. This final expression quantifies the likelihood that the voxel count for the segmented region lies inclusively within the specified bounds.

To incorporate the influence of size, a binary presence flag f_s is set based on whether size information is explicitly referenced in the medical reports:

$$f_s = \begin{cases} 1, & \text{if size information is referenced,} \\ 0, & \text{otherwise.} \end{cases} \tag{6.5}$$

For each nodule, the probability of matching specific characteristics C_i for $i = 1, \ldots, n$ is also evaluated, where n is the total number of characteristics under study. The presence and relevance of each characteristic are represented by a feature array $f_c(i)$, defined as:

$$f_c(i) = \begin{cases} 1, & \text{if characteristic } C_i \text{ is referenced,} \\ 0, & \text{otherwise,} \end{cases} \tag{6.6}$$

and a corresponding weighting array $g_c(i)$ that adjusts the probability based on the values of the characteristics relative to predefined thresholds. These weights, $g_c(i)$, are calculated using rules specific to each characteristic type, ensuring that probabilities reflect the clinical relevance of the characteristic values. When at least one characteristic is present, the array $f_c(i)$ is normalized by dividing by the sum of its elements:

$$f_c(i) = \frac{f_c(i)}{\sum_{i=1}^{n} f_c(i)}, \quad \text{if } \sum_{i=1}^{n} f_c(i) > 0 \tag{6.7}$$

The overall score f_{cum} combines the presence flags of both size and characteristics:

$$f_{\text{cum}} = f_s + \left(\sum_{i=1}^{n} f_c(i) > 0 \right) \tag{6.8}$$

This score f_{cum} adjusts whether size, characteristics, or both are referenced: if only one is referenced, $f_{\text{cum}} = 1$; if both are referenced, $f_{\text{cum}} = 2$. The combined probability for nodule selection is then defined as:

$$P(N) = \frac{1}{f_{\text{cum}}} \left(P(D_{\min} \leq D \leq D_{\max} \mid S) + \sum_{i=1}^{n} f_c(i) \cdot g_c(i) \right) \tag{6.9}$$

where S represents the segmentation prediction of the nodule in voxel space. Dividing by f_{cum} ensures that the influence of the probabilities is normalized based on the presence of either or both references. The candidate nodule with the highest overall probability $P(N)$ is selected as the most likely match. This probabilistic approach ensures that both size and characteristic data are effectively considered, optimizing the identification of the correct nodule within the candidate pool.

For the evaluation, two scenarios were considered: the perfect scenario, in which the NLP module operates perfectly, allowing the performance of the CV module to be assessed independently; and the real scenario, where the prediction results from the NLP module, which may include some identification errors, are passed to the CV module. In both scenarios, precision and recall (Equations 3.2 and 3.3) were used to determine the number of nodules correctly identified in relation to the GT. Additionally, undetected nodules were examined, along with hypothetical nodules that were considered due to errors from the NLP module. The impact on follow-up recommendations was also assessed using Fleischner guidelines in both scenarios, to evaluate whether different nodules per patient lead to varying clinical recommendations.

6.5 Results

6.5.1 Natural Language Processing

The NLP module aims to accurately identify attributes and their relationships to ensure correct information is passed to the CV module. Below is an example of a correct result from the NLP module. To enhance clarity, the original medical report in Portuguese is presented first, followed by its English translation and subsequent NLP processing only for the purposes of this thesis:

Original Input: "Mulher de 67 anos, não fumadora, nódulo de 7mm no segmento anterior do LSE (TC do exterior). Não dispomos de exames prévios para comparação. Calcificações da válvula mitral. Sem adenomegalias hilares ou mediastínicas. Sem derrame pleural ou pericárdico. No parênquima pulmonar observam-se várias áreas de atenuação em vidro despolido, com grande expressão no LID e, em menor grau e em menor grau no LIE e nos lobos superiores (nomeadamente na vertente paramediastínica do LSD), que poderá traduzir intercorrência infecciosa ou eventualmente processo infeccioso em resolução, a merecer reavaliação a curto prazo (3 meses). Não se observam outros nódulos, massas ou consolidações no parênquima pulmonar".

Translated Version: "67-year-old woman, non-smoker, 7mm nodule in the anterior segment of the LUL. We do not have previous exams for comparative study. Mitral valve calcifications. No hilar or mediastinal lymphadenopathy. No pleural or pericardial effusion. Several areas of ground-glass attenuation are observed in the lung parenchyma, predominantly in the right lower lobe (particularly in the paramediastinal region), which may suggest an infectious event or possibly a resolving infectious process, warranting short-term reevaluation (3 months). No nodules, masses, or consolidations are observed in the lung parenchyma".

NER Output – in the format of cluster number (attribute) entity:
– Size: “7mm”; Lesion Type: “nodule”, Relative Location: “anterior”, Absolute Location: “LUL”; Negative: “No”, Lesion Type: “nodules”, Lesion Type: “masses”.

RE Output – with the entities of each cluster:
– Cluster 1: “7mm”, “nodule”, “anterior segment”, “LUL”;
– Cluster 2: “No”, “nodules”, “masses”.

TS Output – with the simplified version of the wording for each entity in each cluster:
– Cluster 1: “7mm”, “nodule”, “anterior”, “LUL”;
– Cluster 2: “No”, “nodule”, “mass”.

Table 6.5 shows the performance in terms of precision, recall, and F1 metrics for NER. Evaluation is done using strict and relaxed criteria: strict requires all tokens of an entity to be correctly identified, while relaxed counts a TP if any token within the GT entity is correctly identified.

From the analysis of Table 6.5, it is evident that the relaxed evaluation shows better performance, as expected. Typically, entities are identified with fewer tokens without significantly affecting their meaning. For instance, instead of returning “nodular body”, it identifies “nodular”; instead of capturing “3 mm in diameter”, it captures just “3 mm”.

Table 6.5: Performance of NER

	Strict Evaluation			Relaxed Evaluation		
	Precision	Recall	F1	Precision	Recall	F1
Absolute Location	0.85	0.81	0.83	0.89	0.85	0.87
Relative Location	0.68	0.71	0.69	0.85	0.88	0.87
Lesion Type	0.90	0.82	0.86	0.94	0.86	0.90
Size	0.84	0.89	0.86	0.91	0.97	0.94
Characteristics	0.78	0.56	0.65	0.84	0.60	0.70
Quantity	0.79	0.59	0.67	0.83	0.63	0.71
Negative	0.79	0.86	0.83	0.88	0.95	0.91
Average	0.83	0.78	0.80	0.90	0.85	0.87

Regarding the attributes, Absolute Location, Relative Location, Lesion Type, Size, and Negative show the best performance. These attributes often have predictable tokens that are automatically identified. For example, Absolute Location frequently includes lobe abbreviations like RUL or ML. The main errors occur with terms that are less predictable in medical reports, which are known for their variability, or when the prediction returns tokens that are very similar to those expected for lung nodules but actually describe other imaging findings.

Regarding RE, an accuracy of 0.69 was achieved. This result stems from the difficulty of linking related entities across different sentences or handling multiple types of related entities within the same sentence. Consequently, there is significant variability that the model must account for,

which is particularly challenging given the limited dataset, making it difficult to achieve robust performance. Finally, regarding TS, there are no relevant errors to mention.

6.5.2 Computer Vision

Figure 6.11 presents three examples of CT scans where the automated method performed correctly. Recognized entities are underlined in the text, with different colors indicating distinct clusters. The color of the bounding box around each nodule corresponds to the color of the entity cluster associated with that nodule in the clinical text. In the first example in Figure 6.11, the nodule description is quite detailed, allowing for a correct match. In the second example, two clusters were identified, with the second cluster highlighted in red in Figure 6.11. In the final example, along with the red cluster representing the nodule, there is also a mention indicating that no additional nodules are present. Interestingly, another nodule is referenced, but as it is not located in the pulmonary region, the method was robust enough to exclude it as an entity.

Nodules that hypothetically match descriptions in medical reports are identified using a predefined file format generated by the NLP module, which serves as the basis for the entire CV module. Table 6.6 presents the precision and recall results for nodule identification in both perfect and real scenarios, where the NLP module is simulated to operate perfectly, and where predictions are derived from the actual NLP module, respectively.

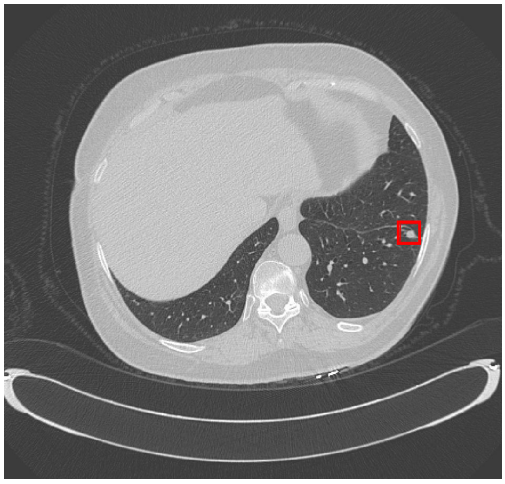
Table 6.6: Performance in the identification of lung nodules.

	Precision	**Recall**		**Precision**	**Recall**
Perfect Scenario	0.75	0.82	Real Scenario	0.54	0.75

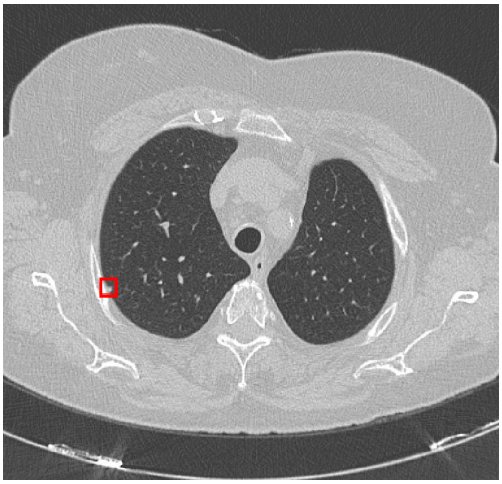
The analysis of Table 6.6 under the real scenario highlights the impact that errors in the NLP module have on the performance of pulmonary nodule identification. The NER affects recall by occasionally failing to identify entities, which increases the False Negatives (FNs), while the RE impacts precision by not successfully relating entities, interpreting them as uncorrelated, leading to a higher number of identified nodules. Even in the perfect scenario of the NLP module, precision and recall do not reach their maximum values, indicating that the lung nodule detection step remains quite challenging and needs further improvement.

Table 6.7 compares the follow-up recommendations according to the Fleischner guidelines between the GT from medical reports and the scenarios where nodules are identified, considering both the perfect and real scenarios within the NLP module, displayed side by side. Higher values are more prominently highlighted in dark blue. Follow-up recommendations are based on all detected nodules, so any FPs or FNs directly impact the outcome of these recommendations. Generally, better performance in nodule identification in the perfect scenario led to more accurate follow-up recommendations compared to the real scenario.

Several lymph node formations in the mediastinum, some with coarse calcifications, of residual nature. Calcifications of the aorta and its branches. The study of the lung parenchyma is overall comparable, with the persisting nodule on reevaluation in the left lower lobe measuring approximately 7 mm in diameter. Small sliding hiatal hernia.



CT scan performed w/o IV contrast. In the subpleural region of the RLL, there is a persisting 12 mm nodular density. In the remaining lung parenchyma, a 3 mm subpleural micronodule is noted in the RUL. Mediastinal and right hilar lymph nodes are infracentimetric, grossly calcified, consistent with sequelae.



Pulmonary parenchyma well-expanded, with stability noted in the 6 mm nodule in the right upper lobe. No other relevant nodules in the pulmonary parenchyma. In the sections intersecting the upper abdomen, gallstones and a subcentimetric hypodense nodule in the left hepatic lobe, too small for further characterization, are observed.

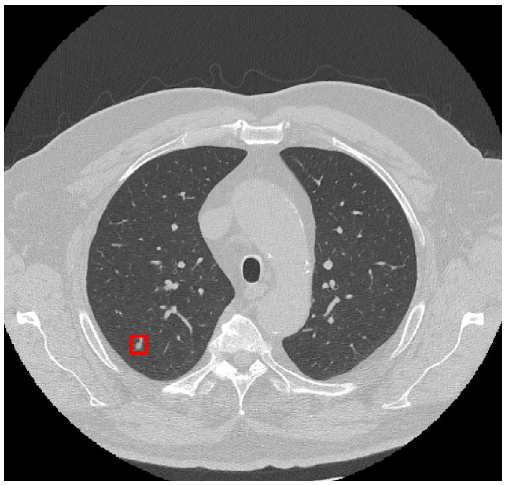


Figure 6.11: Examples of correctly matched nodules with clinical report on the left and CT images with marked nodules on the right. In the text, the recognized entities are underlined, with different colors indicating distinct clusters. The color of the bounding box around each nodule matches the color of the entity cluster associated with that nodule in the clinical report. The medical report is originally in Portuguese and is translated into English for illustration purposes.

Table 6.7: Confusion matrices comparing the Fleischner guidelines' follow-up recommendations between the GT from medical reports and scenarios where nodules are identified, considering both the perfect and real scenarios within the NLP module.

		Perfect Scenario				*Real Scenario*				
		1	2	3	4	1	2	3	4	*Total*
Medical Reports	1	21	1	0	2	20	0	3	1	**24**
	2	1	3	0	0	2	0	0	2	**4**
	3	0	0	7	2	0	1	3	5	**9**
	4	0	0	2	14	0	2	4	10	**16**
	Total	**22**	**4**	**9**	**18**	**22**	**3**	**10**	**18**	**53**

Label: The bluer colors indicate a higher % of representation (□ → ■).

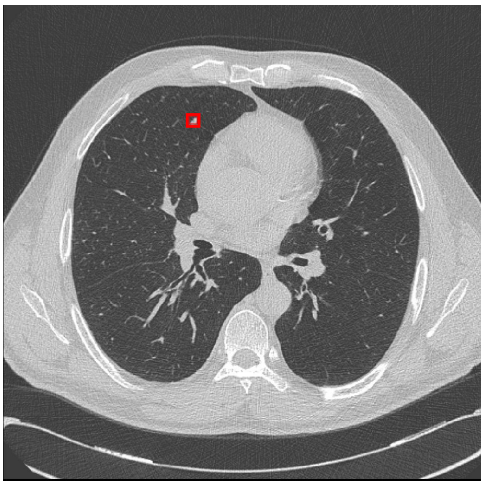
6.5.3 Discussion

Three incorrect examples are shown in Figure 6.12. These correspond respectively to cases where either additional clusters were identified, fewer clusters were recognized, or a nodule was incorrectly detected. Clinical text is displayed on the left, and CT images with circled nodules appear on the right. In the text, recognized entities are underlined, with different colors indicating distinct clusters. The color of the bounding box around each nodule matches the color of the entity cluster associated with that nodule in the clinical text. In the first example, the terms "granuloma", "calcified" and "right lower lobe" were correctly identified as related entities. However, "superior" and "3 mm" were incorrectly grouped into a separate entity cluster, leading to inaccuracies by identifying a finding that isn't even a nodule. In the second example, although the correctly identified nodule appears in the CT image, only one cluster was recognized in the clinical text, when two clusters should have been identified, resulting in an additional undetected nodule in the CT. Finally, in the last example, the textual clustering is accurate, but the nodule is detected with a red bounding box, whereas it should have been identified with a green bounding box.

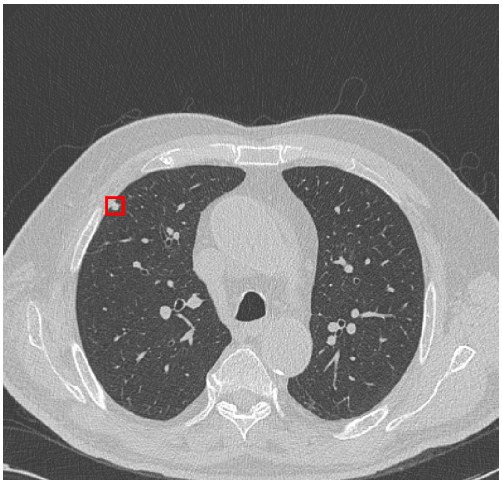
Given that the automatic approach to matching text references with image annotations for pulmonary nodules is modular, it was expected that errors throughout the pipeline would impact the overall performance of nodule identification. The steps that most significantly affected performance were RE and nodule detection. For RE, an accuracy of 0.69 was achieved, indicating limitations in reliably extracting relevant entities from clinical texts. For lung nodule detection, previous studies on the LNDb dataset, including our methodological approach, achieved a precision of 0.7.

Training the RE model from scratch without any pre-training likely had a considerable impact on its performance. The state of the art highlights that pre-training on large, relevant datasets can enhance model accuracy and generalization (Wei et al., 2019), especially in RE tasks. However, given the limited availability of pre-trained resources in European Portuguese, pre-training was not feasible in this case, which may explain the observed limitations in performance. Furthermore, the restricted amount of data poses an inherent challenge; generating synthetic sentences to augment

Calcified granuloma in the right lower lobe, specifically within the superior segment, measuring 3 mm. Mild centrilobular-predominant emphysema. The abdominal sections show no significant alterations on spontaneous imaging, with only slightly rounded liver contours. No suspicious bone lesions.



Two micronodules are observed on the right side, one in the upper lobe measuring approximately 5 mm and another in the middle lobe measuring around 7 mm. There is an increased diameter of the pulmonary artery trunk (37 mm), suggesting pulmonary hypertension. In the abdominal planes included in the study, a probable subcentimetric cyst is identified in the right hepatic lobe, along with a cyst in the right kidney.



Solid nodule with spiculated margins in the apical segment of the RUL, measuring 15 mm. Pulmonary micronodule measuring 5 mm in the superior segment of the RLL. Pulmonary emphysema. Sclerotic nodular lesion measuring 10 mm in the L1 vertebral body, with high density (over 1000 HU), showing no aggressive features - likely an enostosis. A 4 mm lesion is present in the T11 vertebral body.

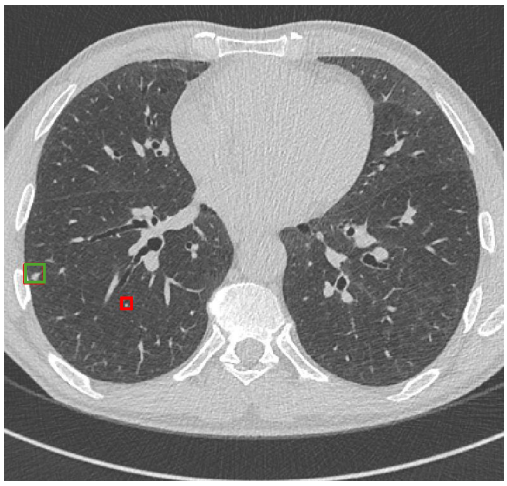


Figure 6.12: Examples of nodule matches with inaccuracies, with clinical report on the left and CT images with marked nodules on the right. In the text, the recognized entities are underlined, with different colors indicating distinct clusters. The color of the bounding box around each nodule corresponds to the color of the entity cluster associated with that nodule in the clinical report. The medical report is originally in Portuguese and is translated into English for illustration purposes.

the dataset could offer a viable solution. For nodule detection, adopting more recent state-of-the-art methods may improve results, as YOLOv3 was used here while YOLOv8, the latest version of this framework, has seen widespread adoption and may provide enhanced accuracy and efficiency.

Interestingly, the NLP module tends to achieve better precision, whereas the CV module shows better recall. For NLP, this occurs because some entities remain undetected due to the variability of terms that can appear in medical writing. In contrast, CV faces the significant challenge of correctly detecting nodules without generating a substantial number of FPs. Notably, 82% of the FPs involve nodules between 3 and 6 mm, suggesting that their size complicates detection. In the literature, methodological proposals for lung nodule detection on the LNDb dataset generally do not achieve precision beyond 0.7.

The introduction of the relaxed evaluation scenario for the NLP module and the perfect scenario for the CV module were prudent choices, as they offer valuable insights into partial token detection and the standalone performance of the CV module, respectively, thus aiding future development efforts.

Although the proof of concept presented here has considerable room for improvement, the results in Table 6.7 — particularly regarding follow-up recommendations — indicate that, despite errors throughout the pipeline, the potential impact on patient care would not lead to significantly different follow-up decisions. The results do not reveal a consistent trend toward either under-recommendation or over-recommendation in follow-up advice, despite variations in nodule identification. The improved accuracy observed in the perfect scenario suggests a potential for enhancing follow-up recommendations as nodule detection becomes optimized. However, even when discrepancies occur, the recommendations tend to fall into adjacent categories, which may mitigate potential risks if this method were implemented in real practice.

As far as the author is aware, this is the first study that enables the extraction of image annotations from medical reports specifically for lung nodules, creating a structured mapping from narrative medical records to spatial markers within CT images. Consequently, no direct comparisons can be made with the existing literature, as similar studies focusing on this type of annotation pipeline are absent, even across individual methodological modules. Moreover, the dataset used here is unique, addressing the specific requirements of this task, which led to the release of the LNDb_v4 precisely for this purpose.

6.6 Concluding Remarks

This chapter aimed to address the complex task of linking references to lung nodules in medical reports with their corresponding image annotations in CT scans, with the goal of enhancing follow-up processes. The study encountered significant challenges due to the asynchronous nature of the image annotations and medical reports, which were completed at different times, resulting in uncertainty when correlating text with image data. To overcome this, a data science approach was employed, integrating clinical insights to maximize the retrieval of annotations and establish a

reliable GT. Validation and evaluation of these pairings were conducted with clinical experts, and once reliable matches were confirmed, an automated method was developed.

The automated method was evaluated on CT scans that had corresponding reports. A further challenge involves extending this approach to future CT scans by interpolating nodule dynamics over time, recognizing that nodules may grow, change, or evolve between scans. This requires identifying nodules in subsequent CT scans using previous medical reports, which can be particularly complex due to the natural variations in nodule appearance and behavior. This forward-looking approach underscores the importance of exploring how earlier textual descriptions can guide the identification and tracking of nodules in later imaging, thereby providing a more comprehensive view of disease progression.

The automated method demonstrated promising results in both NLP and CV tasks. The NLP component was robust in identifying phrases referring to pulmonary nodules, even though the reports also included other findings. The ability of NLP to accurately extract relevant information is critical, given the complexity and variability of language used in medical reports. However, even when the NLP component encountered errors, these did not significantly impact the overall results, as the CV component often correctly identified the nodules regardless. This redundancy between the NLP and CV modules highlights the robustness of the approach, where the strengths of one component can offset the limitations of the other.

Despite these advancements, the study faced several limitations. The relatively small dataset used constrained the model's ability to generalize across diverse clinical settings, potentially limiting the broad applicability of the findings. Additionally, the performance of the CV component, particularly in nodule detection, revealed challenges that require further refinement. These detection difficulties suggest that additional data and more sophisticated training approaches could improve the robustness and accuracy of the nodule detection algorithms. Furthermore, issues related to interobserver variability are also relevant, as different radiologists contributed to the annotation process, with varying degrees of permissiveness in considering lung nodules. This variability can affect the consistency of automatic models. Historically, the LNDb dataset has shown significant challenges in the task of nodule detection, reflecting the inherent complexity of this clinical problem.

The automated tool developed in this research facilitates the creation of annotated image datasets for CV by automating the extraction of image annotations. This approach enhances the training of AI models for medical imaging and offers a pathway to improve the accuracy and efficiency of lung nodule follow-up. By integrating NLP and CV techniques, the tool bridges the gap between textual and imaging data, streamlining the traditionally manual annotation process and supporting clinical workflows with a more cohesive approach to patient data management.

Chapter 7

Conclusions

The contributions made by this document were motivated by the recognized significant gaps in current diagnostic support systems for the follow-up of lung cancer. In the case of lung cancer, the focus has traditionally been on prevention, as its onset is primarily linked to smoking behaviors. Thus, when patients reach a hospital, it becomes imperative to maintain the highest level of accuracy in follow-up procedures, as immediate treatment is not always the first line of action.

The methods for diagnostic support in lung cancer have predominantly focused on tasks related to the detection, segmentation, and characterization of nodules. However, proposals specifically aimed at follow-up are not as common. This may be due to the fact that follow-up processes build upon these earlier tasks, leading to a lack of dedicated methodological studies. Nevertheless, errors can occur at each of these stages, underscoring the importance of health studies that focus on follow-up, particularly those based on guidelines that provide actionable recommendations.

In this thesis, follow-up has been explored through three main dimensions: evaluation, forecasting, and reassessment. For evaluation, a method was developed to automatically predict follow-up recommendations based on Lung-RADS guidelines. In forecasting, a Computer-Aided Diagnosis (CAD) system was proposed to predict the growth of pulmonary nodules. Finally, in reassessment, a method was devised to restore image annotations of pulmonary nodules based on information from medical reports.

The proposed follow-up evaluation compares end-to-end approaches to feature-based methodologies. Both demonstrated similar performance, but the methodologies should be targeted toward different goals. The end-to-end approach, due to its lower explainability, may be more suited for stratification purposes, while the feature-based approach, which allows for easier modification of individual feature components, could be more appropriate for clinical applications.

In predicting nodule growth, different types of data were used, namely, tabular data, images, and a combination of both. Although the tabular data led to better results, this discrepancy may stem from the underutilized complexity of image data. In contrast to tabular data, which has been pre-processed and is thus closer to unlocking its potential utility, image data still holds untapped potential that could be harnessed with larger datasets and more advanced processing techniques.

Reassessing nodules from medical reports for Computed Tomography (CT), with the support

of Artificial Intelligence (AI), introduces a new paradigm. Traditionally, radiologists were required to conduct dedicated image annotation studies, but with the availability of text-image pairs, there is potential for automatically establishing connections between both. The primary challenge remains the nodule detection phase, as achieving robust results for this task has been difficult across different methodological proposals in the state-of-the-art literature.

In addition to the Lung-RADS and Fleischner guidelines, it is also important to consider the British Thoracic Society (BTS) guidelines, as they incorporate patient data, providing more context to assist in follow-up. In terms of lung cancer progression, it is evident that persistent heavy smoking habits compared to transitioning to tobacco abstinence could lead to different outcomes, and guidelines should reflect these behaviors. Lung-RADS, being nodule-centric, and Fleischner, with its more holistic approach, are the most prominent proposals. Moreover, these guidelines are often applied in regions with more influence, such as in the United States of America (USA) and the Europe.

This thesis has explored the uncertainty in annotations, particularly involving multiple annotators, seeking to embed this information into algorithms. Annotation uncertainty revealed a critical role in ensuring that algorithmic predictions are robust and can account for variations in human interpretation. In turn, the use of medical reports to perceive spatial image nodule locations was also studied, addressing the opposite scenario of the absence of image annotations. When analyzing differences in the restoration of image annotations, it is observed that clinical text tends to be more conservative in mentioning findings. However, these differences, compared to dedicated image annotation studies, did not lead to significant divergences in Fleischner guideline recommendations. The nodules most frequently referenced in both types of data are those critical for diagnosis. Using text-to-image annotation could be a solution to reduce the burden on medical professionals in computer vision studies where annotations are required.

This thesis has consistently sought to evaluate results across datasets as a whole, while also focusing on relevant nodules, such as those that grow and those that are malignant. The evidence from these studies suggests that the developed method for predicting nodule size growth could be used for personalized predictions of Volume Doubling Time (VDT) for malignant nodules, surpassing the current reference widely used in the literature.

In summary, it is evident that the integration of AI into follow-up processes is not only a matter of technical enhancements but also a step towards reshaping clinical workflows. AI-driven systems can advance towards more personalized and precise care by enabling dynamic updates to follow-up protocols based on real-time data. The ability to predict nodule growth, automate follow-up recommendations, or simply restore image annotations for comparison in a new exam exemplifies how AI can accelerate diagnostic procedures, increase the accuracy of clinical evaluations, and reduce inefficiencies within healthcare systems.

7.1 Future Trends

The future of lung cancer follow-up is likely to be shaped by several emerging trends. One of the most significant developments will be the increased use of Large Language Models (LLMs) and transformers in the analysis of clinical data. These models, such as Generative Pre-trained Transformer (GPT) and others, are capable of understanding and generating human-like text, which can be crucial in interpreting complex clinical narratives and integrating them with imaging data for more comprehensive follow-up strategies.

Looking forward, the challenges in this field are expected to revolve around leveraging CT as a data source for uncovering knowledge and patterns that may not be readily apparent to the human eye. In the past, radiologists inferred the malignancy of a nodule based solely on visual inspection, knowing that larger, solid nodules have a higher likelihood of being malignant. However, definitive confirmation typically requires a biopsy. By using CT scans as input and biopsy results as labels, the literature has increasingly focused on developing methods to tackle this challenge.

Beyond malignancy, there are ongoing efforts to predict the specific mutations that cause cancer. Different mutations may lead to distinct imaging patterns, and AI could play a crucial role in predicting these mutations. Predicting mutations without invasive methods could be pivotal for personalized treatment plans, aligning with the broader trend of personalized medicine. Furthermore, the integration of AI with genomic data may reveal deeper insights into the etiology of lung cancer, offering new avenues for treatment.

While deep learning remains a dominant trend, radiomics has also regained attention in the literature. It is important to have explainable methods, and radiomics, based on a filter-based approach, offers a level of predictability regarding which filters should be taken to yield the best results. Deep learning, often seen as a black box, may not always be the best option in terms of trust and transparency. Therefore, future research may focus on creating hybrid models that combine the strengths of deep learning and traditional machine learning, balancing accuracy with interpretability.

In addition to focusing on pulmonary nodules, there is now an interest in other findings from CT scans to confirm the presence of lung cancer. For instance, emphysema has recently been suggested as another potential indicator. Additionally, other parenchymal changes, such as fibrosis or air trapping, might also be leveraged as secondary indicators for lung cancer, providing a broader context for disease characterization.

Another promising area of research is the development of complementary metrics alongside deep learning model results. These metrics could provide additional confidence in the reliability of the model's predictions, thereby enhancing clinical trust in AI-driven decisions. In particular, the integration of uncertainty measures could help quantify the confidence levels associated with specific predictions, making these tools more reliable for clinical use. Moreover, incorporating real-time feedback from physicians into these models could lead to continuous improvement, allowing AI systems to adapt and evolve in response to clinical needs.

There is also a growing advocacy to use incidental data in research, as opposed to primarily relying on screening data. This involves using CT scans, initially targeted for other disease purposes, to detect nodules since other findings may also be reported alongside the expected observation. A comprehensive screening for all types of findings in the cardiothoracic context is advised, which could lead to the early detection of a broader range of conditions, improving overall patient care.

It is important to note that the practical implementation of these systems in clinical practice remains challenging, as they often require significant changes to routine processes. Radiologists are already burdened with extensive interaction with technology, and they do not need another tool that could potentially add more workload than benefit. Therefore, these systems should be intuitive by design, serving as a second opinion that physician can consult when necessary, thus minimizing interference while facilitating integration through small, user-friendly features such as buttons or dropdown menus that are invoked only when needed.

Finally, the future of follow-up systems should not only focus on technological advancements but also consider the socio-ethical implications of AI in healthcare. Issues such as data privacy, patient consent, and the potential for algorithmic bias must be addressed to ensure that AI-driven tools contribute positively to healthcare outcomes. Engaging with patients and healthcare professionals in the development process will be crucial for creating systems that are not only effective but also trusted by all stakeholders involved.

References

American College of Radiology. Lung-RADS Assessment Criteria. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads, 2019. [Accessed 31-07-2023].

Guilherme Aresta, Teresa Araújo, Colin Jacobs, Bram van Ginneken, António Cunha, Isabel Ramos, and Aurélio Campilho. Towards an automatic lung cancer screening system in low dose computed tomography. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 310–318. Springer, 2018. doi: 10.1109/JTEHM.2014.2381213.

Guilherme Aresta, Colin Jacobs, Teresa Araújo, António Cunha, Isabel Ramos, Bram van Ginneken, and Aurélio Campilho. iW-Net: an automatic and minimalistic interactive lung nodule segmentation deep network. *Scientific reports*, 9(1):1–9, 2019. doi: 10.1038/s41598-019-48004-8.

Guilherme Aresta, Carlos Ferreira, Joao Pedrosa, Teresa Araújo, Joao Rebelo, Eduardo Negrao, Margarida Morgado, Filipe Alves, Antonio Cunha, Isabel Ramos, and Aurelio Campilho. Automatic Lung Nodule Detection Combined With Gaze Information Improves Radiologists' Screening Performance. *IEEE journal of biomedical and health informatics*, 24(10):2894–2901, 2020. ISSN 21682208. doi: 10.1109/JBHI.2020.2976150.

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. *Medical Physics*, 38 (2):915, 2011. ISSN 00942405. doi: 10.1118/1.3528204.

Gurraj Atwal and Hady Ahmady Phoulady. Automatic lung cancer follow-up recommendation with 3D deep learning. In *International Conference on Image Analysis and Recognition*, pages 406–418. Springer, 2020. doi: 10.1007/978-3-030-50516-5_36.

Joseph Awad, Amir Owrangi, Lauren Villemaire, Elaine O'Riordan, Grace Parraga, and Aaron Fenster. Three-dimensional lung tumor segmentation from X-ray computed tomography using sparse field active models. *Medical physics*, 39(2):851–865, 2012. doi: 10.1118/1.3676687.

Matthew S. Brown, Pechin Lo, Jonathan G. Goldin, Eran Barnoy, Grace Hyun J. Kim, Michael F. McNitt-Gray, and Denise R. Aberle. Toward clinically usable CAD for lung cancer screening with computed tomography. *European Radiology*, 24(11):2719–2728, 11 2014. ISSN 0938-7994. doi: 10.1007/s00330-014-3329-0.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel

Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*, 33:1877–1901, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Michael A Bruno, Eric A Walker, and Hani H Abujudeh. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*, 35(6):1668–1676, 2015. ISSN 1527-1323. doi: 10.1148/rg.2015150023.

M E J Callister, D R Baldwin, A R Akram, S Barnard, P Cane, J Draffan, K Franks, F Gleeson, R Graham, P Malhotra, M Prokop, K Rodger, M Subesinghe, D Waller, and I Woolhouse. British thoracic society guidelines for the investigation and management of pulmonary nodules: accredited by NICE. *Thorax*, 70(Suppl 2):ii1–ii54, June 2015. ISSN 1468-3296. doi: 10.1136/thoraxjnl-2015-207168.

Haichao Cao, Hong Liu, Enmin Song, Chih-Cheng Hung, Guangzhi Ma, Xiangyang Xu, Renchao Jin, and Jianguo Lu. Dual-branch residual network for lung nodule segmentation. *Applied Soft Computing*, 86:105934, 2020. doi: 10.1016/j.asoc.2019.105934.

D. Cascio, R. Magro, F. Fauci, M. Iacomi, and G. Raso. Automatic detection of lung nodules in CT datasets based on stable 3D mass–spring models. *Computers in Biology and Medicine*, 42 (11):1098–1109, 11 2012. ISSN 00104825. doi: 10.1016/j.compbiomed.2012.09.002.

Yung-Chun Chang et al. Deep multi-objective learning from low-dose CT for automatic Lung-RADS report generation. *Journal of Personalized Medicine*, 12(3):417, 2022. doi: 10.3390/jpm12030417.

Sihong Chen, Jing Qin, Xing Ji, Baiying Lei, Tianfu Wang, Dong Ni, and Jie-Zhi Cheng. Automatic scoring of multiple semantic attributes with multi-task feature leverage: A study on pulmonary nodules in CT images. *IEEE transactions on medical imaging*, 36(3):802–814, 2016. doi: 10.1109/TMI.2016.2629462.

Hailan Cheng, Yaping Zhu, and Hong Pan. Modified U-Net block network for lung nodule detection. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pages 599–605. IEEE, 2019. doi: 10.1109/ITAIC.2019.8785445.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Cheng_YOLO-World_Real-Time_Open-Vocabulary_Object_Detection_CVPR_2024_paper.pdf.

W.-J. Choi and T.-S. Choi. Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor. *Computer Methods and Programs in Biomedicine*, 113(1):37–54, 2014. doi: 10.1016/j.cmpb.2013.08.015.

N. A Christakis. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study commentary: Why do doctors overestimate? Commentary: Prognoses should be based on proved indices not intuition. *BMJ*, 320(7233):469–473, February 2000. ISSN 0959-8138. doi: 10.1136/bmj.320.7233.469.

Ozgun Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Sebastien Ourselin, Leo Joskowicz, Mert R Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46723-8. doi: 10.1007/978-3-319-46723-8_49.

João Rodrigo Ferreira da Silva Sousa, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, and Rodolfo Acatauassú Nunes. Methodology for automatic detection of lung nodules in computerized tomography images. *Computer Methods and Programs in Biomedicine*, 98(1):1–14, 2010. ISSN 01692607. doi: 10.1016/j.cmpb.2009.07.006.

Yaojun Dai, Shiju Yan, Bin Zheng, and Chengli Song. Incorporating automatically learned pulmonary nodule attributes into a convolutional neural network to improve accuracy of benign-malignant nodule classification. *Physics in Medicine & Biology*, 63(24), 12 2018. doi: 10.1109/TMI.2016.2629462.

Antonio Oseas de Carvalho Filho, Wener Borges de Sampaio, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, Rodolfo Acatauassú Nunes, and Marcelo Gattass. Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. *Artificial Intelligence in Medicine*, 60(3):165–177, 2014. ISSN 0933-3657. doi: http://doi.org/10.1016/j.artmed.2013.11.002.

Jamshid Dehmeshki, Xujiong Ye, XinYu Lin, Manlio Valdivieso, and Hamdan Amin. Automated detection of lung nodules in CT images using shape-based genetic algorithm. *Computerized Medical Imaging and Graphics*, 31(6):408–417, 2007. ISSN 08956111. doi: 10.1016/j.compmedimag.2007.03.002.

Annemilia Del Ciello, Paola Franchi, Andrea Contegiacomo, Giuseppe Cicchetti, Lorenzo Bonomo, and Anna Rita Larici. Missed lung cancer: when, where, and why? *Diagnostic and Interventional Radiology*, 23(2):118–126, 2017. ISSN 1305-3612. doi: 10.5152/dir.2016.16187.

Ivan Diaz, Sabine Schmidt, Francis R. Verdun, and François O. Bochud. Eye-tracking of nodule detection in lung CT volumetric data. *Medical Physics*, 42(6):2925–2932, 2015. ISSN 00942405. doi: 10.1118/1.4919849.

Jia Ding, Aoxue Li, Zhiqiang Hu, and Liwei Wang. 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, 2017. doi: 10.48550/arXiv.1706.04303.

Xianling Dong, Shiqi Xu, Yanli Liu, Aihui Wang, M Iqbal Saripan, Li Li, Xiaolei Zhang, and Lijun Lu. Multi-view secondary input collaborative deep learning for lung nodule 3D segmentation. *Cancer Imaging*, 20(1):1–13, 2020. doi: 10.1186/s40644-020-00331-0.

Qi Dou, Hao Chen, Lequan Yu, Jing Qin, and Pheng Ann Heng. Multi-level Contextual 3D CNNs for False Positive Reduction in Pulmonary Nodule Detection. *IEEE Transactions on Biomedical Engineering*, PP(99):1558–1567, 2016. ISSN 15582531. doi: 10.1109/TBME.2016.2613502.

Timothy Dozat. Incorporating Nesterov momentum into Adam. In *Proceedings of the 4th International Conference on Learning Representations*, pages 1–4, 2016. URL https://openreview.net/pdf/OM0jvwB8jIp57ZJjtNEZ.pdf.

Trafton Drew, Melissa L.-H. Vo, A. Olwal, Francine Jacobson, Steven E Seltzer, and Jeremy M Wolfe. Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision*, 13(10):3–3, 8 2013. ISSN 1534-7362. doi: 10.1167/13.10.3.

Andinet A Enquobahrie, Anthony P Reeves, David F Yankelevitz, and Claudia I Henschke. Automated Detection of Small Pulmonary Nodules in Whole Lung CT Scans. *Academic Radiology*, 14(5):579–593, 2007. ISSN 1076-6332. doi: http://doi.org/10.1016/j.acra.2007.01.029.

Jeremy J Erasmus, John E Connolly, H Page McAdams, and Victor L Roggli. Solitary pulmonary nodules: Part I. Morphologic evaluation for differentiation of benign and malignant lesions. *Radiographics*, 20(1):43–58, 2000. doi: 10.1148/radiographics.20.1.g00ja0343.

Jiansheng Fang, Jingwen Wang, Anwei Li, Yuguang Yan, Yonghe Hou, Chao Song, Hongbo Liu, and Jiang Liu. Siamese encoder-based spatial-temporal mixer for growth trend prediction of lung nodules on CT scans. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, page 484–494. Springer Nature Switzerland, 2022. ISBN 9783031164316. doi: 10.1007/978-3-031-16431-6_46.

Amal A Farag, Hossam E Abd El Munim, James H Graham, and Aly A Farag. A novel approach for lung nodules segmentation in chest CT using level sets. *IEEE Transactions on Image Processing*, 22(12):5202–5213, 2013. doi: 10.1109/TIP.2013.2282899.

Xinyang Feng, Jie Yang, Andrew F Laine, and Elsa D Angelini. Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 568–576. Springer, 2017. doi: 10.1007/978-3-319-66179-7_65.

Carlos A Ferreira, António Cunha, Ana Maria Mendonça, and Aurélio Campilho. Convolutional neural network architectures for texture classification of pulmonary nodules. In *Iberoamerican Congress on Pattern Recognition*, pages 783–791. Springer, 2018. doi: 10.1007/978-3-030-13469-3_91.

Carlos A. Ferreira, Guilherme Aresta, António Cunha, Ana Maria Mendonça, and Aurélio Campilho. Wide residual network for Lung-RADS™ screening referral. In *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*, pages 1–4, 2019. doi: 10.1109/ENBENG.2019.8692560.

Carlos A Ferreira, Guilherme Aresta, João Pedrosa, João Rebelo, Eduardo Negrão, António Cunha, Isabel Ramos, and Aurélio Campilho. Classification of lung nodules in CT volumes using the Lung-RADS™ guidelines with uncertainty parameterization. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 791–794. IEEE, 2020. doi: 10.1109/ISBI45749.2020.9098586.

Carlos A Ferreira, Isabel Ramos, Miguel Coimbra, and Aurélio Campilho. A comparative study of feature-based and end-to-end approaches for lung nodule classification in CT volumes to Lung-RADS follow-up recommendation. In *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)*. IEEE, 2024a. doi: 10.1109/MELECON56669.2024.10608773.

Carlos A Ferreira, Célia Sousa, Inês Dias Marques, Pedro Sousa, Isabel Ramos, Miguel Coimbra, and Aurélio Campilho. LNDb v4: pulmonary nodule annotation from medical reports. *Scientific Data*, 11(1):512, 2024b. doi: 10.1038/s41597-024-03345-6.

Carlos A Ferreira, Kiran Vaidhya Venkadesh, Colin Jacobs, Miguel Coimbra, and Aurélio Campilho. Towards automatic forecasting of lung nodule diameter with tabular data and CT imaging. *Biomedical Signal Processing and Control*, 96:106625, 2024c. doi: 10.1016/j.bspc.2024.106625.

Adrian Galdran and Hamid Bouchachia. Residual networks for pulmonary nodule segmentation and texture characterization. In *International Conference on Image Analysis and Recognition*, pages 396–405. Springer, 2020. doi: 10.1007/978-3-030-50516-5_35.

Lucília Galha. "O rastreio do cancro do pulmão pode reduzir em mais de 20% a mortalidade desta doença", Aug 2020. URL https://shorturl.at/ijvG1.

Yadong Gang, Xiongfeng Chen, Huan Li, Hanlun Wang, Jianying Li, Ying Guo, Junjie Zeng, Qiang Hu, Jinxiang Hu, and Haibo Xu. A comparison between manual and artificial intelligence–based automatic positioning in CT imaging for COVID-19 patients. *European Radiology*, 31:6049–6058, 2021. doi: 10.1007/s00330-020-07629-4.

Lee W Goldman. Principles of CT: multislice CT. *Journal of nuclear medicine technology*, 36(2): 57–68, 2008. doi: 10.2967/jnmt.107.044826.

L Gonçalves, J Novo, and A Campilho. Hessian based approaches for 3D lung nodule segmentation. *Expert Systems with Applications*, 61:1–15, 2016. ISSN 0957-4174. doi: http://doi.org/10.1016/j.eswa.2016.05.024.

Henry Gray and Warren H Lewis. *Anatomy of the human body*. Philadelphia,Lea & Febiger,, 1918. doi: 10.5962/bhl.title.20311.

T Guerreiro, G Forjaz, L Antunes, J Bastos, A Mayer, P Aguiar, A Araújo, and C Nunes. Lung cancer survival and sex-specific patterns in portugal: A population-based analysis. *Pulmonology*, 29:S70–S79, 2023. doi: 10.1016/j.pulmoe.2021.09.001.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016. doi: 10.1109/cvpr.2016.90.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998. doi: 10.1109/5254.708428.

Carl Heneghan, Ben Goldacre, and Kamal R Mahtani. Why clinical trial outcomes fail to translate into benefits for patients. *Trials*, 18(1):1–7, 2017. ISSN 1745-6215. doi: 10.1186/s13063-017-1870-2.

Marjolein A. Heuvelmans, Rozemarijn Vliegenthart, Harry J. de Koning, Harry J.M. Groen, Michel J.A.M. van Putten, Uraujh Yousaf-Khan, Carla Weenink, Kristiaan Nackaerts, Pim A. de Jong, and Matthijs Oudkerk. Quantification of growth patterns of screen-detected lung cancers: The NELSON study. *Lung Cancer*, 108:48–54, June 2017a. ISSN 0169-5002. doi: 10.1016/j.lungcan.2017.02.021.

Marjolein A Heuvelmans, Joan E Walter, Rozemarijn Vliegenthart, Peter M A van Ooijen, Geertruida H De Bock, Harry J de Koning, and Matthijs Oudkerk. Disagreement of diameter and volume measurements for pulmonary nodule size estimation in CT lung cancer screening. *Thorax*, 73(8):779–781, October 2017b. ISSN 1468-3296. doi: 10.1136/thoraxjnl-2017-210770.

Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4(1):1–13, 2020. doi: 10.1186/s41747-020-00173-2.

Erskine J Holmes and Rakesh R Misra. *Interpretation of emergency head CT: a practical handbook*. Cambridge University Press, 2017. doi: 10.1017/CBO9780511544804.

Benjamin Hou, Georgios Kaissis, Ronald M Summers, and Bernhard Kainz. Ratchet: Medical transformer for chest X-ray diagnosis and reporting. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, pages 293–303. Springer, 2021. doi: 10.1007/978-3-030-87234-2_28.

Jiachen Hou, Chuan Yan, Ru Li, Qingnan Huang, Xiangsuo Fan, and Fangyu Lin. Lung nodule segmentation algorithm with SMR-UNet. *IEEE Access*, 11:34319–34331, 2023. doi: 10.1109/ACCESS.2023.3264789.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. doi: 10.1109/CVPR.2017.243.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, abs/1904.05342, 2019a. URL http://arxiv.org/abs/1904.05342.

Xia Huang, Wenqing Sun, Tzu-Liang Bill Tseng, Chunqiang Li, and Wei Qian. Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks. *Computerized Medical Imaging and Graphics*, 74:25–36, 2019b. doi: 10.1016/j.compmedimag.2019.02.003.

Michael V Huppmann, William B Johnson, and Marcia C Javitt. Radiation risks from exposure to chest computed tomography. In *Seminars in Ultrasound, CT and MRI*, volume 31, pages 14–28. Elsevier, 2010. doi: 10.1053/j.sult.2009.09.003.

Sarfaraz Hussein, Kunlin Cao, Qi Song, and Ulas Bagci. Risk stratification of lung nodules using 3D CNN-based multi-task learning. In *International conference on information processing in medical imaging*, pages 249–260. Springer, 2017. doi: 10.1007/978-3-030-50516-5_34.

IHME, Global Burden of Disease Study. Tracheal, bronchus & lung cancer, 2019. [dataset] – processed by Our World in Data.

John P. A. Ioannidis. Why most clinical research is not useful. *PLoS Medicine*, 13(6):e1002049, 2016. ISSN 1549-1676. doi: 10.1371/journal.pmed.1002049.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, July 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.3301590.

Julip Jung, Helen Hong, and Jin Mo Goo. Ground-glass nodule segmentation in chest CT images using asymmetric multi-phase deformable model and pulmonary vessel removal. *Computers in biology and medicine*, 92:128–138, 2018. doi: 10.1016/j.compbiomed.2017.11.013.

Jayashree Kalpathy-Cramer, Binsheng Zhao, Dmitry Goldgof, Yuhua Gu, Xingwei Wang, Hao Yang, Yongqiang Tan, Robert Gillies, and Sandy Napel. A comparison of lung nodule segmentation algorithms: methods and results from a multi-institutional study. *Journal of digital imaging*, 29(4):476–487, 2016. doi: 10.1007/s10278-016-9859-z.

Krishna Chaitanya Kaluva, Kiran Vaidhya, Abhijith Chunduru, Sambit Tarai, Sai Prasad Pranav Nadimpalli, and Suthirth Vaidya. An automated workflow for lung nodule follow-up recommendation using deep learning. In *International Conference on Image Analysis and Recognition*, pages 369–377. Springer, 2020. doi: 10.1109/JTEHM.2014.2381213.

Or Katz, Dan Presil, Liz Cohen, Yael Schwartzbard, Sarah Hoch, and Shlomo Kashani. Pulmonary-nodule detection using an ensemble of 3D SE-ResNet18 and DPN68 models. In Aurélio Campilho, Fakhri Karray, and Zhou Wang, editors, *Image Analysis and Recognition*, pages 378–385, Cham, 2020. Springer International Publishing. ISBN 978-3-030-50516-5. doi: 10.1007/978-3-030-50516-5_33.

Ishan Devdatt Kawathekar and Anu Shaju Areeckal. Performance analysis of texture characterization techniques for lung nodule classification. In *Journal of Physics: Conference Series*, page 012045. IOP Publishing, 2022. doi: 10.1109/TMI.2016.2629462.

Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartlomiej Papiez, and Thomas Lukasiewicz. Explaining chest X-ray pathologies in natural language. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 701–713. Springer, 2022. doi: 10.1007/978-3-031-16443-9_67.

Ali Nawaz Khan, Hamdan H Al-Jahdali, Carolyn M Allen, Klaus L Irion, Sarah Al Ghanem, and Shyam Sunder Koteyar. The calcified lung nodule: What does it mean? *Annals of thoracic medicine*, 5(2):67–79, 2010. doi: 10.4103/1817-1737.62469.

Nadeem Yousuf Khanday and Shabir Ahmad Sofi. Deep insight: Convolutional neural network and its applications for COVID-19 prognosis. *Biomedical Signal Processing and Control*, 69: 102814, August 2021. ISSN 1746-8094. doi: 10.1016/j.bspc.2021.102814.

Naji Khosravan, Haydar Celik, Baris Turkbey, Elizabeth C. Jones, Bradford Wood, and Ulas Bagci. A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning. *Medical Image Analysis*, 51(1):101–115, 1 2019. ISSN 13618423. doi: 10.1016/j.media.2018.10.010.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computing Research Repository (CoRR)*, abs/1412.6980, 2015. doi: 10.48550/arXiv.1412.6980.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 60. Curran Associates, Inc., 2012. doi: 10.1145/3065386.

Toshiro Kubota, Anna K Jerebko, Maneesh Dewan, Marcos Salganicoff, and Arun Krishnan. Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models. *Medical Image Analysis*, 15(1):133–154, 2011. doi: 10.1016/j.media.2010.08.005.

B. C. Lassen, C. Jacobs, J. M. Kuhnigk, B. Van Ginneken, and E. M. Van Rikxoort. Robust semi-automatic segmentation of pulmonary subsolid nodules in chest computed tomography scans. *Physics in Medicine and Biology*, 60(3):1307–1323, 2015. ISSN 13616560. doi: 10.1088/0031-9155/60/3/1307.

Kelly Latimer and Timothy Mott. Lung cancer: diagnosis, treatment principles, and screening. *American family physician*, 91(4):250–256, 2015. URL https://www.aafp.org/pubs/afp/issues/2015/0215/p250.html.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.48550/arXiv.1901.08746.

Dana Li, Lea Marie Pehrson, Rasmus Bonnevie, Marco Fraccaro, Jakob Thrane, Lea Tøttrup, Carsten Ammitzbøl Lauridsen, Sedrah Butt Balaganeshan, Jelena Jankovic, Tobias Thostrup Andersen, et al. Performance and agreement when annotating chest X-ray text reports—a preliminary step in the development of a deep learning-based prioritization and detection system. *Diagnostics*, 13(6):1070, 2023. doi: 10.3390/diagnostics13061070.

Q Li, F Li, and K Doi. Computerized Detection of Lung Nodules in Thin-Section CT Images by Use of Selective Enhancement Filters and an Automated Rule-Based Classifier. *Academic Radiology*, 15(2):165–175, 2008. doi: 10.1016/j.acra.2007.09.018.

Yamin Li, Jiancheng Yang, Yi Xu, Jingwei Xu, Xiaodan Ye, Guangyu Tao, Xueqian Xie, and Guixue Liu. Learning tumor growth via follow-up volume prediction for lung nodules. In *Lecture Notes in Computer Science*, page 508–517. Springer International Publishing, 2020. ISBN 9783030597252. doi: 10.1007/978-3-030-59725-2_49.

Ri-qiang Liao, An-wei Li, Hong-hong Yan, Jun-tao Lin, Si-yang Liu, Jing-wen Wang, Jian-sheng Fang, Hong-bo Liu, Yong-he Hou, Chao Song, Hui-fang Yang, Bin Li, Ben-yuan Jiang, Song Dong, Qiang Nie, Wen-zhao Zhong, Yi-long Wu, and Xue-ning Yang. Deep learning-based growth prediction for sub-solid pulmonary nodules on CT images. *Frontiers in Oncology*, 12, October 2022. ISSN 2234-943X. doi: 10.3389/fonc.2022.1002953.

Chia-Ying Lin et al. Combined model integrating deep learning, radiomics, and clinical data to classify lung nodules at chest CT. *La radiologia medica*, pages 1–14, 2023. doi: 10.1007/s11547-023-01730-6.

Kehong Liu. STBi-YOLO: A real-time object detection method for lung nodule recognition. *IEEE Access*, 10:75385–75394, 2022. doi: 10.1109/ACCESS.2022.3192034.

Lihao Liu, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Multi-task deep model with margin ranking loss for lung nodule analysis. *IEEE transactions on medical imaging*, 2019. doi: 10.1109/TMI.2019.2934577.

E. Lopez Torres, E. Fiorina, F. Pennazio, C. Peroni, M. Saletta, N. Camarlinghi, M. E. Fantacci, and P. Cerello. Large scale validation of the M5L lung CAD on heterogeneous CT datasets. *Medical Physics*, 42(4):1477–1489, 4 2015. ISSN 0094-2405. doi: 10.1118/1.4907970.

Heber MacMahon, David P. Naidich, Jin Mo Goo, Kyung Soo Lee, Ann N. C. Leung, John R. Mayo, Atul C. Mehta, Yoshiharu Ohno, Charles A. Powell, Mathias Prokop, Geoffrey D. Rubin, Cornelia M. Schaefer-Prokop, William D. Travis, Paul E. Van Schil, and Alexander A. Bankier. Guidelines for management of incidental pulmonary nodules detected on CT images: From the Fleischner society 2017. *Radiology*, 284(1):228–243, July 2017. ISSN 1527-1315. doi: 10.1148/radiol.2017161659.

Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020. doi: 10.1148/radiol.2019191293.

Sonia Marques, Filippo Schiavo, Carlos A Ferreira, Joao Pedrosa, Antonio Cunha, and Aurelio Campilho. A multi-task CNN approach for lung nodule malignancy classification and characterization. *Expert Systems with Applications*, 184:115469, 2021. doi: 10.1016/j.eswa.2021.115469.

Sen Mei, HuiQin Jiang, and Ling Ma. Yolo-lung: A practical detector based on imporved yolov4 for pulmonary nodule detection. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6. IEEE, 2021. doi: 10.1109/CISP-BMEI53629.2021.9624373.

Temesguen Messay, Russell C Hardie, and Timothy R Tuinstra. Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the Lung Image Database Consortium and Image Database Resource Initiative dataset. *Medical Image Analysis*, 22(1):48–62, 2015. ISSN 1361-8415. doi: http://doi.org/10.1016/j.media.2015.02.002.

Hassan Mkindu, Longwen Wu, and Yaqin Zhao. Lung nodule detection in chest ct images based on vision transformer network with bayesian optimization. *Biomedical Signal Processing and Control*, 85:104866, 2023. doi: 10.1016/j.bspc.2023.104866.

Jan Hendrik Moltz, Lars Bornemann, Jan-Martin Kuhnigk, Volker Dicken, Elena Peitgen, Stephan Meier, Hendrik Bolte, Michael Fabel, Hans-Christian Bauknecht, Markus Hittinger, et al. Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in CT scans. *IEEE Journal of selected topics in signal processing*, 3(1):122–134, 2009. doi: 10.1109/JSTSP.2008.2011107.

Yohei Momoki, Akimichi Ichinose, Keigo Nakamura, Shingo Iwano, Shinichiro Kamiya, Keiichiro Yamada, and Shinji Naganawa. Development of automatic generation system for lung nodule finding descriptions. *Plos one*, 19(3):e0300325, 2024. doi: 10.1371/journal.pone.0300325.

Sudipta Mukhopadhyay. A segmentation framework of pulmonary nodules in lung CT images. *Journal of digital imaging*, 29(1):86–103, 2016. doi: 10.1007/s10278-015-9801-9.

Arjun Nair, Anand Devaraj, Matthew E J Callister, and David R Baldwin. The Fleischner society 2017 and British thoracic society 2015 guidelines for managing pulmonary nodules: keep calm and carry on. *Thorax*, 73(9):806–812, June 2018. ISSN 1468-3296. doi: 10.1136/thoraxjnl-2018-211764.

Ron Niehaus, Daniela Stan Raicu, Jacob Furst, and Samuel Armato. Toward understanding the size dependence of shape features for predicting spiculation in lung nodules for computer-aided diagnosis. *Journal of digital imaging*, 28(6):704–717, 2015. doi: 10.1007/978-3-030-50516-5_34.

Mizuho Nishio, Osamu Sugiyama, Masahiro Yakami, Syoko Ueno, Takeshi Kubo, Tomohiro Kuroda, and Kaori Togashi. 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), 2018. doi: 10.1371/journal.pone.0200721.

Omer Onder, Yasin Yarasir, Aynur Azizova, Gamze Durhan, Mehmet Ruhi Onur, and Orhan Macit Ariyurek. Errors, discrepancies and underlying bias in radiology with case examples: a pictorial review. *Insights into Imaging*, 12(1):1–21, 2021. ISSN 1869-4101. doi: 10.1186/s13244-021-00986-8.

Pia Opulencia, David S Channin, Daniela S Raicu, and Jacob D Furst. Mapping LIDC, RadLex™, and lung nodule image features. *Journal of Digital Imaging*, 24(2):256–270, 2011. ISSN 1618-727X. doi: 10.1007/s10278-010-9285-6.

Eugenio Paci, Donella Puliti, Andrea Lopes Pegna, Laura Carrozzi, Giulia Picozzi, Fabio Falaschi, Francesco Pistelli, Ferruccio Aquilini, Cristina Ocello, Marco Zappa, et al. Mortality, survival and incidence rates in the italung randomised lung cancer screening trial. *Thorax*, 72(9):825–831, 2017. doi: 10.1136/thoraxjnl-2016-209825.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper_files/paper/2009/file/1543843a4723ed2ab08e18053ae6dc5b-Paper.pdf.

João Pedrosa, Guilherme Aresta, Carlos Ferreira, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, et al. LNDb: a lung nodule database on computed tomography. Preprint at https://arxiv.org/abs/1911.08434, 2019.

João Pedrosa, Guilherme Aresta, Carlos Ferreira, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, and Aurélio Campilho. LNDb dataset. *Zenodo*, 2020a.

João Pedrosa, Guilherme Aresta, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, and Aurélio Campilho. LNDetector: A flexible gaze characterisation collaborative platform for pulmonary nodule screening. In *XV Mediterranean Conference on Medical and Biological Engineering and Computing–MEDICON 2019: Proceedings of MEDICON 2019, September 26-28, 2019, Coimbra, Portugal*, pages 333–343. Springer, 2020b. doi: 10.1007/978-3-030-31635-8_40.

Joao Pedrosa, Guilherme Aresta, Carlos Ferreira, Gurraj Atwal, Hady Ahmady Phoulady, Xiaoyu Chen, Rongzhen Chen, Jiaoliang Li, Liansheng Wang, Adrian Galdran, et al. Lndb challenge on automatic lung cancer patient management. *Medical image analysis*, 70:102027, 2021. doi: 10.1016/j.media.2021.102027.

João Pedrosa, Guilherme Aresta, Carlos Ferreira, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, and Aurélio Campilho. LNDb dataset. *Zenodo*, 2023.

João Pedrosa, Guilherme Aresta, and Carlos Ferreira. *Computer-aided lung cancer screening in computed tomography: state-of the-art and future perspectives*, chapter 4. IOP Publishing, January 2022. doi: 10.1088/978-0-7503-3355-9ch4.

Anne-Dominique Pham, Aurélie Névéol, Thomas Lavergne, Daisuke Yasunaga, Olivier Clément, Guy Meyer, Rémy Morello, and Anita Burgun. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. *BMC Bioinformatics*, 15(1):1–10, 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-266.

Paul F Pinsky, Timothy R Church, Grant Izmirlian, and Barnett S Kramer. The national lung screening trial: results stratified by demographics, smoking history, and lung cancer histology. *Cancer*, 119(22):3976–3983, 2013. doi: 10.1002/cncr.28326.

Laila Qadan, Adel Ahmed, and Kusum Kapila. Thyroid ultrasound reports: Deficiencies and recommendations. *Medical Principles and Practice*, 28(3):280–283, 2019. ISSN 1423-0151. doi: 10.1159/000497789.

Jianping Qiao, Yanling Fan, Mowen Zhang, Kunlun Fang, Dengwang Li, and Zhishun Wang. Ensemble framework based on attributes and deep features for benign-malignant classification of lung nodule. *Biomedical Signal Processing and Control*, 79:104217, January 2023. ISSN 1746-8094. doi: 10.1016/j.bspc.2022.104217.

Bowen Qiu, Jacob Furst, Alexander Rasin, Roselyne Tchoua, and Daniela Raicu. Learning latent spiculated features for lung nodule characterization. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1254–1257. IEEE, 2020. doi: 10.1007/978-3-030-50516-5_34.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy, September 2017. URL http://aclweb.org/anthology/W17-6523.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Xavier Rafael-Palou, Anton Aubanell, Mario Ceresa, Vicent Ribas, Gemma Piella, and Miguel A. González Ballester. Prediction of lung nodule progression with an uncertainty-aware hierarchical probabilistic network. *Diagnostics*, 12(11):2639, October 2022. ISSN 2075-4418. doi: 10.3390/diagnostics12112639.

Alexandr Rassadin. Deep residual 3D U-Net for joint segmentation and texture classification of nodules in lung. In *International Conference on Image Analysis and Recognition*, pages 419–427. Springer, 2020. doi: 10.1007/978-3-030-50516-5_37.

James V Rawson and Allen L Pelletier. When to order contrast-enhanced CT. *American family physician*, 88(5):312–316, 2013. URL https://www.aafp.org/pubs/afp/issues/2013/0901/p312.html#abstract.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. URL http://arxiv.org/abs/1804.02767.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. doi: 10.1109/CVPR.2016.91.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 6 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2577031.

A Retico, M E Fantacci, I Gori, P Kasae, B Golosio, A Piccioli, P Cerello, G De Nunzio, and S Tangaro. Pleural nodule identification in low-dose and thin-slice lung computed tomography. *Computers in Biology and Medicine*, 39(12):1137–1144, 2009. ISSN 0010-4825. doi: http://doi.org/10.1016/j.compbiomed.2009.10.005.

Marie-Pierre Revel, Alvine Bissery, Marie Bienvenu, Laetitia Aycard, Catherine Lefort, and Guy Frija. Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? *Radiology*, 231(2):453–458, May 2004. ISSN 1527-1315. doi: 10.1148/radiol.2312030167.

Isabel Rio-Torto, Jaime S Cardoso, and Luis Filipe Teixeira. Parameter-efficient generation of natural language explanations for chest X-ray classification. In *Medical Imaging with Deep Learning*, 2024. URL https://openreview.net/pdf?id=EWbMmmQnVy.

Joana Rocha, António Cunha, and Ana Maria Mendonça. Conventional filtering versus U-Net based models for pulmonary nodule segmentation in CT images. *Journal of medical systems*, 44(4):1–8, 2020. doi: 10.1007/s10916-020-1541-9.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Geoffrey D Rubin. Lung nodule and cancer detection in CT screening. *Journal of thoracic imaging*, 30(2):130, 2015. doi: 10.1097/RTI.0000000000000140.

Pablo Sanchez-Salcedo, Juan Berto, Juan P de Torres, Arantzazu Campo, Ana B Alcaide, Gorka Bastarrika, Jesús C Pueyo, Alberto Villanueva, José I Echeveste, Maria D Lozano, et al. Lung cancer screening: fourteen year experience of the Pamplona early detection program (P-IELCAP). *Archivos de Bronconeumología (English Edition)*, 51(4):169–176, 2015. doi: 10.1016/j.arbr.2015.02.015.

Alex Martins Santos, Antonio Oseas De Carvalho Filho, Aristófanes Corrêa Silva, Anselmo Cardoso De Paiva, Rodolfo Acatauassú Nunes, and Marcelo Gattass. Automatic detection of small lung nodules in 3D CT data using Gaussian mixture models, Tsallis entropy and SVM. *Engineering Applications of Artificial Intelligence*, 36:27–39, 2014. ISSN 09521976. doi: 10.1016/j.engappai.2014.07.007.

Ernst Th Scholten, Colin Jacobs, Bram van Ginneken, Sarah van Riel, Rozemarijn Vliegenthart, Matthijs Oudkerk, Harry J de Koning, Nanda Horeweg, Mathias Prokop, Hester A Gietema, et al. Detection and quantification of the solid component in pulmonary subsolid nodules by semiautomatic segmentation. *European radiology*, 25:488–496, 2015. doi: 10.1007/s00330-014-3427-z.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Arnaud Arindra Adiyoso Setio, C Jacobs, J Gelderblom, and B Van Ginneken. Automatic detection of large pulmonary solid nodules in thoracic CT images. *Medical Physics*, 42(10): 5642–5653, 2015. doi: 10.1118/1.4929562.

Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram Van Ginneken. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Transactions on Medical Imaging*, 35(5): 1160–1169, 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2536809.

Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. 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, 2017. doi: 10.1016/j.media.2017.06.015.

Elham Shakibapour, António Cunha, Guilherme Aresta, Ana Maria Mendonça, and Aurélio Campilho. An unsupervised metaheuristic search approach for segmentation and volume measurement of pulmonary nodules in lung ct scans. *Expert Systems with Applications*, 119:415–428, 2019. doi: 10.1016/j.eswa.2018.11.010.

Shiwen Shen, Simon Han, Deni Aberle, Alex Bui, and William Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Systems with Applications*, 128:84–95, 08 2019. doi: 10.1007/978-3-030-50516-5_34.

Wei Shen, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, 61:663–673, 2017. doi: 10.1007/978-3-030-50516-5_34.

Haoqi Shi, Junguo Lu, and Qianjun Zhou. A novel data augmentation method using style-based GAN for robust pulmonary nodule segmentation. In *2020 Chinese Control And Decision Conference (CCDC)*, pages 2486–2491. IEEE, 2020. doi: 10.1109/CCDC49329.2020.9164303.

Francisco Silva, Tania Pereira, Julieta Frade, José Mendes, Claudia Freitas, Venceslau Hespanhol, José Luis Costa, António Cunha, and Hélder P Oliveira. Pre-training autoencoder for lung nodule malignancy assessment using ct images. *applied sciences*, 10(21):7837, 2020. doi: 10.3390/app10217837.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009. doi: 10.1016/j.ipm.2009.03.002.

Marine Soret, Stephen L Bacharach, and Irene Buvat. Partial-volume effect in PET tumor imaging. *Journal of nuclear medicine*, 48(6):932–945, 2007. doi: 10.2967/jnumed.106.035774.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer, 2020. doi: 10.1007/978-3-030-61377-8_28.

Robert L Spitzer, Jacob Cohen, Joseph L Fleiss, and Jean Endicott. Quantification of agreement in psychiatric diagnosis: A new approach. *Archives of General Psychiatry*, 17(1):83–87, 1967. doi: 10.1001/archpsyc.1967.01730250085012.

Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, February 2021. ISSN 1542-4863. doi: 10.3322/caac.21660.

Denis Tack and Nigel Howarth. *Missed Lung Lesions: Side-by-Side Comparison of Chest Radiography with MDCT*, pages 17–26. Springer International Publishing, Cham, 2019. ISBN 978-3-030-11149-6. doi: 10.1007/978-3-030-11149-6_2.

Tina D Tailor, Sarah Bell, and Ruth C Carlos. The impact of downstream procedures on lung cancer screening adherence. *Journal of the American College of Radiology*, 20(10):969–978, 2023. doi: 10.1016/j.jacr.2023.08.003.

Maxine Tan, Rudi Deklerck, Bart Jansen, Michel Bister, and Jan Cornelis. A novel computer-aided lung nodule detection system for CT images. *Medical Physics*, 38(10):5630, 2011. ISSN 00942405. doi: 10.1118/1.3633941.

Yongqiang Tan, Lawrence H. Schwartz, and Binsheng Zhao. Segmentation of lung lesions on CT scans using watershed, active contours, and Markov random field. *Medical Physics*, 40(4), 2013. ISSN 00942405. doi: 10.1118/1.4793409.

Ning Tang, Rui Zhang, Zeliang Wei, Xicheng Chen, Gaoming Li, Qiuyue Song, Dong Yi, and Yazhou Wu. Improving the performance of lung nodule classification by fusing structured and unstructured data. *Information Fusion*, 88:161–174, 2022. doi: 10.1016/j.inffus.2022.07.019.

Guangyu Tao, Li Zhu, Qunhui Chen, Lekang Yin, Yamin Li, Jiancheng Yang, Bingbing Ni, Zheng Zhang, Chi Wan Koo, Pradnya D. Patil, Yinan Chen, Hong Yu, Yi Xu, and Xiaodan Ye. Prediction of future imagery of lung nodule as growth modeling with follow-up computed tomography scans using deep learning: a retrospective cohort study. *Translational Lung Cancer Research*, 11(2):250–262, February 2022. ISSN 2226-4477. doi: 10.21037/tlcr-22-59.

National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, August 2011. ISSN 1533-4406. doi: 10.1056/nejmoa1102873.

Atsushi Teramoto and Hiroshi Fujita. Fast lung nodule detection in chest CT images using cylindrical nodule-enhancement filter. *International Journal of Computer Assisted Radiology and Surgery*, 8(2):193–205, 3 2013. ISSN 1861-6410. doi: 10.1007/s11548-012-0767-5.

Zhenhuan Tian, Yizhuan Jia, Xuejun Men, and Zhongwei Sun. 3DCNN for pulmonary nodule segmentation and classification. In *International Conference on Image Analysis and Recognition*, pages 386–395. Springer, 2020. doi: 10.1007/978-3-030-50516-5_34.

Guofeng Tong, Yong Li, Huairong Chen, Qingchun Zhang, and Huiying Jiang. Improved U-Net network for pulmonary nodules segmentation. *Optik*, 174:460–469, 2018. doi: 10.1016/j.ijleo.2018.08.086.

Xiaoguang Tu, Mei Xie, Jingjing Gao, Zheng Ma, Daiqiang Chen, Qingfeng Wang, Samuel G Finlayson, Yangming Ou, and Jie-Zhi Cheng. Automatic categorization and scoring of solid, part-solid and non-solid pulmonary nodules in CT images with convolutional neural network. *Scientific reports*, 7(1):1–10, 2017. doi: 10.1038/s41598-017-08040-8.

Muhammad Usman, Byoung-Dai Lee, Shi-Sub Byon, Sung-Hyun Kim, Byung-il Lee, and Yeong-Gil Shin. Volumetric lung nodule segmentation using adaptive ROI with multi-view residual learning. *Scientific Reports*, 10(1):1–15, 2020. doi: 10.1038/s41598-020-69817-y.

Carlijn M van der Aalst, Kevin ten Haaf, and Harry J de Koning. Lung cancer screening: latest developments and unanswered questions. *The Lancet Respiratory Medicine*, 4(9):749–761, 2016. doi: 10.1016/S2213-2600(16)30200-4.

Bram Van Ginneken, Samuel G Armato III, Bartjan de Hoop, Saskia van Amelsvoort-van de Vorst, Thomas Duindam, Meindert Niemeijer, Keelin Murphy, Arnold Schilham, Alessandra Retico, Maria Evelina Fantacci, et al. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. *Medical image analysis*, 14(6):707–722, 2010. doi: 10.1016/j.media.2010.05.005.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. doi: 10.48550/arXiv.1706.03762.

Kiran Vaidhya Venkadesh, Arnaud A. A. Setio, Anton Schreuder, Ernst T. Scholten, Kaman Chung, Mathilde M. W. Wille, Zaigham Saghir, Bram van Ginneken, Mathias Prokop, and Colin Jacobs. Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT. *Radiology*, 300(2):438–447, August 2021. ISSN 1527-1315. doi: 10.1148/radiol.2021204433.

B Wang, X Tian, Q Wang, Y Yang, H Xie, S Zhang, and L Gu. Pulmonary nodule detection in CT images based on shape constraint CV model. *Medical Physics*, 42(3):1241–1254, 2015. ISSN 0094-2405. doi: 10.1118/1.4907961.

Hexi Wang, Ning Xiao, Jina Zhang, Wanting Yang, Yulan Ma, Yao Suo, Juanjuan Zhao, Yan Qiang, Jianhong Lian, and Qianqian Yang. Static–dynamic coordinated transformer for tumor longitudinal growth prediction. *Computers in Biology and Medicine*, 148:105922, September 2022. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2022.105922.

Qingfeng Wang, Xuehai Zhou, Chao Wang, Zhiqin Liu, Jun Huang, Ying Zhou, Changlong Li, Hang Zhuang, and Jie-Zhi Cheng. WGAN-based synthetic minority over-sampling technique: Improving semantic fine-grained classification for lung nodules in CT images. *IEEE Access*, 7: 18450–18463, 2019. doi: 10.1109/ACCESS.2019.2896409.

Shuo Wang, Mu Zhou, Zaiyi Liu, Zhenyu Liu, Dongsheng Gu, Yali Zang, Di Dong, Olivier Gevaert, and Jie Tian. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical Image Analysis*, 40:172–183, 8 2017. ISSN 13618415. doi: 10.1016/j.media.2017.06.014.

Weilun Wang, Goutam Chakraborty, and Basabi Chakraborty. 3D multi-scale DenseNet for malignancy grade classification of pulmonary nodules. In *2020 11th International Conference on Awareness Science and Technology (iCAST)*, pages 1–4. IEEE, 2020. doi: 10.1109/TMI.2016.2629462.

Wenzhe Wang, Yifei Lu, Bian Wu, Tingting Chen, Danny Z Chen, and Jian Wu. Deep active self-paced learning for accurate pulmonary nodule segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 723–731. Springer, 2018. doi: 10.1007/978-3-030-00934-2_80.

Ying Wang, Rob J van Klaveren, Geertruida H de Bock, Yingru Zhao, René Vernhout, Anne Leusveld, Ernst Scholten, Johny Verschakelen, Willem Mali, Harry de Koning, et al. No benefit for consensus double reading at baseline screening for lung cancer with the use of semiautomated volumetry software. *Radiology*, 262(1):320–326, 2012. doi: 10.1148/radiol.11102289.

Qiang Wei, Zongcheng Ji, Yuqi Si, Jingcheng Du, Jingqi Wang, Firat Tiryaki, Stephen Wu, Cui Tao, Kirk Roberts, and Hua Xu. Relation extraction from clinical narratives using pre-trained language models. In *AMIA annual symposium proceedings*, volume 2019, page 1236. American Medical Informatics Association, 2019. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7153059/.

Richard Wender, Elizabeth TH Fontham, Ermilo Barrera Jr, Graham A Colditz, Timothy R Church, David S Ettinger, Ruth Etzioni, Christopher R Flowers, G Scott Gazelle, Douglas K Kelsey, et al. American cancer society lung cancer screening guidelines. *CA: a cancer journal for clinicians*, 63(2):106–117, 2013. doi: 10.3322/caac.21172.

Martin CS Wong, Xiang Qian Lao, Kin-Fai Ho, William B Goggins, and LA Shelly. Incidence and mortality of lung cancer: global trends and association with socioeconomic status. *Scientific Reports*, 7(1):14300, 2017. doi: 10.1038/s41598-017-14513-7.

Dag Wormanns, Karl Ludwig, Florian Beyer, Walter Heindel, and Stefan Diederich. Detection of pulmonary nodules at multirow-detector CT: Effectiveness of double reading to improve sensitivity at standard-dose and low-dose chest CT. *European Radiology*, 15(1):14–22, 2005. ISSN 09387994. doi: 10.1007/s00330-004-2527-6.

Botong Wu, Zhen Zhou, Jianwei Wang, and Yizhou Wang. Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1109–1113. IEEE, 2018. doi: 10.1109/ISBI.2018.8363765.

Wenhao Wu, Lei Gao, Huihong Duan, Gang Huang, Xiaodan Ye, and Shengdong Nie. Segmentation of pulmonary nodules in CT images based on 3D-UNet combined with three-dimensional conditional random field optimization. *Medical Physics*, 47(9):4054–4063, 2020. doi: 10.1002/mp.14248.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. doi: 10.1109/TPAMI.2018.2857768.

Ning Xiao, Yan Qiang, Zijuan Zhao, Juanjuan Zhao, and Jianhong Lian. Tumour growth prediction of follow-up lung cancer via conditional recurrent variational autoencoder. *IET Image Processing*, 14(15):3975–3981, December 2020. ISSN 1751-9667. doi: 10.1049/iet-ipr.2020.0496.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M. Summers. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(03):1, July 2018. ISSN 2329-4302. doi: 10.1117/1.jmi.5.3.036501.

Ran Yang, Dongming Hui, Xing Li, Kun Wang, Caiyong Li, and Zhichao Li. Prediction of single pulmonary nodule growth by CT radiomics and clinical features — a one-year follow-up study. *Frontiers in Oncology*, 12, October 2022. ISSN 2234-943X. doi: 10.3389/fonc.2022.1034817.

Xujiong Ye, Gareth Beddoe, and Greg Slabaugh. Automatic graph cut segmentation of lesions in CT using mean shift superpixels. *International journal of biomedical imaging*, 2010, 2010. doi: 10.1155/2010/983963.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. doi: 10.48550/arXiv.1605.07146.

Jing Zhang, Shi Qiu, Xiaohai Cui, and Ting Liang. A pulmonary nodule spiculation recognition algorithm based on generative adversarial networks. *BioMed Research International*, 2022(1): 3341924, 2022. doi: 10.1155/2022/3341924.

Ling Zhang, Le Lu, Ronald M. Summers, Electron Kebebew, and Jianhua Yao. Convolutional invasion and expansion networks for tumor growth prediction. *IEEE Transactions on Medical Imaging*, 37(2):638–648, February 2018. ISSN 1558-254X. doi: 10.1109/tmi.2017.2774044.

Ling Zhang, Le Lu, Xiaosong Wang, Robert M. Zhu, Mohammadhadi Bagheri, Ronald M. Summers, and Jianhua Yao. Spatio-temporal convolutional LSTMs for tumor growth prediction by learning 4D longitudinal patient data. *IEEE Transactions on Medical Imaging*, 39(4): 1114–1126, April 2020. ISSN 1558-254X. doi: 10.1109/tmi.2019.2943841.

Juan Juan Zhao, Guo-hua Ji, Yong Xia, and Xiao-long Zhang. Cavitary nodule segmentation in computed tomography images based on self–generating neural networks and particle swarm optimisation. *International Journal of Bio-Inspired Computation*, 7(1):62–67, 2015. doi: 10.1504/IJBIC.2015.067999.

Sunyi Zheng, Jiapan Guo, Xiaonan Cui, Raymond N.J. Veldhuis, Matthijs Oudkerk, and Peter M.A. Van Ooijen. Automatic Pulmonary Nodule Detection in CT Scans Using Convolutional Neural Networks Based on Maximum Intensity Projection. *IEEE Transactions on Medical Imaging*, 39(3):797–805, 2020. ISSN 1558254X. doi: 10.1109/TMI.2019.2935553.

Guangbiao Zhou. Tobacco, air pollution, environmental carcinogenesis, and thoughts on conquering strategies of lung cancer. *Cancer Biology & Medicine*, 16(4):700, 2019. doi: 10.20892/j.issn.2095-3941.2019.0180.

Dmitriy Zinovev, Jonathan Feigenbaum, Jacob Furst, and Daniela Raicu. Probabilistic lung nodule classification with belief decision trees. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4493–4498. IEEE, 2011. doi: 10.1109/IEMBS.2011.6091114.

Appendix A

List of Publications

Book Chapters

João Pedrosa, Guilherme Aresta, and **Carlos Ferreira**. *Computer-aided lung cancer screening in computed tomography: state-of the-art and future perspectives*, chapter 4. IOP Publishing, January 2022. doi: 10.1088/978-0-7503-3355-9ch4.

International Journals

Guilherme Aresta, **Carlos Ferreira**, Joao Pedrosa, Teresa Araújo, Joao Rebelo, Eduardo Negrao, Margarida Morgado, Filipe Alves, Antonio Cunha, Isabel Ramos, and Aurelio Campilho. Automatic Lung Nodule Detection Combined With Gaze Information Improves Radiologists' Screening Performance. *IEEE journal of biomedical and health informatics*, 24(10):2894– 2901, 2020. ISSN 21682208. doi: 10.1109/JBHI.2020.2976150.

Joao Pedrosa, Guilherme Aresta, **Carlos Ferreira**, Gurraj Atwal, Hady Ahmady Phoulady, Xiaoyu Chen, Rongzhen Chen, Jiaoliang Li, Liansheng Wang, Adrian Galdran, et al. Lndb challenge on automatic lung cancer patient management. *Medical image analysis*, 70:102027, 2021. doi: 10.1016/j.media.2021.102027.

Sonia Marques, Filippo Schiavo, **Carlos A Ferreira**, Joao Pedrosa, Antonio Cunha, and Aurelio Campilho. A multi-task CNN approach for lung nodule malignancy classification and charac- terization. *Expert Systems with Applications*, 184:115469, 2021. doi: 10.1016/j.eswa.2021. 115469.

João Pedrosa, Guilherme Aresta, **Carlos Ferreira**, Catarina Carvalho, Joana Silva, Pedro Sousa, Lucas Ribeiro, Ana Maria Mendonça, and Aurélio Campilho. Assessing clinical applicability of COVID-19 detection in chest radiography with deep learning. *Scientific Reports*, vol. 12, no. 1, pp. 6596, 2022. doi: 10.1038/s41598-022-10568-3.

Carlos A Ferreira, Célia Sousa, Inês Dias Marques, Pedro Sousa, Isabel Ramos, Miguel Coimbra, and Aurélio Campilho. LNDb v4: pulmonary nodule annotation from medical reports. *Scientific Data*, 11(1):512, 2024. doi: 10.1038/s41597-024-03345-6.

Carlos A Ferreira, Kiran Vaidhya Venkadesh, Colin Jacobs, Miguel Coimbra, and Aurélio Campilho. Towards automatic forecasting of lung nodule diameter with tabular data and CT imaging. *Biomedical Signal Processing and Control*, 96:106625, 2024. doi: 10.1016/j.bspc.2024.106625.

Conferences

Carlos A Ferreira, Guilherme Aresta, João Pedrosa, João Rebelo, Eduardo Negrão, António Cunha, Isabel Ramos, and Aurélio Campilho. Classification of lung nodules in CT volumes using the Lung-RADS™ guidelines with uncertainty parameterization. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 791–794. IEEE, 2020. doi: 10.1109/ISBI45749.2020.9098586.

Diego Wanderley, **Carlos Ferreira**, Aurélio Campilho, and Jorge Silva. Ovarian Structures Detection using Convolutional Neural Networks. In *HCist - International Conference on Health and Social Care Information Systems and Technologies 2021*, vol. 196, pp. 542-549, 2022. doi: 10.1016/j.procs.2021.12.047.

Joao Pedrosa, Guilherme Aresta, **Carlos A. Ferreira**, Ana Maria Mendonça, and Aurélio Campilho. Automatic Label Detection in Chest Radiography Images. In *Proceedings of BIOIMAGING 2021*, pp. 63-69, 2022. doi: 10.5220/0010888100003123.

Rui Castro, Inês Sousa, Fábio Nunes, Jennifer Mancio, Ricardo Fontes-Carvalho, **Carlos Ferreira**, and João Pedrosa. Automated Ventral Cavity Segmentation in Computed Tomography. In *Proceedings of the 29th Portuguese Conference on Pattern Recognition (RECPAD)*, pp. 32-33, 2023. URL https://drive.google.com/file/d/1zem-NYHW5BrJN411ep1cRDM-S4cBR4Ks/view.

Rui Castro, Inês Sousa, Fábio Nunes, Jennifer Mancio, Ricardo Fontes-Carvalho, **Carlos Ferreira**, and João Pedrosa. Automated Visceral and Subcutaneous Fat Segmentation in Computed Tomography. In *2024 IEEE 21st International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2024. doi: 10.1109/ISBI56570.2024.10635784.

Carlos A Ferreira, Isabel Ramos, Miguel Coimbra, and Aurélio Campilho. A comparative study of feature-based and end-to-end approaches for lung nodule classification in CT volumes to Lung-RADS follow-up recommendation. In *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)*. IEEE, 2024. doi: 10.1109/MELECON56669.2024.10608773.

Carlos A Ferreira, Kiran Vaidhya Venkadesh, Colin Jacobs, Miguel Coimbra, and Aurélio Campilho. Unveiling Automatic Diameter Forecasting with Tabular Data and CT Imaging: Insights into Malignant Lung Nodules. In *Proceedings of the 30th Portuguese Conference on Pattern Recognition (RECPAD)*, 2024.

Preprints

Guilherme Aresta, **Carlos A Ferreira**, and João Pedrosa. Artificial Intelligence for COVID-19 detection in X-ray Images. *INESC TEC Science & Society*, vol. 1, no. 1, pp. 34-37, 2020. URL https://science-society.inesctec.pt/pt/index.php/inesctecesociedade/article/download/29/37.

Jianning Li, Zongwei Zhou, Jiancheng Yang, et al., **Carlos A Ferreira**, et al. MedShapeNet – A Large-Scale Dataset of 3D Medical Shapes for Computer Vision. Preprint at https://arxiv.org/abs/2308.16139. 2023.

Patents

Carlos A Ferreira, Alexandre Ferreira, Miguel Coimbra, Nuno Guimarães, Alípio Jorge, and Aurélio Campilho. RePick Annotator. *European Patent (Pending)*, 2024.

Datasets

João Pedrosa, Guilherme Aresta, **Carlos Ferreira**, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, and Aurélio Campilho. LNDb dataset. *Zenodo*, 2020. doi: 10.5281/zenodo.6613714.

João Pedrosa, Guilherme Aresta, **Carlos Ferreira**, Catarina Carvalho, Joana Silva, Pedro Sousa, Lucas Ribeiro, Ana Maria Mendonça, and Aurélio Campilho. Clinical applicability of COVID-19 detection in chest radiography with deep learning radiologist annotations. 2022. doi: 10.25747/342B-GF87.

João Pedrosa, Guilherme Aresta, **Carlos Ferreira**, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, and Aurélio Campilho. LNDb_v4 dataset. *Zenodo*, 2024. doi: 10.5281/zenodo.8348419.