Generalization and Fairness of Deep Learning Models for Lung Cancer Screening



Gonçalo Diogo Gomes Melo
Master in Bioinformatics and Computational Biology
Department of Computer Science
Faculty of Sciences of the University of Porto
2026

Generalization and Fairness of Deep Learning Models for Lung Cancer Screening

Gonçalo Diogo Gomes Melo

Dissertation carried out as part of the Master in Bioinformatics and Computational Biology

Department of Computer Science

2026



Supervisor
Prof. Dr. Hélder Filipe Pinto de Oliveira, Associate Professor, Faculty of Sciences of the University of Porto

Co-supervisor
Prof. Dr. Tânia Maria Pereira Lopes, Assistant Professor, Faculty of Engineering of the University of Porto

External Host Supervisor
Margarida Gonçalves Gouveia, Research Assistant, INESC TEC - Institute for Systems and Computer Engineering, Technology and Science

Acknowledgements

This dissertation would not have been possible without the support, guidance, and encouragement of many people, to whom I would like to express my sincere gratitude.

First, I would like to thank INESC TEC, particularly the VCMI group, and the Faculty of Sciences of the University of Porto (FCUP) for providing the research environment, computational resources, and academic support that made this work possible.

I am deeply grateful to my supervisor, Professor Tânia Pereira, for her guidance, encouragement, and valuable scientific insights throughout this project. My sincere thanks also go to my supervisor, Professor Hélder Oliveira, for his support, constructive feedback, and for continually encouraging me to strengthen the scientific quality of this work.

A very special thank you goes to my mentor, Margarida Gonçalves, whose dedication, patience, and constant availability were instrumental throughout this dissertation. Her attention to detail, practical guidance, and willingness to help solve the countless challenges encountered during this research had an invaluable impact on both this work and my learning experience. I am truly grateful for her continuous support.

I would also like to thank all the members of the VCMI group for creating a collaborative and stimulating research environment.

On a personal note, I am especially grateful to my family, particularly my mother, for her unconditional support, encouragement, and belief in me throughout my academic journey. I would also like to thank my friends for their companionship and for providing balance during the more demanding moments of this dissertation. Finally, a special thank you to my feline companions, Mousse, Chita, and Petra, whose constant company made the long hours of work a little easier.

Finally, I would like to thank everyone who takes the time to read this dissertation. Your interest in this work gives purpose to the effort invested in it, and I sincerely hope that the ideas and findings presented here may prove useful for future research in this field.

Funding

This work was funded by the European Union under the Horizon Europe Programme AI4LUNGS project (Grant Agreement No. 101080756).

Resumo

O cancro do pulmão continua a ser uma das principais causas de mortalidade oncológica em todo o mundo, tornando a deteção precoce através do rastreio por tomografia computadorizada de baixa dose (TCBD) essencial para melhorar os resultados clínicos dos doentes. Os recentes avanços em Deep Learning têm demonstrado resultados promissores na classificação automatizada de nódulos pulmonares; no entanto, as preocupações relativas à generalização dos modelos em diferentes contextos clínicos e aos potenciais enviesamentos demográficos continuam a ser os principais obstáculos à sua adoção clínica.

Esta dissertação investiga o desempenho, a generalização e a equidade de modelos de Aprendizagem Profunda para o rastreio do cancro do pulmão através de tomografia computadorizada (TAC) torácica. Foram avaliadas duas famílias de arquiteturas, incluindo redes neuronais convolucionais baseadas em Redes Residuais (ResNets) e Vision Transformers auto-supervisionados baseados em RadioDINO. As experiências foram realizadas utilizando as coortes de rastreio LUNA25 e Duke Lung Cancer Screening (DLCS), com validação externa adicional no coorte LUNGx. O desempenho foi avaliado tanto ao nível dos nódulos como ao nível dos pacientes, enquanto a equidade foi avaliada em vários atributos demográficos, incluindo género, idade, raça, nível de escolaridade e estatuto tabágico e parâmetros físicos de aquisição de imagem.

Os resultados revelaram que os modelos baseados no RadioDINO alcançaram consistentemente o melhor desempenho preditivo, superando as arquiteturas ResNet convencionais na maioria dos cenários de avaliação. A agregação ao nível do doente melhorou o desempenho do rastreio em comparação com as previsões ao nível do nódulo, particularmente em coortes com menor prevalência de malignidade. No entanto, observou-se uma degradação substancial do desempenho na validação externa, destacando os desafios colocados pela mudança de domínio e pelas diferenças na composição dos conjuntos de dados. Estes efeitos foram especialmente evidentes na coorte LUNGx, onde os modelos treinados em populações de rastreio exibiram transferibilidade limitada para um contexto de diagnóstico.

A análise de equidade revelou que as disparidades demográficas dependiam em grande medida tanto da população-alvo como do cenário de implementação. O género surgiu

como a fonte mais proeminente de enviesamento, particularmente nas avaliações relacionadas com o DLCS dataset, enquanto a idade apresentou um comportamento comparativamente estável ao longo das experiências. Além disso, a combinação de conjuntos de dados para treino não melhorou de forma consistente a equidade e, em alguns casos, amplificou as disparidades demográficas, apesar do aumento da diversidade no treino. Adicionalmente, os modelos revelaram-se altamente sensíveis aos parâmetros de aquisição, com a espessura do corte e a voltagem do tubo a degradarem consistentemente a sensibilidade, embora o treino multi-fonte tenha mitigado com sucesso os enviesamentos associados ao fabricante do scanner.

Em geral, os resultados demonstram que um forte desempenho interno, por si só, é insuficiente para estabelecer a fiabilidade clínica. O desenvolvimento de sistemas de inteligência artificial fiáveis para o rastreio do cancro do pulmão requer a consideração simultânea do desempenho preditivo, da generalização entre conjuntos de dados e da equidade demográfica, a fim de garantir uma implementação robusta e equitativa na prática clínica real.

Palavras-chave: Classificação de Nódulos Pulmonares, Deep Learning, Equidade Algorítmica, Generalização de Domínio, Mudança de Domínio, Rastreio do Cancro do Pulmão, Tomografia Computadorizada

Abstract

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, making early detection through Low-Dose Computed Tomography (LDCT) screening essential for improving patient outcomes. Recent advances in deep learning have demonstrated promising results for automated pulmonary nodule classification; however, concerns regarding model generalization across clinical settings and potential demographic biases remain major barriers to clinical adoption.

This dissertation investigates the performance, generalization, and fairness of deep learning models for lung cancer screening using chest CT imaging. Two families of architectures were evaluated, including convolutional neural networks based on Residual Network (ResNet) and self-supervised Vision Transformers based on RadioDINO. Experiments were conducted using the LUNA25 and DLCS screening cohorts, with additional external validation on the LUNGx cohort. Performance was assessed at both nodule and patient levels, while fairness was evaluated across multiple demographic attributes, including gender, age, race, education level, and smoking status and physical image acquisition parameters.

Results showed that RadioDINO-based models consistently achieved the strongest predictive performance, outperforming conventional ResNet architectures in most evaluation settings. Patient-level aggregation improved screening performance relative to nodule-level predictions, particularly in cohorts with lower malignancy prevalence. However, substantial performance degradation was observed under external validation, highlighting the challenges posed by domain shift and differences in dataset composition. These effects were especially evident on the LUNGx cohort, where models trained on screening populations exhibited limited transferability to a diagnostic setting.

The fairness analysis revealed that demographic disparities were highly dependent on both the target population and deployment scenario. Gender emerged as the most prominent source of bias, particularly in DLCS-related evaluations, whereas age showed comparatively stable behaviour across experiments. Furthermore, combining datasets for training did not consistently improve fairness and, in some cases, amplified demographic disparities despite increasing training diversity. Additionally, models proved highly sensitive to physical acquisition parameters, with slice thickness and tube voltage consistently degrading sensitivity, although multi-source training successfully mitigated scanner manufacturer biases.

Overall, the findings demonstrate that strong internal performance alone is insufficient to establish clinical reliability. The development of trustworthy artificial intelligence systems for lung cancer screening requires the simultaneous consideration of predictive performance, cross-dataset generalization, and demographic fairness to ensure robust and equitable deployment in real-world clinical practice.

Keywords: Algorithmic Fairness,Computed Tomography, Deep Learning, Domain Generalization, Domain Shift, Lung Cancer Screening, Lung Nodule Classification

Contents

Contents vii
List of Figures xi
List of Tables xiii
List of Acronyms xiv

1. Introduction 1
   1.1. Context 1
   1.2. Motivation 2
   1.3. Objectives 2
   1.4. Contributions 3
   1.5. Structure of the Thesis 4
2. Background 5
   2.1. Lung Cancer: Clinical Overview 5
      2.1.1. Lung Cancer Sub-Types 5
      2.1.2. Lung Cancer Risk Factors 6
      2.1.3. Early Diagnosis 6
   2.2. Epidemiology of Lung Cancer 8
      2.2.1. Global Incidence and Mortality 8
      2.2.2. Incidence by Sex 9
      2.2.3. Incidence by Age 11
      2.2.4. Incidence by Ethnicity 12
   2.3. Computed Tomography in Lung Cancer Screening 14
      2.3.1. Computed Tomography (CT): General Principles 14
      2.3.2. LDCT: Characteristics and Differences 16
      2.3.3. Pulmonary Nodules in Screening Contexts 17
      2.3.4. AI Systems in Screening Contexts 18
   2.4. Clinical Motivation for Robust and Equitable AI Systems 20
      2.4.1. Clinical Implications of Limited Generalization 20
      2.4.2. Fairness Considerations in Population-Level Screening 22
3. Literature Review 24
   3.1. Lung Nodule Classification 24
      3.1.1. Public and Private Datasets 24
      3.1.2. Model Architectures 25
         3.1.2.1. Standard Convolutional Networks 25
         3.1.2.2. The Dimensional Dilemma: 2D vs. 3D Approaches 28
         3.1.2.3. Vision Transformers and Hybrid Models 30

3.1.3. Performance and Evaluation Metrics 31
3.1.4. Clinical and Technical Limitations 32
3.2. Generalization in Medical Imaging 32
3.2.1. Domain Shift 33
3.2.2. Sources of Domain Shift in Medical Imaging 34
3.2.3. Domain Adaptation and Domain Generalization 35
3.2.4. Methods to Promote Generalization 36
3.3. Fairness in Medical Imaging 37
3.3.1. Sources and Types of Bias 38
3.3.2. Fairness Metrics 40
3.3.3. Methods to Mitigate Bias 41
3.3.4. Clinical Evidence of Algorithmic Bias 42
3.4. Generalization and Fairness in Lung Cancer Screening 42
3.4.1. Empirical Evidence of Generalization Challenges 43
3.4.2. Fairness Audits in Screening Models 44
3.5. Critical Review of the Literature 45
4. Datasets 47
4.1. LUNA25 47
4.2. DLCS 53
4.2.1. Comparison of CT Acquisition Parameters 56
4.3. LUNGx 57
4.4. Data Pre-processing 59
4.4.1. Spatial Alignment 59
4.4.2. Resampling 60
4.4.3. Intensity Normalization 60
4.4.4. Nodule Localization and 2D Patch Extraction 60
4.4.5. Dataset Partitioning and Class Balance 61
5. Methodology 63
5.1. Data Augmentation and Class Balancing 64
5.1.1. Input Adaptation 64
5.1.2. Data Augmentation 64
5.1.3. Class Balancing Strategy 65
5.2. Model Architecture 66
5.2.1. ResNet-based Models 66
5.2.2. RadioDINO and Transformer-based Models 67
5.2.3. Architectural Considerations 68
5.3. Model Optimization 70
5.3.1. Training Procedure 70

5.3.2. Early Stopping and Model Selection 70
5.3.3. Hyperparameter Optimization 71
5.3.4. Gradual Unfreezing Strategy 71
5.3.5. LUNA25: Cross-Validation Strategy 72
5.3.6. DLCS Dataset: Cross-Validation Strategy 73
5.3.7. Combined Dataset (LUNA25 + DLCS): Cross-Validation Strategy 73
5.3.8. Reproducibility 73
5.4. Evaluation Protocol and Metrics 74
5.4.1. Prediction and Metric Computation 74
5.4.2. Patient-Level Aggregation (Pmax) 75
5.4.3. Generalization Evaluation 76
5.4.4. Fairness and Error Evaluation Protocol 77
5.4.5. Statistical Testing of Fairness 79
6. Results and Discussion 81
6.1. Internal Evaluation 81
6.1.1. ResNet Models 81
6.1.1.1. Nodule-Level Performance 82
6.1.1.2. Patient-Level Performance 84
6.1.2. DINO Models 87
6.1.2.1. Nodule-Level Performance 87
6.1.2.2. Patient-Level Performance 89
6.1.3. Impact of the Training Dataset 90
6.1.3.1. LUNA25-Trained Models 90
6.1.3.2. DLCS-Trained Models 91
6.1.3.3. Jointly-Trained Models (LUNA25 + DLCS) 92
6.1.3.4. Comparison with State-of-the-Art Literature 95
6.1.4. Summary 96
6.2. Cross-Dataset Generalization Analysis 96
6.2.1. LUNA25 $\rightarrow$ DLCS 97
6.2.2. LUNA25 $\rightarrow$ LUNGx 98
6.2.3. DLCS $\rightarrow$ LUNA25 99
6.2.4. DLCS $\rightarrow$ LUNGx 100
6.2.5. Jointly-Trained Models (LUNA25 + DLCS) $\rightarrow$ LUNGx 102
6.2.6. Qualitative and Systematic Error Analysis 104
6.2.7. Summary 108
6.3. Fairness Analysis 109
6.3.1. Gender 110
6.3.2. Age 113

6.3.3. Race 114
6.3.4. Education 115
6.3.5. Smoking Status 115
6.3.6. Hardware and Acquisition Bias 116
6.3.6.1. Scanner Manufacturer 117
6.3.6.2. Reconstruction Slice Thickness 118
6.3.6.3. X-Ray Tube Voltage (kVp) 119
6.3.6.4. Reconstruction Kernels 121
6.3.7. Summary 121
6.4. Limitations 122
7. Conclusion 123
References 126

List of Figures

2.1. Overall survival rates of non-small-cell lung cancer stratified by clinical stage. 7
2.2. CT slice region of interest with a marked pulmonary nodule. 15
2.3. The Hounsfield Unit (HU) scale and typical reference densities for clinical CT windowing. 16
2.4. Morphological classification of screening-detected pulmonary nodules in LDCT. 17
2.5. Conceptual diagram illustrating the self-reinforcing feedback loop of algorithmic bias. 21

3.1. Representative full axial CT slices from the primary datasets. 26
3.2. Macro-architecture of a standard ResNet-50 model. 27
3.3. Standard architecture of a Vision Transformer (ViT). 30
3.4. Illustration of domain shift and adaptation. 33
3.5. The Artificial Intelligence development pipeline and potential sources of bias injection. 38

4.1. Age distribution across benign and malignant diagnoses in the LUNA25 dataset. 49
4.2. Patient distribution by sex and diagnosis in the LUNA25 dataset. 49
4.3. Smoking status history across the patient cohort in the LUNA25 dataset. 50
4.4. Patient distribution by educational attainment across the LUNA25 cohort. 50
4.5. Original racial/ethnic distribution across all categories available in the NLST data. 51
4.6. Racial composition after aggregation into White and Non-White groups for the LUNA25 dataset. 52
4.7. Representative examples of benign and malignant nodules from the LUNA25 dataset. 52
4.8. Age distribution across benign and malignant diagnoses in the DLCS dataset. 54
4.9. Patient distribution by sex and diagnosis in the DLCS cohort. 54
4.10. Smoking status history across the patient cohort in the DLCS dataset. 55
4.11. Racial composition illustrating demographic representation in the DLCS dataset. 55
4.12. Representative examples of benign and malignant nodules from the DLCS dataset. . 56
4.13. Comparative EDA of CT acquisition parameters for LUNA25 and DLCS. 57
4.14. Class distribution of benign and malignant nodules in the LUNGx dataset. 58
4.15. Representative examples of benign and malignant nodules from the LUNGx dataset. 59
4.16. Overview of the standardized data pre-processing pipeline applied to all datasets. ... 60

5.1. Unified experimental pipeline detailing the progression from raw data to fairness auditing. 63
5.2. Representative examples of the stochastic data augmentation pipeline. 65
5.3. Structural comparison of the residual blocks used in the CNN architectures. 67
5.4. Architectural overview of the Vision Transformer (ViT) backbone used by RadioDINO. 69

6.1. Comparison of nodule-level AUC performance between LUNA25 and DLCS for ResNet architectures. 84
6.2. Comparison of Nodule vs. Patient-level AUC on LUNA25. 85
6.3. Comparison of Nodule vs. Patient-level AUC on DLCS. 86
6.4. Comparison of nodule-level AUC between ResNet and RadioDINO families on LUNA25. 88
6.5. Comparison of nodule-level AUC between ResNet and RadioDINO families on DLCS. 89

6.6. Qualitative sampling of unanimous model predictions on the LUNA25 cohort. 105
6.7. Systematic error profiling of unanimous model predictions based on physical and spatial metadata. 106
6.8. Heatmap of cross-dataset generalization gaps (Δ AUC). 109
6.9. Bias intensity map of statistically significant disparities across demographic attributes. 110
6.10. Performance collapse of the RadioDINO-S8 architecture across genders. 112
6.11. The impact of dataset merging on gender fairness evaluated on the DLCS cohort. ... 112
6.12. Specificity disparity across racial subgroups for the RadioDINO-Small model. 114
6.13. Sensitivity disparity based on smoking history for ResNet-18 evaluated out-of-distribution. 116
6.14. Hardware fairness scoreboard for the LUNA25 to DLCS cross-dataset scenario. 117
6.15. Hardware fairness scoreboard for the Merged to DLCS cross-dataset scenario. 118
6.16. Sensitivity disparity across scanner manufacturers. 119
6.17. Impact of spatial resolution on model sensitivity. 120
6.18. Impact of X-Ray tube voltage on model sensitivity. 120

List of Tables

3.1. Summary of key public datasets for lung nodule classification. 27
3.2. Comparison of 2D and 3D CNN Architectures for Lung Nodule Classification. 29

4.1. Summary of demographic and clinical characteristics of the LUNA25 dataset. 48
4.2. Summary of demographic and clinical characteristics of the DLCS dataset. 53
4.3. Distribution of annotated nodules per patient in the LUNGx dataset. 58
4.4. Distribution of target label and demographic attributes across train+val and test splits. 62

5.1. Hyperparameter search space for the ResNet models. 71
5.2. Hyperparameter search space for the RadioDINO experiments. 71

6.1. Internal test set performance of ResNet-based models on LUNA25. 81
6.2. Internal test set performance of ResNet-based models on DLCS. 82
6.3. Internal test set performance of RadioDINO-based models on LUNA25. 87
6.4. Internal test set performance of RadioDINO-based models on DLCS. 87
6.5. Summary of the best-performing model per architecture family and training dataset. . 90
6.6. Performance of jointly-trained models evaluated on the LUNA25 internal test set. 92
6.7. Performance of jointly-trained models evaluated on the DLCS internal test set. 93
6.8. Patient-level absolute performance of LUNA25-trained models on the DLCS test set. 97
6.9. Generalization gaps for LUNA25-trained models evaluated on DLCS. 97
6.10. Patient-level absolute performance of LUNA25-trained models on LUNGx. 98
6.11. Generalization gaps for LUNA25-trained models evaluated on LUNGx. 98
6.12. Patient-level absolute performance of DLCS-trained models on the LUNA25 test set. 99
6.13. Generalization gaps for DLCS-trained models evaluated on LUNA25. 99
6.14. Patient-level absolute performance of DLCS-trained models on LUNGx. 101
6.15. Generalization gaps for DLCS-trained models evaluated on LUNGx. 101
6.16. Absolute patient-level performance of jointly-trained models evaluated on the external LUNGx dataset. 102
6.17. Executive summary of cross-dataset generalization performance. 108
6.18. Summary of statistically significant bias detections per demographic attribute and evaluation scenario. 110

List of Acronyms

ADC Adenocarcinoma. 6, 8, 10, 12, 13

AI Artificial Intelligence. 1, 2, 5, 18, 20– 23

AUC Area Under The Curve. 31, 74

CNN Convolutional Neural Network. 2, 24, 25, 27, 30, 66

CT Computed Tomography. xi, 3, 14– 18, 24

DA Domain Adaptation. 35

DG Domain Generalization. 35, 36

DL Deep Learning. 2, 3, 18, 19, 24, 39, 66

DLCS Duke Lung Cancer Screening. iii, v, xi, xiii, 3, 25, 47, 53– 60, 64, 65, 73, 76, 77

HU Hounsfield Unit. 14– 16

INESC TEC Institute for Systems and Computer Engineering, Technology and Science. 3

LCC Large Cell Carcinoma. 6

LDCT Low-Dose Computed Tomography. v, xi, 1, 5, 10, 14, 16– 18

LIDC-IDRI Lung Image Database Consortium and Image Database Resource Initiative. 24, 25, 28, 29, 31

LNDb Lung Nodule Database. 25

ML Machine Learning. 29, 37, 38, 40, 41, 45

NELSON Nederlands-Leuvens Longkanker Screenings Onderzoek. 10

NIAID National Institute of Allergy and Infectious Diseases. 28

NLST National Lung Screening Trial. 7, 14, 25, 28, 47

NSCLC Non-small-cell lung cancer. 6, 28

ResNet Residual Network. v, 66

SCC Squamous Cell Carcinoma. 6, 8, 10

SOTA State of the Art. 4

VCMI Visual Computing and Machine Intelligence. 3

ViTs Vision Transformers. 24, 30, 31, 67

1. Introduction

This chapter introduces the fundamental context and the core problem addressed in this dissertation. It begins by establishing the clinical landscape of lung cancer screening and the emerging challenges of generalization and fairness in Artificial Intelligence (AI). Following this, the motivation for the study is outlined, leading directly into the specific research questions and the primary scientific contributions of this work. Finally, it provides an overview of the document's structure.

1.1. Context

Ideally, lung cancer is a disease that should be found, not felt. Yet, for the majority of patients, diagnosis occurs only when the disease has advanced beyond a cure [1]. This devastating reality cements lung cancer as the leading cause of cancer mortality globally, accounting for approximately 1.8 million deaths annually, a toll heavier than that of breast, brain, and prostate cancers combined [2]. This high mortality is largely due to late diagnosis, often at advanced stages of the disease. This reality reinforces the importance of screening and early detection, as survival improves substantially when the disease is identified at an early stage. LDCT has thus become the central modality in screening programmes, offering high sensitivity for the detection of pulmonary nodules, which may represent malignant or benign lesions [3].

Despite its effectiveness, manual interpretation of LDCT scans has well-documented limitations. The diagnostic process is complicated and susceptible to human error due to radiologist fatigue, the high frequency of benign nodules, overlapping anatomical structures, and variation between observers. Furthermore, histological confirmation by biopsy, although the definitive method, is invasive, carries risks, and may fail due to tumour heterogeneity [4]. Together, these factors place significant pressure on healthcare systems, amplified by a shortage of radiologists and growing demand associated with screening programmes.

Associated with generalization limitations, a growing challenge in the literature is fairness in medical AI. Several studies show that models can perform differently between subgroups, either due to inequality in demographic representation, variability in image acquisition, or reliance on spurious correlations (shortcuts). In particular, there is consistent evidence that fairness measured in a training set does not necessarily transfer to

other clinical or institutional contexts, raising concerns about the safe and equitable use of these models [5]. In the case of pulmonary nodules, recent studies show that the greatest disparities in performance may be associated with intrinsic characteristics of the nodules, such as size or subtype, rather than the demographic attributes of the patient [6].

Thus, the need for robust and equitable systems becomes evident. It is essential to ensure that AI models maintain consistent performance across different populations and in the face of nodule diversity, in order for them to be used for screening and diagnostics. This thesis focuses precisely on this convergence between generalization and fairness, seeking to understand and mitigate the limitations that compromise the reliability and ethical integrity of Deep Learning (DL) models applied to pulmonary nodule classification in the context of lung cancer screening.

1.2. Motivation

In this context, methods based on AI, particularly DL models, emerge as promising solutions to support diagnosis. Convolutional Neural Network (CNN)s, applied to the analysis of pulmonary nodules, demonstrate the ability to automate detection and classification tasks, reducing the burden on specialists and increasing diagnostic accuracy in various clinical scenarios [4]. However, their integration into real clinical environments remains limited by problems of reproducibility, dependence on large amounts of annotated data, and difficulties in generalizing to populations or centres other than those used in training [7].

In lung cancer screening, this issue is further amplified by the size of the target population and the potential impact of biased or inaccurate predictions. Therefore, ensuring that DL models are accurate, robust and equitable across diverse settings is a fundamental requirement for their clinical adoption. The goal of this dissertation is to critically evaluate, study and improve the generalization and fairness properties of DL models applied to lung cancer screening, moving beyond isolated or classic performance metrics towards more reliable, clinically meaningful evaluation frameworks [5].

1.3. Objectives

This thesis aims to investigate the generalization and fairness properties of DL models applied to lung cancer screening. The central objective is to assess whether classification models can maintain reliable and equitable performance when deployed across diverse

populations and clinical environments, specifically evaluating the impact of demographic and technical attributes on model behaviour.

More specifically, this work pursues the following high-level objectives:

1. Develop and evaluate pulmonary nodule malignancy classification models, comparing distinct representation learning paradigms to identify architectures resilient to data-constrained medical environments.
2. Assess algorithmic generalization under distribution shifts, investigating how varying levels of institutional and clinical heterogeneity impact model reliability when transitioning from screening to diagnostic settings.
3. Conduct a comprehensive fairness and bias audit, identifying whether structural performance disparities exist across different demographic subgroups and determining how physical imaging variations influence diagnostic equity.

1.4. Contributions

This thesis was developed at the Visual Computing and Machine Intelligence (VCMI) group at Institute for Systems and Computer Engineering, Technology and Science (INESC TEC) and contributes to the field of medical image analysis and lung cancer screening by providing a systematic empirical study of generalization and fairness issues in DL-based pulmonary nodule classification.

The main contributions of this work are:

1. The implementation of a unified and reproducible evaluation methodology that jointly measures predictive performance, cross-dataset generalization, and subgroup fairness at both the nodule and patient levels, mitigating data leakage through rigorous cross-validation strategies.
2. A comprehensive comparative analysis demonstrating the superior out-of-distribution robustness of self-supervised Vision Transformers (RadioDINO) over standard convolutional networks (ResNets) across three distinct chest CT cohorts (LUNA25, DLCS, and LUNGx).
3. The empirical demonstration that multi-source training (merging LUNA25 and DLCS datasets) functions as a mechanism to mitigate hardware-induced domain shifts

(e.g., scanner manufacturer biases), but fails to act as a universal solution for underlying demographic inequalities.

4. A large-scale algorithmic fairness audit revealing that gender constitutes a prominent source of bias due to ”shortcut learning,” and that variations in physical acquisition parameters (such as slice thickness ≤ 1.5 mm and tube voltage ≤ 100 kVp) systematically degrade diagnostic sensitivity.

1.5. Structure of the Thesis

Besides this introduction, this dissertation thesis has 6 more chapters. Chapter 2 focuses more on the clinical background, addressing the epidemiology of lung cancer with a particular emphasis on the disparities in incidence and mortality by gender, ethnicity, and age. Chapter 3 presents the State of the Art (SOTA), reviewing the methods available for nodule classification in the literature, as well as current concepts and challenges of fairness in medical imaging. Chapter 4 describes the methodology used, detailing the datasets used in the study and the methods implemented. Chapter 5 presents the experimental results obtained, along with the critical discussion that forms the basis for the analysis of model performance and fairness. Finally, Chapter 6 presents a summary of the main conclusions of the study and proposes guidelines for future research.

2. Background

This chapter establishes the clinical and technical basis of the thesis. It begins with an overview of lung cancer epidemiology and the rationale behind LDCT screening. It then examines the characteristics of pulmonary nodules and the challenges of human interpretation. Finally, it discusses the motivation for integrating AI into screening workflows, emphasising the key issues of generalisation and fairness.

2.1. Lung Cancer: Clinical Overview

Lung cancer is a biologically aggressive malignancy characterised by disruptions in the regulatory mechanisms that control cell proliferation and apoptosis. This leads to the formation of abnormal cellular clones with the capacity for tissue invasion and systemic dissemination through lymphatic and haematogenous pathways [8, 9]. This invasive and metastatic potential is the main determinant of morbidity and mortality, distinguishing malignant lung tumours from benign proliferative lesions.

The clinical severity of lung cancer is largely determined by its natural progression. Once tumour cells acquire the ability to spread beyond the primary site, curative treatment becomes much more challenging and long-term survival rates decrease significantly. Consequently, despite advances in surgery, radiotherapy, targeted therapy and immunotherapy, lung cancer continues to be associated with poor overall outcomes. Global survival estimates remain low, with five-year survival rates often reported below 15% in population-level analyses. This reflects both the aggressive nature of the disease and the challenges associated with achieving a timely diagnosis [10].

2.1.1 Lung Cancer Sub-Types

Different histological subtypes exhibit distinct biological behaviours. One of the most aggressive forms is small-cell lung cancer (SCLC), which is characterised by rapid cellular proliferation, early metastatic spread and an extremely poor prognosis. Five-year survival rates are typically below 10% [11]. This aggressive profile is also reflected in its short tumour volume doubling time of around 73 days, which indicates rapid disease progression [12]. Despite its clinical severity, SCLC accounts for a minority of cases, representing approximately 15% of all lung cancers.

Non-small-cell lung cancer (NSCLC) accounts for the majority of cases, representing approximately 85% of lung cancers, and generally grows more slowly, yet still carries a substantial disease burden. Tumour doubling times are longer, typically ranging from around 140 to 223 days, although prognosis remains strongly dependent on the stage at diagnosis [12]. NSCLC comprises three main histological subtypes: Adenocarcinoma (ADC), Squamous Cell Carcinoma (SCC), and Large Cell Carcinoma (LCC). ADC is the most common subtype, accounting for roughly 40–50% of all lung cancers, followed by SCC at approximately 20–30%, while LCC is less frequent [13]. These subtypes exhibit distinct growth patterns and clinical behaviours; for example, LCC is an undifferentiated variant associated with particularly aggressive progression and poorer outcomes compared to other NSCLC subtypes [13].

2.1.2 Lung Cancer Risk Factors

From an epidemiological perspective, the disease is strongly associated with tobacco exposure, accounting for an estimated 85–90% of cases [14]. While tobacco exposure plays a central role in incidence patterns, the clinical impact of lung cancer is ultimately defined by its aggressive progression and the difficulty of achieving early diagnosis.

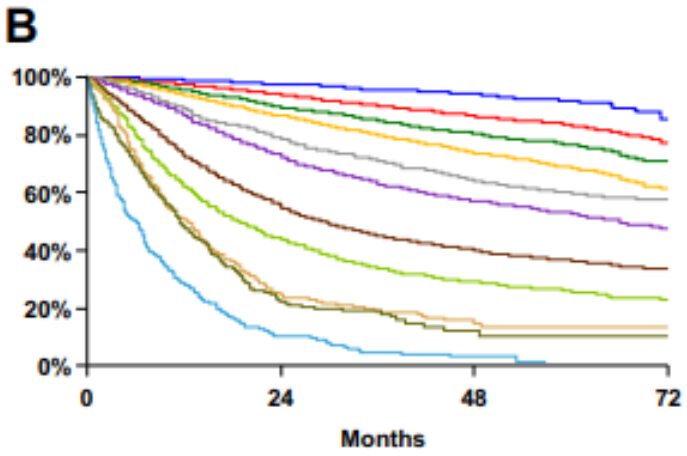
Beyond tobacco, occupational exposure to carcinogens (e.g., asbestos, silica, diesel exhaust, radon) is recognised by the IARC as the second major modifiable risk factor for lung cancer [15]. It is estimated that roughly 10% of lung cancer deaths among men worldwide are attributable to such exposures [15], with reported occupational odds ratios typically ranging between 1.4 and 1.5 after adjustment for smoking [16]. Critically, these hazardous exposures are unevenly distributed across the population: individuals with lower educational attainment are disproportionately employed in manual occupations entailing contact with these agents, and the ICARE study estimated a substantial mediation effect of occupational exposures on the education–lung cancer relationship in less-educated men [17]. This is relevant to the present work, as it offers a plausible explanation for performance disparities observed across educational strata in the screening cohort.

2.1.3 Early Diagnosis

A key factor contributing to the high mortality rate of lung cancer is frequent diagnosis at an advanced stage. It is estimated that a large proportion of patients are identified only after the disease has progressed locally or metastasised, which significantly limits curative treatment options [10]. This late diagnosis, combined with the aggressive nature

of certain subtypes, helps to explain why lung cancer remains one of the leading causes of cancer-related death worldwide.

The rationale for lung cancer screening is primarily based on the well-established link between the stage at which the disease is diagnosed and survival rates. In National Lung Screening Trial (NLST) in particular, early-stage disease is associated with significantly better outcomes, while advanced stages are characterised by limited treatment options and substantially reduced survival (Figure 2.1) [18, 19]. Therefore, detecting and classifying tumours before they spread throughout the body is a critical clinical objective.



Proposed	Events / N	MST	24 Month	60 Month
IA1	68 / 781	NR	97%	92%
IA2	505 / 3105	NR	94%	83%
IA3	546 / 2417	NR	90%	77%
IB	560 / 1928	NR	87%	68%
IIA	215 / 585	NR	79%	60%
IIB	605 / 1453	66.0	72%	53%
IIIA	2052 / 3200	29.3	55%	36%
IIIB	1551 / 2140	19.0	44%	26%
IIIC	831 / 986	12.6	24%	13%
IVA	336 / 484	11.5	23%	10%
IVB	328 / 398	6.0	10%	0%

Figure 2.1: Overall survival by clinical stage according to the International Association for the Study of Lung Cancer (IASLC) 8th Edition proposals. There is a marked stratification in prognosis: early-stage disease (Stage IA1, top blue line) is associated with 5-year survival rates exceeding 90%, whereas metastatic disease (Stage IVB, bottom light blue line) has a dismal prognosis with survival rates near 0%. This disparity highlights the critical need for early detection. Abbreviations and terms: N, total number of patients; MST, median survival time; NR, not reached; Events, number of observed deaths; IA1–IVB, clinical disease stages. [20].

One of the major challenges in achieving an early diagnosis of lung cancer is that it frequently develops without any clear symptoms during the initial stages. ADC, the most common subtype, often arises in peripheral lung regions and can remain clinically silent for extended periods, thereby delaying detection [13]. Even centrally located tumours, such as SCC, tend to produce symptoms only after they have grown large enough to obstruct major airways, meaning that clinical presentation often occurs at a relatively advanced stage [13]. Consequently, relying solely on symptom-based detection is insufficient for identifying the disease at an early stage.

This biological pattern strongly justifies imaging-based screening. As the capacity for metastasis can develop before symptoms appear, proactive detection through screening is one of the few viable strategies for identifying the disease at a stage when it can be cured [19].

Overall, lung cancer poses a significant public health concern due to its high biological aggressiveness, early metastatic potential and low survival rates. The fact that prognosis is strongly dependent on disease stage highlights the importance of detecting tumours before they progress beyond a localised form, which provides the clinical foundation for early detection strategies.

2.2. Epidemiology of Lung Cancer

2.2.1 Global Incidence and Mortality

Lung cancer is currently one of the greatest public health challenges worldwide, being the leading cause of cancer incidence and mortality. It is estimated that there are approximately two million new diagnoses and approximately 1.8 million deaths annually, representing about 18% of all cancer deaths globally [2]. According to GLOBOCAN data from 2018, lung cancer remained the most lethal malignant neoplasm in the world, with more than two million new cases reported that year.

Despite being only the second most diagnosed cancer in men and women, after prostate and breast cancer, respectively - lung cancer has a disproportionately high mortality rate, reflecting its late diagnosis and biological aggressiveness. The overall five-year survival rate remains low, ranging from 19% to 23%, in contrast to other common cancers, such as breast cancer, where the five-year survival rate exceeds 90% [21, 22].

The global burden of the disease varies greatly across regions, with incidence rates differing by more than twentyfold between regions. The highest rates are observed in countries

with high smoking prevalence, including several developing and rapidly industrialising nations such as China, Russia and India. In these contexts, increased access to tobacco, combined with exposure to environmental and occupational pollutants, has contributed to a growing trend in lung cancer incidence and mortality [2]. In contrast, many Western countries have seen a gradual decline in incidence and mortality, in line with progress in smoking cessation policies.

Smoking remains the primary modifiable risk factor, accounting for over 80% of lung cancer cases in the Western world. However, other factors are playing an increasingly significant role in the global disease burden, particularly in low- and middle-income countries. These include environmental and occupational exposures to substances such as radon, asbestos and arsenic, as well as air pollution. The latter is classified as a Group I carcinogen. In addition, comorbidities such as chronic obstructive pulmonary disease, as well as infections such as tuberculosis and HIV, are associated with a significant increase in the risk of developing lung cancer [2, 21].

Consistently, most cases are diagnosed at advanced or metastatic stages globally and across different healthcare systems, where treatment options are limited and survival rates are extremely low. However, when lung cancer is diagnosed at an early, localised stage, the survival rate is substantially higher, highlighting the potential impact of early detection in reducing mortality [2, 21].

Despite therapeutic advances in recent decades, including targeted therapies and immunotherapy, improvements in overall mortality have been modest, primarily due to ongoing obstacles to screening, primary prevention and healthcare. These limitations disproportionately affect socially and economically disadvantaged groups, perpetuating disparities in clinical outcomes and emphasising the importance of integrated global strategies that combine prevention, early detection and equitable access to care [21, 22].

2.2.2 Incidence by Sex

The incidence of lung cancer varies significantly between the sexes, with men being diagnosed at a rate more than twice that of women overall. This pattern largely reflects historical differences in smoking behaviour: men adopted tobacco use earlier and more intensely than women throughout the 20th century [2]. Consequently, men have a substantially higher cumulative lifetime risk (3.8%) than women (1.77%).

Despite this global disparity, there have been significant divergences in temporal trends. In developed countries, there has been a sustained decline in the incidence and mortality of lung cancer among men, which has been paralleled by a reduction in smoking prevalence. In contrast, lung cancer rates among women continue to rise, reflecting their later adoption of smoking and slower cessation. For example, this phenomenon culminated in an epidemiological milestone in Europe in 2017, when lung cancer surpassed breast cancer to become the leading cause of cancer death among women [2].

In addition to behavioural factors, there is growing evidence of relevant biological differences between the sexes. Women are more likely to develop lung ADC, which is often located in the periphery of the lung and less strongly associated with smoking, and is also the most common subtype in never-smokers. In contrast, SCC are more prevalent in men, typically located centrally and strongly associated with tobacco use [2].

The molecular profile of lung cancer also varies significantly between men and women. In particular, women have a higher frequency of actionable driver mutations, including EGFR, ALK, BRAF and ROS1 gene mutations. These differences suggest distinct biological susceptibility, which is potentially modulated by genetic and hormonal factors. The overexpression of oestrogen receptors in lung ADCs, alongside the link between reproductive factors and disease risk, supports the theory that female sex hormones play a role in lung carcinogenesis [2].

A particularly relevant aspect from an epidemiological point of view is the incidence of lung cancer in women who have never smoked. Several studies indicate that women with no history of smoking are more than twice as likely to develop lung cancer compared to men who have never smoked, a phenomenon that is particularly evident in Asian-American populations, where a significant proportion of women diagnosed have never smoked. In these cases, exposures such as second-hand smoke and inhalation of fumes from cooking oils at high temperatures have been identified as potential contributing factors [2, 22].

Even though men continue to have higher absolute mortality rates, women demonstrate, on average, better clinical outcomes, including higher five-year survival rates and a greater likelihood of diagnosis at earlier stages. Preliminary results from screening trials, such as Nederlands-Leuvens Longkanker Screenings Onderzoek (NELSON), also suggest that the relative benefit of LDCT screening may be greater in women, despite their underrepresentation in large clinical trials [21]. These data reinforce the importance of considering

sex as a central variable in the epidemiology of lung cancer and raise important limitations in current screening guidelines, which do not adequately incorporate biological and behavioural differences between men and women.

2.2.3 Incidence by Age

Lung cancer predominantly affects older adults, with age being one of the strongest determinants of incidence. The mean age at diagnosis across populations is approximately 70 years for both sexes, reflecting the long latency period between exposure to carcinogenic factors and the manifestation of clinical disease [2]. Epidemiological data indicate that the majority of cases occur after midlife: approximately 53% of diagnoses arise between the ages of 55 and 74, and a further 37% occur in individuals over the age of 75. By contrast, only around 10% of lung cancer cases are diagnosed in individuals under 55 years of age.

The age-dependent distribution of lung cancer incidence can largely be explained by biological mechanisms associated with cumulative exposure and ageing. Tumourigenesis usually develops decades after initial exposure to mutagenic agents such as tobacco smoke or environmental carcinogens. Age-related biological changes, including telomere shortening, declining NAD+ levels and reduced DNA repair capacity, further increase the vulnerability of older individuals to malignant transformation [2]. Consequently, lung cancer remains substantially more prevalent in populations over the age of 60.

Nevertheless, certain subgroups deviate from this general pattern, presenting with an earlier onset of disease. Individuals from certain racial and socioeconomically disadvantaged groups are diagnosed at a younger age, which highlights the interaction between biological, environmental and structural factors. For instance, in the United States, Black individuals have a lower median age at diagnosis (67 years) than White individuals (70 years), despite typically starting smoking later in life [21]. Similarly, populations residing in rural areas tend to start smoking at a younger age, which could cause the incidence of lung cancer to shift towards earlier decades.

Early-onset lung cancer is particularly prevalent among patients living with HIV, with diagnosis occurring 25 to 30 years earlier than in the general population. In this group, the mean age at diagnosis is between 38 and 57 years, compared to around 70 years in the general population. These findings highlight the importance of immunological and comorbidity-related factors in accelerating disease onset, regardless of chronological ageing [21].

Although relatively rare, lung cancer in younger adults (aged approximately 20-46) is a distinct clinical entity. In these patients, the disease is more often associated with hereditary genetic alterations than with cumulative environmental exposures, and it usually presents as ADC. Despite being diagnosed at more advanced stages, younger patients are more likely to receive aggressive treatment and generally have better survival outcomes than older individuals [2].

Age also plays a critical role in eligibility for lung cancer screening, which has important implications for incidence estimates. A significant number of people who develop lung cancer are deemed ineligible for screening solely because of age restrictions. Data from the Framingham study suggest that almost 30% of cases ineligible for screening fall outside the recommended age range rather than failing to meet the smoking-related criteria [21]. Notably, age-adjusted incidence differences between Black and White smokers peak in the 50-54 age group, which has historically been below the lower age threshold for screening. Lowering the minimum screening age to 50 has been shown to markedly increase eligibility among high-risk groups, and consensus guidelines already recommend this for individuals with additional risk factors, including occupational exposure.

Together, these findings demonstrate that although lung cancer incidence increases dramatically with age, there is substantial heterogeneity between populations. The earlier onset of disease in certain demographic and clinical groups highlights the limitations of age-based risk stratification, emphasising the need for more nuanced approaches that consider not only age, but also biological, environmental and social factors.

2.2.4 Incidence by Ethnicity

The incidence of lung cancer varies significantly across racial and ethnic groups, reflecting a complex interplay of behavioural, environmental, biological and structural determinants. In the United States, some of the most striking disparities in cancer epidemiology are evident in lung cancer, with minority ethnic groups frequently experiencing a disproportionate burden of the disease.

Among men, African American individuals have the highest incidence rates of lung cancer. Age-adjusted estimates indicate that Black men have a higher incidence rate than White men and a substantially higher rate than Hispanic and Asian/Pacific Islander populations. This disparity persists across smoking categories, including among never-smokers, suggesting that factors beyond cumulative tobacco exposure contribute to the increased risk

[2, 21]. African American men also experience higher lung cancer-specific mortality rates, which is consistent with both higher incidence rates and a more advanced stage at diagnosis.

Although incidence patterns differ among women, they remain unequal. White women have the highest rates, followed closely by African American women. Hispanic and Asian/-Pacific Islander women have substantially lower rates, often less than half those of White women. However, certain subgroups deviate from these general trends. For example, women of East Asian origin, particularly Chinese women, have a comparable lung cancer incidence to women in Western Europe, despite significantly lower reported tobacco consumption. This elevated risk has been linked to non-smoking-related exposures, including air pollution and indoor combustion products. This highlights the role of environmental carcinogens in shaping ethnic disparities [2].

Genetic susceptibility also contributes to ethnic differences in the incidence of lung cancer. East Asian populations with non-small cell lung cancer, particularly never-smokers with ADC, exhibit a significantly higher prevalence of actionable driver mutations, such as EGFR alterations, than Western Caucasian populations. Additional population-specific genetic risk loci, including variants identified at 3q28 in Japanese and Han Chinese populations, as well as loci associated with telomere maintenance and nicotine dependence, reinforce the contribution of heritable risk factors that vary in prevalence across ethnic groups [2].

Beyond biological factors, structural and socioeconomic determinants play a critical role in shaping ethnic disparities in incidence. In the United States, tobacco use is strongly associated with racial and ethnic minority status, particularly when combined with low socioeconomic status. However, smoking patterns vary between different groups. For example, Black individuals diagnosed with lung cancer are more likely to be light or intermittent smokers, yet they experience higher levels of carcinogen exposure due to differences in smoking intensity and nicotine metabolism. This helps to explain why traditional exposure metrics, such as cumulative pack-years, fail to fully capture risk in certain ethnic groups [21].

Disparities are further exacerbated by unequal access to preventive services. Although millions of individuals meet the eligibility criteria for lung cancer screening based on their smoking history, participation rates remain extremely low, particularly among racial and ethnic minorities. These populations are disproportionately affected by barriers related

to healthcare access, insurance coverage, health literacy and systemic bias, which limit opportunities for early detection and potentially inflate the observed incidence of advanced disease [2, 21].

Importantly, however, evidence from screening trials suggests that these disparities are not immutable. Secondary analyses of the NLST indicate that Black participants may derive the same or a greater mortality benefit from LDCT screening than White participants, despite being severely under-represented in such studies. These finding highlights that the observed differences in incidence and outcomes are driven not only by intrinsic risk, but also by inequities in eligibility criteria and access to screening [21].

Taken together, these findings show that lung cancer incidence varies substantially across ethnic groups due to a combination of differential exposure, genetic susceptibility and systemic inequity. Aggregating heterogeneous populations into broad racial or ethnic categories further obscures high-risk subgroups, particularly within Asian American and Hispanic populations. This heterogeneity has important implications for epidemiological interpretation and the development of risk prediction models, reinforcing the need for disaggregated data and equity-aware approaches to lung cancer prevention and early detection.

2.3. Computed Tomography in Lung Cancer Screening

2.3.1 Computed Tomography (CT): General Principles

CT is an imaging modality that uses X-rays to generate detailed cross-sectional representations of internal anatomical structures. During acquisition, multiple X-ray projections are obtained from different angles around the patient and computationally reconstructed into tomographic slices. This process enables three-dimensional visualisation of internal tissues and eliminates the structural overlap present in conventional radiography, allowing more precise identification of pulmonary abnormalities such as nodules or masses [23–26].

A fundamental characteristic of CT imaging is its ability to quantitatively represent tissue density using the Hounsfield Unit (HU) scale (Figure 2.3). This standardised scale assigns distilled water a value of 0 HU and air −1000 HU, with other tissues positioned according to their relative radiodensity. For example, lung parenchyma typically appears at low HU values due to its air content, whereas soft tissues and bone present progressively higher

values. This numerical encoding provides a consistent physical basis for tissue characterisation and forms the foundation for both clinical interpretation and computational analysis of CT images [27].

Image interpretation is further influenced by reconstruction and visualisation parameters, particularly windowing settings, which control the range of HU values displayed. In lung imaging, specialised lung window settings enhance the visibility of pulmonary parenchyma and small nodules, whereas mediastinal or bone windows emphasise denser anatomical structures. These representation choices directly influence the visibility of abnormalities and illustrate how acquisition and reconstruction parameters shape the information available for diagnostic and computational analysis [28]. An example of a marked pulmonary nodule is shown in Figure 2.2.

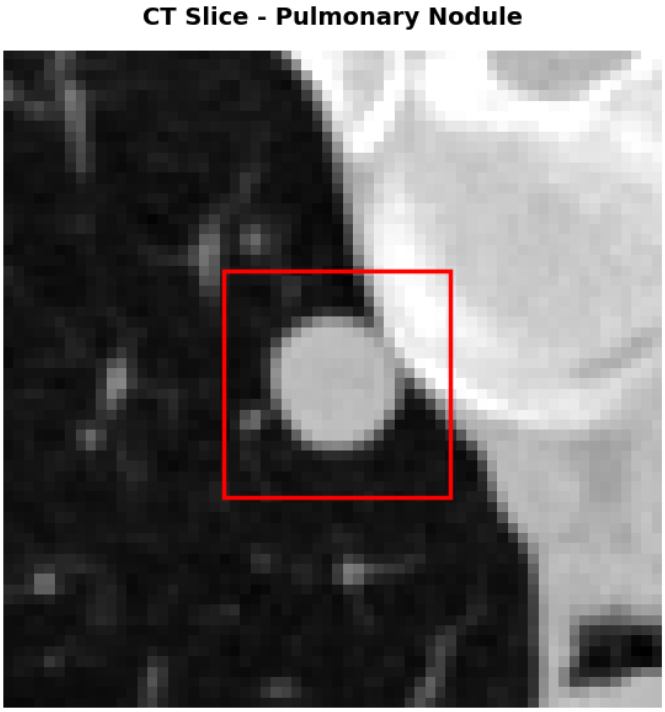


Figure 2.2: Region of Interest (ROI) from a CT slice displayed with lung window settings. The red bounding box highlights a pulmonary nodule, which appears as a denser (whiter) spherical region contrasting with the darker, air-filled lung parenchyma.

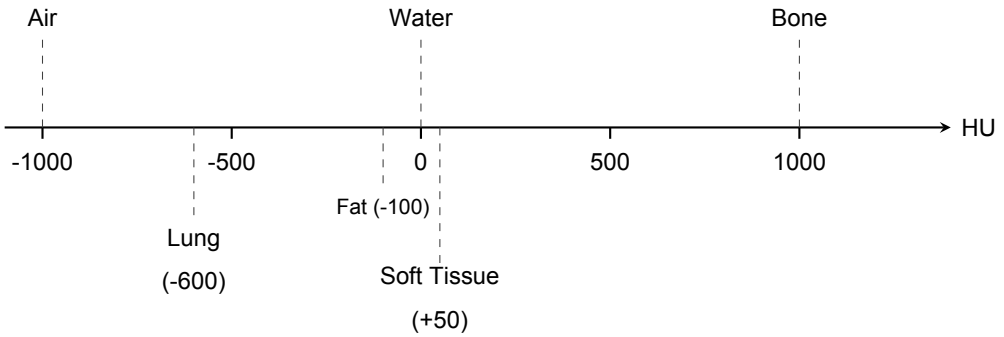


Figure 2.3: The HU scale with labeled tissue types positioned at the centers of common CT window levels. Each label corresponds to typical window centers used to highlight specific tissue densities, such as air, lung, fat, water, soft tissue, and bone. These centers serve as reference points around which different window widths are applied in clinical CT imaging to optimize the visualization of various anatomical structures and pathologies.

2.3.2 LDCT: Characteristics and Differences

LDCT is a CT acquisition protocol specifically designed to reduce radiation exposure while preserving sufficient image quality for lung cancer screening. Compared to standard diagnostic CT, LDCT uses substantially lower radiation doses, enabling repeated imaging of asymptomatic high-risk individuals while minimising cumulative radiation risk. This balance between safety and diagnostic utility has enabled the large-scale implementation of lung cancer screening programmes [29–31].

The reduction in radiation dose directly affects image characteristics. LDCT images typically exhibit increased noise and reduced signal-to-noise ratio compared to standard-dose CT, which can make the detection and characterisation of small pulmonary nodules more challenging. In addition, LDCT screening is generally performed without intravenous contrast, meaning that lesion identification relies entirely on natural differences in tissue density. Image quality may also be affected by artefacts such as respiratory motion or beam hardening, which introduce additional uncertainty in visual interpretation [3, 32].

These acquisition differences are particularly relevant for computational analysis. Variations in radiation dose, noise levels, and reconstruction protocols alter the statistical properties of CT images, introducing heterogeneity between datasets acquired under different conditions. Such variability can affect the robustness and generalization of automated analysis systems, especially when models are applied across different screening programmes or imaging protocols [3, 6].

Despite these limitations, LDCT remains the primary imaging modality for lung cancer

screening due to its demonstrated ability to detect pulmonary nodules at early, potentially curable stages. The clinical adoption of LDCT as a screening tool was supported by large randomized trials demonstrating significant reductions in lung cancer mortality in high-risk populations, supporting its adoption as a central component of early detection strategies [3, 29]. However, its widespread implementation also introduces operational challenges, including increased diagnostic workload and the need for accurate and efficient interpretation, reinforcing the importance of computational methods capable of supporting clinical decision-making in screening contexts [6].

2.3.3 Pulmonary Nodules in Screening Contexts

The primary objective of LDCT screening is to detect and characterise pulmonary nodules, which may be an early sign of lung cancer. The high spatial resolution of CT enables a detailed assessment of the morphological features associated with the risk of malignancy, including the spiculation of the margin and internal density patterns. Based on consistency, nodules are typically classified into three main categories: solid, part-solid, or non-solid (Figure 2.4) [24–26]. As well as analysing the nodule itself, CT provides contextual anatomical information, enabling evaluation of features such as pleural involvement or chest wall invasion. These features are relevant for staging and clinical decision-making.

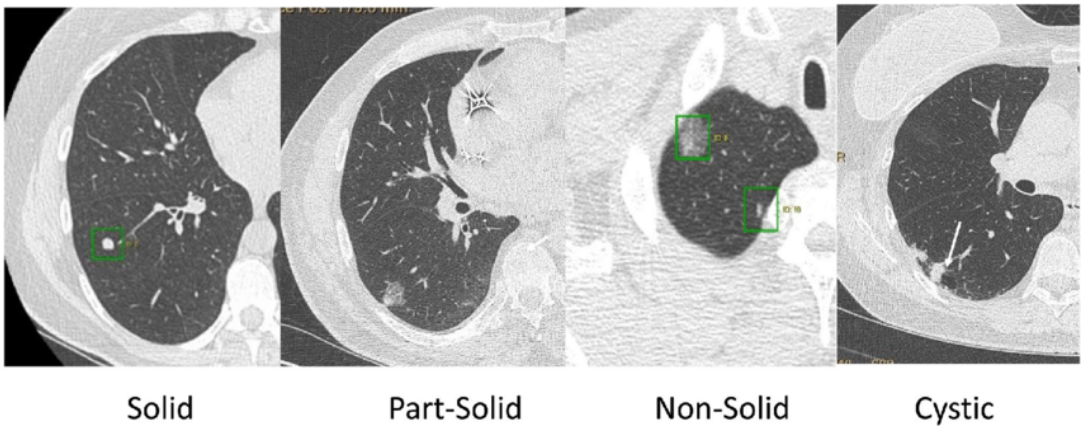


Figure 2.4: Morphological classification of pulmonary nodules in screening LDCT. From left to right: Solid nodule, which obscures the underlying lung architecture; Part-Solid nodule, containing both dense solid components and hazy ground-glass opacity; Non-Solid (or ground-glass) nodule, where vessels remain visible through the lesion density; and Cystic nodule. Non-solid nodules are often harder for automated systems to detect due to their low contrast. Image adapted from [33].

Despite these advantages, nodule interpretation is inherently complex and subject to variability. Studies have shown that agreement among radiologists is not absolute. *Gierada et al.*[34] reported an average concordance of approximately 82% among sixteen specialists

interpreting LDCT scans, indicating a meaningful level of disagreement in detection and classification [34]. Accurate measurement is also challenging. Since growth rate and volume doubling time are key indicators of malignancy, precise segmentation is critical. However, *Gietema et al.*[35] identified significant variability in around 11% of semi-automated measurements, particularly for irregular or spiculated nodules where boundaries are difficult to define [35]. These inconsistencies can influence clinical follow-up decisions and highlight the limits of purely visual assessment.

To address the complexity and inter-observer variability inherent to nodule interpretation, several quantitative risk prediction models have been developed to estimate the probability of malignancy in pulmonary nodules [36, 37]. These models aim to provide a more objective and standardised assessment by integrating both radiological and clinical variables.

Among the most widely adopted models is the Brock University Cancer Prediction Equation (also known as the PanCan model) [36], alongside the Mayo Clinic [37] and Veterans Affairs (VA) [38] models. Rather than relying exclusively on imaging appearance, these calculators combine nodule-specific characteristics — including size, upper-lobe location, spiculation, lobulation, and attenuation type (solid, part-solid, or non-solid) — with patient-level factors [36, 37].

Importantly, demographic variables such as age and sex are consistently incorporated as independent predictors of malignancy risk [36, 37]. Smoking exposure and, in some models, family or prior cancer history are also considered [38]

2.3.4 AI Systems in Screening Contexts

Given the complexity of pulmonary nodule assessment and the documented inter-observer variability, AI systems have been increasingly investigated as decision-support tools in lung cancer screening. DL models, in particular, have demonstrated the ability to automatically detect, segment, and characterise nodules directly from CT images, offering the potential for more consistent and scalable analysis in high-volume screening programmes [39–41].

However, the deployment of such systems introduces additional challenges. Model performance is highly dependent on the characteristics of the training data. While models often achieve high internal performance when evaluated on held-out subsets from the

same training distribution, differences in scanners, acquisition protocols, patient demographics, and institutional practices frequently lead to domain shifts. These shifts can severely degrade performance when models are applied in new clinical environments [6, 7]. Ensuring robust generalisation across heterogeneous populations and imaging settings—beyond baseline internal validation—therefore remains a central challenge in the development of reliable AI-based screening tools, which this study addresses through systematic cross-dataset evaluations.

The variability is further compounded by the nature of the ground truth used in clinical studies and computational research. Discrepancies between experts when defining nodule boundaries create inherently noisy annotations, setting a practical limit to diagnostic performance. Additionally, many studies rely on limited public datasets or institution-specific collections, which may not reflect the full diversity of real-world clinical environments, thereby restricting generalisation [3].

The heterogeneity of nodules themselves also introduces challenges. For example, solid nodules are far more prevalent than sub-solid or ground-glass lesions, creating an imbalance that shapes both clinical experience and data-driven learning. Consequently, rarer but clinically relevant subtypes may be detected less consistently [6]. Furthermore, detection performance is strongly influenced by intrinsic lesion characteristics, such as size and texture, rather than by patient-level demographic factors [6].

At the population level, the composition of the dataset introduces additional variability. Large screening cohorts often exhibit demographic imbalances, and variations in scanners, acquisition protocols and institutional practices further restrict representativeness, thereby reinforcing the need for external validation [6]. Bias can also arise from anatomical variations between patient groups and inconsistencies in human annotations. This can result in noisy datasets and uneven performance across subpopulations [7]. Importantly, defining fairness in medical diagnosis is not straightforward, as equal detection rates may be clinically inappropriate given legitimate differences in disease prevalence by age or sex [7].

From a machine learning perspective, DL models may implicitly encode demographic information from images and rely on it as a shortcut during prediction, potentially creating hidden performance disparities when deployed in new clinical settings [5]. They

may also learn spurious correlations from non-clinical image features, such as scanner-specific artefacts, rather than true pathological patterns [7]. Together, these factors highlight that pulmonary nodule assessment is shaped not only by morphological complexity and inter-observer variability, but also by dataset and acquisition heterogeneity, which directly impacts model generalization and equity.

2.4. Clinical Motivation for Robust and Equitable AI Systems

2.4.1 Clinical Implications of Limited Generalization

The clinical adoption of AI in lung cancer screening hinges on more than just accuracy in controlled conditions, it also requires reliable performance that can be maintained across institutions, populations, and time. When models fail to generalise, the consequences extend beyond technical degradation to directly affecting patient safety. One of the most significant risks in this context is automation bias, which is defined as the tendency to accept automated decisions despite contradictory visual evidence [42]. In clinical practice, this can cause radiologists, particularly those with less experience, to rely on AI outputs even when they are incorrect. Consequently, if a model fails to detect a lesion in an unfamiliar setting, this error may not be corrected by human oversight, but instead reinforced, thereby amplifying diagnostic errors rather than mitigating them.

Model failures are often linked to a distribution shift, where the statistical properties of real-world data differ from those encountered during the training process. For example, systems developed using data from specialty hospitals may perform poorly when deployed in community hospitals, where scanners, acquisition protocols and patient populations differ. From a clinical perspective, even modest reductions in sensitivity can result in cancers being missed, leading to a delayed diagnosis until later, less treatable stages [42]. Unlike random noise, statistical bias produces systematic and reproducible errors. When the internal representation learned by the model does not match the characteristics of the population being screened, missclassification can occur consistently within specific groups, resulting in repeated diagnostic inaccuracies rather than isolated errors.

Hardware variability is another major source of fragility. Models trained on data from a single scanner vendor or acquisition pipeline may malfunction when exposed to images from different systems, which limits their deployment in diverse healthcare environments [42]. Even minor changes, such as software upgrades or protocol adjustments, can alter image characteristics sufficiently to affect model outputs. This instability is particularly

problematic in lung cancer screening, where longitudinal comparisons of scans taken over multiple years are required. If model performance changes over time due to evolving data distributions, the assessment of growth becomes unreliable, which could result in unnecessary biopsies for stable nodules or the failure to identify clinically meaningful progression.

Additional risks arise from the way models are trained and evaluated. Sampling bias is common when datasets are collected from specialised centres where disease prevalence is higher and cases are more advanced. This creates a mismatch with population-level screening settings, where findings are more subtle [42]. Label ambiguity further complicates matters, as inconsistencies in human annotations introduce noise into the ground truth and limit achievable performance. In some cases, performance may be artificially inflated due to data leakage across combined public datasets, creating unrealistic expectations that cannot be realised in real clinical deployment [42].

Automation bias can manifest in different ways. Commission errors occur when clinicians accept false positive suggestions generated by AI, while omission errors arise when radiologists fail to detect abnormalities missed by the system, assuming that the absence of an alert implies normality [42]. These risks are particularly pronounced in high-volume environments or resource-constrained systems, where there is a greater reliance on automated support. Over time, such errors can create feedback loops: incorrect AI-assisted decisions may be incorporated into future training data, thereby reinforcing and perpetuating systematic inaccuracies (Figure 2.5).

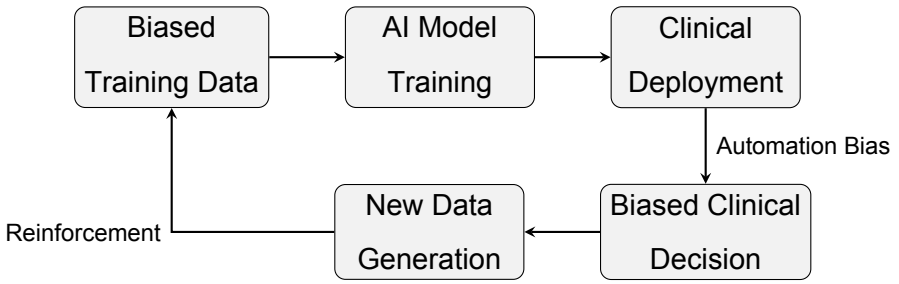


Figure 2.5: The feedback loop of algorithmic bias. If a biased model is deployed (e.g., one that under-diagnoses a specific demographic), and clinicians accept these incorrect suggestions due to automation bias, the resulting clinical decisions become the ”ground truth” for future datasets. This cycle permanently encodes and amplifies the original bias.

For these reasons, robustness cannot be considered a one-time validation criterion. Continuous monitoring after deployment is necessary to detect performance degradation caused

by changes in hardware, protocols or patient demographics [42]. In the context of population-scale screening, limited generalisation is therefore not merely a technical limitation, but also a patient safety issue. This is because systematic model failures can lead to delayed diagnoses and reduced trust in clinical decision support systems.

2.4.2 Fairness Considerations in Population-Level Screening

In addition to robustness, fairness is a fundamental requirement for safely integrating AI into screening programmes targeting large and diverse populations. If models perform unevenly across demographic or socioeconomic groups, they risk reinforcing existing disparities in access to early cancer detection. Training data are often derived from large urban academic centres, where minority populations, rural communities and safety-net patients may be underrepresented. When such models are deployed on a large scale, these gaps in representation can result in unequal diagnostic accuracy, leading to a situation where certain groups systematically receive lower-quality screening [42, 43].

The clinical implications of these disparities are significant. If these features were not adequately captured during training, differences in anatomical characteristics, tissue density, and disease presentation may reduce sensitivity in underrepresented populations. This can result in higher rates of missed or delayed diagnoses, which can have direct consequences for survival in screening contexts where early detection is critical [42]. At the same time, fairness is not limited to detection rates. If false positive rates are higher in certain subgroups, patients in these groups may be disproportionately exposed to follow-up imaging, radiation, anxiety, and invasive diagnostic procedures [44]. Therefore, ensuring equity also involves guaranteeing that the burden of unnecessary interventions is not unevenly distributed.

Bias in clinical AI often reflects the structural inequalities that are embedded within healthcare systems. For example, historical differences in referral patterns, access to imaging and follow-up care can become encoded in training data, causing models to treat these patterns as if they were biological signals [42]. Similarly, AI systems may learn latent demographic information directly from imaging data and use it as a shortcut for prediction, thereby introducing hidden disparities in performance across groups. Over time, such effects may be reinforced through feedback loops, in which biased predictions influence clinical decisions and future datasets.

Structural factors further shape equity. Systems optimised for high-quality scans from modern equipment may not perform as reliably in settings with limited resources, which limits their effectiveness in centres with older hardware and exacerbates existing disparities in care [43]. Therefore, addressing fairness requires more than simply removing sensitive attributes from the data. It is essential to ensure adequate representation of underserved populations and to evaluate performance across diverse subgroups, in order to prevent AI from unintentionally amplifying existing disparities in population-level screening.

3. Literature Review

This chapter provides a comprehensive review of the existing literature fundamental to the problem under study. The transition from traditional radiomics to DL has revolutionized lung nodule classification; however, as these models move closer to real-world clinical deployment, new challenges have emerged. To understand this landscape, this review is structured around three critical pillars: predictive performance, cross-dataset generalization, and algorithmic fairness.

First, Section 3.1 explores the evolution of automated lung nodule classification, detailing the primary public datasets and comparing state-of-the-art DL architectures, from CNNs to Vision Transformers (ViTs). Section 3.2 then delves into the challenge of model generalization in medical imaging, explaining the mechanics of domain shift and how variations in physical acquisition parameters degrade model robustness. Building on this, Section 3.3 introduces the concept of fairness in artificial intelligence, identifying how demographic imbalances and shortcut learning introduce systemic biases into medical algorithms. Section 3.4 connects these domains, analyzing empirical evidence of how generalization failures and fairness disparities manifest specifically within lung cancer screening cohorts. Finally, Section 3.5 provides a critical synthesis of the literature, highlighting the current limitations and open methodological gaps that this dissertation aims to address.

3.1. Lung Nodule Classification

Lung nodule classification aims to distinguish benign from malignant nodules detected in chest CT examinations. Over the past decade, deep learning has become the dominant paradigm for this task, progressively replacing traditional radiomics and hand-crafted feature approaches. Recent systematic reviews consistently report substantial performance improvements driven by convolutional and transformer-based architectures [3, 4, 45, 46].

3.1.1 Public and Private Datasets

The selection of datasets is fundamental to the training and evaluation of lung nodule classification models. Historically, the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) database [47] has been the most prominent benchmark. It contains 1,018 diagnostic thoracic CT scans annotated by four experienced radiologists. However, a significant limitation of LIDC-IDRI is that it relies solely on

radiologist-noted malignancy suspicions as a diagnostic proxy, lacking histological confirmation from biopsies for the vast majority of cases. LUNA16 [48], a subset of LIDC-IDRI excluding scans with a slice thickness exceeding 2.5 mm, provides 888 cases specifically tailored for nodule detection and false-positive reduction.

For large-scale screening and longitudinal analyses, the NLST [24] is a critical resource, containing data from over 26,000 participants. However, a major limitation of the raw NLST data is the lack of localized nodule coordinates, providing only scan-level cancer annotations. To address this limitation, this dissertation utilizes LUNA25 [49], a recent dataset derived from the NLST that provides over 6,000 annotated nodules across approximately 4,000 scans, complete with 3D center coordinates and malignancy labels. Another complementary resource is the Lung Nodule Database (LNDb)[50], which offers 294 annotated cases from a European cohort. However, similarly to LIDC-IDRI, it shares the limitation of relying on clinician annotations without systematic histological confirmation.

In addition to LUNA25, this study utilizes the LUNGx and DLCS datasets for model training and validation. LUNGx [51, 52] provides 73 nodules with histologically confirmed benign or malignant statuses, offering highly reliable ground truth. The Duke Lung Cancer Screening dataset [53] further supports this study, offering 1,613 patients and 2,487 nodules marked with 3D bounding boxes and detailed diagnostic outcomes based on the Lung-RADS lexicon.

Despite the growth of these public repositories, a substantial proportion of studies still rely on private institutional datasets. Training models on inaccessible private data prevents the scientific community from reproducing results, conducting direct cross-study comparisons, and performing external validation. This lack of transparency restricts the generalization of models to real-world clinical environments. The characteristics of the primary datasets discussed are summarized in Table 3.1.

3.1.2 Model Architectures

3.1.2.1 Standard Convolutional Networks

CNNs remain the foundational architecture for lung nodule classification. Their ability to automatically learn hierarchical spatial representations from grid-structured image data makes them particularly well-suited to CT-based malignancy prediction. Standard CNN

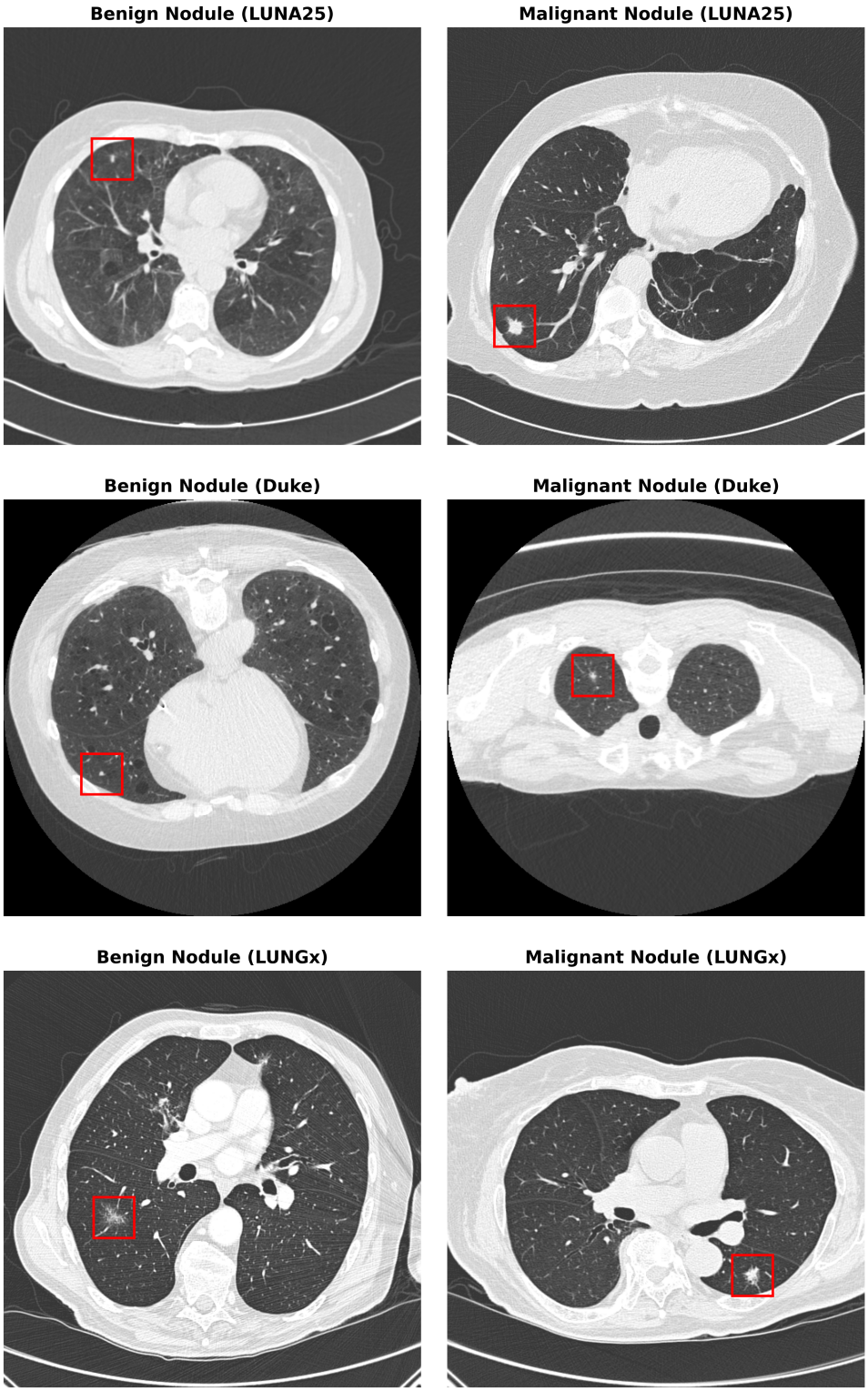


Figure 3.1: Representative full axial CT slices from the three primary datasets utilized in this study (Top: LUNA25, Middle: DLCS/Duke, Bottom: LUNGx). Red bounding boxes indicate the location of annotated benign (left column) and malignant (right column) pulmonary nodules, illustrating the challenge of identifying small lesions within the broader pulmonary context.

Table 3.1: Summary of key public datasets for lung nodule classification. The entries in the 'Download Portal' column are interactive; click on the portal name to access the respective dataset repository.

Dataset	Patients / Scans	Access	Annotations	Download Portal
LIDC-IDRI [47]	1,010 / 1,018	Public	4 Radiologists (No systematic biopsy)	TCIA LIDC-IDRI
LUNA16 [48]	888 scans	Public	Radiologist consensus (subset)	Grand Challenge
NLST [24]	∼26,722 / ∼75,000	Restricted	Scan-level only (No coordinates)	CDAS Portal
LUNA25 [49]	∼4,000 scans	Public	>6,000 nodules (3D Coordinates)	Grand Challenge
LUNGx [54]	60 / 73	Public	73 nodules (Histology confirmed)	TCIA LUNGx
DLCS [53]	1,613 patients	Public	2,487 nodules (3D Bounding boxes)	Zenodo (Duke DLCS)
LNDb [50]	294 cases	Public	Coordinates & Fleischner labels	Grand Challenge

pipelines have a hierarchical structure consisting of stacked convolutional layers, nonlinear activation functions such as ReLU, and pooling operations for spatial downsampling. They also have fully connected classification heads with Softmax output layers, which produce final class probabilities.

Advanced CNN variants such as VGG-16, ResNet, DenseNet and ResNeXt have been widely adopted [4, 45, 46]. ResNet's residual connections mitigate vanishing gradient problems and enable deeper networks without performance degradation, while DenseNet promotes feature reuse through dense connectivity patterns.

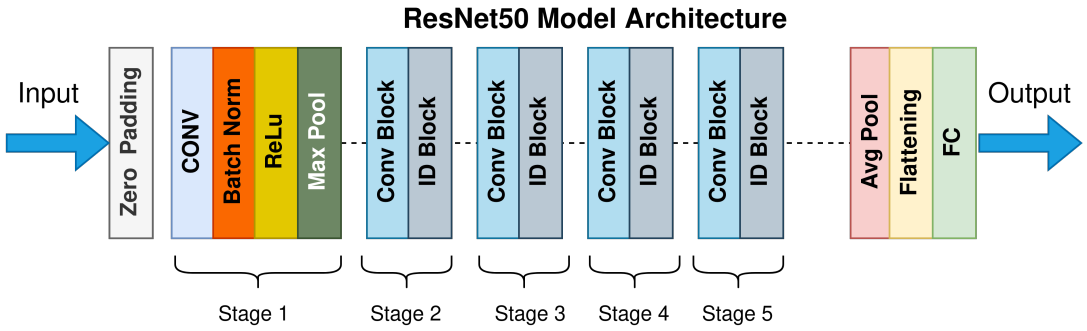


Figure 3.2: Macro-architecture of a standard ResNet-50 model, illustrating the progression through five distinct stages, including convolutional blocks, identity blocks, and pooling layers. Adapted from the original framework by He *et al.* [55].

To illustrate the practical application of these architectures, several studies have proposed concrete methodologies with notable results. For instance, *Tan et al.*[56] adapted a 2D

VGG-16 architecture utilizing transfer learning to distinguish malignant from benign nodules using a combined dataset from the NLST and the National Institute of Allergy and Infectious Diseases (NIAID) portal. By leveraging its 13 convolutional and 4 pooling layers for efficient feature extraction, their model achieved a classification accuracy of 90.40% and an F1-score of 90.10% [56]. Exploring architectural efficiency, *Zhang et al.*[57] implemented a DenseNet-121 model. By utilizing identity connections that grant each layer direct access to preceding feature maps, they enhanced information flow without increasing model complexity, achieving an accuracy of 87.67% and an AUC of 93.79% on the LIDC-IDRI dataset [57].

3.1.2.2 The Dimensional Dilemma: 2D vs. 3D Approaches

A central methodological debate concerns the so-called "dimensional dilemma" between 2D and 3D models [4]. 2D slice-based models analyze individual CT slices. These models are computationally efficient and benefit from transfer learning from natural image datasets such as ImageNet [58]. However, they risk losing volumetric context, which is clinically relevant for malignancy assessment. In contrast, 3D CNNs process full nodule volumes or CT patches, capturing richer spatial information. Multi-model 3D ensembles and volumetric CNNs have demonstrated improved discriminative capacity but require substantially higher computational resources and larger annotated datasets. Hybrid approaches attempt to balance this trade-off by incorporating Recurrent Neural Networks (RNNs). For example, *Moitra and Mandal*[59] proposed a deep ensemble combining 2D CNNs with bidirectional LSTMs to model inter-slice spatial coherence without the full computational burden of pure 3D networks, achieving 96% accuracy and an AUC of 99% on the NSCLC Radiogenomics dataset [59]. Capsule networks and multiview architectures further aim to preserve hierarchical spatial relationships across planes [4]. A summary of the advantages and limitations of these dimensional approaches is presented in Table 3.2.

Given the computational advantages of slice-based approaches, models tailored for specific input resolutions are particularly relevant for clinical deployment. Focusing specifically on 64×64 pixel 2D nodule patches, *Ali et al.*[60] proposed an efficient transferable texture 2D CNN that reached an accuracy of 97% and an AUC of 99% on the LIDC-IDRI benchmark [60]. Similarly, operating on the exact same 64×64 input dimensions, *Halder et al.*[61] utilized an adaptive morphology-aided two-pathway 2D CNN. By integrating distinct feature extraction pathways, their framework yielded 96% accuracy and 99% AUC

Table 3.2: Comparison of 2D and 3D CNN Architectures for Lung Nodule Classification.

Architecture Type	Key Advantages	Main Limitations
2D Slice-based	Computationally efficient; easily leverages transfer learning from natural image datasets (e.g., ImageNet).	Loses critical volumetric context; ignores inter-slice spatial relationships.
3D Volumetric	Captures rich, real-world spatial information; generally provides superior discriminative power.	High computational cost; heavily data-dependent; prone to overfitting on small datasets.
Hybrid (2D + RNN)	Balances spatial coherence with lower computational demands than pure 3D networks.	Increased architectural complexity; training dynamics can be unstable.

[61]. These specific implementations demonstrate that while deep 3D networks offer rich spatial context, highly optimized 2D architectures employing transfer learning, feature reuse, and multi-path processing remain highly competitive and computationally efficient.

In addition to standard architectures, the literature suggests several specialized CNN-based models tailored specifically for lung nodule classification. These include multi-crop CNNs, which are designed to extract local, nodule-specific features by cropping distinct regions directly from intermediate convolutional feature maps; and multiview residual networks, which analyze the nodule's broader context by processing the axial, coronal, and sagittal planes simultaneously. Furthermore, researchers have successfully implemented Squeeze-and-Excitation (SE)-enhanced networks that incorporate channel-wise attention mechanisms to improve feature representation. Hybrid methodologies have also proven highly effective. For example, *Da Nóbrega et al.* [62] combined deep learning with traditional Machine Learning (ML) by utilizing a ResNet50 architecture to extract deep features, which were subsequently classified using an RBF-SVM classifier. This hybrid approach achieved an accuracy of 88.41% and an AUC of 93.19% on the LIDC-IDRI dataset [62]. Overall, reported performance for these specialized architectures on benchmark datasets such as LIDC-IDRI and LUNA16 often achieves AUC values above 0.90, with several models exceeding 95% accuracy [45, 46].

To address the inherent morphological and dimensional variability of pulmonary nodules, the literature frequently identifies multi-view and multi-scale architectures as specialized, highly effective subgroups. Multi-view networks attempt to capture 3D spatial context without the heavy computational burden of full volumetric convolutions by processing 2D slices simultaneously across the axial, coronal, and sagittal planes. For instance, *Chang*

et al.[63] proposed a Multiview Residual Selective Kernel Network (MRSKNet) that integrated multi-plane data, achieving a classification accuracy of 93.6% and an AUC of 97.1% on the LIDC-IDRI dataset [63]. Conversely, multi-scale models tackle the dramatic variance in nodule size by extracting features at different receptive fields. Demonstrating the efficacy of this approach, *Zhao et al.*[64] developed a multi-scale, multi-task 3D CNN that combined volumetric features of different sizes, achieving 94% accuracy, 93% sensitivity, and an AUC of 98% [64].

3.1.2.3 Vision Transformers and Hybrid Models

More recently, ViTs have been incorporated into frameworks for classifying lung cancer. Unlike CNNs, ViTs use self-attention mechanisms to model long-range dependencies between different parts of an image. Their architecture usually comprises patch embedding, positional encoding, multi-head self-attention blocks and an MLP classification head.

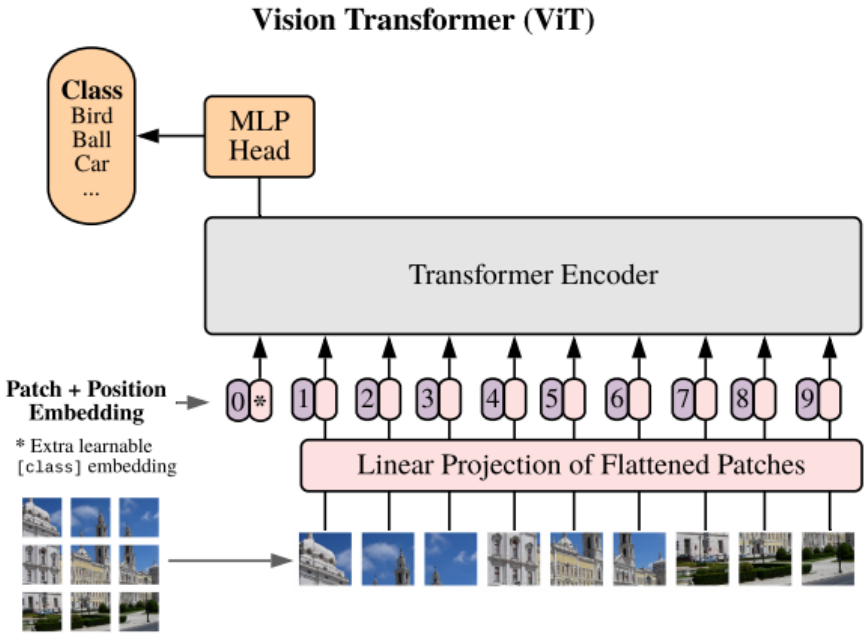


Figure 3.3: Standard architecture of a Vision Transformer (ViT). The input image is split into fixed-size patches, linearly embedded, and combined with position embeddings before being processed by the Transformer Encoder. Adapted from Dosovitskiy *et al.* [65].

While Transformers offer improved global contextual modelling, they require substantial amounts of data and computing power. Reviews consistently note that ViTs require substantially larger annotated datasets and are less data-efficient than CNNs in scenarios

with limited data [45]. Furthermore, they lack inherent translational invariance, rendering positional encoding essential. Hybrid CNN-Transformer models, such as TransUNet and other combined architectures, offer a compromise between local feature extraction (CNNs) and global contextual modelling ViTs. These hybrid systems have achieved high classification performance in controlled settings, sometimes exceeding 98% accuracy [45, 46]. To illustrate this, recent implementations have successfully adapted transformer architectures for nodule malignancy prediction. *Khademi et al.*[66] proposed the CAET-SWin framework, which utilizes parallel spatial and temporal self-attention mechanisms to simultaneously extract inter-slice and intra-slice features from unthinned CT scans [66]. Furthermore, *Mkindu et al.*[67] introduced the 3D Multi-Scale Vision Transformer (3D-MSViT), completely discarding traditional CNNs in favor of local and global transformer blocks. This architecture effectively captured multi-scale features and achieved a remarkable sensitivity of 97.81% [67]. These examples confirm that while ViTs are computationally demanding, their ability to model long-range spatial and temporal dependencies offers a powerful alternative to standard convolutional methods.

3.1.3 Performance and Evaluation Metrics

Model performance is usually evaluated using accuracy, sensitivity (recall), specificity, precision, the F1 score, ROC curves and Area Under The Curve (AUC). Recent literature reports extremely high performance metrics. Binary classification accuracy values often exceed 90%, with some studies reporting values as high as 98–99% when evaluated internally on public benchmarks such as the LIDC-IDRI or LUNA16 datasets [45, 46]. AUC values above 0.95 are common in these top-performing models. However, multiple reviews explicitly warn against performance optimism. Reported metrics are often influenced by dataset homogeneity, a lack of multicentre validation, class rebalancing strategies that distort clinical prevalence, and evaluation on internal test splits rather than external cohorts.

Models trained exclusively on public benchmarks like the LIDC-IDRI, for example, have demonstrated substantial performance drops when evaluated on independent institutional datasets. Accuracy can decrease from approximately 98% to 76% under different scanner protocols and acquisition settings [45]. This generalisation drop highlights the discrepancy between experimental performance and real-world robustness.

3.1.4 Clinical and Technical Limitations

Despite substantial reported improvements, important limitations persist. Limited generalizability remains a central challenge. A small fraction of studies validate their models on external or strictly independent datasets [45]. Furthermore, even when utilizing multi-institutional databases like LIDC-IDRI or NLST, researchers frequently fail to stratify their performance analysis across different scanners or acquisition protocols, overlooking a critical dimension of model robustness. Moreover, public datasets predominantly represent Western populations, raising concerns about demographic bias and restricted applicability to underrepresented groups [45, 68].

Annotation variability further constrains achievable performance. As previously highlighted in Section 3.1.1, the reliance on subjective radiologist consensus introduces inherent label noise. Because deep learning models optimize toward these subjective annotations, their reliability is intrinsically tied to the quality and consistency of this ground truth.

The widespread use of private datasets and the absence of publicly available source code contribute to a broader reproducibility crisis [4]. External replication is frequently impossible, undermining scientific transparency. Additionally, the literature remains heavily concentrated on binary classification, while clinically relevant multi-class staging and subtype prediction tasks remain comparatively underexplored [3]. Although explainability tools are increasingly proposed, systematic integration into clinical workflows remains limited, and the "black box" nature of complex models continues to hinder adoption. Finally, a critical methodological gap persists in data curation and evaluation: studies frequently fail to disclose their precise patient selection criteria when utilizing subsets of larger databases, and the reporting of formal fairness metrics across distinct demographic groups remains conspicuously absent from standard evaluation pipelines.

3.2. Generalization in Medical Imaging

Deep learning models have demonstrated remarkable performance in controlled experimental settings. However, their deployment in real clinical environments often reveals substantial performance degradation. A primary cause of this limitation is the presence of *domain shift*, where the statistical distribution of the data encountered during deployment differs from the one used during training. In medical imaging, such shifts are particularly

common due to variations in imaging devices, acquisition protocols, and patient populations across healthcare institutions.

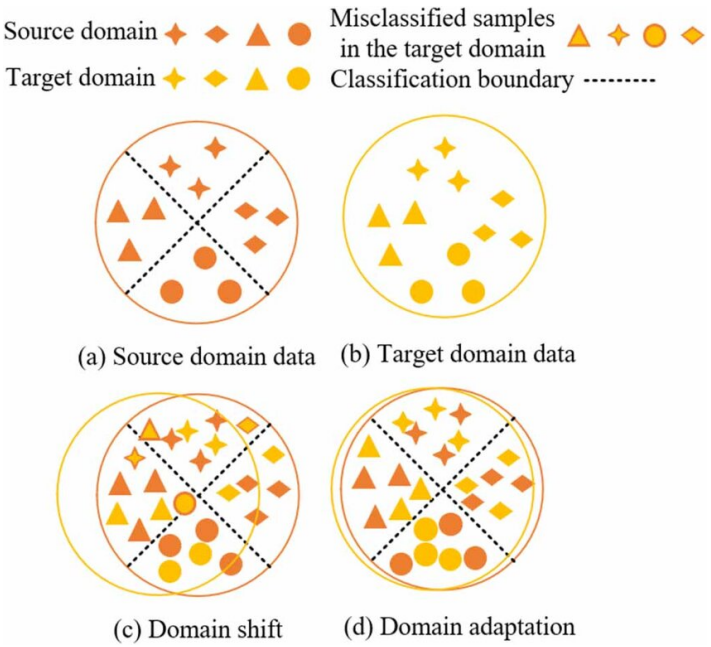


Figure 3.4: Illustration of domain shift and adaptation. (a) A classifier is trained on source domain data. (b) The target domain presents a different data distribution. (c) Domain shift causes the original decision boundary to misclassify samples in the target domain. (d) Domain adaptation adjusts the boundary to perform accurately across distributions.

3.2.1 Domain Shift

Formally, a domain can be defined by a joint probability distribution P_{XY} over the feature space $\mathcal{X}$ and label space $\mathcal{Y}$ [69]. In the context of medical image analysis, $\mathcal{X}$ typically corresponds to the input image (e.g., a 2D cropped patch of a lung nodule extracted from a CT scan), while $\mathcal{Y}$ represents the associated clinical label (e.g., benign or malignant).

A *domain shift* occurs when the joint distributions of the training and test data differ:

$$P_{train}(X, Y) \neq P_{test}(X, Y) \tag{3.1}$$

Even when the predictive task remains identical, this discrepancy can significantly impact model performance. As illustrated in Figure 3.4, a classifier trained on the source domain data (**a**) may fail when deployed on target domain data (**b**), leading to misclassified samples (**c**) due to an inadequate decision boundary.

Two common forms of dataset shift are frequently discussed in the literature. The first is *covariate shift*, which occurs when the marginal distribution of the input features changes

between training and testing scenarios while the conditional relationship between the input and the label remains constant [70]. This situation can be expressed as:

$$P_{train}(X) \neq P_{test}(X), \quad P_{train}(Y|X) = P_{test}(Y|X) \tag{3.2}$$

In medical imaging, covariate shift commonly arises when models are deployed on images acquired using different scanners or acquisition protocols. Although the visual characteristics of the images may change (e.g., contrast, resolution, or noise), the underlying biological relationship between image features and disease remains unchanged.

Another common form is *label shift*, also known as prior probability shift. In this case, the class distribution differs between training and testing environments:

$$P_{train}(Y) \neq P_{test}(Y), \quad P_{train}(X|Y) = P_{test}(X|Y) \tag{3.3}$$

In clinical practice, this often corresponds to variations in disease prevalence across populations. For example, a model trained in a specialized oncology center may encounter a significantly lower prevalence of malignant cases when deployed in a population-level screening program.

3.2.2 Sources of Domain Shift in Medical Imaging

In medical image analysis, domain shift is particularly prevalent due to the inherent heterogeneity of clinical data. Variability may arise from multiple sources, including differences in scanner manufacturers, acquisition protocols, reconstruction parameters, and patient populations across institutions [71].

Variations in imaging protocols, such as slice thickness, reconstruction kernels, or spatial resolution, can substantially alter image characteristics. Similarly, differences in exposure levels may produce distinct intensity distributions and noise patterns, particularly when comparing standard-dose and low-dose CT acquisitions. Hardware limitations and patient motion may also introduce artifacts that further modify the visual appearance of images. In addition, demographic differences between healthcare centers can lead to variations in anatomical characteristics and disease prevalence.

These factors make it difficult for deep learning models to identify domain-invariant anatomical patterns, often causing performance degradation when models are deployed outside the distribution of the training data.

3.2.3 Domain Adaptation and Domain Generalization

To address the challenges posed by domain shift, the literature typically distinguishes between two main paradigms: *Domain Adaptation* (DA) and *Domain Generalization* (DG).

Domain Adaptation (DA) assumes that the feature space and prediction task remain identical across domains, but the underlying distributions differ. Formally, let S and T denote the source and target domains with joint distributions P_S and P_T, respectively [72]. Given labeled samples from the source domain

$$D_S = \{(x_i^S, y_i^S)\}_{i=1}^{n_S} \tag{3.4}$$

and unlabeled or partially labeled samples from the target domain

$$D_T = \{x_j^T\}_{j=1}^{n_T}, \tag{3.5}$$

the goal of domain adaptation is to leverage the information learned from S to improve predictive performance in T, effectively adjusting the decision boundary as depicted in Figure 3.4(d).

Depending on the availability of labeled data in the target domain, DA can be categorized into supervised, semi-supervised, or unsupervised settings [72]. In medical imaging applications, domain adaptation is often used to transfer models across hospitals or imaging devices while utilizing a limited amount of data from the target institution.

In contrast, Domain Generalization (DG) addresses a more challenging scenario in which no information from the target domain is available during training. Instead, the model is trained using one or multiple source domains and is expected to generalize to completely unseen environments [69, 73].

Formally, consider K source domains:

$$D_S = \{D^k = \{(x_i^k, y_i^k)\}_{i=1}^{n_k}\}_{k=1}^{K} \tag{3.6}$$

where each domain is associated with its own joint distribution P_{XY}^k. Because these distributions differ across domains,

$$P_{XY}^{k} \neq P_{XY}^{k'}, \quad k \neq k' \tag{3.7}$$

the objective is to learn a predictive function

$$f : \mathcal{X} \rightarrow \mathcal{Y} \tag{3.8}$$

that minimizes the prediction error on an unseen target domain T with distribution P_{XY}^{T} [73]. Compared with domain adaptation, DG is considerably more challenging because the model must learn representations that remain valid under previously unseen distribution shifts.

Depending on the number of source domains available during training, DG can be further categorized into two main settings, which significantly dictate the applicable methods [69]:

- **Multi-Source Domain Generalization (MSDG):** This setting utilizes training data from multiple distinct source domains ($K > 1$). By observing real distributional variations across sources, the model can explicitly learn domain-invariant features. Consequently, classical methods applied in this scenario typically involve domain alignment and adversarial training [71, 73].
- **Single-Source Domain Generalization (SSDG):** In this more restrictive setting, the model is trained on only one source domain ($K = 1$). This is a highly realistic scenario in clinical applications, where collecting data from multiple institutions is often infeasible due to privacy regulations. Because the model lacks direct exposure to domain diversity, SSDG methods typically rely on advanced data augmentation (e.g., frequency-based mixing or domain randomization) to artificially synthesize unseen target distributions [74].

3.2.4 Methods to Promote Generalization

To mitigate the impact of domain shift, several methodological strategies have been proposed to improve the robustness and generalization capabilities of deep learning models.

A common approach involves *data-level strategies*, which aim to increase the diversity of the training data. Techniques such as data augmentation, intensity normalization, and histogram matching modify the input images to reduce site-specific biases and simulate

variations that may occur across different clinical environments [71]. More advanced approaches employ generative models or adversarial techniques to synthesize realistic images that mimic variations in acquisition protocols or scanner characteristics.

Another category of approaches focuses on *feature-level alignment*. Instead of modifying the raw images, these methods attempt to learn latent representations whose statistical distributions remain consistent across domains. By minimizing distribution discrepancies between features extracted from different datasets, models are encouraged to focus on domain-invariant anatomical structures rather than scanner-specific artifacts [71].

Adversarial learning represents a related strategy, often implemented through Domain-Adversarial Neural Networks (DANN). In this framework, an auxiliary discriminator attempts to identify the domain from which an image originates, while the feature extractor is trained to prevent the discriminator from succeeding. This adversarial objective encourages the network to remove domain-specific information from its learned representations [71, 73].

Another promising direction involves *self-supervised learning*, where models are pre-trained using large collections of unlabeled medical images through auxiliary tasks such as image reconstruction or rotation prediction. By learning general anatomical representations before fine-tuning on labeled datasets, self-supervised approaches can improve robustness to domain shifts [71].

Finally, *test-time adaptation* methods attempt to adjust model parameters dynamically during inference using small batches of unlabeled test data. This strategy allows models to adapt to new environments without requiring extensive retraining or labeled target-domain datasets [71].

3.3. Fairness in Medical Imaging

As discussed in previous sections, domain shifts can significantly reduce the effectiveness of ML models in medical imaging. When these shifts are correlated with patient demographics, such as sex, race, age or socioeconomic status, the consequences of model failure extend beyond technical limitations to raise ethical and clinical concerns. These disparities are commonly referred to as algorithmic bias. Therefore, ensuring fairness in medical imaging systems requires identifying where biases originate throughout the ML pipeline and understanding how they can influence clinical decision-making.

3.3.1 Sources and Types of Bias

ML models are not inherently objective. Because they learn from data generated and curated by humans, they can reproduce and even amplify historical inequalities present in the data [75]. In the context of medical image analysis, for example, biases can emerge at multiple stages of the artificial intelligence development pipeline. *Drukker et al.* [76] identify numerous potential sources of bias across the full AI roadmap, including data collection, data preparation and annotation, model development, evaluation, and deployment [76]. These biases can broadly be grouped into four categories: data-driven, human and annotation, algorithmic, and evaluation and deployment, as illustrated in Figure 3.5

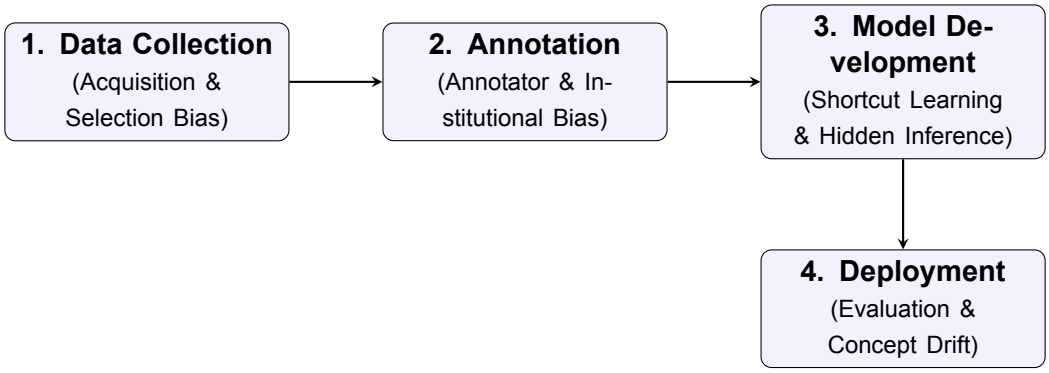


Figure 3.5: The Artificial Intelligence development pipeline and potential sources of bias injection [76].

Data-driven biases usually arise when the training dataset does not accurately reflect the target clinical population. A common example is data acquisition bias, which occurs when datasets are collected from a limited number of institutions, imaging devices or acquisition protocols. Models trained on such restricted data may fail to generalize to images produced by different scanner manufacturers or clinical environments [76]. Selection and sampling bias can also arise when datasets are collected without appropriate randomization, leading to distributions that do not reflect real-world clinical populations [75]. In medical imaging, this can cause models to learn spurious correlations or shortcuts rather than genuine pathological features. Reporting bias represents another issue, as datasets often contain disproportionate numbers of severe or unusual cases while common or early-stage manifestations of disease remain underrepresented [75]. Additionally, historical bias may be embedded in the data if previous medical practices systematically underdiagnosed or undertreated specific patient populations, thereby transferring those inequities into the model [75]. *Drukker et al.*[76] also highlight activity bias, where wealthier regions are more capable of producing large-scale datasets due to greater access to imaging infrastructure and data storage systems, leading to models trained primarily on

patients from high-resource healthcare systems [76]. Temporal bias can further emerge as clinical practices, disease characteristics, and imaging technologies evolve over time, causing models trained on historical data to perform poorly on contemporary patient populations [76].

Human and annotation-related biases constitute another major category. In supervised learning pipelines, human annotators define the ground truth labels used to train the model, meaning subjective judgements inevitably influence the resulting dataset. Annotator bias may occur due to differences in expertise, training background, or clinical interpretation among radiologists [76]. These inconsistencies can introduce systematic labelling errors that propagate into the learned model. Content production bias may arise from differences in how clinicians document findings or describe imaging features, including variations in terminology or reporting standards across institutions [76]. Presentation bias can also occur when the design of annotation interfaces influences how clinicians interpret images or identify suspicious regions [76]. Furthermore, implicit and confirmation biases may influence how annotators interpret imaging findings, particularly if they hold preconceived beliefs about how certain diseases manifest in specific patient populations [75]. Institutional bias may also arise when hospital practices or policies systematically affect how different demographic groups are diagnosed or treated. For instance, if hospital management decisions such as ICU admission are used as ground truth labels, any existing inequities in resource allocation may be encoded directly into the model [76].

Even when the data collection and annotation processes are carefully controlled, algorithmic and model development biases can arise. Clinical AI systems often involve multistage pipelines including preprocessing, segmentation, and classification, meaning that errors or biases introduced in early stages can propagate through the entire system [76]. In transfer learning settings, biases present in the large public datasets used for pretraining may also be inherited by the final medical model. Another challenge is the lack of universally reliable ground truth in many medical tasks. Diagnostic reference standards may rely on different clinical tests or procedures depending on availability, cost, or institutional policy, resulting in inconsistent training targets [76]. DL models may also develop shortcut learning behaviors, where easily identifiable patterns such as demographic characteristics are used as proxies for the clinical outcome instead of learning the intended pathology-related features [76]. Furthermore, recent studies have demonstrated that neural networks can infer sensitive attributes, such as race, directly from medical images, even when these attributes are removed from metadata. This renders simple

demographic 'blinding' strategies ineffective.

Bias can also be introduced or amplified during model evaluation and clinical deployment. Evaluation bias may occur when models are assessed solely using aggregate performance metrics across an entire dataset without stratifying results by demographic subgroups [76]. Such evaluations can mask poor performance in minority populations despite high overall accuracy. Deployment bias arises when the intended use of an algorithm differs from how it is actually used in clinical practice. For example, an AI tool designed for quantitative measurement may be incorrectly used as a diagnostic decision aid, leading to unreliable predictions [76]. Finally, concept drift can cause initially unbiased models to become unfair over time as patient populations, imaging technologies or disease patterns evolve.

3.3.2 Fairness Metrics

Evaluating fairness in ML systems requires more than just reporting aggregate metrics such as accuracy, precision and recall. High overall performance may conceal poor performance on minority subgroups, resulting in unequal clinical outcomes [75]. Fairness-aware metrics quantify disparities between demographic groups to address this issue.

One commonly used metric is demographic parity (DP), which evaluates whether the proportion of positive predictions is equal across different demographic groups. Under DP, individuals from different groups should have the same probability of receiving a positive prediction regardless of group membership [75]. While this metric ensures balanced representation among predicted positive cases, it does not account for differences in the underlying distribution of true outcomes between groups. Consequently, DP can lead to misleading conclusions when disease prevalence varies across populations.

Equality of opportunity (EO) offers an alternative definition of fairness by focusing specifically on the model's true positive rate. This metric requires that individuals who truly belong to the positive class, for example, patients who actually have a disease, have the same probability of being correctly identified regardless of their demographic group [75]. EO is particularly relevant in medical diagnosis, where failing to detect disease in certain populations can have serious clinical consequences. However, this metric assumes the existence of a clearly defined preferred label and focuses only on positive outcomes. In situations where both positive and negative predictions are equally important, the related concept of equalized odds may be more appropriate.

Another definition of fairness is counterfactual fairness. This concept evaluates whether a model would produce the same prediction for two identical individuals if only a sensitive attribute, such as race or gender, were hypothetically changed [75]. Counterfactual fairness is particularly useful when demographic information is incomplete or unavailable for the entire dataset. By analysing pairs of examples that differ only in the sensitive attribute, this approach can identify prediction inconsistencies at an individual level. However, counterfactual analysis typically examines only small subsets of the data and therefore may fail to capture broader systemic disparities across entire populations [75].

In practice, researchers often combine multiple fairness metrics to gain a more comprehensive understanding of potential bias within a model.

3.3.3 Methods to Mitigate Bias

Once bias has been identified, mitigation strategies can be applied at various stages of the ML process. These strategies are commonly categorized as pre-processing, in-processing, and post-processing interventions [75].

Pre-processing methods aim to address bias directly within the training data. Ideally, this would involve collecting additional data to ensure that the dataset accurately represents all relevant demographic groups. However, acquiring new annotated data is often difficult in medical imaging due to the high cost of imaging procedures, the time required for expert annotation, and strict patient privacy regulations [75]. Consequently, researchers often rely on computational techniques such as oversampling underrepresented groups or using generative models to synthesise additional training examples.

In-processing methods modify the model training procedure itself. Standard loss functions, such as cross-entropy or log loss, optimize predictive accuracy without considering demographic disparities [75]. Fairness-aware training approaches introduce additional constraints or penalties to reduce performance gaps between groups. For example, distributional penalty methods such as MinDiff penalize the model when prediction distributions differ significantly across demographic subgroups [75]. Another approach, counterfactual logit pairing (CLP), forces the model to produce consistent predictions when sensitive attributes are artificially modified in the input representation [75]. This encourages the model to rely on clinically relevant features rather than demographic proxies.

Post-processing methods are applied after model training and focus on adjusting the final

predictions. These approaches typically modify decision thresholds for different demographic groups in order to balance fairness metrics across populations. Although post-processing can improve fairness metrics without retraining the model, it does not address the underlying sources of bias within the data or the model itself.

3.3.4 Clinical Evidence of Algorithmic Bias

Empirical studies have demonstrated that concerns about fairness are not merely theoretical, but arise in real-world medical AI systems. A notable example is the CheXclusion study, which analyzed fairness gaps in deep learning classifiers trained on major chest X-ray datasets such as MIMIC-CXR, ChestX-ray8, and CheXpert [77]. The study revealed significant disparities in true positive rates across multiple protected attributes, including patient sex, race, age, and insurance status.

In particular, female patients were consistently identified as the least favored subgroup across datasets, experiencing lower true positive rates despite representing nearly half of the training data [77]. Racial disparities were also observed. Within the MIMIC-CXR dataset, for example, Hispanic patients experienced negative diagnostic disparities in the majority of disease categories, while white patients, who represented the majority demographic, generally achieved the highest true positive rates. For certain conditions, such as pneumonia, the difference in true positive rate between the most and least favoured racial groups was as high as 44%. Socioeconomic disparities were also identified, with patients insured through Medicaid showing consistently lower diagnostic performance across multiple disease labels [77].

Importantly, the study found that these disparities were not strongly correlated with the disease burden of the affected subgroups, indicating that simply increasing the number of samples from underrepresented groups may not fully eliminate bias [77]. However, combining data from multiple institutions and datasets was shown to reduce true positive rate disparities, suggesting that diverse, multi-source datasets can play a crucial role in improving fairness in medical imaging models.

3.4. Generalization and Fairness in Lung Cancer Screening

Ensuring that deep learning models perform reliably across diverse clinical settings is a critical hurdle for their implementation in lung cancer screening. Despite the proliferation of high-performing models on internal benchmarks, a critical analysis of the literature

reveals a significant scarcity of studies that rigorously evaluate both generalization and fairness in realistic, out-of-distribution scenarios.

3.4.1 Empirical Evidence of Generalization Challenges

While many models report high accuracy on internal tests (e.g., exceeding 90% AUC on LIDC-IDRI), performance frequently degrades when applied to external datasets that differ in acquisition protocols or population characteristics. Recent studies have begun to expose this gap by performing multi-center external validations specifically for lung nodule detection:

- **Cross-Population Validation:** *Cui et al.*[78] developed a deep learning model on a large Chinese screening cohort and performed external validation using the LUNA public database (derived from the US-based LIDC-IDRI). Their work highlights the critical challenge of maintaining accuracy across distinct geographic and demographic populations [78].
- **Institutional Shift:** *Hendrix et al.*[79] evaluated a malignancy detection system across CT scans acquired in two distinct Dutch hospitals (internal and external validation). They demonstrated that while AI can achieve sensitivities comparable to radiologists, rigorous external testing across different clinical centers is essential to validate real-world reliability [79].
- **Multi-Scanner Benchmarking:** *Tushar et al.*[80] provided a comprehensive benchmark of detection and malignancy classification models across the DLCS, LUNA16, and NLST datasets. Their findings emphasize that dataset characteristics strongly influence AI performance, demonstrating that models trained on curated sets (like LUNA16) suffer significant sensitivity drops when evaluated externally on screening-naive datasets like NLST [80].
- **Patient-Level Data Leakage:** Generalization issues also stem from poor methodological practices. *Ashames et al.*[81] demonstrated that many deep learning studies artificially inflate accuracy by randomly shuffling data without strict patient-wise separation. When models trained with this ”data leakage” are evaluated on entirely new patients, their performance degrades drastically, underscoring the need for strict patient-level isolation to ensure true interpretability and generalization [81].

3.4.2 Fairness Audits in Screening Models

In clinical AI, cross-center generalization refers to a model's capacity to maintain high predictive performance when deployed across different medical institutions, imaging centers, and scanner vendors, without succumbing to performance degradation caused by domain shifts. Beyond cross-center generalization, the few existing studies focusing on algorithmic fairness reveal severe disparities in how models treat different patient subgroups. Fairness in lung cancer screening is not merely about demographic parity; it involves identifying systematic biases in both population groups and nodule morphologies.

- **Demographic Disparities:** *Gaur et al.*[82] performed a comprehensive fairness audit of lung cancer risk estimation models trained on the NLST dataset. Their analysis revealed a severe demographic disparity in model sensitivity. For instance, the Venkadesh21[82] lung nodule classifier achieved a sensitivity of 69% for White participants at a fixed specificity of 90%, but dropped significantly to only 39% for Black participants [82]. These disparities persisted even after controlling for clinical confounders, which the authors attribute primarily to ”health data poverty” and the extreme demographic imbalance in the NLST training data (approximately 93% White).

- **Clinical versus Demographic Bias:** In a large-scale evaluation of pulmonary nodule detection algorithms using the SUMMIT screening dataset, *McCabe et al.*[6] initially observed apparent performance differences between ethnic groups. However, after controlling for confounding variables (particularly sex imbalance within subgroups), the demographic disparities largely disappeared [6]. Instead, the authors found that the most significant biases were clinical. Models systematically underperformed on rarer but clinically critical nodule subtypes, such as non-solid or ground-glass lesions, while achieving high detection rates for common solid nodules.

Ultimately, these concrete examples demonstrate that state-of-the-art models for lung cancer screening are highly susceptible to both institutional domain shifts and demographic/clinical biases. The literature clearly shows a ”local versus global” paradox: models optimized for high performance on specific benchmarks often lack the robustness required for equitable real-world deployment.

3.5. Critical Review of the Literature

The literature on deep learning for lung nodule classification in lung cancer screening has shown substantial methodological progress over the past decade. Early approaches relied mainly on radiomics pipelines combined with traditional ML models, whereas more recent studies increasingly employ deep learning architectures such as ResNet, DenseNet, and other convolutional neural networks. More recently, transformer-based models and hybrid CNN–Transformer architectures have also been explored. Despite these developments, most approaches rely on similar methodological foundations, with innovations typically focusing on architectural refinements rather than fundamentally new paradigms for nodule analysis.

A major limitation across the literature is the strong dependence on a small number of benchmark datasets. Public datasets such as LIDC-IDRI and its derived collections have been widely used to develop and evaluate models, but their repeated use raises concerns regarding dataset-specific bias and limited population diversity. In addition, many studies rely on private institutional datasets that are not publicly accessible, which restricts reproducibility and limits independent validation. This lack of diverse and openly available data complicates the assessment of how models perform in real clinical environments.

Another persistent challenge concerns the generalization ability of deep learning models. Models trained on data from a single dataset or institution often exhibit significant performance degradation when evaluated on external cohorts. Differences in CT acquisition protocols, scanner manufacturers, and patient populations introduce domain shifts that alter the distribution of imaging data. As a result, models may learn representations that are strongly tied to dataset-specific characteristics rather than underlying pathological features.

Closely related to the issue of generalization is the growing concern regarding algorithmic fairness in medical imaging systems. Recent studies show that deep learning models can encode sensitive demographic attributes, such as race, sex, and age, directly from medical images. When training datasets are demographically imbalanced, these representations may lead to disparities in diagnostic performance across population subgroups. In addition, bias may also arise from the uneven representation of different nodule types, potentially affecting the detection of rare but clinically relevant patterns.

In summary, while the literature demonstrates the strong potential of deep learning for lung nodule classification in lung cancer screening, important challenges remain regarding

model robustness and equitable performance. In particular, the relationship between domain shift, model generalization, and algorithmic fairness remains insufficiently explored. Existing studies often investigate either generalization across imaging domains or fairness across demographic groups in isolation, with limited work examining how these factors interact. Addressing these challenges requires systematic evaluation frameworks capable of assessing both generalization and fairness across diverse datasets and patient populations.

4. Datasets

This chapter details the three independent medical imaging datasets utilized in this study and describes the standardized pre-processing pipeline applied to prepare them for model development. Section 4.1 introduces the LUNA25 dataset, while Section 4.2 describes the DLCS dataset. These two screening cohorts form the core of the experimental design: both are utilized symmetrically as independent training and internal validation sets, and each serves as an external evaluation cohort for the other to assess cross-dataset generalization and the transferability of algorithmic fairness. Importantly, they provide complementary demographic profiles essential for robust subgroup disparity analyses. Section 4.3 details the LUNGx dataset, a smaller diagnostic cohort used strictly for out-of-distribution evaluation to assess model resilience against severe label and acquisition shifts. Finally, Section 4.4 outlines the uniform data pre-processing methodology applied across all three cohorts, encompassing spatial alignment, resampling, intensity normalization, and 2D patch extraction.

4.1. LUNA25

The LUNA25 dataset [49, 83–85], derived from the NLST [24], constitutes the primary dataset used in this study. It provides annotated pulmonary nodules with associated clinical and demographic metadata, enabling both classification and subgroup-based analyses. The dataset comprises a total of 2,120 unique patients and approximately 6,166 annotated nodules. Regarding annotation, malignant nodules were confirmed through histopathological examination, while benign cases were verified based on their radiological stability over a median follow-up period of at least six years [85]. Nodule localization and bounding-box annotations were performed by an experienced radiologist (with more than five years of experience in lung cancer screening CT reading) and two medical students, after a dedicated training session. The hidden test cohort was independently annotated by two experienced thoracic radiologists through double reading during screening [85].

The LUNA25 release provides two patient-level features directly: age and gender. All remaining demographic attributes used in this study, namely smoking status, self-reported race/ethnicity, and educational attainment, were obtained by cross-referencing patient identifiers with the broader NLST clinical data repository [24].

The class distribution reflects a typical lung cancer screening scenario, with a relatively low prevalence of malignant cases. A summary of the key demographic and clinical characteristics of the cohort is presented in Table 4.1.

Table 4.1: Summary of demographic and clinical characteristics of the LUNA25 dataset ($N = 2{,}120$ patients).

Characteristic	**Category**	**N (%)**
Diagnosis	Benign	1,787 (84.3%)
	Malignant	333 (15.7%)
Age (years)	Mean $\pm$ SD	62.7 ± 5.1
	Range	55–76
Sex	Male	1,211 (57.1%)
	Female	909 (42.9%)
Smoking status	Current smoker	1,073 (50.6%)
	Former smoker	1,047 (49.4%)
Education	8th grade or less	29 (1.4%)
	9–11th grade	122 (5.8%)
	High school graduate/GED	519 (24.5%)
	Post high school training	297 (14.0%)
	Some college	467 (22.0%)
	College graduate	348 (16.4%)
	Postgraduate/professional	299 (14.1%)
	Unknown	39 (1.8%)
Race (aggregated)	White	1,975 (93.2%)
	Non-White	139 (6.6%)
	Unknown	6 (0.3%)

The demographic characteristics of the cohort highlight specific variations across subgroups. The mean patient age is 62.7 years (standard deviation: 5.1), with ages ranging from 55 to 76 years, as illustrated in Figure 4.1. In terms of sex distribution, the dataset includes 1,211 male patients and 909 female patients (Figure 4.2).

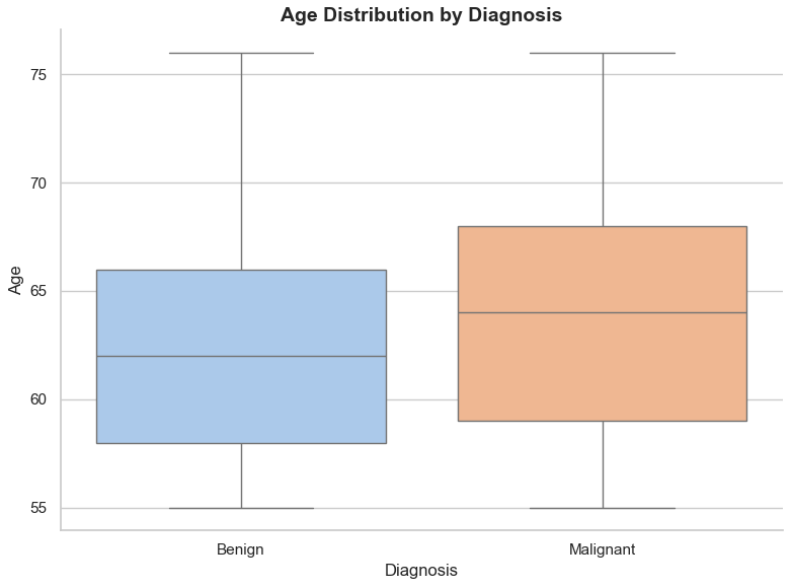


Figure 4.1: Age distribution across benign and malignant diagnoses in the LUNA25 dataset.

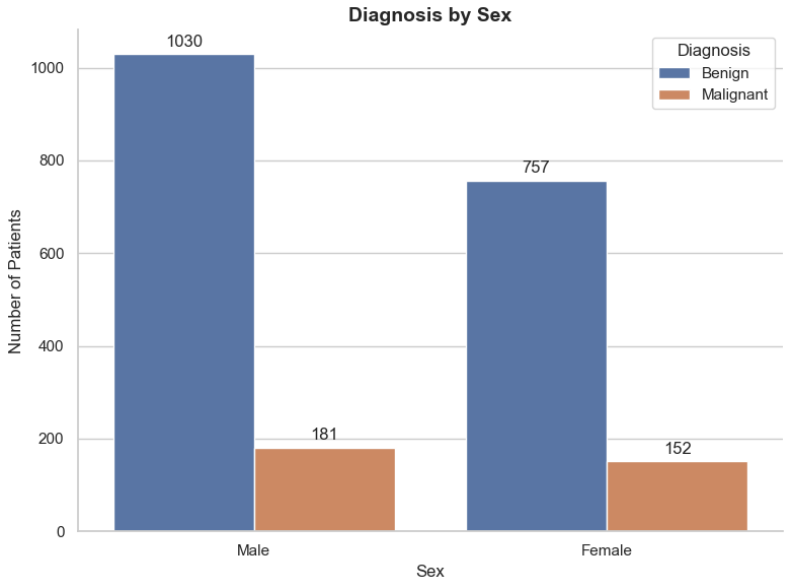


Figure 4.2: Patient distribution by sex and diagnosis in the LUNA25 dataset.

Regarding smoking history, the cohort is composed of 1,073 current smokers and 1,047 former smokers. No patients are labeled as never smokers in the available metadata (Figure 4.3). Furthermore, regarding educational attainment, the dataset presents a varied distribution of socioeconomic backgrounds. As extracted from the NLST metadata, the cohort encompasses a wide spectrum of educational levels, with the largest segments being high school graduates (519 patients) and individuals with some college education

(467 patients). A detailed breakdown of all available educational categories across the cohort is illustrated in Figure 4.4.

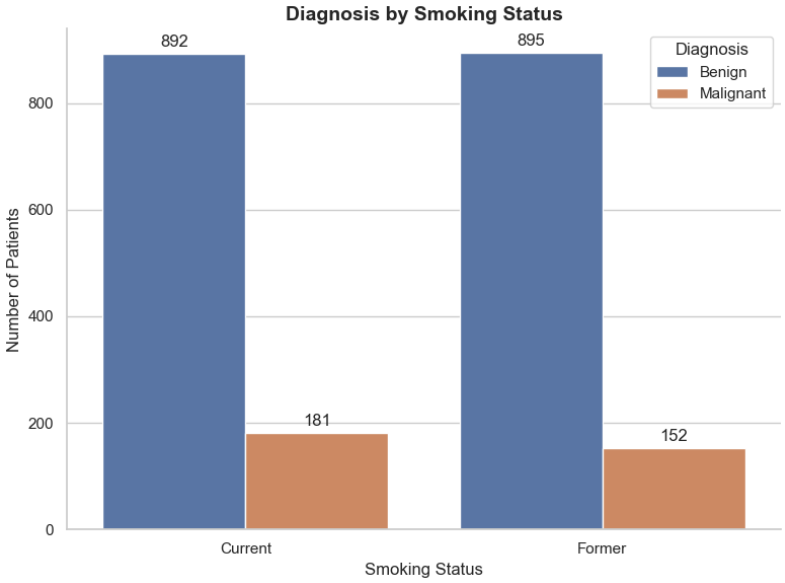


Figure 4.3: Smoking status history across the patient cohort in the LUNA25 dataset.

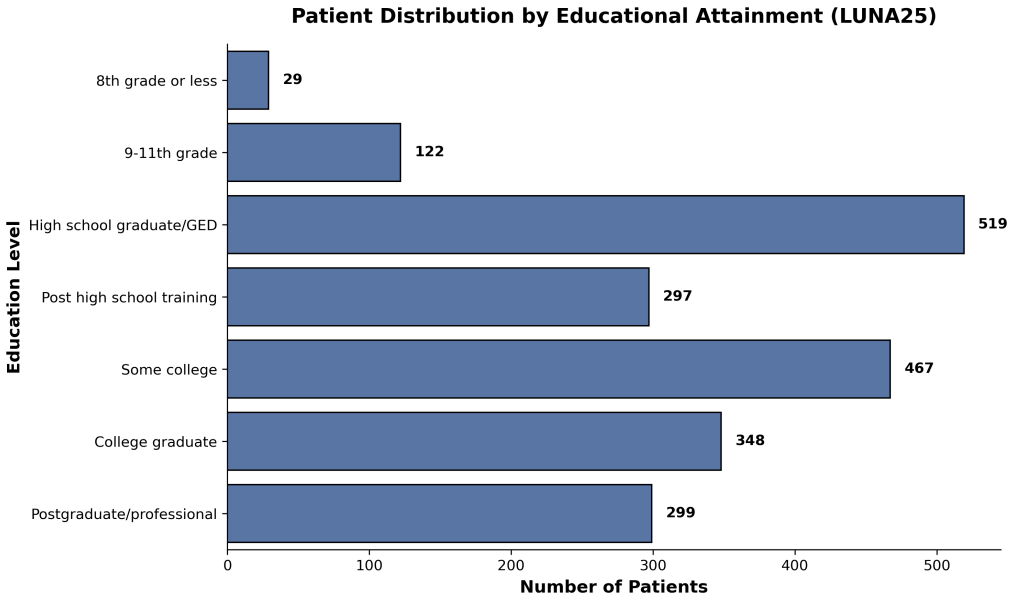


Figure 4.4: Patient distribution by educational attainment across the LUNA25 cohort.

The racial composition of the dataset is notably imbalanced. As shown in Figure 4.5, the NLST data provides seven distinct racial/ethnic categories. However, most minority groups are represented by very small numbers of patients, for example, Asian (5 patients), Black or African-American (17 patients), Native Hawaiian or Other Pacific Islander (1 patient), More than one race (1 patient), and American Indian or Alaskan Native (1 patient) — which would render any per-group statistical analysis unreliable. To ensure meaningful

subgroup comparisons, these categories were therefore aggregated into a single Non-White group, yielding a binary racial split of White and Non-White, with a small number of cases of unknown race excluded. The resulting distribution, consisting of 1,975 White patients (93.2%) and 139 Non-White patients (6.6%), is illustrated in Figure 4.6. This strong predominance of a single racial group is a known characteristic of NLST recruitment and is relevant for the fairness analyses presented in subsequent chapters.

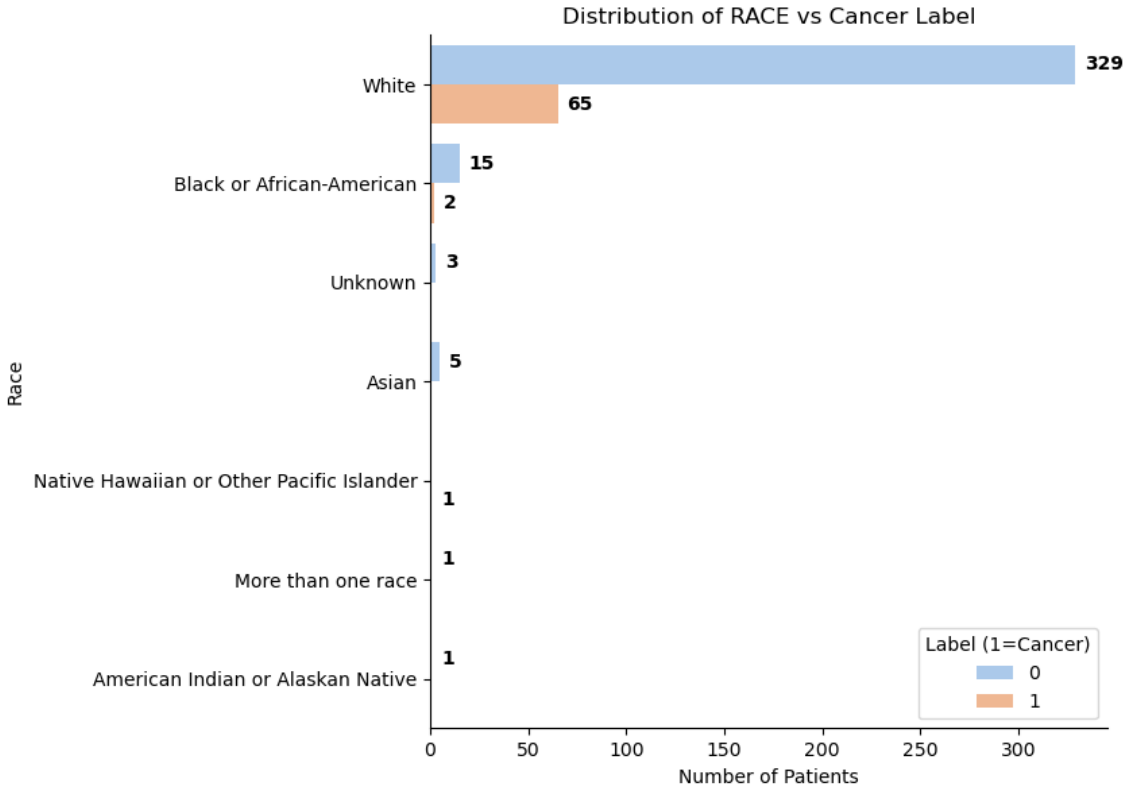


Figure 4.5: Original racial/ethnic distribution across all seven categories available in the NLST data, broken down by cancer label. The low patient counts in minority groups motivate the binary White/Non-White aggregation used throughout this study.

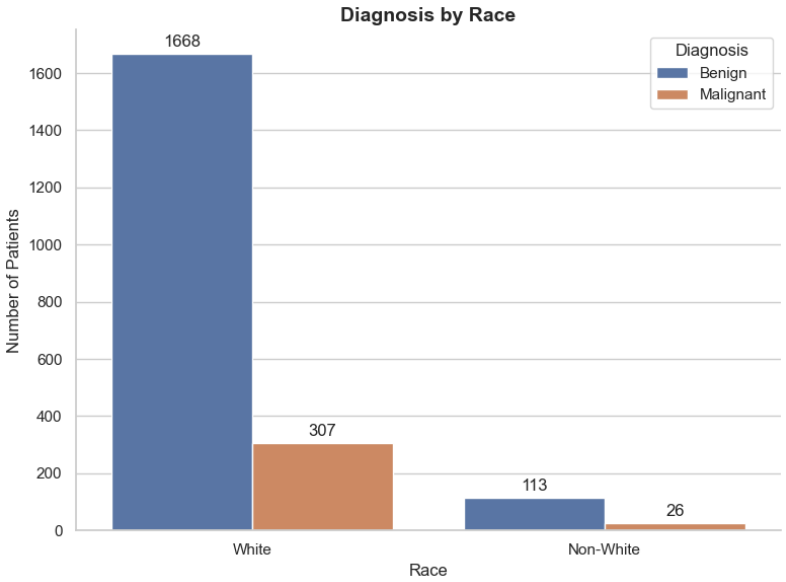


Figure 4.6: Racial composition after aggregation into White and Non-White groups, highlighting the significant demographic imbalance in the LUNA25 dataset.

These demographic attributes are relevant for subsequent subgroup-based evaluation, particularly in the context of fairness analysis.

To provide a qualitative understanding of the input data, Figure 4.7 illustrates representative examples of benign and malignant nodules extracted from the LUNA25 dataset. These examples highlight the visual variability and subtle morphological differences that characterize the classification task.

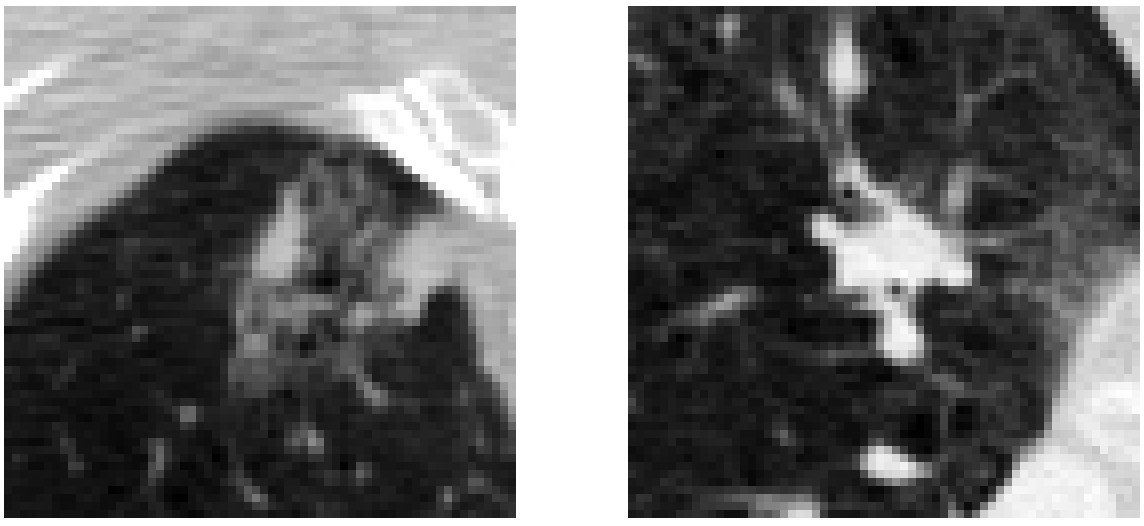


Figure 4.7: Representative examples of benign (left) and malignant (right) nodules from the LUNA25 dataset.

4.2. DLCS

The DLCS dataset [86–88] is used as an additional cohort to complement the primary dataset. It contains a total of 2,483 annotated nodules corresponding to 1,611 unique patients.

A summary of the key demographic and clinical characteristics of the DLCS cohort is presented in Table 4.2, followed by their detailed distributions in Figures 4.8–4.11.

Table 4.2: Summary of demographic and clinical characteristics of the DLCS dataset ($N = 1{,}611$ patients).

Characteristic	**Category**	**N (%)**
Diagnosis	Benign	1,468 (91.1%)
	Malignant	143 (8.9%)
Age (years)	Mean $\pm$ SD	66.7 ± 6.2
	Range	50–89
Sex	Male	808 (50.2%)
	Female	803 (49.8%)
Smoking status	Current smoker	827 (51.3%)
	Former smoker	701 (43.5%)
	Unknown	83 (5.2%)
Race (aggregated)	White	1,191 (73.9%)
	Non-White	420 (26.1%)

The prevalence of malignant cases in this dataset is lower than in LUNA25, with 8.9% of patients labeled as malignant and 91.1% as benign. The mean age is 66.7 years (standard deviation: 6.2), with a broader range spanning from 50 to 89 years, as illustrated in Figure 4.8.

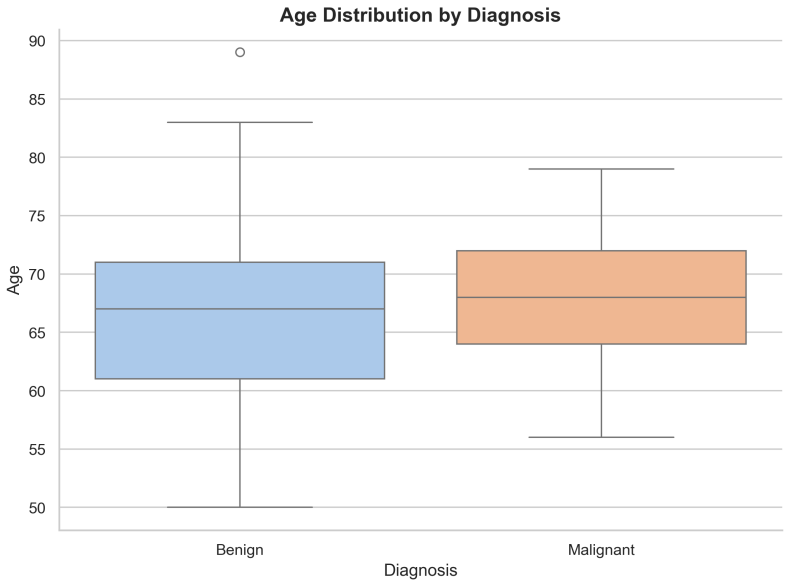


Figure 4.8: Age distribution across benign and malignant diagnoses in the DLCS dataset.

In contrast to LUNA25, the sex distribution in the DLCS dataset is nearly perfectly balanced, comprising 808 male and 803 female patients (Figure 4.9).

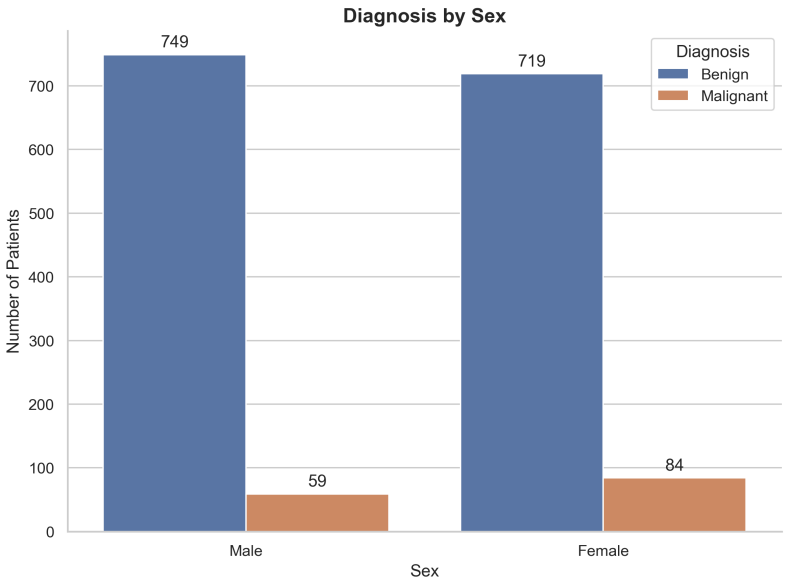


Figure 4.9: Patient distribution by sex and diagnosis in the DLCS cohort.

The dataset also includes information on smoking status. A total of 827 patients are current smokers, 701 are former smokers, and 83 cases are labeled as unknown (Figure 4.10).

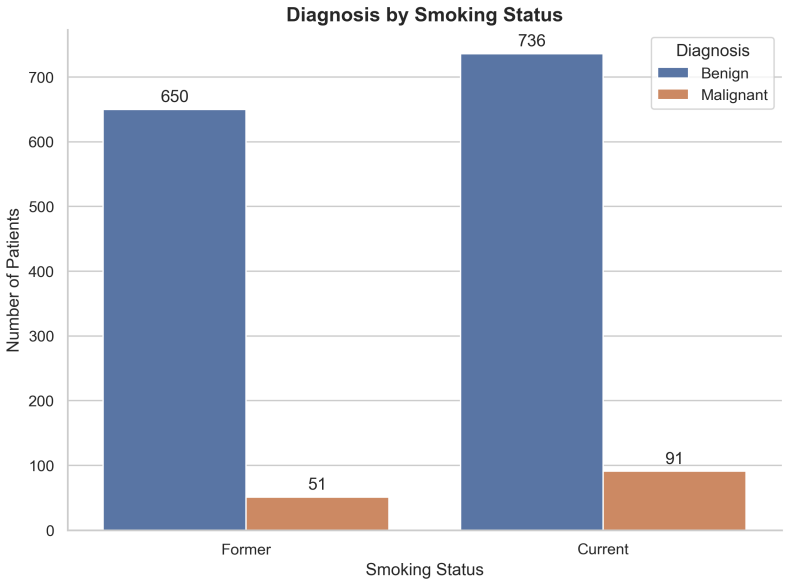


Figure 4.10: Smoking status history across the patient cohort in the DLCS dataset.

Crucially, the racial distribution is more diverse compared to LUNA25. The dataset includes 1,191 White patients and 420 Non-White patients (Figure 4.11), resulting in a significantly higher representation of minority groups (approximately 26% of the cohort).

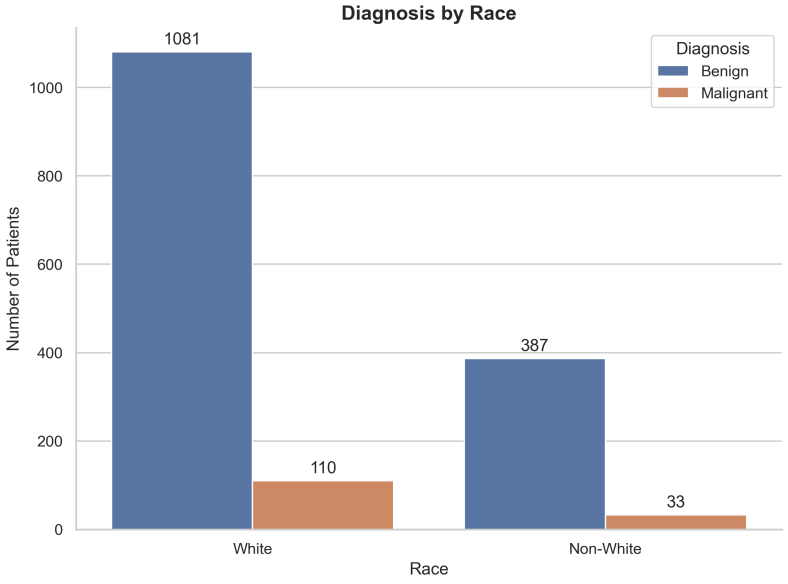


Figure 4.11: Racial composition illustrating a more diverse representation of minority groups compared to the primary dataset.

These characteristics provide complementary demographic variability relative to the primary dataset, establishing the DLCS cohort as a robust external validation set for both generalization and fairness evaluations.

Similarly, Figure 4.12 presents representative benign and malignant nodules from the DLCS dataset.

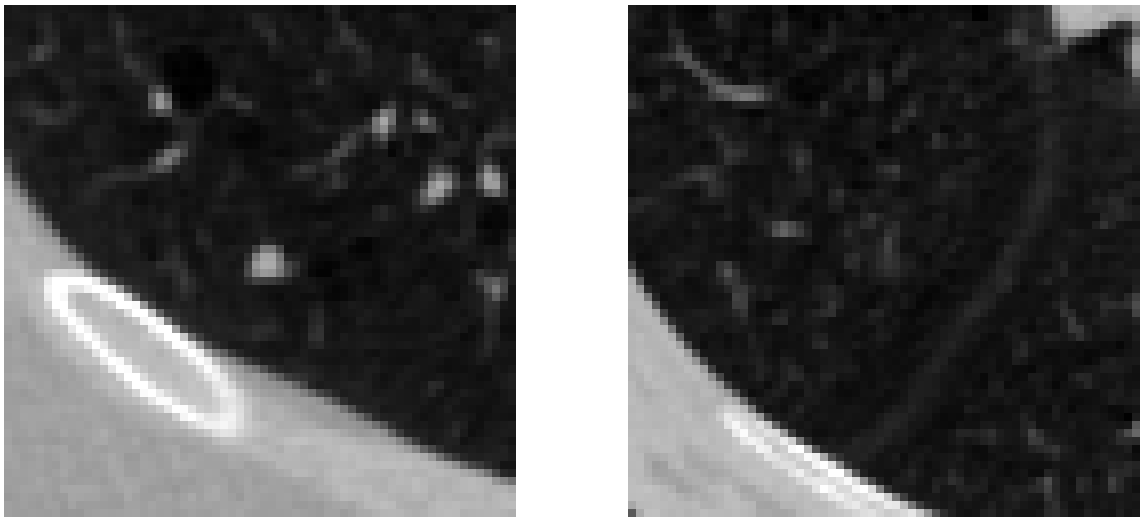


Figure 4.12: Representative examples of benign (left) and malignant (right) nodules from the DLCS dataset.

4.2.1 Comparison of CT Acquisition Parameters

Beyond patient demographics, the technical heterogeneity of the CT acquisition process introduces a significant source of domain shift. To provide further context on the institutional differences between the primary screening cohorts, Figure 4.13 presents a comparative exploratory data analysis of the scan-level acquisition parameters across the LUNA25 and DLCS datasets.

Notable differences exist in the hardware and protocol distributions. While both datasets prominently feature Siemens and GE Medical Systems scanners, DLCS exhibits a higher concentration of GE devices. More critically, the reconstruction slice thickness reveals a severe distribution shift: DLCS is heavily dominated by ultra-thin slices (<= 1.5 mm), whereas LUNA25 predominantly consists of standard screening thicknesses (1.5-3.0 mm). Variations are also evident in the X-Ray tube voltage (kVp), with DLCS utilizing a larger proportion of lower energy scans (<= 100 kVp) compared to the 101-125 kVp majority in LUNA25. Regarding reconstruction kernels, both datasets favor smooth kernels, although LUNA25 includes a visible subset of sharp reconstructions.

The distribution of these heterogeneous acquisition parameters visually confirms the presence of complex covariate shifts between the two institutions, forming the basis for the image-level fairness and domain shift audits presented in Chapter 6.

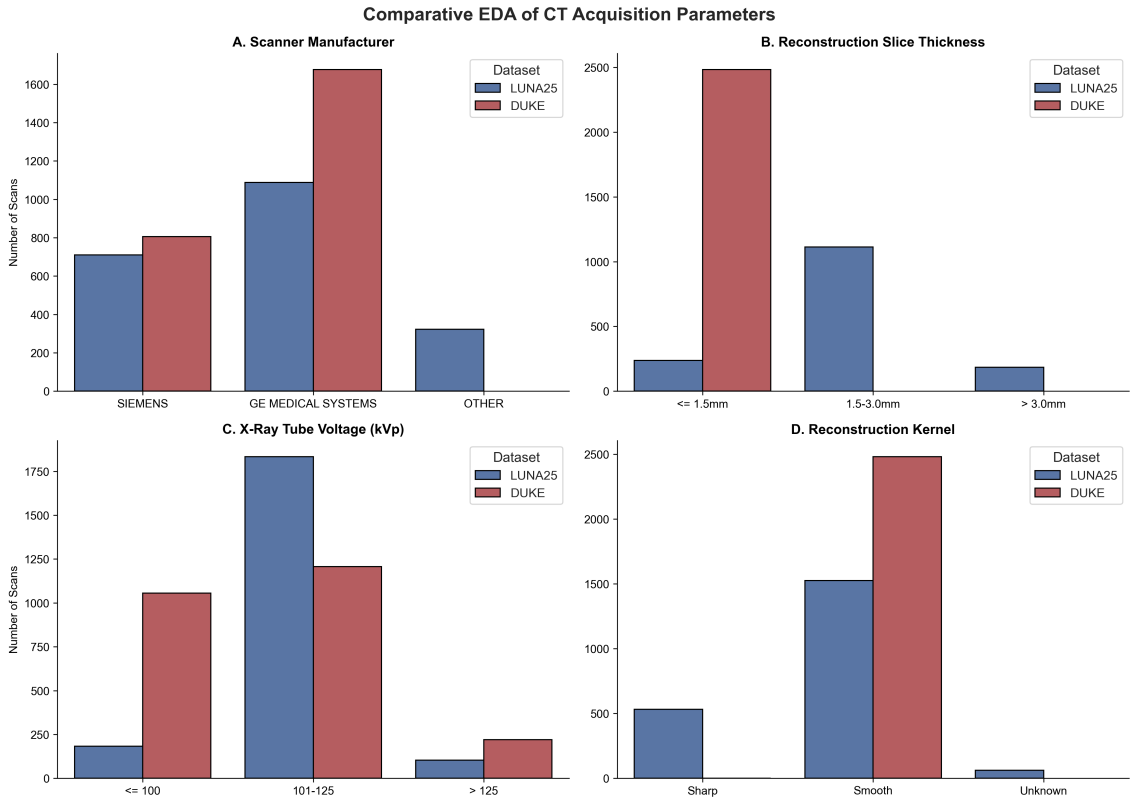


Figure 4.13: Comparative exploratory data analysis of CT acquisition parameters across the LUNA25 and DLCS datasets. The distributions highlight significant institutional shifts in (A) scanner manufacturer, (B) reconstruction slice thickness, (C) X-Ray tube voltage, and (D) reconstruction kernels.

4.3. LUNGx

The LUNGx dataset [51, 52] is a publicly available benchmark dataset designed for lung nodule classification. Unlike LUNA25 and DLCS, which were collected as part of lung cancer screening programmes targeting high-risk asymptomatic individuals, LUNGx consists of CT scans acquired in a diagnostic clinical setting, where patients were referred for workup of a pulmonary lesion [51, 52]. This distinction is clinically relevant: diagnostic CT scans are typically acquired at higher dose and with different acquisition parameters compared to low-dose screening CT, and the underlying patient population may present with more advanced or symptomatic disease. As a consequence, LUNGx serves exclusively as an out-of-distribution test set to probe model generalisation under a simultaneous shift in acquisition protocol and patient population, rather than as a representative evaluation of screening performance.

The dataset does not provide demographic metadata; therefore it is not used for subgroup-based fairness analyses.

The dataset contains 83 annotated nodules from a total of 70 unique patients. Table 4.3 summarises the distribution of annotated nodules per patient: 57 patients (81.4%) present a single annotated nodule, while the remaining 13 patients (18.6%) present two annotated nodules.

Table 4.3: Distribution of annotated nodules per patient in the LUNGx dataset ($N = 70$ patients, 83 nodules).

Nodules per patient	N patients	%
1	57	81.4%
2	13	18.6%
Total	**70**	**100%**

In contrast to typical screening datasets like LUNA25 and DLCS, the class distribution in LUNGx is approximately balanced. Specifically, 41 nodules (49.4%) are labeled as malignant, while 42 nodules (50.6%) are benign (Figure 4.14). This deliberate 50/50 balance differs substantially from the malignancy prevalence observed in screening populations (approximately 9–16% in LUNA25 and DLCS), and constitutes a severe label shift that is exploited in the out-of-distribution evaluation.

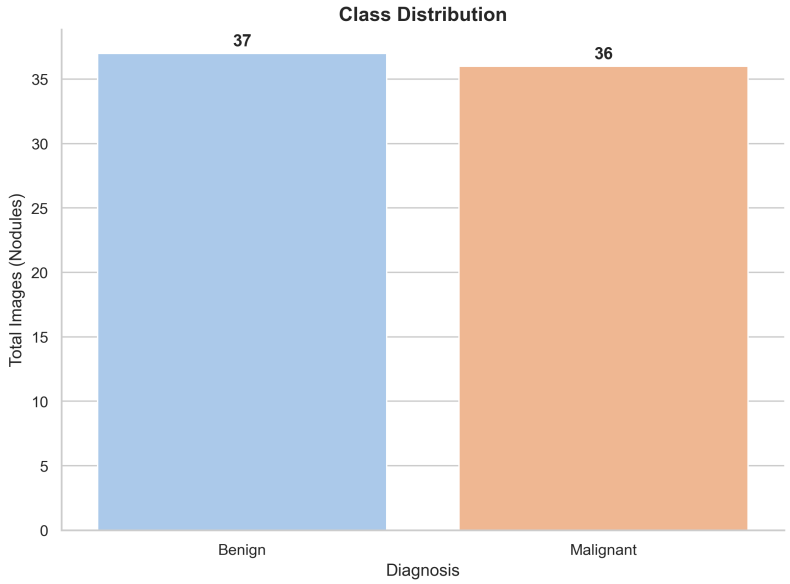


Figure 4.14: Class distribution of benign and malignant nodules in the LUNGx dataset.

Given its diagnostic (non-screening) acquisition context, deliberate class balance, and distinct data origin, LUNGx is utilised strictly for out-of-distribution evaluation to assess model generalisation and resilience against simultaneous acquisition and label shift.

Figure 4.15 shows representative examples of benign and malignant nodules from the LUNGx dataset. Despite the absence of demographic metadata, the dataset provides visually diverse cases that are useful for assessing model generalisation under different data distributions.

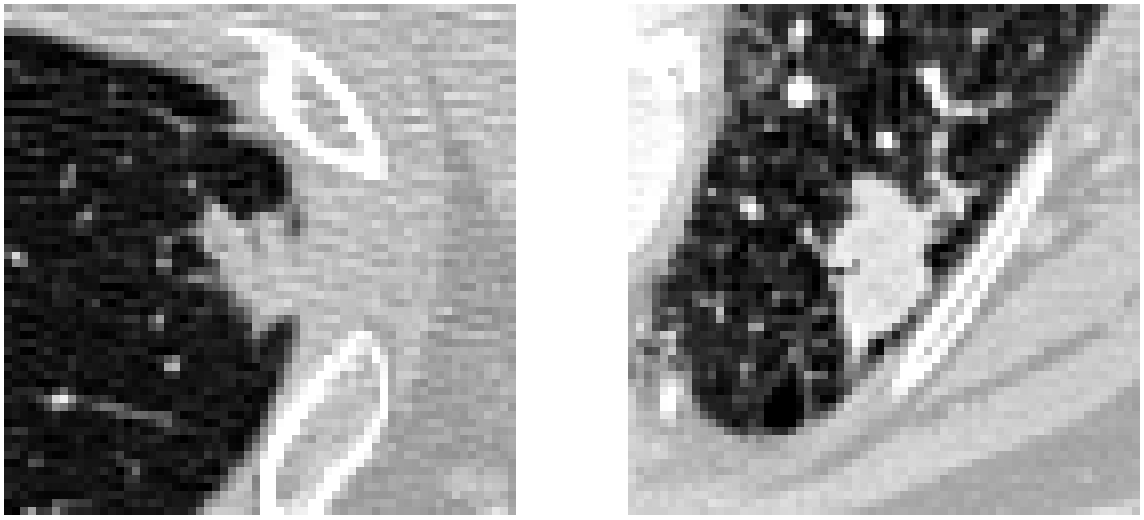


Figure 4.15: Representative examples of benign (left) and malignant (right) nodules from the LUNGx dataset.

4.4. Data Pre-processing

Prior to model development, all datasets were subjected to a standardized pre-processing pipeline, originally established by *Silva et al.* [89], during dataset curation, ensuring consistency across cohorts and mitigating variability arising from heterogeneous CT acquisition protocols. This pipeline, illustrated in Figure 4.16, summarizes the sequence of transformations applied to the raw CT scans, from spatial alignment and resampling to intensity normalization and the extraction of standardized nodule-centered regions of interest. The process was applied uniformly to the LUNA25, DLCS, and LUNGx datasets, with minor dataset-specific spatial adjustments introduced where necessary.

4.4.1 Spatial Alignment

Because CT scans from different institutions often present varying coordinate systems, spatial alignment was enforced during dataset preparation. Axis transposition and, where applicable, axis flipping protocols were applied (notably in the DLCS and LUNGx datasets) to ensure a consistent anatomical orientation across all volumes.

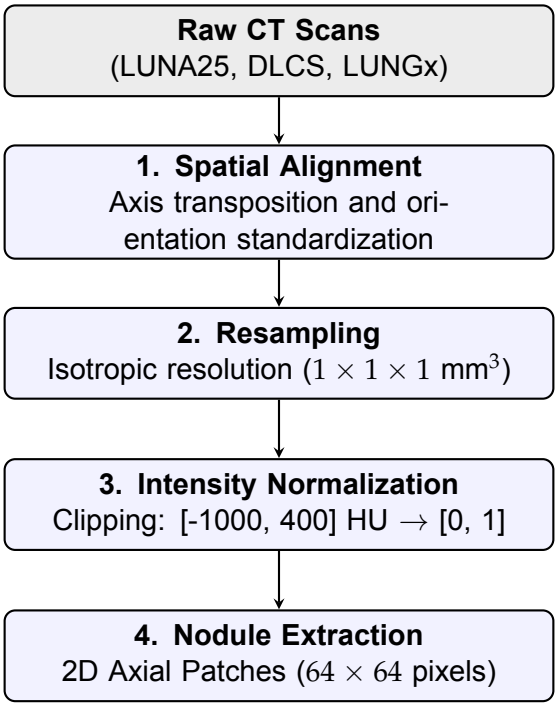


Figure 4.16: Overview of the standardized data pre-processing pipeline applied to all datasets prior to model development.

4.4.2 Resampling

To address variability introduced by differences in scanner characteristics, particularly pixel spacing and slice thickness, all CT volumes were resampled to a uniform isotropic resolution of $1 \times 1 \times 1$ mm^3 using interpolation.

4.4.3 Intensity Normalization

To standardize intensity distributions and reduce inter-scan variability, all images were clipped to the range [-1000, 400] HU. This interval captures the relevant spectrum of lung tissue and nodule structures while excluding extreme values corresponding to background air and dense bone.

Following clipping, voxel intensities were linearly transformed to the range [0, 1] using min-max normalization. Values originally below -1000 HU were mapped to 0, and values above 400 HU were mapped to 1. This normalization step ensures numerical stability during the training of deep learning models [48].

4.4.4 Nodule Localization and 2D Patch Extraction

Pulmonary nodules were localized using bounding box annotations provided by each dataset. In this study, a 2D representation of lung nodules was adopted. Specifically,

each nodule was represented by a single axial slice centered on the annotated centroid, extracted as a fixed-size region of interest (ROI) of 64×64 pixels [56].

For nodules located near image boundaries, bounding box center shifting was applied to preserve the expected spatial dimensions without introducing padding artifacts. While volumetric 3D approaches capture richer spatial context, the use of standardized 2D patches significantly reduces computational complexity while preserving the critical morphological features required for malignancy classification.

4.4.5 Dataset Partitioning and Class Balance

To evaluate model performance and fairness without data leakage, both LUNA25 and DLCS datasets were partitioned into development (train+validation) and held-out test splits. For LUNA25, a stratified group split strategy was employed, whereas for DLCS, the official benchmark split structure was preserved.

As detailed in Table 4.4, the label prevalence (malignancy rate) and demographic distributions remain highly consistent between the development and test splits across both cohorts, confirming that the evaluation sets accurately reflect the underlying population distributions without introducing artificial selection bias.

Table 4.4: Comparison of target label prevalence and key demographic attributes between the development (*Train+Val*) and evaluation (*Test*) splits for LUNA25 and DLCS.

	LUNA25		DLCS	
Attribute / Category	**Train+Val (%)**	**Test (%)**	**Train+Val (%)**	**Test (%)**
Malignancy Prevalence				
Benign	84.3%	84.4%	91.2%	90.9%
Malignant	15.7%	15.6%	8.8%	9.1%
Age (Mean years)	62.7	62.8	66.7	66.8
Sex / Gender				
Female	42.4%	44.8%	48.8%	57.1%
Male	57.6%	55.2%	51.1%	42.9%
Race / Ethnicity				
Caucasian / White	93.2%	93.2%	74.3%	70.2%
Black / African American	3.3%	3.1%	22.3%	26.3%
Other / Unknown	3.5%	3.7%	3.2%	3.5%
Smoking Status				
Current Smoker	50.9%	49.3%	51.4%	50.5%
Former Smoker	49.1%	50.7%	43.3%	44.9%
Unknown	–	–	5.1%	4.5%
Education Level				
High school graduate/GED	23.3%	29.0%	–	–
Associate degree/some college	22.1%	21.9%	–	–
Bachelor's degree	16.7%	15.3%	–	–
Graduate school	15.0%	10.6%	–	–
Post high school training (non-college)	14.4%	12.5%	–	–
9th-11th grade	5.5%	6.8%	–	–
8th grade or less	1.4%	1.2%	–	–
Other / Unknown	1.7%	2.6%	–	–

5. Methodology

This chapter details the methods and experimental design adopted in this thesis. Figure 5.1 presents a high-level overview of the complete experimental pipeline, spanning from data augmentation and class balancing to model optimization and evaluation.

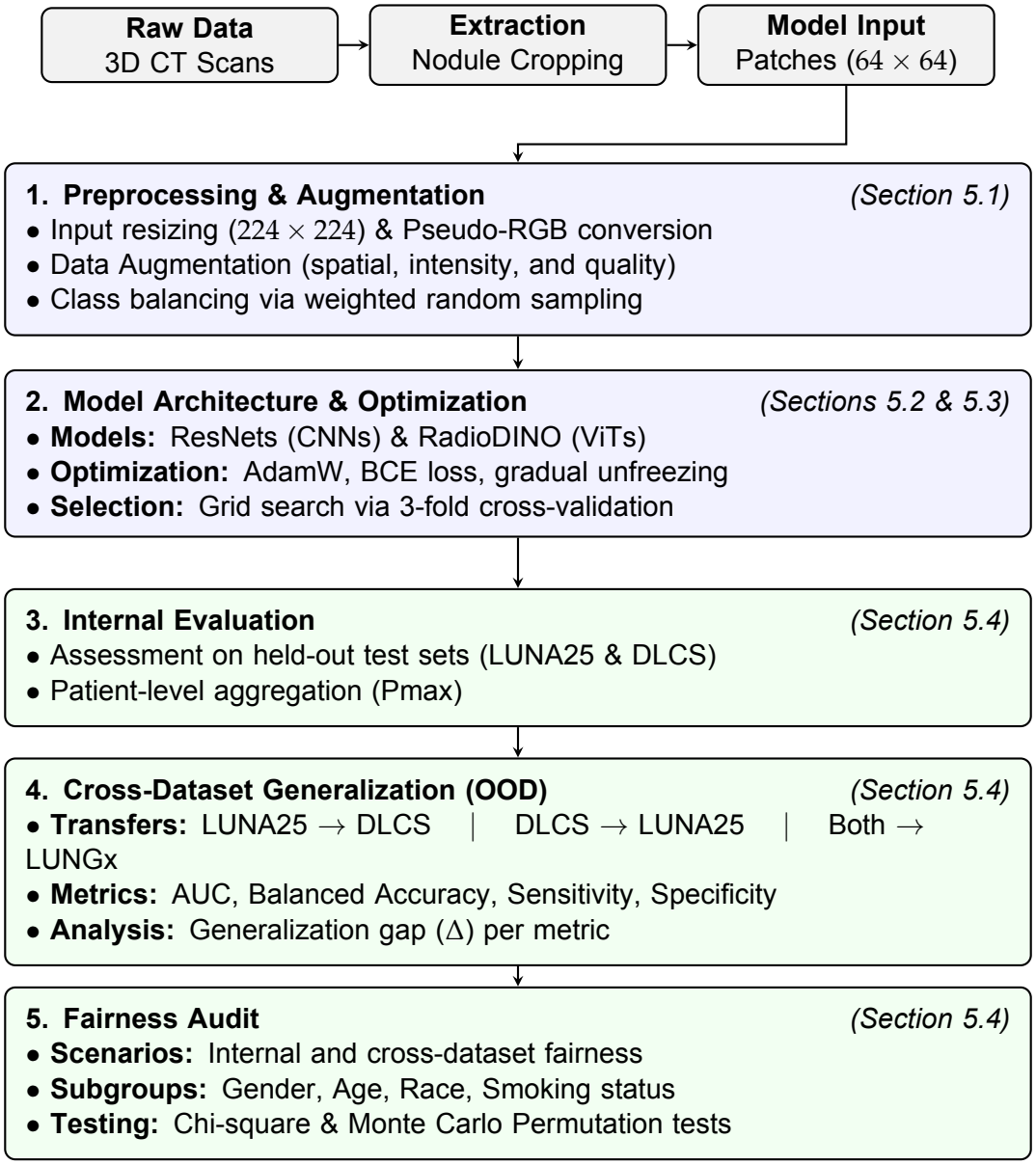


Figure 5.1: Unified experimental pipeline detailing the progression from raw data input through training, evaluation, and fairness auditing. The framework separates the initial model development phases (Steps 1 and 2) from the downstream testing and disparity analysis scenarios (Steps 3 to 5).

5.1. Data Augmentation and Class Balancing

Deep learning models require sufficiently diverse training data to achieve robust generalization. In medical imaging scenarios, this requirement is often challenged by both limited dataset size and pronounced class imbalance, as observed in the LUNA25 and DLCS cohorts. To mitigate these issues, this work employs a combination of stochastic data augmentation and class-aware sampling during the training phase.

5.1.1 Input Adaptation

Prior to augmentation and network ingestion, input adaptation was required. Although the nodule patches were originally extracted as 64×64 single-channel grayscale images, the deep learning architectures utilized in this study (both CNNs and Vision Transformers) rely on pre-trained feature extractors that expect standard pseudo-RGB inputs. Consequently, all input images were converted to a 3-channel format by duplicating the grayscale channel along the color axis. Subsequently, the images were resized to a uniform spatial resolution of 224×224 pixels and normalized using standard ImageNet statistics (mean: $[0.485, 0.456, 0.406]$, standard deviation: $[0.229, 0.224, 0.225]$) to ensure optimal compatibility with the ImageNet [58] pre-trained weights.

5.1.2 Data Augmentation

Data augmentation was applied exclusively to the training data, while validation and test sets remained unaltered to ensure an unbiased evaluation of model performance. All transformations were implemented using the `Albumentations` library [90]. For each training sample, a sequence of stochastic transformations was applied with a fixed probability of $p = 0.5$ per operation. The specific augmentation parameters were adopted following established protocols in prior literature [91], simulating both anatomical variability and acquisition-related noise:

- **Spatial transformations:** random rotations (up to $20°$), random scaling (up to $\pm 15\%$), and elastic deformations ($\alpha = 1$, $\sigma = 10$), promoting invariance to orientation, size, and local morphological variations.
- **Image quality perturbations:** Gaussian noise, Gaussian blur (kernel size between 3 and 5, $\sigma \in [0.25, 0.5]$), and a sharpening operation ($\alpha \in [0.1, 0.2]$, lightness $\in [0.5, 1.0]$) to simulate differences in scanner quality and reconstruction artifacts.

- **Intensity transformations:** random brightness and contrast adjustments (up to $\pm 10\%$), accounting for variability in acquisition conditions across institutions.

To visually demonstrate the effect of these operations, Figure 5.2 illustrates the stochastic augmentation pipeline applied to two representative nodule patches.

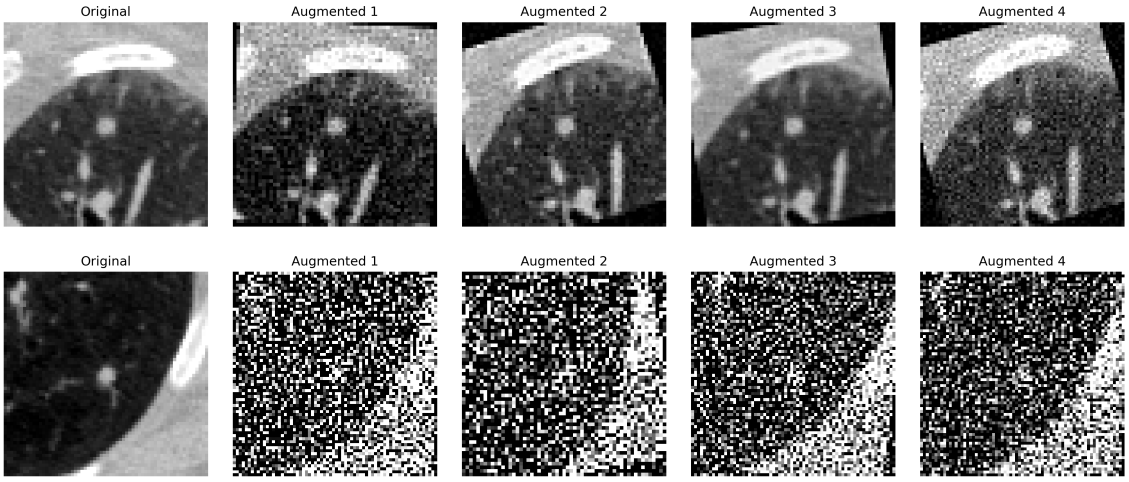


Figure 5.2: Representative examples of the stochastic data augmentation pipeline applied to two sample nodules from the DLCS dataset (top and bottom rows). Both nodules are subjected to the exact same stochastic pipeline. The original images (left) are subjected to sequences of spatial, image quality, and intensity transformations to generate diverse training variants (Augmented 1-4). Because the parameters are randomized in each execution, the resulting transformations can range from subtle morphological shifts (as seen in the top row) to extreme, noise-heavy perturbations (as seen in the bottom row). This high variance is intentional, forcing the network to maintain robust feature extraction even under severely degraded image quality.

5.1.3 Class Balancing Strategy

Both LUNA25 and DLCS datasets exhibit significant class imbalance, with malignant nodules representing a distinct minority of the samples (approximately 15.7% in LUNA25 and 8.9% in DLCS, as detailed in Table 4.1 and Table 4.2). Training directly on such skewed distributions can bias the model toward the majority class, leading to poor sensitivity in detecting malignant cases.

To address this issue, class imbalance was handled at the sampling level using a weighted random sampling approach. To determine the sampling probability for each instance, let N_c denote the total number of samples belonging to class c in the training split. Each training sample is assigned an inverse-frequency weight w_c, defined as:

$$w_c = \frac{1}{N_c} \tag{5.1}$$

Batches were subsequently constructed by sampling with replacement according to these weights. This strategy results in an approximately balanced class distribution within each training epoch, effectively oversampling the minority class without modifying the underlying dataset. By enforcing balanced exposure to both classes during training, this approach promotes the learning of discriminative features for malignant nodules, improving sensitivity while maintaining overall predictive performance.

5.2. Model Architecture

This work employs two distinct families of DL architectures for lung nodule classification: CNNs, represented by ResNets [55], and transformer-based foundation models, represented by RadioDINO [92]. These architectures differ fundamentally in how they extract and represent visual features, providing complementary inductive biases for medical image analysis.

5.2.1 ResNet-based Models

ResNets [55] were selected due to their ability to facilitate stable training of deep architectures through residual learning, defined as:

$$y = F(x, \{W_i\}) + x \tag{5.2}$$

where $F(x, \{W_i\})$ represents the residual mapping and x is propagated through identity shortcut connections.

The models used in this work include ResNet18, ResNet34, and ResNet50, covering both standard residual blocks and bottleneck architectures. While ResNet18 and ResNet34 rely on basic residual blocks, ResNet50 employs a bottleneck design consisting of 1×1–3×3–1×1 convolutions to reduce computational complexity while maintaining representational capacity.

To adapt all ResNet variants for the binary classification task, the final fully connected layer was replaced with a dropout-regularized linear layer:

$$\hat{y} = \text{Linear}(\text{Dropout}(z)) \tag{5.3}$$

where z denotes the feature vector produced by the backbone. A single output neuron was utilized, maintaining consistency with the `BCEWithLogitsLoss` objective. Regarding the input representations, the networks ingested pseudo-RGB 224×224 images, as detailed

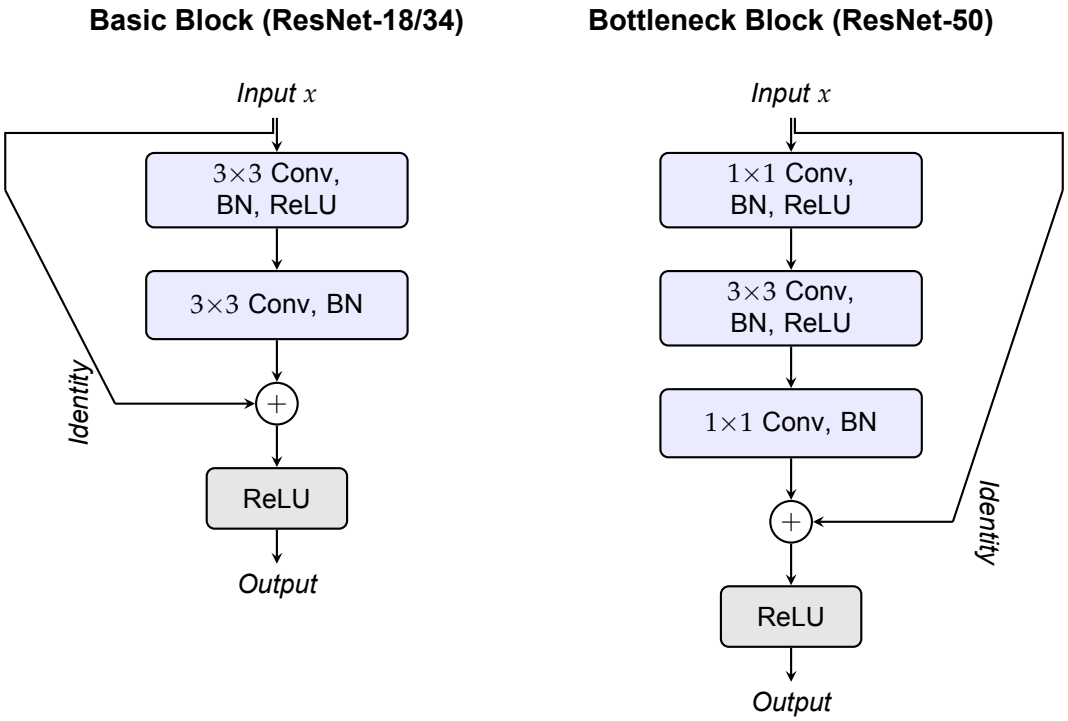


Figure 5.3: Structural comparison of the residual blocks used in the CNN architectures. The Basic Block (left) applies two stacked 3×3 convolutions with a direct identity shortcut, suitable for shallower networks. The Bottleneck Block (right) reduces channel dimensionality with an initial 1×1 convolution before the 3×3 convolution and restores it afterwards, significantly reducing computational cost while preserving representational capacity in deeper networks.

in Section 5.1. To ensure architectural comparability, all models were trained using this uniform input format, avoiding modifications to the initial convolutional layers and reducing potential sources of bias in cross-model evaluation.

To assess the impact of pretraining, ResNet models were trained under two initialization regimes: from scratch with random weight initialization, and using ImageNet [58] pre-trained weights. In both cases, all network parameters were optimized on the target datasets, enabling adaptation to the medical imaging domain.

5.2.2 RadioDINO and Transformer-based Models

RadioDINO [92] is a transformer-based foundation model built upon ViTs. Input images are divided into non-overlapping patches, which are embedded as tokens and processed by the transformer layers.

These models capture global relationships through self-attention mechanisms, defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.4}$$

RadioDINO is pre-trained on large-scale radiological data, building upon the self-supervised knowledge distillation frameworks introduced in DINOv1 [93] and DINOv2 [94]. In this paradigm, a student network learns to match the outputs of a momentum-updated teacher network across multiple augmented views of the same image. This training strategy enables the model to learn rich, robust semantic representations without requiring manual annotations. This architecture was included to evaluate whether large-scale self-supervised pretraining on radiological data improves generalization and fairness compared to standard CNN-based approaches.

For model adaptation, the RadioDINO backbones (Small, Base, and parameters-reduced variants) were imported via the `timm` library [95] using their pre-trained weights. The original classification head was discarded (`num_classes=0`) and substituted with a task-specific head:

$$\hat{y} = \text{Linear}(\text{Dropout}(z)) \tag{5.5}$$

where z corresponds to the global feature representation produced by the transformer. The entire model was fine-tuned end-to-end under the exact same pseudo-RGB input format and training protocol applied to the CNN-based architectures. Unlike the ResNet models, no from-scratch training was performed for RadioDINO, as the study focuses on evaluating its pre-trained representations.

5.2.3 Architectural Considerations

The architectures explored in this work embody two fundamentally different representation learning paradigms.

ResNet models rely on convolutional operations and strong inductive biases such as locality and translation invariance, which are well-suited for capturing fine-grained spatial patterns such as nodule texture and shape. Their hierarchical feature extraction process progressively builds complex representations from local structures.

In contrast, transformer-based models such as RadioDINO rely on global self-attention, enabling direct modeling of long-range dependencies across the image. This allows the model to integrate contextual information more effectively, which may be beneficial in medical imaging scenarios where subtle patterns depend on broader anatomical context.

From a generalization perspective, CNNs tend to perform well in data-constrained settings due to their strong priors, while transformer-based foundation models leverage large-scale

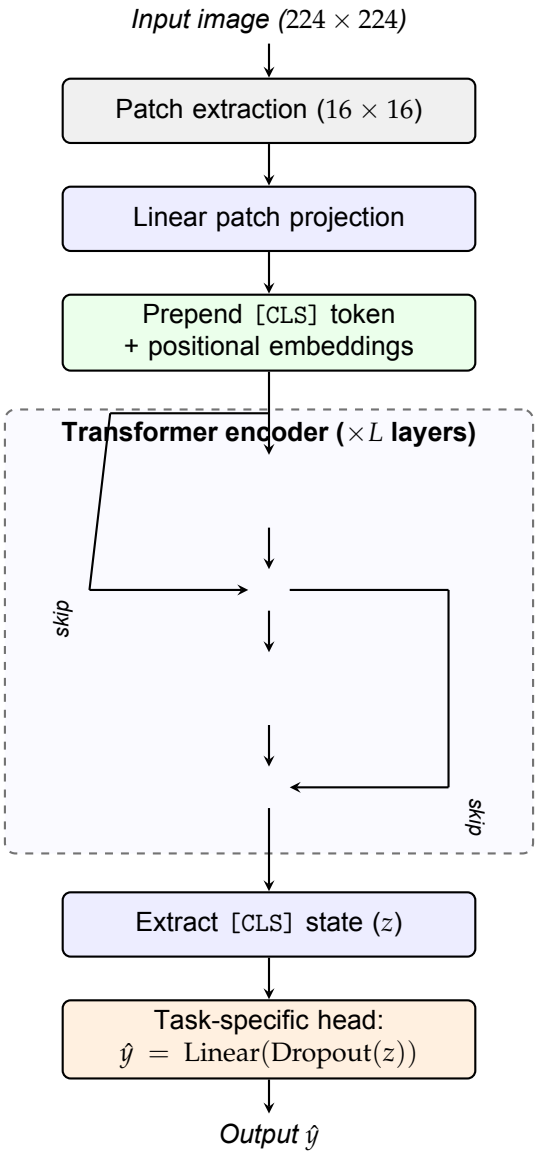


Figure 5.4: Architectural overview of the Vision Transformer (ViT) backbone used by RadioDINO. The input image is divided into non-overlapping 16×16 patches, linearly projected into token embeddings, and prepended with a learnable [CLS] token and positional encodings. The resulting sequence is processed by L stacked encoder blocks, each comprising a layer-normalised multi-head self-attention (MHSA) module and a layer-normalised MLP feed-forward network, both with residual skip connections. The global representation z is extracted from the [CLS] token at the final encoder layer and forwarded to a task-specific classification head.

pretraining to learn transferable representations. Evaluating both paradigms enables a direct comparison of their impact on cross-dataset generalization and fairness, which are central objectives of this work.

5.3. Model Optimization

This section describes the training and optimization procedures adopted for all models, including hyperparameter selection, training strategies, and dataset-specific protocols. The optimization pipeline was designed to ensure robust convergence, fair model comparison, and effective adaptation of pre-trained architectures to the medical imaging domain.

5.3.1 Training Procedure

All models were trained using mini-batch stochastic optimization. Given a batch of input images X and corresponding binary labels $y \in \{0, 1\}$, the model outputs continuous logits $f_\theta(x_i)$. These logits are mapped to predicted probabilities $p_i = \sigma(f_\theta(x_i))$ via the sigmoid function. The optimization objective was to minimize the standard Binary Cross-Entropy (BCE) loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \tag{5.6}$$

where N represents the batch size, y_i is the ground truth label for the i-th sample, and p_i is the predicted probability of malignancy.

Optimization was performed using the AdamW optimizer across all architectures (both ResNet-based and Transformer-based). AdamW decouples weight decay from gradient updates, which consistently improves generalization. The optimizer was applied only to parameters with `requires_grad=True`, allowing dynamic control over trainable layers.

For each epoch, the model was updated through standard forward and backward passes, followed by gradient-based parameter updates. The average training loss over all batches was used to monitor optimization dynamics, while validation performance was used for model selection.

5.3.2 Early Stopping and Model Selection

All models were trained for a maximum of 100 epochs using a fixed batch size of 16. To prevent overfitting, early stopping was employed using validation AUC as the primary monitoring metric. Specifically, training was prematurely interrupted if no improvement in the validation AUC was observed for 15 consecutive epochs (defined as the patience

parameter). Consequently, the model checkpoint corresponding to the highest validation AUC was saved as the final model for downstream evaluation.

Model checkpointing was performed by saving the weights corresponding to the best validation performance. In particular, the model achieving the highest validation AUC was retained for subsequent evaluation, ensuring consistency across all experiments.

5.3.3 Hyperparameter Optimization

Hyperparameters were optimized using grid search over predefined ranges. The explored hyperparameter ranges were model-specific to account for the different architectural requirements and pretraining paradigms.

The broader grid utilized for the convolutional networks (ResNets) is presented in Table 5.1, while the search space for the transformer-based architectures (RadioDINO) is detailed in Table 5.2.

Table 5.1: Hyperparameter search space for the ResNet models.

Hyperparameter	**Search Space**
Learning rate	$\{10^{-2}, 10^{-3}, 10^{-4}\}$
Weight decay	$\{10^{-2}, 10^{-3}, 10^{-4}\}$
Dropout rate	$\{0.1, 0.3, 0.5\}$

Table 5.2: Hyperparameter search space for the RadioDINO experiments.

Hyperparameter	**Search Space**
Learning rate	$\{10^{-4}, 10^{-5}\}$
Weight decay	$\{0.05, 0.01\}$
Dropout rate	$\{0.1, 0.3\}$

5.3.4 Gradual Unfreezing Strategy

For pre-trained models, a gradual unfreezing strategy was employed to stabilize training and preserve useful pre-trained representations. Instead of fine-tuning all layers simultaneously, model parameters were progressively unfrozen according to a predefined epoch schedule:

- **Warm-up phase:** Only the classification head is trained;

- **Intermediate phases:** Higher-level layers are progressively unfrozen;
- **Final phase:** The entire network is fine-tuned end-to-end.

For convolutional architectures such as ResNet, this corresponds to sequentially unfreezing deeper residual blocks (layer4 $\rightarrow$ layer3 $\rightarrow$ layer2). For transformer-based models like RadioDINO, unfreezing is performed progressively from the top transformer blocks downwards.

Whenever the set of trainable parameters changed during phase transitions, the optimizer was re-instantiated to accurately register the new set of active weights and their respective momentum statistics. Models trained from scratch were exempt from this procedure, as all parameters were fully optimized from the beginning.

5.3.5 LUNA25: Cross-Validation Strategy

For the LUNA25 dataset, model optimization was performed using stratified group 3-fold cross-validation. This approach ensures that multiple nodules originating from the same patient are not split across training and validation sets, preventing data leakage.

The dataset was initially divided into training/validation and held-out test subsets. The training/validation portion was further split into internal folds using `StratifiedGroupKFold`, preserving both class distribution and patient-level grouping.

For each hyperparameter configuration:

- Models were trained independently on each fold;
- Validation performance (AUC) was computed for each fold;
- The mean validation AUC across folds was used to select the optimal hyperparameters.

Following hyperparameter selection, models trained within the cross-validation procedure were independently evaluated on the held-out test set. Specifically, each of the K models obtained from the cross-validation folds was used to perform inference on the test cohort. Predictions were aggregated at patient level using the Pmax strategy described in Section 5.4.

Final test performance was reported as the mean and standard deviation of the evaluation metrics across the K independently trained fold models. This protocol provides a robust estimate of model generalization under variability induced by data splitting.

5.3.6 DLCS Dataset: Cross-Validation Strategy

Although the DLCS dataset provides a predefined benchmark split (training, validation, and test sets), model optimization and hyperparameter tuning were conducted using the exact same 3-fold stratified group cross-validation strategy previously described for the LUNA25 dataset. The combined training and validation subsets were restructured for this cross-validation process, while the predefined test set was kept strictly isolated and used solely for final independent evaluation.

Following hyperparameter selection, the $K = 3$ models obtained from the cross-validation were retained and independently evaluated on the held-out test set. To ensure complete methodological consistency with the LUNA25 evaluation protocol, final test performance was reported as the mean and standard deviation of the metrics across these three independently trained fold models.

5.3.7 Combined Dataset (LUNA25 + DLCS): Cross-Validation Strategy

The LUNA25 and DLCS training and validation cohorts were merged into a single development dataset to increase sample diversity and expose the models to a broader range of acquisition settings and population characteristics during training. A 3-fold Stratified Group Cross-Validation strategy was then applied to the merged dataset, using patient-level grouping to prevent data leakage and preserve class balance across folds. Hyperparameter optimization was performed within this cross-validation procedure, and the best-performing configuration for each architecture was selected based on the mean validation AUC across folds.

5.3.8 Reproducibility

To ensure experimental reproducibility, a fixed random seed of 42 was consistently used across all experiments. This seed controlled data splitting procedures, weight initialization, and stochastic components of training, including sampling strategies and data shuffling. This setup ensures a high degree of experimental reproducibility, although minor

variations may still arise due to non-deterministic GPU operations and stochastic sampling procedures.

5.4. Evaluation Protocol and Metrics

This section describes the evaluation protocol adopted to assess model performance, generalization, and fairness. The evaluation pipeline was designed to reflect both standard machine learning practices and clinically relevant assessment criteria, including patient-level aggregation and statistical testing of subgroup disparities.

5.4.1 Prediction and Metric Computation

Model performance is evaluated using a comprehensive set of metrics computed over the full dataset, namely: AUC, Recall (Sensitivity), Specificity, F1-score, and Balanced Accuracy.

To derive these metrics, model predictions are initially obtained as logits and converted to probabilities using the sigmoid function:

$$p = \sigma(f_\theta(x)) \tag{5.7}$$

For metrics requiring discrete class assignments, binary predictions are derived using a fixed decision threshold of 0.5:

$$\hat{y} = \mathbb{1}(p > 0.5) \tag{5.8}$$

It is acknowledged that in imbalanced clinical datasets, a threshold of 0.5 may not be optimal. However, since the primary evaluation metrics, AUC and Balanced Accuracy, are either threshold-independent or explicitly correct for class imbalance, the choice of threshold does not undermine the validity of the reported results.

Due to the inherent class imbalance present in lung cancer screening Cohorts (for instance, approximately 84.3% of nodules in LUNA25 are benign), standard accuracy alone can be a misleading performance indicator, as a trivial classifier predicting the majority class would achieve high accuracy without any diagnostic utility. Consequently, AUC and Balanced Accuracy are treated as the primary evaluation metrics: AUC provides a threshold-independent measure of discriminative ability, while Balanced Accuracy accounts for class imbalance by averaging sensitivity and specificity. Furthermore, in clinical screening contexts, maximising Recall (Sensitivity) is of particular importance, as false

negatives correspond to missed malignant nodules with potentially severe clinical consequences [45].

These metrics are derived from the elements of the confusion matrix—True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)—and are defined as follows:

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{5.9}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5.10}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.11}$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{5.12}$$

All metrics are computed on validation and test sets using accumulated predictions across batches.

5.4.2 Patient-Level Aggregation (Pmax)

Since multiple nodules may exist per patient, evaluation is performed at both the nodule level and the patient level to provide a comprehensive assessment of model behaviour. To ensure temporal consistency for patients with longitudinal imaging history, only the baseline screening scan (Year 0) was selected for each patient. Consequently, all nodules evaluated for a given individual are derived exclusively from this initial time point, with follow-up scans explicitly excluded from the analysis.

Nodule-Level Evaluation. At the most granular level, each nodule is treated as an independent instance. Standard classification metrics, including AUC, precision, recall, specificity, F1-score, and balanced accuracy, are computed using individual nodule predictions p_i and corresponding labels y_i. This evaluation reflects the intrinsic ability of the model to discriminate malignant from benign nodules without aggregation effects.

Patient-Level Aggregation (Pmax). For clinical relevance, patient-level predictions are derived using a maximum probability aggregation strategy (Pmax), under the assumption that the presence of at least one malignant nodule is sufficient for a positive diagnosis. Formally:

$$p_{\text{patient}} = \max_{i \in \mathcal{N}_p} p_i \tag{5.13}$$

where $\mathcal{N}_p$ denotes the set of nodules belonging to patient p.

This aggregation aligns with a realistic screening scenario, where detection of any high-risk lesion triggers further clinical investigation.

Evaluating both nodule-level and patient-level performance is essential for fairness assessment. Subgroup disparities may manifest differently across these two levels: for example, demographic groups may exhibit similar patient-level performance while differing significantly at the nodule level due to variations in lesion count, lesion characteristics, or class imbalance within patients. Therefore, fairness analysis is conducted consistently at both levels to avoid masking subgroup-specific biases introduced by the aggregation process.

5.4.3 Generalization Evaluation

To assess model robustness beyond the training distribution, both internal and external generalization were systematically evaluated. Specifically, every trained model—whether optimized exclusively on the LUNA25 domain, the DLCS domain, or the merged multi-source combination of both—was exhaustively evaluated across all three target domains: LUNA25 (internal), DLCS (internal cross-dataset), and LUNGx (external out-of-distribution). This comprehensive evaluation matrix ensures that all baseline performances, cross-dataset transfers, and severe domain shifts are fully captured and comparable.

Internal Evaluation: Each model was first assessed on the held-out test set drawn from its own training distribution (LUNA25 models on the LUNA25 test set; DLCS models on the DLCS test set). This provides the internal reference performance, denoted M_{int}, against which external performance is compared. Here, performance is reported using multiple clinically relevant metrics, including Area Under the ROC Curve (AUC), Balanced Accuracy, Sensitivity, and Specificity.

Cross-Dataset Generalization: To probe robustness under covariate shift, the models trained on one dataset were subsequently applied to the full test set of the other: LUNA25-trained models were evaluated on DLCS, and DLCS-trained models were evaluated on LUNA25. For each cross-dataset scenario, a generalization gap was computed for each metric M as:

$$\Delta_{\text{gap}}^{M} = M_{\text{int}} - M_{\text{ext}} \tag{5.14}$$

where M_{ext} denotes performance on the external target dataset. A large positive gap indicates a significant degradation in performance under distribution shift, while a gap close to zero suggests robust generalization.

Out-of-Distribution Evaluation (LUNGx): All models, those trained on LUNA25, those trained on DLCS, and, as an additional reference point, models trained jointly on the combined LUNA25 and DLCS cohorts, were further evaluated on the LUNGx dataset. As described in Section 4.3, LUNGx introduces a simultaneous and severe domain shift: it consists of diagnostic clinical scans rather than routine low-dose screening images, and it exhibits a fundamentally different class prevalence (an approximately balanced distribution of benign and malignant nodules, in stark contrast to the highly imbalanced screening cohorts). This scenario therefore constitutes the most demanding generalization test. Evaluation on this dataset reports AUC, Balanced Accuracy, Sensitivity, and Specificity to fully characterise performance under out-of-distribution conditions.

The combined LUNA25 + DLCS training condition was included specifically to provide a reference upper bound on the amount of source-domain data available prior to out-of-distribution transfer.

Aggregation across folds: For all cross-validated models (trained on either LUNA25 or DLCS), predictions from each fold model were evaluated independently on the target dataset. Final performance was reported as the mean and standard deviation across the K fold models, ensuring that the reported generalization estimates reflect variability induced by training data splits rather than a single arbitrary split.

5.4.4 Fairness and Error Evaluation Protocol

Fairness was evaluated by analysing performance disparities across demographic subgroups. As discussed in Section 3.3.2, high aggregate performance may conceal clinically

significant disparities affecting minority populations [75]. The following demographic attributes were considered:

- Gender (Male vs Female)
- Race (White vs Non-White)
- Smoking Status (Current vs Former)
- Age (grouped into ≤ 59, 60–65, and > 65)
- Education

In addition to patient demographics, the evaluation protocol was extended to audit potential biases stemming from physical acquisition parameters and scanner hardware, as these constitute critical sources of domain shift in medical imaging. The following image-related attributes were analysed:

- Scanner Manufacturer (e.g., Siemens vs. GE Medical Systems)
- Reconstruction Slice Thickness (e.g., ultra-thin ≤ 1.5 mm vs. standard 1.5–3.0 mm)
- X-Ray Tube Voltage (kVp)
- Reconstruction Kernels (Smooth vs. Sharp)

Finally, to systematically profile unanimous model failures and identify morphological vulnerabilities, the evaluation incorporated an analysis of nodule-level physical and spatial metadata. This included:

- Relative Nodule Scale (derived from bounding box volume)
- Maximum Attenuation Density (measured in Hounsfield Units, HU)
- Spatial Location within the pulmonary cavity (Z-slice index)

To ensure statistical validity, samples with missing or undefined metadata were excluded from the respective analyses. Furthermore, due to the inherent annotation heterogeneity between independent screening cohorts, specific variables (such as educational attainment or standardized radiodensity metrics) were systematically evaluated exclusively on the datasets where such metadata was natively available and harmonized.

5.4.5 Statistical Testing of Fairness

Fairness was assessed using hypothesis testing to determine whether performance differences across subgroups are statistically significant.

Sensitivity and Specificity: Differences in sensitivity and specificity across groups were evaluated using the chi-square test of independence. Contingency tables were constructed by aggregating true positives/false negatives and true negatives/false positives across all folds. Notably, evaluating disparities in Sensitivity across demographic groups corresponds to assessing Equality of Opportunity (EO), as defined in Section 3.3.2: a model satisfying EO must assign equal true positive rates across groups, ensuring that patients with malignant nodules are not systematically under-diagnosed based on demographic characteristics [68, 75].

AUC and Balanced Accuracy: For AUC and balanced accuracy, statistical significance of observed performance disparities across demographic groups was assessed using a non-parametric permutation test (also referred to as a randomization test) [96, 97].

This method evaluates the null hypothesis (H_0) that model performance is independent of demographic group membership. Under H_0, group labels are exchangeable, meaning that any observed performance gap is attributable to random variation rather than systematic bias.

Let Δ denote the observed performance gap:

$$\Delta = \max_g M_g - \min_g M_g \tag{5.15}$$

where M_g represents the metric value (AUC or balanced accuracy) for group g.

To approximate the null distribution, a Monte Carlo permutation procedure is performed:

- Group labels are randomly permuted $N = 1000$ times, breaking any association between demographic attributes and model predictions;
- For each permutation i, the performance gap $\Delta^{(i)}$ is recomputed;
- The empirical p-value is defined as:

$$p = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\Delta^{(i)} \geq \Delta) \tag{5.16}$$

where $\mathbb{1}(\cdot)$ is the indicator function. A p-value below 0.05 indicates that the observed disparity is unlikely under the null hypothesis, suggesting statistically significant performance differences across demographic groups.

6. Results and Discussion

This chapter presents the empirical results obtained from the experimental framework and discusses their clinical and technical implications. The analysis is structured to progressively evaluate model performance from internal baseline validation to complex out-of-distribution generalization. Furthermore, demographic fairness is assessed to determine whether performance remains equitable across diverse patient subgroups.

6.1. Internal Evaluation

This section analyses the internal, in-distribution performance of the evaluated architectures using held-out test splits derived from their respective training datasets (LUNA25 and DLCS). The analysis initially investigates architectural behaviour and the influence of ImageNet pre-training at the nodule level. Subsequently, it examines the impact of patient-level aggregation (Pmax) and the underlying dataset characteristics on overall classification performance.

6.1.1 ResNet Models

Tables 6.1 and 6.2 report the internal test set performance of the six ResNet configurations on the LUNA25 and DLCS datasets, respectively. Results are presented at both the nodule level and the patient level following Pmax aggregation, and are expressed as the mean and standard deviation across the cross-validation folds.

Table 6.1: Internal test set performance of ResNet-based models on LUNA25. Values are reported as mean ± standard deviation across cross-validation folds. **Bold** denotes the best value per column.

	Nodule-Level						Patient-Level (Pmax)					
Model	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**
ResNet-18 (Scratch)	0.827 ± 0.013	0.671 ± 0.123	0.808 ± 0.061	0.740 ± 0.031	0.392	0.443 ± 0.002	0.792 ± 0.015	0.671 ± 0.128	0.762 ± 0.074	0.717 ± 0.027	0.343	0.463 ± 0.009
ResNet-18 (ImageNet)	**0.842** ± 0.007	0.648 ± 0.120	**0.865** ± 0.046	0.756 ± 0.041	**0.470**	**0.495** ± 0.033	**0.814** ± 0.011	0.647 ± 0.115	**0.827** ± 0.054	0.737 ± 0.037	**0.409**	**0.509** ± 0.037
ResNet-34 (Scratch)	0.830 ± 0.001	0.671 ± 0.042	0.836 ± 0.020	0.754 ± 0.011	0.431	0.473 ± 0.006	0.800 ± 0.004	0.671 ± 0.038	0.794 ± 0.023	0.733 ± 0.010	0.376	0.492 ± 0.009
ResNet-34 (ImageNet)	0.822 ± 0.026	0.729 ± 0.062	0.809 ± 0.034	**0.769** ± 0.017	0.414	0.473 ± 0.018	0.791 ± 0.032	0.734 ± 0.056	0.761 ± 0.040	**0.747** ± 0.014	0.362	0.496 ± 0.018
ResNet-50 (Scratch)	0.798 ± 0.005	0.662 ± 0.038	0.787 ± 0.009	0.725 ± 0.021	0.365	0.417 ± 0.022	0.762 ± 0.007	0.657 ± 0.038	0.737 ± 0.011	0.697 ± 0.017	0.316	0.436 ± 0.017
ResNet-50 (ImageNet)	0.821 ± 0.010	**0.762** ± 0.038	0.732 ± 0.055	0.747 ± 0.009	0.344	0.419 ± 0.027	0.789 ± 0.013	**0.758** ± 0.038	0.674 ± 0.059	0.716 ± 0.011	0.301	0.444 ± 0.021

Table 6.2: Internal test set performance of ResNet-based models on DLCS. Values are reported as mean ± standard deviation across cross-validation folds. **Bold** denotes the best value per column.

	Nodule-Level						Patient-Level (Pmax)					
Model	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**
ResNet-18 (Scratch)	0.626 ± 0.043	0.323 ± 0.186	0.816 ± 0.126	0.570 ± 0.032	0.149	0.212 ± 0.060	0.713 ± 0.022	0.463 ± 0.228	0.772 ± 0.145	0.618 ± 0.044	0.169	0.238 ± 0.033
ResNet-18 (ImageNet)	0.639 ± 0.018	0.222 ± 0.076	0.902 ± 0.047	0.562 ± 0.016	0.185	0.216 ± 0.035	0.648 ± 0.046	0.333 ± 0.120	0.865 ± 0.064	0.599 ± 0.033	**0.198**	0.242 ± 0.043
ResNet-34 (Scratch)	0.627 ± 0.027	**0.424** ± 0.171	0.710 ± 0.218	0.567 ± 0.035	0.128	0.247 ± 0.057	0.700 ± 0.057	**0.593** ± 0.131	0.650 ± 0.229	0.621 ± 0.059	0.145	0.272 ± 0.098
ResNet-34 (ImageNet)	0.660 ± 0.021	0.364 ± 0.236	0.780 ± 0.187	0.572 ± 0.029	0.142	0.225 ± 0.034	0.699 ± 0.020	0.444 ± 0.236	0.726 ± 0.212	0.585 ± 0.018	0.140	0.218 ± 0.023
ResNet-50 (Scratch)	0.630 ± 0.002	0.354 ± 0.076	0.819 ± 0.068	**0.586** ± 0.013	0.164	**0.255** ± 0.017	**0.733** ± 0.027	0.574 ± 0.069	0.756 ± 0.085	**0.665** ± 0.015	0.191	**0.294** ± 0.041
ResNet-50 (ImageNet)	**0.688** ± 0.065	0.202 ± 0.087	**0.920** ± 0.027	0.561 ± 0.030	**0.202**	0.210 ± 0.064	0.718 ± 0.058	0.315 ± 0.114	**0.885** ± 0.036	0.600 ± 0.039	0.215	0.247 ± 0.050

6.1.1.1 Nodule-Level Performance

Across both datasets, the ResNet architectures demonstrated markedly different operating characteristics depending on model depth and the use of ImageNet pre-training.

On LUNA25, all ResNet configurations achieved relatively strong discriminative performance, with nodule-level AUC values ranging from 0.798 to 0.842 (Table 6.1). Among the evaluated configurations, ResNet-18 with ImageNet pre-training achieved the highest nodule-level AUC (0.842). In contrast, the ResNet-50 configurations produced the lowest overall AUC values, particularly the scratch-trained variant (0.798). Overall, ImageNet pre-training produced modest but relatively consistent improvements across the ResNet family on the LUNA25 cohort.

Despite the relatively similar AUC values, pretrained and scratch-trained models frequently exhibited different operating profiles. In general, pretrained configurations tended to favour specificity, whereas scratch-trained models often produced a more equitable balance between sensitivity and specificity. For example, ResNet-18 (ImageNet) achieved the highest nodule-level specificity (0.865), while ResNet-50 (ImageNet) achieved the highest sensitivity (0.762).

ResNet-34 (ImageNet) achieved the highest nodule-level balanced accuracy (0.769), indicating the most balanced classification behaviour among the evaluated ResNet configurations. Conversely, ResNet-50 (Scratch) produced the weakest overall nodule-level performance, with an AUC of 0.798 and a balanced accuracy of 0.725.

In contrast, the DLCS experiments produced substantially lower performance across all

ResNet configurations compared with the LUNA25 results (Table 6.2). Nodule-level AUC values ranged from 0.626 to 0.688.

Unlike the behaviour observed on LUNA25, ImageNet pre-training generally produced larger improvements on DLCS, particularly for deeper architectures. ResNet-50 (ImageNet) achieved the highest nodule-level AUC (0.688), outperforming the scratch-trained ResNet-50 configuration by approximately 0.058 AUC points. Similar improvements were observed for ResNet-34.

However, evaluating these diverse metrics reveals a severe vulnerability driven by the extreme class imbalance of the DLCS cohort (8.9% malignancy prevalence). The pretrained configurations frequently demonstrated a critical collapse in sensitivity. ResNet-50 (ImageNet), despite achieving the top nodule-level AUC of 0.688, generated a highly skewed operating profile with a sensitivity of only 0.202 and a specificity of 0.920. A similar diagnostic polarization was observed for ResNet-18 (ImageNet), indicating that these models maximized AUC by overwhelmingly predicting the majority benign class.

Among the scratch-trained models, ResNet-50 (Scratch) achieved the highest nodule-level balanced accuracy (0.586), indicating a comparatively more balanced operating profile.

Overall, the ResNet experiments suggest that the effect of ImageNet pre-training was highly dependent on the training cohort. While pre-training provided only marginal gains on LUNA25, it produced clearer AUC improvements on DLCS, although systematically at the cost of reduced clinical sensitivity. This macro-level performance divergence between the two datasets is visually synthesized in Figure 6.1.

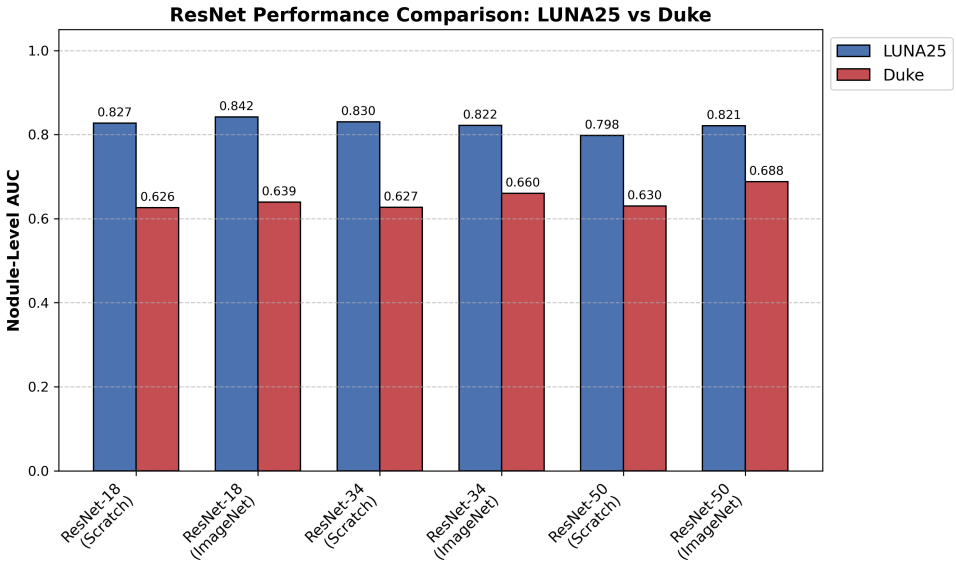


Figure 6.1: Comparison of nodule-level AUC performance between LUNA25 and DLCS across all evaluated ResNet architectures. The systemic performance degradation on the DLCS cohort illustrates the severe impact of extreme class imbalance and data heterogeneity.

6.1.1.2 Patient-Level Performance

At the patient level, predictions were derived using Pmax aggregation—where the highest malignancy probability among all nodules within a single patient dictates the overall patient diagnosis, as outlined in Section 5.4.

Comparing performance between the nodule and patient levels provides critical insights into whether patient-level aggregation can compensate for individual nodule misclassifications. Theoretically, Pmax aggregation can rescue nodule-level false negatives: if a patient possesses multiple nodules and the model successfully detects at least one malignant lesion, the patient is still correctly flagged for follow-up. However, this same mechanism makes the model more susceptible to false positives, as a single highly scored benign nodule will falsely flag the entire patient.

This dynamic manifested differently depending on the training dataset, a behaviour that can be directly attributed to the distinct cohort characteristics and nodule multiplicity outlined in Chapter 4. Statistically, the LUNA25 dataset contains 6,166 annotated nodules across 2,120 patients, yielding an average of approximately 2.9 nodules per patient. In contrast, the DLCS dataset contains 2,483 nodules across 1,611 patients, averaging roughly 1.5 nodules per patient. In LUNA25, the higher multiplicity of nodules per patient statistically increases the probability that at least one benign nodule is falsely classified as malignant. This exacerbates the false-positive penalty inherent to Pmax aggregation,

which explains the slight reductions in overall AUC for several models, as nodule-level performance was already robust and had little room for false-negative rescue.

Conversely, in the DLCS experiments, the lower average number of nodules per patient inherently reduced this compounding false-positive effect. Furthermore, as established in Section 4.2, the DLCS cohort is characterized by higher clinical and demographic heterogeneity, including a broader age range and a significantly larger representation of minority racial groups (26.1% Non-White, compared to 6.6% in LUNA25). This diversity inherently increased the difficulty of the baseline nodule-level classification, leading to generally poorer nodule-level sensitivity. Consequently, patient-level aggregation in DLCS successfully functioned as a rescue mechanism for individual missed detections without being overwhelmingly penalized by multiple false positives, substantially improving sensitivity and overall AUC. For instance, ResNet-50 (Scratch) increased its balanced accuracy from 0.586 at the nodule level to 0.665 at the patient level.

To visually contextualize these dynamics, Figures 6.2 and 6.3 illustrate the side-by-side comparison of the AUC metrics before and after Pmax aggregation for the LUNA25 and DLCS cohorts, respectively.

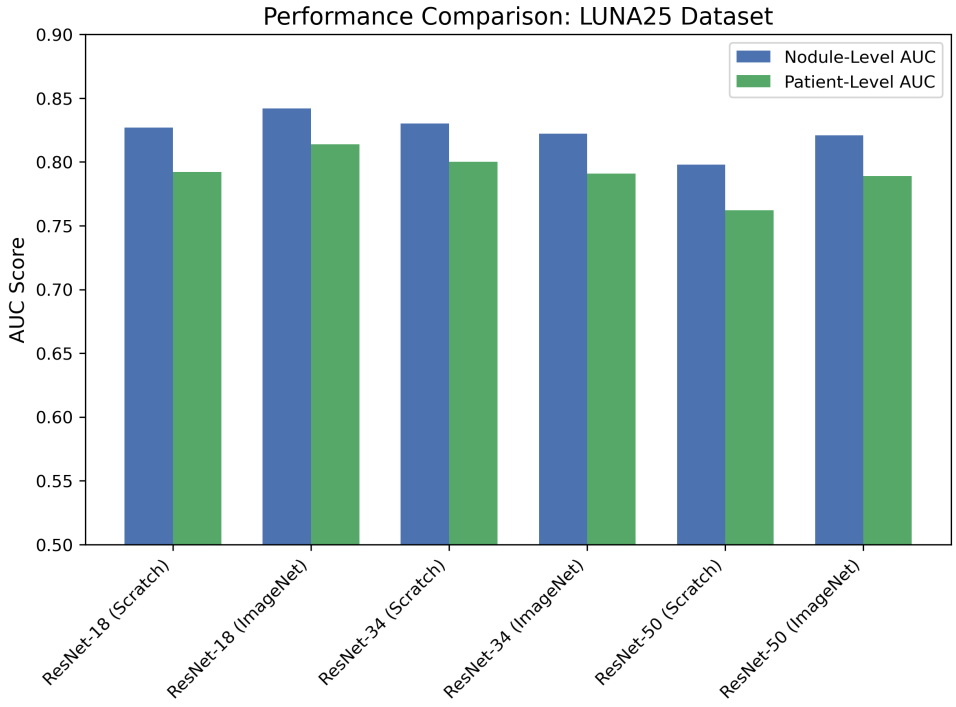


Figure 6.2: Comparison of Nodule vs. Patient-level AUC across ResNet architectures on the LUNA25 dataset. The high nodule multiplicity inherently penalizes the patient-level AUC due to the compounding risk of false positives.

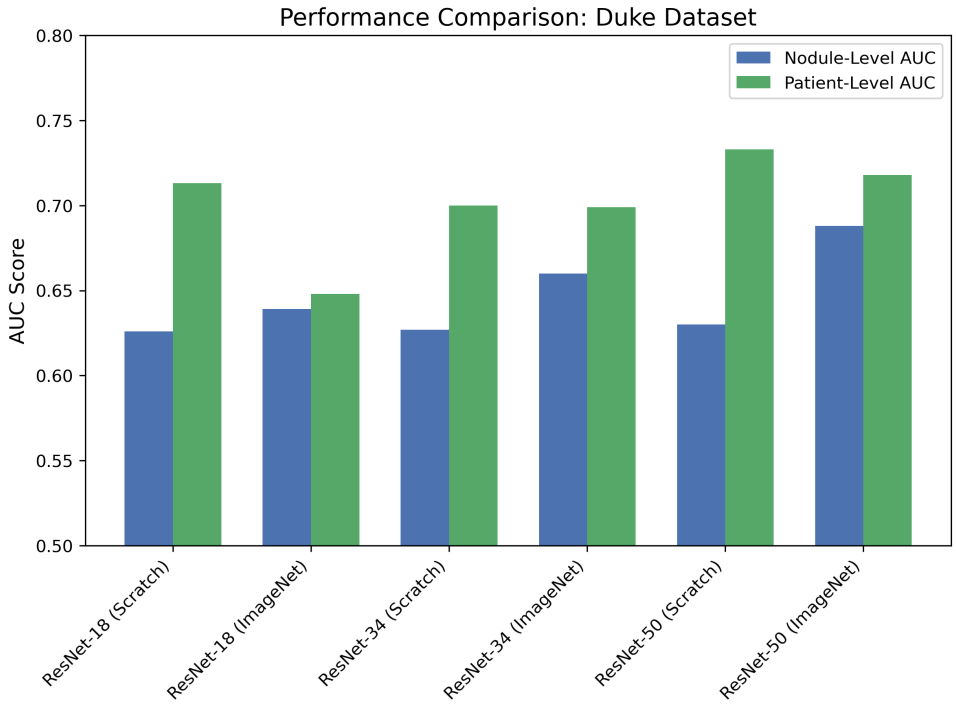


Figure 6.3: Comparison of Nodule vs. Patient-level AUC across ResNet architectures on the DLCS dataset. Pmax aggregation successfully acts as a rescue mechanism for individual missed detections, substantially improving the overall patient-level AUC.

Regarding specific model rankings, for the LUNA25-trained models, ResNet-18 (ImageNet) achieved the highest patient-level AUC (0.814), whereas ResNet-50 (ImageNet) achieved the highest patient-level sensitivity (0.758), although with comparatively reduced specificity (0.674).

In the DLCS experiments, ResNet-50 (Scratch) achieved the highest patient-level AUC (0.733) and balanced accuracy (0.665), outperforming the pretrained ResNet-50 configuration despite its lower nodule-level AUC. Compared with the LUNA25 experiments, the DLCS results also exhibited larger standard deviations across several evaluation metrics, indicating increased variability across folds. This behaviour likely reflects the smaller cohort size and higher heterogeneity of the DLCS dataset, which were previously discussed in Chapter 4.

Overall, the transition from nodule to patient-level evaluation demonstrated that while aggregation can effectively mitigate poor nodule-level sensitivity (as seen in DLCS), improvements in patient-level AUC do not necessarily correspond to improved class balance, as several pretrained configurations continued to favour specificity over sensitivity.

6.1.2 DINO Models

Tables 6.3 and 6.4 summarise the internal test set performance of the three RadioDINO configurations on LUNA25 and DLCS, respectively. As with the ResNet experiments, metrics are reported at both the nodule level and the patient level.

Table 6.3: Internal test set performance of RadioDINO-based models on LUNA25. Values are reported as mean ± standard deviation across cross-validation folds. Precision is reported as a derived point estimate based on mean cohort prevalence, mean sensitivity, and mean specificity. **Bold** denotes the best value per column.

	Nodule-Level						Patient-Level (Pmax)					
Model	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**
RadioDINO-S8	0.846 ± 0.005	0.710 ± 0.049	0.789 ± 0.026	0.749 ± 0.012	0.383	0.443 ± 0.005	0.820 ± 0.007	0.715 ± 0.048	0.740 ± 0.029	0.727 ± 0.010	0.337	0.468 ± 0.001
RadioDINO-Small	0.861 ± 0.004	**0.733** ± 0.007	0.812 ± 0.024	**0.773** ± 0.009	0.419	0.480 ± 0.024	**0.835** ± 0.004	**0.729** ± 0.007	0.769 ± 0.028	**0.749** ± 0.012	0.368	**0.501** ± 0.023
RadioDINO-Base	**0.862** ± 0.010	0.567 ± 0.099	**0.893** ± 0.038	0.730 ± 0.031	**0.495**	**0.485** ± 0.009	0.833 ± 0.011	0.570 ± 0.096	**0.859** ± 0.050	0.715 ± 0.023	**0.430**	0.496 ± 0.008

Table 6.4: Internal test set performance of RadioDINO-based models on DLCS. Values are reported as mean ± standard deviation across cross-validation folds. Precision is reported as a derived point estimate based on mean cohort prevalence, mean sensitivity, and mean specificity. **Bold** denotes the best value per column.

	Nodule-Level						Patient-Level (Pmax)					
Model	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**
RadioDINO-S8	0.645 ± 0.058	0.323 ± 0.287	0.773 ± 0.175	0.548 ± 0.057	0.125	0.171 ± 0.070	0.693 ± 0.069	0.426 ± 0.292	0.728 ± 0.189	0.577 ± 0.057	0.135	0.191 ± 0.047
RadioDINO-Small	0.636 ± 0.043	0.313 ± 0.100	**0.874** ± 0.046	0.593 ± 0.027	**0.201**	0.265 ± 0.029	0.695 ± 0.025	**0.463** ± 0.114	**0.830** ± 0.053	**0.646** ± 0.036	**0.213**	**0.291** ± 0.034
RadioDINO-Base	**0.659** ± 0.032	**0.333** ± 0.086	0.867 ± 0.061	**0.600** ± 0.012	0.200	**0.278** ± 0.007	**0.714** ± 0.064	0.444 ± 0.079	0.820 ± 0.080	0.632 ± 0.001	0.198	0.284 ± 0.035

6.1.2.1 Nodule-Level Performance

The RadioDINO architectures achieved strong overall nodule-level performance on LUNA25 (Table 6.3). Nodule-level AUC values ranged from 0.846 to 0.862.

RadioDINO-Base achieved the highest nodule-level AUC (0.862) and specificity (0.893), although with a comparatively lower sensitivity (0.567). In contrast, RadioDINO-Small achieved the highest balanced accuracy (0.773) and a more balanced sensitivity-specificity trade-off, while obtaining a nearly identical AUC (0.861).

RadioDINO-S8 produced slightly lower overall performance than the other DINO configurations, although it still achieved competitive discriminative performance, with an AUC of 0.846 and a balanced accuracy of 0.749.

When compared directly to the ResNet family, the DINO architectures consistently yielded superior discriminative performance on LUNA25, as visually summarized in Figure 6.4. This advantage is particularly notable in terms of overall AUC stability and balanced accuracy. These results align with previous studies discussed in Chapter 3, where self-supervised vision transformer approaches demonstrated improved representation learning capabilities for medical imaging tasks.

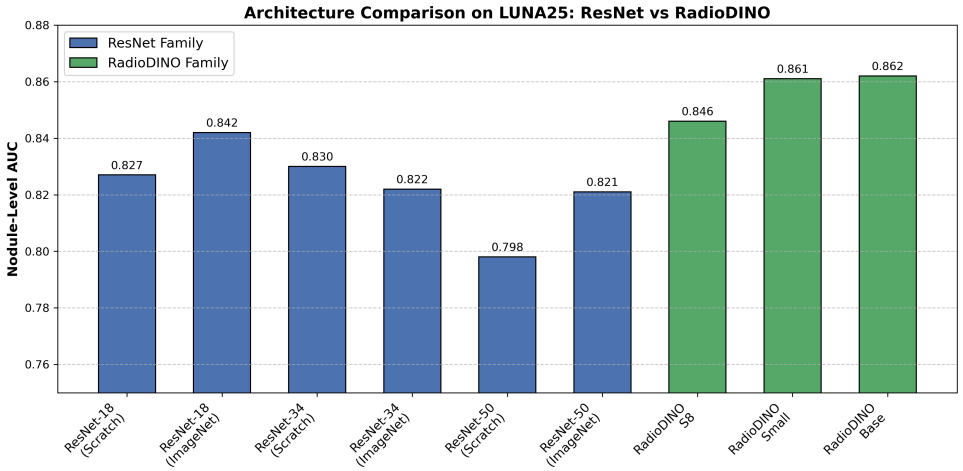


Figure 6.4: Comparison of nodule-level AUC between the ResNet and RadioDINO families on the LUNA25 internal test set. The self-supervised vision transformers consistently outperform the convolutional architectures, achieving higher and more stable discriminative performance.

On the DLCS dataset, however, all RadioDINO configurations demonstrated substantially lower performance compared with their corresponding LUNA25 results (Table 6.4). Nodule-level AUC values ranged from 0.636 to 0.659.

Among the evaluated configurations, RadioDINO-Base achieved the strongest overall DLCS performance, obtaining the highest nodule-level AUC (0.659), balanced accuracy (0.600), and F1-score (0.278).

Similarly to the ResNet experiments, the DINO models frequently demonstrated specificity-oriented behaviour. For example, RadioDINO-Small achieved a specificity of 0.874 while producing a sensitivity of only 0.313. RadioDINO-Base exhibited a similar operating profile.

The DLCS experiments also demonstrated substantially higher variability across folds compared with LUNA25, particularly for sensitivity metrics. This behaviour was especially noticeable for RadioDINO-S8, which showed considerable inter-fold variability.

This loss of performance advantage over the convolutional models on the DLCS cohort is visually summarized in Figure 6.5.

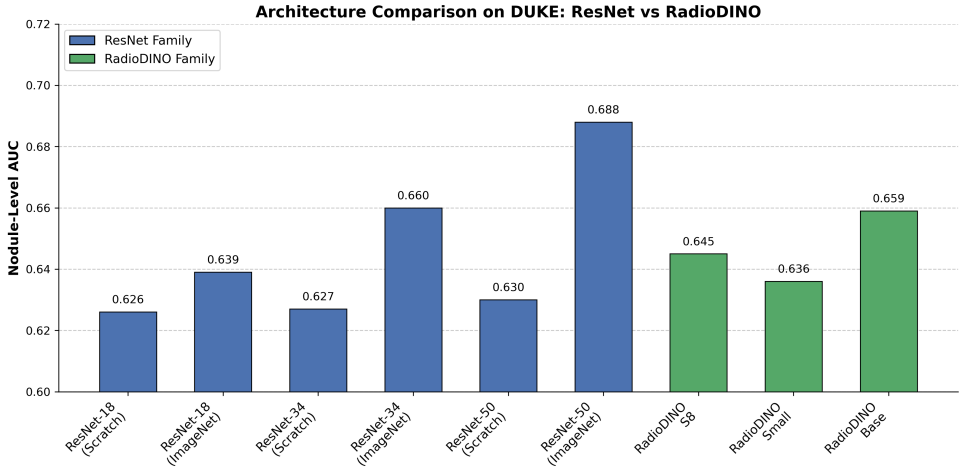


Figure 6.5: Comparison of nodule-level AUC between the ResNet and RadioDINO families on the DLCS internal test set. Unlike the LUNA25 results, the self-supervised vision transformers do not exhibit a clear discriminative advantage over the convolutional architectures, with both families struggling to maintain robust performance.

6.1.2.2 Patient-Level Performance

At the patient level, the DINO models generally preserved the same relative ranking observed at the nodule level.

For the LUNA25 experiments, RadioDINO-Small achieved the highest patient-level balanced accuracy (0.749), while RadioDINO-Base and RadioDINO-Small achieved nearly identical patient-level AUC values (0.833 and 0.835, respectively). RadioDINO-Base again demonstrated a specificity-oriented operating profile, achieving a specificity of 0.859 but a lower sensitivity of 0.570.

For the DLCS cohort, RadioDINO-Base achieved the highest patient-level AUC (0.714) and RadioDINO-Small achieved the highest balanced accuracy (0.646). However, similarly to the nodule-level experiments, all DINO configurations demonstrated relatively modest sensitivity values compared with specificity.

Overall, unlike the behaviour observed on LUNA25, the DINO architectures no longer demonstrated a clear performance advantage over the ResNet family on DLCS. While RadioDINO-Small achieved the strongest overall performance on LUNA25, the best DINO configuration on DLCS remained below the best-performing ResNet configuration in terms

of nodule-level AUC. This performance shift can be largely attributed to the substantial difference in dataset sizes. Vision Transformers typically require larger amounts of training data to learn effective representations compared to CNNs, which benefit from built-in inductive biases [65]. Since the DLCS dataset is significantly smaller than LUNA25 (2,483 versus 6,166 nodules), it likely provides insufficient volume for the DINO architectures to fully optimize, allowing the ResNets to maintain a comparative advantage in this data-constrained regime.

6.1.3 Impact of the Training Dataset

The previous subsections focused on architectural behaviour within each dataset. This section instead analyses the impact of the training dataset itself, comparing how the best-performing models behaved at both the nodule and patient levels after Pmax aggregation.

To facilitate this comparison, Table 6.5 summarises the best-performing model from each architecture family for each training dataset.

Table 6.5: Summary of the best-performing model per architecture family and training dataset at the nodule and patient levels. Models are selected based on the highest nodule-level Balanced Accuracy. Δ_{AUC} denotes the difference in AUC between the patient level and nodule level (positive values indicate improved patient-level performance after Pmax aggregation).

		Nodule-Level				Patient-Level (Pmax)				
Train	**Model**	**AUC**	**Sens.**	**Spec.**	**BAcc**	**AUC**	**Sens.**	**Spec.**	**BAcc**	Δ_{AUC}
LUNA25	ResNet-34 (ImageNet)	0.822 ± 0.026	0.729 ± 0.062	0.809 ± 0.034	0.769 ± 0.017	0.791 ± 0.032	0.734 ± 0.056	0.761 ± 0.040	0.747 ± 0.014	-0.031
	RadioDINO-Small	0.861 ± 0.004	0.733 ± 0.007	0.812 ± 0.024	0.773 ± 0.009	0.835 ± 0.004	0.729 ± 0.007	0.769 ± 0.028	0.749 ± 0.012	-0.027
DLCS	ResNet-50 (Scratch)	0.630 ± 0.002	0.354 ± 0.076	0.819 ± 0.068	0.586 ± 0.013	0.733 ± 0.027	0.574 ± 0.069	0.756 ± 0.085	0.665 ± 0.015	$+0.103$
	RadioDINO-Base	0.659 ± 0.032	0.333 ± 0.086	0.867 ± 0.061	0.600 ± 0.012	0.714 ± 0.064	0.444 ± 0.079	0.820 ± 0.080	0.632 ± 0.001	$+0.055$

6.1.3.1 LUNA25-Trained Models

Models trained on LUNA25 achieved the strongest overall performance across both architecture families. At the nodule level, the best ResNet configuration (ResNet-34, ImageNet) achieved an AUC of 0.822 and a balanced accuracy of 0.769, while RadioDINO-Small achieved the highest overall balanced accuracy (0.773) with an AUC of 0.861.

These results suggest that the LUNA25 cohort provided a comparatively more favourable training environment for deep learning models. As discussed in Chapter 4, LUNA25 is

substantially larger than DLCS and presents more standardised acquisition characteristics, which may have contributed to improved optimisation stability and representation learning.

When predictions were aggregated at the patient level using Pmax, all LUNA25-trained models showed a small reduction in AUC. ResNet-34 (ImageNet) decreased from 0.822 to 0.791 ($\Delta_{\text{AUC}} = -0.031$), while RadioDINO-Small decreased from 0.861 to 0.835 ($\Delta_{\text{AUC}} = -0.027$).

This reduction was accompanied by relatively small shifts in sensitivity and specificity. For example, RadioDINO-Small showed a slight reduction in sensitivity from 0.733 to 0.729, while specificity decreased from 0.812 to 0.769. Similar behaviour was observed for the ResNet configuration.

Overall, despite the slight reduction in AUC after aggregation, patient-level performance remained strong for the LUNA25-trained models, and the relative ranking between architectures was preserved, with RadioDINO-Small remaining the best-performing configuration at both evaluation levels.

6.1.3.2 DLCS-Trained Models

In contrast to the LUNA25 experiments, models trained on DLCS achieved substantially lower performance across all evaluation metrics. At the nodule level, the best ResNet configuration (ResNet-50, Scratch) achieved an AUC of 0.630 and a balanced accuracy of 0.586, while RadioDINO-Base achieved an AUC of 0.659 and a balanced accuracy of 0.600.

The lower overall performance observed for DLCS-trained models may be associated with several factors discussed previously in Chapter 4 such as the smaller cohort size and higher variability in imaging characteristics and acquisition protocols.

Unlike the behaviour observed for the LUNA25-trained models, Pmax aggregation improved patient-level performance for the DLCS-trained configurations. ResNet-50 (Scratch) increased from a nodule-level AUC of 0.630 to a patient-level AUC of 0.733 ($\Delta_{\text{AUC}} = +0.103$), while RadioDINO-Base increased from 0.659 to 0.714 ($\Delta_{\text{AUC}} = +0.055$).

Sensitivity and balanced accuracy also improved after aggregation. For example, ResNet-50 (Scratch) increased balanced accuracy from 0.586 to 0.665 and sensitivity from 0.354 to 0.574.

These results indicate that the effect of patient-level aggregation differed substantially between datasets. While aggregation slightly reduced performance in the LUNA25 experiments, it improved performance in the DLCS cohort, particularly in terms of sensitivity and balanced accuracy.

One possible explanation is that the DLCS-trained models produced less stable nodule-level predictions, making patient-level aggregation beneficial by reducing the impact of individual false-negative nodules. In contrast, the stronger nodule-level performance observed on LUNA25 may have limited the potential benefit of Pmax aggregation.

6.1.3.3 Jointly-Trained Models (LUNA25 + DLCS)

To further investigate the influence of dataset composition on model behaviour, an additional set of experiments was conducted using jointly-trained cohorts combining the LUNA25 and DLCS datasets. Unlike the previous sections, where models were trained and evaluated within a single dataset, these experiments assessed how the combined training strategy affected performance when evaluated independently on the LUNA25 and DLCS internal test sets.

Tables 6.6 and 6.7 summarise the performance of all jointly-trained configurations on the LUNA25 and DLCS cohorts, respectively.

Table 6.6: Performance of jointly-trained models (LUNA25 + DLCS) evaluated on the **LUNA25** internal test set. Results are reported as mean ± standard deviation across the 3 cross-validation folds. Precision and F1-score are reported as derived point estimates based on mean cohort prevalence, mean sensitivity, and mean specificity. **Bold** denotes the best value per column.

	Nodule-Level						Patient-Level (Pmax)					
Model	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**
ResNet-18 (Scratch)	0.800 ± 0.014	0.638 ± 0.095	0.810 ± 0.054	0.724 ± 0.025	0.383	0.479	0.763 ± 0.015	0.638 ± 0.092	0.756 ± 0.063	0.697 ± 0.022	0.327	0.432
ResNet-18 (ImageNet)	0.832 ± 0.002	0.771 ± 0.121	0.705 ± 0.191	0.738 ± 0.035	0.326	0.458	0.803 ± 0.002	0.768 ± 0.123	0.650 ± 0.209	0.709 ± 0.043	0.288	0.419
ResNet-34 (Scratch)	0.815 ± 0.003	0.657 ± 0.023	0.801 ± 0.014	0.729 ± 0.007	0.379	0.481	0.783 ± 0.005	0.652 ± 0.024	0.754 ± 0.014	0.703 ± 0.006	0.329	0.437
ResNet-34 (ImageNet)	0.822 ± 0.011	0.724 ± 0.018	0.779 ± 0.017	0.751 ± 0.007	0.378	0.497	0.788 ± 0.013	0.720 ± 0.018	0.718 ± 0.021	0.719 ± 0.007	0.321	0.444
ResNet-50 (Scratch)	0.809 ± 0.013	0.738 ± 0.047	0.751 ± 0.037	0.744 ± 0.027	0.354	0.478	0.775 ± 0.017	0.734 ± 0.048	0.690 ± 0.040	0.712 ± 0.032	0.305	0.431
ResNet-50 (ImageNet)	0.833 ± 0.008	**0.781** ± 0.097	0.673 ± 0.153	0.727 ± 0.030	0.306	0.440	0.798 ± 0.007	**0.778** ± 0.099	0.613 ± 0.161	0.696 ± 0.033	0.271	0.402
RadioDINO-Small	0.849 ± 0.010	0.752 ± 0.029	0.785 ± 0.014	0.769 ± 0.011	0.393	**0.516**	0.821 ± 0.011	0.759 ± 0.030	0.732 ± 0.017	0.745 ± 0.009	0.344	0.473
RadioDINO-S8	0.846 ± 0.005	0.667 ± 0.018	**0.824** ± 0.016	0.745 ± 0.002	**0.412**	0.509	0.815 ± 0.007	0.676 ± 0.018	**0.771** ± 0.020	0.724 ± 0.001	**0.354**	0.465
RadioDINO-Base	**0.863** ± 0.008	0.743 ± 0.070	0.796 ± 0.016	**0.770** ± 0.027	0.402	0.522	**0.837** ± 0.011	0.744 ± 0.065	0.748 ± 0.018	**0.746** ± 0.024	0.353	**0.479**

Table 6.7: Performance of jointly-trained models (LUNA25 + DLCS) evaluated on the **DLCS** internal test set. Results are reported as mean ± standard deviation across the 3 cross-validation folds. Precision and F1-score are reported as derived point estimates based on mean cohort prevalence, mean sensitivity, and mean specificity. **Bold** denotes the best value per column.

	Nodule-Level						Patient-Level (Pmax)					
Model	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**	**AUC**	**Sens.**	**Spec.**	**BAcc**	**Prec.**	**F1**
ResNet-18 (Scratch)	0.632 ± 0.028	0.354 ± 0.136	**0.829** ± 0.066	0.591 ± 0.036	0.171	0.231	0.709 ± 0.031	0.500 ± 0.120	**0.778** ± 0.071	0.639 ± 0.027	0.184	0.269
ResNet-18 (ImageNet)	0.672 ± 0.046	0.475 ± 0.200	0.678 ± 0.249	0.577 ± 0.025	0.129	0.203	0.718 ± 0.024	0.667 ± 0.164	0.624 ± 0.265	0.645 ± 0.058	0.150	0.245
ResNet-34 (Scratch)	0.685 ± 0.036	0.434 ± 0.087	0.811 ± 0.026	0.623 ± 0.036	0.187	0.261	0.743 ± 0.030	0.574 ± 0.052	0.757 ± 0.030	0.666 ± 0.011	0.191	0.287
ResNet-34 (ImageNet)	0.685 ± 0.053	0.465 ± 0.080	0.793 ± 0.011	0.629 ± 0.045	0.183	0.263	0.721 ± 0.014	0.630 ± 0.026	0.722 ± 0.012	0.676 ± 0.018	0.185	0.286
ResNet-50 (Scratch)	0.684 ± 0.041	0.455 ± 0.066	0.800 ± 0.022	0.627 ± 0.023	0.186	0.264	0.741 ± 0.027	0.593 ± 0.026	0.735 ± 0.030	0.664 ± 0.010	0.183	0.280
ResNet-50 (ImageNet)	0.704 ± 0.033	**0.636** ± 0.089	0.673 ± 0.106	0.655 ± 0.024	0.163	0.259	0.757 ± 0.024	**0.815** ± 0.069	0.594 ± 0.109	0.705 ± 0.044	0.167	0.277
RadioDINO-Small	0.664 ± 0.012	0.556 ± 0.014	0.754 ± 0.027	0.655 ± 0.015	0.184	0.277	0.742 ± 0.007	0.722 ± 0.045	0.674 ± 0.032	0.698 ± 0.021	0.181	0.290
RadioDINO-S8	0.712 ± 0.005	0.485 ± 0.025	0.801 ± 0.027	0.643 ± 0.002	0.196	0.279	0.731 ± 0.022	0.593 ± 0.026	0.735 ± 0.032	0.664 ± 0.006	0.183	0.280
RadioDINO-Base	**0.723** ± 0.019	0.606 ± 0.025	0.757 ± 0.037	**0.682** ± 0.013	**0.199**	**0.299**	**0.775** ± 0.014	0.741 ± 0.052	0.689 ± 0.044	**0.715** ± 0.006	**0.192**	**0.305**

At the nodule level, the jointly-trained models evaluated on LUNA25 achieved substantially stronger performance than the same jointly-trained models evaluated on DLCS. This behaviour closely mirrors the trends previously observed for the single-dataset training experiments and further reinforces the strong influence of dataset characteristics on model performance.

On the LUNA25 cohort (Table 6.6), all jointly-trained configurations achieved relatively high discriminative performance. Nodule-level AUC values ranged from 0.800 to 0.863 across all evaluated models.

Among the ResNet family, the pretrained configurations again achieved the strongest overall AUC values. ResNet-50 (ImageNet) achieved the highest sensitivity among all ResNet models (0.781), while ResNet-34 (ImageNet) achieved the highest balanced accuracy (0.751), indicating the most balanced operating profile within the ResNet family.

Similarly to the previous LUNA25-only experiments, the pretrained ResNet configurations frequently favoured sensitivity at the expense of specificity. For example, ResNet-50 (ImageNet) achieved a specificity of only 0.673 despite its comparatively high sensitivity.

The jointly-trained RadioDINO configurations again outperformed the ResNet family in terms of overall discriminative performance on LUNA25. RadioDINO-Base achieved the

highest overall nodule-level AUC (0.863), balanced accuracy (0.770), and F1-score (0.522), while RadioDINO-Small achieved the highest sensitivity among the DINO models (0.752).

Compared with the DLCS-only experiments discussed previously, the jointly-trained configurations demonstrated substantially improved performance and reduced variability when evaluated on LUNA25. In particular, the jointly-trained DINO models recovered performance levels that were highly comparable to the original LUNA25-trained experiments.

At the patient level, the relative performance trends established prior to Pmax aggregation were largely preserved. The RadioDINO-Base architecture achieved the highest patient-level AUC (0.837), balanced accuracy (0.746), and F1-score (0.479). Conversely, the RadioDINO-Small configuration attained the highest overall patient-level sensitivity (0.759).

Unlike the behaviour previously observed for DLCS-trained models, patient-level aggregation on LUNA25 did not substantially improve performance for the jointly-trained configurations. Most models instead exhibited small reductions in AUC after aggregation. For example, RadioDINO-Base decreased from a nodule-level AUC of 0.863 to a patient-level AUC of 0.837, while ResNet-34 (ImageNet) decreased from 0.822 to 0.788.

In contrast, the jointly-trained models evaluated on the DLCS cohort (Table 6.7) achieved considerably lower overall performance compared with their corresponding LUNA25 results. Nevertheless, the jointly-trained configurations consistently outperformed the original DLCS-only experiments across nearly all evaluation metrics.

At the nodule level, the jointly-trained ResNet models achieved AUC values ranging from 0.632 to 0.704. ResNet-50 (ImageNet) achieved the highest nodule-level AUC among the ResNet family (0.704), while also achieving the highest sensitivity (0.636). However, similarly to the previous DLCS experiments, the pretrained configurations frequently demonstrated specificity-oriented behaviour. ResNet-18 (Scratch) achieved the highest specificity among the ResNet models (0.829), while ResNet-50 (ImageNet) achieved a lower specificity of 0.673 despite its stronger sensitivity.

Among the DINO architectures, RadioDINO-Base achieved the strongest overall DLCS performance, obtaining the highest nodule-level AUC (0.723), balanced accuracy (0.682), and F1-score (0.299). Compared with the original DLCS-trained DINO experiments, the jointly-trained DINO configurations demonstrated clear performance improvements across all evaluation metrics.

Importantly, the jointly-trained models substantially reduced the performance gap previously observed between the ResNet and DINO families on DLCS. While the original DLCS-only experiments showed relatively similar performance between both architecture families, the jointly-trained DINO models demonstrated clearer advantages in terms of overall discriminative performance and balanced accuracy.

At the patient level, Pmax aggregation again improved performance for most jointly-trained models evaluated on DLCS. RadioDINO-Base achieved the highest patient-level AUC (0.775) and balanced accuracy (0.715), while ResNet-50 (ImageNet) achieved the highest patient-level sensitivity (0.815).

Similarly to the original DLCS-only experiments, patient-level aggregation generally increased sensitivity and balanced accuracy for the jointly-trained models. For example, RadioDINO-Base improved from a nodule-level AUC of 0.723 to a patient-level AUC of 0.775, while ResNet-50 (ImageNet) improved from 0.704 to 0.757.

Overall, the jointly-trained experiments suggest that combining LUNA25 and DLCS produced substantially more robust models than DLCS-only training while largely preserving the strong discriminative performance observed on LUNA25. Furthermore, the jointly-trained setting appeared particularly beneficial for the DINO-based architectures, which consistently achieved the strongest overall performance across both evaluation cohorts.

6.1.3.4 Comparison with State-of-the-Art Literature

The performance dynamics observed across the internal and external evaluation scenarios strongly align with the challenges identified in the current literature (Chapter 3). During internal testing on the LUNA25 cohort, the top-performing models achieved AUC values of approximately 0.86. While recent systematic reviews note that internal evaluations on LIDC-derived benchmarks frequently report AUCs exceeding 0.95 [45, 46], the literature explicitly warns that such optimism is often driven by dataset homogeneity or patient-level data leakage [81]. The baseline metrics reported in this study, therefore, reflect a rigorous, strictly patient-isolated evaluation.

Furthermore, the substantial performance degradation observed during cross-dataset generalization, particularly the sensitivity collapse when transferring from the curated LUNA25 cohort to the more heterogeneous DLCS dataset—is directly corroborated by the benchmarking performed by Tushar et al. [53]. Their study demonstrated an identical vulnerability when models trained on curated sets are evaluated on screening-naive data.

Ultimately, the performance shift from internal baselines to external out-of-distribution cohorts (such as LUNGx) mirrors the well-documented generalization drop in medical imaging, where models routinely experience severe accuracy reductions (e.g., from 98% to 76%) under acquisition and population shifts [45].

6.1.4 Summary

In summary, the internal evaluation established that baseline performance is highly contingent on both architectural paradigm and cohort characteristics. The self-supervised RadioDINO architectures consistently outperformed the standard convolutional ResNet models, demonstrating superior discriminative stability, particularly when trained on the larger LUNA25 dataset.

The application of patient-level aggregation (Pmax) revealed contrasting dynamics: it successfully mitigated false negatives in the challenging DLCS cohort—boosting overall sensitivity—but slightly penalized AUC in LUNA25 due to the compounding risk of false positives in patients with high nodule multiplicity.

Crucially, the experiments demonstrated that standard ImageNet pre-training frequently skewed the CNN operating profiles towards specificity, causing critical drops in clinical sensitivity under high class imbalance. However, joint training (LUNA25 + DLCS) effectively stabilized performance, allowing the Vision Transformers to achieve robust and balanced baselines that align with the realistic expectations established by recent critical literature.

6.2. Cross-Dataset Generalization Analysis

This section evaluates model robustness under distributional shift by applying models trained on one dataset to a different target cohort. All results are reported exclusively at the patient level following Pmax aggregation. For each scenario, two complementary tables are presented: one reporting absolute performance values (mean $\pm$ standard deviation across folds) and one reporting generalization gaps (Δ = cross-dataset value $-$ internal value), computed by subtracting each model's internal patient-level performance on its respective source dataset from its cross-dataset result. Positive Δ values indicate improvement relative to internal performance; negative values indicate degradation. The key cross-dataset patterns are summarised in the discussion following each pair of tables.

Table 6.8: Patient-level absolute performance of LUNA25-trained models on the DLCS test set. Values are reported as mean ± standard deviation across folds. **Bold** text denotes the highest value achieved for each evaluation metric.

Model	AUC	BAcc	Sens.	Spec.
ResNet-18 (Scratch)	0.721 ± 0.001	0.666 ± 0.046	0.574 ± 0.170	0.757 ± 0.085
ResNet-18 (ImageNet)	0.734 ± 0.019	$\mathbf{0.691 \pm 0.014}$	0.593 ± 0.140	0.789 ± 0.114
ResNet-34 (Scratch)	0.730 ± 0.021	0.657 ± 0.026	0.537 ± 0.064	0.778 ± 0.020
ResNet-34 (ImageNet)	0.710 ± 0.024	0.667 ± 0.049	0.556 ± 0.147	0.778 ± 0.060
ResNet-50 (Scratch)	0.718 ± 0.043	0.641 ± 0.036	0.537 ± 0.085	0.744 ± 0.015
ResNet-50 (ImageNet)	0.739 ± 0.014	0.669 ± 0.021	$\mathbf{0.685 \pm 0.128}$	0.652 ± 0.096
RadioDINO-S8	0.745 ± 0.024	0.689 ± 0.030	0.667 ± 0.056	0.711 ± 0.015
RadioDINO-Small	$\mathbf{0.746 \pm 0.030}$	0.665 ± 0.055	0.556 ± 0.111	0.774 ± 0.018
RadioDINO-Base	0.739 ± 0.038	0.642 ± 0.061	0.426 ± 0.160	$\mathbf{0.857 \pm 0.059}$

Table 6.9: Generalization gaps (Δ = cross-dataset − internal) for LUNA25-trained models evaluated on DLCS. Positive values indicate gains over internal performance; negative values indicate degradation.

Model	Δ AUC	Δ BAcc	Δ Sens.	Δ Spec.
ResNet-18 (Scratch)	-0.101	-0.079	-0.223	$+0.064$
ResNet-18 (ImageNet)	-0.106	-0.065	-0.161	$+0.030$
ResNet-34 (Scratch)	-0.099	-0.100	-0.265	$+0.065$
ResNet-34 (ImageNet)	-0.119	-0.097	-0.280	$+0.086$
ResNet-50 (Scratch)	-0.072	-0.089	-0.260	$+0.080$
ResNet-50 (ImageNet)	-0.087	-0.058	-0.180	$+0.062$
RadioDINO-S8	-0.097	-0.067	-0.178	$+0.044$
RadioDINO-Small	-0.111	-0.109	-0.280	$+0.062$
RadioDINO-Base	-0.113	-0.115	-0.270	$+0.039$

6.2.1 LUNA25 → DLCS

When LUNA25-trained models were evaluated on the DLCS test set, all configurations showed consistent AUC and BAcc degradation, with ΔAUC ranging from -0.072 to -0.119 and ΔBAcc from -0.058 to -0.115. The dominant driver of this degradation was a uniform collapse in sensitivity (ΔSens = -0.161 to -0.280), while specificity marginally increased across all models (ΔSpec = $+0.030$ to $+0.086$). This pattern indicates that models calibrated on LUNA25's 15.7% malignancy prevalence systematically over-predicted benign cases when exposed to DLCS's lower prevalence of 8.9%, shifting their effective operating point towards higher specificity at the expense of sensitivity.

Despite this degradation, all LUNA25-trained models achieved higher cross-dataset AUC values (range 0.710–0.746) than DLCS-trained models achieved on their own internal test set (range 0.648–0.733), demonstrating that LUNA25 provides a stronger source

domain. The RadioDINO models consistently outperformed ResNets in this transfer scenario, with RadioDINO-Small achieving the highest AUC (0.746) and RadioDINO-S8 the highest BAcc (0.689), reproducing the performance hierarchy established during internal evaluation.

6.2.2 LUNA25 $\rightarrow$ LUNGx

Table 6.10: Patient-level absolute performance of LUNA25-trained models on LUNGx (mean $\pm$ std across folds). **Bold** text denotes the highest value achieved for each evaluation metric

Model	AUC	BAcc	Sens.	Spec.
ResNet-18 (Scratch)	0.513 ± 0.030	0.539 ± 0.019	0.700 ± 0.208	0.378 ± 0.222
ResNet-18 (ImageNet)	0.621 ± 0.043	$\mathbf{0.594 \pm 0.026}$	0.689 ± 0.135	0.500 ± 0.120
ResNet-34 (Scratch)	0.567 ± 0.001	0.533 ± 0.033	0.722 ± 0.038	0.344 ± 0.077
ResNet-34 (ImageNet)	0.588 ± 0.056	0.539 ± 0.035	0.811 ± 0.051	0.267 ± 0.033
ResNet-50 (Scratch)	0.584 ± 0.021	0.561 ± 0.067	0.822 ± 0.038	0.300 ± 0.145
ResNet-50 (ImageNet)	0.619 ± 0.021	0.550 ± 0.017	0.856 ± 0.019	0.244 ± 0.019
RadioDINO-S8	0.634 ± 0.036	0.589 ± 0.026	$\mathbf{0.911 \pm 0.019}$	0.267 ± 0.058
RadioDINO-Small	0.630 ± 0.033	0.578 ± 0.035	0.767 ± 0.033	0.389 ± 0.084
RadioDINO-Base	$\mathbf{0.636 \pm 0.031}$	$\mathbf{0.594 \pm 0.025}$	0.567 ± 0.115	$\mathbf{0.622 \pm 0.069}$

Table 6.11: Generalization gaps (Δ = cross-dataset $-$ internal) for LUNA25-trained models evaluated on LUNGx.

Model	Δ AUC	Δ BAcc	Δ Sens.	Δ Spec.
ResNet-18 (Scratch)	-0.309	-0.206	-0.097	-0.315
ResNet-18 (ImageNet)	-0.219	-0.162	-0.065	-0.259
ResNet-34 (Scratch)	-0.262	-0.224	-0.080	-0.369
ResNet-34 (ImageNet)	-0.241	-0.225	-0.025	-0.425
ResNet-50 (Scratch)	-0.206	-0.169	$+0.025$	-0.364
ResNet-50 (ImageNet)	-0.207	-0.177	-0.009	-0.346
RadioDINO-S8	-0.208	-0.167	$+0.066$	-0.400
RadioDINO-Small	-0.227	-0.196	-0.069	-0.323
RadioDINO-Base	-0.216	-0.163	-0.129	-0.196

The LUNA25 $\rightarrow$ LUNGx scenario produced the most severe generalization failure of any evaluated configuration, with ΔAUC ranging from -0.206 to -0.309 and ΔBAcc from -0.162 to -0.225 across all models. Critically, the failure mode is qualitatively distinct from the LUNA25 $\rightarrow$ DLCS direction: here, specificity collapsed catastrophically (ΔSpec = -0.196 to -0.425), while sensitivity remained comparatively stable or even slightly improved in some models (ΔSens = -0.129 to $+0.066$). This reversal may be associated with the direction of the label shift: whereas DLCS's lower prevalence pushed models toward conservative predictions, LUNGx's near-balanced class distribution ($\sim$50%) causes

models calibrated on LUNA25's 15.7% prevalence to aggressively over-predict malignancy. The extreme case is RadioDINO-S8, which achieved a sensitivity of 0.911 but a specificity of only 0.267, classifying virtually every patient as malignant.

The only model to partially resist this collapse was RadioDINO-Base (AUC 0.636, BAcc 0.594, sensitivity 0.567, specificity 0.622), whose intrinsically conservative decision boundary partially compensated for the prevalence mismatch. Its ΔSpec of -0.196 was the smallest observed, suggesting that more conservative operating profiles may be less sensitive to positive label shift. Nevertheless, no LUNA25-trained model achieved clinically viable performance on LUNGx, underscoring the fundamental incompatibility between screening-calibrated models and diagnostic-context data with balanced prevalence.

6.2.3 DLCS $\rightarrow$ LUNA25

Table 6.12: Patient-level absolute performance of DLCS-trained models on the LUNA25 test set (mean $\pm$ std across folds). **Bold** text denotes the highest value achieved for each evaluation metric.

Model	AUC	BAcc	Sens.	Spec.
ResNet-18 (Scratch)	0.651 ± 0.011	0.608 ± 0.012	0.498 ± 0.154	0.718 ± 0.157
ResNet-18 (ImageNet)	0.680 ± 0.044	0.631 ± 0.030	0.459 ± 0.090	0.804 ± 0.034
ResNet-34 (Scratch)	0.642 ± 0.020	0.585 ± 0.009	0.575 ± 0.215	0.594 ± 0.209
ResNet-34 (ImageNet)	0.699 ± 0.039	0.631 ± 0.046	0.570 ± 0.164	0.691 ± 0.203
ResNet-50 (Scratch)	0.688 ± 0.013	0.629 ± 0.015	0.560 ± 0.092	0.697 ± 0.116
ResNet-50 (ImageNet)	0.697 ± 0.023	0.629 ± 0.007	0.382 ± 0.045	$\mathbf{0.877 \pm 0.033}$
RadioDINO-S8	0.707 ± 0.047	0.617 ± 0.033	0.507 ± 0.277	0.726 ± 0.211
RadioDINO-Small	0.737 ± 0.008	$\mathbf{0.690 \pm 0.012}$	$\mathbf{0.589 \pm 0.076}$	0.791 ± 0.057
RadioDINO-Base	$\mathbf{0.749 \pm 0.017}$	0.661 ± 0.059	0.536 ± 0.205	0.786 ± 0.088

Table 6.13: Generalization gaps (Δ = cross-dataset $-$ internal) for DLCS-trained models evaluated on LUNA25.

Model	Δ AUC	Δ BAcc	Δ Sens.	Δ Spec.
ResNet-18 (Scratch)	-0.062	-0.010	$+0.035$	-0.054
ResNet-18 (ImageNet)	$+0.031$	$+0.032$	$+0.126$	-0.061
ResNet-34 (Scratch)	-0.059	-0.037	-0.018	-0.056
ResNet-34 (ImageNet)	-0.000	$+0.045$	$+0.126$	-0.035
ResNet-50 (Scratch)	-0.045	-0.036	-0.014	-0.059
ResNet-50 (ImageNet)	-0.021	$+0.029$	$+0.067$	-0.008
RadioDINO-S8	$+0.014$	$+0.040$	$+0.081$	-0.002
RadioDINO-Small	$+0.042$	$+0.044$	$+0.126$	-0.039
RadioDINO-Base	$+0.036$	$+0.029$	$+0.092$	-0.035

The DLCS $\rightarrow$ LUNA25 experiments produced mixed generalization behaviour across the evaluated architectures. As shown in Table 6.13, several models achieved modest improvements after transfer to LUNA25, particularly among the RadioDINO configurations,

where ΔAUC values ranged from $+0.014$ to $+0.042$. However, unlike the behaviour observed in earlier experiments, performance improvements were not uniform across models. Multiple ResNet configurations exhibited reductions in AUC and balanced accuracy after transfer, particularly the scratch-trained variants.

Despite the heterogeneous AUC behaviour, most models demonstrated increased sensitivity when evaluated on LUNA25. Sensitivity improvements ranged from $+0.035$ to $+0.126$ for most configurations, with the largest gains observed for ResNet-18 (ImageNet), ResNet-34 (ImageNet), and RadioDINO-Small. These sensitivity increases were frequently accompanied by reductions in specificity, indicating a shift toward less conservative prediction behaviour on the LUNA25 cohort.

One possible explanation for this behaviour is the difference in prevalence and dataset composition between DLCS and LUNA25. As discussed previously in Chapter 4, DLCS contains a lower malignancy prevalence and fewer nodules per patient compared with LUNA25. Models trained on DLCS may therefore learn more conservative decision boundaries that partially relax when evaluated on the denser LUNA25 cohort, improving sensitivity while reducing specificity.

Among all evaluated configurations, the RadioDINO architectures demonstrated strongest overall cross-dataset robustness in this direction. RadioDINO-Base achieved the highest absolute cross-dataset AUC (0.749), while RadioDINO-Small achieved the highest balanced accuracy (0.690) and sensitivity (0.589). In contrast, several ResNet scratch-trained models demonstrated clear performance degradation after transfer, suggesting greater sensitivity to dataset shift.

Overall, the DLCS $\rightarrow$ LUNA25 experiments indicate that transfer from the smaller and more heterogeneous DLCS cohort to LUNA25 remains feasible, particularly for the DINO-based architectures, although the previously observed uniformly positive transfer behaviour was no longer maintained under the updated evaluation setting.

6.2.4 DLCS $\rightarrow$ LUNGx

DLCS-trained models generalized substantially better to LUNGx than their LUNA25-trained counterparts, with AUC values ranging from 0.549 to 0.649 and ΔAUC from -0.038 to -0.184, compared to -0.206 to -0.309 for LUNA25-trained models. A structurally important difference from the LUNA25 $\rightarrow$ LUNGx scenario is that ΔSens was *positive* for every DLCS-trained model without exception ($+0.034$ to $+0.229$), while ΔSpec remained

Table 6.14: Patient-level absolute performance of DLCS-trained models on LUNGx (mean ± std across folds). **Bold** text denotes the highest value achieved for each evaluation metric

Model	AUC	BAcc	Sens.	Spec.
ResNet-18 (Scratch)	0.551 ± 0.064	0.517 ± 0.029	0.522 ± 0.350	0.511 ± 0.301
ResNet-18 (ImageNet)	0.610 ± 0.037	0.567 ± 0.033	0.367 ± 0.153	$\mathbf{0.767 \pm 0.145}$
ResNet-34 (Scratch)	0.610 ± 0.031	0.556 ± 0.054	0.689 ± 0.250	0.422 ± 0.342
ResNet-34 (ImageNet)	0.601 ± 0.020	0.550 ± 0.017	0.644 ± 0.204	0.456 ± 0.236
ResNet-50 (Scratch)	0.549 ± 0.024	0.550 ± 0.017	$\mathbf{0.722 \pm 0.139}$	0.378 ± 0.135
ResNet-50 (ImageNet)	$\mathbf{0.649 \pm 0.053}$	0.578 ± 0.025	0.544 ± 0.117	0.611 ± 0.069
RadioDINO-S8	0.634 ± 0.083	0.589 ± 0.067	0.611 ± 0.283	0.567 ± 0.353
RadioDINO-Small	0.640 ± 0.030	$\mathbf{0.633 \pm 0.017}$	0.611 ± 0.051	0.656 ± 0.069
RadioDINO-Base	0.634 ± 0.023	0.578 ± 0.063	0.511 ± 0.241	0.644 ± 0.217

Table 6.15: Generalization gaps (Δ = cross-dataset − internal) for DLCS-trained models evaluated on LUNGx.

Model	Δ AUC	Δ BAcc	Δ Sens.	Δ Spec.
ResNet-18 (Scratch)	-0.162	-0.101	$+0.059$	-0.261
ResNet-18 (ImageNet)	-0.038	-0.032	$+0.034$	-0.098
ResNet-34 (Scratch)	-0.090	-0.065	$+0.096$	-0.228
ResNet-34 (ImageNet)	-0.098	-0.035	$+0.200$	-0.270
ResNet-50 (Scratch)	-0.184	-0.115	$+0.148$	-0.378
ResNet-50 (ImageNet)	-0.069	-0.022	$+0.229$	-0.274
RadioDINO-S8	-0.059	$+0.012$	$+0.185$	-0.161
RadioDINO-Small	-0.055	-0.013	$+0.148$	-0.174
RadioDINO-Base	-0.080	-0.054	$+0.067$	-0.176

negative (-0.098 to -0.378). This means DLCS-trained models did not suffer the extreme specificity collapse that characterized LUNA25-trained models on LUNGx; instead, they arrived at a more calibrated trade-off, gaining sensitivity while accepting a moderate specificity reduction. This behaviour may be related to the closer proximity between DLCS's training prevalence (8.9%) and LUNGx's balanced distribution ($\sim$50%), which limits the severity of the calibration mismatch relative to LUNA25.

Among RadioDINO models, RadioDINO-Small achieved the highest BAcc (0.633) and the smallest ΔAUC (-0.055), confirming its robustness to distributional shift as the most generalizable DINO configuration across all evaluated scenarios. On the ResNet side, ResNet-50 (ImageNet) achieved the highest AUC (0.649) with the most balanced ΔSens / ΔSpec profile among ResNets ($+0.229/-0.274$). The large standard deviations for ResNet-18 (Scratch) (± 0.350 in sensitivity) and RadioDINO-S8 (± 0.283) reflect evaluation instability stemming from the small LUNGx test set (70 patients) combined with the variance of the DLCS training procedure, and should be interpreted with caution.

6.2.5 Jointly-Trained Models (LUNA25 + DLCS) $\rightarrow$ LUNGx

To complement the analysis of single-dataset generalization, this section evaluates the robustness of the architectures trained under the combined protocol (LUNA25 + DLCS) when transferred to the external LUNGx cohort. This scenario evaluates whether integrating heterogeneous screening data from multiple institutions broadens the source domain representation sufficiently to mitigate target-domain degradation. Table 6.16 details the absolute performance values obtained at patient levels after Pmax aggregation.

Table 6.16: Absolute patient-level performance of jointly-trained models (LUNA25 + DLCS) evaluated on the external LUNGx dataset. Values are reported as mean $\pm$ standard deviation across the 3 cross-validation folds. **Bold** text denotes the highest value achieved for each evaluation metric.

Model	AUC	Sens.	Spec.	BAcc	F1
ResNet-18 (Scratch)	0.579 $\pm$ 0.040	0.600 $\pm$ 0.047	**0.433** $\pm$ 0.109	0.516 $\pm$ 0.036	0.554 $\pm$ 0.016
ResNet-18 (ImageNet)	0.620 $\pm$ 0.066	0.744 $\pm$ 0.134	0.333 $\pm$ 0.196	0.539 $\pm$ 0.034	0.614 $\pm$ 0.027
ResNet-34 (Scratch)	0.581 $\pm$ 0.080	0.822 $\pm$ 0.042	0.233 $\pm$ 0.000	0.528 $\pm$ 0.021	0.635 $\pm$ 0.022
ResNet-34 (ImageNet)	0.571 $\pm$ 0.011	0.833 $\pm$ 0.027	0.333 $\pm$ 0.027	**0.583** $\pm$ 0.027	0.667 $\pm$ 0.022
ResNet-50 (Scratch)	0.592 $\pm$ 0.051	0.900 $\pm$ 0.027	0.211 $\pm$ 0.042	0.556 $\pm$ 0.031	**0.670** $\pm$ 0.022
ResNet-50 (ImageNet)	0.621 $\pm$ 0.038	**0.911** $\pm$ 0.042	0.178 $\pm$ 0.057	0.544 $\pm$ 0.008	0.666 $\pm$ 0.007
RadioDINO-S8	0.577 $\pm$ 0.008	0.733 $\pm$ 0.098	0.422 $\pm$ 0.110	0.578 $\pm$ 0.021	0.632 $\pm$ 0.031
RadioDINO-Small	0.666 $\pm$ 0.012	0.800 $\pm$ 0.000	0.322 $\pm$ 0.016	0.561 $\pm$ 0.008	0.646 $\pm$ 0.004
RadioDINO-Base	**0.669** $\pm$ 0.014	0.767 $\pm$ 0.000	0.378 $\pm$ 0.016	0.572 $\pm$ 0.008	0.642 $\pm$ 0.004

The empirical findings from the jointly-trained configurations reveal substantial insights into how multi-domain source data modulates out-of-distribution transfer. The integration of LUNA25 and DLCS data led to a visible stabilization of patient-level discriminative performance, with RadioDINO-Base and RadioDINO-Small achieving the highest overall absolute AUC values (0.669 ± 0.014 and 0.666 ± 0.012, respectively). This performance reflects a distinct improvement compared to the single-dataset experiments, where the maximum patient-level AUC achieved on LUNGx by DLCS-trained models remained noticeably lower for the DINO family.

However, an analysis of the operating profiles demonstrates that the jointly-trained models did not entirely bypass the prevalence-driven calibration mismatch that characterizes screening-to-diagnostic domain transfers. The architectures consistently exhibited a heavy skew toward sensitivity at the expense of specificity. For instance, ResNet-50 (ImageNet) reached an outstanding sensitivity of 0.911 ± 0.042, but at the cost of a catastrophic specificity collapse to 0.178 ± 0.057. Similarly, RadioDINO-Small polarized its operating point, yielding a sensitivity of 0.800 ± 0.000 and a specificity of 0.322 ± 0.016.

This prevalent behavior indicates that despite the architectural and source-domain variations introduced by merging datasets, the overarching screening context of LUNA25 and DLCS—characterized by low baseline malignancy rates—forces the networks to overcompensate when encountering the balanced diagnostic prevalence ($\sim$50%) of LUNGx, aggressively over-predicting malignancy.

Comparing these joint-training dynamics directly against the isolated single-dataset results reveals a highly architecture-dependent impact of data merging. For the vision transformer family, the combined training protocol proved to be a major catalyst for out-of-distribution robustness. RadioDINO-Base expanded its patient-level AUC to 0.669 ± 0.014 (up from 0.636 when trained on LUNA25 and 0.634 on DLCS), while RadioDINO-Small advanced to 0.666 ± 0.012 (up from 0.630 on LUNA25 and 0.640 on DLCS). This behavior demonstrates that self-supervised architectures possess a superior capacity to absorb institutional diversity and construct highly generalized feature spaces when exposed to scaled, multi-centric data pools.

Conversely, the CNN family exhibited a distinct vulnerability to data merging during out-of-distribution transfer. The pre-trained ResNet-50 configuration, which had previously emerged as the top individual generalizer in the DLCS $\rightarrow$ LUNGx transfer scenario with a patient-level AUC of 0.649 ± 0.053, suffered a performance contraction to 0.621 ± 0.038 under the joint protocol. This regression, paired with a severe drop in specificity from 0.611 down to 0.178, strongly implies that the massive sample volume of the imbalanced LUNA25 subset effectively overwhelmed the fine-grained multi-institutional alignment. Consequently, the convolutional models over-indexed on LUNA25's domain statistics, completely wiping out the more balanced operating profile they were capable of achieving when optimized on DLCS data alone.

Furthermore, the overall predictive dynamics of the combined cohort indicate that the

sheer size and standardizing pressure of the LUNA25 subset dominated the loss landscape during joint training. Unlike the single-dataset DLCS models, which managed to strike a somewhat calibrated balance on LUNGx, the jointly-trained models inherited the exact predictive behavior of the LUNA25 domain—specifically, the aggressive over-prediction of malignancy—confirming that simple data aggregation is insufficient to counter extreme class imbalances.

In terms of architectural resilience, the self-supervised foundation backbones consistently out-stabilized the convolutional frameworks under multi-domain conditions. While the best-performing convolutional architecture, ResNet-50 (ImageNet), provided a solid patient-level AUC of 0.621 ± 0.038, the RadioDINO architectures achieved significantly tighter standard deviations across all validation folds, achieving a flat 0.000 variance in patient-level sensitivity for both the Small and Base variants. This reinforces that massive self-supervised radiological pretraining provides a foundational semantic alignment that is highly resilient to data splits, even when the model is simultaneously forced to negotiate structural institutional disparities and severe clinical context shifts.

6.2.6 Qualitative and Systematic Error Analysis

To contextualize the aggregated performance metrics and investigate whether model failures occur randomly or follow specific patterns, a qualitative and systematic error profiling was conducted. This analysis isolated the unanimous AI predictions on the LUNA25 internal test set. Specifically, the dataset was filtered to include only the subset of nodules where all 9 evaluated models (the 6 ResNet configurations and the 3 RadioDINO variants) produced the exact same classification outcome—meaning every single architecture simultaneously registered a True Positive, True Negative, False Positive, or False Negative.

Figure 6.6 illustrates a qualitative sampling of these unanimous outcomes, providing visual examples of concordant successes and discordant failures.

To move beyond visual inspection, a quantitative profiling of these unanimous errors was performed by crossing the consensus outcomes with the underlying physical and spatial metadata of the LUNA25 nodules. The analysis evaluated three key metrics extracted directly from the provided LUNA25 metadata annotations: nodule bounding scale (computed from the annotated bounding box dimensions), maximum attenuation density (Hounsfield Units, HU), and spatial location (Z-slice centroid). The resulting distributions

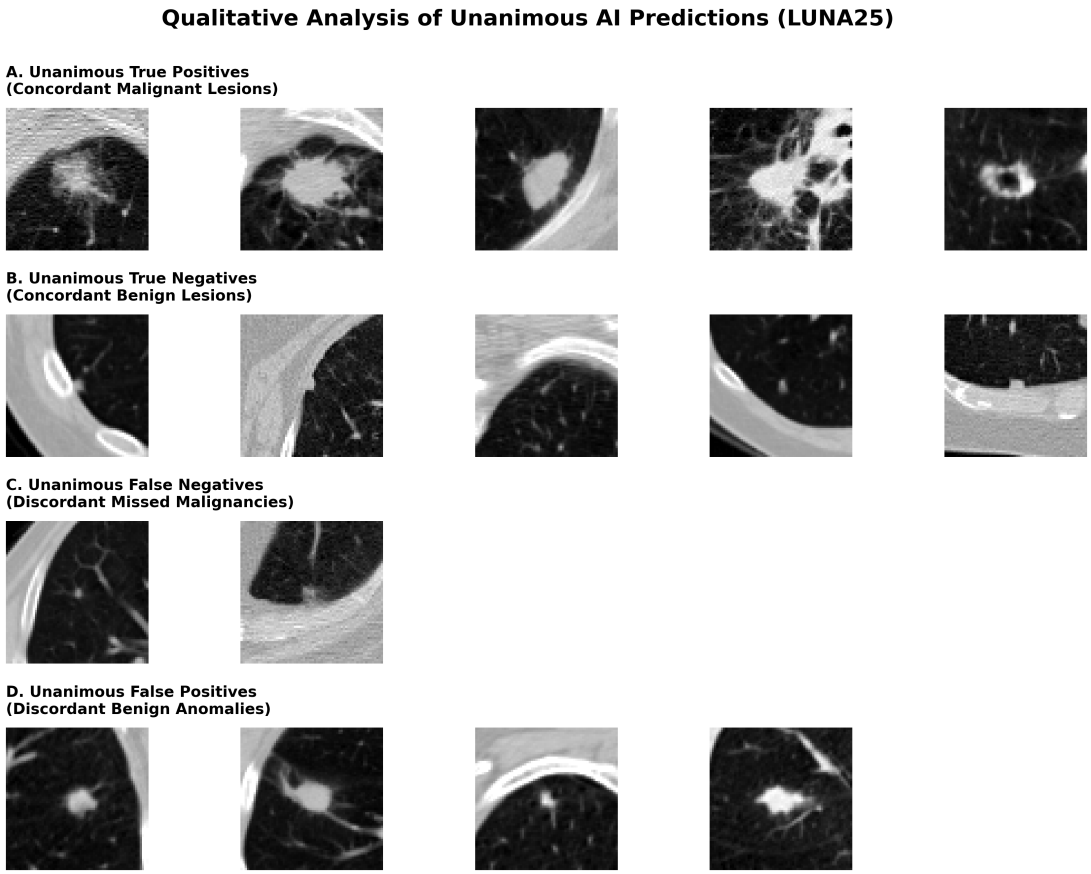


Figure 6.6: Qualitative sampling of unanimous model predictions on the LUNA25 cohort. True Positives (A) and True Negatives (B) exhibit clear morphological features. Conversely, unanimous False Negatives (C) often lack distinct dense cores, while False Positives (D) involve benign structures that visually mimic malignant opacities.

are presented in Figure 6.7. The empirical data demonstrates that unanimous AI failures are not random, but rather adhere to strict geometric and anatomical profiles.

The Profile of False Negatives (Missed Malignancies): Unanimous false negatives, where all models failed to identify a true malignancy—represent the most critical clinical vulnerability in a screening context. The systematic analysis revealed three shared characteristics within this cohort:

- **Reduced Relative Dimensions:** Missed malignancies exhibited the lowest bounding scale distribution across the study cohort.
- **Lower Attenuation Density:** The average maximum density for false negatives (2273.50 HU) was substantially lower than that of correctly identified malignancies (2554.75 HU). This confirms the hypothesis that the networks struggle to segment

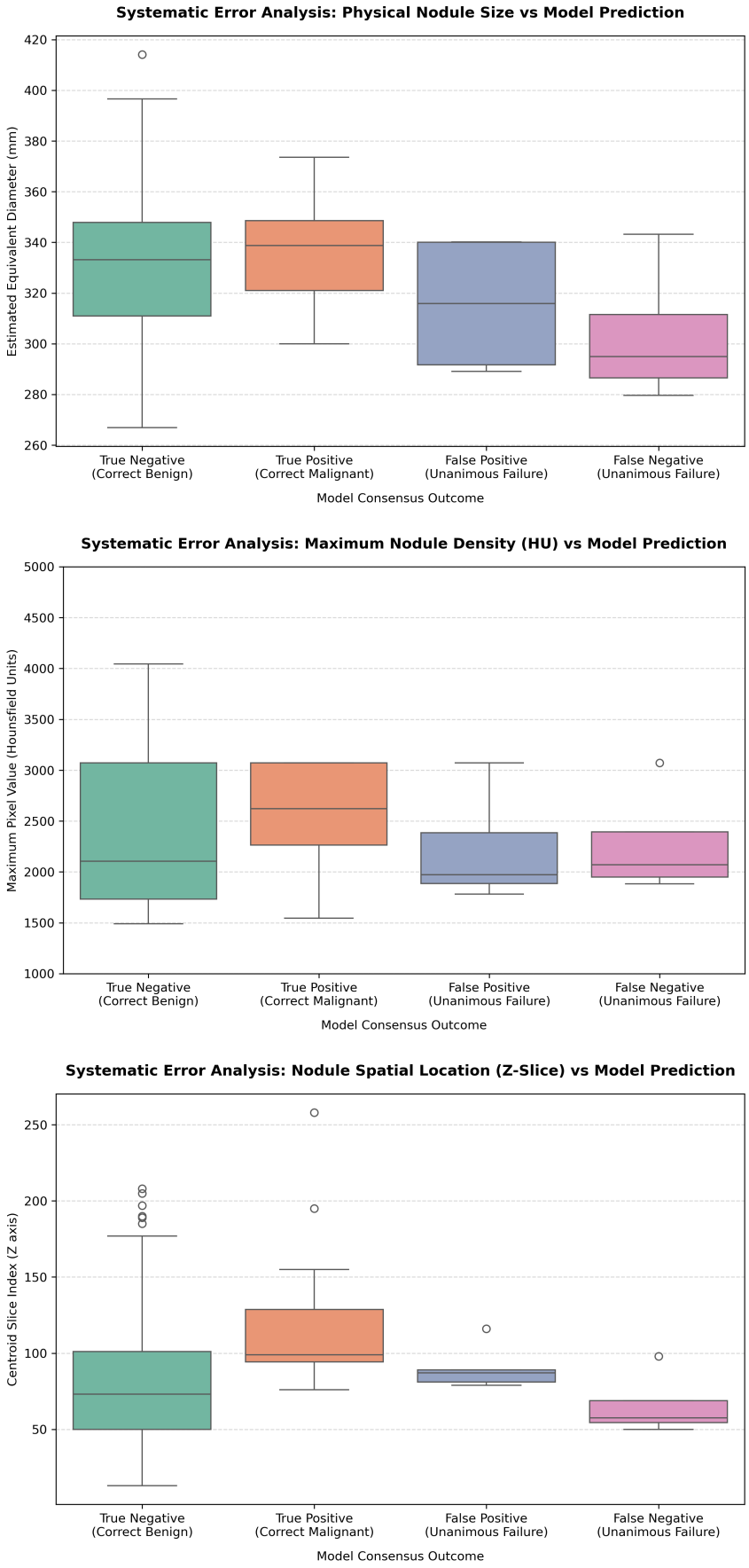


Figure 6.7: Systematic error profiling of unanimous model predictions based on physical and spatial metadata. Top: Relative nodule scale (derived from the transformed bounding box volume). Middle: Maximum pixel attenuation (HU). Bottom: Spatial location (Z-slice index). Note: The estimated equivalent diameter reflects the scaled bounding box dimensions after preprocessing, serving as a relative spatial context indicator rather than an absolute clinical tumor measurement.

and classify sub-solid or ground-glass opacities, where low radiologic contrast obscures boundary definition.

- **Basal Localization:** The spatial distribution indicated a strong concentration in the lower pulmonary slices (mean Z-slice: 65.75). This basal localization exposes the nodules to respiratory motion artifacts and complex anatomical transitions near the diaphragm, heavily degrading feature extraction.

The Profile of False Positives (Spurious Alarms): Unanimous false positives occurred when benign structures triggered confident malignancy predictions across all models. This cohort exhibited a highly specific signature:

- **Intermediate Density:** These anomalies presented the lowest overall maximum density (2219.60 HU), distinguishing them from highly calcified benign granulomas which models easily reject.
- **Strict Anatomical Confinement:** The spatial distribution of false positives was remarkably narrow, clustered tightly around the central-lower thorax (mean Z-slice: 90.40). This suggests these spurious detections are localized to specific anatomical regions, likely corresponding to hilar structures or mediastinal vascular cross-sections that geometrically mimic malignant nodularity.

The Optimal Detection Zone (True Concordance): Conversely, the models consistently succeeded when lesions fell within an optimal physical profile. Correctly identified benign and malignant nodules shared higher overall densities (2487.64 and 2554.75 HU, respectively), providing robust structural contrast. Furthermore, true positives were predominantly located in the central and upper pulmonary regions (mean Z-slice: 123.50), where the lung parenchyma is relatively unobstructed by dense surrounding tissues, allowing both convolutional and self-supervised networks to leverage clear morphological features.

Importantly, this granular profiling was restricted to the LUNA25 dataset. As previously detailed during the exploratory data analysis in Chapter 4, while spatial coordinates exist for the external DLCS cohort, the dataset lacks harmonized nodule-level radiodensity metadata. Furthermore, the high variance in scan-level acquisition protocols established in Section 4.2 precludes a direct and fair geometric comparison between the two cohorts.

6.2.7 Summary

To synthesize the extensive cross-dataset evaluation, Table 6.17 provides an executive summary of the transfer scenarios. It highlights the best-performing models per architecture family and categorizes the dominant predictive shift observed relative to their internal baselines.

Table 6.17: Executive summary of cross-dataset generalization. The table highlights the best-performing models per architecture family across all transfer scenarios, alongside the dominant operating shift observed relative to their internal baselines.

Scenario	**Best ResNet**	**Best RadioDINO**	**Dominant Shift Profile**		**Robustness**
(Source → Target)	**(Patient AUC)**	**(Patient AUC)**	**Δ Sens.**	**Δ Spec.**	
LUNA25 → DLCS	ResNet-50 (ImgNet) [**0.739**]	RadioDINO-Small [**0.746**]	↓↓ Severe Drop	↑ Slight Gain	Poor (Sens. Collapse)
LUNA25 → LUNGx	ResNet-18 (ImgNet) [**0.621**]	RadioDINO-Base [**0.636**]	≈ Stable	↓↓ Severe Drop	Critical (Spec. Collapse)
DLCS → LUNA25	ResNet-34 (ImgNet) [**0.699**]	RadioDINO-Base [**0.749**]	↑ Moderate Gain	↓ Slight Drop	Moderate (DINO leads)
DLCS → LUNGx	ResNet-50 (ImgNet) [**0.649**]	RadioDINO-Small [**0.640**]	↑ Moderate Gain	↓↓ Severe Drop	Moderate (Balanced)
Merged → LUNGx	ResNet-50 (ImgNet) [**0.621**]	RadioDINO-Base [**0.669**]	↑↑ Extreme Gain	↓↓ Severe Drop	Improved ViT stability

As summarized above, the direction of the domain transfer dictates the failure mode. Transferring from a high-prevalence to a low-prevalence screening domain (LUNA25 → DLCS) triggers sensitivity collapse, whereas transferring to a balanced diagnostic domain (LUNGx) forces a severe drop in specificity. Ultimately, the self-supervised RadioDINO architectures demonstrated superior resilience across the majority of severe domain shifts compared to the supervised convolutional baselines.

To visually complement this summary, Figure 6.8 illustrates the Δ AUC generalization gaps across all single-source transfer scenarios, clearly highlighting the architecture-dependent resilience.

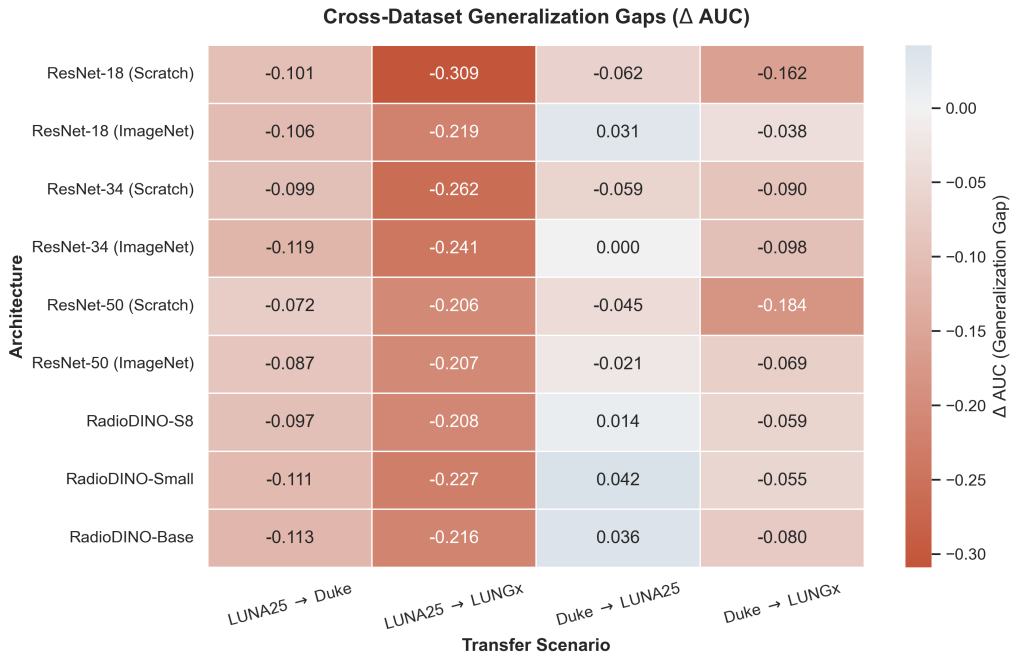


Figure 6.8: Heatmap of cross-dataset generalization gaps (Δ AUC). Red hues indicate performance degradation relative to the internal baseline evaluation, while blue hues indicate performance gains. The self-supervised RadioDINO models demonstrate comparatively greater stability (lighter colours and positive gaps) across the evaluation scenarios.

6.3. Fairness Analysis

This section evaluates demographic fairness across the evaluated architectures under both internal and external evaluation settings.

To facilitate interpretation, fairness results are organised by demographic attribute, combining internal and cross-dataset evaluations within the same subsection. This structure enables each demographic attribute to be analysed consistently across different deployment scenarios.

Overall, gender was the demographic attribute associated with the highest number of significant disparities, particularly in DLCS-related evaluations, while age produced the fewest significant findings across all scenarios. Table 6.18 provides an aggregate overview of statistically significant bias detections per attribute and scenario.

To further illustrate the concentration of demographic disparities, Figure 6.9 visualizes the data from Table 6.18 as a bias intensity map. This representation immediately highlights gender and cross-dataset deployment (particularly Merged $\rightarrow$ DLCS) as the primary failure modes of the evaluated architectures.

Table 6.18: Summary of statistically significant bias detections ($p < 0.05$) per demographic attribute and evaluation scenario. Each cell reports biased tests out of the total performed (9 models $\times$ 4 metrics per attribute). N/A indicates the attribute was not available for that target dataset.

Attribute	LUNA25 (int.)	DLCS (int.)	LUNA25→DLCS	DLCS→LUNA25	Merged→LUNA25	Merged→DLCS
Gender	4 / 36	14 / 36	15 / 36	13 / 36	6 / 36	**19 / 36**
Age	0 / 36	1 / 36	1 / 36	3 / 36	2 / 36	3 / 36
Race	1 / 36	0 / 36	5 / 36	2 / 36	0 / 36	0 / 36
Education	3 / 36	N/A	N/A	1 / 36	3 / 36	N/A
Smoking Status	1 / 36	4 / 36	2 / 36	6 / 36	3 / 36	3 / 36
Total	**9 / 180**	**19 / 144**	**23 / 144**	**25 / 180**	**14 / 180**	**25 / 144**

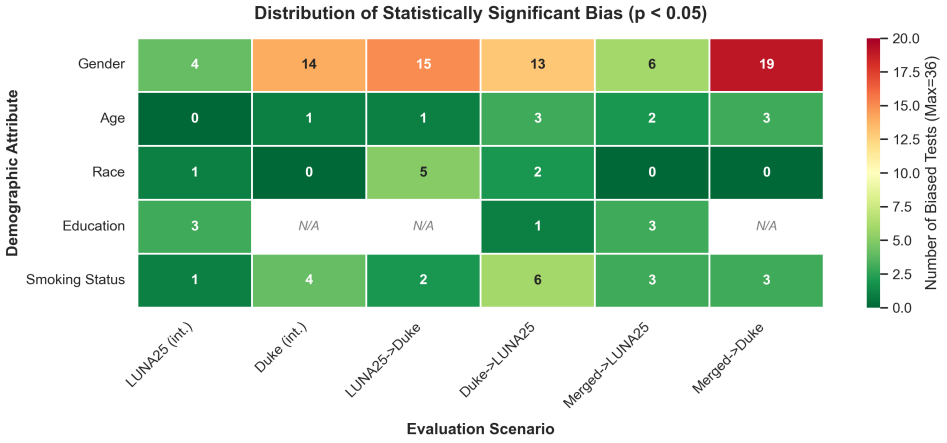


Figure 6.9: Bias intensity map illustrating the number of statistically significant disparities ($p < 0.05$) across demographic attributes and deployment scenarios. Warmer colours indicate a higher concentration of biased models.

6.3.1 Gender

Gender represented the most prevalent source of demographic disparity across the evaluated scenarios. While only limited gender-related bias was observed during internal evaluation on LUNA25, substantially larger disparities emerged in DLCS-related evaluations, particularly under cross-dataset and merged-training settings.

For internal evaluation on LUNA25, only 4 out of 36 gender-related tests reached statistical significance, mainly affecting scratch-trained ResNet models through sensitivity or specificity differences between male and female patients. In contrast, all ImageNet-pretrained ResNets and all RadioDINO variants showed no significant gender disparities on this dataset.

A markedly different pattern was observed on DLCS, where 14 out of 36 tests were significant. Multiple models exhibited large performance gaps between male and female patients, particularly for AUC. For example, RadioDINO-S8 achieved an AUC of 0.357 for male patients compared with 0.791 for female patients. Similar patterns were observed

for RadioDINO-Base and RadioDINO-Small. These disparities likely reflect the strong prevalence imbalance present in the DLCS cohort, compounded by a critical clinical-morphological mismatch. As male patients are more susceptible to central, solid squamous cell carcinomas—whereas females develop more peripheral, sub-solid adenocarcinomas [2, 13]—the reduction in sensitivity for males suggests that models optimized on screening distributions may preferentially adapt to imaging patterns more prevalent in female patients, potentially reflecting reduced generalization to male-associated imaging patterns.

To contextualize these disparities, it is essential to examine the underlying gender distributions of the cohorts. Intuitively, algorithmic bias is often attributed to the numerical underrepresentation of a specific subgroup. However, demographic analysis reveals that the LUNA25 training cohort is actually male-predominant (57.1% male vs. 42.9% female), and the DLCS evaluation cohort is nearly perfectly balanced (50.2% male vs. 49.8% female). The fact that the networks exhibit significantly lower discrimination capacity for male patients—despite males being adequately or overly represented in the training data—demonstrates that this performance collapse is not a simple artifact of sample size imbalance. Instead, it strongly reinforces the hypothesis that male-associated nodule morphologies present inherently more heterogeneous or radiologically complex features that the architectures fail to systematically learn, whereas female-associated patterns are more readily generalized across domains.

Cross-dataset evaluation further amplified these disparities. In the LUNA25 $\rightarrow$ DLCS scenario, 15 out of 36 gender-related tests became significant, despite several of these models showing no significant gender bias internally on LUNA25. RadioDINO-S8 was particularly affected, reaching significance across all four evaluated fairness metrics simultaneously. To illustrate this disparity, RadioDINO-S8 achieved an AUC of 0.838 and a sensitivity of 0.786 for female patients, but decreased to an AUC of 0.444 and a sensitivity of just 0.250 for male patients. The DLCS $\rightarrow$ LUNA25 direction also showed substantial gender disparities (13/36 tests), mainly affecting ResNet-18 variants and RadioDINO-S8.

The merged-training setting produced the strongest overall gender bias, particularly for Merged $\rightarrow$ DLCS, where 19 out of 36 tests were significant. This represents the highest number of fairness violations observed across any attribute or evaluation scenario. Notably, merged training produced more gender bias on DLCS than DLCS-only training.

Overall, the results indicate that gender-related fairness issues are strongly influenced

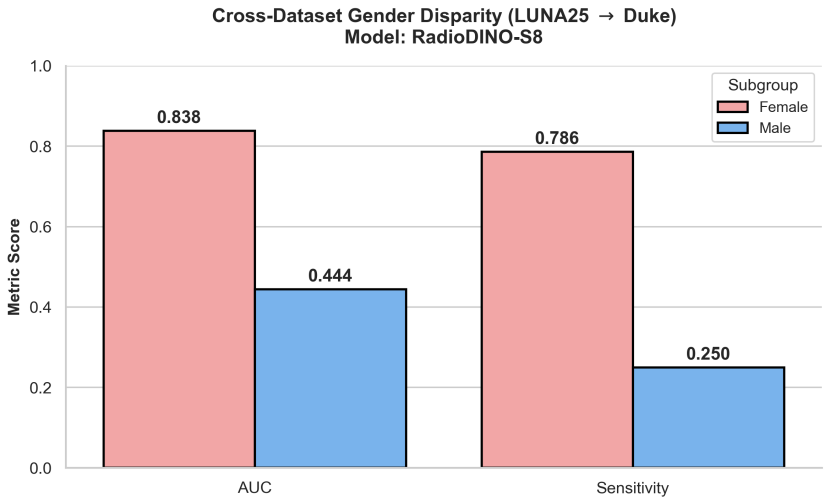


Figure 6.10: Performance collapse of the RadioDINO-S8 architecture when trained on LUNA25 and evaluated on DLCS. The model maintains substantially higher performance for female patients, while exhibiting markedly reduced discrimination capability for male patients, including AUC values below random-chance level (0.5).

by dataset composition and subgroup prevalence differences rather than by architecture choice alone. This suggests that simply increasing training data through dataset merging does not necessarily improve fairness and may instead amplify subgroup prevalence imbalances already present in the cohorts, a paradox visualised in Figure 6.11.

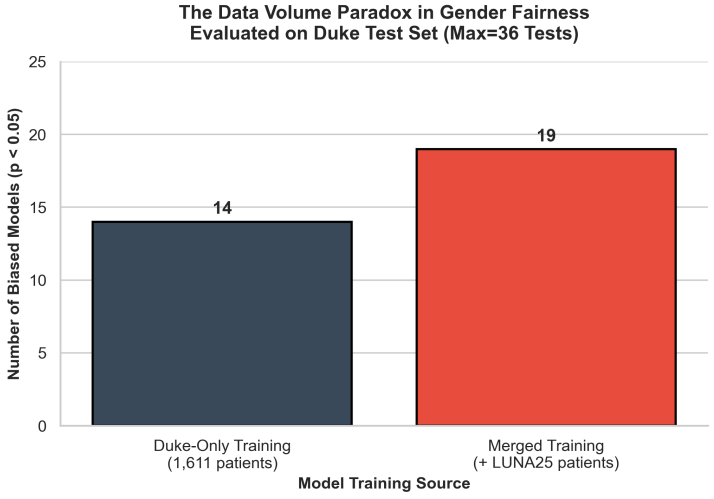


Figure 6.11: The impact of dataset merging on gender fairness when evaluated on the DLCS cohort. Contrary to the assumption that larger training datasets inherently improve fairness, introducing the LUNA25 distribution exacerbated the number of models exhibiting statistically significant gender bias.

These findings are strongly corroborated by recent literature. *Gaur et al.* [82] demonstrated that deep learning architectures for lung cancer prediction exhibit significant performance degradation in male patients, despite males constituting the majority of the training cohort [82]. This validates the conclusion that gender disparities are not driven solely by numerical underrepresentation. Instead, it suggests that deep learning models are highly susceptible to ”shortcut learning,” where the networks identify biological sex directly from the imaging data and overfit to gender-specific morphological patterns rather than universally generalizing true pathological features [82].

6.3.2 Age

Age showed the most stable fairness behaviour across all evaluated attributes, with very few statistically significant findings.

No significant age-related disparities were detected during internal evaluation on LUNA25 (0/36 tests). On DLCS, only a single significant result was identified, corresponding to an AUC disparity for ResNet-18 (Scratch).

Cross-dataset evaluation produced only isolated findings. The LUNA25 $\rightarrow$ DLCS scenario resulted in one significant sensitivity disparity for ResNet-34 (ImageNet), while the DLCS $\rightarrow$ LUNA25 setting produced three significant specificity disparities affecting ResNet-18 and ResNet-34 variants.

Similarly, the merged-training experiments showed only a small number of age-related disparities. Three significant findings were detected for Merged $\rightarrow$ DLCS, mainly involving sensitivity differences between younger and older patients.

Overall, age exhibited the most stable fairness behaviour in this study. However, this apparent stability must be interpreted within the context of lung cancer screening protocols. Because screening eligibility is inherently age-restricted, both the LUNA25 and DLCS datasets present a narrow age distribution tightly concentrated on older adults (e.g., 55 to 76 years for LUNA25, and 50 to 89 years for the DLCS cohort). Consequently, the lack of observed bias likely reflects this limited demographic variance, rather than a universal algorithmic resilience across all age brackets.

6.3.3 Race

Race-related disparities were generally limited during internal evaluations but became more visible during external validation.

On LUNA25, only one statistically significant race-related result was identified, corresponding to a specificity disparity for RadioDINO-Base. No significant race-related findings were detected during internal evaluation on DLCS.

The strongest racial disparities emerged in the LUNA25 $\rightarrow$ DLCS scenario, where 5 out of 36 tests reached statistical significance. These findings primarily involved sensitivity or specificity differences affecting both ResNet and RadioDINO models. For instance, RadioDINO-Small exhibited a significant specificity disparity ($p = 0.012$), dropping from 0.805 in White patients to 0.703 in Non-White patients. This vulnerability is directly explained by the demographic characteristics detailed in Chapter 4. Specifically, the LUNA25 training cohort is highly homogeneous, with White patients comprising 93.2% of the data and Non-White populations severely underrepresented (6.6%). Consequently, models optimized on this imbalanced distribution lack the necessary exposure to generalize robustly when evaluated on the DLCS cohort, which features a significantly higher Non-White prevalence (26.1%). In contrast, models trained on the more racially diverse DLCS dataset exhibited fewer disparities when evaluated on LUNA25, confirming that exposure to more representative training data is critical to mitigating racial bias.

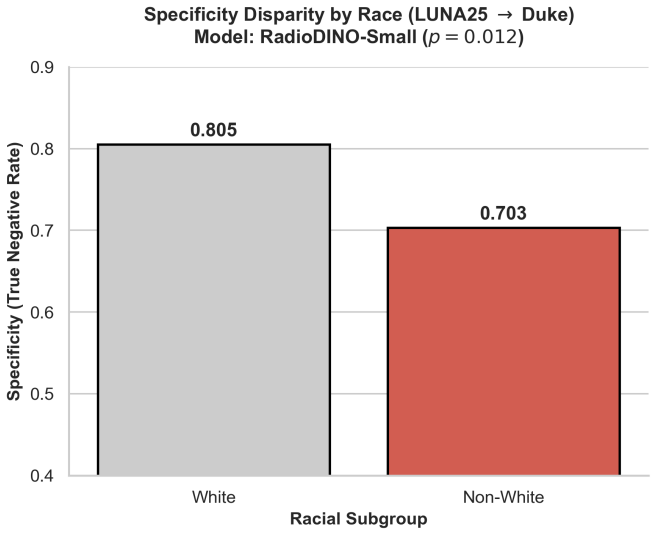


Figure 6.12: Specificity disparity across racial subgroups for the RadioDINO-Small model (LUNA25 $\rightarrow$ DLCS). The model generates a significantly higher rate of false positives for Non-White patients, suggesting reduced calibration robustness for underrepresented populations.

Overall, race-related fairness issues were considerably less frequent than gender-related disparities, but they became more evident under external validation settings.

6.3.4 Education

Education level was only available for LUNA25-based evaluations. Most significant findings corresponded to specificity disparities between patients with higher and lower educational attainment.

During internal LUNA25 evaluation, three significant education-related disparities were identified, affecting ResNet-50 (Scratch), ResNet-34 (ImageNet), and ResNet-18 (ImageNet). In all cases, patients with lower educational attainment showed lower specificity values, indicating a higher false-positive rate. This degradation in specificity may indicate increased sensitivity to benign structural anomalies—such as apical scarring, calcifications, or granulomas from prior respiratory infections—which may be more prevalent in lower socioeconomic cohorts due to differences in occupational exposure or healthcare access, leading the network to trigger spurious malignancy alerts [2, 21, 22, 42].

The same general pattern persisted under merged-training evaluation on LUNA25, where three additional specificity disparities were observed. In contrast, the DLCS $\rightarrow$ LUNA25 scenario produced only one significant education-related result.

Although less prevalent than gender bias, these findings suggest that socioeconomic factors associated with educational level may indirectly influence imaging characteristics or screening outcomes.

6.3.5 Smoking Status

Smoking status produced relatively few disparities during internal evaluation on LUNA25 but became more relevant in DLCS-related scenarios.

As established in Chapter 4, the LUNA25 cohort is composed exclusively of current and former smokers, with no never-smokers represented. Within this strict binary distribution, only one statistically significant smoking-related result was observed internally on LUNA25. However, DLCS internal evaluation produced four significant findings, primarily affecting RadioDINO-Base and involving sensitivity differences between current and former smokers.

The strongest smoking-related disparities emerged in the DLCS → LUNA25 scenario, where 6 out of 36 tests reached significance. Multiple models achieved substantially lower sensitivity for current smokers compared with former smokers, an important disparity illustrated in Figure 6.13. This vulnerability is well-documented in recent literature, which attributes such algorithmic failures to smoking-related parenchymal alterations. Conditions such as emphysema, chronic inflammation, and interstitial fibrosis introduce architectural distortion and background noise in CT scans. These alterations effectively mask nodular features and degrade the feature extraction capabilities of deep learning models compared to the relatively clearer lung parenchyma typically found in former smokers [98–100].

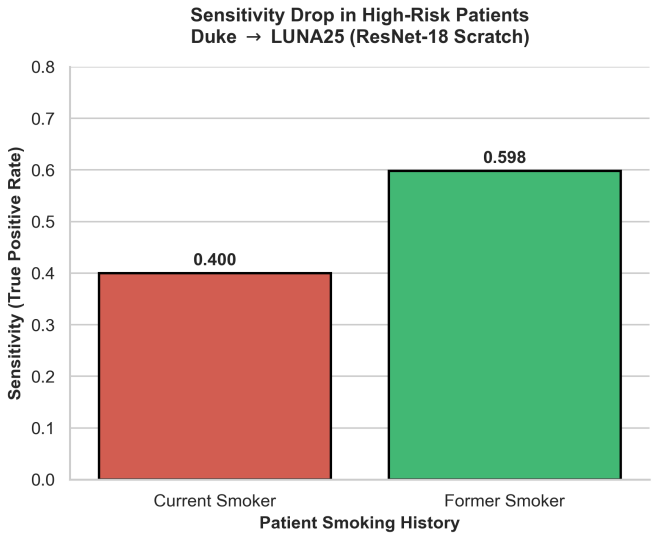


Figure 6.13: Sensitivity disparity based on smoking history for ResNet-18 evaluated out-of-distribution (DLCS → LUNA25). The model shows reduced sensitivity for malignancies in current smokers, the primary target demographic for lung cancer screening.

Merged-training experiments also showed persistent smoking-related disparities across both target datasets, indicating that these effects were not restricted to a single training distribution.

Given that current smokers represent the primary target population for lung cancer screening, these disparities are clinically relevant and highlight the importance of evaluating subgroup sensitivity independently from overall predictive performance.

6.3.6 Hardware and Acquisition Bias

Beyond patient demographics, the influence of physical acquisition parameters and scanner hardware was evaluated to assess model robustness against domain shifts. Statistical

tests were systematically applied to scanner manufacturer, reconstruction slice thickness, X-ray tube voltage (kVp), and reconstruction kernels across all deployment scenarios. To provide a general overview of these vulnerabilities, detailed hardware scoreboards were generated for each evaluation setting, categorizing statistical tests as either fair ($p \geq 0.05$) or biased ($p < 0.05$) across the evaluated architectures. This macroscopic analysis revealed that models are highly susceptible to the physics of image acquisition, with cross-dataset scenarios exposing severe hardware dependencies.

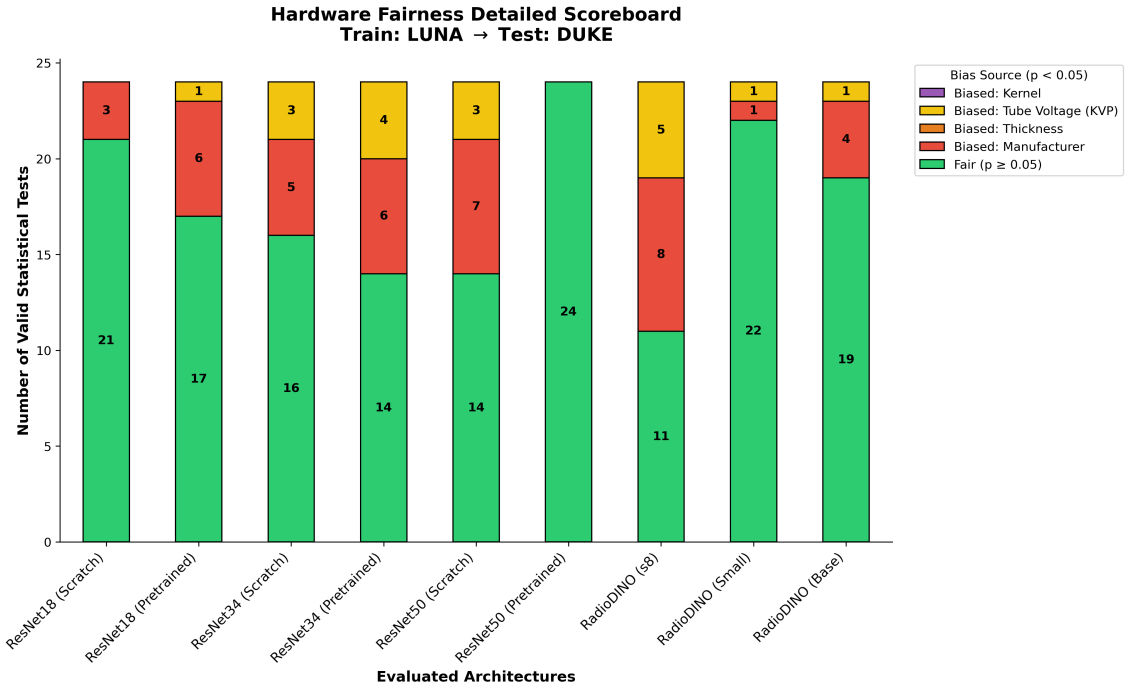


Figure 6.14: Hardware fairness scoreboard for the LUNA25 → DLCS cross-dataset scenario. The red segments highlight a massive proportion of models exhibiting statistically significant bias related to the scanner manufacturer.

6.3.6.1 Scanner Manufacturer

The choice of scanner manufacturer (Siemens vs. GE Medical Systems) emerged as the most critical source of performance disparity in single-source models. When models trained exclusively on the LUNA25 cohort—where over 85% of the scans originate from just two manufacturers (GE Medical Systems and Siemens), as established in Section 4.2.1—were evaluated on DLCS, they exhibited a severe sensitivity collapse across different vendors, as evidenced by the dominant red segments in Figure 6.14.

However, unlike demographic disparities, this hardware-induced bias was successfully mitigated through dataset merging. As illustrated in the Merged → DLCS scoreboard

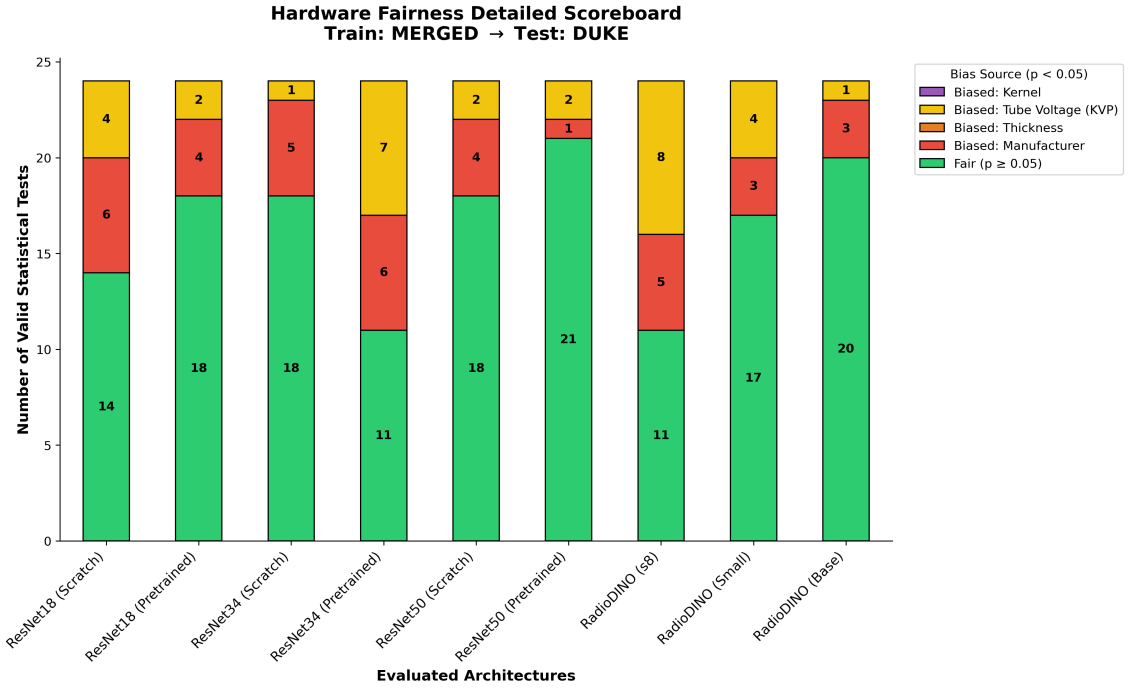


Figure 6.15: Hardware fairness scoreboard for the Merged → DLCS scenario. Compared to single-source training, multi-source exposure substantially increases the number of fair tests (green) and mitigates manufacturer bias (red).

(Figure 6.15), the frequency of manufacturer bias dropped significantly, replaced by a much higher proportion of fair statistical tests.

This mitigation is further detailed in Figure 6.16. Single-source models trained on DLCS demonstrated large sensitivity gaps between GE Medical Systems and Siemens scans ($p < 0.05$). In contrast, merged-training models learned to generalize across hardware textures, rendering the sensitivity differences statistically non-significant ($p > 0.05$). For instance, RadioDINO-Small achieved highly balanced sensitivities across manufacturers ($p = 0.313$), indicating that multi-source training effectively forces the network to ignore manufacturer-specific noise patterns.

6.3.6.2 Reconstruction Slice Thickness

Spatial resolution, specifically reconstruction slice thickness, also severely impacted model discrimination, consistently appearing as a source of bias (orange segments in the scoreboards) across multiple scenarios. As demonstrated by the permutation tests, performance degraded significantly in ultra-thin slices (≤ 1.5 mm) compared to standard thicknesses ($1.5 - 3.0$ mm). For example, evaluating the RadioDINO-Base model internally on LUNA25 revealed a significant performance drop, with the AUC falling from 0.933 in standard slices to 0.768 in thin slices ($p = 0.004$).

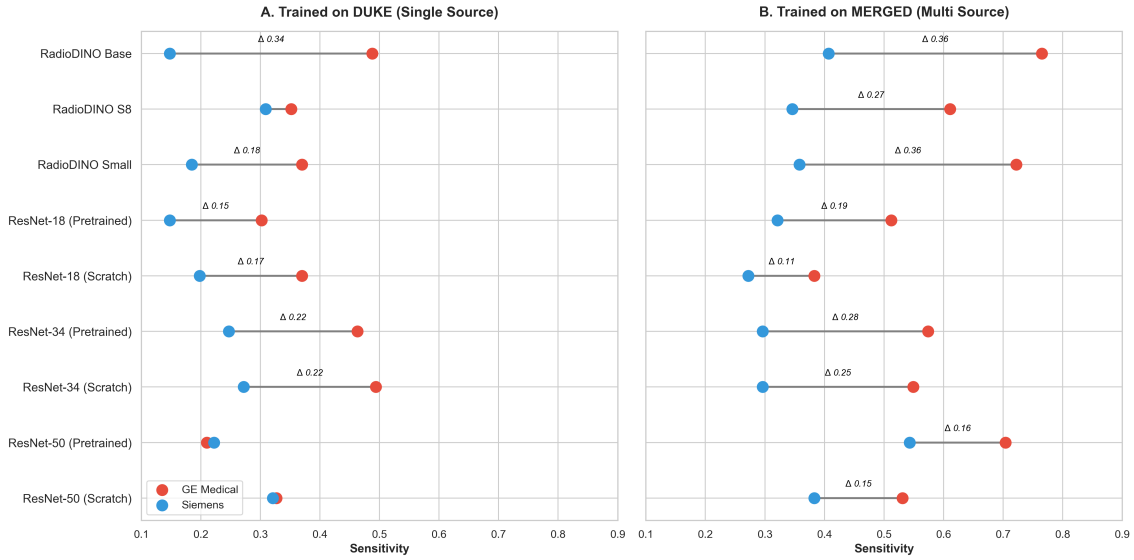


Figure 6.16: Sensitivity disparity across scanner manufacturers (GE Medical vs. Siemens). Panel A demonstrates significant performance gaps for models trained solely on DLCS. Panel B illustrates how merged training effectively mitigates this hardware bias.

It is crucial to note that although all CT volumes were computationally resampled to an isotropic $1 \times 1 \times 1$ mm^3 resolution during pre-processing (Section 4.4), this spatial interpolation cannot recover the signal-to-noise ratio inherent to the original physical acquisition. This degradation, visualized in Figure 6.17, is intrinsically linked to the physics of computed tomography: ultra-thin reconstructions systematically increase high-frequency image noise, which confounds the convolutional feature extractors despite the standardized voxel dimensions.

6.3.6.3 X-Ray Tube Voltage (kVp)

Variations in X-ray tube voltage (kVp) introduced measurable, physics-driven disparities as well, particularly visible as yellow segments in DLCS-related evaluations. As shown in Figure 6.18, lower voltage acquisitions (≤ 100 kVp) systematically resulted in reduced average sensitivity compared to standard and high-energy protocols. Permutation tests confirmed this trend, with multiple models showing statistically significant drops in predictive performance under low-kVp conditions. This aligns with the physical expectation that lower energy protocols yield higher quantum mottle, increasing the morphological ambiguity of pulmonary nodules.

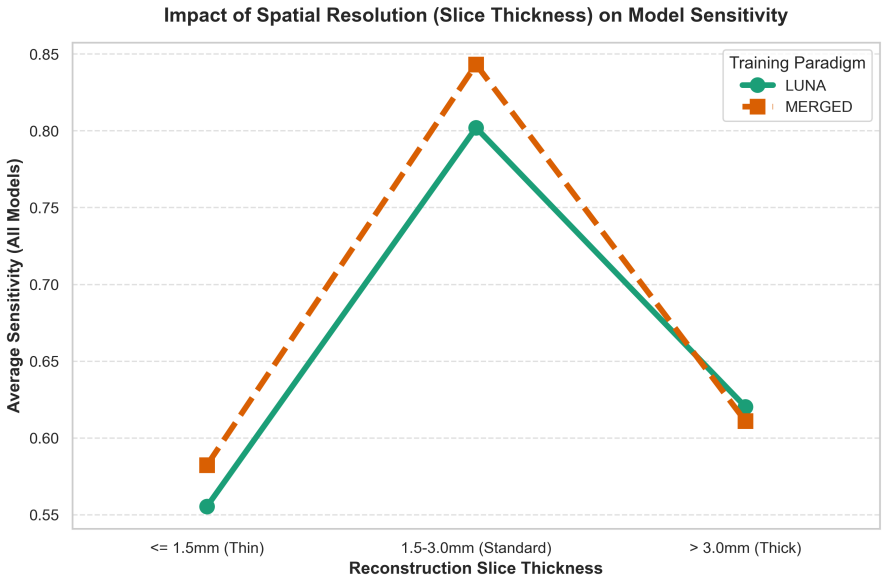


Figure 6.17: Impact of spatial resolution on model sensitivity. A sharp degradation in performance is observed for ultra-thin slices (≤ 1.5 mm) across training paradigms, reflecting the algorithmic vulnerability to high-frequency image noise.

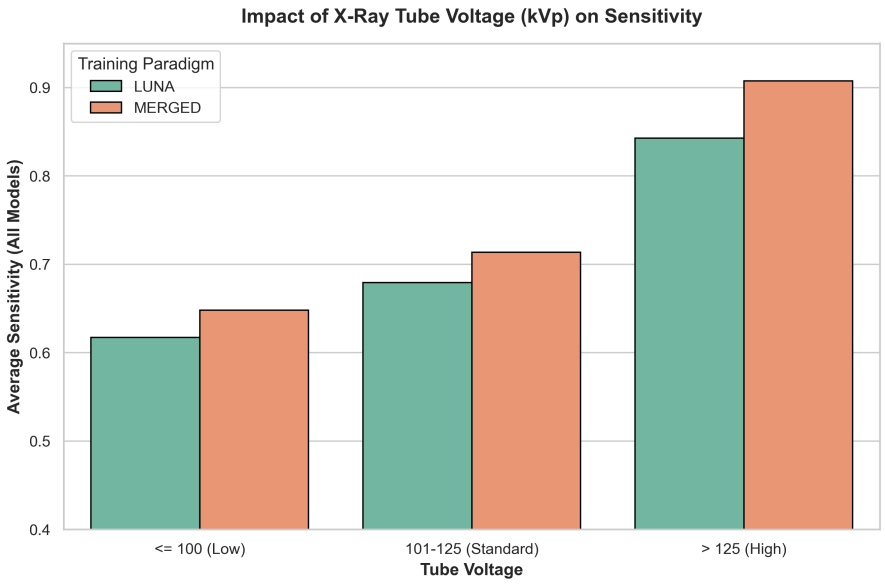


Figure 6.18: Impact of X-Ray tube voltage (kVp) on model sensitivity. Lower energy acquisitions (≤ 100 kVp) systematically reduce diagnostic sensitivity due to increased quantum noise.

6.3.6.4 Reconstruction Kernels

Finally, attempts were made to evaluate the impact of reconstruction kernels (Smooth vs. Sharp). However, the lack of kernel diversity within the external datasets—which are overwhelmingly dominated by standard/smooth reconstructions—prevented robust statistical validation across subgroups, with the majority of tests resulting in non-computable states (represented by the scarcity of purple segments in the scoreboards). This limitation highlights the necessity for highly heterogeneous, multi-center datasets in future studies to fully audit kernel-induced biases.

6.3.7 Summary

Overall, the fairness analysis revealed that algorithmic disparities were highly dependent on both the target dataset and the deployment setting, encompassing both demographic inequalities and acquisition-based domain shifts.

Gender emerged as the most problematic demographic attribute, particularly in DLCS-related evaluations, where strong subgroup prevalence imbalances produced large differences in model behaviour between male and female patients. Importantly, merged training did not reduce these demographic disparities and, in some cases, amplified them. Age showed the most stable behaviour across all experiments, with only isolated statistically significant findings. Race-related disparities remained limited during internal evaluation but became more apparent under external validation, particularly when models trained on the predominantly White LUNA25 cohort were evaluated on DLCS. Smoking status also produced important disparities, especially in cross-dataset settings, with several models showing lower sensitivity for current smokers.

Crucially, the extended hardware analysis demonstrated that deep learning architectures are highly sensitive to physical imaging parameters. Single-source models suffered from massive manufacturer-induced domain shifts when evaluated across external cohorts. However, unlike demographic traits, this specific hardware bias was successfully resolved through multi-source merged training. Conversely, physical acquisition constraints, such as ultra-thin slice thicknesses and low tube voltages, consistently degraded model sensitivity across all training paradigms due to the inherent physical increase in tomographic image noise.

Finally, no architecture consistently achieved the best fairness performance across all scenarios. The results therefore highlight the critical importance of evaluating predictive

performance, demographic fairness, and hardware robustness independently, particularly in multi-cohort medical imaging studies involving highly heterogeneous patient populations and scanner configurations.

6.4. Limitations

Several limitations should be considered when interpreting the findings of this dissertation.

First, all experiments were conducted using 2D axial nodule patches extracted from CT scans. While this approach enables computationally efficient training and facilitates the use of pretrained computer vision models, it does not fully exploit the three-dimensional nature of pulmonary nodules. Volumetric information and inter-slice relationships may provide additional morphological cues that could improve classification performance and potentially enhance robustness under cross-dataset deployment.

Second, some fairness analyses were constrained by the demographic composition of the available public cohorts. In particular, the LUNA25 dataset is highly homogeneous, with White patients representing more than 93% of the cohort. As a result, certain subgroup comparisons involved relatively small sample sizes, reducing the statistical power of fairness analyses and limiting the strength of conclusions regarding underrepresented populations.

Third, the primary datasets used for model development and internal validation (LUNA25 and DLCS) originate from lung cancer screening programmes and therefore represent highly selected populations. Eligibility criteria typically target individuals at elevated risk, most commonly current or former smokers within specific age ranges. Consequently, the findings and performance baseline reported in this dissertation should be interpreted primarily within the context of screening populations and may not directly generalize to broader diagnostic settings (as evidenced by the domain shifts observed during the LUNGx evaluation) or to populations with substantially different clinical characteristics.

Finally, although this work evaluated a diverse set of demographic and acquisition-related factors, several potentially relevant sources of bias could not be investigated due to limited data availability across cohorts. Variables such as socioeconomic status, environmental exposures, comorbidities, family history, and detailed smoking characteristics were not consistently available for all datasets. Future studies incorporating richer and more diverse clinical information may provide a more comprehensive assessment of fairness and model robustness.

7. Conclusion

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, making early detection through screening programmes a critical component of improving patient outcomes. In this context, the present dissertation investigated the generalization and fairness of deep learning models for pulmonary nodule classification across multiple screening cohorts. Beyond predictive performance, particular emphasis was placed on evaluating how models behave when deployed on external datasets and whether demographic disparities emerge under different deployment scenarios.

The experimental results demonstrated that self-supervised foundation models, particularly the RadioDINO architectures, consistently achieved the strongest overall predictive performance on the primary evaluation cohort. Across most experiments, pretrained models outperformed their scratch-trained counterparts, highlighting the importance of transfer learning when developing deep learning systems for medical imaging tasks. Furthermore, patient-level aggregation generally improved performance relative to nodule-level evaluation, suggesting that clinically meaningful information can be recovered through appropriate aggregation strategies.

Despite strong internal performance, the generalization experiments revealed a substantial degradation when models were evaluated on external datasets. This trend was particularly evident on the LUNGx cohort, where models experienced marked performance drops despite achieving strong results on the screening datasets. Models that achieved excellent discrimination on their training cohort frequently experienced notable performance drops under cross-dataset evaluation, confirming that domain shift remains a major obstacle to the reliable deployment of artificial intelligence systems in lung cancer screening. Differences in patient populations, acquisition protocols, scanner characteristics, and dataset composition were all shown to influence model behaviour, reinforcing the importance of external validation beyond traditional internal testing procedures.

The fairness analysis further demonstrated that predictive performance alone is insufficient to characterize model reliability. Gender emerged as the demographic attribute most frequently associated with significant disparities, particularly in DLCS-related evaluations and cross-dataset deployment settings. In contrast, age showed relatively stable behaviour across all experiments, while race and smoking status produced disparities that

became more apparent under external validation. Importantly, the results showed that fairness is highly dependent on the target population and deployment scenario, with several biases only becoming visible after distribution shifts. These findings highlight a close relationship between generalization and fairness, suggesting that demographic robustness should be considered alongside predictive performance during model evaluation.

One of the most interesting findings of this work was that increasing the amount of training data through dataset merging did not consistently improve fairness. Although merged training sometimes improved predictive performance, it could also amplify demographic disparities, particularly for gender-related evaluations on the DLCS cohort. This observation challenges the common assumption that larger datasets inherently lead to fairer models and emphasizes the importance of explicitly assessing demographic behaviour rather than relying solely on aggregate performance metrics.

Equally critical to the demographic audit was the systematic evaluation of physical acquisition parameters, which revealed that the architectures are highly susceptible to scanner hardware and imaging protocols. Single-source models suffered from severe manufacturer-induced domain shifts, demonstrating massive sensitivity gaps between distinct scanner brands. However, in stark contrast to the demographic disparities, this hardware-specific bias was successfully resolved through multi-source merged training, demonstrating that models can learn to generalize across distinct scanner noise patterns when exposed to sufficient institutional diversity. Conversely, physical acquisition constraints, such as ultra-thin reconstruction slice thicknesses and low X-ray tube voltages, consistently degraded diagnostic sensitivity across all training paradigms due to the inherent amplification of tomographic image noise, an effect that computational resampling alone could not overcome.

Several limitations should be acknowledged. First, the study relied on 2D nodule patches, which may not fully capture the three-dimensional contextual information available in CT examinations. Second, fairness analyses for some demographic attributes were constrained by relatively small subgroup sample sizes, limiting statistical power. Additionally, the screening cohorts used in this work represent highly selected populations defined by specific eligibility criteria, many of whom were current or former smokers. The availability and consistency of detailed clinical information regarding participant selection varied across datasets, which may influence both generalization and fairness analyses. Consequently, the findings should be interpreted within the context of lung cancer screening

populations rather than the broader general population.

Building upon these findings and limitations, several avenues for future research emerge. First, transitioning from 2D axial patches to fully 3D volumetric architectures could allow models to leverage richer spatial context and inter-slice relationships, potentially improving diagnostic robustness. Second, future studies should investigate dynamic or adaptive decision thresholds, moving beyond the fixed 0.5 probability cutoff to better accommodate the extreme class imbalances inherent to screening populations. Furthermore, incorporating diverse, multi-center cohorts with multi-modal clinical metadata (such as detailed socio-economic factors and longitudinal smoking history) will be crucial for conducting more granular fairness audits. Finally, exploring algorithmic debiasing techniques directly during the optimization phase could help mitigate the demographic and hardware disparities identified in this work before clinical deployment.

In conclusion, this dissertation demonstrates that evaluating deep learning models for lung cancer screening requires a multidimensional perspective that extends beyond predictive accuracy. While foundation models such as RadioDINO show considerable promise for pulmonary nodule classification, substantial challenges remain regarding cross-dataset generalization and demographic fairness. The results reinforce the need for rigorous external validation, diverse evaluation cohorts, and fairness-aware assessment protocols before these systems can be confidently deployed in real-world clinical environments. Ultimately, the clinical viability of artificial intelligence for lung cancer screening depends not only on predictive accuracy, but also on the ability of models to remain robust, generalizable, and fair across diverse populations and deployment settings.

References

[1] World Health Organization. (2023) Lung cancer fact sheet. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/lung-cancer [Cited on page 1.]

[2] K. C. Thandra, A. Barsouk, K. Saginala, J. S. Aluru, and A. Barsouk, "Epidemiology of lung cancer," *Contemporary Oncology*, vol. 25, no. 1, pp. 45–52, 2021. [Cited on pages 1, 8, 9, 10, 11, 12, 13, 14, 111, and 115.]

[3] F. Silva, T. Pereira, I. Neves, J. Morgado, C. Freitas, M. Malafaia, J. Sousa, J. Negrão, and B. F. de Lima, "Towards machine learning-aided lung cancer clinical routines: Approaches and open challenges," *Journal of Personalized Medicine*, vol. 12, no. 3, p. 480, 2022. [Cited on pages 1, 16, 17, 19, 24, and 32.]

[4] S. Tomassini *et al.*, "Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: A survey," *Computers in Biology and Medicine*, 2022. [Cited on pages 1, 2, 24, 27, 28, and 32.]

[5] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi, "The limits of fair medical imaging ai in real-world generalization," *Nature Medicine*, 2024. [Cited on pages 2 and 19.]

[6] J. McCabe, D. Cheng, A. Bhamani, M. Mullin, T. Patrick, and A. Nair, "Exploring fairness in state-of-the-art pulmonary nodule detection algorithms," in *Ethics and Fairness in Medical Imaging*. Springer, 2024. [Cited on pages 2, 16, 17, 19, and 44.]

[7] Z. Xu, J. Li, Q. Yao, H. Li, M. Zhao, and S. K. Zhou, "Addressing fairness issues in deep learning-based medical image analysis: a systematic review," *npj Digital Medicine*, vol. 7, p. 286, 2024. [Cited on pages 2, 19, and 20.]

[8] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011. [Cited on page 5.]

[9] A. Patel, "Benign vs malignant tumors," *JAMA Oncology*, vol. 6, no. 9, p. 1488, 2020. [Cited on page 5.]

[10] F. R. Hirsch, G. V. Scagliotti, J. L. Mulshine, R. Kwon, W. J. C. Jr., Y.-L. Wu, and L. Paz-Ares, "Lung cancer: Current therapies and new targeted treatments," *The Lancet*, vol. 389, no. 10066, pp. 299–311, 2017. [Cited on pages 5 and 6.]

[11] C. M. Rudin, E. Brambilla, C. Faivre-Finn, and J. Sage, "Small-cell lung cancer," *Nature Reviews Disease Primers*, vol. 7, no. 1, p. 3, 2021. [Cited on page 5.]

[12] B. Jiang, D. Han, C. M. van der Aalst *et al.*, "Lung cancer volume doubling time by computed tomography: A systematic review and meta-analysis," *European Journal of Cancer*, vol. 212, p. 114339, 2024. [Cited on pages 5 and 6.]

[13] W. D. Travis, "Pathology of lung cancer," *Clinics in Chest Medicine*, vol. 32, no. 4, pp. 669–692, 2011. [Cited on pages 6, 8, and 111.]

[14] Y. Mao, D. Yang, J. He, and M. J. Krasna, "Epidemiology of lung cancer," *Surgical Oncology Clinics of North America*, vol. 25, no. 3, pp. 439–445, 2016. [Cited on page 6.]

[15] C. S. Dela Cruz, L. T. Tanoue, and R. A. Matthay, "Lung cancer: epidemiology, etiology, and prevention," *Clinics in chest medicine*, vol. 32, no. 4, pp. 605–644, 2011. [Cited on page 6.]

[16] I. Brüske-Hohlfeld, M. Mühner, H. Pohlabeln, W. Ahrens, U. Bolm-Audorff, L. Kreienbrock, M. Kreuzer, I. Jahn, H.-E. Wichmann, and K.-H. Jückel, "Occupational lung cancer risk for men in germany: Results from a pooled case-control study," *American Journal of Epidemiology*, vol. 151, no. 4, pp. 384–395, 02 2000. [Online]. Available: https://doi.org/10.1093/oxfordjournals.aje.a010218 [Cited on page 6.]

[17] G. Menvielle, J. Franck, L. Radoï *et al.*, "Quantifying the mediating effects of smoking and occupational exposures in the relation between education and lung cancer: the icare study," *European journal of epidemiology*, vol. 31, pp. 1213–1221, 2016. [Cited on page 6.]

[18] R. Wender, E. T. Fontham, E. B. Jr. *et al.*, "American cancer society lung cancer screening guidelines," *CA: A Cancer Journal for Clinicians*, vol. 63, no. 2, pp. 107–117, 2013. [Cited on page 7.]

[19] C. Gridelli, A. Rossi, D. P. Carbone *et al.*, "Non-small-cell lung cancer," *Nature Reviews Disease Primers*, vol. 1, p. 15009, 2015. [Cited on pages 7 and 8.]

[20] P. Goldstraw, K. Chansky, J. Crowley *et al.*, “The iaslc lung cancer staging project: Proposals for revision of the tnm stage groupings in the forthcoming (eighth) edition of the tnm classification for lung cancer,” *Journal of Thoracic Oncology*, vol. 11, no. 1, pp. 39–51, 2016. [Cited on page 7.]

[21] D. N. Haddad, K. L. Sandler, L. M. Henderson, M. P. Rivera, and M. C. Aldrich, “Disparities in lung cancer screening: A review,” *Annals of the American Thoracic Society*, vol. 17, no. 4, pp. 399–405, 2020. [Cited on pages 8, 9, 10, 11, 12, 13, 14, and 115.]

[22] S. Monestime, M. Rigney, N. Phipps, C. Carlson, T. Alongi, M. Redding, and J. C. King, “Health inequities across the lung cancer care continuum in ten marginalized populations: a narrative review,” *Journal of Thoracic Disease*, vol. 15, no. 11, pp. 6326–6334, 2023. [Cited on pages 8, 9, 10, and 115.]

[23] European Society for Medical Oncology (ESMO). (2025) Living guidelines: Non-small cell lung cancer. [Online]. Available: https://www.esmo.org/guidelines/living-guidelines?tumour_sites=Non-Small+Cell+Lung+Cancer [Cited on page 14.]

[24] The National Lung Screening Trial Research Team, “The national lung screening trial: Overview and study design,” *Radiology*, vol. 258, no. 1, pp. 243–253, 2011. [Cited on pages 17, 25, 27, and 47.]

[25] Y. R. Zhao, X. Xie, H. J. de Koning, W. P. Mali, R. Vliegenthart, and M. Oudkerk, “Nelson lung cancer screening study,” *Cancer Imaging*, vol. 11, no. 1A, pp. S79–S84, 2011.

[26] M. M. W. Wille, A. Dirksen, H. Ashraf *et al.*, “Results of the randomized danish lung cancer screening trial with focus on high-risk profiling,” *American Journal of Respiratory and Critical Care Medicine*, vol. 193, no. 5, pp. 542–551, 2016. [Cited on pages 14 and 17.]

[27] K. Greenway. (2025) Hounsfield unit. Radiopaedia.org. [Online]. Available: https://radiopaedia.org/articles/hounsfield-unit [Cited on page 15.]

[28] Radiology Cafe. (2017) Ct windowing. [Online]. Available: https://www.radiologycafe.com/medical-students/radiology-basics/ct-windowing/ [Cited on page 15.]

[29] H. J. D. Koning, C. M. van der Aalst, P. A. de Jong *et al.*, "Reduced lung-cancer mortality with volume CT screening in a randomized trial," *New England Journal of Medicine*, vol. 382, no. 6, pp. 503–513, 2020. [Cited on pages 16 and 17.]

[30] H.-U. Kauczor, A.-M. Baird, T. G. Blum *et al.*, "ESR/ERS statement paper on lung cancer screening," *European Radiology*, vol. 30, pp. 3277–3294, 2020.

[31] G. Veronesi, D. R. Baldwin, C. I. Henschke *et al.*, "Recommendations for implementing lung cancer screening with low-dose computed tomography in europe," *Cancers*, vol. 12, no. 6, p. 1672, 2020. [Cited on page 16.]

[32] D. Cuete. (2025) Ct artifacts. Radiopaedia.org. [Online]. Available: https://radiopaedia.org/articles/ct-artifacts [Cited on page 16.]

[33] S. Lam, H. Bryant, L. Donahoe, A. Domingo *et al.*, "Management of screen-detected lung nodules: A canadian partnership against cancer guidance document," *Canadian Journal of Respiratory, Critical Care, and Sleep Medicine*, vol. 4, no. 4, pp. 236–265, 2020. [Cited on page 17.]

[34] D. S. Gierada, T. K. Pilgram, M. Ford *et al.*, "Lung cancer: interobserver agreement on interpretation of pulmonary findings at low-dose ct screening," *Radiology*, vol. 246, no. 1, pp. 265–272, 2008. [Cited on pages 17 and 18.]

[35] H. A. Gietema, Y. Wang, D. Xu *et al.*, "Pulmonary nodules detected at lung cancer screening: interobserver variability of semiautomated volume measurements," *Radiology*, vol. 241, no. 1, pp. 251–257, 2006. [Cited on page 18.]

[36] A. McWilliams, M. C. Tammemagi, J. R. Mayo, H. Roberts, G. Liu, K. Soghrati, K. Yasufuku, S. Martel, F. Laberge, M. Gingras *et al.*, "Probability of cancer in pulmonary nodules detected on first screening ct," *New England Journal of Medicine*, vol. 369, no. 10, pp. 910–919, 2013. [Cited on page 18.]

[37] S. J. Swensen, M. D. Silverstein, D. M. Ilstrup, C. D. Schleck, and E. S. Edell, "The probability of malignancy in solitary pulmonary nodules: Application to small radiologically indeterminate nodules," *Archives of Internal Medicine*, vol. 157, no. 8, pp. 849–855, 1997. [Cited on page 18.]

[38] M. K. Gould, L. Ananth, P. G. Barnett, and V. A. S. C. S. Group, "A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules," *Chest*, vol. 131, no. 2, pp. 383–391, 2007. [Cited on page 18.]

[39] S. J. Adams, R. C. Kornas, M.-L. Nguyen, E. Seto, S. Urschel, A. O'Riordan, M. Cheung, N. S. Paul, and H. C. Roberts, "Artificial intelligence and machine learning in lung cancer screening," *Thoracic Surgery Clinics*, vol. 33, no. 4, pp. 389–399, 2023, review on AI applications in lung cancer screening, including nodule detection and challenges in generalization. [Cited on page 18.]

[40] C. Gao, L. Wu, W. Wu, Y. Huang, X. Wang, Z. Sun, M. Xu, and C. Gao, "Deep learning in pulmonary nodule detection and segmentation: a systematic review," *European Radiology*, vol. 35, pp. 255 – 266, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271086044

[41] A. A. Peters, W. H. Mares, M. van Assen *et al.*, "Impact of artificial intelligence assistance on pulmonary nodule detection and localization in chest ct: a comparative study among radiologists of varying experience levels," *Scientific Reports*, vol. 14, no. 1, p. 22147, 2024, study showing AI improves nodule detection rates, especially for less experienced readers; relevant for motivation in high-volume screening. [Cited on page 18.]

[42] A. S. Tejani, Y. S. Ng, Y. Xi, and J. C. Rayan, "Understanding and mitigating bias in imaging artificial intelligence," *RadioGraphics*, vol. 44, no. 2, p. e230026, 2024. [Cited on pages 20, 21, 22, and 115.]

[43] I. Dankwa-Mullan, K. Ndoh, D. Akogo *et al.*, "Artificial intelligence and cancer health equity: Bridging the divide or widening the gap," *Current Oncology Reports*, vol. 27, pp. 1–12, 2025. [Cited on pages 22 and 23.]

[44] A. Mbakwe, I. Lourentzou, L. Celi, and J. T. Wu, "Fairness metrics for health ai: we have a long way to go," *EBioMedicine*, vol. 88, p. 104440, 2023. [Cited on page 22.]

[45] H. T. Gayap and M. A. Akhloufi, "Deep machine learning for medical diagnosis, application to lung cancer detection: A review," *BioMedInformatics*, vol. 4, no. 1, pp. 236–284, 2024. [Online]. Available: https://www.mdpi.com/2673-7426/4/1/15 [Cited on pages 24, 27, 29, 31, 32, 75, 95, and 96.]

[46] ——, "Deep machine learning for medical diagnosis, application to lung cancer detection: A review," 2024, perception, Robotics, and Intelligent Machines (PRIME), Department of Computer Science, Université de Moncton, Moncton, NB, Canada. [Cited on pages 24, 27, 29, 31, and 95.]

[47] F. Prior, K. Smith, A. Sharma, J. Kirby, L. Tarbox, K. Clark, W. Bennett, T. Nolan, and J. Freymann, "The public cancer radiology imaging collections of the cancer imaging archive," *Sci. Data*, vol. 4, pp. 1–7, 2017. [Cited on pages 24 and 27.]

[48] A. Setio, A. Traverso, T. De Bel, M. Berens, C. Bogaard, P. Cerello, H. Chen, Q. Dou, M. Fantacci, B. Geurts *et al.*, "Lung nodule analysis 2016 (luna16) dataset," https://luna16.grand-challenge.org/, 2016, (accessed on 27 March 2023). [Cited on pages 25, 27, and 60.]

[49] LUNA25 Challenge Organizers, "LUNA25: Lung Nodule Analysis 2025 Challenge," https://luna25.grand-challenge.org/, 2025, accessed: 2026-03-20. [Cited on pages 25, 27, and 47.]

[50] J. Pedrosa, G. Aresta, C. Ferreira *et al.*, "Lndb: a lung nodule database on computed tomography," *arXiv preprint arXiv:1911.08434*, 2019. [Cited on pages 25 and 27.]

[51] S. G. Armato, K. Drukker, F. Li, L. Hadjiiski, G. D. Tourassi, R. M. Engelmann, M. L. Giger, G. Redmond, K. Farahani, J. S. Kirby, and L. P. Clarke, "LUNGx Challenge for computerized lung nodule classification," *Journal of Medical Imaging*, vol. 3, no. 4, p. 044506, Dec. 2016. [Cited on pages 25 and 57.]

[52] ——, "SPIE-AAPM-NCI Lung Nodule Classification Challenge Dataset," 2015. [Cited on pages 25 and 57.]

[53] F. I. Tushar *et al.*, "Duke lung cancer screening (dlcs) dataset," Sep. 2024. [Online]. Available: https://zenodo.org/records/13799069 [Cited on pages 25, 27, and 95.]

[54] S. G. Armato III *et al.*, "Prostatex/lungx challenge datasets," *Journal of Medical Imaging*, 2015. [Cited on page 27.]

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385 [Cited on pages 27 and 66.]

[56] H. Tan, J. Bates, and C. Kinsey, "Discriminating tb lung nodules from early lung cancers using deep learning," *BMC Medical Informatics and Decision Making*, vol. 22, pp. 1–7, 2022. [Cited on pages 27, 28, and 61.]

[57] Q. Zhang, H. Wang, S. Yoon, D. Won, and K. Srihari, “Lung nodule diagnosis on 3d computed tomography images using deep convolutional neural networks,” *Procedia Manufacturing*, vol. 39, pp. 363–370, 2019. [Cited on page 28.]

[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” pp. 248–255, 2009. [Cited on pages 28, 64, and 67.]

[59] D. Moitra and R. K. Mandal, “Prediction of non-small cell lung cancer histology by a deep ensemble of convolutional and bidirectional recurrent neural network,” *Journal of Digital Imaging*, vol. 33, no. 4, pp. 895–902, 2020. [Cited on page 28.]

[60] I. Ali, M. Muzammil, I. Haq, A. Khaliq, and S. Abdullah, “Efficient lung nodule classification using transferable texture convolutional neural network,” *IEEE Access*, vol. 8, pp. 175 859–175 870, 2020. [Cited on page 28.]

[61] A. Halder, S. Chatterjee, and D. Dey, “Adaptive morphology aided 2-pathway convolutional neural network for lung nodule classification,” *Biomedical Signal Processing and Control*, vol. 72, p. 103347, 2022. [Cited on pages 28 and 29.]

[62] R. V. M. Da Nóbrega, S. A. Peixoto, S. P. P. da Silva, and P. P. Rebouças Filho, “Lung nodule classification via deep transfer learning in ct lung images,” in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2018, pp. 244–249. [Cited on page 29.]

[63] H.-H. Chang, C.-Z. Wu, and A. Gallogly, “Pulmonary nodule classification using a multiview residual selective kernel network,” *Journal of Imaging Informatics in Medicine*, vol. 37, pp. 347–362, 2024. [Cited on page 30.]

[64] W. Zhao *et al.*, “3d multi-scale multi-task cnn for lung nodule classification,” *Magnetic Resonance Imaging*, 2020. [Cited on page 30.]

[65] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021. [Cited on pages 30 and 90.]

[66] S. Khademi, S. Heidarian, P. Afshar, F. Naderkhani, A. Oikonomou, K. Plataniotis, and A. Mohammadi, “Spatio-temporal hybrid fusion of cae and swin transformers for

lung cancer malignancy prediction," *arXiv preprint arXiv:2210.15297*, 2022. [Cited on page 31.]

[67] H. Mkindu, L. Wu, and Y. Zhao, "3d multi-scale vision transformer for lung nodule detection in chest ct images," *Signal, Image and Video Processing*, vol. 17, pp. 2473–2480, 2023. [Cited on page 31.]

[68] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature Medicine*, vol. 27, no. 12, pp. 2176–2182, 2021. [Cited on pages 32 and 79.]

[69] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, April 2023. [Cited on pages 33, 35, and 36.]

[70] A. Storkey, "When training and test sets are different: characterizing learning transfer," in *Dataset shift in machine learning*. MIT Press, 2009, pp. 3–28. [Cited on page 34.]

[71] J. S. Yoon, K. Oh, Y. Shin, M. A. Mazurowski, and H.-I. Suk, "Domain generalization for medical image analysis: A review," *Proceedings of the IEEE*, vol. 112, no. 10, pp. 1763–1793, 2024. [Cited on pages 34, 36, and 37.]

[72] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, March 2022. [Cited on page 35.]

[73] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, May 2022, arXiv:2103.03097. [Cited on pages 35, 36, and 37.]

[74] H. Li, H. Li, J. Chen, Z. Qiu, H. Fu, L. Wang, Y. Hu, and J. Liu, "Rafflesdg: Random frequency filtering enabled single-source domain generalization for medical image segmentation," *arXiv preprint arXiv:2405.01228*, 2024. [Cited on page 36.]

[75] Google Developers, "Machine learning crash course: Fairness," 2024. [Online]. Available: https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias [Cited on pages 38, 39, 40, 41, 78, and 79.]

[76] K. Drukker, W. Chen, J. Gichoya, N. Gruszauskas, J. Kalpathy-Cramer, S. Koyejo, K. Myers, R. C. Sá, B. Sahiner, H. Whitney, Z. Zhang, and M. Giger, "Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment," *Journal of Medical Imaging*, vol. 10, no. 6, p. 061104, 2023. [Cited on pages 38, 39, and 40.]

[77] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "Chexclusion: Fairness gaps in deep chest x-ray classifiers," in *Pacific Symposium on Biocomputing*, 2020. [Cited on page 42.]

[78] S. Cui, S. Ming, Y. Lin, F. Chen, Q. Shen, H. Li, G. Chen, X. Gong, and H. Wang, "Development and clinical application of deep learning model for lung nodules screening on ct images," *Scientific Reports*, vol. 10, no. 1, p. 13657, 2020. [Cited on page 43.]

[79] W. Hendrix, N. Hendrix, E. T. Scholten, M. Mourits, J. Trap-de Jong, S. Schalekamp, M. Korst, M. van Leuken, B. van Ginneken, M. Prokop *et al.*, "Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest ct scans," *npj Digital Medicine*, vol. 6, no. 1, p. 229, 2023. [Cited on page 43.]

[80] F. I. Tushar, A. Wang, L. Dahal, E. Samei, M. R. Harowicz, J. Kalpathy-Cramer, K. J. Lafata, T. D. Tailor, C. Rudin, and J. Y. Lo, "Reproducible benchmarking for lung nodule detection and malignancy classification across multiple low-dose ct datasets," *arXiv preprint arXiv:2405.04605*, 2024. [Cited on page 43.]

[81] M. M. Ashames, A. Demir, O. N. Gerek, M. Fidan, M. B. Gulmezoglu, S. Ergin, R. Edizkan, M. Koc, A. Barkana, and C. Calisir, "Are deep learning classification results obtained on ct scans fair and interpretable?" *Signal, Image and Video Processing*, pp. 1–11, 2024. [Cited on pages 43 and 95.]

[82] S. Gaur, M. Vitale, A. Hering, J. Kwisthout, C. Jacobs *et al.*, "Fairness evaluation of risk estimation models for lung cancer screening," *Machine Learning for Biomedical Imaging*, 2025. [Cited on pages 44 and 113.]

[83] "Peeters, D., Obreja, B., Antonissen, N., & Jacobs, C. (2025). The LUNA25 Challenge: Public Training and Development set - Imaging Data (1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14223624." [Cited on page 47.]

[84] “Peeters, D., Obreja, B., Antonissen, N., & Jacobs, C. (2025). The LUNA25 Challenge: Public Training and Development set - Annotation Data (1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14673658.”

[85] “Peeters, D., Obreja, B., Antonissen, N., & Jacobs, C. (2025). Benchmarking of Artificial Intelligence and Radiologists for Lung Cancer Screening in CT: The LUNA25 Challenge. Medical Image Computing and Computer Assisted Intervention 2025 (MICCAI). Zenodo. https://doi.org/10.5281/zenodo.15094631.” [Cited on page 47.]

[86] A. Wang, F. I. Tushar, M. R. Harowicz, K. J. Lafata, T. D. Tailor, and J. Y. Lo, “Duke Lung Cancer Screening Dataset 2024,” Mar. 2024. [Cited on page 53.]

[87] A. J. Wang, F. I. Tushar, M. R. Harowicz, B. C. Tong, K. J. Lafata, T. D. Tailor, and J. Y. Lo, “The Duke Lung Cancer Screening (DLCS) Dataset: A Reference Dataset of Annotated Low-Dose Screening Thoracic CT,” *Radiology: Artificial Intelligence*, vol. 7, no. 4, p. e240248, Jul. 2025.

[88] F. I. Tushar, A. Wang, L. Dahal, E. Samei, M. R. Harowicz, J. Kalpathy-Cramer, K. J. Lafata, T. D. Tailor, C. Rudin, and J. Y. Lo, “Reproducible Benchmarking for Lung Nodule Detection and Malignancy Classification Across Multiple Low-Dose CT Datasets,” Feb. 2026, arXiv:2405.04605 [cs]. [Cited on page 53.]

[89] E. M. Rodrigues, M. Gouveia, H. P. Oliveira, and T. Pereira, “Efficient-proto-caps: A parameter-efficient and interpretable capsule network for lung nodule characterization,” *IEEE Access*, vol. 13, pp. 56 616–56 630, 2025. [Cited on page 59.]

[90] E. K. V. I. I. A. Buslaev, A. Parinov and A. A. Kalinin, “Albumentations: fast and flexible image augmentations,” *ArXiv e-prints*, 2018. [Cited on page 64.]

[91] M. Gouveia, J. Araújo, H. P. Oliveira, and T. Pereira, “Domain-specific data augmentation for lung nodule malignancy classification,” in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025, pp. 1–7. [Cited on page 64.]

[92] L. Zedda, A. Loddo, and C. D. Ruberto, “Radio dino: A foundation model for advanced radiomics and ai-driven medical imaging analysis,” *Computers in Biology and Medicine*, vol. 195, p. 110583, 2025. [Cited on pages 66 and 67.]

[93] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *Proceedings of*

the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660. [Cited on page 68.]

[94] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Soyer, P. Goyal, Q. Fernandez, A. Seba *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023. [Cited on page 68.]

[95] R. Wightman, "Pytorch image models," https://github.com/rwightman/pytorch-image-models, 2019. [Cited on page 68.]

[96] C. DiCiccio, S. Vasudevan, K. Basu, K. Kenthapadi, and D. Agarwal, "Evaluating fairness using permutation tests," 2020. [Online]. Available: https://arxiv.org/abs/2007.05124 [Cited on page 79.]

[97] S. Wastvedt, J. D. Huling, and J. Wolfson, "An intersectional framework for counterfactual fairness in risk prediction," *Biostatistics*, vol. 25, no. 3, p. 702–717, Aug. 2023. [Online]. Available: http://dx.doi.org/10.1093/biostatistics/kxad021 [Cited on page 79.]

[98] N. Sourlos, G. Pelgrim, H. J. Wisselink, X. Yang, G. de Jonge, M. Rook, M. Prokop, G. Sidorenkov, M. van Tuinen, R. Vliegenthart, and P. M. A. van Ooijen, "Effect of emphysema on ai software and human reader performance in lung nodule detection from low-dose chest ct," *European Radiology Experimental*, vol. 8, 2024. [Cited on page 116.]

[99] N. Sourlos, J. Wang, Y. Nagaraj, P. {van Ooijen}, and R. Vliegenthart, "Possible bias in supervised deep learning algorithms for ct lung nodule detection and classification," *Cancers*, vol. 14, no. 16, Aug. 2022, funding Information: The position of the first author is supported by a research grant from the Netherlands Organisation for Scientific Research/Innovative Medical Device Initiative. Publisher Copyright: © 2022 by the authors.

[100] D. Penha, E. Pinto, B. Hochhegger, C. Monaghan, E. Marchiori, L. Taborda-Barata, and K. Irion, "The impact of lung parenchyma attenuation on nodule volumetry in lung cancer screening," *Insights into Imaging*, vol. 12, no. 1, p. 84, 2021. [Cited on page 116.]