FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

AI-based Models for Lung Cancer Characterization: a Multimodal Approach

Maria Sofia Brandão Porto Carvalho Gonçalves

DISSERTAÇÃO



Mestrado em Engenharia Informática e Computação

Supervisor: Tania Maria Pereira Lopes

Co-Supervisor: Hélder Filipe Pinto de Oliveira

Co-Supervisor: Joana Vale Amaro de Sousa

July 30, 2025

AI-based Models for Lung Cancer Characterization: a Multimodal Approach

Maria Sofia Brandão Porto Carvalho Gonçalves

Mestrado em Engenharia Informática e Computação

Approved by . . . :

President: Henrique Lopes Cardoso

Referee: Tomé Albuquerque

Supervisor: Tania Maria Pereira Lopes

July 30, 2025

Resumo

O cancro do pulmão é um dos cancros mais mortais do mundo, responsável por aproximadamente 1,8 milhões de mortes anualmente. A elevada taxa de mortalidade deve-se, em grande parte, ao diagnóstico tardio, o que reforça a necessidade de deteção precoce. A tomografia computorizada de baixa dose demonstrou ser eficaz para este fim. No entanto, a sua interpretação manual pode ser demorada, subjetiva e limitada quando utilizada isoladamente. Os dados clínicos estruturados, por si só, podem não fornecer informações espaciais e morfológicas essenciais para um diagnóstico preciso. Uma abordagem multimodal que combine dados imagiológicos e clínicos pode captar padrões complementares, melhorando, em última análise, a precisão diagnóstica para além do que cada método pode alcançar individualmente.

Esta tese centra-se no desenvolvimento de modelos multimodais baseados em IA que combinam imagens de tomografia computorizada torácica com dados clínicos estruturados para auxiliar na classificação binária de nódulos pulmonares como cancerosos ou não cancerosos. A abordagem proposta utiliza dados do National Lung Screening Trial (NLST).

Para a modalidade clínica foram avaliados vários modelos de *machine learning*, incluindo *logistic regression*, *random forest*, *gradient boosting algorithms (LightGBM* e *XGBoost)* e *multilayer perceptrons* (MLPs). Estes modelos passaram por uma extensa otimização de hiperparâmetros para melhorar o seu desempenho. Na modalidade de imagem, foram utilizadas as redes neuronais convolucionais ResNet-18 e ResNet-34, sendo testadas com e sem aumento de dados para identificar as configurações mais eficazes. As representações unimodais foram combinadas utilizando estratégias de *intermediate fusion* e *late fusion*. As experiências revelaram que os modelos multimodais superaram consistentemente as *baselines* unimodais, alcançando uma área sob a curva ROC (AUC) de 0.9138, uma precisão de 0.8424 e um *score* F1 de 0.8422 na melhor configuração. Estes resultados confirmam o valor adicional de diagnóstico da integração de dados clínicos e de imagem estruturados, sublinhando o potencial das abordagens de aprendizagem multimodal na melhoria da oncologia de precisão.

Abstract

Lung cancer is one of the deadliest cancers worldwide, responsible for approximately 1.8 million deaths annually. The high mortality rate is largely due to late-stage diagnoses, highlighting the need for early detection. Low-dose computed tomography (CT) has been shown to be effective for this purpose. However, its manual interpretation can be time-consuming, subjective and limited when used in isolation. Structured clinical data on its own may not provide essential spatial and morphological information needed for accurate diagnosis. A multimodal approach that combines both imaging and clinical data can capture complementary patterns, ultimately improving diagnostic accuracy beyond what each method can achieve individually.

This thesis focuses on developing AI-based multimodal models that combine chest CT imaging with structured clinical data to assist in the binary classification of lung nodules as either cancerous or non-cancerous. The proposed approach utilizes data from the National Lung Screening Trial (NLST).

For the clinical modality, a variety of machine learning models were evaluated, including logistic regression, random forests, gradient boosting algorithms (LightGBM and XGBoost), and multilayer perceptrons (MLPs). These models underwent extensive hyperparameter optimization to enhance their performance. In the imaging modality, ResNet18 and ResNet34 convolutional neural networks were used, testing both with and without data augmentation to identify the most effective settings. The unimodal representations were combined using both intermediate and late fusion strategies. The experiments revealed that the multimodal models consistently outperformed unimodal baselines, achieving an area under the ROC curve (AUC) of 0.9138, an accuracy of 0.8424, and an F1-score of 0.8422 in the best configuration. These results confirm the additional diagnostic value of integrating structured clinical and imaging data, underscoring the potential of multimodal learning approaches in enhancing precision oncology.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework to achieve a better and more sustainable future for all. It includes 17 goals to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

This thesis contributes to this mission by addressing healthcare innovation through the use of AI models for lung cancer detection using multimodal data. The approach integrates imaging and clinical information, aiming to enhance diagnostic accuracy and support timely medical interventions.

Lung cancer is a major contributor to global morbidity and mortality, often diagnosed at advanced stages due to limitations in current diagnostic procedures. By improving early detection, this work supports global efforts to reduce preventable deaths, enhance healthcare infrastructure, and promote innovation in medical technologies. The combination of low-dose CT and structured clinical data into AI frameworks contributes directly to the development of intelligent, reliable, and interpretable decision-support systems in clinical environments.

The specific Sustainable Development Goals addressed by this thesis are:

SDG 3: Ensure healthy lives and promote well-being for all at all ages.

SDG 9: Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation.

SDG	Target	Contribution	Performance Indicators and Metrics
3	3.4	Development of AI-based models contributes to reducing premature mortality from non-communicable diseases through early cancer detection.	Improvements in classification accuracy, AUC, and F1-score in binary lung cancer prediction.
	3.d	Reinforces diagnostic capacity and preparedness in healthcare systems through intelligent data integration and predictive modeling.	Integration of imaging and clinical data; enhanced decision support capabilities.
9	9.5	Encourages scientific research and strengthens technological capabilities in medical diagnostics via AI and machine learning.	Deployment and benchmarking of multimodal fusion strategies and architectures.
	9.b	Supports the development of high-end technology in the health sector, enabling scalable and interpretable solutions for cancer screening.	Reproducibility of multimodal pipelines; extensibility to other cancers and modalities.

Acknowledgements

Primeiro de tudo, gostaria de agradecer à Professora Tania Lopes por todo o apoio e acompanhamento que me deu ao longo da elaboração desta tese. Agradeço também ao Professor Hélder Oliveira pela disponibilidade neste processo. Um obrigada do fundo do coração à Joana Sousa, pela (muita) paciência, simpatia, compreensão, humanidade, por me rever tudo o que escrevi vezes sem conta e por me chamar à atenção quando precisei. Consegui fazer esta tese graças a toda a dedicação da Joana. Obrigada também à Margarida Gouveia pelos valiosos inputs durante as reuniões. Muito obrigada à Maria Inês Amaro por toda a paciência que também teve para mim e toda a ajuda preciosa que me deu, tornando todo este processo mais feliz. Com esta tese, acaba o meu percurso académico. Tenho a sorte de ter ótimos amigos que me acompanham desde o início, outros fui fazendo pelo caminho. Não poderia deixar de mencionar as minhas amigas de sempre e para sempre: a minha abécula Ana Beatriz, a Ana e a Ribeiro. Gostaria de dedicar este culminar dos meus anos de aluna à minha família e madrinha Carla. Ao Vô Mário, que nasceu para ser avô. Ao Vô Nelo, que tinha um orgulho gigante em todos os netos (e que teria adorado ver-me a acabar o curso), à Vó Luísa, a minha eterna companheira de chamadas à noite e à Vó Dete, que me ensinou a ser uma borboleta social. Obrigada a todos os tios e primos, em especial ao Tio Paulo, que nunca nos deixou sentir sozinhas. Aos meus amigos da faculdade: foram das melhores coisas que me aconteceu na vida. Ao Ávila, ao Dani, ao Dinis, à Lara, à Matilde, ao Miguel, ao Pedro, ao Fonseca, à Sofia e ao Tim. Aceitamo-nos uns aos outros exatamente como somos, sem tirar nem pôr. O quão bom e raro isso é. A muitos mais verões em Arouca. Um especial obrigada ao Dani. Foi graças à tua companhia diária que fiz o curso. Obrigada, Beatriz, por todo o apoio que me deste desde que nasci. Um forte aperto de mão com respeito ao Nuno por entrar agora nesta aventura. Ao André, por me ter mostrado que estava a ver a vida a preto e branco até ele entrar nela e dar-lhe cor. Estás sempre ao meu lado, a dar-me a mão, em todas as fases da vida (e para todas as que vierem). A Feup deu-me amigos, um curso e mais, mas a melhor coisa que me deu foi, de longe, tu. Ao meu pai, Luís, que não me pôde ver sequer a acabar o 8º ano, mas que, nos poucos anos que tivemos juntos, me ensinou a como encarar a vida (e até a morte). É graças a ele que tenho ambição, que adoro ler e um milhão de outras coisas que me definem. É fantástico como me continua a fazer aperceber e apreciar tanta coisa mesmo passados tantos anos. Foi ao exemplo dele que me agarrei sempre que senti que precisava de força. Uma das coisas que ele mais pena tinha era não saber que curso eu iria tirar. Sei que, esteja ele onde estiver, estará muito orgulhoso por eu ser engenheira. E, finalmente, a ti, Mamã. Teria de escrever toda uma tese sobre ti. Como posso pôr em algumas palavras tudo o que tu és? Absolutamente tudo o que tu fazes é por mim e pela Beatriz. És quem mais sofre comigo as minhas dores e quem mais rejubila com os meus sucessos. Na vida só há certeza de 3 coisas: a morte, os impostos e que tu és a melhor mãe do mundo. Tudo o que fazes surpreende-me. Ao longo dos anos, usei grandes personalidades como modelos a seguir. Agora, apercebo-me que a melhor pessoa a ter como exemplo foi a que sempre viveu comigo e me deu a vida. Tudo o que fiz, faço e farei, é graças a ti.

This work is co-financed by Component 5 - Capitalization and Business Innovation, integrated

in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project HfPT, with reference 41."This work is co-financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project HfPT, with reference 41.

SOFIA

"How lucky am I to have something that makes saying goodbye so hard."

Winnie the Pooh

Contents

Acknowledgements v

1 Introduction 1
- 1.1 Context . . . 1
- 1.2 Motivation . . . 2
- 1.3 Problem Description . . . 2
- 1.4 Research Questions . . . 3
- 1.5 Hypothesis . . . 3
- 1.6 Contributions . . . 3
- 1.7 Structure . . . 4

2 Background 5
- 2.0.1 Lung Anatomy and Physiology . . . 5
- 2.0.2 Epidemiology and Histology of Lung Cancer . . . 5
- 2.0.3 Major Risk Factors . . . 7
- 2.0.4 Chest CT Imaging in Lung Cancer . . . 7
- 2.0.5 Structured Clinical Data in Multimodal Studies . . . 7

3 State of the Art 8
- 3.1 Multimodal Data Fusion . . . 8
- 3.2 Advanced Image-Based Deep Learning for Diagnosis . . . 20
- 3.3 Feature Extraction and Fusion Techniques . . . 23
- 3.4 Temporal and Longitudinal Data Integration . . . 27
- 3.5 Data Challenges in Multimodal Systems . . . 29
- 3.6 Comparative Studies on Multimodal Approaches . . . 30
- 3.7 Summary . . . 32

4 Data 33
- 4.1 National Lung Screening Trial . . . 33
- 4.2 Data selection . . . 34
- 4.3 Imaging data pre-processing . . . 35
  - 4.3.1 Choice of DICOM series . . . 35
  - 4.3.2 Resampling . . . 38
  - 4.3.3 Normalization . . . 39
  - 4.3.4 Trimming and padding . . . 39
  - 4.3.5 Segmentation . . . 39
  - 4.3.6 Manual mask correction . . . 41
  - 4.3.7 Standardization of masks . . . 41

4.4 Clinical data pre-processing . . . 42
4.4.1 Feature selection . . . 43
4.4.2 Missing data handling, normalization and scaling . . . 43
4.4.3 Dimensionality reduction . . . 44

5 Single and Multimodal Approaches **46**
5.1 Single-Modality Models . . . 46
5.1.1 Imaging data . . . 46
5.1.2 Clinical data baseline . . . 50
5.2 Multimodal Fusion Models . . . 51
5.2.1 Imaging Feature Vectors . . . 51
5.2.2 Fusion Strategies . . . 52

6 Results and Discussion **56**
6.1 Single–Modality Baselines . . . 56
6.1.1 Clinical Data . . . 56
6.1.2 Imaging Data . . . 57
6.2 Multi-modality Baselines . . . 59
6.2.1 Intermediate Fusion . . . 59
6.2.2 Late Fusion . . . 61
6.3 Summary of Best Configurations . . . 63

7 Conclusions and Future Work **64**
7.1 Conclusions . . . 64
7.1.1 Study Limitations . . . 65
7.1.2 Future Work . . . 65
7.1.3 Final Remarks . . . 66

References **67**

A Appendix **73**
A.1 Columns Kept . . . 73
A.2 Columns After Clinical Data Preprocessing . . . 74
A.3 Clinical Data Hyperparameter tuning . . . 75

List of Figures

2.1 Lung Anatomy . . . 6
2.2 Global lung cancer incidence, mortality, and 5-year prevalence by continent. . . . 6

3.1 Proposed Multimodal Fusion Deep Neural Network (MFDNN). . . . 9
3.2 Classification results for the CT Scans . . . 11
3.3 Proposed feature-based framework for lung cancer classification . . . 13
3.4 LungNet ARchitecture . . . 14
3.5 AUC, sensitivity and MCC results . . . 16
3.6 Framework of a deep learning-based system for lung nodule classification and detection, including preprocessing, voxel coordinate conversion, image rescaling, data augmentation, patch extraction, and training/testing of deep learning models 17
3.7 Performance results . . . 18
3.8 3D CNN with Visual Insights . . . 18
3.9 Proposed architecture . . . 19
3.10 Proposed MIXCAPS architecture . . . 22
3.11 Imbalanced Data for Lung Cancer Classification . . . 24
3.12 Lung Cancer Classification . . . 26
3.13 Lung Nodule Malignancy Prediction . . . 29
3.14 Hybrid Model for Lung Cancer . . . 30

4.1 CT scans showing cancerous and non-cancerous lung nodules. . . . 35
4.2 Patient selection flowchart from the NLST dataset for imaging analysis. This diagram illustrates the step-by-step inclusion criteria applied to identify eligible CT scan series, distinguishing between patients diagnosed and not diagnosed with lung cancer, and filtering based on radiological availability, abnormality matching, and image presence in the NLST dataset. . . . 36
4.3 Scan pre-processing pipeline, using slice thickness resizing, Hounsfield unit normalization and slice trimming/padding to achieve standard slice depth. . . . 37
4.4 Example of an axial slice of CT scan. . . . 40
4.5 Lung CT Scan with segmentation mask. . . . 40
4.6 Lung CT scans showing the segmentation mask before and after manual correction. An arrow points to the lung nodule that has not been correctly automatically segmented. . . . 41
4.7 Clinical Data Pipeline, showing column choice, missing data handling, normalization, scaling and dimensionality reduction. . . . 45

5.1 Visual impact of each augmentation applied independently to the same axial CT slice. The original image (a) is followed by six deterministic transformations used. 50
5.2 Intermediate fusion pipeline, showing the two different approaches. . . . 53

5.3 Late fusion pipeline, showing the two different weight strategies. 54

List of Tables

4.1 Summary of patient and CT scan selection from the NLST dataset. 34
4.2 Kernel preference hierarchy used in Step 2 (lower rank → higher priority). 38

5.1 Grid search configuration for CNN hyperparameters. Each combination was tested across folds to identify the most effective settings for classification performance. 47

6.1 Performance of the best clinical baseline using the Random Forest model. 57
6.2 Grid of hyperparameter combinations evaluated for single-modality CNN models (optimizer fixed as Adam). 57
6.3 Optimal hyperparameter combinations per architecture (excluding optimizer, learning rate, and epochs). 58
6.4 Performance metrics for the best-performing configuration of each single-modality model, averaged across the three cross-validation folds. 59
6.5 Intermediate fusion – mean ± std. Best values in bold. 60
6.6 Late fusion: mean ± std over the three validation folds. 62
6.7 Mean AUC of the best pipelines (single-modality and multimodal). 63

A.1 Outer–fold performance of every hyper-parameter configuration (mean ± SD. . . 76

Abbreviations and Symbols

3D CNN	3D Convolutional Neural Network
AFF	Atypical Femur Fractures
AI	Artificial Intelligence
ALK	Anaplastic Lymphoma Kinase
AUC	Area Under the Curve
CAN	Convolutional Attention Network
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
COPD	Chronic Obstructive Pulmonary Disease
CT Scan	Computed Tomography Scan
DCNN	Deep Convolutional Neural Network
DCT	Discrete Cosine Transform
DFFT	Discrete Fast Fourier Transform
DWT	Discrete Wavelet Transform
EBT	Ensemble Bagging Tree
EHR	Electronic Health Records
GCmIW	Gamma Correction Max Intensity Weights
GDPR	General Data Protection Regulation
Grad-CAM	Gradient-weighted Class Activation Mapping
HGLCM	Haralick Gray Level Co-occurrence Matrix
IARC	International Agency for Research on Cancer
ISUP	International Society of Urological Pathology
LASSO	Least Absolute Shrinkage and Selection Operator
LBP	Local Binary Patterns
LDCT	Low-dose Computed Tomography
MC-ECNN	Multi-Channel Ensemble Convolutional Neural Network
MFDNN	Multimodal Fusion Deep Neural Network
MKL	Multiple Kernel Learning
MLP	Multilayer Perceptron
MSER	Maximally Stable Extremal Region
NFF	Normal Femur Fractures
NLST	National Lung Screening Trial
NRRD	Nearly Raw Raster Data
NSCLC	Non-small Cell Lung Cancer
PCA	Principal Component Analysis
PID	Patient Identifier
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RCC	Renal Cell Carcinoma
SCCA	Serial Canonical Correlation Analysis
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TCIA	The Cancer Imaging Archive
WNCA	Weighted Neighborhood Component Analysis

Chapter 1

Introduction

1.1 Context

There are many factors that can contribute to lung cancer development. The most significant ones are: smoking (and inhaling passive smoke), exposure to radon and asbestos, air pollution and family history of lung cancer [1].

It is estimated that one in five people will develop some form of cancer in their lifetime [2]. Lung cancer is the second most common type of cancer in the world, and it is predicted that around 1.8 million people die from this disease each year [3].

Imaging techniques like chest X-rays and Computed Tomography (CT) scans, which can identify abnormal masses or nodules in the lungs, are commonly used in the identification of lung cancer. CT scans offer more detailed images and can detect smaller nodules or abnormalities that X-rays might overlook. Therefore, they are generally preferred over standard chest X-rays for lung cancer detection, particularly in high-risk individuals. Given that low-dose CT scans are more successful at early detection in at-risk populations and expose less radiation than regular CT scans, they are particularly useful for screening [4].

A biopsy, in which tissue samples are obtained from the lung to check for the presence of cancer cells, may be done if the CT scans reveal areas of concern. Blood tests and sputum cytology, which examines mucus, can also aid in the diagnosis [5].

With advancements in medical technology, cancer research is at the forefront of innovation, introducing new tools for the treatment and diagnosis of the disease. Additionally, due to the high prevalence and mortality rates associated with lung cancer, it has become one of the most closely monitored and rapidly evolving types of cancer in this revolution [6]. Thus, currently, doctors have access to endless amounts of medical data. However, using all this information from different sources and formats in the best possible way can prove to be an arduous task, especially in an already frenetic hospital environment [7].

Even so, as promised in the Hippocratic Oath, doctors must always seek the best for their patients, and this is where computing and Artificial Intelligence (AI) come in hand. Sousa et al. [8] concluded that the type of data that revealed most important for a correct lung cancer

diagnosis is imaging data, since clinical features are not worth much when used alone because they are information that is complementary to the diagnosis. Despite this, a crucial conclusion was reached: a multimodal approach to lung cancer diagnosis is something that has a great potential, with its development being something that cannot, by all means, be overlooked. Therefore, there seems to be a great potential of success with the creation of models that combine both clinical and imaging data.

This approach will contribute to the development of more accurate AI-based tools for lung cancer diagnosis and treatment, supporting clinicians in making more informed decisions and ultimately improving patient outcomes, which is, ultimately, the main goal of Medicine.

1.2 Motivation

Lung cancer is a silent disease. This means that the illness's symptoms only begin to show up when the cancer is metastasized. This translates itself in a 5-year survival rate of approximately 15% to 6% in the later stages. If the disease is diagnosed in earlier stages, this 5-year survival rate is around 40% to 65% [9]. Many deaths could be prevented, since 90% of lung cancer cases in men and 70 to 80% cases in women are due to smoking [10].

Cancer is the twentieth century's disease [11], and this statement has not changed for the current century. In order to fight this illness, an ever growing investment in new technologies shall be made. While the use of AI and machine learning (ML) models in oncology is expanding, most of these models solely focus on prediction accuracy and don't explain the relationships between patient-specific characteristics and outcomes. Investigating a multimodal approach (integrating several kinds of patient data besides just imaging) is essential for improving cancer research and treatment. Although the majority of AI applications in cancer mostly employs imaging data, doctors frequently use a wide range of data, such as genetic profiles, test results, patient histories, and lifestyle factors in their day-to-day life. Physicians are able to comprehend each patient's condition completely due to this integration. Similar to this, a multimodal AI model might use a variety of data sources to improve diagnose accuracy and therapy tailoring while also potentially providing intricate, patient-specific insights. Researching this method may help close the gap between clinically useful decision-making and forecasts based on data. Ultimately, this thesis is centered on the binary classification of lung cancer using a multimodal approach. It aims to move beyond fragmented, data-fitting models by developing integrated frameworks that offer a more comprehensive understanding of the disease and its underlying pathological relationships.

1.3 Problem Description

All different kinds of clinical information gathered from medical records, such as lab data, comorbidities, medications, demographics, are all evaluated separately in current clinical procedures, suffering from fragmented data analysis [12]. This lack of integration makes it more difficult to

understand how complex lung cancer is and makes individualized treatment plans more difficult to implement.

Considering the major problem presented in the previous section concerning the lack of the AI models explainability, their practical application is therefore diminished by this lack of interpretability, since doctors need more than simply predictions [13]. Reasonable motives should be given to the physician on why a patient needs to have surgery or a certain treatment, since they might not operate or treat a patient based only on the output of a model.

With the above considerations in mind, this thesis is centered on the following objectives:

- **Development of AI baselines using multimodal data**: This work investigates the integration of imaging data and structured clinical records to construct predictive models for lung cancer diagnosis. Although various data types exist, this thesis focuses specifically on CT images and clinical variables from the National Lung Screening Trial.
- **Evaluation of multimodal fusion strategies**: Different architectural approaches (intermediate and late fusion) are tested to determine how best to combine imaging and clinical information for binary classification of lung nodules as cancerous or non-cancerous.

1.4 Research Questions

1. **RQ1**: What are the difficulties and opportunities of preprocessing and aligning multimodal data sources, such as imaging and medical records, so that they can be used effectively in models for cancer diagnosis?
2. **RQ2**: Which feature fusion methods work best in multimodal AI frameworks to integrate CT imaging and patient-specific data?
3. **RQ3**: How can multimodal approaches improve the performance of AI models for binary classification of cancer patients compared to unimodal approaches?

1.5 Hypothesis

Integrating multimodal information—specifically clinical records and CT imaging into AI models will lead to more accurate predictions for lung cancer diagnosis. This approach is expected to improve patient stratification and support earlier and more precise clinical decision-making.

1.6 Contributions

This dissertation improves lung cancer binary classification in several significant ways:

- It establishes a comprehensive preprocessing pipeline that integrates low-dose CT scans with structured clinical data from NLST.

- It creates robust single-modality baselines for both image-based models (ResNet-18 and ResNet-34) and clinical models (logistic regression, random forest, gradient boosting, and multilayer perceptron).
- It introduces two fusion approaches (intermediate and late fusion) that effectively combine image features with clinical information.
- It demonstrates that these multimodal models outperform those based on a single type of data.

These findings indicate that the integration of imaging and clinical data results in more accurate and reliable detection of lung cancer.

1.7 Structure

This document is constituted by seven chapters. Chapter 1 introduces the clinical and technological background of lung cancer screening, formulates the research questions and hypotheses, and highlights the need for an AI-driven multimodal approach. Chapter 2 gives a concise yet comprehensive overview of the biological, epidemiological, and technological foundations of lung cancer, covering lung anatomy, major risk factors, CT imaging, and the role of structured clinical data in multimodal research. Chapter 3 reviews previous work on image-based deep learning, clinical risk modeling, and data-fusion strategies, identifying the scientific gap that this study aims to address. Chapter 4 describes the NLST cohort, outlines the inclusion criteria, and explains the image and clinical preprocessing pipelines that generate harmonized inputs for modeling. Chapter 5 details the architecture and training protocols for the unimodal baselines, as well as the intermediate and late fusion systems that combine them. Chapter 6 presents the experimental results, compares the fusion strategies, analyzes error patterns, and discusses the clinical implications. Chapter 7 summarizes the main findings, outlines the study's limitations, and suggests directions for future research. An Appendix A provides supplementary variable lists.

Chapter 2

Background

Lung cancer is the most common and deadly type of cancer worldwide [14]. Understanding its biological foundations and diagnostic scenary is crucial for creating effective tools for early detection and treatment. This chapter offers a concise overview of lung anatomy, major epidemiological and histological patterns of lung cancer, key risk factors, the role of CT in clinical workflows, and the integration of structured clinical data in contemporary multimodal research.

2.0.1 Lung Anatomy and Physiology

The lungs occupy most of the thoracic cavity and are divided into five lobes: three on the right and two on the left [15], as shown in figure 2.1. They are further partitioned into 18 bronchopulmonary segments. Each segment ends in clusters of *alveoli*, which feature an approximately 0.5 µm air–blood barrier [16], allowing for rapid diffusion of oxygen into the blood and carbon dioxide out of the pulmonary capillary network. Two epithelial cell types are predominant: type I pneumocytes, which make up around 95% of the surface area and facilitate gas exchange, and type II pneumocytes, which secrete surfactant and help to repair damaged epithelium [17]. The lungs receive dual circulations: low-pressure pulmonary arteries for gas exchange and high-pressure bronchial arteries for supplying the lung tissue with nutrients, and are drained by an extensive lymphatic plexus that converges on the hilar and mediastinal lymph nodes. Resident immune cells, including alveolar macrophages, dendritic cells, and tissue-resident T cells, patrol this interface [18]. However, chronic exposure to inhaled toxins can overwhelm these defenses and initiate carcinogenesis.

2.0.2 Epidemiology and Histology of Lung Cancer

Lung cancer remains the leading cause of cancer mortality worldwide, with approximately 2.5 million new cases and 1.8 million deaths reported in 2022 [20]. The five-year overall survival rate is less than 20%, largely because over 50% of tumors are diagnosed at an advanced stage. Two broad histological classes are recognized: non-small-cell lung cancer (NSCLC), accounting for

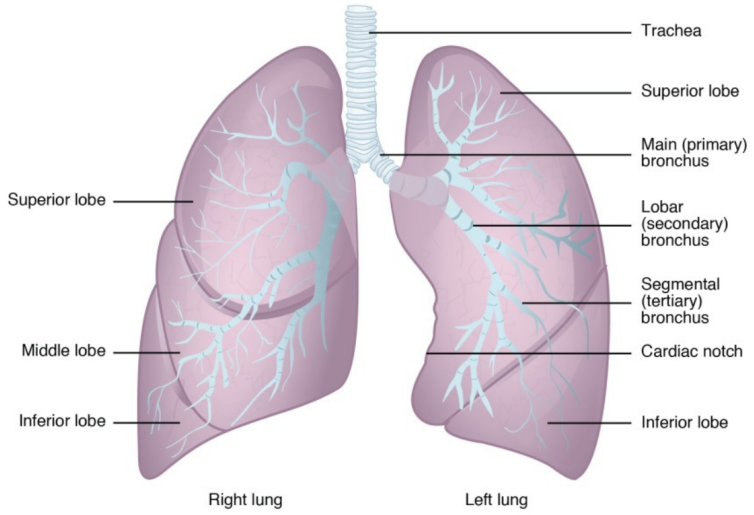


Figure 2.1: Lung anatomy, adapted from [19].

around 85% of cases, which includes adenocarcinoma, squamous cell carcinoma and large-cell carcinoma, and small-cell lung cancer (SCLC), making up about 15%.

The global distribution of lung cancer burden varies significantly by continent. As shown in Figure 2.2 [21], taken from International Agency for Research on Cancer (IARC)'s official site [20], Asia accounts for the majority of global incidence (63.1%), mortality (62.9%), and five-year prevalence (63.9%) of lung cancer, followed by Europe and North America. This geographic disparity reflects differences in population size, smoking prevalence, environmental exposures, and access to early detection and treatment services.

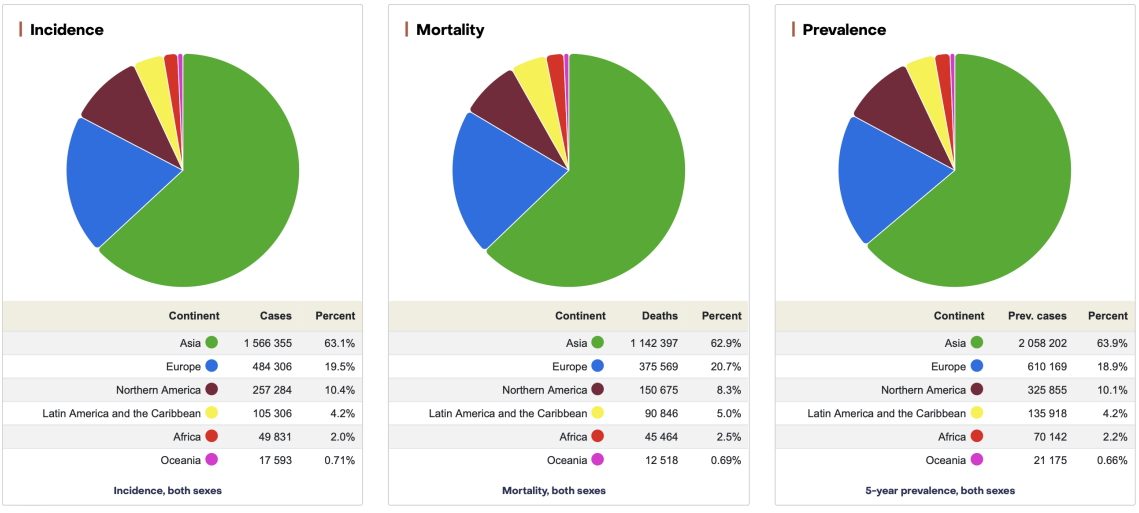


Figure 2.2: Global lung cancer incidence, mortality, and 5-year prevalence by continent (both sexes, 2022) [20].

2.0.3 Major Risk Factors

Cigarette smoking accounts for 80-90% of lung cancer cases. The risk increases with cumulative exposure and declines slowly after quitting, as shown in [22]. Other established carcinogens include residential radon, asbestos, crystalline silica, diesel exhaust, and ambient particulate matter. Host factors such as chronic obstructive pulmonary disease (COPD), pulmonary fibrosis, primary immunodeficiency, and germline mutations in DNA repair genes can influence individual susceptibility.

2.0.4 Chest CT Imaging in Lung Cancer

CT is a non-invasive medical imaging technique that uses a series of X-ray measurements taken from various angles to create cross-sectional images of the body. In the diagnosis of lung cancer, CT provides high-resolution visualization of pulmonary structures, enabling the detection of small nodules, masses, or tissue abnormalities that might be overlooked in conventional chest X-rays.

During a CT scan, the machine rotates around the patient, emitting X-rays that pass through the body. Detectors measure the amount of radiation absorbed in each slice, and this information is processed by computer algorithms to produce detailed 2D or 3D images. The resulting scan represents tissue densities in Hounsfield Units (HU), a standardized scale where air measures approximately -1000 HU, water is 0 HU, and dense bone is around +1000 HU [23]. Lung tissue, which is mostly air, appears dark (with low HU values), while denser tissues such as blood vessels, tumors, or fibrosis show lighter regions (with higher HU values) [24].

The contrast in a CT image arises from variations in tissue density. Malignant nodules often exhibit different attenuation characteristics, such as higher or more heterogeneous HU distributions, compared to benign nodules, allowing radiologists and AI models to distinguish between the two. Additionally, parameters such as nodule size, shape, edge definition, and growth patterns over time are key indicators used in diagnosis and risk stratification.

Low-dose CT (LDCT) is a variant specifically optimized for screening. It uses significantly reduced radiation while maintaining adequate image quality for nodule detection. LDCT screening significantly reduces lung cancer mortality—approximately 20% in the National Lung Screening Trial (NLST) [25, 26] and around 24% in the Dutch–Belgian NELSON trial [27].

2.0.5 Structured Clinical Data in Multimodal Studies

Imaging alone rarely captures the full biological complexity of lung cancer. Electronic health record (EHR) variables, such as demographics (age, sex, ethnicity), exposure metrics (pack-years, occupational hazards, radon exposure by postcode), comorbidities (COPD, interstitial lung disease, cardiovascular disease), and genomic drivers, actually provide very valuable complementary information [28]. Linking longitudinal clinical data with CT features can strengthen prognostic and predictive models, but it also demands careful management of missing data, class imbalances, time-related biases, and strict adherence to ethics guidelines.

Chapter 3

State of the Art

This chapter provides an overview of the State of the Art in AI-based models for lung cancer characterization, organized into key methodological categories. It begins with Multimodal Data Fusion (Section 3.1), exploring techniques that integrate imaging, clinical, and molecular data. Next, Advanced Image-Based Deep Learning for Diagnosis (Section 3.2) highlights imaging-driven approaches leveraging cutting-edge deep learning architectures. The chapter also examines Feature Extraction and Fusion Techniques (Section 3.3) and approaches leveraging Temporal and Longitudinal Data Integration (Section 3.4). Finally, it discusses Data Challenges in Multimodal Systems (Section 3.5), concluding with insights from Comparative Studies on Multimodal Approaches (Section 3.6).

3.1 Multimodal Data Fusion

In order to improve the precision and reliability of lung cancer diagnosis and prognosis, this section examines strategies that combine many data modalities, including imaging, clinical, and molecular data.

Deep Fusion Network for Lung Cancer Classification

In the paper "An enhanced multimodal fusion deep learning neural network for lung cancer classification" [29], the authors define that a Multimodal Fusion Deep Neural Network (MFDNN) architecture would be good in integrating information from different modalities.

In this research, data was integrated from multiple sources (The Cancer Imaging Archive and The Cancer Genome Atlas). The data was integrated using consistent patient identifiers, with missing values handled via imputation or sample removal, and outliers managed statistically. Features from medical images are extracted (e.g., texture, shape) and normalized (e.g., min-max scaling), while dimensionality reduction (PCA, t-SNE) and encoding methods (e.g., one-hot encoding for categorical variables) are applied to ensure data consistency and usability.

The proposed network extracts features from lung cancer images at various scales using CNNs and combines them through mechanisms such as attention layers, gating mechanisms, or straightforward concatenation.

To fuse this information from multiple sources (consisting of medical images, genomics and clinical data, all in different levels of granularity, according to the authors), the modalities are combined, using a concatenation operation. Then, a neural network is applied to the fused features so that it learns a joint representation. After that, a classification layer calculates the class probability of being cancerous or not. These probabilities are usually obtained through a softmax activation function, and the binary cross-entropy loss is employed to calculate binary classification. In order to minimize the loss function, backpropagation and optimization techniques are used to update the network's weights. Finally, the probability of an image having a cancerous node or not is computed during inference.

Figure 3.1 shows the MFDNN proposed by the authors. It consists on a set of mathematical operations, neural network architecture details and the choice of fusion mechanisms.

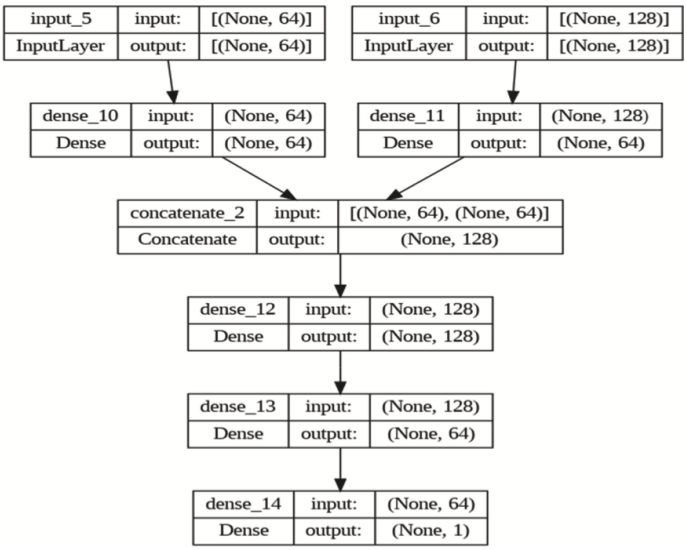


Figure 3.1: Proposed Multimodal Fusion Deep Neural Network (MFDNN) used by the authors in [29].

The experimental results show that the overall accuracy of the MFDNN for classifying correctly lung cancer images regarding if they are, in fact, cancerous or not, is 92,5%. MFDNN achieves a precision of 87.4%, a recall of 86.4%, and an F1-score of 86.2%, reflecting its ability to balance accuracy and sensitivity in cancer diagnosis, which is crucial for minimizing false positives and false negatives. This means that the MFDNN is actually a great way to improve early lung cancer diagnosis, thus decreasing the mortality rate of those who are affected by this disease.

Regarding its limitations, the authors bring the reader's attention to the fact that it is of major importance to emphasize the urgent need for thorough validation and regulatory guidelines, as well as the ethical issues surrounding the use of AI in medical contexts.

For the research at hand, this information is of uttermost importance, as it is in accordance with the goal of this thesis: to integrate data from various modalities, enabling a more accurate prediction if a patient has lung cancer or not. The fact that MFDNN have such good results makes it mandatory to keep this strategy in mind when considering what the best approach might be to predict lung cancer.

Heterogeneous Feature Learning for Nodule Diagnosis

Traditional lung nodule classifications often focus solely on image features, ignoring valuable heterogeneous clinical data that can significantly improve diagnostic performance. This problem is addressed in "Pulmonary Nodule Classification Based on Heterogeneous Features Learning" [30].

The researchers proposed a 3D-ResNet for extracting deep image features from CT scans and used a Support Vector Machine (SVM) with Multiple Kernel Learning (MKL) to combine image features with patient clinical data. This approach takes advantage of the complementary strengths of both types of features for the classification of nodule malignancies.

The study used two data sets: the LIDC-IDRI dataset with 1,018 annotated cases for pre-training and a private dataset with 14.415 CT images and clinical information from 61 patients for fine-tuning. The combined model achieved an accuracy of 89.72% on the private dataset, significantly outperforming models using only image features.

The methodology involves using a 34-layer 3D-ResNet to extract deep image features from CT scans, pretrained on the LIDC-IDRI dataset and fine-tuned on a private dataset for nodule classification. Clinical data, including patient attributes such as age and smoking history, were encoded and combined with the image features using a SVM with MKL. The MKL approach utilized Radial Basis Function (RBF) kernels for image features and polynomial kernels for clinical features, effectively integrating heterogeneous data. The fused features were then used by the SVM to classify the nodules as benign or malignant, using the complementary strengths of deep image and clinical data for improved diagnostic performance.

The 34-layer 3D-ResNet utilized residual connections for efficient feature extraction, while the SVM with MKL handled the fusion of heterogeneous features.

Fusion was performed at the feature level by combining the deep image features extracted from the 3D-ResNet with clinical features using MKL, which optimally weighted different kernels for feature integration.

The combination of image features and clinical data significantly improved classification accuracy, demonstrating the value of heterogeneous feature learning for pulmonary nodule diagnosis. These results are shown in the figure 3.2.

The private dataset was relatively small, and the clinical features were manually extracted, limiting automation and scalability.

Future efforts shall focus on expanding the dataset, automating clinical feature extraction, and developing an end-to-end nodule diagnosis system integrating detection, feature extraction, and

Methods	LIDC-IDRI			Private Data Set		
	Accuracy(%)	Sensitivity(%)	Specification(%)	Accuracy(%)	Sensitivity(%)	Specification(%)
DNN	79.01	80.19	78.03	62.35	76.47	41.18
SAE	76.36	76.41	76.31	58.82	68.63	44.12
CNN	77.08	71.54	81.71	67.06	70.59	61.76
MC-CNN	87.14	77.00	93.00	-	-	-
3D-CNN(Ours)	**91.29**	**91.01**	91.40	**84.70**	**86.05**	83.33
ResNet-34(Ours)	89.68	75.28	**95.47**	83.52	81.39	**85.71**

Figure 3.2: Classification results for the CT Scans in the study[30].

classification.

This study is useful because it demonstrates the effectiveness of integrating image and clinical data using deep learning and feature fusion techniques.

ALK Status Prediction via Clinical and Imaging Data

The study "Predicting anaplastic lymphoma kinase rearrangement status in patients with non-small cell lung cancer using a machine learning algorithm that combines clinical features and CT images" [31] focuses on predicting Anaplastic Lymphoma Kinase (ALK) rearrangement status in NSCLC patients, addressing the need for a non-invasive method that combines clinical features and CT imaging. Traditional methods often rely on invasive biopsies and are limited in capturing tumor heterogeneity, highlighting the need for improved diagnostic tools.

The researchers developed a machine learning framework combining radiomic features extracted from CT scans and clinical data to predict ALK rearrangement status.

The study demonstrated that combining clinical and radiomic features significantly improved predictive performance for ALK rearrangement status. In the primary cohort, the combined model achieved an accuracy of 94.3%, an AUC of 0.965 (95% CI: 0.826–0.882), a precision of 98.5%, an F1 score of 91.1%, and a recall of 85.1%. In the validation cohort, the combined model achieved an accuracy of 84.9%, an AUC of 0.914 (95% CI: 0.804–0.891), a precision of 93.2%, an F1 score of 74.7%, and a recall of 63.0%. These results consistently outperformed models trained on clinical features alone (AUCs of 0.805 and 0.735) or radiomic features alone (AUCs of 0.953 and 0.890), highlighting the efficacy of feature fusion in enhancing diagnostic accuracy.

Radiomic features were extracted from segmented CT tumor regions using PyRadiomics, capturing texture, histogram, and local binary pattern characteristics. Clinical features such as age, smoking history, and tumor stage were included. Feature selection was performed using the Least Absolute Shrinkage and Selection Operator (LASSO) technique with 10-fold cross-validation. The SVM classifier fused clinical and radiomic features at the feature level, leveraging a combination of kernels to optimize classification performance.

The approach utilized a classical ML pipeline rather than deep learning. Radiomic features were combined with clinical data through kernel-based fusion in the SVM model, which offered flexibility and interpretability.

Feature-level fusion was implemented by concatenating selected clinical and radiomic features into a single vector for SVM classification. This allowed the model to leverage complementary

information from both data types.

The combined model demonstrated superior predictive performance, suggesting that integrating radiomic and clinical features can serve as a reliable non-invasive approach for detecting ALK rearrangements in NSCLC.

The study was limited by a small dataset and lack of external validation, which may impact the model's generalizability. Additionally, the manual segmentation of CT images could limit scalability.

Future studies will aim to validate the model on larger and external datasets, automate the segmentation process, and explore deep learning approaches for improved accuracy and efficiency.

Multimodal Prognosis Modeling in Renal Cancer

The study "Multimodal deep learning for personalized renal cell carcinoma prognosis: Integrating CT imaging and clinical data" [32] addresses the challenge of predicting renal cell carcinoma (RCC) survival outcomes by integrating clinical and CT imaging data using a multimodal deep-learning approach.

The researchers proposed a methodology which includes three modules. Starting with CT image feature extraction using a modified 3D EfficientNet-B7 CNN, trained to classify RCC tumors based on International Society of Urological Pathology (ISUP) grades, followed by clinical variable selection using Spearman correlation and random forest importance scores to identify the most relevant variables. Finally, survival modeling through a discrete-time network with three fully connected layers, trained with LogisticHazard loss to predict survival probabilities over 15 discrete intervals. The model, shown in figure 3.3 integrates radiomic features and selected clinical data through feature-level fusion to improve survival prediction accuracy.

The framework was validated using the KiTS21 dataset, which includes 244 RCC patients with clinical and CT imaging data. The best experimental configuration achieved a C-index of 0.84, an AUC of 0.80, and an Integrated Brier Score (IBS) of 0.18, outperforming prior methods in RCC prognosis modeling.

The multimodal approach demonstrated the value of integrating CT imaging features with selected clinical data for accurate survival predictions. The personalized, non-proportional survival curves generated by the model have potential applications in optimizing RCC treatment strategies.

The reliance on a limited dataset with high censoring and the need for annotated segmentation for feature extraction may hinder generalizability and scalability. Furthermore, the inclusion of multiple clinical variables increases data dependency for accurate predictions.

Future work includes merging tumor grading and survival prediction into a unified framework, automating feature extraction to reduce dependence on manual annotations, and expanding to datasets with diverse modalities such as genomic and histopathological data.

This study is important, as it effectively demonstrates a multimodal deep-learning approach combining CT imaging and clinical data for individualized survival predictions. Its insights into

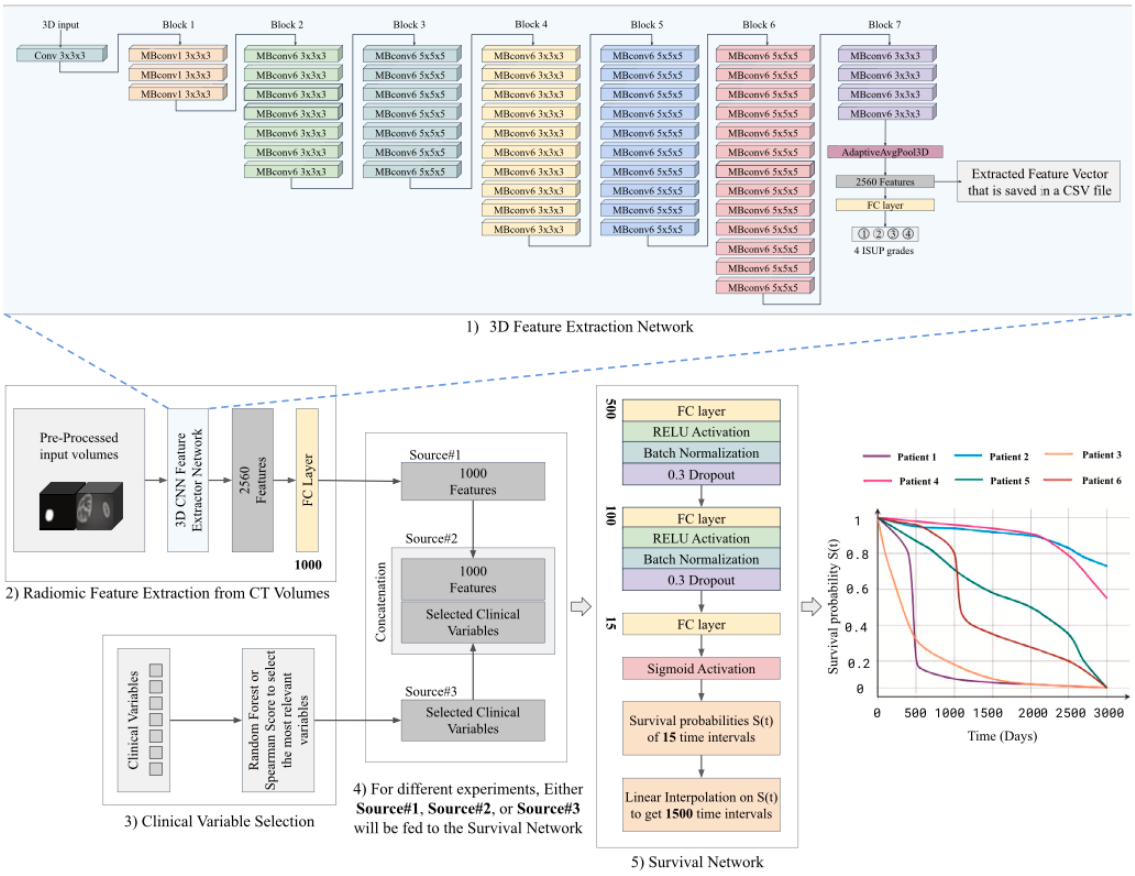


Figure 3.3: Proposed feature-based framework for lung cancer classification in the study[32].

feature extraction, fusion, and survival modeling are very informative for the development of multimodal frameworks for lung cancer classification and prognosis.

Hybrid CNN for Data Integration

The study "LungNet: A hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data" [33] highlights the challenges of early lung cancer detection due to ambiguous nodule features in CT scans and the limited accuracy of existing diagnostic systems. It aims to improve early stage lung cancer classification and diagnosis through the development of LungNet, a novel hybrid CNN-based model that integrates CT imaging and wearable Medical IoT (MIoT) data.

LungNet (figure 3.4) employs a hybrid deep learning framework that integrates a 22-layer CNN for extracting features from CT images with data from wearable MIoT devices capturing physiological indicators like heart rate, body temperature, and breathing patterns. CT images are preprocessed using resizing, color-space conversion, segmentation, and feature enhancement. Augmented images improve training, and image features are combined with normalized MIoT data through a weighted decision support system. This fusion occurs at the feature level, optimizing

predictions using a novel hybrid classifier that supports classification into five main classes and four sub-classes based on cancer stage.

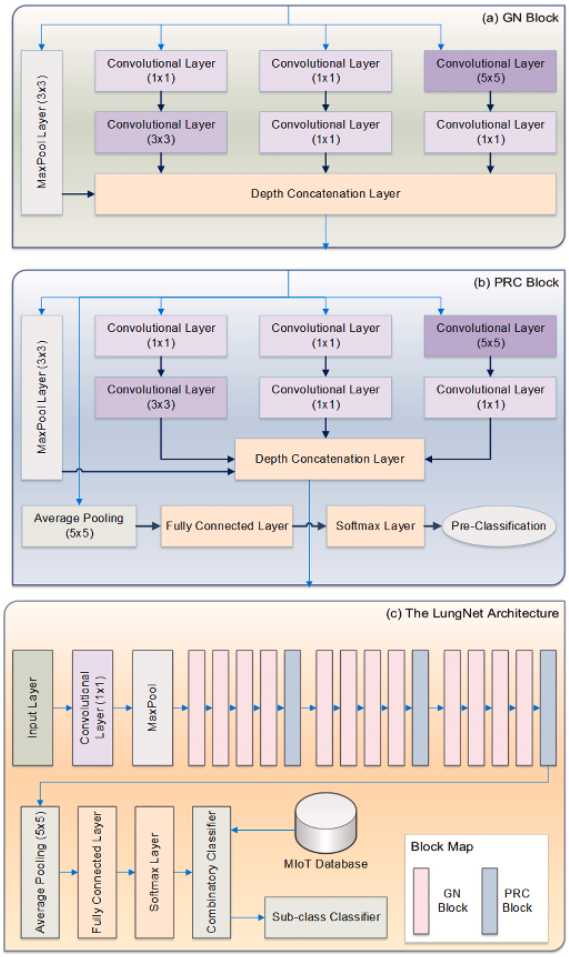


Figure 3.4: The LungNet ARchitecture, as discussed in [33].

The model was trained and tested using two publicly available datasets, LIDC-IDRI and LUNGx Challenge, consisting of 525,000 augmented images. LungNet achieved a classification accuracy of 96.81% on LIDC-IDRI and 91.4% on the LUNGx Challenge dataset, outperforming similar models in sensitivity, specificity, and overall accuracy.

The integration of image and physiological data via LungNet significantly improves diagnostic reliability, achieving superior performance compared to existing models. Its ability to sub-classify stages enhances clinical relevance, demonstrating potential for widespread clinical use.

LungNet's performance relies heavily on image augmentation and MIoT data accuracy, and the computationally intensive nature of its image preprocessing pipeline may hinder real-time scalability. Additionally, external validation across more diverse datasets is needed.

Future work includes automating feature extraction, improving image augmentation strategies, and developing a complete graphical interface to expand accessibility. Integration with genomic and other modalities could further refine diagnostic precision.

This study is useful for this thesis, as it demonstrates a practical implementation of multimodal fusion combining imaging and physiological data for cancer classification.

Fusion of EHR and Radiographic Data for Fracture Prediction

Atypical femur fractures (AFF) are a rare type of fracture that are often difficult to differentiate from normal femur fractures (NFF) based on radiological images alone. Misdiagnoses significantly hinder the ability to provide timely and appropriate treatment. Given the high prevalence of osteoporosis-related fragility fractures in the elderly population, improving the accuracy of AFF identification is critical to addressing this clinical gap. That is the focus of "Fusion of Electronic Health Records and Radiographic Images for a Multimodal Deep Learning Prediction Model of Atypical Femur Fractures" [34].

To tackle this challenge, the study proposes using deep learning models capable of classifying fractures as either AFF or NFF. The approach integrates radiographic imaging data with tabular data extracted from electronic health records (EHR), which includes information on patient demographics and clinical variables. A dataset consisting of 159 AFF patients and 914 NFF patients was created for this purpose. By leveraging multimodal data, the proposed deep learning framework mimics real-world diagnostic scenarios where multiple data sources are considered simultaneously.

In this study, researchers developed a multimodal deep learning approach for predicting atypical femur fractures by integrating EHR and radiographic images. The architecture employed a dense neural network for processing clinical data and a CNN for analyzing medical images, utilizing three fusion techniques: probability fusion, where independent predictions from imaging and tabular models were merged; feature fusion, which concatenated extracted feature vectors into a single input for final prediction; and learned feature fusion, where features from both modalities were jointly trained and integrated into a dense network. By concatenating and adaptively aligning features from these heterogeneous data sources, the model demonstrated an innovative method of synthesizing clinical and imaging information, showcasing the potential of deep learning to enhance predictive diagnostics by leveraging multiple data modalities and improving the accuracy and comprehensiveness of medical risk assessment.

The integrated approach demonstrated notable improvements in diagnostic accuracy. Figure 3.5 represents the results obtained by the research.

The study acknowledges several limitations. High-resolution radiographs were down-sampled and zero-padded to fit the input requirements of pre-trained CNNs, potentially discarding valuable information. Variability in image quality and labeling due to differences in radiology equipment and procedures across 72 hospitals may have also influenced the results. Additionally, the dataset was highly imbalanced, with only 15% of cases being AFF, which may affect model attention to the minority class. Furthermore, while a dense network was used to process tabular data, alternative methods like random forest or XGBoost might yield better performance. The lack of diversity in the dataset, which predominantly includes individuals of Caucasian ethnicity, raises concerns

Data	AUC	Sensitivity	MCC	Specificity
Only images, all images	0.9655	0.7963	0.8135	0.9797
Only images, two images per patient	0.9490	0.7274	0.7739	0.9833
Only tabular data	0.8968	0.8100	0.6049	0.8721
Probability fusion				
fusion of all images + tabular data	0.9866	0.9025	0.8467	0.9641
fusion of two images + tabular data	0.9844	0.8489	0.8367	0.9772
Feature fusion				
fusion of all images + tabular data	0.9701	0.8350	0.8415	0.9805
fusion of two images + tabular data	0.9573	0.7941	0.8087	0.9799
Learned feature fusion				
fusion of all images + tabular data	0.9820	0.8450	0.8309	0.9737
fusion of two images + tabular data	0.9718	0.8015	0.8233	0.9833

Figure 3.5: AUC, sensitivity and MCC results [34].

about the generalizability of the findings to other ethnic groups, particularly populations with a higher risk of AFF. Future work will focus on combining CNNs or vision transformers with more robust tabular data models, such as random forests or XGBoost, to improve predictive metrics. This will include utilizing all 45 tabular variables to achieve better performance by overcoming the limitations of dense networks in handling uninformative features.

The fact that this research used multimodal data is of interest for this thesis. The most important parts are when it is explained how the data from different modalities was fused and the limitations faced, since they might be similar with the ones that this thesis will have to deal with.

Feature-Level Fusion for Nodule Classification and Detection

The fact that great breakthroughs have been made in biomedical imaging technologies translates itself in great opportunities for the healthcare sector. With that, a great problem arises: how complicated and time-consuming it is to collect, measure and label correctly medical data. "Automated Pulmonary Nodule Classification and Detection Using Deep Learning Architectures" [35] dives deep into that.

The authors suggest using state-of-the-art detection architectures for detecting and, posteriorly, classifying lung nodules. It is stated the three architectures used were for the nodules' classification: VGG-16, ResNet-50 and MobileNet. For the detection, the ones used were: Faster-RCNN, Ssd and YOLOv3. The datset LIDC-IDRI was used for the study.

The study integrates radiographic image data with tabular data through a feature-level fusion

approach. This fusion combines extracted features into a unified representation, enabling the model to leverage complementary information from both modalities. The integration enhances diagnostic performance by ensuring discriminative features from both modalities are utilized effectively.

Figure 3.6 shows the framework of how the detection, classification of lung nodules was done in the research. Besides that, it demonstrates how the data was handled. Figure 3.7 shows the results of the study.

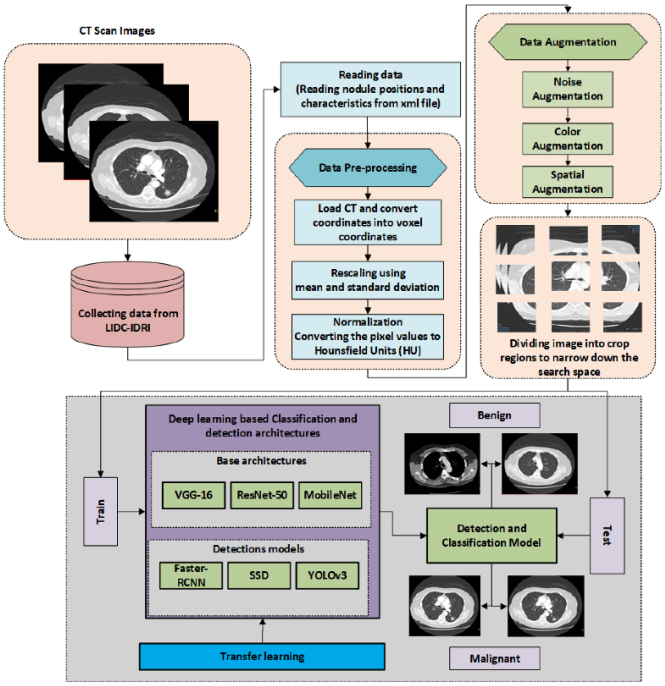


Figure 3.6: Framework of a deep learning-based system for lung nodule classification and detection, including preprocessing, voxel coordinate conversion, image rescaling, data augmentation, patch extraction, and training/testing of deep learning models [35].

It is concluded that all deep learning models have achieved great accuracy scores. Future work will consist on using different architectures, other cancer images and also utilizing other benchmark datasets to generalize how the proposed system performs.

The framework used for the data processing and the detection and classification of the nodules is of great importance for the scope of this thesis, in addition with the fact that the study shows the best combination of detection and classification architectures.

3D CNN with Visual Insights

The study "3D CNN with Visual Insights for Early Detection of Lung Cancer Using Gradient-Weighted Class Activation" [36] talks about current diagnostic methods, particularly CT-based lung nodule classification, are limited by radiologist variability and the complexity of data. This

Deep Learning Architecture	Accuracy	TPR	FPR
Faster-RCNN (VGG-16)	0.90	0.91	0.2
Faster-RCNN (Resnet50)	0.91	0.92	0.15
Faster-RCNN (Mobilenet)	0.91	0.92	0.14
SSD (VGG-16)	0.92	0.92	0.15
SSD (Resnet50)	0.925	0.93	0.10
SSD (Mobilenet)	0.94	0.93	0.10
YOLOv3 (VGG-16)	0.93	0.93	0.08
YOLOv3 (Resnet50)	0.935	0.935	0.08
YOLOv3 (Mobilenet)	0.94	0.94	0.06

Figure 3.7: Performance results of [35].

study seeks to enhance diagnostic accuracy through an automated system that provides both classification and visual explanations, addressing the need for precision and interpretability to gain the trust of clinicians.

The paper proposes an improvised 3D convolutional neural network (3D CNN) based on AlexNet architecture to classify lung nodules as benign or malignant using CT scan data from the LUNA 16 Dataset. The model incorporates gradient-weighted class activation mapping (Grad-CAM) to provide visual insights, highlighting regions contributing to the decision-making process. The LUNA 16 dataset is used to train and evaluate the model, with steps including preprocessing, data augmentation and iterative model optimization. Figures 3.8 and 3.9 demonstrate these architectures.

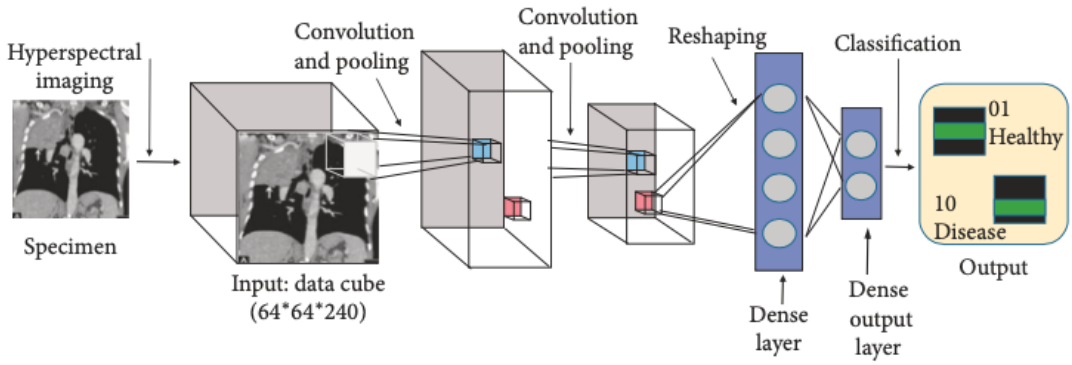


Figure 3.8: 3D CNN with Visual Insights proposed 3D architecture by [36].

The model achieves superior performance, with an AUC of 0.970 and an accuracy of 0.972, and high recall and precision compared to other approaches such as 2D CNNs (AUC 0.94) and standard 3D AlexNet (AUC 0.95). The Grad-CAM analysis adds interpretability, making the model's decisions transparent and aiding clinicians in validating predictions. These results demonstrate the model's ability to enhance diagnostic accuracy significantly.

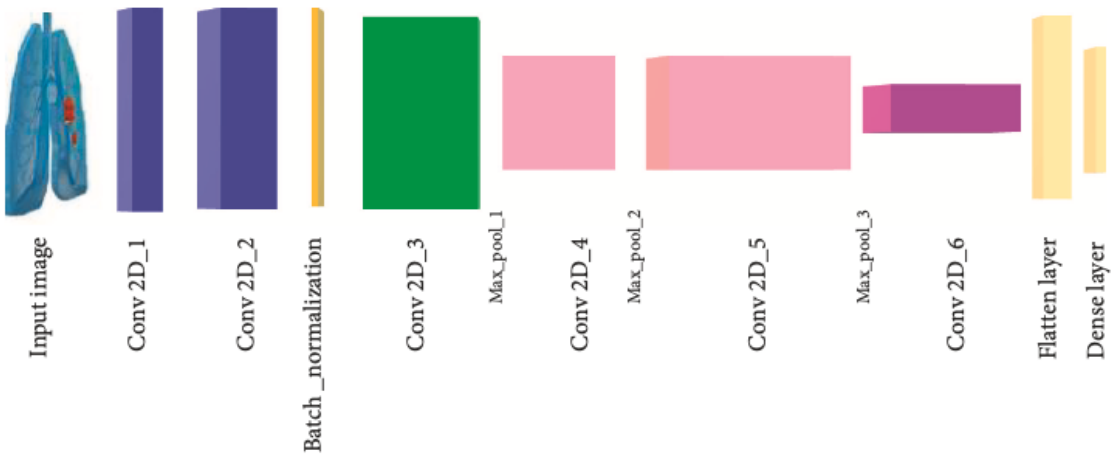


Figure 3.9: Proposed architecture by [36].

The proposed 3D CNN outperforms existing models for lung nodule classification, offering both high precision and interpretability through Grad-CAM. This balance of performance and usability positions the system as a valuable tool for early detection of lung cancer.

Despite its high performance, the study notes that max-pooling operations in CNNs may result in loss of significant data, potentially limiting the system's robustness. The dataset used is relatively small and imbalanced, which might impact generalizability. Additionally, CNNs lack class equivalence, making them less aligned with human visual perception.

Future research will explore capsule neural networks, which preserve spatial hierarchies and require less data, potentially enhancing accuracy while imitating human vision. Expanding the dataset and refining the Grad-CAM framework to handle challenging cases are also highlighted as critical next steps.

The most important point showed in this study for this thesis is the proposed 3D architecture, since it stands out from the usual ones. Despite that, this paper is not of fulcral importance for the thesis at hands and refers that the next step to be taken would be to use capsule neural networks, which has already been done in [37].

Multimodal Dermatology AI

"Artificial intelligence-assisted dermatology diagnosis: From unimodal to multimodal" [38] tackles the problem that exists with AI: although it is progressively being more widely used in the medical field, it has a major flaw - the fact that AI models are, usually, unimodal and reliant on large volumes of accurately labeled data and single data type usage, this way not being enough to actually assist dermatological diagnosis.

The authors found a possible solution for this problem: these models must be augmented with textual data resulting from patient narratives, laboratory reports and image data of skin lesions,

dermascopy and pathologies. This is made possible by using Transformers architecture, Natural Language Processing, Recurrent Neural Networks and Long Short-term Memory.

Several tests integrating many datasets, including ISIC and other specific dermatological datasets, are used to validate the suggested methodology. By combining complimentary data from many modalities, these models show improved diagnostic accuracy. Since this paper is a review and not the result of a single research, many different values are presented for accuracy and recall, based on validations from other papers analyzed.

The study acknowledges many limitations, such as data scarcity, privacy concerns, computational costs, bias and robustness.

Concerning future work, the paper defends that it will be focused on the development of unified platforms for secure, privacy-preserving data sharing (so it is easier to train on different datasets), the exploration of advanced multimodal architectures, the optimization of federated learning techniques to balance privacy, efficiency and accuracy, and on expanding datasets to include a broader range of dermatological conditions and improving dataset availability through collaborative efforts.

The study then concludes that the integration of large-scale pre-trained multimodal models marks a significant step forward in AI-assisted dermatological diagnosis.

Although the thesis at hand is not focused on dermatology, but on lung cancer, this papers brings up many interesting (and fairly helpful) points for the elaboration if this research, such as the utilization of Natural Languaging Processing and the combination of data of different modalities.

3.2 Advanced Image-Based Deep Learning for Diagnosis

This section highlights cutting-edge deep learning methods focused on extracting and analyzing diagnostic insights from imaging data.

Deep Learning for CT Diagnosis

The study "Diagnosis of Lung Cancer from Computed Tomography Scans with Deep Learning Methods" [39] addresses the inefficiencies of traditional lung cancer diagnostic methods, which are labor-intensive and prone to errors due to radiologist workload.

The researchers proposed a deep convolutional neural network (DCNN) alongside other architectures such as ResNet-152V2, AlexNet and VGG-19. These models were optimized for binary classification of CT scans as cancerous or non-cancerous, leveraging preprocessing techniques (e.g., Sobel filtering, watershed segmentation) and advanced transfer learning approaches.

The study utilized the LIDC-IDRI dataset, consisting of 10,991 images (5,491 cancerous, 5,500 non-cancerous), so it can be considered balanced. Models were evaluated using metrics such as accuracy, sensitivity, specificity, F1 score, precision, and AUC. The DCNN and ResNet-152V2 achieved the highest accuracy (99.1%), demonstrating superior diagnostic performance.

The methodology involved: data preprocessing (uniform resolution was achieved through Sobel filtering and watershed segmentation), model training (models employed binary cross-entropy loss, sigmoid activation, and Adam optimization) and evaluation (performance was assessed with 10-fold cross-validation to ensure robustness).

The proposed DCNN included six convolutional layers (32–64 kernels) with max-pooling layers, dropout for regularization and fully connected layers for classification. ResNet-152V2 used transfer learning, fine-tuned with new dense layers, and batch normalization for improved performance.

The proposed DCNN and ResNet-152V2 outperformed other models, offering high accuracy, sensitivity, and specificity. These results highlight the potential of advanced deep learning architectures in improving lung cancer diagnostics.

The study was limited by the size of the dataset and computational resources. Additionally, no external datasets were used for validation, potentially limiting generalizability.

According to the authors, future research will focus on expanding the dataset, exploring deeper architectures, and integrating precise tumor localization capabilities to enhance diagnostic precision and usability.

This study provides valuable insights into advanced deep learning architectures and robust methodologies for lung cancer diagnosis. Despite this, since the main focus of this thesis is the application of multimodal approaches for the binary classification of lung nodules in cancerous or non-cancerous, this paper is not crucial. The fact that it explains the methodology of image analysis may be helpful, since one kind of multimodal data that will be used is imaging data.

MIXCAPS

The authors of "MIXCAPS: A capsule network-based mixture of experts for lung nodule malignancy prediction" [37] focus on improving lung nodule malignancy prediction, addressing the limitations of traditional diagnostic approaches, such as their reliance on hand-crafted features and annotated datasets. The variability and complexity of CT scan features necessitate advanced solutions like capsule networks for robust classification.

The MIXCAPS framework, present in Figure 3.10 consists of two capsule network experts and a convolutional gating network, designed for binary classification of lung nodules (benign vs. malignant). Input patches from the LIDC-IDRI dataset are preprocessed and padded to standard dimensions. The gating network dynamically assigns data subsets to capsule experts based on learned probabilities, ensuring specialization. The outputs of the capsule experts are combined through a weighted summation to produce the final prediction. This hierarchical mixture of experts approach leverages capsule networks' strengths, such as handling spatial relationships and working with smaller datasets, to achieve high accuracy.

The model was validated using the LIDC-IDRI dataset, containing CT scans of 1,018 patients. MIXCAPS achieved an average accuracy of 90.7%, sensitivity of 89.5%, specificity of 93.4%, and

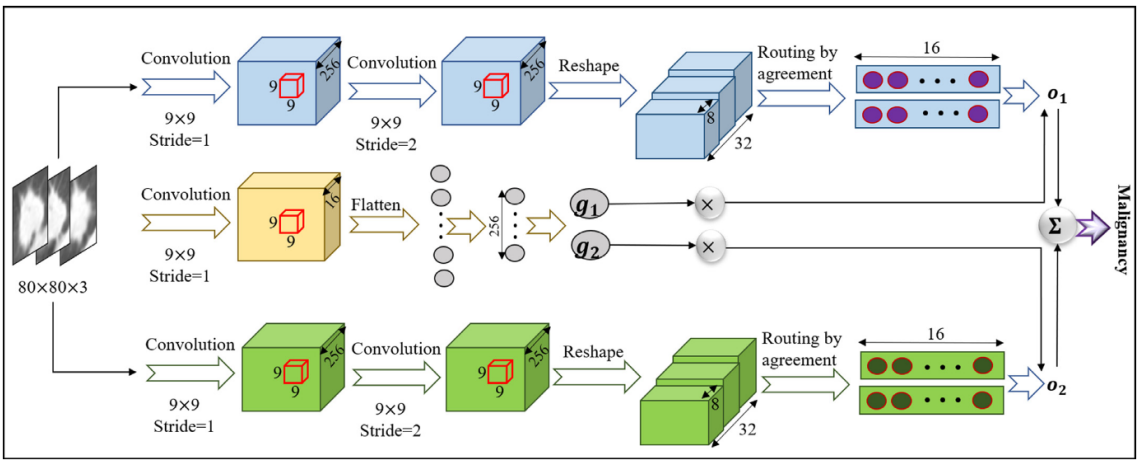


Figure 3.10: Proposed MIXCAPS architecture in [37].

AUC of 0.956, outperforming single capsule networks, mixtures of CNNs, and other benchmark models.

MIXCAPS demonstrates superior classification performance by leveraging capsule networks' ability to handle small datasets and spatial relationships. Its use of a gating mechanism to dynamically assign data subsets improves overall accuracy and interpretability.

The framework requires computational resources due to the mixture of experts' complexity and lacks external validation on diverse datasets. Additionally, capsule networks' high training cost might limit real-time applications.

Future research will explore reducing computational overhead through model distillation, testing generalizability with diverse datasets, and incorporating capsule gating networks for improved efficiency.

This study is relevant to this thesis as it presents an innovative use of capsule networks and a gating mechanism for medical data imaging.

AI for Chest Image Analysis

"The application of artificial intelligence to chest medical image analysis" [40] focuses on improving the accuracy and efficiency of chest medical image analysis using AI, particularly deep learning, to address the growing demand for non-invasive diagnostic tools. Traditional methods struggle with handling large datasets, inconsistent interpretations, and heavy reliance on medical professionals.

The study reviews and evaluates various DL architectures and learning strategies for classification, detection, and segmentation tasks in chest imaging. It highlights the application of DL to diverse modalities for diagnosing diseases like lung cancer, pneumonia, and pulmonary embolism.

The methodology includes applying state-of-the-art DL architectures such as 2D and 3D CNNs, ResNets, and U-Nets for feature extraction, classification, and segmentation tasks on chest

imaging datasets. Data preprocessing involved normalization and augmentation, while feature-level fusion integrated imaging and clinical data to enhance predictions. Training employed supervised learning strategies, optimizing using methods like Adam and stochastic gradient descent, and evaluating with metrics such as AUROC and precision.

Public datasets such as LIDC-IDRI, LUNA16, and ChestX-ray8 were extensively used to train and validate the DL models.

DL methods demonstrated superior performance in chest imaging tasks, enabling accurate and scalable diagnostics. The integration of multi-modal data, including imaging and clinical features, showed potential to further enhance diagnostic reliability.

The study notes challenges such as data heterogeneity, the need for larger and more diverse datasets, and difficulties in deploying models in real-world clinical settings

Future works is based on improving dataset diversity, developing explainable AI for medical applications, and exploring multi-modal fusion to integrate imaging and clinical data for comprehensive diagnostics.

This study is highly relevant for this thesis as it explores the use of DL for chest image analysis and multimodal integration of clinical and imaging data.

3.3 Feature Extraction and Fusion Techniques

This section focuses on sophisticated methods that improve the representational capability of multimodal data by extracting and integrating features from data sources.

Imbalanced Data for Lung Cancer Classification

Many current lung cancer diagnostic methods struggle with imbalanced datasets and insufficient analysis of spatial and frequency domain data. These limitations can lead to overfitted results and hinder the generalization of classification models.

The paper "A multichannel analysis of imbalanced computed tomography data for lung cancer classification" [41] proposes a Multi-Channel Ensemble Convolutional Neural Network (MC-ECNN) that combines spatial and frequency domain analysis. The approach utilizes preprocessing techniques such as resizing, Gaussian blur smoothing, and Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance. Input data includes grayscale CT images, as well as their frequency-domain transformations using Discrete Fast Fourier Transform (DFFT) and Discrete Cosine Transform (DCT). Features extracted from each domain are processed through individual CNN channels and integrated using ensemble learning to improve classification performance across benign, malignant, and normal classes. Figure 3.11 shows the architecture used in this study.

The model was trained and evaluated using the IQ-OTHNCCD Version 4 dataset, which contains 1,097 lung CT images (120 benign, 561 malignant, and 416 normal). After applying SMOTE,

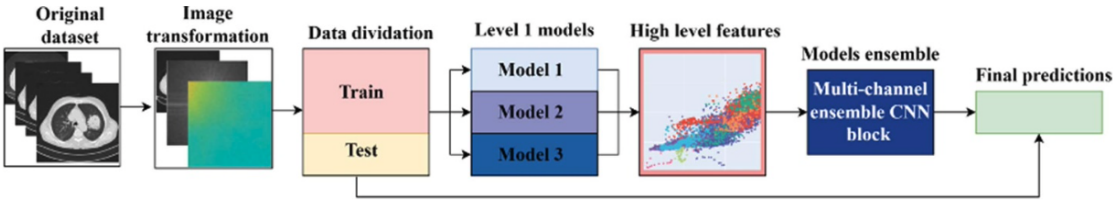


Figure 3.11: Imbalanced Data for Lung Cancer Classification CNN architecture that is used in this study. [41]

the dataset was balanced to 1,680 images (560 per class). Using a 10-fold cross-validation approach, the MC-ECNN achieved an accuracy of 99.60%, outperforming state-of-the-art methods in precision, recall and F1-score. The multi-channel approach, compared to single-channel analysis, demonstrated superior performance in capturing subtle variations in the data, especially when classifying benign and malignant samples.

The proposed MC-ECNN effectively addresses data imbalance and improves lung cancer classification by integrating spatial and frequency domain features. Its ensemble learning framework not only enhances accuracy but also ensures robustness and generalizability, establishing a new benchmark for lung cancer detection models.

The study acknowledges limitations in computational complexity due to multi-domain analysis and high-resolution imaging. While the SMOTE technique mitigates data imbalance, further improvements in the benign class accuracy are necessary. The reliance on the IQ-OTHNCCD dataset, which may not reflect broader population diversity, could limit the generalization of the model.

Future work will focus on integrating advanced transformation techniques, such as wavelet analysis, to further refine spatial and frequency domain feature extraction. Additionally, incorporating automated anomaly measurement capabilities and validating the model across diverse datasets will aim to enhance both accuracy and practical applicability.

This paper is highly relevant to this thesis as it addresses handling imbalanced datasets and integrating spatial and frequency domain features through a multichannel ensemble CNN.

Form Factors to Differentiate Lung lesions

The study "Form Factors as Potential Imaging Biomarkers to Differentiate Benign vs. Malignant Lung Lesions on CT Scans" [42] aims to improve the differentiation of benign versus malignant lung nodules on CT scans, addressing the critical need for accurate and early diagnosis, which is crucial for maximizing treatment outcomes. Traditional diagnostic approaches are subjective and often lack reproducibility, motivating the use of quantitative imaging biomarkers.

The study explores form factors, including elongation, flatness, and compactness, as potential malignancy indicators, combining them with conventional imaging metrics such as size and surface area to enhance diagnostic accuracy.

Two datasets were used: SSR-1 (109 nodules: 38 benign, 71 malignant) and LUNGx (83 nodules: 42 benign, 41 malignant), assessed through manual delineation and histological validation. The predictive models, using Support Vector Machines (SVM), demonstrated that adding form factors to conventional features improved diagnostic accuracy by up to 14.5%.

The approach involved several systematic steps to enhance lung nodule classification using CT scan data. Initially, lung nodules were manually delineated in the CT images to create precise regions of interest for analysis. Next, features were extracted, including conventional metrics such as volume and diameter, as well as advanced form factors like compactness and flatness, providing a more comprehensive quantitative representation of the nodules. Linear SVM models were then trained using two distinct feature sets: a base set comprising conventional metrics alone and an extended set that combined the base features with form factors. Finally, the models underwent rigorous evaluation through intra-dataset validation (within the same dataset) and inter-dataset validation (across different datasets) to assess their accuracy and robustness. This methodical approach demonstrated the potential of integrating form factors to improve diagnostic performance.

The study concludes that form factors significantly improve the differentiation of benign and malignant nodules, with malignant lesions tending to be more compact and isotropic. This approach highlights the clinical potential of shape-based biomarkers in lung cancer diagnosis.

The study is limited by its retrospective design, small sample size, and lack of prospective validation, which may affect generalizability.

The authors believe that future efforts will focus on larger and prospective datasets, as well as investigating biological correlations between shape features and tumor growth patterns.

This thesis would probably may benefit since the paper's rigorous methodology, including intra- and inter-dataset validation using SVM models, offers a robust framework that could guide this thesis model evaluation process.

Lung Cancer in CT Images

"Lungs cancer classification from CT images: An integrated design of contrast based classical features fusion and selection" [43] addresses the limitations of conventional lung cancer diagnostic methods, such as low tumor visibility and high false-negative rates in CT images. The aim is to enhance classification accuracy through improved feature extraction, fusion, and selection techniques for lung cancer classification.

The researchers proposed a three-step method. First, CT images are preprocessed using Gamma Correction Max Intensity Weights (GCmIW) to enhance contrast and improve tumor visibility. Second, multiple features, including Local Binary Patterns (LBP), Discrete Wavelet Transform (DWT)-Geometric features, Haralick Gray Level Co-occurrence Matrix (HGLCM) texture features, and Maximally Stable Extremal Region (MSER) orientation features, are extracted from the contrast-enhanced images. Third, feature fusion is performed using Serial Canonical Correlation Analysis (SCCA) to combine the extracted features into a unified vector, followed by

Weighted Neighborhood Component Analysis (WNCA) to select the most relevant features by removing zero or irrelevant values. The selected features are then classified into cancerous or non-cancerous categories using an Ensemble Bagging Tree (EBT) classifier, achieving high accuracy and computational efficiency. Figure 3.12 shows the proposed framework by the authors.

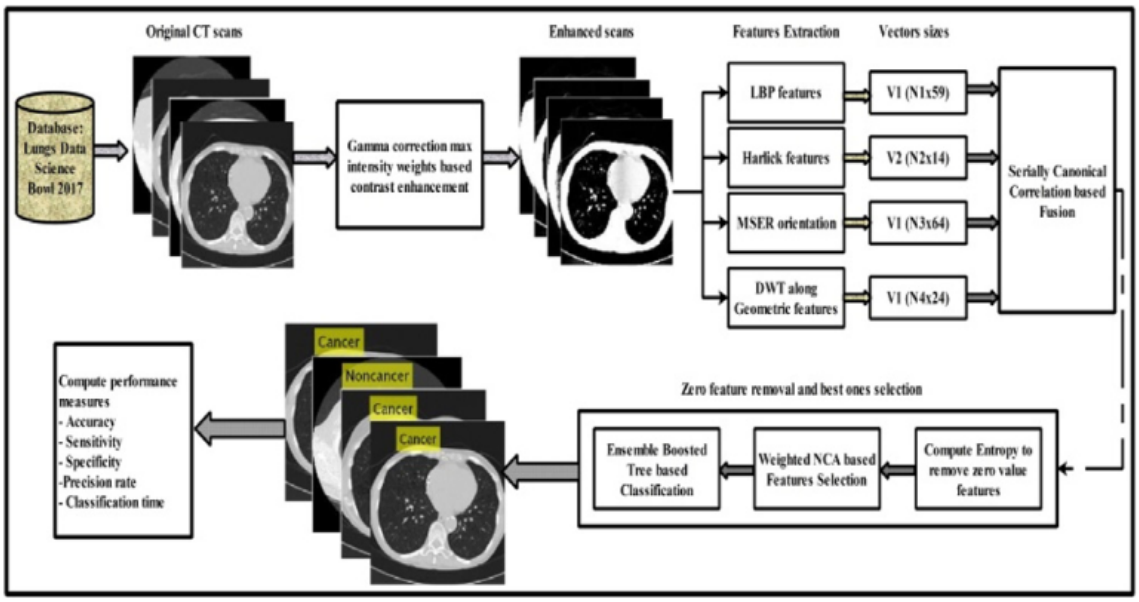


Figure 3.12: Proposed feature-based framework for lung cancer classification in the study[43].

The method was validated on the publicly available Lung Data Science Bowl 2017 dataset, which includes 2,101 CT scans of cancerous and non-cancerous lungs. The proposed approach achieved a classification accuracy of 99.5%, outperforming existing methods in sensitivity, specificity, and precision.

The integrated design significantly improved lung cancer classification, achieving superior accuracy by leveraging contrast enhancement, multi-feature fusion, and optimal feature selection. The approach demonstrates the potential for automated diagnostic systems to assist clinicians in early detection and classification of lung cancer.

The model's reliance on a single dataset and manual steps, such as segmentation and feature extraction, may limit its generalizability and scalability. Computational time also increases during the fusion process, impacting real-time applicability.

Future directions include applying deep learning methods to reduce computational complexity, automating feature extraction processes, and validating the model on larger and more diverse datasets to enhance its clinical utility.

This study is relevant to the thesis, as it demonstrates advanced feature fusion techniques, including SCCA and WNCA, combined with a robust classification framework. The approach can provide valuable insights for integrating multimodal data and improving classification performance.

3.4 Temporal and Longitudinal Data Integration

This section examines approaches that incorporate temporal or longitudinal data to capture changes in lung nodules over time, providing dynamic insights for improved classification and prognosis. Temporal data refers to information collected at specific time points, while longitudinal data tracks the same entity across multiple time intervals, enabling the analysis of trends, progression, and causality in patient outcomes.

Fusion Techniques Review

"A review of deep learning-based information fusion techniques for multimodal medical image classification" [44] explores how multimodal medical image fusion, leveraging deep learning techniques, can improve diagnostic accuracy. The motivation lies in overcoming the limitations of single-modality diagnostics, which often fail to provide sufficient or complementary information for complex diseases like cancer and neurological disorders.

This review underscores the critical role of multimodal data in improving lung cancer diagnosis and prognosis. Traditional approaches that rely solely on CT imaging often struggle with the variability in nodule appearance and the complexity of the disease. By combining CT imaging data with clinical and molecular profiles, diagnostic models can achieve significantly better accuracy and robustness, particularly in malignancy classification and tumor staging.

The review emphasizes the efficacy of intermediate fusion techniques, especially attention-based mechanisms, for lung cancer applications. These methods dynamically weight and prioritize relevant features from different data sources, enabling more accurate predictions. For example, CT imaging features can be aligned with patient metadata, such as smoking history or genetic mutations, to enhance the model's focus on high-impact variables. This fusion strategy is particularly useful in handling the inherent heterogeneity of lung cancer data.

It also identifies key challenges in lung cancer diagnostics, including the need for larger multimodal datasets and methods to address inconsistencies in data resolutions and formats. These limitations highlight the importance of robust preprocessing techniques, such as segmentation and feature extraction, to standardize input data for reliable fusion and analysis.

The review further provides examples of how multimodal fusion techniques can be applied to specific lung cancer tasks. These include combining radiomic features from CT scans with clinical variables to classify nodule malignancy or integrating molecular profiles for survival predictions. Such applications demonstrate the practical value of integrating diverse data modalities for more comprehensive and accurate diagnostics.

The study points toward future directions in lung cancer research, such as leveraging transformer architectures for better data alignment and scalability. It also emphasizes the need for improved multimodal datasets that combine imaging, clinical, and molecular data to facilitate more effective model training and validation.

This review offers actionable insights into how multimodal fusion techniques can advance lung cancer diagnostics, providing valuable guidance for developing robust AI frameworks in this research.

Lung Nodule Malignancy Prediction

The study "Lung Nodule Malignancy Prediction From Longitudinal CT Scans With Siamese Convolutional Attention Networks" [45] addresses the limitations of single-time-point lung nodule classification methods by utilizing longitudinal CT scans to capture temporal changes, which are crucial for malignancy prediction. Existing approaches often overlook growth progression and temporal patterns, which can provide critical diagnostic insights.

A Siamese Convolutional Attention Network (CAN) is proposed, leveraging shared 2D CNN feature extractors combined with an attention mechanism to dynamically weight relevant slices. This architecture efficiently processes multi-time-point CT scans to improve the prediction of nodule malignancy.

The NLSTx dataset was used, consisting of 857 nodules (207 malignant, 650 benign) with biopsy-confirmed diagnoses across multiple time points (t0, t1, t2).

The proposed Siamese CAN outperformed baseline models across single and multi-time-point classifications. For single-time-point classification on the NLSTx dataset, the CAN achieved an AUC of 0.962, Average Precision (AP) of 0.969, and Log Loss (NLL) of 0.256, compared to the 3D CNN baseline, which achieved a ROC-AUC of 0.931, AP of 0.943, and NLL of 0.372. For multi-time-point classification, the two-time-point CAN (CAN2-2D) showed further improvement with a ROC-AUC of 0.979 and AP of 0.984, while the three-time-point CAN (CAN3-2D) maintained similar performance with a ROC-AUC of 0.978 and AP of 0.983. These results demonstrate the effectiveness of integrating longitudinal CT scan data using the attention-based Siamese architecture.

The methodology begins with preprocessing, where images are resampled to a uniform resolution and enhanced through data augmentations like rotations and Gaussian noise. Next, feature extraction is performed using a pre-trained 2D CNN to process slices individually. A slice-wise attention mechanism dynamically emphasizes relevant slices while suppressing irrelevant ones. For fusion, attention-weighted features across multiple time points are concatenated to capture temporal information. Finally, fully connected layers classify the nodules as benign or malignant. Figure 3.13 represents the proposed framework by the authors.

The Siamese CAN architecture employs shared 2D CNN branches with an attention mechanism, significantly reducing parameters compared to 3D CNNs. The network processes single-time-point data and concatenates features for multi-time-point inputs.

The proposed Siamese CAN efficiently combines temporal features, outperforming conventional 3D CNNs and single-time-point models. It demonstrates that integrating temporal data with attention mechanisms can significantly enhance malignancy prediction performance.

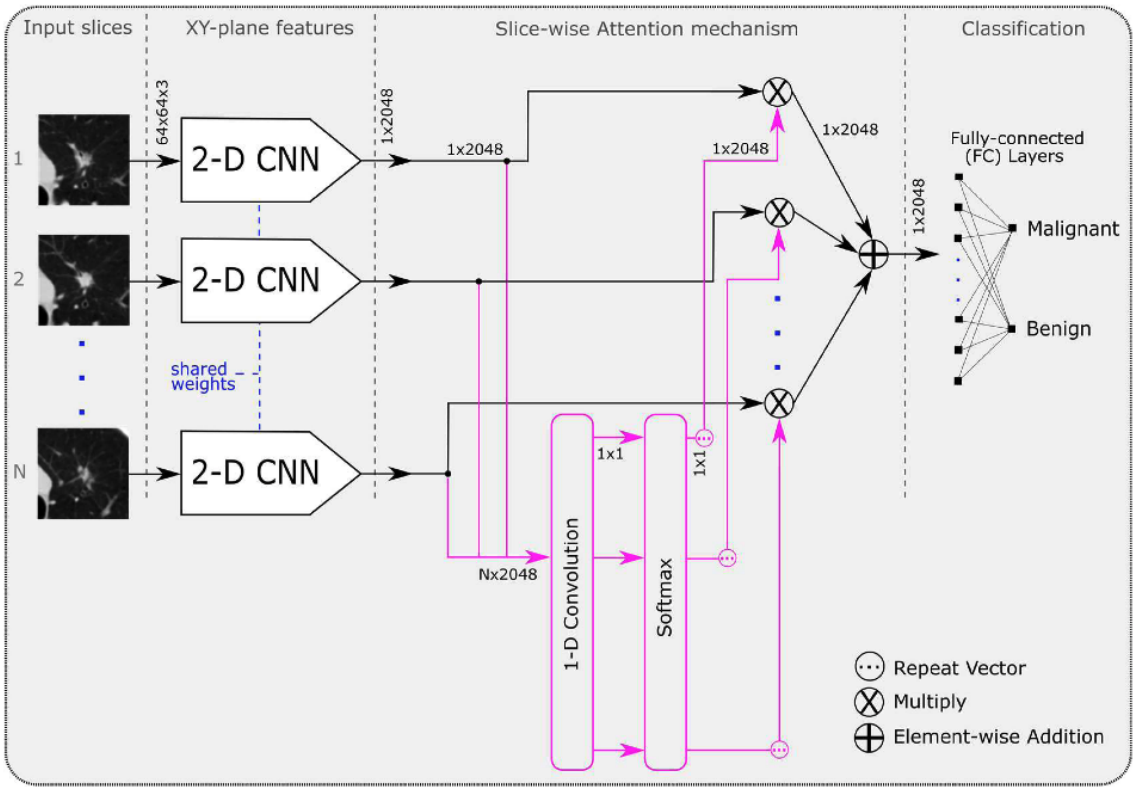


Figure 3.13: Diagram of the proposed framework in the study[45].

The study is limited by the size of the NLSTx dataset and the availability of longitudinal scans, restricting generalizability to larger and more diverse populations.

According to the authors, future work must focus on expanding the dataset, refining the attention mechanism, and applying the approach to other medical imaging tasks to improve its clinical applicability.

This study is highly relevant to the thesis as it demonstrates an innovative temporal data fusion approach using a Siamese attention network for lung nodule classification.

3.5 Data Challenges in Multimodal Systems

This section examines techniques intended to address issues like noise, heterogeneity, and data imbalance in multimodal datasets, guaranteeing stable and dependable model performance.

Hybrid Model for Lung Cancer

The study "A hybrid model for lung cancer prediction using patch processing and deep learning on CT images" [46] seeks to improve lung cancer detection using CT images by addressing the limitations of existing methods, such as low accuracy and high processing times, through a novel approach combining patch processing and deep learning.

A hybrid model is proposed that employs CNN for feature extraction and classification, combined with preprocessing techniques like median filtering and patch processing for noise reduction and image enhancement. Figure 3.14 displays the proposed system.

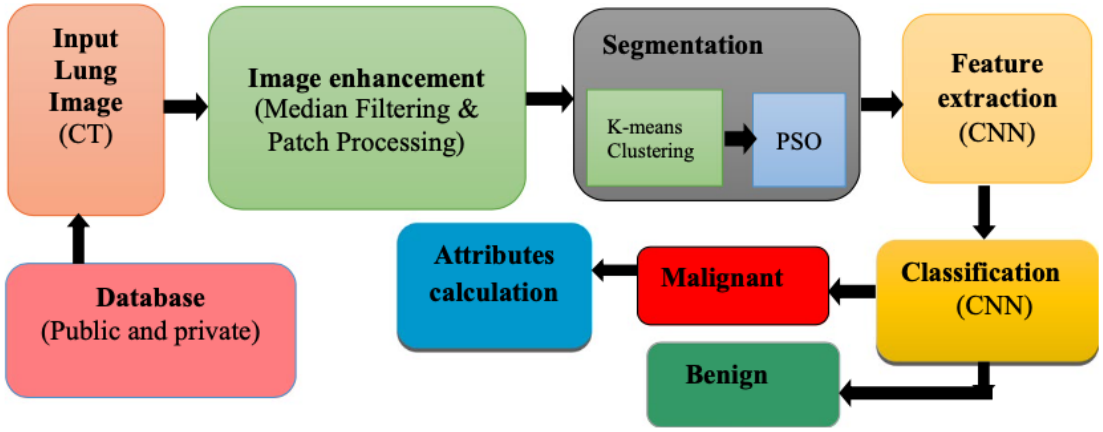


Figure 3.14: Hybrid Model for Lung Cancer system in the study[46].

The method was validated using CT images from the LIDC and Kaggle datasets, comprising a total of 90 images. The model demonstrated an accuracy of 99.96%, outperforming other existing methods.

The preprocessing stage involves applying median filtering for noise reduction and enhancing the images using a patch-based method. Segmentation is then performed using a combination of K-means clustering and Particle Swarm Optimization (PSO) to identify regions of interest. Finally, a CNN is used to extract features and classify tumors as benign or malignant.

This hybrid model significantly improves detection and classification of lung tumors, achieving superior accuracy and efficiency compared to traditional methods. This approach highlights the potential of integrating preprocessing and deep learning techniques for reliable cancer diagnosis.

The small dataset size (90 images) and lack of external validation limit the model's generalizability and robustness across diverse populations.

Future research would involve training on larger datasets, employing 3D CNN architectures, and improving neural network configurations for enhanced accuracy and scalability.

This study is relevant as it demonstrates a different hybrid approach combining preprocessing techniques and deep learning for lung cancer detection. Its methodology and results can inform the development of this thesis multimodal diagnostic framework, particularly in handling CT images.

3.6 Comparative Studies on Multimodal Approaches

This section summarizes comparative research that assesses and compares different multimodal strategies, including information on their relative advantages and disadvantages.

AI in Cancer: a Review

The review "Artificial intelligence in cancer diagnosis and therapy: Current status and future perspective" [47] emphasizes the complexity of lung cancer diagnosis and treatment, identifying it as one of the most lethal cancers due to its late detection and aggressive progression. Traditional diagnostic methods, which often rely on manual interpretation of imaging data, are noted to be prone to errors and inefficiencies, highlighting the growing need for advanced technological solutions.

The review highlights the pivotal role of imaging, particularly CT scans, in lung cancer diagnostics. CT imaging is described as indispensable for detecting lung nodules and assessing malignancy. However, the variability in nodule size, shape, and appearance is cited as a major barrier to accurate early detection, even for experienced radiologists. These challenges demonstrate the need of improving diagnostic tools to better handle the difficulties of lung cancer presentation.

AI, particularly deep learning, is presented as a transformative tool for lung cancer diagnosis and treatment planning. Techniques such as CNNs and U-Nets are shown to be highly effective in tasks like nodule segmentation, malignancy classification, and outcome prediction. The integration of imaging data with clinical and genomic information is also identified as a promising avenue for creating multimodal frameworks that enable more comprehensive and accurate diagnostics. These AI applications are demonstrated to enhance sensitivity, specificity, and overall diagnostic accuracy while reducing false positives.

This review provides valuable insights into how AI-driven innovations address the limitations of traditional methods in lung cancer diagnosis. These advancements are reported to have the potential to revolutionize early detection, personalized treatment planning, and prognostic assessments, contributing to significant improvements in patient outcomes.

Medical image identification

The review "Medical image identification methods: A review" [48] emphasizes the growing importance of AI-driven methods in lung cancer diagnosis, particularly through medical image classification and segmentation. Lung cancer remains a leading cause of cancer-related deaths, and early detection plays a critical role in improving survival rates. The review identifies challenges such as variability in tumor appearance, class imbalance in datasets, and the high demand for annotated data, all of which limit the efficacy of traditional diagnostic tools.

The review highlights the potential of CNNs and DL methods to improve accuracy in lung cancer diagnostics. Examples include classification models using CT scans to detect and segment pulmonary nodules, with architectures like U-Net and transfer learning approaches demonstrating significant advancements. Additionally, the review emphasizes the use of datasets such as LIDC-IDRI for benchmarking, with some models achieving high accuracy and sensitivity, making them suitable for helping radiologists in clinical workflows.

AI models integrating multimodal data, including imaging and clinical metadata, are noted for

their potential to enhance diagnostic precision. Feature fusion strategies, particularly those using attention mechanisms and ensemble learning, are recognized as key for handling the complexity and heterogeneity of lung cancer data. However, the review acknowledges the challenges of data scarcity, labeling inconsistencies, and the need for external validation across diverse populations.

This review offers valuable insights for this thesis by highlighting cutting-edge methodologies and their applications to lung cancer.

3.7 Summary

The revolutionary potential of combining several data sources for more precise and understandable results is shown by the recent developments in multimodal AI for medical diagnostics, especially in the characterisation of lung cancer. Imaging data has historically been used to diagnose lung cancer, but it is becoming more widely acknowledged that combining clinical, molecular, and temporal data into coherent frameworks is an essential first step toward more comprehensive and individualized healthcare solutions. The application of sophisticated models, like MFDNN, which integrate genomic data, clinical data, and CT imaging to enhance early detection, has been investigated in this field of study. Although there are still issues like reduced datasets and the requirement for automated feature extraction, these approaches improve prediction performance by utilizing feature fusion techniques, such as concatenation and support vector machines with multiple kernel learning.

The integration of temporal and longitudinal data further enriches diagnostic capabilities. Techniques like Siamese CAN effectively process multi-time-point CT scans, dynamically weighting relevant slices to predict malignancy while reducing model complexity. Despite these advances, challenges persist, including dataset imbalances, the computational demands of integrating heterogeneous data, and the limited external validation of models. Future research directions focus on automating feature extraction processes, expanding the diversity and size of datasets, and incorporating real-time, interpretable AI frameworks to meet clinical demands.

The application of multimodal AI frameworks in lung cancer diagnosis has potential for major improvements in early detection, individualized care, and overall patient outcomes by tackling these issues. This development provides a solid basis for personalized medicine's future, which will be fueled by strong AI systems and data-rich insights.

This thesis will concentrate on tackling important issues like the successful integration of multimodal data, such as clinical records (e.g., demographics, comorbidities), imaging data (e.g., CT scans, nodule characteristics), and molecular/omics data (e.g., genetic profiles), into a single AI framework, while keeping in mind the limitations encountered by previous efforts in AI-based models for lung cancer characterization. This research will focus on creating reliable, broadly applicable AI models to improve their actual clinical usefulness in finding out if a patient has lung cancer or not. By giving physicians useful information, diminishing diagnostic errors, and facilitating individualized treatment plans, this strategy will seek to improve lung cancer detection and treatment planning.

Chapter 4

Data

This chapter offers a detailed overview of the data that supports this study. It begins by introducing the dataset used, outlining the criteria for selecting the study population, and detailing the preprocessing steps applied to both clinical and imaging data. Additionally, the chapter describes the standardization procedures and the segmentation technique implemented to create lung area binary masks. Collectively, these processes create a strong and reproducible empirical foundation for the analyses and models presented in the following chapters.

4.1 National Lung Screening Trial

The National Lung Screening Trial (NLST) [49] was a randomized clinical trial organized by the National Cancer Institute between 2002 and 2010. Involving 53,454 current or former heavy smokers aged 55-74, the study took place across 33 medical centers in the United States. Participants had a smoking history of at least 30 pack-years. At enrollment, participants showed no signs, symptoms, or history of lung cancer, and were randomly assigned to receive either low-dose helical CT scans or standard chest X-rays. The aim was to evaluate the effectiveness of these screening methods for the early detection of lung cancer [50].

The NLST dataset is rich in baseline demographic information, detailed smoking history, and comprehensive medical records for all participants. For participants diagnosed with lung cancer, the dataset includes thorough documentation on histology, cancer stage, treatment strategies, and patient outcomes [51].

The NLST is an ideal dataset for implementing a multimodal approach in lung cancer research due to its extensive and diverse data collection, comprising over 70,000 CT scans [52]. This combination of imaging and clinical data helps create models that analyze the links between radiological findings in the CT scans and patient history. The NLST's scale, depth, and multimodal composition make it an extremely valuable resource for advancing diagnostics and prognostics in lung cancer.

4.2 Data selection

For this thesis, patients were selected from the NLST dataset using strict inclusion criteria to ensure the consistency and reliability of the imaging and diagnostic data. The selected patients can be divided into two parts:

- **Lung cancer positive cases:**
  - Individuals diagnosed with lung cancer during the radiological screening phase.
  - The reported abnormality location matched the confirmed site of cancer diagnosis.
  - Only cases with complete and accurately labeled CT images available in the dataset were included.
- **Lung cancer negative cases:**
  - Individuals not diagnosed with lung cancer but had type 51 abnormalities ("Non-calcified nodule or mass (opacity $\geq$ 4 mm diameter)" [51]) reported in their final screening exam within the last year of imaging.
  - Required to have corresponding CT images accurately represented in the dataset.

Figure 4.1a shows three representative CT slices with confirmed cancerous nodules, while Figure 4.1b presents slices from negative cases, meaning that the nodules found are non-cancerous.

The selection process aimed to ensure high-quality data for both cancer and non-cancer cases, with clearly defined imaging and diagnostic information.

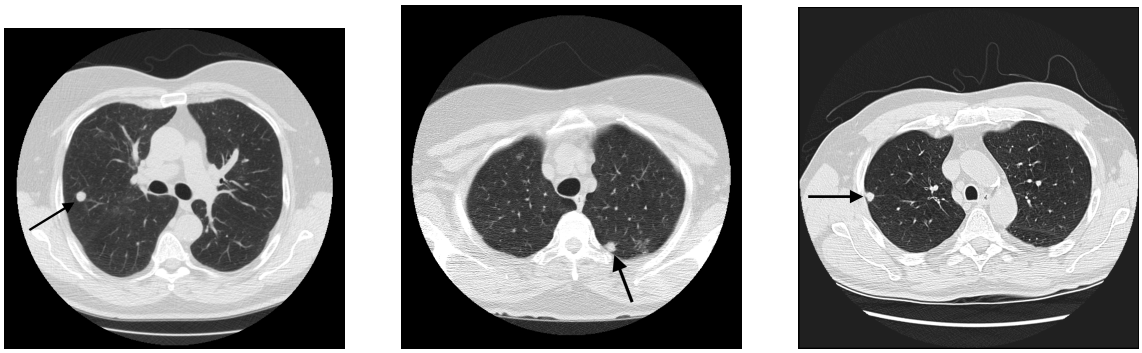
Patients with cancer had only one node, while patients with non-cancerous tumors had one or more nodules.

Figure 4.2 represents the patients' selection process. Due to time and resource limitations, only around 700 nodule slices (half of them cancerous, half benign) were manually checked and used for this thesis.

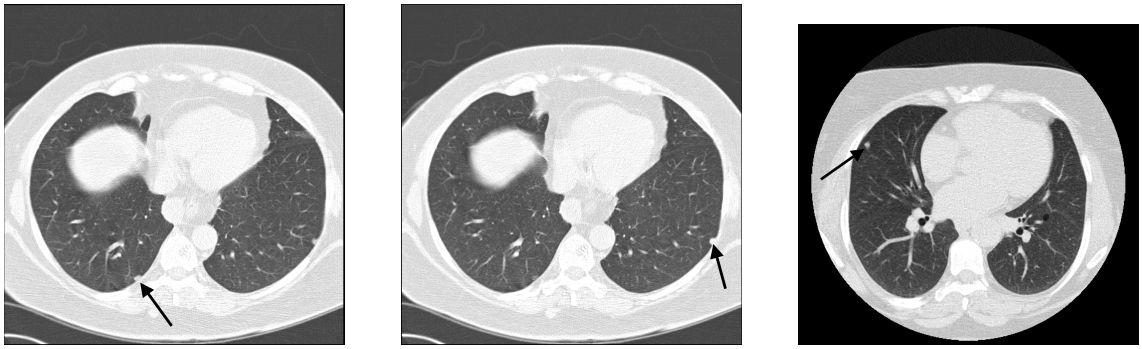
Table 4.1 provides a concise summary of the final number of patients with valid CT scans used in this study. After applying all inclusion and exclusion criteria to the NLST dataset, 400 patients with confirmed lung cancer and matched abnormalities, and 1,149 patients with non-cancerous nodules (type 51) were retained. This yielded a final total of 1,549 patients with diagnostically usable CT scans.

Table 4.1: Summary of patient and CT scan selection from the NLST dataset.

Category	**Patients with valid CTs**
Diagnosed with Lung Cancer	400
Not Diagnosed with Lung Cancer (Type 51 only)	1,149
Total Patients with Valid CTs	**1,549**



(a) CT slices with cancerous nodules (indicated by arrows).



(b) CT slices with non-cancerous lung nodules (indicated by arrows).

Figure 4.1: CT scans showing cancerous and non-cancerous lung nodules.

4.3 Imaging data pre-processing

The preprocessing of imaging data is crucial for preparing CT scans for reliable input into models. This section details the steps taken to standardize and normalize the raw DICOM files. Figure 4.3 shows the used pipeline for the imaging pre-processing.

4.3.1 Choice of DICOM series

Each CT scan visit generates a series of DICOM files (one per axial slice) resulting in multiple scan series per patient across time or due to differing reconstruction parameters. Because the imaging task in this thesis focuses on binary classification of lung nodules (cancerous vs. non-cancerous), it is critical to ensure that only the most diagnostically relevant and technically consistent series is selected for each patient. Selecting the optimal series improves input homogeneity, reduces variability introduced by protocol differences, and enhances the model's ability to learn discriminative features.

A key factor influencing CT image quality is the reconstruction kernel, which governs how raw projection data is processed into axial images. Different kernels emphasize different image properties, such as sharpness for bone edges or smoothness for soft tissues. Kernels tailored for lung imaging strike a balance between edge detail and noise suppression, preserving the subtle parenchymal structures necessary for accurate nodule detection. Following the methodology of

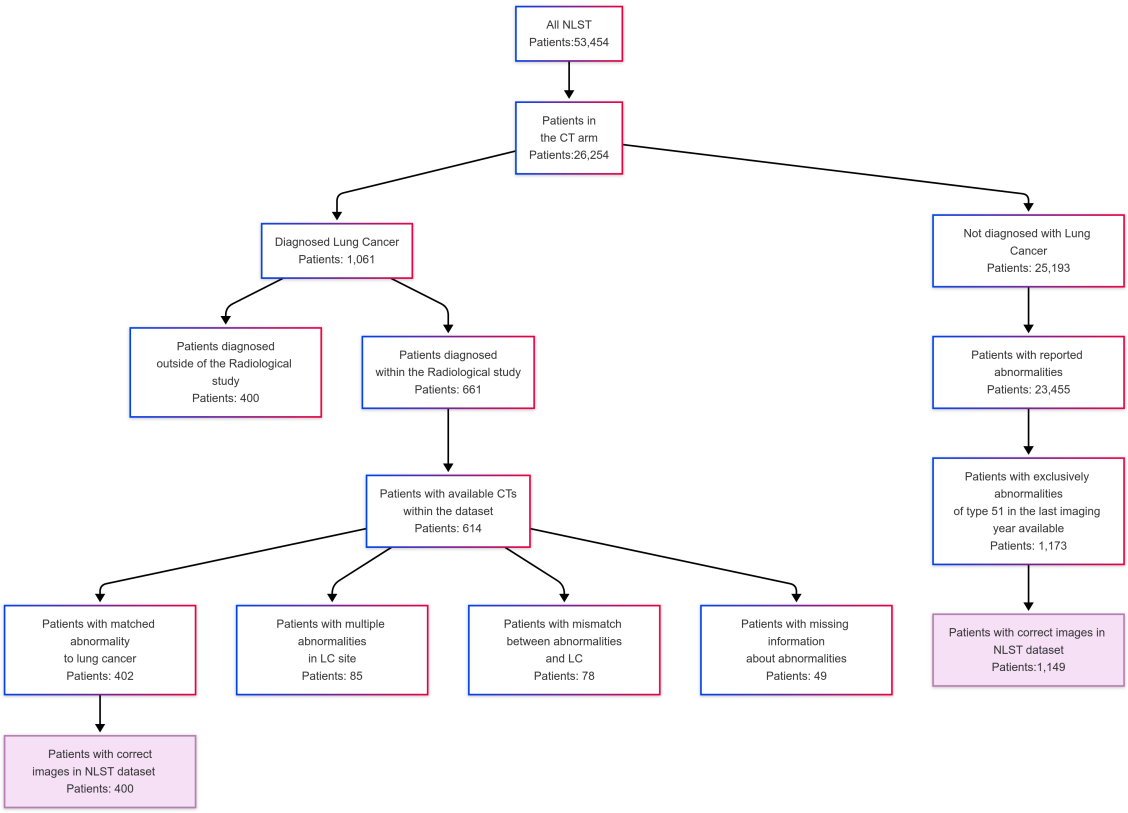


Figure 4.2: Patient selection flowchart from the NLST dataset for imaging analysis. This diagram illustrates the step-by-step inclusion criteria applied to identify eligible CT scan series, distinguishing between patients diagnosed and not diagnosed with lung cancer, and filtering based on radiological availability, abnormality matching, and image presence in the NLST dataset.

[53], who demonstrated improved model performance when using high-quality, lung-specific kernels, this thesis adopts a structured kernel and manufacturer ranking protocol.

In addition to kernel type, scanner manufacturer is also considered. Image characteristics vary across vendors due to differences in hardware, post-processing algorithms, and default reconstruction settings. Empirical evidence from previous studies and qualitative inspection of NLST data suggest that Siemens scanners tend to produce more consistent and diagnostically useful images, followed by GE, Philips, and Toshiba. The ranking procedure described below reflects these considerations.

The following criteria were applied to each candidate series:

1. **Slice-thickness filter.** Any series with `SliceThickness` $> 5\,\text{mm}$ was discarded, as thicker slices reduce spatial resolution and can obscure small nodules.

2. **Kernel ranking.** The remaining series were ranked according to a predefined hierarchy shown in Table 4.2. Manufacturers were prioritized in the following order: Siemens, GE, Philips, and Toshiba. Within each manufacturer, kernels were ordered from most to least

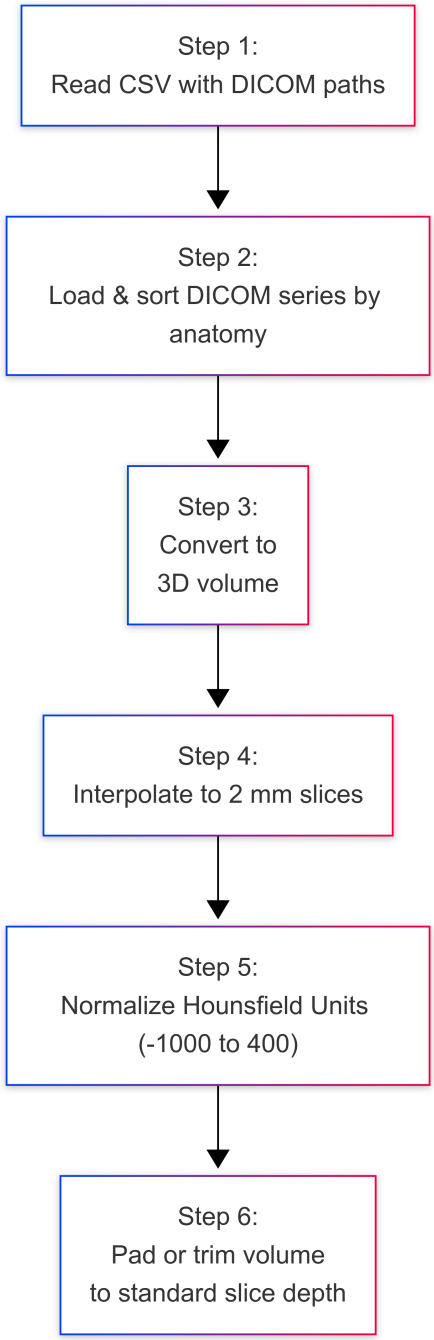


Figure 4.3: Scan pre-processing pipeline, using slice thickness resizing, Hounsfield unit normalization and slice trimming/padding to achieve standard slice depth.

desirable, favoring those optimized for lung parenchyma. If a series' manufacturer–kernel pair did not appear in the table, it was assigned an infinite rank and excluded.

3. **Tie-breaking.** If two or more series shared the same (lowest) kernel rank, the series with the highest DICOM `SeriesNumber` was selected. This heuristic favors reconstructions that span the largest z-extent or were generated most recently by the scanner.

Table 4.2: Kernel preference hierarchy used in Step 2 (lower rank → higher priority).

Manufacturer	**Kernel order (most → least preferred)**
Siemens	B50f → B45f → B50s → B40f → B41s → B60f → B60s → B70f → B36f → B35f → B30f → B31s
GE Medical Systems	LUNG → BONE → BODY FILTER/BONE → STANDARD → BODY FILTER/-STANDARD → SOFT → EXPERIMENTAL7 → BODY FILTER/EXPERIMEN-TAL7
Philips	D → C → B → A
Toshiba	FC51 → FC50 → FC52 → FC53 → FC30 → FC11 → FC10 → FC82 → FL04 → FC02 → FC01 → FL01

This procedure ensures that each subject contributes one high-resolution, lung-appropriate CT series selected according to reproducible and empirically motivated criteria. The number of slices and their thickness were recorded for each selected series, and the most common values across the dataset were later used to standardize scan dimensions during preprocessing. Overall, this approach minimizes protocol-induced variability and enhances the robustness and generalizability of the downstream deep learning models.

4.3.2 Resampling

Resampling is a critical preprocessing step when working with CT data, as it addresses the variability in acquisition parameters across patients and scanners. In particular, CT scans often differ in slice thickness along the axial (z) dimension, which leads to inconsistencies in the spatial resolution between slices. This heterogeneity poses challenges for 3D convolutional models, which assume uniform spacing between voxels in all three dimensions.

In this study, in-plane resolution (x and y axes) was consistent across all scans at 512×512 pixels, so resampling in those dimensions was not necessary. However, slice thickness along the z-axis varied significantly between patients. To standardize the volumetric resolution and ensure uniform input dimensions for the model, all scans were resampled along the axial axis to a fixed slice thickness of 2 mm using linear interpolation.

The value of 2 mm was chosen because it was the most frequently occurring slice thickness among the selected CT series. Choosing the mode value as the resampling target offers several advantages:

- It minimizes the number of volumes that require interpolation, thereby preserving original data fidelity for a larger portion of the dataset.
- It avoids unnecessary upsampling (which can introduce noise) or aggressive downsampling (which may lose critical anatomical details).
- It ensures that the majority of the dataset remains close to its native resolution, reducing the variability introduced by interpolation artifacts.

Standardizing slice thickness allows the model to interpret inter-slice spatial relationships consistently across samples. Without this step, a 5 mm nodule could appear across a different number

of slices depending on the original resolution, potentially confusing the model and degrading its ability to learn shape, volume, and context.

By resampling all scans to a common 2 mm axial spacing, the dataset achieves a harmonized spatial structure that enhances the comparability of input volumes, improves model generalization, and aligns with best practices in medical image preprocessing.

4.3.3 Normalization

CT images are measured in Hounsfield Units (HU), which can vary significantly based on tissue type and scanner configuration [54]. To standardize the intensity values and improve the stability of model training, a normalization step was implemented. Specifically, voxel intensities were clipped to a range of -1000 to 400 HU, which corresponds roughly to the values for air and soft tissue, respectively. This range was selected to suppress irrelevant high-density structures, such as bone, while preserving clinically meaningful tissue contrasts. After clipping, the values were linearly scaled to fit within the range of [0, 1], resulting in a normalized representation suitable for input into neural networks. This process ensures that all images share a consistent intensity distribution, minimizing variability due to differences in scanners or protocols.

4.3.4 Trimming and padding

Additionally, to ensure uniform input dimensions across the dataset, each volume is either trimmed or zero-padded to match the most common slice count observed, thereby standardizing the depth dimension of the scans.

4.3.5 Segmentation

Segmentation was performed to eliminate structures outside the pulmonary parenchyma, allowing all subsequent processing to concentrate on the lung region. Several studies have shown that using a lung mask can reduce the false-positive rate and enhance the sensitivity of nodule detection and malignancy classification networks [55, 56]. Consequently, a binary mask of the left and right lungs was created for each CT volume, and the masked images were subsequently processed by the downstream models.

The CT scans were organized and preprocessed based on information provided in a CSV file that included patient identifiers (PID), the year the study was conducted, and the paths to the DICOM folders.

To segment lung regions from CT scans, a deep learning-based segmentation pipeline was employed, utilizing a Residual U-Net architecture developed in [57]. This model builds upon the classical U-Net design, a well-established standard for biomedical image segmentation, by integrating residual connections that improve gradient propagation and training stability in deeper networks. The decision to adopt this specific model was both methodological and pragmatic: the architecture had already been trained and validated on thoracic imaging data, ensuring its suitability for the lung segmentation task. Furthermore, the model was developed by one of the

thesis co-supervisors, which facilitated seamless integration into the pipeline and eliminated the need for extensive retraining. This combination of architectural robustness, prior validation, and domain expertise made it a highly appropriate choice for accurate and efficient lung segmentation in this study.

Each scan was loaded into a function that read a series of DICOM slices and assembled them into a 3D NumPy array. Once the scan was loaded, segmentation was performed by applying the trained Residual U-Net model to the entire scan volume. The resulting mask was saved in two formats: a NumPy array and a Nearly Raw Raster Data (NRRD) file using SimpleITK, which is suitable for visualization and analysis in tools like 3D Slicer [58]. For the NRRD export, the predicted mask is resized to match the dimensions of the original DICOM slices (512x512). Figure 4.5 is a representation of a CT scan with an automated segmentation mask.

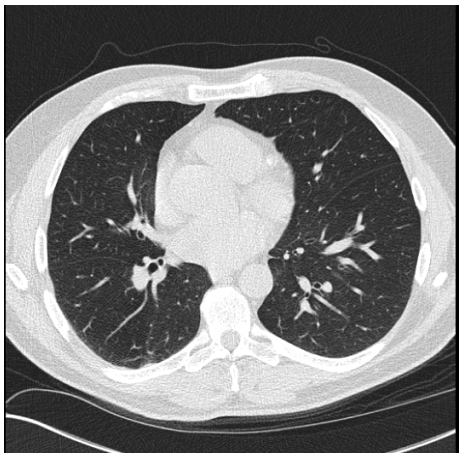


Figure 4.4: Example of an axial slice of CT scan.

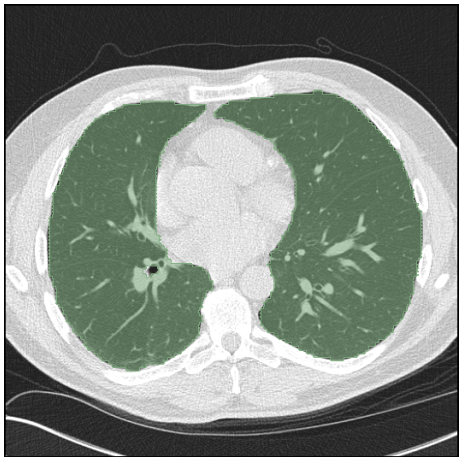


Figure 4.5: Lung CT Scan with segmentation mask.

4.3.6 Manual mask correction

It is important to emphasize that a manual verification process was conducted to ensure that each slice of the scan was accurately segmented using the 3D Slicer tool. In cases where automatic segmentation did not capture the tumor, particularly those located at the edges, or any substantial part of the lung, corrections were made using the manual painting tool available within 3D Slicer. Figure 4.4 shows the example of a lung CT scan from NLST in 3D-Slicer. The corrected mask was then saved. Due to external circumstances, the verification was performed manually and not by a radiologist, making it understandable that achieving 100% accuracy in lung segmentation was not possible. Figures 4.6 is an example of an automated mask that did not segment correctly all the lung area. The area signaled with an arrow shows a tumor that was not considered as part of the lung, thus needing a manual correction. The picture on the right in that same figure shows that same segmentation mask after a manual correction was done.

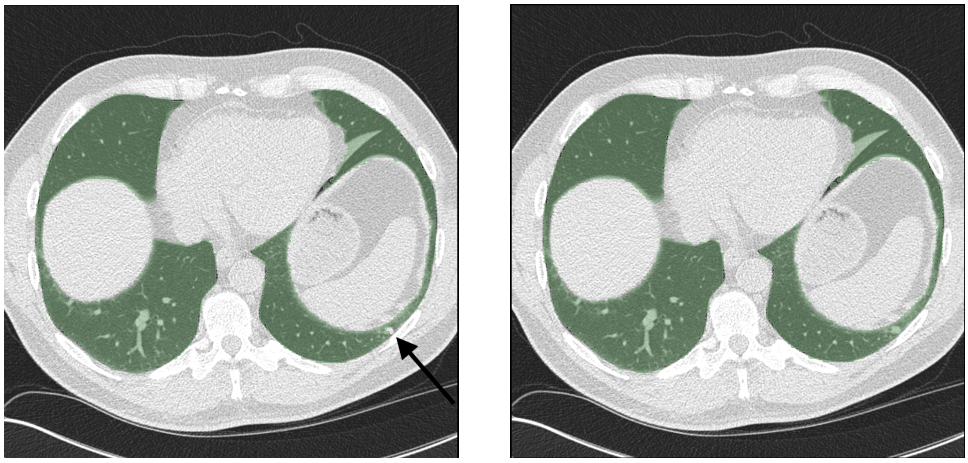


Figure 4.6: Lung CT scans showing the segmentation mask before and after manual correction. An arrow points to the lung nodule that has not been correctly automatically segmented.

4.3.7 Standardization of masks

Following the segmentation process, a mask standardization step is implemented to ensure that all lung masks are geometrically aligned with their corresponding CT scans. This spatial alignment is crucial for accurate model training, visualization, and further quantitative analysis. Without standardization, variations in slice thickness, in-plane resolution, and volume shape could sabotage the comparability of samples across the dataset.

The standardization process begins by interpolating the mask volume along the z-axis to match the slice thickness of the associated CT scan. This is done by comparing the original spacing metadata of the mask with the target spacing of the scan. The interpolation is focused solely on the z-dimension, preserving the x and y resolution during this step. This method ensures that the anatomical proportions of the mask remain accurate in the reconstructed 3D volume.

If the x and y axis of the interpolated mask do not match the axes of the CT scan, a resizing operation is performed. Using PyTorch's transformation utilities, each 2D slice of the mask is resized to the target shape using nearest-neighbor interpolation. This interpolation method is specifically chosen to retain the binary nature of the segmentation, ensuring that the labels remain intact during the resizing process.

Next, the number of slices in the mask volume is standardized to exactly match the number of slices in the corresponding CT scan. This is achieved by either trimming excess slices from the top and bottom of the volume or by symmetrically padding the mask with zero-valued slices. The result is a mask that aligns spatially with the scan and shares its full volumetric shape.

After resizing and standardizing the volume, the mask is validated to confirm that it is binary, containing only zero and one values. If other values are present, a fixed global threshold at 0.5 is applied (voxels with values >0.5 are set to 1 and all others to 0), before casting the result to an 8-bit unsigned integer mask.

By enforcing a uniform spatial structure across all mask volumes, the standardization pipeline greatly enhances the reliability and reproducibility of downstream analyses. It mitigates the effects of scan machine variability and ensures that all masks and scans are directly comparable in shape, resolution and spatial orientation.

4.4 Clinical data pre-processing

This thesis employs a multimodal approach that includes the analysis of clinical data. The clinical data referenced is from the "Participant Dataset" found in [51]. Each row in this CSV dataset represents a single trial participant and is characterized by 326 variables organized into different thematic sections.

Key sections of the dataset include:

- Demographics: This includes variables such as age at randomization, sex, race/ethnicity, and weight;

- Smoking History: This section captures current or former smoking status, cumulative pack-years, and cessation age;

- Screening Exams: This includes the dates and modalities of each CT scan or chest X-ray;

- Lung Cancer: a section that summarizes all confirmed lung cancer events recorded in the trial. For each participant, it indicates whether a case was verified, the broad histologic category, tumor stage and grade and whether initial treatment information was done;

- Extended Follow-Up Death: continues monitoring beyond the active screening period by documenting the days from randomization to death or last known contact. It provides information for the underlying cause of death and indicates whether lung cancer was the certified cause of death;

Additional sections, such as Medical History (which records comorbidities and prior tumor diagnoses), Family Cancer History, and Occupational Exposures, provide valuable covariates for risk stratification.

4.4.1 Feature selection

Before proceeding, a selective review of the columns was conducted. This is necessary because, as mentioned in the section above, certain sections of the dataset are directly correlated with the target (the "cancer" column, which indicates if a patient has either a cancerous tumor or non-cancerous tumor/tumors), specifically the cancer column added through an inner join on the PID from another CSV file, which contains abnormalities and is also available at [51]. Some of these columns provide explicit hints about the target. For example, the column "de_stag" refers to "Lung Cancer Stage," which is a significant indicator of the outcome of a patient.

If a simple thresholding approach were applied, such as eliminating all columns with very high correlation to the target, it would lead to the exclusion of important columns that naturally have a high correlation with the target column. For instance, the column "cigsmok," which indicates whether a person is a current or former smoker, directly influences the likelihood of developing lung cancer and would unjustly be dropped from the analysis.

A thorough analysis was then performed to identify the columns that are clinically relevant without inadvertently revealing the target column. Every variable that explicitly revealed the screening result, such as ‘can_scr‘ or ‘de_stag‘, was removed. In contrast, baseline predictors deemed clinically relevant, such as age, sex, smoking history and COPD status, were retained.

In the beginning, there are 326 columns. After the manual choice of the variables, 152 are kept. The selected columns to retain are available at Appendix A.1. After all the preprocessing, the remaining 82 columns can be found at Appendix A.2.

4.4.2 Missing data handling, normalization and scaling

The dataset exhibited a significant amount of missing values, necessitating an effective imputation strategy. Initially, features with over 80% missing data were identified and excluded to maintain data integrity and reliability. Additionally, threshold values of 75%, 70%, and 65% were tested. Ultimately, it was found that applying a 70% threshold led to a substantial improvement in the overall values. There were no improvements between the 70% threshold and the 65% threshold, so the 70% one was left unchanged. After applying the 70% threshold, 82 variables remained.

Missing values of the maintained columns were addressed through an imputation approach that calculated feature-specific statistics for each class separately during the training phase. The classes were defined as "Cancer" and "Non-cancer", as available on the CSV, recognizing that patients may have different clinical information depending on their cancer status. Therefore, it would be inappropriate to ignore the class and impute the information without consideration of this distinction. Missing values in numeric features were imputed using class-specific means to preserve the internal distribution within each class [59]. The resulting values were then standardized using the `StandardScaler` [60] from the `scikit-learn` library. For categorical features, missing values were imputed using class-specific modes. These features were subsequently encoded using one-hot encoding via scikit-learn’s `OneHotEncoder` [61].

During validation, missing data was filled in using statistics from the training set only, to prevent any data leakage between the two sets. At test time (or in validation folds), global averages were used as a fallback, since the true label is unknown. This procedure ensures that imputation does not introduce bias while still utilizing class-specific patterns observed in the training set.

4.4.3 Dimensionality reduction

The primary features were utilized largely as they were, after undergoing cleaning and encoding. Dimensionality reduction was explored as a form of feature engineering by including an optional Principal Component Analysis (PCA) step in the modeling pipeline. This step aimed to reduce feature dimensionality while retaining most of the variance. PCA with a retention of 95% variance was tested to compress correlated features into fewer components. The motivation behind this approach was to determine whether reducing noise and colinearity through PCA could enhance model generalization. Figure 4.7 represents the pipeline of the clinical data preprocessing.

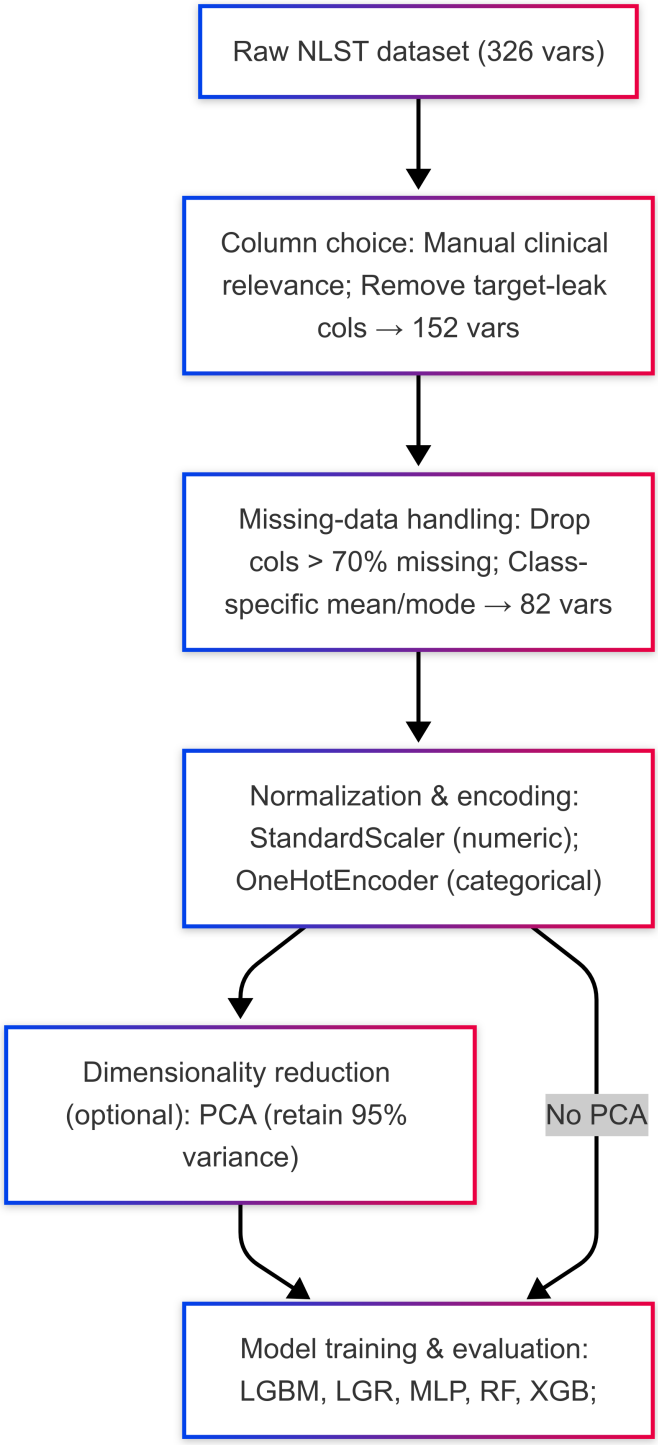


Figure 4.7: Clinical Data Pipeline, showing column choice, missing data handling, normalization, scaling and dimensionality reduction.

Chapter 5

Single and Multimodal Approaches

This chapter describes the experimental methods used to develop and evaluate both single-modality and multimodal models for lung cancer classification. Section 5.1 focuses on models trained exclusively on either CT imaging or clinical data, while Section 5.2 presents fusion strategies designed to integrate both data types.

5.1 Single-Modality Models

This section describes models created using either CT imaging or structured clinical data.

5.1.1 Imaging data

5.1.1.1 Architectures

The ResNet-18 convolutional neural network architecture [62] was utilized for the binary classification of lung CT slices as either cancerous or non-cancerous. The canonical 2-D variant was adopted, so every convolution and pooling layer operates only in the axial plane; each slice is processed independently rather than as part of a 3-D volume. ResNet-18 is a compact residual network consisting of 18 layers, designed to address the vanishing gradient problem by incorporating identity-based skip connections. This design allows for effective training of deeper architectures without compromising performance.

The ResNet-34 architecture [62] was likewise implemented in its 2-D form, with kernels sliding across height and width only. It was also used to perform binary classification of lung CT slices into cancerous and non-cancerous categories. ResNet-34 is a deeper variant within the residual network family, consisting of 34 layers. Similar to ResNet-18, it utilizes identity-based skip connections to enhance gradient flow, which allows for more effective training of deeper networks. This kind of ResNet was implemented so that its results could be compared with the ones from ResNet-18.

5.1.1.2 Optimization Process

Three-fold cross-validation was adopted to balance computational efficiency with robust performance estimation. In a k-fold scheme, each sample is used as validation data exactly once and forms part of the training set during the remaining $k-1$ iterations, thereby providing a comprehensive assessment of model generalizability [63]. For each fold, the model was trained for a fixed number of 50 epochs, which was empirically determined to allow sufficient convergence while avoiding overfitting and keeping training time manageable.

While 5- or 10-fold cross-validation is commonly used for small to medium datasets, the imaging pipeline in this study is computationally intensive due to 3D CT scan preprocessing, segmentation masks, and data augmentation routines. Increasing the number of folds would result in significantly longer training times without proportionate gains in performance estimation reliability. Empirical testing showed that 3-fold CV offered stable and consistent performance metrics, while keeping model training and evaluation within reasonable time and hardware constraints.

Additionally, for imaging data, it is essential that all slices originating from the same patient remain within a single fold. Otherwise, nearly identical tumor appearances may be present in both training and validation sets, causing data leakage that inflates accuracy and AUC [64]. Accordingly, `StratifiedGroupKFold` [65] was employed to enforce patient-level grouping while preserving class stratification. A fixed random seed (`random_state = 42`) ensured reproducibility. An auxiliary verification step confirmed that each patient identifier appeared in only one split.

Before training the model, a comprehensive grid search was conducted to assess the impact of key hyperparameters on the binary classification of lung nodules. Table 5.1 details the specific hyperparameter values that were tested, which include variations in learning rate, weight decay, batch size, optimizer type, and the use of pretrained ImageNet weights. These selections represent widely explored configurations in CNN training, as they significantly affect model convergence and generalization performance.

Table 5.1: Grid search configuration for CNN hyperparameters. Each combination was tested across folds to identify the most effective settings for classification performance.

Hyperparameter	**Values Tested**
Learning Rate	`1e-4`, `1e-3`
Optimizer	Adam
Weight Decay	`0`, `1e-4`
Batch Size	`16`, `32`
Pretrained Weights	True (ImageNet), False (random init)

All possible combinations of the defined hyperparameter values were evaluated using a three-fold cross-validation strategy, as mentioned above. This approach ensured a robust performance estimation and allowed for the selection of the most effective configuration across different data

partitions. Metrics such as accuracy, precision, recall, F1-score, and validation loss were also recorded for each run. Model selection was based on AUC values.

Each training example is one axial CT slice stored as a $3 \times 512 \times 512$ tensor:

- Channel 1 – raw slice (grayscale intensities);
- Channel 2 – same slice masked pixel-wise by the binary mask, so voxels outside the lungs are set to 0;
- Channel 3 – the binary mask itself ($\{0, 1\}$).

The third channel specifically encodes the binary lung mask, which serves as a spatial guide to differentiate lung tissue from surrounding structures. By including the mask as a separate channel instead of just using it to eliminate irrelevant areas, the model gains valuable structural context. This approach enables the network to learn the spatial distribution and shape of the lung boundaries, which can indirectly help in distinguishing between malignant and benign regions. This decision was made intentionally to expose the network not only to filtered intensity values but also to the binary anatomical segmentation, providing an additional feature map to assist the early convolutional filters during training.

The 2-D matrices are concatenated along the channel axis and passed unchanged to the original three-channel `Conv2d` layer of ResNet-18. No spatial cropping, padding, or depth stacking is applied. Because every convolution in the network is a `Conv2d`, kernels slide only across height and width, making the backbone a strictly 2-D architecture. The additional channels merely enrich each pixel with complementary information about the same slice and do not introduce a third spatial dimension.

The final fully connected layer was replaced with a single-node output layer appropriate for binary classification. During inference, a sigmoid activation function was applied to convert the output logits into probabilities. Binary predictions were generated by applying a threshold of 0.5 to the sigmoid output.

Two initialization strategies were evaluated: random initialization and initialization with ImageNet-pretrained weights, as referenced in the hyperparameter tuning paragraph. This comparison aimed to assess the impact of transfer learning on classification performance.

Model training was carried out using the binary cross-entropy loss function (`BCEWithLogitsLoss`), which is suitable for binary classification involving raw logits.

5.1.1.3 Use of Multi-Channel Inputs with Segmentation Masks

A key design decision in the imaging model pipeline involved the use of a three-channel input tensor per CT slice. This strategy incorporates not only the raw slice (grayscale image), but also two mask-related channels: one where only the lung area is preserved (masked slice) and another that explicitly encodes the binary lung segmentation mask. The goal of this setup was to enrich the input representation with structural context. Including the binary mask as an input channel provides spatial cues that help the model focus on anatomically relevant regions while avoiding

irrelevant background noise such as ribs, muscle, or artifacts. The masked image channel reinforces this focus by filtering out voxels outside the lung region, which may otherwise bias feature extraction.

This approach enables early convolutional layers to learn more localized, lung-specific filters and to implicitly incorporate shape and region priors into the feature maps. Rather than relying solely on raw intensities, the network can leverage the spatial boundaries of the lungs, which may support better discrimination of peripheral nodules or exclude confounding regions. This is especially useful when nodules are located near pleural surfaces or boundaries, where noise and ambiguity are more likely.

Alternative single-channel configurations (e.g., raw image only or masked image only) were considered, but preliminary ablation experiments showed that the three-channel approach led to more stable convergence and better generalization performance across folds. However, the use of segmentation masks also increases preprocessing overhead and relies on the accuracy of the segmentation pipeline. Future work could evaluate whether learned attention mechanisms could replace the need for explicit anatomical masks.

5.1.1.4 Augmentation techniques

Both the ResNet-18 and ResNet-34 experiments were trained with an identical set of augmentation techniques.

To improve generalization and reduce overfitting [66], extensive data augmentation was incorporated into the training pipeline. This strategy aimed to simulate the variability commonly found in real-world clinical CT scans and to artificially enhance the diversity of training examples without the need for additional annotated data.

The augmentation process utilized the Albumentations library [67], which offers a highly efficient and flexible framework for image transformation.

The applied transformations included:

- Geometric Distortions: Random horizontal flips (with a probability of 0.5) and affine transformations such as shifting, scaling, and rotation, with configurable limits.

- Photometric Changes: Controlled variations in brightness and contrast, along with the application of Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance local contrast.

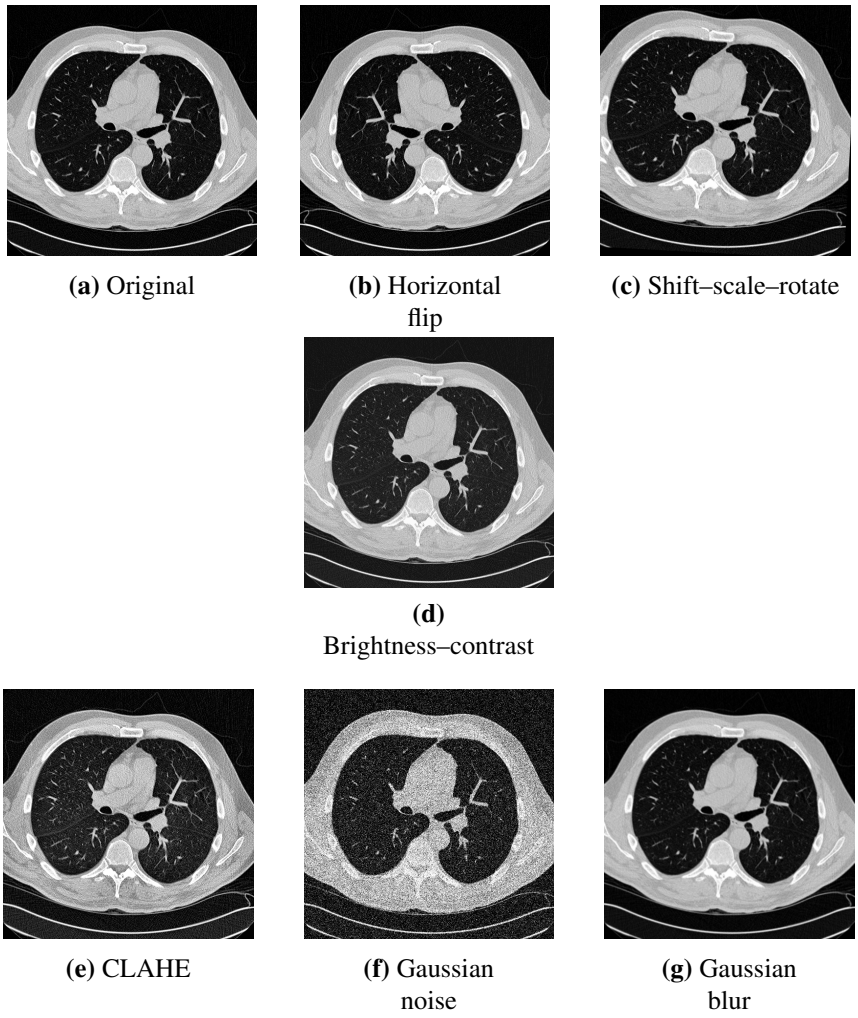
- Noise Injection: Stochastic application of Gaussian noise and blur operations (including motion blur, median blur and Gaussian blur) to simulate sensor or reconstruction noise.

- Conversion: The augmented images were converted into PyTorch tensors using `'ToTensorV2()'`.

The different augmentations performed can be observed in Figure 5.1.

All augmentations were applied during training only. A separate, deterministic preprocessing pipeline was defined for validation. This validation pipeline utilized only CLAHE and tensor conversion to ensure consistent evaluation and minimize distributional shifts during inference.

This augmentation strategy was systematically evaluated as part of a comprehensive hyperparameter search. Experiments were conducted across the three cross-validation folds, and performance was monitored using the same evaluation metrics as the ResNet-18 baseline model.



(a) Original **(b)** Horizontal flip **(c)** Shift–scale–rotate

(d) Brightness–contrast

(e) CLAHE **(f)** Gaussian noise **(g)** Gaussian blur

Figure 5.1: Visual impact of each augmentation applied independently to the same axial CT slice. The original image (a) is followed by six deterministic transformations used.

5.1.2 Clinical data baseline

Hyperparameter optimization for each classification model was conducted using the `GridSearchCV` utility from the `scikit-learn` library [68]. This approach allowed for an exhaustive evaluation of predefined hyperparameter combinations within a stratified cross-validation framework. For logistic regression, the regularization strength parameter C was varied across a logarithmic range (e.g., 0.01 to 10) to assess the impact of L_2 regularization. Tree-based models, including Random Forest and gradient-boosted classifiers (XGBoost, LightGBM), were tuned by adjusting parameters such as the number of estimators, maximum tree depth, learning rate (for boosting), and minimum samples per leaf. The multilayer perceptron (MLP) was included in the search space with variations in the number of hidden units (e.g., 50, 100, 200), number of layers, and L_2 penalty term (α), along with adjustments to the learning rate. `GridSearchCV` systematically evaluated all parameter configurations using internal cross-validation. This process ensured that each model was trained and assessed under consistent conditions, reducing selection bias and enabling fair

comparative evaluation. The different models and parameters can be found at A.3.

5.1.2.1 Logistic Regression

Logistic regression is widely utilized in lung cancer risk modeling and serves as a well-calibrated baseline [69]. To prevent overfitting, an L2-regularized logistic regression was employed, testing different regularization strengths.

5.1.2.2 Random Forest

Random Forests demonstrate resilience to outliers and effectively model complex relationships. They have been used successfully to analyze clinical risk factors associated with lung cancer [70]. In the experiments, the number of trees (estimators) and the depth of the trees were adjusted. The ensemble approach of Random Forests also offers an implicit ranking of feature importance.

5.1.2.3 Gradient Boosted Trees (XGBoost and LightGBM)

These models create an additive ensemble of weak learners, optimizing performance progressively. Previous research demonstrates that boosting methods can effectively predict lung cancer risk by leveraging their capability to handle heterogeneous data and address imbalances [70]. Hyperparameters such as the learning rate, maximum tree depth, and the number of leaves were fine-tuned.

5.1.2.4 Multilayer Perceptron

The MLP was relatively small given the limited number of input features. Neural networks can, in theory, model complex patterns, but they require careful regularization with limited data [71]. Early stopping, through a fixed iteration cap, and L2 penalties were applied to the MLP.

5.2 Multimodal Fusion Models

This chapter explains how information from structured clinical records and CT images was transformed into numerical form and subsequently combined to aid in the binary classification of lung cancer. All low-level preprocessing steps, including handling missing values, normalization, and CT resampling, are detailed in Chapter 4 and will not be repeated here.

The clinical feature vectors were created using relevant attributes for each participant, such as age, sex, smoking behavior, comorbidities, and screening outcomes, and were represented as a single row vector of real numbers. After standardization, this vector was passed unchanged to the fusion models.

5.2.1 Imaging Feature Vectors

Only the CT slices with tumors were processed. Therefore, any patient with multiple nodules contributed with the corresponding number of slices. Each of those slice was processed through

a CNN selected from the ResNet family [72]. Vectors are therefore created at the nodule level, a strategy shown to preserve nodule-level heterogeneity in large thoracic CT datasets [73].

For each ResNet model (ResNet-18 and ResNet-34), two versions of feature extraction were performed: one using raw slices and another where the same slices were fine-tuned with data augmentation applied via Albumentations. This resulted in four sets of imaging vectors: ResNet-18, ResNet-18 (Aug), ResNet-34, and ResNet-34 (Aug). Augmentation was used only during training. During the final forward pass used to extract embeddings for evaluation, augmentation was disabled to ensure consistency [53].

After the last residual block in each network, a global average pooling layer was applied. This produced a 512-dimensional feature vector. The 512-dimensional output is a standard design of both ResNet-18 and ResNet-34, determined by their final convolutional layer. Keeping this default size enables the use of ImageNet-pretrained weights and provides a good balance between representation quality and model complexity. Increasing this dimensionality typically does not improve performance on small to medium-sized medical datasets and may lead to overfitting and higher memory consumption [74].

5.2.2 Fusion Strategies

Two complementary strategies are explored for combining the imaging and clinical vectors: intermediate fusion and late fusion.

Intermediate fusion — concatenation strategy. In this approach, imaging and clinical features are extracted independently as described in Sections 5.2 and 5.2.1. Each nodule is represented by two frozen embeddings: a 512-dimensional vector produced by the ResNet encoder and a 113-dimensional clinical vector obtained after median/mode imputation, standardization, and one-hot encoding of categorical attributes. After the missing-value filter, 82 clinical variables remain. Each numeric variable keeps one column, whereas every categorical variable is split into several binary indicators by one-hot encoding. This expansion adds 31 extra columns, so the clinical branch ends with 113 features in total.

Concatenating these embeddings yields a 625-dimensional feature vector that is processed by a classification head consisting of a fully connected layer with 128 units, a ReLU activation, dropout with a rate of 0.3 [75], and a final linear layer producing a single logit. This design follows the operation-based feature-level fusion paradigm described at [76], where fusion is performed via non-trainable operations and all learnable parameters reside in the classification head.

Intermediate fusion — attention strategy. The attention-based variant reuses the same frozen clinical and imaging vectors described previously. Before fusion, a lightweight channel-wise attention mechanism modulates the image features based on the clinical embedding. Specifically, the clinical vector is passed through a linear layer to produce a gating vector, which is activated with a sigmoid function and applied element-wise to the image embedding to reweight its channels.

The gated image vector is then concatenated with the original clinical vector and passed through the same classification head used in the concatenation strategy. This architecture implements a simple form of cross-modal attention, where one modality (in this case, clinical) influences the representation of another (the imaging). The design is consistent with the channel-wise attention mechanisms surveyed by [76], in which non-image modalities are used to adaptively recalibrate visual features to improve multimodal integration.

Figure 5.2 represents the two different intermediate fusion approaches used.

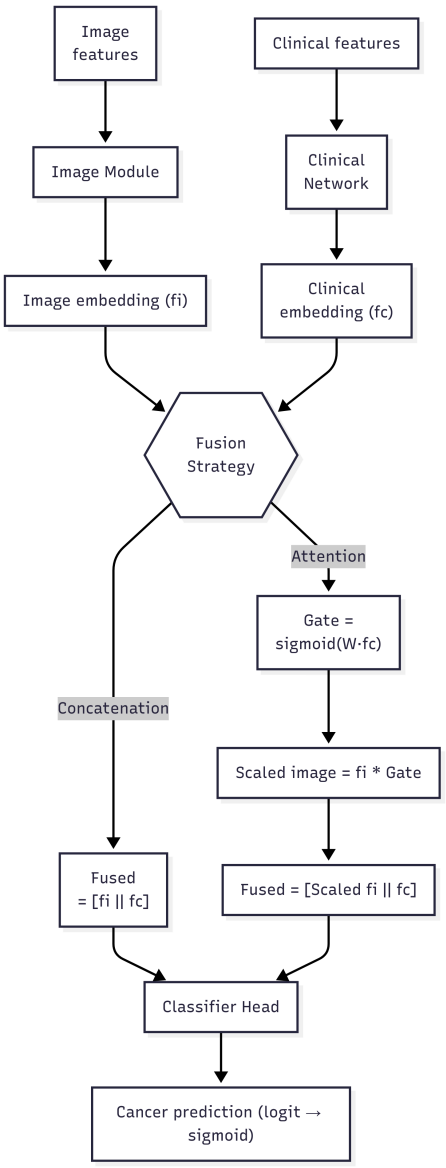


Figure 5.2: Intermediate fusion pipeline, showing the two different approaches.

Late fusion: In the late-fusion scheme, the imaging and clinical feature vectors are first processed separately by lightweight MLPs that each produce a single logit score. These two logits are subsequently merged through a weighted average governed by a coefficient α (fixed at 0.5 for

equal trust or learned during training). The fused score is transformed into a probability with the sigmoid function and optimized via binary cross-entropy. Because the modalities are kept apart until this final step, the method is agnostic to differing feature dimensions.

Figure 5.3 represents the two different fusion paradigms employed.

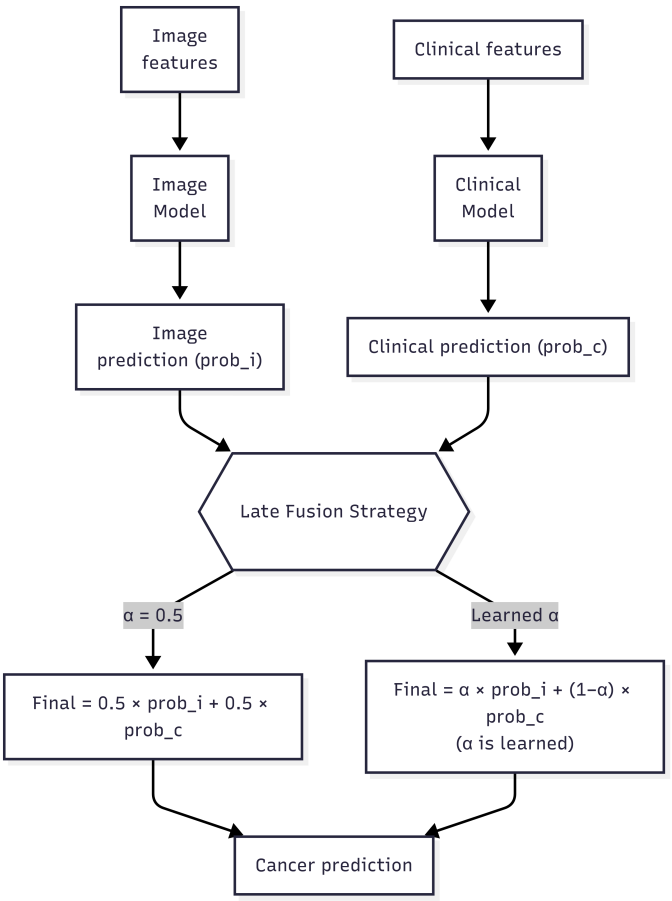


Figure 5.3: Late fusion pipeline, showing the two different weight strategies.

Summary

This chapter outlined the experimental methodology used for the development and evaluation of both single-modality and multimodal models aimed at binary lung cancer classification. The primary goal is to assess the performance of models trained independently on either CT images or structured clinical data, as well as to explore approaches that integrate both modalities to improve predictive accuracy.

Single-Modality Models. For the imaging-only models, CNNs utilizing the ResNet-18 and ResNet-34 architectures were employed. A three-fold `StratifiedGroupKFold` cross-validation scheme was implemented to maintain patient-level grouping and prevent data leakage. Grid search

techniques were applied to optimize key hyperparameters. Model performance was evaluated using metrics such as AUC, accuracy, precision, recall, F1-score, and loss. To improve generalizability, data augmentation techniques were employed.

For the models trained on clinical data, a variety of classifiers were assessed, including logistic regression, random forests, gradient-boosted trees, and MLP. Hyperparameter tuning for these classifiers was performed using `GridSearchCV`.

Multimodal Fusion Models. Imaging features were extracted from the trained ResNet models using global average pooling, resulting in 512-dimensional embeddings. Clinical features, including age, sex, smoking status, and comorbidities, were standardized and presented as 113-dimensional vectors. Each nodule-level CT slice produced one image feature vector, thereby preserving heterogeneity within patients.

Two fusion strategies were explored. The first was an intermediate fusion approach, which included two methods. The first method involved concatenating the image and clinical embeddings into a single 625-dimensional vector that was then passed through a fully connected classification head. The second method utilized channel-wise attention, where the clinical vector gated the image embedding via a learned sigmoid-modulated linear transformation. This gated image vector was then concatenated with the clinical vector and classified similarly.

In the late fusion strategy, each modality was processed independently by a lightweight MLP to generate logits. These logits were combined using a weighted average, either with a fixed coefficient ($\alpha = 0.5$) or a learned value, and then passed through a sigmoid activation function to yield the final prediction. This approach maintained modality independence until the final decision step and is agnostic to the dimensionality of the input vectors.

Chapter 6

Results and Discussion

This chapter reports the quantitative performance of all pipelines developed in this work. Unless stated otherwise, every model was evaluated with the patient–stratified, three-fold cross-validation protocol described in the previous chapter.

6.1 Single–Modality Baselines

This section presents the results of models trained using only one type of data, either clinical variables or CT images. These unimodal baselines serve as a reference point for assessing the benefits of combining both data types in the multimodal fusion models that will be discussed later.

6.1.1 Clinical Data

The performance of the best clinical-only baseline is presented in Table 6.1. A Random Forest classifier yielded the highest overall results among the clinical models evaluated, outperforming other classical baselines such as logistic regression, gradient-boosted trees, and MLPs.

Considering the relatively moderate number of input features in this clinical dataset, limited to tens of variables after preprocessing, PCA did not significantly reduce the dimensionality while maintaining 95% variance. Instead, it primarily served as a comparison to the full feature set.

Across the three folds of cross-validation, the Random Forest demonstrated strong generalization, achieving a mean AUC of 0.8954 and a standard deviation of 0.0149. The model also achieved a mean F1 score of 0.7832, with a precision of 0.8541 and recall of 0.7352, reflecting a conservative decision boundary that favored minimizing false positives over maximizing sensitivity.

The cross-entropy loss averaged 0.4606, reflecting moderate model confidence. While this value is not directly interpretable in isolation, it allows comparison across different modalities and fusion strategies, as done in subsequent sections.

Although clinical data alone yielded strong predictive performance, particularly in terms of AUC and precision, the comparatively lower recall value suggest a conservative decision boundary that prioritizes avoiding false positives, possibly at the cost of missing some positive cases. These

results serve as a unimodal baseline for evaluating the added value of imaging data and multimodal fusion explored in later sections.

The clinical baseline was implemented using a *Random Forest* model with three different hyper-parameter configurations: (1) `max_depth=20`, `n_estimators=200`; (2) `max_depth=10`, `n_estimators=200`; and (3) `max_depth=None`, `n_estimators=200`.

Table 6.1: Performance of the best clinical baseline using the Random Forest model.

Fold	Loss	F1	Precision	Recall	Acc	AUC
0	0.4917	0.7016	0.8933	0.5776	0.7664	0.8772
1	0.4504	0.8390	0.8600	0.8190	0.8778	0.9138
2	0.4397	0.8091	0.8091	0.8091	0.8409	0.8952
Mean	0.4606	0.7832	0.8541	0.7352	0.8284	**0.8954**
Std	0.0224	0.0590	0.0346	0.1116	0.0463	0.0149

6.1.2 Imaging Data

Table 6.3 summarizes the optimal hyperparameter combinations that yielded the best performance for each CNN configuration. These results were selected from an initial grid of 16 configurations, presented in Table 6.2.

Table 6.2: Grid of hyperparameter combinations evaluated for single-modality CNN models (optimizer fixed as Adam).

Combination ID	Learning Rate	Weight Decay	Batch Size	Pretrained
1	1×10^{-3}	0	16	True
2	1×10^{-3}	0	16	False
3	1×10^{-3}	0	32	True
4	1×10^{-3}	0	32	False
5	1×10^{-3}	1×10^{-4}	16	True
6	1×10^{-3}	1×10^{-4}	16	False
7	1×10^{-3}	1×10^{-4}	32	True
8	1×10^{-3}	1×10^{-4}	32	False
9	1×10^{-4}	0	16	True
10	1×10^{-4}	0	16	False
11	1×10^{-4}	0	32	True
12	1×10^{-4}	0	32	False
13	1×10^{-4}	1×10^{-4}	16	True
14	1×10^{-4}	1×10^{-4}	16	False
15	1×10^{-4}	1×10^{-4}	32	True
16	1×10^{-4}	1×10^{-4}	32	False

All models were trained for 50 epochs using the Adam optimizer. The learning rate of 1×10^{-4} consistently delivered the best results across experiments. Most models benefited from using ImageNet-pretrained weights. However, the best-performing ResNet-34 (Aug) model achieved its highest AUC with random initialization, suggesting that data augmentation can sometimes offset the advantages of transfer learning.

The hyperparameter settings shown in Table 6.3 correspond to the final configurations used to produce the performance metrics reported in Table 6.4. The optimizer used, the learning rate and the number of epochs trained were a constant throughout the best hyperparameter combinations.

Table 6.3: Optimal hyperparameter combinations per architecture (excluding optimizer, learning rate, and epochs).

Model	**Batch Size**	**Weight Decay**	**Pretrained**	**Augmentation**
ResNet-18	16	0	Yes	No
ResNet-18 (Aug)	16	1e-4	Yes	Yes
ResNet-34	16	0	Yes	No
ResNet-34 (Aug)	32	0	No	Yes

A small weight decay of 1e-4 was applied in ResNet-18 with augmentation, reflecting the need for stronger regularization when heavy augmentation is utilized. These choices align with the methodology outlined previously.

Two convolutional backbones, ResNet-18 and ResNet-34, were trained on CT images independently, both with and without data augmentation, using the three-fold validation scheme described in previous chapters. Table 6.4 shows that data augmentation consistently improved performance. For instance, the ResNet-18 model trained with Albumentations augmentation achieved an AUC of 0.7583, compared to 0.7169 without augmentation. Similarly, ResNet-34, which is a deeper network, produced a higher AUC than ResNet-18 in comparable settings (0.7361 versus 0.7169 without augmentation). The augmented ResNet-34 model recorded the highest single-modality AUC of 0.7600.

The detailed performance of each top-performing single-modality model is presented in Table 6.4. The lower AUCs observed in the imaging-only pipelines are possibly a result of using the entire 512×512 CT slice instead of focusing on nodule-centered crops with bounding boxes. By including the full slice, anatomical background is introduced, which acts as noise. This noise dilutes lesion-specific information and reduces the system's ability to discriminate effectively. This issue occurred because radiologist-defined bounding boxes were not available for this study. The augmented ResNet-34 achieved the highest average AUC of 0.7600 and the best precision at 0.9053. However, the highest F1 score was actually produced by the augmented ResNet-18, which reached 0.7216. Augmentation enhanced the precision-recall balance for both architectures. For instance, ResNet-18's precision increased from 0.7722 to 0.8059, while the recall only dipped slightly, and showed improvements in F1 and AUC scores. The trend of lower loss continued for ResNet-18, but this was not the case for ResNet-34, suggesting that augmentation may sometimes sacrifice cross-entropy in exchange for better calibration. Overall, the augmented models demonstrated

modest but consistent improvements in AUC and F1, establishing a reliable imaging-only baseline before incorporating clinical fusion.

Table 6.4: Performance metrics for the best-performing configuration of each single-modality model, averaged across the three cross-validation folds.

Model	Recall	Precision	Loss	F1	Acc	AUC
ResNet-18	1.0000	0.7722	5.6241	0.6773	0.6193	0.7169 ± 0.0212
ResNet-18 (Aug)	0.9409	0.8059	1.6185	0.7216	0.7063	0.7583 ± 0.0472
ResNet-34	0.9592	0.7573	2.1300	0.6941	0.6896	0.7361 ± 0.0183
ResNet-34 (Aug)	0.9967	0.9053	2.9262	0.7055	0.6804	**0.7600 ± 0.0314**

6.2 Multi-modality Baselines

This section discusses the performance of fusion models that combine clinical and imaging features. Both intermediate and late fusion strategies are assessed to evaluate how multimodal learning can improve classification compared to single-modality models.

6.2.1 Intermediate Fusion

Intermediate fusion combines a 512-dimensional image embedding with a 32-dimensional clinical feature vector, either by straightforward concatenation or by a lightweight, clinically guided attention gate. As shown in Table 6.5, the highest mean AUC is achieved by ResNet-34 (Aug) with attention (0.9138 ± 0.0113), followed closely by the same backbone using plain concatenation (0.9108 ± 0.0116).

For the shallower ResNet-18 without augmentation, concatenation slightly outperforms attention (0.9085 vs. 0.8989). When strong augmentation is applied, however, the attention gate provides a clear boost with ResNet-18, raising AUC from 0.7764 to 0.8061, an absolute gain of 0.0297. ResNet-34 also benefits from attention even without augmentation (0.7999 vs. 0.7877), and the advantage, though smaller, persists after augmentation (0.9138 vs. 0.9108).

Across all backbones, incorporating the 82 clinical features consistently improves accuracy and F1 scores compared to using CT images alone. For example, the AUC for ResNet-34 (Aug) increased from 0.7600 (CT-only) to 0.9108 with concatenation and 0.9138 with attention-based fusion. These results answer Research Question 3 by demonstrating that feature-level fusion, particularly when guided by attention, leverages complementary clinical information that single-modality networks overlook.

Table 6.5: Intermediate fusion – mean ± std. Best values in bold.

Backbone	Fusion	Recall	Precision	Loss↓	F1	Acc	AUC
ResNet-18	Concatenation	**0.9541 ± 0.0649**	0.9099 ± 0.0790	0.4151 ± 0.0576	0.8423 ± 0.0218	0.8346 ± 0.0227	0.9085 ± 0.0150
ResNet-18	Attention	0.9299 ± 0.0809	0.8606 ± 0.0420	0.4243 ± 0.0664	0.8338 ± 0.0226	0.8314 ± 0.0216	0.8989 ± 0.0259
ResNet-18 (Aug)	Concatenation	0.7102 ± 0.0749	0.7231 ± 0.0632	0.6675 ± 0.0468	0.7105 ± 0.0316	0.7034 ± 0.0286	0.7764 ± 0.0513
ResNet-18 (Aug)	Attention	0.7344 ± 0.0612	0.7616 ± 0.0656	0.6154 ± 0.0316	0.7424 ± 0.0384	0.7380 ± 0.0385	0.8061 ± 0.0526
ResNet-34	Concatenation	0.8637 ± 0.0843	0.7255 ± 0.1255	0.6220 ± 0.1506	0.7612 ± 0.0773	0.7324 ± 0.0962	0.7877 ± 0.0963
ResNet-34	Attention	0.8252 ± 0.0294	0.7354 ± 0.1266	0.5708 ± 0.1220	0.7620 ± 0.0757	0.7340 ± 0.0934	0.7999 ± 0.0886
ResNet-34 (Aug)	Concatenation	0.8406 ± 0.0189	0.8720 ± 0.0239	0.5311 ± 0.0365	0.8414 ± 0.0205	**0.8424 ± 0.0169**	0.9108 ± 0.0116
ResNet-34 (Aug)	Attention	0.8375 ± 0.0193	**0.9062 ± 0.0138**	**0.4504 ± 0.0284**	**0.8422 ± 0.0167**	**0.8424 ± 0.0159**	**0.9138 ± 0.0113**

6.2.2 Late Fusion

In late fusion, the image and clinical branch prediction scores are combined by simple averaging, optionally with a learned weighting factor. As shown in Table 6.6, the best result is obtained with ResNet-18 without augmentation, giving a mean AUC of 0.9104 ± 0.0142 when the weights are learned and 0.9103 ± 0.0145 with fixed 50–50 weights. For ResNet-18 (Aug), the AUC scores were 0.7649 and 0.7648, while ResNet-34 (Aug) achieved 0.8834 and 0.8848, depending on whether fixed or learned fusion weights were used. In both cases, the difference between weighting strategies was negligible ($\Delta\text{AUC} \leq 0.002$), indicating that the contribution of each modality remained balanced regardless of the fusion method. This near-parity implies that the two modalities contribute in nearly equal measure.

For every backbone, late fusion falls short of intermediate fusion by two to four AUC points. For example, ResNet-34 (Aug) achieves 0.9138 with attention-based intermediate fusion but only 0.8848 with late fusion. Averaging logits therefore appears less capable of exploiting cross-modal correlations than feature-level fusion. Nevertheless, even this simple ensemble delivers high accuracy, confirming that CT and clinical information each add value to the final prediction.

Table 6.6: Late fusion: mean ± std over the three validation folds.

Imaging Backbone	Learned	Loss	Accuracy	Precision	Recall	F1	AUC
ResNet-18	False	**0.3894 ± 0.0444**	**0.8377 ± 0.0169**	0.8674 ± 0.0264	**0.8652 ± 0.0115**	**0.8415 ± 0.0210**	0.9103 ± 0.0145
ResNet-18	True	0.3901 ± 0.0449	**0.8377 ± 0.0169**	0.8669 ± 0.0264	0.8621 ± 0.0159	**0.8415 ± 0.0210**	**0.9104 ± 0.0142**
ResNet-18 (Aug)	False	0.6038 ± 0.0180	0.6829 ± 0.0062	0.7126 ± 0.0461	0.7015 ± 0.0965	0.6875 ± 0.0251	0.7649 ± 0.0247
ResNet-18 (Aug)	True	0.6040 ± 0.0179	0.6829 ± 0.0062	0.7116 ± 0.0468	0.7015 ± 0.0965	0.6875 ± 0.0251	0.7648 ± 0.0247
ResNet-34	False	0.5603 ± 0.1147	0.7262 ± 0.0888	0.7192 ± 0.1119	0.8221 ± 0.0324	0.7564 ± 0.0748	0.7860 ± 0.0879
ResNet-34	True	0.5592 ± 0.1166	0.7278 ± 0.0910	0.7238 ± 0.1184	0.8251 ± 0.0364	0.7581 ± 0.0772	0.7865 ± 0.0886
ResNet-34 (Aug)	False	0.4924 ± 0.0419	0.8001 ± 0.0417	0.8407 ± 0.0298	0.8066 ± 0.0343	0.8003 ± 0.0448	0.8834 ± 0.0255
ResNet-34 (Aug)	True	0.4928 ± 0.0418	0.8048 ± 0.0380	**0.8438 ± 0.0296**	0.8096 ± 0.0347	0.8053 ± 0.0412	0.8848 ± 0.0259

6.3 Summary of Best Configurations

The clinical variables alone provide a strong model with an AUC of 0.8954. In contrast, the CT images alone achieve only about 0.7600 AUC, even with extensive augmentation. However, when merged, the two modalities using intermediate fusion and apply a small attention gate to weigh the clinical information, the performance reaches its highest point of around 0.914 AUC. A more relaxed late fusion approach, which averages the scores from both branches, also improves performance but falls short of the intermediate level fusion results. In summary, the data indicate that while clinical information sets a solid baseline, the addition of imaging provides extra value, and the most significant benefit arises from fusing both data sources at the feature level rather than at the decision stage.

This pattern is reflected in Table 6.7, which summarizes the mean AUC scores for all models. It highlights the performance gap between single-modality pipelines and the gains achieved through multimodal fusion. The table also shows that intermediate fusion with attention consistently outperforms all other configurations, confirming it as the most effective approach.

Table 6.7: Mean AUC of the best pipelines (single-modality and multimodal).

Category	**Pipeline**	**Model**	**Mean AUC**
Single modality	Clinical only	Random Forest	0.8954
	Imaging only	ResNet-34 (Aug)	0.7600
Intermediate fusion	Clinical + Imaging	ResNet-34 (Aug) + attention	**0.9138**
Late fusion	Clinical + Imaging	ResNet-18 + learned weights	0.9104

Chapter 7

Conclusions and Future Work

This final chapter summarizes the key findings of this thesis, discusses its main limitations, and outlines directions for future research aimed at advancing multimodal AI systems for lung cancer diagnosis.

7.1 Conclusions

This dissertation set out to determine whether fusing CT information with clinical variables could improve the binary classification of lung nodules malignancy over single-modality approaches. Three research questions framed the work.

- **RQ1 – Multimodal preprocessing.** A fully reproducible pipeline was developed to align NLST imaging with tabular records and extract consistent feature vectors (512-dimensional for images and 32-dimensional for clinical data). The protocol proved effective in training deep networks without overfitting, as indicated by a low standard deviation across folds (≤ 0.015 AUC for the best models).
- **RQ2 – Fusion strategy.** Two families of fusion methods were benchmarked. Intermediate fusion yielded the best results, with ResNet-34 (Aug) combined with a lightweight clinical-gated attention block achieving a mean AUC of 0.9138. This performance surpassed all unimodal baselines and outperformed all late-fusion variants.
- **RQ3 – Added value of multimodality.** The clinical-only Random Forest model achieved a solid AUC of 0.8954, while the best CT-only model only reached 0.7600 AUC, despite extensive augmentation. The multimodal model, however, achieved an AUC of 0.9138, which represents a relative improvement of approximately 20.24% over the CT-only approach and 2.06% over the clinical-only approach. This confirms the hypothesis that imaging and structured data encode complementary information.

The findings indicate that while clinical variables provide a solid foundation, they do not fully capture the complete diagnostic picture. Relying solely on imaging fails to provide reliable

screening results within this dataset. However, by combining both clinical variables and imaging through a thoughtfully designed intermediate approach, advantage can be taken of their complementary strengths. This study confirms all three research questions and demonstrates that using different modalities offers a practical and effective way to enhance the accuracy of early-stage lung cancer decision support systems.

7.1.1 Study Limitations

- **Annotation quality.** It was not possible for the lung masks to be reviewed by experts or radiologists. Although quality checks were performed, the absence of a radiologist's expertise may have introduced label noise and boundary inaccuracy.
- **Datasets.** All experiments were conducted exclusively on the NLST dataset. While NLST is a large and well-documented resource, it has limitations in terms of demographic diversity, clinical protocols, and imaging equipment. This raises concerns about the generalizability of the models to external populations, especially those from different geographic regions or healthcare systems. Additionally, relying on a single source restricts the model's ability to learn domain-invariant features, which are crucial for robust deployment in real-world scenarios. Incorporating external datasets from other institutions or countries would help validate the approach and assess its broader applicability.
- **Fusion simplicity.** Only concatenation and a single attention gate were evaluated in the case of intermediate fusion.

7.1.2 Future Work

Future research can explore several promising directions beyond the current study. One area of focus could be the development of more advanced fusion mechanisms, such as transformer-based multimodal encoders or cross-attention modules. These methods may better capture the rich interactions between clinical and imaging modalities.

Another promising direction is the use of temporal data. Incorporating sequential CT scans and time-stamped clinical records could help model disease progression, which is often key in assessing malignancy. For example, a growing nodule or worsening clinical indicators over time may signal cancer more clearly than a single observation. Future models could combine CNN-based image features and longitudinal clinical data using LSTMs or Transformers to capture these trends and improve prediction accuracy.

In addition to architectural advancements, applying explainability techniques specifically designed for multimodal systems would be valuable. This would enable clinicians to understand how each modality contributes to the final decision-making process, thereby building trust and promoting clinical adoption.

Finally, it is essential to assess the generalizability of the model across multiple institutions and diverse populations. This could be achieved through federated learning frameworks or by benchmarking the model on various datasets.

7.1.3 Final Remarks

This thesis shows that even with a modest group of CT scans, combining imaging data with clinical attributes enhances predictive performance compared to traditional models that rely on a single source. Addressing the identified limitations, particularly the need for rigorous annotation and extensive external validation, is a crucial step toward developing tools that are ready for deployment and can effectively support radiologists ain the early detection and management of lung cancer.

References

[1] Centers for Disease Control and Prevention. Lung cancer risk factors. https://www.cdc.gov/lung-cancer/risk-factors/index.html, 2024. Accessed: 2024-11-04.

[2] World Health Organization. Global cancer burden growing, amidst mounting need for services. https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services, 2024. Accessed: 2024-10-06.

[3] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray. Global Cancer Observatory: Cancer Today. https://gco.iarc.fr/today/home, 2020. Accessed: 2024-10-06.

[4] Centers for Disease Control and Prevention (CDC). Lung cancer screening. https://www.cdc.gov/lung-cancer/screening/index.html, Mar 2024. Accessed: 2024-12-19.

[5] Health.com Editorial Team. How is non-small cell lung cancer diagnosed? https://www.health.com/non-small-cell-lung-cancer-diagnosis-8673728, 2024. Accessed: 2024-12-19.

[6] Yawei Li, Xin Wu, Ping Yang, Guoqian Jiang, and Yuan Luo. Machine learning applications in lung cancer diagnosis, treatment and prognosis, 2022.

[7] C.J. Kelly, A. Karthikesalingam, M. Suleyman, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, 2019.

[8] Joana Vale Sousa, Pedro Matos, Francisco Silva, Pedro Freitas, Hélder P Oliveira, and Tania Pereira. Single modality vs. multimodality: What works best for lung cancer screening? *Sensors*, 23(12):5597, 2023.

[9] Cancer Research UK. Lung cancer survival statistics. https://www.cancerresearchuk.org/about-cancer/lung-cancer/survival, 2024. Accessed: 2024-10-06.

[10] Prabhat Jha, Chinthanie Ramasundarahettige, Victoria Landsman, Brian Rostron, Michael Thun, Robert N. Anderson, Tim McAfee, and Richard Peto. 21st-century hazards of smoking and benefits of cessation in the united states. *New England Journal of Medicine*, 368(4):341–350, 2013.

[11] M. Khatami. Cancer; an induced disease of twentieth century! induction of tolerance, increased entropy and 'dark energy': loss of biorhythms (anabolism v. catabolism). *Clinical and Translational Medicine*, 7(1):20, 2018.

[12] B. I. Helgheim, R. Maia, J. C. Ferreira, and A. L. Martins. Merging data diversity of clinical medical records to improve effectiveness. *International Journal of Environmental Research and Public Health*, 16(5):769, 2019.

[13] Lisa Keikes, Simon Medlock, Dorien J van de Berg, Shuchen Zhang, Oscar R Guicherit, Cornelis J A Punt, and Martijn G H van Oijen. The first steps in the evaluation of a "black-box" decision support tool: a protocol and feasibility study for the evaluation of watson for oncology. *Journal of Clinical and Translational Research*, 3(Suppl 3):411–423, July 2018.

[14] International Agency for Research on Cancer. Lung cancer, 2024. Fact page, accessed 25 Jun 2025.

[15] Raheel Chaudhry, Adekunle E. Omole, and Bruno Bordoni. Anatomy, thorax, lungs, 2024. Updated 20 Apr 2024; accessed 25 Jun 2025.

[16] Giuseppe Miserocchi. The impact of heterogeneity of the air–blood barrier on control of lung extravascular water and alveolar gas exchange. *Frontiers in Network Physiology*, 3:1142245, 2023.

[17] Josiah P. Brandt and Pujyitha Mandiga. Histology, alveolar cells. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023. Updated 28 Aug 2023; accessed 25 Jun 2025.

[18] Dimitra E. Zazara, Ioannis Belios, Jöran Lücke, Tao Zhang, and Anastasios D. Giannou. Tissue-resident immunity in the lung: a first-line defense at the environmental interface. *Seminars in Immunopathology*, 44(6):827–854, 2022.

[19] J. Gordon Betts, Kelly A. Young, James A. Wise, Eddie Johnson, Brandon Poe, Dean H. Kruse, Oksana Korol, Jody E. Johnson, Mark Womble, and Peter DeSaix. 22.2 the lungs. In *Anatomy and Physiology 2e*. OpenStax, Houston, TX, 2022. Module 6 “The Respiratory System,” accessed 25 Jun 2025.

[20] International Agency for Research on Cancer. Cancer today—globocan 2022: Estimated cancer incidence, mortality and prevalence worldwide. https://gco.iarc.fr/today, 2022. Accessed 17 Jun 2025.

[21] International Agency for Research on Cancer (IARC). Trachea, bronchus and lung cancer: Globocan 2022 fact sheet. https://gco.iarc.who.int/media/globocan/factsheets/cancers/15-trachea-bronchus-and-lung-fact-sheet.pdf, 2024. Accessed 17 Jun 2025.

[22] Richard Doll, Richard Peto, Jillian Boreham, and Isabelle Sutherland. Mortality in relation to smoking: 50 years’ observations on male british doctors. *BMJ*, 328(7455):1519, 2004.

[23] G. N. Hounsfield. Computerized transverse axial scanning (tomography): Part 1. description of system. *British Journal of Radiology*, 46(552):1016–1022, 1973.

[24] Diana E. Litmanovich, Phillip M. Boiselle, and Alexander A. Bankier. Ct of the chest: Principles and applications. *Journal of Thoracic Imaging*, 24(4):216–230, 2009.

[25] Denise R. Aberle, Arthur M. Adams, Christine D. Berg, and et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

[26] National Cancer Institute. National lung screening trial (nlst): Study results. https://www.cancer.gov/types/lung/research/nlst, 2011. Accessed 17 Jun 2025.

[27] Harry J. de Koning, Carlijn M. van der Aalst, Paul A. de Jong, and et al. Reduced lung-cancer mortality with volume ct screening in a randomized trial. *New England Journal of Medicine*, 382(6):503–513, 2020.

[28] Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports*, 12(1):17981, October 2022.

[29] Sangeetha S.K.B, Sandeep Kumar Mathivanan, P. Karthikeyan, Hariharan Rajadurai, Basu Dev Shivahare, Saurav Mallik, and Hong Qin. An enhanced multimodal fusion deep learning neural network for lung cancer classification. *Systems and Soft Computing*, 6:200068, December 2024.

[30] C Tong, BY Liang, Q Su, MB Yu, JX Hu, AK Bashir, and ZG Zheng. Pulmonary nodule classification based on heterogeneous features learning. 39(2):574–581.

[31] P Hao, BY Deng, CT Huang, J Xu, F Zhou, ZX Liu, W Zhou, and YK Xu. Predicting anaplastic lymphoma kinase rearrangement status in patients with non-small cell lung cancer using a machine learning algorithm that combines clinical features and CT images. 12.

[32] M Mahootiha, HA Qadir, J Bergsland, and I Balasingham. Multimodal deep learning for personalized renal cell carcinoma prognosis: Integrating CT imaging and clinical data. 244.

[33] Nuruzzaman Faruqui, Mohammad Abu Yousuf, Md Whaiduzzaman, A.K.M. Azad, Alistair Barros, and Mohammad Ali Moni. LungNet: A hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data. 139:104961.

[34] Jörg Schilcher, Alva Nilsson, Oliver Andlid, and Anders Eklund. Fusion of electronic health records and radiographic images for a multimodal deep learning prediction model of atypical femur fractures. 168:107704.

[35] Imran Ahmed, Abdellah Chehri, Gwanggil Jeon, and Francesco Piccialli. Automated pulmonary nodule classification and detection using deep learning architectures. 20(4):2445–2456.

[36] E.S. Neal Joshua, D. Bhattacharyya, M. Chakkravarthy, and Y.-C. Byun. 3d CNN with visual insights for early detection of lung cancer using gradient-weighted class activation. 2021.

[37] Parnian Afshar, Farnoosh Naderkhani, Anastasia Oikonomou, Moezedin Javad Rafiee, Arash Mohammadi, and Konstantinos N. Plataniotis. MIXCAPS: A capsule network-based mixture of experts for lung nodule malignancy prediction. 116:107942.

[38] Nan Luo, Xiaojing Zhong, Luxin Su, Zilin Cheng, Wenyi Ma, and Pingsheng Hao. Artificial intelligence-assisted dermatology diagnosis: From unimodal to multimodal. *Computers in Biology and Medicine*, 165:107413, October 2023.

[39] FB Seyrek and H Yigit. Diagnosis of lung cancer from computed tomography. 30(8):1089–1111.

[40] Feng Liu, Jie Tang, Jiechao Ma, Cheng Wang, Qing Ha, Yizhou Yu, and Zhen Zhou. The application of artificial intelligence to chest medical image analysis. 1(3):104–117.

[41] M Sohaib, MJ Hasan, and ZL Zheng. A multichannel analysis of imbalanced computed tomography data for lung cancer classification. 35(8).

[42] F Bianconi, I Palumbo, ML Fravolini, M Rondini, M Minestrini, G Pascoletti, S Nuvoli, A Spanu, M Scialpi, C Aristei, and B Palumbo. Form factors as potential imaging biomarkers to differentiate benign vs. malignant lung lesions on CT scans. 22(13).

[43] MA Khan, S Rubab, A Kashif, MI Sharif, N Muhammad, JH Shah, YD Zhang, and SC Satapathy. Lungs cancer classification from CT images: An integrated design of contrast based classical features fusion and selection. 129:77–85.

[44] Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. A review of deep learning-based information fusion techniques for multimodal medical image classification. 177:108635.

[45] BP Veasey, J Broadhead, M Dahle, A Seow, and AA Amini. Lung nodule malignancy prediction from longitudinal CT scans with siamese convolutional attention networks. 1:257–264.

[46] C Venkatesh, J Chinnababu, A Kiran, CH Nagaraju, and M Kumar. A hybrid model for lung cancer prediction using patch processing and deeplearning on CT images.

[47] Muhammad Sufyan, Zeeshan Shokat, and Usman Ali Ashfaq. Artificial intelligence in cancer diagnosis and therapy: Current status and future perspective. 165:107356.

[48] Juan Li, Pan Jiang, Qing An, Gai-Ge Wang, and Hua-Feng Kong. Medical image identification methods: A review. 169:107777.

[49] National Lung Screening Trial Research Team. Data from the national lung screening trial (nlst) (version 3) [data set]. https://doi.org/10.7937/TCIA.HMQ8-J677, 2013.

[50] National Cancer Institute. National lung screening trial (nlst). https://www.cancer.gov/types/lung/research/nlst, 2022. Accessed: 2025-06-03.

[51] National Cancer Institute. National lung screening trial (nlst) datasets. https://cdas.cancer.gov/datasets/nlst/, 2025. Accessed: 2025-06-03.

[52] National Cancer Institute. Access publicly available imaging data from the national lung screening trial. https://datascience.cancer.gov/news-events/news/access-publicly-available-imaging-data-national-lung-screening-trial, November 2021. Accessed: 2025-06-03.

[53] Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bowen Choi, Jason J. Reicher, Lily Peng, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6):954–961, June 2019.

[54] Ritu Lamba, John P. McGahan, Michael T. Corwin, Chien-Wei S. Li, Tammy Tran, J. Anthony Seibert, and John M. Boone. Ct hounsfield numbers of soft tissues on unenhanced abdominal ct scans: Variability between two different manufacturers' mdct scanners. *AJR. American Journal of Roentgenology*, 203(5):1013–1020, November 2014.

[55] Weihua Liu, Xiabi Liu, Huiyu Li, Mincan Li, Xinming Zhao, and Zheng Zhu. Integrating lung parenchyma segmentation and nodule detection with deep multi-task learning. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3073–3081, 2021.

[56] Hui Zhang, Yanjun Peng, and Yanfei Guo. Pulmonary nodules detection based on multi-scale attention networks. *Scientific Reports*, 12(1):1466, 2022.

[57] Joana Sousa, Tânia Pereira, Inês Neves, Francisco Silva, and Hélder P. Oliveira. The influence of a coherent annotation and synthetic addition of lung nodules for lung segmentation in ct scans. *Sensors*, 22(9):3443, 2022. Open Access.

[58] 3D Slicer Community. 3d slicer. https://www.slicer.org, 2025. Accessed: 2025-06-03.

[59] Shafiq Alam, Muhammad Sohaib Ayub, Sakshi Arora, and Muhammad Asad Khan. An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*, 9(1):100341, 2023. Open access under CC BY-NC-ND 4.0.

[60] scikit-learn developers. sklearn.preprocessing.standardscaler, 2024. Accessed: 2025-06-11.

[61] scikit-learn developers. sklearn.preprocessing.onehotencoder, 2024. Accessed: 2025-06-11.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[63] Vinod Chugani. A comprehensive guide to k-fold crossvalidation, June 2024.

[64] Moritz Roberts, Dennis Driggs, Matthew Thorpe, James Gilbey, Michael Yeung, Simon Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Conor McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

[65] scikit-learn developers. Stratifiedgroupkfold - scikit-learn 1.4.2 documentation, 2024. Accessed: 2025-06-13.

[66] Furkan Berk Seyrek and Halil Yiğit. Diagnosis of lung cancer from computed tomography scans with deep learning methods. *Journal of Universal Computer Science*, 30(8):1089–1111, 2024.

[67] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[68] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. $sklearn.model_selection.GridSearchCV.scikit-learn, 2011.Accessed: 2025-06-11.$

[69] Riqiang Gao, Yucheng Tang, Mirza S. Khan, Kaiwen Xu, Alexis B. Paulson, Shelbi Sullivan, Yuankai Huo, Stephen Deppen, Pierre P. Massion, Kim L. Sandler, and Bennett A. Landman. Cancer risk estimation combining lung screening ct with clinical data elements. *Radiology: Artificial Intelligence*, 3(6):e210032, 2021.

[70] Satya Prakash Maurya, Pushpendra Singh Sisodia, Rahul Mishra, and Devesh Pratap Singh. Performance of machine learning algorithms for lung cancer prediction: A comparative approach. *Scientific Reports*, 14(1):18562, 2024.

[71] Farzad Mirzakhani. Detection of Lung Cancer using Multilayer Perceptron Neural Network. *Index Copernicus Master List*, 2017. 32 samples, 57 variables, 3 classes; reported accuracy 90%.

[72] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[73] Arnaud A. A. Setio, Francesco Ciompi, Geert Litjens, and et al. Pulmonary nodule detection in CT images: False positive reduction using multi–view convolutional networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169, 2016.

[74] Fuyong Han, Ruixuan Chen, and Pingkun Yan. Multimodal deep learning in medical imaging: A review of the state of the art. *Medical Image Analysis*, 76:102305, 2022.

[75] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[76] Can Cui, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A Coburn, Keith T Wilson, Bennett A Landman, and Yuankai Huo. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering*, 5(2):022002, 2023.

Appendix A

Appendix

A.1 Columns Kept

```
columns = [
    "cancer", "split", "cen", "dataset_version", "elig", "ineligible", "pid",
    "rndgroup", "study", "age", "educat", "ethnic", "gender", "height", "marital",
    "race", "weight", "age_quit", "cigar", "cigsmok", "pipe", "pkyr", "smokeage",
    "smokeday", "smokelive", "smokework", "smokeyr", "sct_image_has", "sct_image_years",
    "contactstatus", "fup_days", "wdlost", "resasbe", "resbaki", "resbutc", "reschem",
    "rescoal", "rescott", "resfarm", "resfire", "resflou", "resfoun", "reshard",
    "respain", "ressand", "resweld", "wrkasbe", "wrkbaki", "wrkbutc", "wrkchem",
    "wrkcoal", "wrkcott", "wrkfarm", "wrkfire", "wrkflou", "wrkfoun", "wrkhard",
    "wrkpain", "wrksand", "wrkweld", "yrsasbe", "yrsbaki", "yrsbutc", "yrschem",
    "yrscoal", "yrscott", "yrsfarm", "yrsfire", "yrsflou", "yrsfoun", "yrshard",
    "yrspain", "yrssand", "yrsweld", "ageadas", "ageasbe", "agebron", "agechas",
    "agechro", "agecopd", "agediab", "ageemph", "agefibr", "agehear", "agehype",
    "agepneu", "agesarc", "agesili", "agestro", "agetube", "diagadas", "diagasbe",
```

```
    "diagbron", "diagchas", "diagchro", "diagcopd", "diagdiab", "diagemph",
    "diagfibr", "diaghear", "diaghype", "diagpneu", "diagsarc", "diagsili",
    "diagstro", "diagtube", "ageblad", "agebrea", "agecerv", "agecolo", "ageesop",
    "agekidn", "agelary", "agelung", "agenasa", "ageoral", "agepanc", "agephar",
    "agestom", "agethyr", "agetran", "cancblad", "cancbrea", "canccerv", "canccolo",
    "cancesop", "canckidn", "canclary", "canclung", "cancnasa", "cancoral", "cancpanc",
    "cancphar", "cancstom", "cancthyr", "canctran", "fambrother", "famchild",
    "famfather", "fammother", "famsister", "acrin_alc_curr", "acrin_alc_ever",
    "acrin_drink24hr", "acrin_drinknum_curr", "acrin_drinknum_form",
    "acrin_drinkyrs_curr", "acrin_drinkyrs_form", "acrin_lastdrink",
    "lss_alcohol_freq", "lss_alcohol_num", "anyscr_has_nodule"
]
```

A.2 Columns After Clinical Data Preprocessing

```
    remaining_columns = ['cancer', 'cen', 'dataset_version', 'elig', 'rndgroup', 'study', 'age', 'educat', 'ethnic', 'gender', 'height', 'marital', 'race', 'weight', 'age_quit', 'cigar', 'cigsmok', 'pipe', 'pkyr', 'smokeage', 'smokeday', 'smokelive', 'smokework', 'smokeyr', 'sct_image_has', 'sct_image_years', 'contactstatus', 'fup_days', 'wdlost', 'wrkasbe', 'wrkbaki', 'wrkbutc', 'wrkchem', 'wrkcoal', 'wrkcott', 'wrkfarm', 'wrkfire', 'wrkflou', 'wrkfoun', 'wrkhard', 'wrkpain', 'wrksand', 'wrkweld', 'diagadas', 'diagasbe', 'diagbron', 'diagchas', 'diagchro', 'diagcopd', 'diagdiab', 'diagemph', 'diagfibr', 'diaghear', 'diaghype', 'diagpneu', 'diagsarc', 'diagsili', 'diagstro', 'diagtube', 'cancblad', 'cancbrea', 'canccerv', 'canccolo', 'cancesop', 'canckidn', 'canclary', 'canclung', 'cancnasa', 'cancoral', 'cancpanc', 'cancphar', 'cancstom', 'cancthyr', 'canctran', 'fambrother', 'famchild', 'famfather', 'fammother', 'famsister', 'lss_alcohol_freq', 'lss_alcohol_num', 'anyscr_has_nodule']
```

A.3 Clinical Data Hyperparameter tuning

Table A.1: Outer–fold performance of every hyper-parameter configuration (mean ± SD.

Model	ParamKey	Log-loss	AUC	F1	Precision	Recall	Accuracy
LogisticRegression	C=0.01	0.446 ± 0.016	0.892 ± 0.004	0.794	0.863	0.735	0.838
LogisticRegression	C=0.1	0.446 ± 0.043	0.887 ± 0.007	0.794	0.808	0.780	0.829
LogisticRegression	C=1	0.591 ± 0.110	0.875 ± 0.011	0.782	0.773	0.792	0.814
RandomForestClassifier	max_depth=10; n_estimators=200	0.493 ± 0.042	0.870 ± 0.040	0.760	0.772	0.765	0.793
MLPClassifier	activation=relu; hidden_layer_sizes=(100, 100)	1.061 ± 0.181	0.870 ± 0.011	0.764	0.743	0.790	0.794
LogisticRegression	C=10	0.818 ± 0.276	0.870 ± 0.016	0.783	0.776	0.791	0.815
RandomForestClassifier	max_depth=None; n_estimators=200	0.497 ± 0.044	0.868 ± 0.040	0.764	0.761	0.785	0.793
RandomForestClassifier	max_depth=20; n_estimators=200	0.497 ± 0.044	0.868 ± 0.040	0.764	0.761	0.786	0.794
RandomForestClassifier	max_depth=10; n_estimators=100	0.492 ± 0.044	0.868 ± 0.042	0.749	0.768	0.751	0.784
XGBClassifier	learning_rate=0.1; max_depth=3; n_estimators=200	0.674 ± 0.403	0.867 ± 0.026	0.745	0.688	0.837	0.746
RandomForestClassifier	max_depth=None; n_estimators=100	0.496 ± 0.048	0.866 ± 0.046	0.757	0.755	0.775	0.788
RandomForestClassifier	max_depth=20; n_estimators=100	0.496 ± 0.048	0.866 ± 0.046	0.758	0.756	0.777	0.788
MLPClassifier	activation=tanh; hidden_layer_sizes=(100,)	0.868 ± 0.157	0.865 ± 0.014	0.758	0.739	0.781	0.790
MLPClassifier	activation=relu; hidden_layer_sizes=(100,)	0.856 ± 0.173	0.864 ± 0.010	0.757	0.732	0.785	0.787
XGBClassifier	learning_rate=0.1; max_depth=3; n_estimators=100	0.619 ± 0.319	0.864 ± 0.023	0.734	0.672	0.833	0.735
MLPClassifier	activation=tanh; hidden_layer_sizes=(100, 100)	1.222 ± 0.104	0.855 ± 0.016	0.749	0.715	0.788	0.777
XGBClassifier	learning_rate=0.1; max_depth=6; n_estimators=200	0.828 ± 0.682	0.853 ± 0.039	0.734	0.669	0.840	0.732
XGBClassifier	learning_rate=0.1; max_depth=6; n_estimators=100	0.785 ± 0.621	0.847 ± 0.042	0.728	0.663	0.835	0.725
XGBClassifier	learning_rate=0.01; max_depth=3; n_estimators=200	0.546 ± 0.056	0.843 ± 0.019	0.649	0.678	0.696	0.701
LGBMClassifier	learning_rate=0.1; n_estimators=200; num_leaves=64	1.281 ± 1.148	0.842 ± 0.075	0.730	0.698	0.799	0.737
LGBMClassifier	learning_rate=0.1; n_estimators=200; num_leaves=31	1.281 ± 1.148	0.842 ± 0.075	0.730	0.698	0.799	0.737
LGBMClassifier	learning_rate=0.1; n_estimators=100; num_leaves=64	0.902 ± 0.797	0.828 ± 0.099	0.736	0.686	0.822	0.737
LGBMClassifier	learning_rate=0.1; n_estimators=100; num_leaves=31	0.902 ± 0.797	0.828 ± 0.099	0.736	0.686	0.822	0.737
LGBMClassifier	learning_rate=0.01; n_estimators=200; num_leaves=64	0.647 ± 0.286	0.824 ± 0.072	0.659	0.705	0.694	0.706

Continued on next page

Table A.1 continued from previous page

Model	ParamKey	Log-loss	AUC	F1	Precision	Recall	Accuracy
LGBMClassifier	learning_rate=0.01; n_estimators=200; num_leaves=31	0.647 ± 0.286	0.824 ± 0.072	0.659	0.705	0.694	0.706
XGBClassifier	learning_rate=0.01; max_depth=3; n_estimators=100	0.586 ± 0.069	0.818 ± 0.025	0.627	0.665	0.679	0.676
XGBClassifier	learning_rate=0.01; max_depth=6; n_estimators=200	0.667 ± 0.264	0.797 ± 0.076	0.635	0.646	0.721	0.666
LGBMClassifier	learning_rate=0.01; n_estimators=100; num_leaves=64	0.634 ± 0.182	0.762 ± 0.121	0.643	0.701	0.666	0.696
LGBMClassifier	learning_rate=0.01; n_estimators=100; num_leaves=31	0.634 ± 0.182	0.762 ± 0.121	0.643	0.701	0.666	0.696
XGBClassifier	learning_rate=0.01; max_depth=6; n_estimators=100	0.638 ± 0.125	0.759 ± 0.059	0.628	0.671	0.669	0.678

Note: The final clinical baseline was selected based on cross-validation performance across hyper-parameter configurations. Although a linear Logistic Regression model with strong ℓ_2 regularization achieved the highest average AUC, the observed improvement (ΔAUC $=$ 0.013) was not statistically significant. Moreover, this configuration exhibited poorer and more variable cross-entropy performance. In contrast, the Random Forest model delivered comparable discriminative ability while producing more reliable and better-calibrated probability estimates. Consequently, the Random Forest was retained as the clinical reference model in all subsequent experiments.