FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Information Fusion-Based Model for Lung Nodule Characterization

João António Maricato Malva

Mestrado em Engenharia Informática e Computação

Supervisor: Hélder Oliveira

Co-Supervisor: Tânia Pereira

Co-Supervisor: Eduardo Rodrigues

July 9, 2025

Information Fusion-Based Model for Lung Nodule Characterization

João António Maricato Malva

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Rui Rodrigues
Referee: Prof. Joel Arrais
Referee: Prof. Tânia Pereira

July 9, 2025

Resumo

O cancro do pulmão continua a ser uma das principais causas de mortalidade relacionada com o cancro em todo o mundo, principalmente devido aos desafios do diagnóstico tardio e às complexidades envolvidas na diferenciação entre nódulos pulmonares benignos e malignos. Esta tese introduz uma nova estrutura para a caraterização de nódulos pulmonares que integra técnicas de fusão de informação, combinando representações de aprendizagem profunda com caraterísticas radiológicas artesanais. O objetivo é melhorar o desempenho da classificação, a interpretabilidade do modelo e a fiabilidade clínica.

A metodologia proposta utiliza uma estratégia de fusão que combina caraterísticas profundas extraídas de CNNs com caraterísticas superficiais, como descritores de textura, forma e intensidade. Estas caraterísticas são integradas em vários níveis da arquitetura da rede para avaliar os efeitos de várias fases de fusão no desempenho. O sistema é rigorosamente avaliado utilizando o conjunto de dados LIDC-IDRI através de experimentação extensiva, incluindo estudos de ablação, seleção de caraterísticas e análises de explicabilidade do modelo utilizando Grad-CAM e SHAP.

Os resultados indicam que a integração de caraterísticas artesanais e profundas conduz a um melhor desempenho de classificação em diferentes arquitecturas, com melhorias consistentes nas métricas AUC. Além disso, as análises de explicabilidade destacam que as caraterísticas artesanais - especialmente os descritores de forma - desempenham um papel significativo na previsão de malignidade. Em geral, os resultados sugerem que esta abordagem baseada na fusão não só aumenta a precisão da previsão, como também melhora a transparência do modelo, proporcionando uma ferramenta mais fiável e interpretável para o diagnóstico do cancro do pulmão assistido por computador.

Esta investigação alinha-se com os esforços globais para promover a deteção precoce e garantir o acesso equitativo aos cuidados de saúde, contribuindo para o desenvolvimento de ferramentas de diagnóstico alimentadas por IA que ajudam na tomada de decisões clínicas informadas e promovem os objectivos de cuidados de saúde sustentáveis.

Abstract

Lung cancer continues to be one of the leading causes of cancer-related mortality worldwide, primarily due to the challenges of late diagnosis and the complexities involved in differentiating between benign and malignant pulmonary nodules. This thesis introduces a novel framework for lung nodule characterisation that integrates information fusion techniques, combining deep learning representations with handcrafted radiomic features. The goal is to enhance classification performance, model interpretability, and clinical reliability.

The proposed methodology employs a fusion strategy that blends deep features extracted from CNNs with shallow features, such as texture, shape, and intensity descriptors. These features are integrated at multiple levels of the network architecture to evaluate the effects of various fusion stages on performance. The system is rigorously assessed using the LIDC-IDRI dataset through extensive experimentation, including ablation studies, feature selection, and model explainability analyses utilising Grad-CAM and SHAP.

The results indicate that the integration of handcrafted and deep features leads to improved classification performance across different architectures, with consistent enhancements in AUC metrics. Additionally, the explainability analyses highlight that handcrafted features - especially shape descriptors - play a significant role in predicting malignancy. Overall, the findings suggest that this fusion-based approach not only boosts predictive accuracy but also enhances model transparency, providing a more reliable and interpretable tool for computer-aided lung cancer diagnosis.

This research aligns with global efforts to promote early detection and ensure equitable access to healthcare, contributing to the development of AI-powered diagnostic tools that aid in informed clinical decision-making and advance the objectives of sustainable healthcare.

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. There are 17 goals addressing the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice [59].

This research contributes to SDG 3 by developing and validating information fusion-based models that combine deep and shallow features for lung cancer characterization. The aim of studying these models is to improve early diagnosis, decrease mortality, and support more informed clinical decision-making. This work aligns with health equity by promoting access to reliable diagnostic tools, supporting global well-being commitments, and reducing health disparities.

The specific Sustainable Development Goals referenced in this work include:

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

SDG	Target	Contribution	Performance Indicators and Metrics
3	3.4	Development of models for early and accurate lung cancer detection, supporting the reduction of mortality from non-communicable diseases.	Improvement in model performance defined by machine learning metrics; Decrease in misdiagnosis rates.
	3.8	Enhancing the reliability and accessibility of Computer-Aided Diagnosis (CAD) systems through robust model design, thereby facilitating access to quality essential healthcare services.	Increase in diagnostic consistency; Reduction in healthcare disparities.
	3.b	Promoting innovation in medical diagnostics using Artificial Intelligence (AI) and information fusion, contributing to the development of advanced technologies for non-communicable disease management.	Technology adoption in clinical practice.
	3.d	Supporting the deployment of diagnostic tools in underserved areas, contributing to strengthened national and global health risk management.	The number of low-resource environments adopting the technology.

Acknowledgements

I want to express my heartfelt gratitude to everyone who contributed, both directly and indirectly, to the completion of this work.

First and foremost, I am profoundly thankful to the professors who have guided me throughout my academic journey. They have played a crucial role in my intellectual and personal development. I extend my deep appreciation to the INESC TEC team, and particularly to Eduardo Rodrigues, for their invaluable guidance, collaboration, and unwavering support throughout the course of this research.

Additionally, I acknowledge the presence and contributions of all the students who have passed through this faculty. Their shared experiences fostered a stimulating academic and social environment. A special note of recognition goes to the Academic Traditions, which I wholeheartedly embraced. These traditions significantly enriched my university experience and deepened my sense of belonging to this academic community.

I am immensely grateful to my girlfriend for her steadfast support, patience, and encouragement during the most challenging moments of my academic journey. To my friends who stood by me over the past five years, providing companionship, motivation, and balance: thank you. I leave with my heart full of our long conversations, unforgettable memories, and the strength drawn from both the joyful and difficult times we shared together. Finally, I express my deep appreciation to my family, whose unwavering support has always been my foundation. I am especially grateful to my father, whose words have consistently reminded me of the value of resilience: “A vida é dura para quem é mole”

The fight against cancer, which serves as the primary motivation for this research, arises from many years of personal struggle and progress in my health journey. I am profoundly grateful to have reached a point where I can contribute, albeit in a modest way, to alleviating the hardship faced by families impacted by this disease.

This work would not have been achievable without the unwavering support of numerous individuals. Success is never a solitary endeavour, and this dissertation stands as a testament to that truth.

As the Orfeão Universitário do Porto sings: "Quero ficar sempre estudante"
With love,
Engineering!

João Malva

Institutional Acknowledgements

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.

João Malva

"On ne voit bien qu'avec le coeur. L'essentiel est invisible pour les yeux."

Antoine de Saint-Exupéry

Contents

Acknowledgements iv

1 Introduction 1
- 1.1 Context . . . 1
- 1.2 Problem . . . 2
- 1.3 Hypothesis . . . 2
- 1.4 Motivation . . . 2
- 1.5 Research Questions . . . 2

2 Theoretical Background 4
- 2.1 Medical and Clinical Context . . . 4
  - 2.1.1 Lung Cancer . . . 4
  - 2.1.2 Lung Nodule . . . 5
- 2.2 Medical Imaging and CT Scans . . . 5
  - 2.2.1 Computed Tomography . . . 5
  - 2.2.2 Screening Challenges . . . 6
- 2.3 Handcrafted Features in Radiomics . . . 6
  - 2.3.1 First-Order Features . . . 7
  - 2.3.2 Texture-based Features . . . 10
  - 2.3.3 2D Shape-based Features . . . 14
  - 2.3.4 Transform-based Features . . . 17
- 2.4 Machine Learning . . . 18
  - 2.4.1 Supervised Learning for Classification . . . 18
  - 2.4.2 Model Evaluation Metrics . . . 19
  - 2.4.3 Convolutional Neural Networks . . . 20
- 2.5 Model Architectures . . . 21
  - 2.5.1 Linear SVM . . . 21
  - 2.5.2 ResNet . . . 21
  - 2.5.3 EfficientNet . . . 22
  - 2.5.4 ConvNeXt . . . 23
- 2.6 Fusion Strategies in Machine Learning . . . 23
  - 2.6.1 Early Fusion . . . 23
  - 2.6.2 Middle Fusion . . . 23
  - 2.6.3 Late Fusion . . . 24
- 2.7 Explainable Artificial Intelligence . . . 24
  - 2.7.1 SHapley Additive exPlanations . . . 25
  - 2.7.2 Class Activation Mapping and Variations . . . 25

3 State of the Art **26**
3.1 Fusion Techniques . . . 26
3.1.1 Decision-Level Fusion . . . 26
3.1.2 Feature Level Fusion . . . 29
3.1.3 Decision And Feature Fusion . . . 31
3.2 Remarks . . . 32

4 Datasets **33**
4.1 LIDC-IDRI . . . 33
4.2 LUNA16 . . . 34
4.3 NLST . . . 34
4.4 ANODE09 . . . 34
4.5 LUNGx . . . 34
4.6 Limitations . . . 35

5 Methodology **36**
5.1 Dataset . . . 36
5.1.1 Data Annotations . . . 36
5.1.2 Data Preprocessing . . . 36
5.1.3 Data Labelling . . . 37
5.2 Handcrafted Feature Extraction . . . 37
5.3 Evaluation . . . 40
5.4 Experimental Procedures . . . 40
5.4.1 Hyperparameter Optimization . . . 40
5.4.2 Baseline Selection . . . 41
5.4.3 Single Fusion Evaluation . . . 41
5.4.4 Feature Selection . . . 42
5.4.5 Multi-Feature Fusion Combinations . . . 42
5.4.6 Dataset Ablation . . . 42
5.4.7 Explainability . . . 43

6 Results **44**
6.1 Baseline Model Performance . . . 44
6.2 Impact of Individual Radiomics Feature Fusion . . . 45
6.3 Radiomics Feature Selection . . . 47
6.3.1 Predictive Power . . . 47
6.3.2 Settings and Stages Stability . . . 47
6.3.3 Selection . . . 49
6.4 Multi-Feature Fusion . . . 49
6.5 Dataset Ablation . . . 51
6.6 Model Explainability . . . 52
6.6.1 Grad-CAM . . . 52
6.6.2 SHAP . . . 52

7 Conclusions **54**
7.1 Research Questions . . . 54
7.2 Hypothesis . . . 56
7.3 Main Contributions . . . 56
7.4 Future Work . . . 57

References **59**

8 Best AUC and Stage for Fusion **66**

List of Figures

6.1 Grad-CAM Heatmap Comparison 52
6.2 Local SHAP Analyses 53

List of Tables

2.1 Common Machine Learning Evaluation Metrics with Descriptions and Equations 19

5.1 Summary of Extracted First-Order Features. 38
5.2 Summary of Experimental Configuration. 41
5.3 Summary of LIDC-IDRI Dataset Variants Used in Experiments. 43

6.1 Comparison of Baseline Architectures on the Base Dataset. 45
6.2 Performance of Single Feature Fusion with Different ResNet Backbone Stages. . 46
6.3 Performance of Individual Radiomics Feature Vectors with Linear SVM 47
6.4 Best AUC and Stage for FOF Fusion 48
6.5 Best AUC and Stage for LBP Fusion 48
6.6 Comparative Performance of ResNet-18 Backbone with Fused Features. 49
6.7 Comparative Performance of EfficientNet-B0 Backbone with Fused Features. . . 50
6.8 Comparative Performance of ConvNeXt-Tiny Backbone with Fused Feature Combinations. 50
6.9 Comparison of ResNet-18 vs. Fused ResNet-18 Across Subsets of Data. 51

8.1 Best AUC and Stage for HOG Fusion 66
8.2 Best AUC and Stage for Gabor Fusion 66
8.3 Best AUC and Stage for Shape Fusion 67
8.4 Best AUC and Stage for Haralick Fusion 67

Abbreviations

ACRIN	American College of Radiology Imaging Network	34
AGV	Absolute Gradient Value	17
AGM	Absolute Gradient Matrix	17
AI	Artificial Intelligence	iii
ALN	Axillary Lymph Node	28
ANODE09	Automatic Nodule Detection 2009	34
ASM	Angular Second Momentum	11
AUC	Area Under the Curve	26
AUC-ROC	Area Under the ROC	40
AVG-Predict	Average Prediction Score	27
BB	Bounding Box	38
BPNN	Back Propagation Neural Network	26
CAD	Computer-Aided Diagnosis	iii
CAM	Class Activation Mapping	25
CBAM	Convolutional Block Attention Module	28
CCA	Canonical Correlation Analysis	30
CNN	Convolutional Neural Network	20
CT	Computed Tomography	1
DCE-MRI	Dynamic Contrast-Enhanced Magnetic Resonance Imaging (MRI)	28
DCNN	Deep Convolutional Neural Network	26
DL	Deep Learning	1
DNN	Deep Neural Network	2
EL	Energy Layer	27
FDA	Food and Drug Administration	33
FLOPS	Floating Point Operations per Second	22
FN	False Negative	19
FNIH	Foundation for the National Institutes of Health	33
FOF	First-Order Features	7
FP	False Positive	19
FC	Fully Connected	21

GCN	Graph Convolutional Network	31
GELU	Gaussian Error Linear Unit	23
GLCM	Gray-Level Co-Occurrence Matrix	10
GLPP	Global-Local Pyramid Pattern	30
GPU	Graphics Processing Unit	40
GTSDM	Gray-Tone Spatial Dependence Matrix	11
HOG	Histogram of Oriented Gradients	17
HU	Hounsfield Units	6
IDM	Inverse Difference Momentum	12
IMC	Information Measures of Correlation	14
IQR	Interquartile Range	8
KNN	K-Nearest Neighbor	30
LBP	Local Binary Patterns	17
LIDC-IDRI	Lung Image Database Consortium Image Collection	26
LNVA	Lung Nodule Visual Attribute	36
LNM	Lung Nodule Malignancy	36
LSS	Lung Screening Study group	34
LSTM	Long Short-Term Memory	30
LTP	Local Trinary Pattern	28
LUNA16	Lung Nodule Analysis 2016	30
MAD	Mean Absolute Deviation	9
MAX-VOTE	Majority Voting	27
MCC	Maximal Correlation Coefficient	14
ML	Machine Learning	18
MRI	Magnetic Resonance Imaging	xii
NCI	National Cancer Institute	33
NELSON	Nederlands–Leuvens Longkanker Screenings Onderzoek	34
NLST	National Lung Screening Trial	30
PCA	Principal Component Analysis	29
PET	Positron Emission Tomography	4
ReLU	Rectified Linear Unit	20
rMAD	Robust Mean Absolute Deviation	9
RMS	Root Mean Squared	9
ROI	Region of Interest	6
SDG	Sustainable Development Goal	iii
SHAP	SHapley Additive exPlanations	25
SVM	Support Vector Machine	21

SVM-FFCAT SVM - Feature Fusion by Concatenation 27
TCIA The Cancer Imaging Archive 28
T-LSTM Time-Modulated Long Short-Term Memory 30
TNM Tumor-Nodules-Metastasis 4
TN True Negative 19
TP True Positive 19
XAI Explainable Artificial Intelligence 24

Chapter 1

Introduction

1.1 Context

Lung cancer is the leading cause of cancer-related deaths and is often diagnosed at an advanced stage, contributing to a low 5-year survival rate of less than 10%, which occurs in 70% of cases. However, if detected early, the survival rate could exceed 90% [46]. In 2022, lung cancer had the highest incidence and mortality rates of all cancers worldwide [21]. In particular, in upper-middle-income countries, there has been a significant increase in lung cancer-related deaths, with a rise of 442,000 deaths, more than 2.5 times the increase in deaths of the combined three other income groups [62].

Efforts to reduce lung cancer mortality by screening have been hampered by the aggressive and diverse nature of the disease [45]. For example, low-dose Computed Tomography (CT) screening helps diagnose lung cancer more precisely and produces a reduction of 20% in mortality. Today, the classification of a pulmonary nodule depends on measuring its growth rate from multiple CT scans and following it for approximately two years to avoid performing a biopsy, which entails risks for patients and additional costs for healthcare entities. However, one significant drawback of conducting slice-by-slice CT scans in lung cancer detection is that they are challenging for doctors since the process of obtaining this data is time-consuming, expensive, prone to reader bias, and requires a high degree of competence and concentration [55].

As medical data becomes more complex, there is a growing need for models that can effectively integrate and analyse this data to support clinical decision-making [22]. CAD is increasingly being investigated as an alternative and complementary approach to conventional reading, as it avoids many of these issues. Automated nodule diagnostic systems can save both time and money while avoiding the risks of invasive surgical procedures. The noninvasive CAD system for lung nodule diagnosis is promising and has achieved high accuracy from a single CT scan [55].

The combined gains in medical imaging and Deep Learning (DL) complement new approaches that are accurate and allow safer disease recognition. DL models can overcome projections that show how medical images are analysed to locate and determine the type of lung abnormalities that commonly cause cancer.

1.2 Problem

This thesis addresses the need for a more accurate and reliable diagnostic tool. Existing diagnostic systems, primarily based on DL models, have certain limitations regarding the accuracy and generalisation of medical image datasets. These models are often based on deep features from neural networks, which can overshadow superficial features such as texture and shape that are important for accurately classifying nodules.

In addition, the lack of explainability of the model poses a challenge in clinical contexts, which can limit its reliability in making critical medical decisions. The inability to provide interpretable information hinders the adoption of these methods for diagnosing lung cancer.

1.3 Hypothesis

Feature extraction is critical for the characterisation task. Although Deep Neural Networks (DNNs), are widely used to extract deep features, shallow features - such as texture and shape - can also be derived using traditional extractors. State-of-the-art works show that combining shallow and deep features can improve the effectiveness of DL models in lung nodule characterization [64]. These advancements highlight the need for further research in DL and information fusion to prompt early detection, reduce mortality, and provide more effective treatment strategies for lung cancer patients.

We hypothesise that information fusion-based model approaches, with shallow and deep features, will result in a more accurate and reliable model for lung cancer characterisation, making it better suited for early detection and precise diagnosis. We seek to overcome the current state-of-the-art limitations in automatic lung cancer diagnostics, offering a solution that not only improves prediction accuracy but also has the potential to assist in clinical decision-making and medical practice.

1.4 Motivation

Promoting the improvement of human life and health through the early detection of diseases continues to be a concern worldwide. Lung cancer is an enemy of public health care, and the development of early and accurate diagnostic tools will help improve survival rates. This research aims to contribute to the goal of promoting health by harnessing modern technologies to address one of the most significant diagnostic issues in oncology today. By developing models that support more precise care, this dissertation aligns with the global imperative to promote well-being for all.

1.5 Research Questions

We will break it down into three research questions to bring clarity and precision to the hypothesis. These questions will guide the investigation, helping us understand our hypothesis' main points.

1. **Does fusing information from shallow and deep feature extractors improve classification or generalisation performance when compared to using a deep approach only?**

2. **How does the fusion approach behave under varying dataset conditions, such as different sample sizes, bounding-box definitions, and image representations?**

3. **In what ways does information fusion contribute to the explainability of lung nodule malignancy predictions?**

Chapter 2

Theoretical Background

2.1 Medical and Clinical Context

2.1.1 Lung Cancer

Most of the lung cancer cases diagnosed at a symptomatic stage are related to the primary or metastatic disease or some paraneoplastic syndrome. The process of patient evaluation involves physical examination, CT scans of the thorax and abdomen, pulmonary function tests, laboratory tests, monitoring patient weight loss and retrieving tumour tissue to perform histologic diagnosis, in order to determine the stage of the disease based on the international Tumor-Nodules-Metastasis (TNM) staging system [39, 61].

Four major histologic lung cancer types comprise the majority of them, including small cell lung cancer and three non-small cell lung cancer types [58]. Small cell lung cancer and squamous cell carcinoma arise mainly in the central airways, while adenocarcinomas are located more peripherally. The large-cell carcinoma is less differentiated from the other non-small cell lung cancer types and arises from metaplastic changes resulting from smoking.

Patients with small-cell lung cancer confined to the chest undergo treatment that includes thoracic radio and chemotherapy. In contrast, patients with more advanced stages of the disease receive chemotherapy along with various other drug combinations. Treatments normally result in tumour shrinkage, symptom relief, and an increase in median survival [39].

On the other hand, patients with non-small cell lung cancer whose disease is confined to the chest undergo surgical evaluation of the mediastinum to check for lymph node involvement. More recently, they have also been subjected to Positron Emission Tomography (PET) scans, a functional imaging technique that detects radiotracer uptake to assess metabolic activity, providing complementary information [8]. In the cases where there is no evidence of mediastinal lymph node involvement or distant metastatic disease, patients have surgical resection of the primary tumour. Those with mediastinal lymph node involvement are submitted either to (1) preoperative chemotherapy followed by an attempt at surgical resection, as in the previous case, or (2) a combination of chemotherapy and chest radiotherapy administered as part of a curative treatment

strategy. These therapies may not cure the condition, but they can relieve symptoms and extend life by 2 to 10 months [39].

2.1.2 Lung Nodule

According to Baum et al. [7] and Loverdos et al. [34], a lung nodule can be defined "as a more or less round, well-demarcated lesion measuring up to 3 cm in diameter".

Lung nodules in CT imaging are classified into three distinct categories based on their attenuation: solid nodules, the most common form, characterised by a consistent homogeneous soft-tissue attenuation; ground-glass nodules, which present a non-uniform appearance with a hazy increase in local attenuation of the lung parenchyma, while still clearly displaying the underlying bronchial and vascular structures; and part-solid nodules, which include both solid and ground-glass attenuation components [34].

Incidental detection of pulmonary nodules is prevalent, occurring in up to 75% of CT scans performed for other clinical reasons, with more than half of patients presenting with two or more nodules. In this context, morphological assessment by CT plays a crucial role in stratifying the risk of malignancy. Specific benign lesions show characteristic morphological patterns that allow for non-invasive diagnosis. For example, granulomas often show central, laminar, or diffuse calcifications, while the presence of "popcorn" calcifications is pathognomonic of hamartomas. In addition, well-defined perifissural nodules are generally benign lymph nodes with an extremely low risk of malignancy [7].

The increasing occurrence of incidental findings in asymptomatic individuals undergoing imaging poses notable challenges in clinical practice. One major issue is the overestimation of minor nodule volumes due to the partial volume effect, which can result in inaccurate assessments and potential misclassification [26]. Additionally, distinguishing between benign and malignant nodules is complex, potentially resulting in delayed diagnoses or unnecessary invasive procedures. This diagnostic uncertainty is worsened by the absence of standardised and reliable management algorithms, which are essential for the timely detection and treatment of malignant lesions [34].

By effectively addressing these challenges and accurately interpreting imaging characteristics to assess the risk of malignancy, healthcare providers can improve the precision of screening practices. This strategy significantly decreases the need for unnecessary invasive diagnostic procedures, ultimately leading to safer, more efficient, and cost-effective clinical follow-up for patients.

2.2 Medical Imaging and CT Scans

2.2.1 Computed Tomography

Computed Tomography (CT) is an imaging modality that uses X-ray technology to generate detailed cross-sectional images of the body by measuring the attenuation of X-ray beams as they pass through different tissues. The attenuation caused by the absorption and scattering of some photons prevents them from reaching the detector, resulting in a record of this variation at different angles

during a complete rotation. The projections are then reconstructed in a 3D volume made up of volumetric pixels (voxels). Each pixel is assigned a CT number, measured in Hounsfield Units (HU), which reflects the local attenuation of the X-rays corresponding to the density of the tissue (air has -1000 HU, water has 0 HU, and values above 400 HU indicate the density of hard tissues) [50]. In spatial resolution, each voxel shares the respective dimension with the pixels present in the X and Y axes of the image, while the slice thickness determines the Z dimension [8, 9, 38].

The process of generating a CT scan involves several key steps: scanning of an object with a specific configuration and parameters (such as magnification, orientation, X-ray energy); reconstruction of 2D projections in a 3D vortex matrix; determining a threshold value for accurate segmentation; generating surface or volume data; and conducting dimensional measurements or geometric analysis [9].

2.2.2 Screening Challenges

When tasked with evaluating the malignancy of lung nodules, the first step is to segment the lung parenchyma[1], followed by the segmentation of the nodules. Large solid nodules ($> 10mm$) present a unique challenge for identification, as they have a different intensity range compared to smaller lesions. Predicting lung cancer at an early stage using CT images poses significant challenges for radiologists. This process requires a significant investment of time and resources, and it is susceptible to errors. Screening requires a high level of concentration and expertise due to various factors, including low contrast variation, heterogeneity, and the visual similarities between benign and malignant nodules [23].

Accurate detection of lung nodules is a difficult task in the field of medical imaging. The nodules are often extremely unbalanced, exhibiting high intra-class variance, and the complex structure of the lungs further complicates this issue.

2.3 Handcrafted Features in Radiomics

Radiomics is a topic of much discussion in the fields of nuclear medicine and medical imaging. It aims to extract quantitative and reproducible insights from diagnostic images by capturing complex patterns that are often difficult for the human eye to recognise or measure [9].

These features can be classified into the following categories: statistical, which includes histogram-based and texture-based; model-based; transform-based; and shape-based, derived from the type of information that is extracted and the way it is extracted [1, 9]. To enhance readability, we use the term Region of Interest (ROI) to refer to both 2D areas and 3D volumes.

In the following sections, we will provide an overview of the features used in this thesis, as well as others that help to understand them.

[1] The functional tissue of an organ as distinguished from the connective and supporting tissue.

2.3.1 First-Order Features

First-Order Features (FOF), often known as histogram-based features or intensity features, describe the distribution of voxel intensities within a region of interest and are derived from the histogram of an image. These features are computed using the grey-level histogram and provide the simplest information extracted from the image. They specifically reflect individual voxel intensities without taking into account any relationships with neighbouring voxels [1].

We will assume the following foundational premises [60]:

- $\mathbf{X}$ is the set of N_p voxels included in the ROI.
- $\mathbf{P}(i)$ is the first-order histogram with N_g discrete intensity levels, where N_g is the number of non-zero bins, equally spaced from 0 with a width defined by the `binWidth` parameter.
- $p(i)$ is the normalized first-order histogram, defined as $p(i) = \frac{\mathbf{P}(i)}{N_p}$.

Energy

Energy serves as a quantitative measure of the magnitude of voxel values within an image. Higher voxel values indicate a greater cumulative total of the squares of these values, reflecting enhanced intensity characteristics within the image data.

$$energy = \sum_{i=1}^{N_p} (\mathbf{X}(i))^2 \tag{2.1}$$

Total Energy

Total Energy represents the quantitative measure of the Energy feature, adjusted by the volume of the voxel, expressed in cubic millimetres.

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i))^2 \tag{2.2}$$

Entropy

Entropy is a quantitative measure of the uncertainty and randomness inherent in image voxel values, capturing the average information content required for the encoding of those values.

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \big(p(i)\big) \tag{2.3}$$

Minimum

This feature refers to the smallest grey level intensity value within the ROI, representing the lowest voxel value detected in the analysed area.

$$minimum = \min(\mathbf{X}) \tag{2.4}$$

10th percentile

This value indicates that 10% of the grey level intensities within the ROI fall below this threshold. This metric offers insight into the lower end of the intensity distribution, assisting in characterising the presence of darker regions.

90th percentile

This value signifies that 90% of the grey level intensities in the ROI are below this mark. The 90th percentile underscores the upper range of intensity distribution, highlighting the presence of brighter regions.

Maximum

This denotes the largest grey level intensity value found within the ROI, reflecting the highest voxel value identified in the analysed area and representing the brightest point within that region.

$$maximum = \max(\mathbf{X}) \tag{2.5}$$

Mean

The mean refers to the average grey level intensity calculated from all voxel values within the given ROI.

$$mean = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbf{X}(i) \tag{2.6}$$

Median

The median refers to the middle value of grey-level intensities within the ROI when the values from the voxels are ordered.

Interquartile Range

The Interquartile Range (IQR) is defined as the difference between the 75th percentile ($\mathbf{P}_{75}$) and the 25th percentile ($\mathbf{P}_{25}$) of a data set.

$$IQR = \mathbf{P}_{75} - \mathbf{P}_{25} \tag{2.7}$$

Range

The range of a dataset is calculated as the difference between the maximum and minimum values of the observed data.

$$range = \max(\mathbf{X}) - \min(\mathbf{X}) \tag{2.8}$$

Mean Absolute Deviation

The Mean Absolute Deviation (MAD) quantifies the average distance of all intensity values from the overall mean value of the image array. This statistic provides insights into the degree of variation present within the intensity levels.

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}| \tag{2.9}$$

Robust Mean Absolute Deviation

The Robust Mean Absolute Deviation (rMAD) measures the average distance of intensity values from the mean calculated from a subset of the image array. This subset consists of grey levels that fall within the range of the 10th to the 90th percentiles.

$$rMAD = \frac{1}{N_{10-90}} \sum_{i=1}^{N_{10-90}} |\mathbf{X}_{10-90}(i) - \bar{X}_{10-90}| \tag{2.10}$$

Root Mean Squared

Root Mean Squared (RMS) is defined as the square root of the mean of the squared intensity values. This metric serves as an alternative measure of the overall magnitude of the image values.

$$RMS = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i))^2} \tag{2.11}$$

Skewness

Skewness quantifies the asymmetry of a distribution in relation to its mean. It indicates whether the distribution's tail extends more towards the higher or lower values, which results in positive or negative skewness depending on the concentration of the data.

$$skewness = \frac{\mu_3}{\sigma^3} = \frac{\frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^3}{\left(\sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2} \right)^3} \tag{2.12}$$

Having μ_3 as the 3rd central moment.

Kurtosis

Kurtosis is a statistical measure that describes the 'peakedness' of the distribution of values within the image ROI. A higher kurtosis indicates that the distribution's mass is concentrated towards the tails rather than the mean. Conversely, a lower kurtosis suggests that the mass is more focused

around a spike near the mean value.

$$kurtosis = \frac{\mu_4}{\sigma^4} = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^4}{\left(\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2\right)^2} \tag{2.13}$$

Having μ_4 as the 4^{th} central moment.

Variance

Variance is defined as the mean of the squared distances of each intensity value from the overall mean value. By definition, variance is represented as σ^2.

$$variance = \frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2 \tag{2.14}$$

Uniformity

Uniformity quantifies the sum of the squares of intensity values, serving as an indicator of an image array's homogeneity. Higher uniformity signifies greater homogeneity and a reduced range of discrete intensity values.

$$uniformity = \sum_{i=1}^{N_g} p(i)^2 \tag{2.15}$$

2.3.2 Texture-based Features

In the words of Kaur, Nazir, and Manik [24], "A textural feature is nothing more than a recurrent arrangement of information in a visual representation".

Textures provide valuable information about an object's material properties, spatial organisation, and surrounding environment. Since the textural properties of images contain helpful information for distinguishing between different objects, it is essential to incorporate texture features into the analysis [18].

By extracting these texture features, we can improve the ability of algorithms to differentiate between various classes of images and identify specific patterns within them.

2.3.2.1 Gray-Level Co-Occurrence Matrix

The Gray-Level Co-Occurrence Matrix (GLCM) is a statistical method used to analyse textures by considering the spatial relationships between pixels. This technique characterises an image's texture by quantifying how frequently pairs of pixels with specific values occur in a defined spatial arrangement [69].

In the generation of a GLCM, each element located at position (i, j) represents the aggregated count of occurrences where a pixel with the intensity value i is found in a specific spatial relationship to a pixel with the intensity value j within the given input image. This matrix captures

the frequency of pixel value pairs at a defined distance and orientation, thereby providing insights into the spatial distribution of pixel intensities. Essentially, for each pair of pixel values (i, j), the GLCM tallies how many times pixels of value i are adjacent to pixels of value j according to the chosen configuration, such as horizontal, vertical, or diagonal arrangements [69].

2.3.2.2 Haralick Features

Based on the GLCM, also referred to in earlier literature as the Gray-Tone Spatial Dependence Matrix (GTSDM), Haralick, Shanmugam, and Dinstein [18] proposed a comprehensive suite of 28 textural features that could be extracted from each image matrix. They collected four values for each of the 14 measures and then calculated the average and range for each measure based on these four values [18, 69]. This resulted in a total of 28 features that could be used as inputs for a classifier.

Let us first establish the needed notation for the calculation of the 14 features:

- $p(i,j)$: Normalized value of the GLCM at row i and column j.
- $p_x(i)$: i^{th} entry in the marginal-probability matrix obtained by summing the rows of $p(i,j)$, $p_x(i) = \sum_{j=1}^{N_g} p(i,j)$.
- N_g: Number of distinct grey levels in the quantised image (size of the GLCM)
- $\sum_i$ and $\sum_j$, shorthand for $\sum_{i=1}^{N_g}$ and $\sum_{j=1}^{N_g}$, respectively.
- $p_y(j) = \sum_{i=1}^{N_g} p(i,j)$
- $p_{x+y}(k) = \sum_{\substack{i=1 \\ j=1 \\ i+j=k}}^{N_g} p(i,j), \quad k = 2,3,\ldots,2N_g$
- $p_{x-y}(k) = \sum_{\substack{i=1 \\ j=1 \\ |i-j|=k}}^{N_g} p(i,j), \quad k = 0,1,\ldots,N_g-1$

The presented equations define the proposed features:

1. **Angular Second Moment:**
   Angular Second Momentum (ASM) quantifies the homogeneity of an image and is defined as [18]:
   $$\text{ASM} = \sum_i \sum_j p(i,j)^2 \tag{2.16}$$
   When the pixels present in a given image are very similar, the ASM value will be significant. In a quantisation scheme with (N_g) grey levels, a uniform image will have only one entry in the GLCM, resulting in a maximum ASM value of 1. In contrast, given an image that is filled completely randomly, all entries of the $(N_g \times N_g)$ GLCM matrix will be equally represented, with the probability of each entry defined by $p(i,j) = 1/N_g^2$. This results in the minimum ASM value being $1/N_g^2$ [47].

High energy values are indicative of a constant or periodic grey-level distribution. Energy is defined within a normalised range, and the GLCM of an image with lower homogeneity will display numerous small entries [47].

2. **Contrast:**
   Contrast measures the local variations present in the image and is defined as follows [18]:

$$\text{Contrast} = \sum_{n=0}^{N_g-1} n^2 \left[\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \quad \text{where} \quad |i-j| = n \right] \tag{2.17}$$

3. **Correlation:**
   Correlation explains the relationship between a reference pixel and its neighbour: 0 signifies no correlation, while 1 indicates perfect correlation. This measure is analogous to the known Pearson correlation coefficient [47].

$$\text{Correlation} = \frac{\sum_i \sum_j (i-\mu_i)(j-\mu_j) p(i,j)}{\sigma_i \sigma_j} \tag{2.18}$$

   μ_i: Mean of row marginals: $\mu_i = \sum_{i=1}^{N} i \cdot \sum_{j=1}^{N} p(i,j)$
   μ_j: Mean of column marginals: $\mu_j = \sum_{j=1}^{N} j \cdot \sum_{i=1}^{N} p(i,j)$
   σ_i: Standard deviation of row marginals.
   σ_j: Standard deviation of column marginals.

4. **Variance:**
   Variance, also known as the Sum of Squares, measures how the values are dispersed around the mean of reference and neighbouring pixels. Its value increases when the grey-level values deviate from their mean [47].

$$\text{Variance} = \sum_i \sum_j (i-\mu)^2 p(i,j) \tag{2.19}$$

   μ: Mean intensity of the GLCM: $\mu = \frac{1}{N_g} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)$

5. **Inverse Difference Moment:**
   Inverse Difference Momentum (IDM) measures the homogeneity of an image by giving larger values to smaller grey-tone differences between pairs of elements. This metric is particularly sensitive to near-diagonal elements in the GLCM. IDM reaches its maximum value when all elements in the image are identical [47]. Its definition is as follows:

$$\text{IDM} = \sum_{i=1} \sum_{j=1} \frac{p(i,j)}{1+(i-j)^2} \tag{2.20}$$

6. **Sum Average:**
   Represent the sum of the average of all grey levels.

$$\text{Sum Average} = \sum_{i=2}^{2N_g} i \cdot p_{x+y}(i) \tag{2.21}$$

7. **Sum Variance:**
   It represents a heterogeneity measure that assigns higher weights to intensity level pairs from neighbours that deviate more from the mean [47].

$$\text{Sum Variance} = \sum_{i=2}^{2N} (i - \text{Sum Average})^2 p_{x+y}(i) \tag{2.22}$$

8. **Sum Entropy:**
   Sum entropy measures the non-uniformity present in an image, and in this case, the complexity of the texture [47].

$$\text{Sum Entropy} = -\sum_{i=2}^{2N} p_{x+y}(i) \log p_{x+y}(i) \tag{2.23}$$

   Since some of the probabilities may be zero, and $log(0)$ is not defined, Haralick, Shanmugam, and Dinstein [18] recommended that the term $log(p+\varepsilon)$ be used in place of $log(p)$ in entropy computations, having ε as an arbitrarily small positive constant.

9. **Entropy:**
   Entropy measures the complexity or randomness within an image by analysing the distribution of values in the GLCM. It reaches its maximum when all entries in the GLCM are equally probable, indicating a state of maximum disorder. Also known as "Shannon entropy," this metric is high when the matrix contains a wide variety of small, non-uniform values, which reflects complex and heterogeneous textures [47].

$$\text{Entropy} = -\sum_i \sum_j p(i,j) \log p(i,j) \tag{2.24}$$

   Since entropy is inversely related to ASM (or Energy), images with intricate, non-repetitive patterns tend to exhibit high entropy and low Energy.

10. **Difference Variance:**
    Represents the variance of the difference between grey levels.

$$\text{Difference Variance} = \sum_{k}^{2N_g} (k - \text{DA})^2 p_{x-y}(k) \tag{2.25}$$

Where Difference Average: $\text{DA} = \sum_k^{2N_g} k p_{x-y}(k)$

11. **Difference Entropy:**
    Difference entropy is a quantitative measure used to assess the level of randomness or disorder present in the contrast of an image [47].

$$\text{Difference Entropy} = -\sum_{i=0}^{N_g-1} p_{x-y}(i)\log p_{x-y}(i) \tag{2.26}$$

12. 13. **Information Measures of Correlation**
    Information Measures of Correlation (IMC) involves analysing different parameters using various techniques. Mutual information is normalised, and information correlation is set to infinity when the denominator is zero [47].

$$\text{IMC1} = \frac{HXY - HXY1}{\max(HX, HY)} \tag{2.27}$$

$$\text{IMC2} = \sqrt{1 - \exp[-2(HXY2 - HXY)]} \tag{2.28}$$

    Where:
    HXY: Entropy of $P(i,j)$: $HXY = -\sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j)\log p(i,j)$
    HX: Entropy of $p_x(i) = \sum_j p(i,j)$: $HX = -\sum_{i=1}^{N} p_x(i)\log p_x(i)$
    HY: Entropy of $p_y(j) = \sum_i p(i,j)$: $HY = -\sum_{j=1}^{N} p_y(j)\log p_y(j)$
    $HXY1$: $HXY1 = -\sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j)\log\left(p_x(i)\cdot p_y(j)\right)$
    $HXY2$: $HXY2 = -\sum_{i=1}^{N}\sum_{j=1}^{N} p_x(i)\cdot p_y(j)\log\left(p_x(i)\cdot p_y(j)\right)$

14. **Maximal Correlation Coefficient:**

$$\text{MCC} = \sqrt{\text{second largest eigenvalue of} Q} \tag{2.29}$$

    Where Matrix with $Q(i,j) = \frac{p(i,j)}{p_x(i)p_y(j)}$

    The matrix Q can be seen as the transition matrix for a Markov chain that represents the grey levels of neighbouring pixels. The Maximal Correlation Coefficient (MCC) indicates the rate at which the Markov chain converges; it serves as a measure of the texture's complexity, whose values range from 0 to 1, inclusive [47].

2.3.3 2D Shape-based Features

In this collection of shape features, descriptions of the 2D dimension and form of the ROI were incorporated. These characteristics are not influenced by the grey level intensity distribution and are consequently computed solely on the original image and mask. We will assume the following foundational premises [60]:

- N_p represents the number of pixels included in the region of interest

- N_f represents the number of lines defining the perimeter of the Mesh.
- A is the surface area of the Mesh in mm^2
- P is the perimeter of the Mesh in mm

Mesh Surface

To calculate the surface area, first, the signed areas, A_i, of each triangle are computed. The total surface area is then obtained by summing all of these calculated sub-areas. The use of signed areas ensures that we accurately determine the surface area - negative areas from triangles that fall outside the ROI will cancel out the excess area contributed by triangles that are partially inside and partially outside the ROI.

$$A_i = \frac{1}{2}\mathrm{Oa}_i \times \mathrm{Ob}_i \text{ (1)} \tag{2.30}$$

$$A = \sum_{i=1}^{N_f} A_i \text{ (2)} \tag{2.31}$$

Where $O_i a_i$ are edges of the i^{th} triangle present in the Mesh, formed by the a_i, b_i of the perimeter and the origin O.

Pixel Surface

The surface area of the ROI A_{pixel} is estimated by multiplying the number of pixels in the ROI by the surface area of a single pixel A_k.

$$A_{pixel} = \sum_{k=1}^{N_v} A_k \tag{2.32}$$

Perimeter

The perimeter of each individual line within the Mesh is systematically calculated, and the overall perimeter is derived by summing these values.

$$P_i = \sqrt{(\mathrm{a}_i - \mathrm{b}_i)^2} \text{ (1)} \tag{2.33}$$

$$P = \sum_{i=1}^{N_f} P_i \text{ (2)} \tag{2.34}$$

Where a_i and b_i are vertices of the i^{th} line in the perimeter.

Perimeter to Surface ratio

The perimeter-to-surface ratio is not dimensionless, which results in it being dependent on the surface area of the ROI. With this feature, lower values are indicative of a more compact shape (circle-like).

$$perimeter\ to\ surface\ ratio = \frac{P}{A} \tag{2.35}$$

Sphericity

Sphericity is a ratio between the perimeter of the tumour region and the perimeter of a circle with the same surface area as the tumour region, resulting in a measure of the roundness of the tumour shape relative to a circle. The possible values range from 0 to 1, where 1 is indicative of a perfect circle shape.

$$sphericity = \frac{2\pi R}{P} = \frac{2\sqrt{\pi A}}{P} \tag{2.36}$$

Where R is the radius of the circle with the same surface as the ROI, and equal to $\sqrt{\frac{A}{\pi}}$

Maximum 2D diameter

The maximum diameter refers to the greatest pairwise Euclidean distance between the vertices of the tumour surface mesh.

Major Axis Length

This feature indicates the length of the largest axis of the ellipsoid that encompasses the ROI, computed from the largest principal component, λ_{major}.

$$major\ axis = 4\sqrt{\lambda_{major}} \tag{2.37}$$

Minor Axis Length

Similar to the previous feature, the minor axis length indicates the length of the second-largest axis of the ellipsoid that encompasses the ROI, computed from the largest principal component, λ_{minor}.

$$minor\ axis = 4\sqrt{\lambda_{minor}} \tag{2.38}$$

Elongation

Elongation is the result of the relation between the two largest principal components in the ROI shape. The resultant values range between 1 (non-elongated) and 0 (maximally elongated, e.g., a $1D$ line).

$$elongation = \sqrt{\frac{\lambda_{minor}}{\lambda_{major}}} \tag{2.39}$$

2.3.4 Transform-based Features

2.3.4.1 Absolute gradient matrix

Facing the fact that first, second, and higher-order features cannot capture the variation degree of intensity across images, the image gradient has plenty of space to address and overcome this challenge [1].

Considering this problem, the Absolute Gradient Matrix (AGM) is able to provide information about spatial intensity changes across images, as high and low gradients are representative of abrupt and smooth variations of intensity, respectively.

Given an absolute gradient matrix, the $AGM(i, j)$ contains information about intensity variations in a particular neighbourhood. Here, i and j represent the number of central pixels in an $n \times n$ window, corresponding to the vertical and horizontal directions in an image of dimensions $m \times m$, where $i = j = m - (n - 1)$. Each AGM element is an Absolute Gradient Value (AGV) determined by the gradients in the "x" (Gx) and "y" (Gy) directions [1].

2.3.4.2 Histogram of Oriented Gradients

Since AGM features only describe the magnitude of gradients within images and do not provide the direction of gradients, Dalal and Triggs [12] introduced the Histogram of Oriented Gradients (HOG) to determine both the magnitude and direction of gradients within images.

HOG is analogous to AGM but specifically measures gradients using G_x and G_y, where $\theta = \arctan(G_y/G_x)$. It outlines the direction of edges in images, enabling HOG features to function as descriptors for local objects and respective shapes [1].

These features are derived by binning orientations to ascertain the gradient magnitudes at specific angles. For unsigned and signed gradients, the bins extend from 0° to 180° and 0° to 360°, respectively. When utilising eight bins, gradients are represented at 22.5° and 45° intervals.

HOG features reflect the magnitudes of gradients at predetermined orientations, with the number of parameters contingent upon the bin count.

2.3.4.3 Local Binary Pattern

Local Binary Patterns (LBP) operator functions as a grey-scale invariant texture measure based on a definition of texture within a local neighbourhood. For each pixel, a binary code is created by comparing its value to that of the central pixel, and a histogram is compiled to record the occurrences of various binary patterns [1, 49].

The notation employed for the LBP operator is denoted as $\mathrm{LBP}_{(P,R)}^{u2}$. In this context, the subscript signifies the application of the operator within a **(P, R)** neighbourhood configuration. The superscript **u2** indicates the utilisation of uniform patterns exclusively, categorising all non-uniform patterns under a singular label. Following the computation of the LBP-labelled image, denoted as $f_l(x,y)$, one can subsequently define the LBP histogram, which provides a representation of the distribution of these labelled patterns within the image [48].

$$H_i = \sum_{x,y} I\{f_l(x,y) = i\}, \quad i = 0,\ldots,n-1 \tag{2.40}$$

To compare histograms of image patches with different sizes, they must be normalised for a consistent description [48]:

$$N_i = \frac{H_i}{\sum_{j=0}^{n-1} H_j} \tag{2.41}$$

This texture characteristic has been extensively employed in the description of regions of interest, demonstrating its capability to elucidate the nuanced details of the surface. It systematically computes a local representation of the texture at the specified point of interest, thereby facilitating a more comprehensive understanding of the material's texture [24].

2.3.4.4 Gabor Filters

Gabor filters are widely used in the computer vision literature, especially in face recognition. A two-dimensional Gabor filter is a Gaussian kernel function modulated by a complex sinusoidal plane wave as:

$$G(x,y) = \frac{f^2}{\pi\gamma\eta} \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(j2\pi f x' + \varphi\right), \tag{2.42}$$

with $x' = x\cos\theta + y\sin\theta, \quad y' = -x\sin\theta + y\cos\theta$, where f is the frequency of the sinusoidal factor, the phase offset is φ, and γ is the spatial aspect ratio [13].

The response of a filter is determined by either convolution or multiplication of the image with the filter, depending on whether the work is done in the spatial or frequency domain, respectively (as stated by the convolution theorem). The Gabor filter's response highlights edges — areas with the most significant intensity variations - in specific directions and frequencies [1].

2.4 Machine Learning

Machine Learning (ML) is an Artificial Intelligence (AI) branch focused on analytical prediction building. The ML model learns from the given data, classifies general patterns, and makes decisions with negligible human interaction. This is principally employed when there is a complex problem that requires data analysis to be solved. Depending on requirements, ML can produce an efficient solution, surpassing complex problems [42]. Although ML mainly uses two types of learning techniques - (1) Supervised Learning and (2) Unsupervised Learning - we will only focus on the supervised technique.

2.4.1 Supervised Learning for Classification

Supervised ML is a computational technique where the algorithms attempt to analyse and understand the relationships between observations in a given training dataset. Then, a predictive

model is created, based on the learned relationships, that can infer outcomes for unseen received data [42].

In the supervised technique, there is an input, X, and a target, output variable, Y. Any used algorithm expressed as $f(X) = Y$ tries to approximate f as close as possible to reality - an ideal classifier.

2.4.2 Model Evaluation Metrics

Now that we understand what could be a model that addresses our problem, how can we determine if it is a good solution or compare it to existing alternatives? Evaluation metrics come into play to help us with this issue.

These performance metrics, in the binary classification context - where the target is $y \in 0, 1$ -, are typically derived from the confusion matrix. This matrix is constructed based on the following four possible combinations of the ground truth, defined by the provided label, (y), and predicted (y') classes:

- True Positive (TP): both y and y' are positive $(y = y' = 1)$;
- False Positive (FP): y' is positive, but y is negative $(y = 0, y' = 1)$;
- True Negative (TN): both y and y' are negative $(y = 0, y' = 0)$;
- False Negative (FN): the y' is negative, but y is positive $(y = 1, y' = 0)$.

Table 2.1: Common Machine Learning Evaluation Metrics with Descriptions and Equations

Metric	Description	Equation
Accuracy	Proportion of correct predictions	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	Correct positive predictions among all predicted positives	$\frac{TP}{TP+FP}$
Recall/Sensitivity	Correct positive predictions among all actual positives	$\frac{TP}{TP+FN}$
F1 Score	Harmonic mean of precision and recall	$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
Specificity	Correct negative predictions among all actual negatives	$\frac{TN}{TN+FP}$
AUC-ROC	Area under the ROC curve	(Graphical measure, computed numerically)

When comparing the defined labels with the predicted ones, we can count the outcomes and, from them, calculate several performance metrics, such as the ones represented in Table 2.1. These consist of a set of statistical indicators that we use to measure the effectiveness and suitability of

classifiers in relation to the classification data being modelled. The choice of metrics is crucial and depends on the specific nature of the problem.

These help us quantify and aggregate the quality of the trained model when it is validated against unseen data. The results of these evaluation metrics will indicate whether the classifier has performed optimally or if further refinement is needed. Note that when a dataset's labels are imbalanced, particularly concerning a specific class, the resulting classification model often exhibits a bias towards that class. [42].

2.4.3 Convolutional Neural Networks

The foundation of Convolutional Neural Networks (CNNs) was established in 1980 when Japanese researcher Fukushima [14] introduced the Neocognitron model. The proposed model was a deep-structured neural network that mimicked the visual cortex and was considered one of the earliest DL algorithms.

The conventional CNN architecture, conceived to extract features from grid-like matrix datasets, is presented in a distinctive structure characterised by specialised hidden layers, including convolutional and pooling, which are unique to them, and fully connected layers. [11, 27].

2.4.3.1 Convolutional Layer

The convolutional layer, designed to learn spatial hierarchies, is responsible for extracting features from the input data by applying a set of learnable filters known as kernels. During the training process, the CNN adjusts the kernel values to enhance feature extraction.

Convolutional kernels are small matrices that function similarly to sliding windows, moving through the input image to perform convolution operations. They compute the dot product between the kernel weights and the corresponding patches of the image, producing feature maps — representations that capture specific patterns or features, often called activation maps. In a convolutional layer, each unit is organised into feature maps and connects to local regions of the previous layer's feature maps through a shared set of weights, referred to as a filter bank. All units within a feature map utilise the same filter bank, while different feature maps in a layer employ distinct filter banks.

The stride defines the size of the step that the kernel takes as it moves across the input. A stride of 1 allows for detailed feature extraction and results in larger output maps, while higher stride values decrease both the output size and computational cost. This trade-off between spatial resolution and efficiency underscores the importance of stride as a critical parameter in the design of CNNs.

Following the convolution, a non-linear activation function, such as Rectified Linear Unit (ReLU), is usually applied to enable the model to understand complex patterns in the data.

This way, convolutional layers are vital for the effectiveness of CNNs in image classification, object detection, and semantic segmentation tasks, making them a powerful tool in the deep learning field.

2.4.3.2 Pooling Layer

Typically, pooling layers follow convolutional layers in the architecture of the network. Their main purpose is to decrease the spatial dimensionality of the input feature maps while keeping the most important information, which accelerates computation, reduces memory usage, and mitigates overfitting.

There are several kinds of pooling methods, with max pooling and average pooling being two common ones.

2.4.3.3 Fully Connected Layer

The Fully Connected (FC) layer, commonly referred to as the linear layer or fully linked layer, is characterised by each neuron connecting to every neuron in the previous layer; therefore, the name. Often used after convolution and pooling, it takes the features extracted from the previous layer as inputs and produces the final output for classification or regression tasks.

2.5 Model Architectures

2.5.1 Linear SVM

A Linear Support Vector Machine (SVM) can be implemented using the PyTorch framework by employing a single FC layer paired with a hinge loss function. This method enables us to integrate the SVM into the same infrastructure used for training and testing DL models while still preserving its fundamental learning principles.

The linear component of the FC layer is defined as follows:

$$f(\mathbf{x}) = \mathbf{W}\mathbf{x} + b \tag{2.43}$$

where $W \in \mathbb{R}^{1\times d}$ and $b \in \mathbb{R}$ are trained parameters, and d represents the input of the feature dimension. When used in conjunction with the hinge loss function, we achieve a linear outcome:

$$\mathscr{L}_{\text{hinge}}(\mathbf{x}, y) = \max\left(0, 1 - y \cdot f(\mathbf{x})\right) \tag{2.44}$$

where $y \in \{-1, 1\}$ - resulting from the mapping of the previous defined label $\{0, 1\}$ - , and the $f(x)$ being the output of the linear model.

Even though this model is not the focus of the research work carried out, it is still possible to use it to try to understand the predictive power of each feature vector merged into the DL models.

2.5.2 ResNet

Residual neural networks, commonly known as ResNet, were first introduced in 2015 by He et al. [19] to address the gradient vanishing and exploding problems that occur in DNNs as their depth

increases. These issues can cause the gradient to either become zero or grow excessively large. Consequently, rather than improving, the network's performance often stagnates or even declines.

Since the deeper the models, with more layers, the more capable they are of capturing more complex patterns, we expected them to perform better than a model with fewer layers. However, the gradient issue demonstrates that models struggle when needing to approximate identity mappings by multiple non-linear layers. The architecture presented by the authors brings a method designated "skip connections" that connects the activations of one layer to deeper layers without being interfered with in between layers, resulting in the formation of a residual block.

The secret behind ResNets comes from piling these residual blocks concurrently, as instead of learning the underlying mapping through layers, $H(x)$, the network fits the residual map, $F(x) = H(x) - x$ [56]. Solvers are able to reduce the weights of multiple non-linear layers to near zero to achieve identity mappings. The result is that if a layer negatively impacts network performance, it will be skipped, making training fast and overcoming the gradient issue.

The distinguishing feature of each existing ResNet architecture is its depth. For our research, we used ResNet-18, ResNet-50, and ResNet-101, which, as the name suggests, have 18, 50, and 101 layers, respectively. Additionally, in ResNets, we can define stages, which refer to a group of layers where the spatial resolution remains constant and usually consists of multiple residual blocks. In architecture, a change in dimension marks the start of a new stage.

2.5.3 EfficientNet

The following family of the CNNs was presented in 2019 by Tan and Le [57] as it introduces compound scaling, which, instead of scaling one network dimension at a time, uses a compound coefficient, ϕ, to scale on depth, number of channels, and resolution simultaneously.

The authors stated that different dimensions are dependent on each other. Naturally, we acknowledge that higher-resolution images would need increased depth and larger receptive fields to capture similar features effectively, the same way we need to include more pixels in larger images.

Let ϕ be a user-defined coefficient that regulates the availability of additional resources for model scaling, while α, β, and γ define how to allocate these extra resources to the network dimensions, such as $\alpha \geq 1, \beta \geq 1, \gamma \geq 1$:

$$depth : d = \alpha^{\phi}, \quad width : w = \beta^{\phi}, \quad resolution : r = \gamma^{\phi} \tag{2.45}$$

Where the subject is constrained in order to maintain the Floating Point Operations per Second (FLOPS) budget:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \tag{2.46}$$

The establishing of one baseline structure for the EfficientNet architecture starts with the EfficientNet-B0 where the compound scaling is applied: first by fixing $\phi = 1$, assuming there is one extra resource available, and then by performing a small grid search to identify the most optimal values for the constants, α, β, γ, respecting the defined constant in 2.46.

The EfficientNet family consists of models ranging from EfficientNet-B1 to EfficientNet-B7, which are derived by varying the value of $\phi \in \mathbf{Z}^+$. Each version is essentially a scaled-up version of the baseline model, EfficientNet-B0.

2.5.4 ConvNeXt

Inspired by Vision Transformers, Liu et al. [33] introduced a novel CNN architecture termed ConvNeXt, which integrates several innovations from this type of model. ConvNeXt retains the hierarchical, stage-wise structure characteristic of traditional CNNs. However, instead of employing the typical 7×7 convolution followed by max pooling at the input stage, it utilises a 4×4 convolution with a stride of 4, akin to the patch embedding mechanism used in Vision Transformers.

A significant shift from earlier CNNs is the adoption of Layer Normalisation in place of Batch Normalisation, which enhances stability and permits training with smaller batch sizes. Additionally, ConvNeXt features Gaussian Error Linear Unit (GELU) activations, replacing the commonly utilised ReLU, and omits the final activation within each block, which is in line with practices established in transformer models.

The architecture also incorporates techniques such as stochastic depth and residual connections, which contribute to improved regularisation and model depth. Collectively, these design choices yield a convolutional architecture capable of achieving competitive accuracy on benchmarks like ImageNet while maintaining computational efficiency and inductive biases. This makes CNNs particularly effective for visual tasks.

2.6 Fusion Strategies in Machine Learning

While CNN, using just one type of data, has shown outstanding results in recognising activities, there is even more potential when we combine different data sources. By bringing together multiple modalities, we can clear up ambiguities that might come from relying on a single source. This combination of information not only enables us to better understand things but also enhances the accuracy and reliability of recognition systems [15].

2.6.1 Early Fusion

Methods that rely only on early fusion first extract various unimodal features and then combine them in a single feature vector before feeding them into a single ML model for training [20]. The modalities could be fused in different processes, such as concatenation, pooling, or the application of a gated unit.

2.6.2 Middle Fusion

Middle fusion refers to the procedure of combining feature representations learned from middle layers of neural networks with features learned from other modalities as input to a final model. The

key distinction with early fusion is that loss is propagated back to the neural networks that participated in feature extraction during training, thereby allowing for improved feature representations for each training iteration [15].

2.6.3 Late Fusion

Late fusion involves using predictions from multiple models to make a final decision, which is why it is often called decision-level fusion, as it is applied at the model decision. Typically, various modalities are employed to train distinct models, with the ultimate decision being derived through an aggregation function that combines the predictions from these models.

Common examples of aggregation functions include averaging, where the result is the mean of model outputs; majority voting, where the most frequent class wins; weighted voting, similar to majority voting, but the value of each vote depends on its weight or the use of a meta-classifier to analyse the predictions generated by each model, that bring a new model to the outputs that learn how to combine the predictions. The choice of aggregation function is mainly empirical and tailored to the specific application and input modalities involved [15].

2.7 Explainable Artificial Intelligence

With the growth of ML solutions, complex problems can be addressed more easily. However, the adoption of these methods raises an important issue: the need to understand the "reasoning" behind the decisions made by these black box algorithms.

The issue presented has led to a surge of Explainable Artificial Intelligence (XAI) as a research field. The main objective is to provide transparency and understanding without compromising AI performance.

XAI is particularly important in high-risk domains, such as healthcare, where the decisions made by ML models may significantly impact patients' lives. If the models provided clear, human-understandable explanations for their decisions, our confidence would shift towards the insights provided by the models.

Explainability is essential in building trust and encouraging the adoption of AI systems among medical professionals and specialists, as it significantly enhances the reliability of models while ensuring fairness, robustness, and interpretability. By allowing end users, such as physicians, to comprehend the decisions made by deep neural networks, explainability instils confidence that AI is making accurate and impartial choices grounded in factual evidence, thereby aiding experts in addressing challenges more effectively. In an era of increasingly stringent regulations, such as the European guidelines for trustworthy AI, the capacity to elucidate model decisions is vital for compliance and for advancing scientific discovery, especially in sensitive domains like healthcare [4].

2.7.1 SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP) provides a unified, model-agnostic approach to interpreting the output of complex machine learning models by attributing the contribution of each input feature to a given prediction. Grounded in cooperative game theory, SHAP values represent the average marginal contribution of a feature across all possible coalitions of features.

Practical estimators such as Kernel SHAP approximate these values by sampling feature coalitions and fitting a weighted linear model to the resulting perturbed predictions, thus enabling application to any differentiable or non-differentiable model [35]. Furthermore, SHAP supports both local explanations—clarifying why a model made a specific prediction—and global explanations—summarising overall feature importance across a dataset—by aggregating individual SHAP values to reveal model behaviour patterns and potential biases.

2.7.2 Class Activation Mapping and Variations

Class Activation Mapping (CAM) and its extensions provide spatial explanations for convolutional neural networks by identifying image regions that most influence a particular class score. The original CAM method requires replacing the final fully connected layers with a global average pooling layer followed by a linear classifier; the resulting class activation map for class c is obtained as

$$M_c(x,y) = \sum_k w_k^c \, a_k(x,y),$$

where $a_k(x,y)$ is the k-th feature map at spatial location (x,y) and w_k^c is the corresponding weight for class c [68]. This formulation directly links feature map activations to class scores, yielding interpretable heat maps without backpropagation.

To accommodate arbitrary CNNs architectures, Grad-CAM generalises CAM by computing importance weights α_k^c as the spatial average of the gradients of the class score with respect to feature maps:

$$\alpha_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial y^c}{\partial a_k(i,j)},$$

and forming the class activation map via a weighted combination of feature maps followed by a ReLU nonlinearity, $M_c = \mathrm{ReLU}\left(\sum_k \alpha_k^c \, a_k\right)$ [52]. Grad-CAM++ further refines this approach by introducing pixel-wise weighting of gradients to better handle multiple instances of the same class in an image, resulting in sharper and more localised maps [10]. Subsequent variants—such as Score-CAM and Smooth-Grad-CAM++—enhance robustness to noise and improve localisation by leveraging class score perturbations and gradient smoothing techniques. Collectively, these methods form a theoretical framework for visual interpretability in CNNs, firmly grounded in gradient-based attribution and feature pooling operations.

Chapter 3

State of the Art

Our purpose with the state-of-the-art chapter is to provide a solid base in the research field of this work. With that in mind, we present a broader analysis of the techniques used in diagnosing medical conditions through imaging. Precisely characterise lung nodules through CT scans for diagnosis.

3.1 Fusion Techniques

Information fusion techniques are particularly relevant in the context of medical image analysis. Its study seeks to integrate data from different sources or natures to enhance the quality of classification or detection models, where features extracted using various methods (e.g., handcrafted and DL) can complement existing approaches to improve diagnosis methodologies.

We aim to provide a comprehensive understanding of the advancements in nodule characterisation, placing our focus on the role of fusion techniques in enhancing diagnostic performance, particularly in terms of accuracy, sensitivity, and robustness against imaging artefacts.

3.1.1 Decision-Level Fusion

The features are first used in decision-level fusion to train independent classifiers. Each classifier produces a classification decision or probability based on its features. Each classifier's decisions or probabilities are combined using weighted voting, majority voting, or probability averaging methods. The final decision is made based on the results of each classifier.

Xie et al. [64] gave us an algorithm, Fuse-TSD, that automatically takes texture, shape, and deep features to classify lung nodules in chest CT images. It uses a texture descriptor based on the GLCM, a Fourier shape descriptor, and a Deep Convolutional Neural Network (DCNN) to extract features. Then, classifiers, such as AdaBoosted Back Propagation Neural Network (BPNN), are applied to each feature, and a decision is made by fusion of the respective results. Evaluated on the Lung Image Database Consortium Image Collection (LIDC-IDRI) dataset, Fuse-TSD achieved an Area Under the Curve (AUC) of 96.65% when nodules with a composite malignancy rate of 3 were discarded (D1), 94.45% when they were considered benign (D2), and 81.24% when they

were deemed malignant (D3). If for each D1, D2, and D3, we compare the fusion with the best AUC results of the test set that does not use fusion, we obtain increments of 0,41, 1,54%, and 4,58%, respectively.

Muzammil et al. [41] investigates different fusion approaches based on deep features fusion and ensemble learning to classify lung nodules in CT scans. The authors propose two heterogeneous fusion techniques: fusion based on the Average Prediction Score (AVG-Predict) and fusion based on Majority Voting (MAX-VOTE). The results showed that the MAX-VOTE technique, combining the predictions of twelve individual classifiers, achieved the highest accuracy in binary classification, with 95.59% ± 0.27% against the 91.12% achieved by the AdaBoostM2 classifier, without fusion and trained on deep features from AlexNet. While in multi-classification, the SVM - Feature Fusion by Concatenation (SVM-FFCAT) method achieved superior performance, with an accuracy of 96.89%, an AUC of 99.21%, and a specificity of 97.70%. These results emphasise that fusion features with ensemble learning can significantly enhance the performance of lung nodule classification.

Ali et al. [3] propose a Transferable Texture CNN for lung nodule classification, whose architecture consists of three convolutional layers and an Energy Layer (EL), omitting pooling layers to reduce trainable parameters and computational complexity. The EL preserves texture information and learns during both forward and backward propagation. The model was evaluated on the LIDC-IDRI and LUNGx Challenge datasets. The texture CNN achieved an accuracy of 96.69% ± 0.72% and an error rate of 3.30% ± 0.72% on LIDC-IDRI. Transfer learning improved accuracy on LUNGx from 86.14% to 90.91%.

The study by Ali et al. [2] evaluated the performance of SVMs and AdaBoostM2 algorithms using deep features from VGG-16, VGG-19, GoogLeNet, Inception-V3, ResNet-18, ResNet-50, ResNet-101, and InceptionResNet-V2 by identifying the optimal layers. Their results showed that SVM was more efficient for deep features than AdaBoostM2. The proposed decision-level fusion technique demonstrates better results in terms of accuracy (90.46 ± 0.25%), recovery (90.10 ± 0.44%), and AUC (94.46 ± 0.11%). Although it was ranked second in specificity (92.56 ± 0.18%), the deviation is notably lower than the Texture CNN approach [3]. Furthermore, the classification accuracy based on the simple average of the prediction scores is also computed at 89.10%, which highlights the robustness and effectiveness of the decision fusion technique compared to other methods. For reference, the highest accuracy achieved for the non-fusion classifications tested was 86.28%, with ResNet-101 through SVM.

The CAD system proposed by Shaffie et al. [55] uses an appearance feature descriptor comprising a Histogram of Oriented Gradients, Multi-view Analytical Local Binary Patterns, and a Markov Gibbs Random Field. In addition, it employs a shape feature descriptor that includes Multi-view Peripheral Sum Curvature Scale Space, Spherical Harmonics Expansion, and a set of fundamental morphological features. Then, a stacked auto-encoder followed by a soft-max classifier is applied to generate the initial malignancy probability. The resultant probabilities are fed to the last network that returns the diagnosis. When comparing with a previous study, they note that the increase in the accuracy is slight (from 93.97% to 94.73%), which is predictable since the

features used to model the same nodule characteristics. However, the increase in system sensitivity from 90.48% to 93.97% represents a notable improvement, demonstrating that the new system, with additional features, is less affected by the segmentation process and image artefacts.

Li, Yang, and Jiao [28] evaluated the effectiveness of fusion models in predicting Axillary Lymph Node (ALN) metastases in breast cancer, comparing traditional radiomics models, DL radiomics models, and fusion models using Dynamic Contrast-Enhanced MRI (DCE-MRI) images. The imaging data were sourced from The Cancer Imaging Archive (TCIA) via the Duke-Breast-Cancer-MRI project. Handcrafted radiomic features and deep features were extracted from 3062 DCE-MRI images, with feature selection performed using mutual information algorithms and recursive feature elimination. The study found that the decision fusion model, integrating radiomic and deep features, outperforms traditional and DL models in all metrics, with increments of 2.81% and 2.58% in accuracy over them, respectively. Adding clinical features to the decision fusion model further increased the AUC. The findings demonstrate the efficacy of fusion models in predicting ALN metastases, with the decision fusion model showing significant potential to aid clinical decision-making in early-stage breast cancer treatment.

Alksas et al. [5] employs an approach that modifies the Local Trinary Pattern (LTP) to use three levels instead of two and a new pattern identification algorithm to capture the heterogeneity and morphology of the nodule. Then, the features were given as training data to a classification architecture based on hyper-tuned stacked generalisation to classify nodules, achieving an overall accuracy of 96.17%, with 97.14% sensitivity and 95.33% specificity. On the other hand, the original LBP and other classification structures resulted in lower performance when compared to the proposed approach. We can see an improvement of 4.56%, 4.12%, and 4.95% in accuracy, sensitivity, and specificity, respectively, over the DL-based model used on the same proposed two-stage stacking-based classification. As for radiomic features, we observed an improvement of 5.89%, 6.66%, and 5.22% on the same metrics.

Liu et al. [31] presents a novel method for classifying benign and malignant lung nodules by combining shallow visual features and deep features. The approach utilises separate pipelines for feature extraction and classification. Shallow features, including texture and morphology, are extracted using statistical 3D data analysis and Haralick's texture model, while morphological features are derived from parameters such as size and shape. SVMs are employed to classify these extracted features. The DL branch uses neural architecture search to design a deep model with three sub-branches and integrates a Convolutional Block Attention Module (CBAM) for enhanced feature learning. The classification results from both shallow and deep approaches are fused using a weighted voting method, achieving an accuracy of 91,21%, sensitivity of 90,27%, a specificity of 91,98% and an F1-score of 91,04%, evaluated with LIDC-IDRI data. It represents a 1.27% accuracy improvement over the best standalone branch, along with the highest specificity among all approaches.

3.1.2 Feature Level Fusion

Feature fusion in lung nodule characterisation involves the integration of information from multiple sources, which may include various imaging modalities, anatomical perspectives, or a mix of handcrafted and deep features into a cohesive representation. This integration can be accomplished through several strategies, such as pooling, attention mechanisms, or learned fusion networks. These methods enable the model to capture complementary insights from a diverse array of features. By leveraging fused features, classifiers can identify more complex relationships and patterns, leading to improved accuracy and robustness in malignancy predictions within state-of-the-art systems.

Farag et al. [13] explored feature fusion by extracting texture descriptors (Gabor filters and LBP) and shape (signed distance transform fused with LBP). They showed that Gabor filters, when implemented on a two-level cascaded framework with SVMs classifiers, obtained the best performance: AUC of 99% and an F1-score of 97.5%. Although this approach does not conclude that feature fusion is optimal, it does encourage the hypothesis that feature fusion, particularly with Gabor filters, can improve classification. We can also take from this study the possibility of carrying out the classification tasks separately, into nodule or non-nodule, and benign or malignant, to improve the cascade classifier.

Shaffie et al. [54] proposed a framework to accurately diagnose lung nodules by integrating two features: appearance features from a seventh-order Gibbs random field model that captures spatial heterogeneity in nodules and geometric features, defining their shape. Then, a deep autoencoder classifier uses these features to distinguish between malignant and benign nodules. Evaluated with data from the LIDC-IDRI, which included 727 nodules from 467 patients, the system demonstrated potential for lung cancer detection, achieving a classification accuracy of 91.20%. It reflects a clear improvement over using only geometric or appearance features, which just achieved an accuracy of 85.83% and 90.51%, respectively.

Saba et al. [51] proposed a method for early-stage lung nodule detection consisting of three main phases: nodule segmentation using Otsu's thresholding and morphological operations, feature extraction of geometric, texture, and deep features to select optimal features and serial fusion of the optimal features for classifying nodules as malignant or benign. The study experiments with the LIDC-IDRI dataset, using Otsu's algorithm and morphological erosion for segmentation. Handcrafted geometric and texture features are combined with deep features extracted using a VGG-19 model. Feature optimisation is performed using Principal Component Analysis (PCA), and the fused features are classified using multiple classifiers. Experimental results show that the proposed method outperforms existing approaches, achieving an accuracy of 99.0%, a sensitivity of 99.0%, and a specificity of 100%, applying fused features. It was able to surpass by 1% and 2% the sensitivity and specificity of the work done in [43], which, despite using segmentation, did not use fusion.

The CAD system presented by Yuan, Wu, and Dai [65] uses a multi-branch classification network with an effective attention mechanism (3D ECA-ResNet) to extract features from 3D

images of nodules, adapting dynamically to improve the extraction of key information. Structured data, such as diameter and other radiological characteristics, is transformed into a feature vector. The experimental results show that the system achieves an accuracy of 94.89%, a sensitivity of 94.91%, and an F1-score of 94.65%, with a false positive rate of 5.55%. The increase becomes evident if we compare the results obtained with the baseline defined for accuracy in 2D [64] and 3D [66] methods: 89.53% and 93.92%. The study concludes that the combination of multimodal data increases the effectiveness of the CAD system, making it more likely to assist doctors in diagnosing pulmonary nodules.

The study by Liu, Wang, and Aftab [32] emphasises the need to consider the temporal aspect in analysing pulmonary nodules. It employs a Faster R-CNN to generate ROI and extract temporal and spatial features from lung nodule data. A 3D CNN fuses these features, and a Time-Modulated Long Short-Term Memory (T-LSTM) model analyses trends and predicts the evolution and malignancy of lung lesions, incrementing the accuracy to 92.8% when compared with the Long Short-Term Memory (LSTM) (91.1%), RNN (87.1%) and SVM (81.2%) methods.

Zhao et al. [67] proposed a lung nodule detection method that integrates multi-scale feature fusion. Candidate nodules are detected using a Faster R-CNN with multi-scale features, achieving a sensitivity of 98.6%, a 10% improvement over single-scale models. For false positive reduction, a 3D CNN based on multi-scale fusion achieved 90.5% sensitivity at four false positives per scan.

Munoz et al. [40] used a predictive model, such as XGBoost, based on morphological characteristics extracted from CT scans, an approach called "3D-MORPHOMICS". Its premise is that morphological changes can be quantified and used in the diagnostic process since irregularities in the nodules are indicators of malignancy. The classification model, using only 3D-morphomic features, achieved an AUC of 96.4% on the National Lung Screening Trial (NLST) test set, and the combination with radiomic features resulted in even better performance, with an AUC of 97.8% on the NLST test set and 95.8% on the LIDC-IDRI dataset.

Based on Hybrid DL models, Li et al. [29] proposed a CAD system that integrates DL techniques for feature extraction and feature fusion. The system extracts relevant features using VGG16 and VGG19 networks with a CBAM. These features are reduced using PCA and fused via Canonical Correlation Analysis (CCA) to create effective representations. The final analysis uses an optimised Multiple Kernel Learning SVM-Improved Particle Swarm Optimisation (MKL-SVM-IPSO). The proposed system achieved 99.56% accuracy, 99.3% sensitivity, and an F1-score of 99.65% on the Lung Nodule Analysis 2016 (LUNA16) dataset, surpassing the respective baseline algorithms of other lung CAD systems by 3.68% in accuracy and by 7.33% in sensitivity. These results demonstrate its competitiveness in reducing false positives and negatives in nodule detection.

Iqbal et al. [22] presents an innovative technique for classifying medical image modalities by combining visual and textural features. A pre-trained CNN extracts deep features, while manual methods like Zernike moments, Haralick features, and Global-Local Pyramid Pattern (GLPP) capture relevant textural and statistical attributes. These fused features train ML classifiers such as SVM, K-Nearest Neighbor (KNN), and Decision Trees. The proposed approach outperformed

standalone pre-trained CNNs, from 93.32% to 96.08% in accuracy and from 93.34% to 96.31% in sensitivity.

3.1.3 Decision And Feature Fusion

The model proposed by Ma et al. [36], RGD, is an algorithm for nodule characterisation that makes use of radiomic features and Graph Convolutional Network (GCN) in multiple CNN architectures to achieve a complete characterisation, combining predictions for robust decision-making. Its process can be divided into two phases, incorporating the previously described fusion levels.

- **Feature Level:** The RGD model extracts radiomic features through LBP, HOG, and GLCM and simultaneously uses five distinct CNN architectures (AlexNet, GoogLeNet, VGG, ResNet, and AttentionNet) to extract deep features independently. The features extracted by the five CNNs are then aggregated by a GCN, which learns over the features in a graph structure. It creates a representation of the features that incorporates not only what the CNNs has learned but also how these features relate to each other, something that would not be captured by simply concatenating or merging fully connected layers. The radiomic features and characteristics learned by the GCN are then combined with the highest level CNN representation learned at the output layer of each 3D CNN to generate decision scores.

- **Decision Level:** Each CNN model is trained independently to produce a probability of a nodule being malignant or benign. Instead of making a decision based on a single model, the proposed model combines the decisions of the five models through a weighted average, where the accuracy of each model determines the weights. This ensemble learning approach allows the model to make use of the information from each of the classifiers.

The proposed solution with the radiomics, GCN, integrated with the CNN, yields substantial performance improvements over the original CNN models. Additionally, the ensemble of the CNNs achieves a higher average accuracy compared to a single CNN model, implying that employing multiple and diverse CNN architectures enhances the extraction of discriminative features.

The use of information fusion allowed RGD to achieve superior performance in the classification task when compared to models that use only one type of fusion (e.g. [64]) or none at all, with a mean accuracy of 93.25% ± 0.021, a sensitivity of 89.22% ± 0.045, a specificity of 95.82% ± 0.032, F1-score of 0.9114 ± 0.029, and AUC of 0.9629 ± 0.018, having a 1.3% accuracy increase over the previous integration of radiomics, GCN and CNN.

Those results with minor standard deviations show the effectiveness and robustness of the proposed method on lung nodule classification. By integrating the features extracted independently by each CNN and their relationships modelled in non-Euclidean space by the GCN, the model can capture more complete and adequate representations of the nodules. In addition, the fusion of decisions from multiple models results in a more robust classification.

3.2 Remarks

We all recognise that AI has shown potential for enhancing diagnostic, reducing false positives, and optimising the management of pulmonary nodules. However, generalisation, interpretability, and clinical integration remain major obstacles [30]. It is essential that these tools are validated on larger datasets that are representative of the population to be applied and that they are integrated into clinical workflows [63].

Standalone DL approaches still fail to overcome many challenges. For example, if the node segmentation is not accurate, the model may not be able to extract the features correctly, leading to inaccurate classification [17, 54]. Despite this, feature extraction has shown better accuracy and lower false-positive rates. Textural features, such as GLCM, are promising for differentiating nodules. Combining feature extraction methods with neural networks also optimises diagnosis [37].

On this thought, fusion-based techniques showed potential in classifying lung nodules, addressing the limitations of autonomous feature extraction methods. By using supplementary information from various feature domains, these approaches increase the accuracy and reliability of the diagnosis. The studies reviewed here highlight the promise of information fusion as a critical enabler of advanced CAD systems, leading the way for better clinical decision-making.

A clear example of this improvement is the study by Xie et al. [64], which showed AUC increases of 0.41%, 1.54%, and 4.58% when compared to results without fusion, depending on the treatment of nodules with malignancy rates 3 (D1, D2, and D3, respectively). Other studies have emphasised the benefits of combining complementary features. Shaffie et al. [55] improved accuracy by integrating appearance and geometric features, while Saba et al. [51] reached an accuracy of 99.0% with feature fusion. Yuan, Wu, and Dai [65] enhanced accuracy to 94.89% through multimodal fusion, and Liu, Yang, and Tsai [30] reported a striking 99.56% accuracy with a CAD system based on fusion. Additionally, Iqbal et al. [22] increased accuracy from 93.32% to 96.08% by combining visual and textural features.

These collective results, many of which show significant accuracy increases, highlight feature fusion's robust advantages in improving diagnostic accuracy and overall performance. While consistently indicating the benefits of information fusion, it demonstrates the importance of the approach for more efficient and reliable CAD systems in diagnosing pulmonary nodules.

Chapter 4

Datasets

This section presents some of the most well-known and widely used datasets in lung nodule detection and characterisation tasks. State-of-the-art studies used these datasets, which are public and complete and can be easily accessed through TCIA or the Grand Challenge.

The malignancy annotations across these datasets are derived either from the judgment of radiologists or from definitive clinical evidence. The LIDC-IDRI (and, by extension, LUNA16) relies exclusively on the subjective assessments of experienced radiologists without any histopathologic confirmation. In contrast, both the NLST and LUNGx datasets offer accurate ground truth labels, which are based on pathology reports or longitudinal follow-up examinations. Similarly, ANODE09 functions solely as a detection challenge and does not incorporate classifications distinguishing benign from malignant cases.

4.1 LIDC-IDRI

The LIDC-IDRI is a comprehensive collection of CT scans of the thorax designed for diagnosing lung cancer and detecting visualised lesions. This internationally accessible database is a valuable resource for developing CAD systems focused on lung cancer diagnosis and evaluation. Launched by the National Cancer Institute (NCI) and further developed by the Foundation for the National Institutes of Health (FNIH), with support from the Food and Drug Administration (FDA), this public-private partnership exemplifies the success of a consensus-based consortium.

The creation of this data registry involved collaboration among seven academic research centers and eight major medical imaging companies, resulting in a total of 1018 cases. Each case includes clinical thoracic CT scan images for individual subjects and an XML file detailing the results of a two-phase image annotation process. In the first phase, four radiologists independently reviewed CT images and annotated lesions into one of three categories: "nodule >=3 mm", "nodule <3 mm", and "non-nodule >=3 mm." During the second phase, the radiologists reviewed their annotations alongside the anonymised annotations of their peers to reach a consensus. This process was designed to allow for the accurate tallying of lung nodules on a CT scan with minimal human intervention, without requiring forced agreement among the radiologists [6].

4.2 LUNA16

The Lung Nodule Analysis 2016 (LUNA16) dataset utilises the publicly available LIDC-IDRI database mentioned earlier. Scans with a slice thickness greater than 2.5 mm were excluded from the dataset. In total, there are 888 CT scans included. The reference standard for this challenge consists of all 3 mm or larger nodules, which were accepted by at least 3 out of 4 radiologists. Annotations not part of the reference standard, such as non-nodules, nodules smaller than 3 mm, and nodules annotated by only 1 or 2 radiologists, are classified as irrelevant findings [53].

4.3 NLST

The National Lung Screening Trial (NLST) was a randomised controlled trial conducted by the Lung Screening Study group (LSS) and the American College of Radiology Imaging Network (ACRIN). The purpose of the trial was to evaluate whether screening for lung cancer with low-dose helical CT reduces mortality compared to screening with chest radiography in high-risk individuals. Approximately 54,000 participants were enrolled between August 2002 and April 2004. Data collection for the study has concluded, with the final information gathered by December 31, 2009, including low-dose CT scans from 26,254 of these subjects [44, 45].

4.4 ANODE09

The Automatic Nodule Detection 2009 (ANODE09) dataset gathers 55 anonymised CT scans provided by the University Medical Center Utrecht from the Nederlands–Leuvens Longkanker Screenings Onderzoek (NELSON) screening program. Five of those exams are supplemented by radiologists' annotations and serve as examples for optimisations. At the same time, the remaining 50 scans are reserved for testing only, with their reference annotations not publicly available. The dataset primarily includes scans from current and former heavy smokers aged 50–75, acquired using 16- or 64-slice CT scanners set for low-dose readings.

The dataset was randomly selected from a small subset chosen among the scans with the highest number of annotations. Scans with evident interstitial lung disease were excluded to avoid minimal nodular findings. Although ANODE09 emphasises larger nodules, unlike other datasets, it contains fewer scans, making it more representative of findings in asymptomatic heavy smokers. Given its design, the dataset is intended exclusively for testing and is not recommended for training CAD algorithms [16].

4.5 LUNGx

The LUNGx Challenge aimed to compare the performance of participants' computerised methods for lung nodule characterisation with six radiologists who participated in an observer study performing the same tasks on the same dataset. The scans were obtained from the clinical archive

at The University of Chicago with Institutional Review Board approval, removing all protected health information before being uploaded to TCIA.

The challenge required participants to use pre-trained algorithms to proceed with classification tasks. Ten scans were given for calibration purposes - five benign, five malignant - and were followed by a test set of 60 scans with 73 nodules (containing 37 benign and 36 malignant). Nodule sizes were measured through Response Evaluation Criteria in Solid Tumours guidelines, with benign nodules averaging 15.8 mm and malignant nodules 18.6 mm in diameter. This design balanced the nodule's size since it could be a malignancy indicator. In addition, spatial coordinates of each nodule were provided, without its sizes or diagnoses [25].

4.6 Limitations

It is important to recognise that medical imaging datasets significantly contribute to developing and validating CAD systems for lung cancer diagnosis. Nevertheless, even the most commonly used datasets, such as LIDC-IDRI and LUNA16, have some limitations that can impact model performance and generalizability. These challenges include a lack of annotated data, subjectivity in labelling, inter-observer variability, and potential biases within the datasets. These factors complicate the creation of robust and broadly applicable models. Additionally, we acknowledge that annotating medical images is time-consuming and costly, often leading to size-limited datasets. This limitation is further exacerbated by patient data privacy regulations, which restrict access to larger datasets [17].

Chapter 5

Methodology

5.1 Dataset

The experiments conducted in this research utilised the previously mentioned in section 4.1 LIDC-IDRI dataset.

5.1.1 Data Annotations

The dataset annotations provided include Lung Nodule Visual Attribute (LNVA) and Lung Nodule Malignancy (LNM) scores, along with the outlines of nodules that are 3mm or larger.

The LNM scores range from 1 to 5 (Highly Unlikely, Moderately Unlikely, Indeterminate, Moderately Suspicious, and Highly Suspicious, respectively). They are representative of a subjective assessment of the malignancy likelihood for a 60-year-old male smoker. In this regard, LNM score 4 is approximately three times more frequent than scores 1 and 5 and twice as frequent as score 2.

Although LNVAs such as subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, and texture pertain to node analysis, our study will focus exclusively on the images and the respective ROIs along with the LNMs scores.

5.1.2 Data Preprocessing

We filtered the data to exclude lung nodules, where the mean score of the respective LNMs was 3 (Indeterminate), and also those annotated by fewer than three radiologists.

The LIDC-IDRI dataset consists of lung CT scans from multiple institutions, leading to imaging variability from different scanners and protocols. To standardise the data, preprocessing transformations were applied [50]:

- HU values above 400 were capped at 400, and those below -1000 were set to -1000, corresponding to the values of hard tissues and air, respectively, as mentioned in section 2.2.1.

- HU values were normalized by rescaling them from the range $[-1000, 400]$ to $[0, 1]$ using linear transformation;
- The slice thickness and pixel spacing were adjusted from $1.91 \pm, 0.73$ mm, and 0.68 ± 0.08 mm - mean ± standard deviation calculated across the entire dataset -, respectively, to 1.0 mm.
- Initially, 2D representations were extracted from the CT images by cropping square patches centred on the lung nodule. Two spatial resolutions were considered: 32×32 and 64×64 pixels.
- Additionally, to the 32×32 2D representation, a simplified three-dimensional variant (2.5D) was also generated. This 2.5D consisted of three orthogonal anatomical planes - axial, sagittal, and coronal - each forming one channel of the resulting image, all centred on the nodule.

5.1.3 Data Labelling

To simplify the complexities of the multi-label classification task, we transformed it into a binary classification problem. We defined the labels as either 0 or 1, corresponding to benign and malignant nodules, respectively. Nodules with a LNM mean score below 3 were assigned a label of 0, while those with a score above were labelled as 1. This serves as the primary reference set for our experiments.

To comprehensively evaluate the performance of our methods, we prepared several subsets of the LIDC-IDRI dataset, each focusing on different labels of the data. These subsets are described below:

- **Extreme Scores**:
  In this subset, we further select nodules considered to be highly differentiated in terms of malignancy assessment. Contains 2D, single-channel images of 32×32 pixels. Specifically, we include only nodules with a mean malignancy score (LNM) strictly less than 2 (benign) or strictly greater than 4 (malignant). This selection reduces ambiguity by focusing on nodules with the most clear-cut diagnoses.
- **Central Scores**:
  Includes 2D, single-channel images of 32×32 pixels, but only for nodules with intermediate malignancy (mean $2 \leq$ LNM ≤ 4), representing cases with less definitive radiological assessment.

5.2 Handcrafted Feature Extraction

A comprehensive set of handcrafted radiomic features was extracted to characterise lung nodules from multiple complementary perspectives, including intensity, texture, morphology, gradient,

and frequency. For each feature category, predefined hyperparameters ensured consistency, reproducibility, and clinical relevance. The extraction of lung nodule Bounding Boxs (BBs) was guided by their corresponding binary masks, both represented as tensors of shape (C,H,W), where C is the number of channels, H and W are the image height and width, respectively.

Prior to feature computation, all input underwent a standardised preprocessing pipeline:

1. **Mask application**: Isolation of the ROI using the binary mask;
2. **NaN handling**: Replacement of invalid values with zero;
3. **Tensor standardization**: Output feature vectors were formatted as $(C,1,N)$, where N is the feature vector number of elements.

For multi-channel inputs $(C > 1)$, feature extraction was performed independently for each channel, and the resulting vectors were subsequently concatenated along the channel dimension.

First-Order Features

First-Order Features (FOF) were computed over the pixel intensities within the nodule mask, resulting in a one-dimensional feature vector that consists of 18 distinct features. The extracted features included measures of central tendency (mean, median, root mean squared), dispersion (variance, interquartile range, range, mean absolute deviation, robust mean absolute deviation), distribution shape (skewness and kurtosis), energy metrics (energy, total energy, uniformity), an entropy-based descriptor (entropy), and percentile-based statistics (minimum, maximum, 10th percentile, and 90th percentile). These features effectively capture the fundamental intensity distributions that reflect lesion heterogeneity.

Table 5.1: Summary of Extracted First-Order Features.

Category	**Features**
Central Tendency	Mean, Median, Root Mean Squared
Dispersion	Variance, Interquartile Range, Range, Mean Absolute Deviation, Robust Mean Absolute Deviation
Distribution Shape	Skewness, Kurtosis
Energy Metrics	Energy, Total Energy, Uniformity
Entropy-Based Descriptor	Entropy
Percentile-Based Statistics	Minimum, Maximum, 10th Percentile, 90th Percentile

Local Binary Pattern

The texture information was encoded using a uniform LBP approach, with a radius of 1 pixel and 8 sampling points. The 'uniform' method was employed to ensure rotation invariance while maintaining computational efficiency. The resulting LBP histogram, computed over the masked region and normalised to create a density distribution, produced a feature vector of size 10, corresponding to the 10 uniform binary patterns.

Histogram of Oriented Gradients

Edge orientation features were extracted using the HOG descriptor. The configuration employed 9 gradient orientation bins, with each cell size measuring 8×8 pixels and a block size of 2×2 cells. To improve robustness against intensity shifts, L2-Hys normalisation was employed. The final 324-dimensional HOG feature vector was obtained by flattening the concatenated histograms across all cells and blocks.

Gabor Filter Features

Frequency-domain texture characterisation was conducted using a Gabor filter bank consisting of 12 filters. These filters were created by combining three spatial frequencies (0.1, 0.2, 0.3) with four orientations (0°, 45°, 90°, 135°). Each filter was implemented with a 15×15 kernel. After convolving the masked image with these filters, the mean response within the nodule region was calculated for each filter, resulting in a vector of size 12 that captures texture patterns specific to both scale and orientation.

Shape-based Features

The morphological properties of the nodules were quantified using shape-based descriptors derived from binary masks. Eight features were extracted: boundary complexity was described using mesh surface, perimeter, and perimeter-surface ratio; size characteristics were captured by maximum diameter, major axis length, and minor axis length. Additionally, roundness and symmetry were evaluated using sphericity and elongation. These geometric characteristics form a feature vector of size 8 and represent clinical criteria that are often utilised in the evaluation of malignancy in radiology.

Haralick Texture Features

Texture features based on the GLCM were calculated to describe the spatial relationships between pixel intensities. The images were quantised to 8 bits, and GLCMs were created in four directions (0°, 45°, 90°, 135°) to ensure rotation invariance. Directional averages were then used to derive a final set of 13 Haralick descriptors, including angular second moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, and two information measures of correlation. The 13-sized vector effectively captures local intensity patterns and textural regularity.

Final Feature Composition

These experiments comprised 18 first-order statistics, 10 LBP features, 324 or 1764 HOG features - depending on whether the lung nodules BBs were 32×32 or 64×64 -, 12 Gabor filter responses, 8 shape descriptors, and 13 Haralick texture features.

5.3 Evaluation

In the context of health and medicine, for this thesis, we evaluated the models using several of the metrics previously mentioned in 2.4.2: (1) accuracy, which provides an overall model performance; (2) precision, aimed at minimising the risk of overdiagnosing; (3) sensitivity, to ensure we identify patients with the disease; and (4) AUC (or Area Under the ROC (AUC-ROC)), which captures the balance between sensitivity and specificity.

To ensure reliable and robust evaluations of predictions while preventing overfitting, we utilised a 5-fold cross-validation strategy. In this process, we partitioned the data for each fold into 80% for training and validation (comprising 72% for training and 8% for validation) and 20% for testing. The validation set constitutes 8% of the total dataset, representing 10% of the 80% designated for both training and validation. We computed the mean and standard deviation of the selected metrics, gathering insights from all test sets to illustrate our findings more clearly.

5.4 Experimental Procedures

5.4.1 Hyperparameter Optimization

A standardised configuration was established to ensure fair comparisons and reproducibility across all experimental runs, including baseline models, fusion implementations, and ablation studies. Models were trained using hyperparameter grid search with batch sizes of 32, 64, and 128, and learning rates set to 10^{-3}, 10^{-4}, and 10^{-5}. The Adam optimiser was employed with β_1 and β_2 defined as 0.9 and 0.999, respectively, and the loss function used was binary cross-entropy. Class weights, derived from the training set distribution, were incorporated to address class imbalance. Training continued for up to 100 epochs, with early stopping implemented if the validation metric did not improve for 25 consecutive epochs. All models were initialised using pre-trained weights from ImageNet.

Experiments were performed on CUDA-enabled Graphics Processing Units (GPUs) with mixed precision to accelerate training. The implementation utilised the PyTorch and PyTorch Lightning frameworks. To promote reproducibility, all random seeds were fixed for every stochastic operation, and three different seeds were employed to verify the robustness of the results. We retained the best-performing model checkpoints based on validation loss and validation AUC, as well as the checkpoint from the last epoch for evaluation purposes. From all the resulting versions, we present the one with the best performance.

The experimental pipeline was designed as a systematic investigation to comprehensively evaluate architecture selection, feature fusion effectiveness, and performance across varied dataset conditions. This structured approach ensured methodological rigour while building upon previous findings at each successive phase.

Table 5.2: Summary of Experimental Configuration.

Category	Details
Batch sizes	32, 64, 128
Learning rates	10^{-3}, 10^{-4}, 10^{-5}
Optimizer	Adam
Loss function	Binary cross-entropy
Training epochs	100 (early stopping, patience = 25)
Weight initialization	ImageNet pre-trained weights
Hardware	CUDA-enabled GPU, mixed precision
Framework	PyTorch & PyTorch Lightning
Reproducibility	Three fixed random seeds

5.4.2 Baseline Selection

The initial phase established optimal baseline architectures for both ResNet and EfficientNet families through evaluation on the Base dataset. For ResNet architectures, three variants spanning different depths were systematically compared: ResNet-18 representing a lightweight configuration, ResNet-50 providing moderate complexity, and ResNet-101 offering maximum representational capacity. Each architecture underwent training using the defined standardised experimental configuration.

Parallel evaluation was conducted for EfficientNet architectures, comparing EfficientNet-B0, EfficientNet-B1, and EfficientNet-B2 variants under identical protocols — this compound scaling evaluation aimed to identify the optimal balance between model complexity and lung nodule classification performance.

In addition to ResNet and EfficientNet, the ConvNeXt architecture was evaluated as well. Due to its considerably larger parameter count and computational requirements - over three times that of EfficientNet-B2 - only the ConvNeXt-Tiny variant was examined. This decision facilitated a fair comparison within reasonable constraints while still enabling an assessment of ConvNeXt's potential in relation to the other architectures.

The architecture selection established separate optimal baselines for all model families, providing the foundation for subsequent fusion experiments.

5.4.3 Single Fusion Evaluation

Building upon ResNet-18, the optimal architecture identified in the first phase, the second phase, the second phase systematically evaluated individual radiomic feature categories through dedicated fusion experiments on the Base dataset. We tested six distinct feature fusions, each integrating a single feature type with the ResNet backbone: FOF, LBP features, HOG features, Gabor Features (12 dimensions), 2D Shape Features, and Haralick Features. To identify the optimal fusion point, feature map fusion was performed after each ResNet-18 block to determine the optimal

fusion point within the network. Auxiliary features were projected to match spatial dimensions and fused via element-wise addition with the ResNet feature maps.

This phase provided crucial insights into the individual contribution of each radiomics feature category and established the foundation for identifying complementary feature combinations in subsequent phases.

5.4.4 Feature Selection

The third phase involved the selection of the three most effective radiomics feature categories based on the results from the second phase, along with predictive power evaluation through classification by SVM. Selection criteria encompassed multiple performance dimensions: improvement in metrics relative to baseline performance, consistency of optimal layer through hyperparameter choice, and complementary nature of the extracted radiomics information. This multi-criteria approach ensured that selected features not only provided individual performance gains but also offered diverse perspectives on nodule characterisation that could potentially synergise in combination experiments.

5.4.5 Multi-Feature Fusion Combinations

The fourth phase explored synergistic effects through the systematic combination of the three top-performing features identified in the third phase. Two-feature combinations were first evaluated through all possible pairings of the three selected features, generating three distinct combination models. Subsequently, a comprehensive three-feature combination model integrating all selected radiomics categories was implemented and evaluated. Each combination maintained the established fusion architecture principles, with feature injection at the now-defined optimal fusion point. The optimal combination was determined through testing, providing the best-performing fusion model for comprehensive dataset evaluation.

In addition to the primary fusion architecture, we also evaluated backbone models based on EfficientNet and ConvNeXt-Tiny, utilising the same three selected features. For these architectures, we systematically explored all configurations, including combinations of single features, two features, and the full three-feature set. However, since neither model underwent feature selection at an optimal layer, we implemented late fusion at the final layer of the network instead of middle fusion.

This approach enabled a comparative assessment of fusion models against their corresponding non-fused counterparts across different backbone architectures, while ensuring consistency in the feature combination protocols.

5.4.6 Dataset Ablation

The final ResNet experimentation conducted a comprehensive evaluation of both the baseline ResNet architecture and the optimal fusion model identified in the first and fourth phases, respectively, across all dataset variants. This comparative analysis included the Base dataset, Ex-

treme Scores subset, Central Scores subset, High Resolution variant - single-channel images of size 64×64 -, and 2.5D Multi-Plane representation - three-channel images of size 32×32. This systematic evaluation across diverse data conditions provided comprehensive insights into fusion effectiveness under varying nodule characteristics, spatial resolutions, and dimensional representations.

Table 5.3: Summary of LIDC-IDRI Dataset Variants Used in Experiments.

Set Name	**Dimension**	**Image Shape**	**Channels**	**Filter**
Base	2D	32×32	1	None
Extreme Scores	2D	32×32	1	LNM < 2 or LNM > 4
Central Scores	2D	32×32	1	$2 \leq$ LNM ≤ 4
High Resolution	2D	64×64	1	None
2.5D Multi-Plane	2.5D	32×32	3	None

5.4.7 Explainability

To study interpretability, we employed Grad-CAM visualisation for both the ResNet baseline and the fused ResNet model. Grad-CAM was used to generate class activation maps, offering insights into the specific regions of lung nodule images that affected the classification decisions of each model. This approach facilitated a direct comparison of their attention patterns and highlighted the interpretability improvements achieved through the integration of radiomics features.

Furthermore, SHAP was utilised on the fused ResNet model to further elucidate the contributions of individual radiomic features to the model's predictions. The SHAP analysis offered a quantitative understanding of feature importance and interactions, complementing the insights gained from Grad-CAM. Importantly, SHAP was exclusively applied to the fusion model, as it specifically addresses the interpretability of the integrated radiomics features within the network.

Chapter 6

Results

6.1 Baseline Model Performance

The summary of the baseline comparison results in Table 6.1 provides insight into the performance of each architecture family, without fusion, on the Base dataset, evaluated using the defined classification metrics.

Within the used variants of the ResNet family, ResNet-18 stood out with an overall performance, achieving an AUC of 82% and an accuracy of 82%. The findings indicate that its lightweight architecture was not only competitive but also surpassed deeper models in this specific context. In contrast, ResNet-50 and ResNet-101, despite their increased depth and capacity, did not exhibit a significant advantage; in fact, their AUC, accuracy, and precision metrics were somewhat lower. Consequently, we establish the ResNet-18 as the baseline for subsequent fusion information studies.

The EfficientNet variants demonstrated comparable results, with AUC and accuracy values consistently ranging between 79% and 80%. Notably, EfficientNet-B1 achieved the highest precision among the three variants at 81%, although its sensitivity was slightly lower at 74% compared to B0 and B2. The findings indicate that all three EfficientNet variants demonstrated strong and consistent performance, with only minor improvements noted when scaling up the models for this dataset. Given these minor performance differences, we opted to use the least complex model, the B0 architecture, for fusion.

The ConvNeXt (Tiny), while imposing higher computational demands, achieved an AUC of 78% and an accuracy of 79%. These results are marginally lower than those of the leading ResNet and EfficientNet models. Its precision was similar, but the sensitivity was notably lower, indicating that the model may be less efficient at accurately identifying positive cases in its current configuration.

The results indicate that lighter-weight architectures may perform competitively in lung nodule classification on the LIDC-IDRI dataset, achieving balanced performance across all metrics. In contrast, deeper or more complex models did not show significant improvements in performance and, in some instances, even performed worse. The inferior performance of deeper models may be

Table 6.1: Comparison of Baseline Architectures on the Base Dataset.

Model	AUC	Accuracy	Precision	Sensitivity
ResNet-18	**0.82 ± 0.01**	**0.82 ± 0.02**	**0.82 ± 0.05**	0.77 ± 0.04
ResNet-50	0.81 ± 0.02	0.81 ± 0.02	0.80 ± 0.05	0.78 ± 0.07
ResNet-101	0.79 ± 0.02	0.80 ± 0.02	0.76 ± 0.04	**0.79 ± 0.05**
EfficientNet-B0	0.80 ± 0.03	0.80 ± 0.03	0.78 ± 0.04	0.76 ± 0.07
EfficientNet-B1	0.80 ± 0.03	0.80 ± 0.04	0.81 ± 0.08	0.74 ± 0.05
EfficientNet-B2	0.79 ± 0.04	0.79 ± 0.04	0.77 ± 0.06	0.76 ± 0.07
ConvNeXt-Tiny	0.78 ± 0.02	0.79 ± 0.02	0.79 ± 0.03	0.71 ± 0.06

explained by the relatively small size of the BB, which may not provide sufficient spatial context for deeper architectures to leverage their increased capacity effectively. This emphasises the need to tailor model complexity to the specific dataset and task. Since ResNet-18 yielded the best results, we used it as the foundation for our feature fusion analyses.

6.2 Impact of Individual Radiomics Feature Fusion

Founded on the results retrieved from this phase, exposed in Table 6.2, the systematic evaluation of individual radiomics feature fusion highlights several key trends. These trends relate to the utility of specific feature categories and the effects of the fusion stage within the ResNet-18 architecture.

Shape features consistently demonstrated the highest performance among all tested radiomics groups, achieving an AUC of up to 84% and an accuracy of 84% when integrated at the earliest stage of the ResNet architecture - after the first ResNet-18 block. This finding underscores the significance of shape-based descriptors, as they provide particularly valuable information for the classification task when incorporated early in the feature fusion pipeline.

Both FOF and LBP also exhibited strong performance, particularly at the first and second fusion stages, after the first and second ResNet-18 blocks, achieving AUC scores of 83%. This underscores the importance of basic intensity and texture information in fusion when fusion occurs at an earlier stage. On the other hand, HOG, Gabor, and Haralick features produced slightly lower, yet still robust results, with AUCs in the range of 80–82% across fusion stages. While their impact was not as pronounced as that of shape features, their integration consistently enhanced or matched baseline performance.

For most feature types, the most favourable results were observed when fusion occurred after either the first or second ResNet block. This suggests that fusion at an early stage may promote a more effective joint representation, likely because the network can more effectively propagate and refine the complementary information provided by radiomics features through subsequent layers. In contrast, fusion at later stages generally yielded marginally lower metrics, potentially due to reduced opportunities for the model to integrate and utilise the auxiliary features as it approached the final decision layers.

Sensitivity values showed an upward trend when fusion was also conducted at earlier stages. Precision tended to be greatest for Shape, FOF, and LBP features during early fusion stages, demonstrating a favourable balance between false positives and false negatives.

Across all feature categories, integrating radiomics features with deep ResNet representations typically resulted in modest but steady improvements in classification metrics when compared to the baseline. This proposes that when radiomics features are integrated correctly, they supply additional information that may boost the model's ability to distinguish between classes. These findings lay the groundwork for future experiments focused on investigating multi-feature fusion strategies and optimising the integration of both handcrafted and deep features to boost classification efficacy.

Table 6.2: Performance of Single Feature Fusion with Different ResNet Backbone Stages.

Feature Vector	**Fusion Stage**	**AUC**	**Accuracy**	**Precision**	**Sensitivity**
FOF	1	0.83 ± 0.03	0.83 ± 0.03	0.82 ± 0.06	0.79 ± 0.08
FOF	2	0.82 ± 0.03	0.82 ± 0.02	0.80 ± 0.03	0.80 ± 0.03
FOF	3	0.81 ± 0.01	0.81 ± 0.02	0.79 ± 0.06	0.80 ± 0.06
FOF	4	0.81 ± 0.01	0.81 ± 0.01	0.80 ± 0.05	0.78 ± 0.07
LBP	1	0.83 ± 0.01	0.83 ± 0.01	0.82 ± 0.06	0.79 ± 0.06
LBP	2	0.82 ± 0.00	0.82 ± 0.00	0.81 ± 0.02	0.79 ± 0.02
LBP	3	0.81 ± 0.02	0.81 ± 0.02	0.84 ± 0.04	0.72 ± 0.02
LBP	4	0.81 ± 0.01	0.81 ± 0.01	0.78 ± 0.02	0.79 ± 0.05
HOG	1	0.82 ± 0.01	0.83 ± 0.01	0.81 ± 0.03	0.80 ± 0.03
HOG	2	0.80 ± 0.01	0.80 ± 0.01	0.79 ± 0.03	0.76 ± 0.03
HOG	3	0.80 ± 0.01	0.80 ± 0.01	0.75 ± 0.03	**0.81 ± 0.05**
HOG	4	0.80 ± 0.01	0.80 ± 0.01	0.77 ± 0.04	0.80 ± 0.08
Gabor	1	0.82 ± 0.01	0.82 ± 0.01	0.80 ± 0.04	0.79 ± 0.07
Gabor	2	0.81 ± 0.01	0.82 ± 0.01	0.83 ± 0.04	0.75 ± 0.05
Gabor	3	0.81 ± 0.01	0.81 ± 0.01	0.79 ± 0.02	0.80 ± 0.01
Gabor	4	0.82 ± 0.02	0.82 ± 0.02	0.81 ± 0.06	0.79 ± 0.07
Shape	1	**0.84 ± 0.02**	**0.84 ± 0.02**	0.84 ± 0.04	0.80 ± 0.05
Shape	2	0.83 ± 0.02	0.83 ± 0.02	0.81 ± 0.02	0.80 ± 0.04
Shape	3	0.82 ± 0.02	0.82 ± 0.02	0.80 ± 0.05	0.80 ± 0.02
Shape	4	0.82 ± 0.02	0.83 ± 0.01	**0.85 ± 0.05**	0.75 ± 0.08
Haralick	1	0.82 ± 0.02	0.82 ± 0.01	0.81 ± 0.03	0.80 ± 0.06
Haralick	2	0.81 ± 0.01	0.82 ± 0.01	0.82 ± 0.03	0.76 ± 0.03
Haralick	3	0.81 ± 0.01	0.82 ± 0.01	0.81 ± 0.04	0.77 ± 0.05
Haralick	4	0.80 ± 0.01	0.81 ± 0.00	0.81 ± 0.01	0.75 ± 0.03

6.3 Radiomics Feature Selection

The aim of this phase was to select the three most impactful categories of radiomics features, taking into account not just raw classification outcomes but also the stability across different hyperparameter configurations and the diversity in feature representation.

6.3.1 Predictive Power

Table 6.3: Performance of Individual Radiomics Feature Vectors with Linear SVM

Feature Vector	AUC	Accuracy	Precision	Sensitivity
FOF	0.57 ± 0.08	0.53 ± 0.11	0.51 ± 0.10	0.85 ± 0.13
LBP	0.55 ± 0.06	0.54 ± 0.09	0.42 ± 0.23	0.65 ± 0.37
HOG	0.57 ± 0.06	0.55 ± 0.09	0.52 ± 0.09	0.79 ± 0.16
Gabor	**0.65 ± 0.03**	**0.65 ± 0.04**	**0.61 ± 0.05**	0.64 ± 0.09
Shape	0.62 ± 0.12	0.58 ± 0.14	0.54 ± 0.11	**0.93 ± 0.07**
Haralick	0.57 ± 0.13	0.54 ± 0.15	0.42 ± 0.25	0.77 ± 0.39

We believed the results presented in Table 6.3 indicate that the evaluation of individual radiomic feature vectors using a linear SVM could reveal important findings.

Gabor features showed promise, achieving the highest AUC of 0.65 ± 0.03 and an accuracy of 0.65 ± 0.04 in the linear SVM, indicating a relatively stronger predictive capability.

Shape features also exhibited strong performance, with an AUC of 0.62 ± 0.12 and notably high sensitivity (0.93 ± 0.07), highlighting their exceptional ability to identify true positive cases correctly. In contrast, LBP, FOF, HOG, and Haralick features yielded lower AUCs, ranging from 0.55 to 0.57. Among these, LBP, FOF stood out due to their lower standard deviations, suggesting more stable and consistent performance.

Feature types that reflect clinically significant changes, like lesion shape, are likely to show better predictive performance due to their strong correlation with morphological differences. In contrast, texture-based features, while capturing acceptable intensity variations, may not align with the key visual cues needed for differentiation, especially with linear classifiers like the linear SVM. This may result in lower and less consistent predictive performance compared to shape-based descriptors.

6.3.2 Settings and Stages Stability

A comprehensive evaluation of each radiomic feature fusion was conducted across multiple batch sizes and learning rates to assess the robustness of their performance under varying training conditions. The results, detailed in Tables 6.4–6.5 and in 8.1–8.4, reveal that most features maintain stable AUC values with low standard deviations across different training settings.

Table 6.4: Best AUC and respective Stage for FOF Fusion with ResNet-18 Across Learning Rates and Batch Sizes.

Batch Size	**Learning Rate**					
	$\mathbf{10^{-3}}$		$\mathbf{10^{-4}}$		$\mathbf{10^{-5}}$	
	AUC	**Stage**	**AUC**	**Stage**	**AUC**	**Stage**
32	0.82 ± 0.02	1	0.81 ± 0.01	2	**0.80 ± 0.01**	1
64	**0.83 ± 0.03**	1	**0.82 ± 0.02**	2	**0.80 ± 0.01**	2
128	0.82 ± 0.02	1	0.80 ± 0.01	1	0.79 ± 0.01	2

This analysis importantly highlights distinct patterns concerning the optimal fusion stage, specifically, the ResNet layer, where feature fusion produces the best performance. The consistent stability observed in both the AUC and the optimal fusion stage for Shape, LBP, and FOF features emphasises their significance. They not only deliver strong performance but also enhance the clarity of model design by reliably identifying the most effective fusion stage.

Table 6.5: Best AUC and respective Stage for LBP Feature Fusion with ResNet-18 Across Learning Rates and Batch Sizes.

Batch Size	**Learning Rate**					
	$\mathbf{10^{-3}}$		$\mathbf{10^{-4}}$		$\mathbf{10^{-5}}$	
	AUC	**Stage**	**AUC**	**Stage**	**AUC**	**Stage**
32	**0.83 ± 0.01**	1	0.81 ± 0.02	1	**0.81 ± 0.01**	1
64	0.82 ± 0.02	1	**0.82 ± 0.00**	2	**0.81 ± 0.01**	1
128	0.81 ± 0.01	4	**0.82 ± 0.00**	1	0.81 ± 0.02	2

6.3.3 Selection

Shape features were selected as they consistently achieved the best AUC and outstanding sensitivity, making it a robust and reliable choice for capturing the geometrical characteristics of nodules. LBP was also chosen for its low standard deviation and consistent optimal performance when integrated at earlier stages of ResNet, indicating stability and reproducibility across various conditions, as shown in Table 6.5. Consequently, we have excluded Haralick features because they both convey texture information.

Although FOF did not lead in AUC, it demonstrated consistent results across layers and hyperparameters, as demonstrated in Table 6.4, coupled with lower variability, which makes it a dependable candidate for synergy in multi-feature fusion.

While certain features like Gabor may present isolated peaks in performance, we believe that the combination of high AUC, stability, and complementary information makes Shape, LBP, and FOF more suitable features for subsequent combination experiments. This multi-criteria selection aimed to identify features that not only perform well individually but also possess the potential for synergistic gains when integrated.

6.4 Multi-Feature Fusion

The results presented in Tables 6.6–6.8 illustrate distinct patterns in the efficacy of radiomics feature fusion across the ResNet-18, EfficientNet-B0, and ConvNeXt-Tiny backbone architectures.

For ResNet-18, fusion was implemented at empirically determined optimal stages, with LBP and FOF features fused at the first optimal stage and Shape features incorporated at the second. This systematic combination of the three top-performing radiomics features yielded significant synergistic benefits. Single-feature fusion resulted in incremental gains over the baseline, with Shape features making the most substantial contribution. Pairwise combinations further enhanced performance. and the highest overall metrics were achieved through the fusion of the three selected features. This highlights that, when guided by optimal feature selection and optimal placement, multi-feature fusion can significantly improve the model's discriminative ability and reliability.

Table 6.6: Comparative Performance of ResNet-18 Backbone with Fused Features.

Fused Features	**AUC**	**Accuracy**	**Precision**	**Sensitivity**
Baseline	0.82 ± 0.01	0.82 ± 0.02	0.82 ± 0.05	0.77 ± 0.04
FOF	0.83 ± 0.03	0.83 ± 0.03	0.82 ± 0.06	0.79 ± 0.08
LBP	0.83 ± 0.01	0.83 ± 0.01	0.82 ± 0.06	0.79 ± 0.06
Shape	0.84 ± 0.02	0.84 ± 0.02	0.84 ± 0.04	0.80 ± 0.05
FOF + LBP	0.83 ± 0.01	0.84 ± 0.01	0.83 ± 0.03	0.79 ± 0.02
FOF + Shape	0.84 ± 0.02	0.84 ± 0.03	0.86 ± 0.07	0.78 ± 0.02
LBP + Shape	0.84 ± 0.02	0.85 ± 0.02	0.85 ± 0.03	**0.80 ± 0.02**
FOF + LBP + Shape	**0.86 ± 0.01**	**0.86 ± 0.01**	**0.89 ± 0.03**	0.79 ± 0.04

Table 6.7: Comparative Performance of EfficientNet-B0 Backbone with Fused Features.

Fused Features	**AUC**	**Accuracy**	**Precision**	**Sensitivity**
Baseline	0.80 ± 0.03	0.80 ± 0.03	0.78 ± 0.04	0.76 ± 0.07
FOF	0.81 ± 0.03	0.82 ± 0.02	0.83 ± 0.01	0.74 ± 0.06
LBP	0.81 ± 0.03	0.81 ± 0.03	0.79 ± 0.06	0.78 ± 0.01
Shape	0.81 ± 0.03	0.81 ± 0.03	0.80 ± 0.06	0.76 ± 0.06
FOF + LBP	0.82 ± 0.02	0.82 ± 0.02	0.80 ± 0.02	0.80 ± 0.05
FOF + Shape	0.82 ± 0.03	0.82 ± 0.03	0.80 ± 0.03	**0.80 ± 0.04**
LBP + Shape	**0.83 ± 0.03**	**0.84 ± 0.02**	**0.86 ± 0.02**	0.75 ± 0.06
FOF + LBP + Shape	0.81 ± 0.02	0.81 ± 0.02	0.81 ± 0.03	0.75 ± 0.04

Table 6.8: Comparative Performance of ConvNeXt-Tiny Backbone with Fused Feature Combinations.

Fused Features	**AUC**	**Accuracy**	**Precision**	**Sensitivity**
Baseline	0.78 ± 0.02	0.79 ± 0.02	0.79 ± 0.03	0.71 ± 0.06
FOF + LBP	0.78 ± 0.01	0.79 ± 0.01	0.79 ± 0.04	0.73 ± 0.04
FOF + Shape	0.79 ± 0.02	0.79 ± 0.02	0.76 ± 0.04	**0.77 ± 0.04**
LBP + Shape	**0.79 ± 0.01**	0.79 ± 0.01	0.76 ± 0.01	**0.77 ± 0.04**
FOF + LBP + Shape	**0.79 ± 0.0**1	**0.80 ± 0.01**	**0.83 ± 0.05**	0.69 ± 0.05

In contrast, the EfficientNet-B0 and ConvNeXt-Tiny backbones utilised a late fusion strategy, integrating features at the final layer, since the study of an optimal stage would be highly time-consuming due to the higher number of possibilities. For EfficientNet-B0, the fusion of two features—especially LBP and Shape—yielded the best performance - AUC: 0.83, Accuracy: 0.84, Precision: 0.86. However, incorporating all three features did not lead to further improvement, indicating that the advantages of fusion may plateau or even decline without careful selection of the fusion stage and configurations. Similarly, ConvNeXt-Tiny displayed only modest gains from feature fusion, with the three-feature model achieving an AUC of 0.79, which is only slightly above the baseline.

These findings emphasise that the effectiveness of radiomics feature fusion is highly contingent upon both the backbone architecture and the integration strategy employed. A fusion at an optimal stage, as observed with ResNet-18, could be perceived as more effective, likely due to a better alignment with network representation hierarchies. In architectures where we only employed late fusion, there are still some performance gains, but they tend to be more modest, making the selection of feature combinations more critical.

6.5 Dataset Ablation

The primary objective of this study phase was to compare the performance of the standard ResNet-18 model with that of the fused ResNet-18 architecture across various dataset variations, each representing distinct labelling protocols, imaging resolutions, or data formats. The data obtained from this phase of the research is summarised in Table 6.9.

In the Extreme Scores subset, as anticipated, both models achieve high performance due to the well-separated nodules, which are easier to identify. However, the fused model still exhibits a measurable advantage. When compared directly, the fused model shows an improvement in AUC by 1%, as well as enhancements in accuracy and precision by 2%. Although sensitivity remains unchanged at 97%, this underscores an increased robustness relative to the baseline.

On a more ambiguous subset, the Central Scores subset, we can perceive more contrasting results. The AUC increases from 0.76 to 0.80, accuracy from 0.76 to 0.81, and precision from 0.72 to 0.81, indicating that the fused model is notably better.

The objective with the High Resolution and the 2.5D Multi-Plane sets was to test if the fused model, when subjected to more information, would still achieve an advantage over the non-fused approach and over the respective base set versions. Even though they did not perform better than the fused version trained on the base set, they did surpass the corresponding non-fused model. In the High Resolution variant, the AUC and accuracy increased by 4%, while precision and sensitivity improved by 6% and 2%, respectively. In the 2.5D variant, the AUC increases by 3%, accuracy by 2%, precision by 4%, and sensitivity by 3%. This shows that the advantage of fusion persists even in more complex, multi-plane representations.

In summary, across all the dataset variants, the fused model consistently outperforms the respective baseline. The more noticeable gains are over AUC, accuracy, and precision, while sensitivity shows more robustness improvements. These findings induce the belief that integrating shallow features within DL representations under diverse image conditions could bring major benefits.

Table 6.9: Comparison of ResNet-18 vs. Fused ResNet-18 Across Subsets of Data.

Dataset	Model	AUC	Accuracy	Precision	Sensitivity
Base	Baseline	0.82 ± 0.01	0.82 ± 0.02	0.82 ± 0.05	0.77 ± 0.04
	Fusion	**0.86 ± 0.01**	**0.86 ± 0.01**	**0.89 ± 0.03**	**0.79 ± 0.04**
Extreme Scores	Baseline	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.03	0.97 ± 0.02
	Fusion	**0.98 ± 0.01**	**0.98 ± 0.01**	**0.99 ± 0.02**	**0.97 ± 0.00**
Central Scores	Baseline	0.76 ± 0.04	0.76 ± 0.05	0.72 ± 0.07	0.73 ± 0.05
	Fusion	**0.80 ± 0.01**	**0.81 ± 0.01**	**0.81 ± 0.03**	**0.73 ± 0.04**
High Resolution	Baseline	0.81 ± 0.02	0.81 ± 0.02	0.80 ± 0.02	0.77 ± 0.03
	Fusion	**0.85 ± 0.03**	**0.85 ± 0.02**	**0.86 ± 0.03**	**0.79 ± 0.04**
2.5D Multi-Plane	Baseline	0.81 ± 0.02	0.82 ± 0.02	0.80 ± 0.02	0.77 ± 0.05
	Fusion	**0.84 ± 0.01**	**0.84 ± 0.01**	**0.84 ± 0.04**	**0.80 ± 0.03**

6.6 Model Explainability

6.6.1 Grad-CAM

The CAM generated by applying Grad-CAM to five representative cases offers a visual analysis of how the baseline and fused models emphasise different areas of the image.

The ResNet model without fusion tends to produce more diffuse activation patterns, with heatmaps that sometimes highlight significant regions in the background of the nodule. In contrast, while the heatmaps from the fused model may not always focus directly on the centre of the nodule, they more consistently capture areas that closely correspond to the nodule or its edges.

These findings indicate that merging radiomics with deep learning representations could not only enhance predictive performance but also improve model interpretability. This shows that the fused model is more adept at identifying the salient regions of lung nodules, thereby providing more meaningful and explainable visual evidence for its decisions when compared to the baseline model.

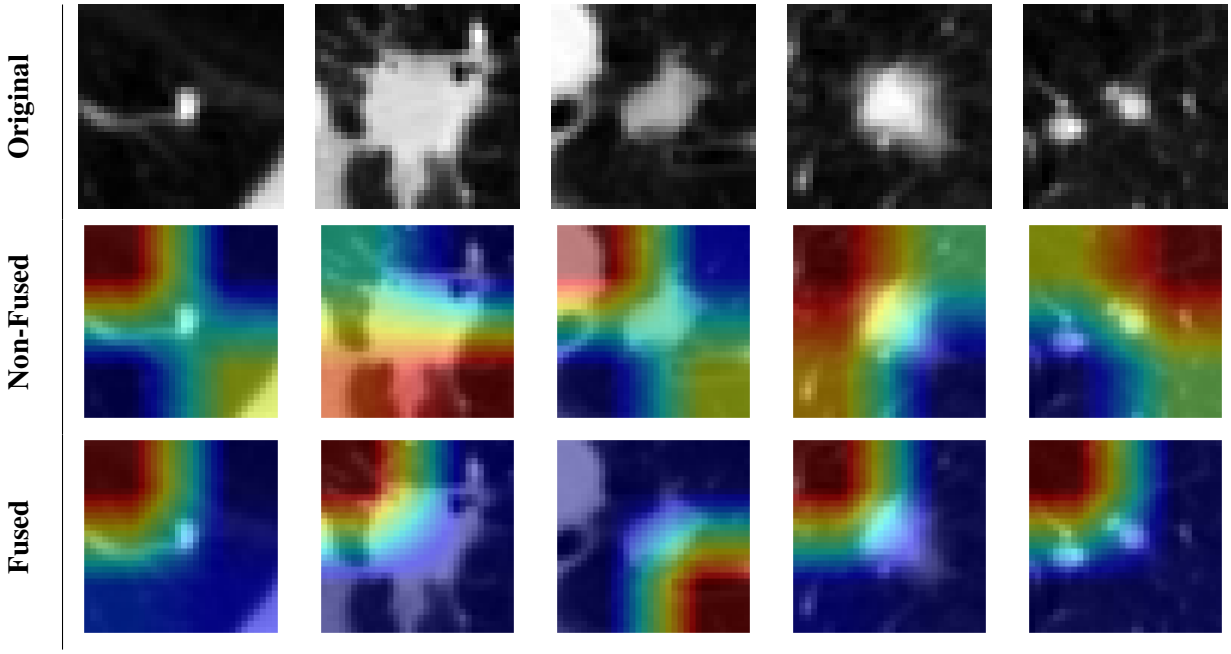
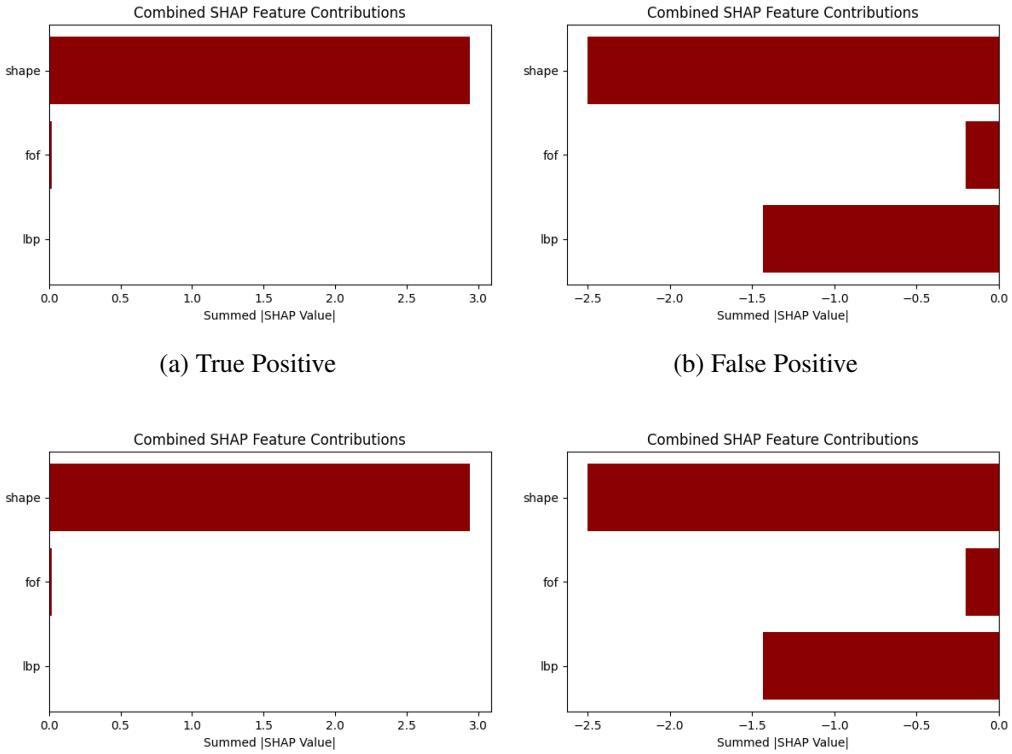


Figure 6.1: Comparison of Grad-CAM Heatmaps for five samples: Original images (top row), Non-Fused model (middle row), and Fused model (bottom row).

6.6.2 SHAP

Notably, SHAP offers local, case-specific explanations, clarifying not only which features are significant on average but also how they specifically affect individual prediction outcomes within the confusion matrix. The findings from the SHAP analysis, illustrated in Figure 6.2, offer valuable insights into the influence of specific radiomics features on the predictions made by the fused ResNet model across the four categories of the confusion matrix.



(a) True Positive

(b) False Positive

(c) False Negative

(d) True Negative

Figure 6.2: Local SHAP Analyses for images representative of confusion matrix categories.

The results were derived by summing the SHAP values for each feature. Higher positive SHAP values indicate a greater influence on positive predictions, while negative values indicate influence on negative predictions.

Across all categories, the shape feature vector emerges as the primary contributor to the model's decisions. For both true positives and false negatives, the SHAP values associated with Shape are significantly positive, suggesting that the presence of shape-related characteristics strongly reinforces the model's prediction of the positive class.

In the cases of false positives and true negatives, the SHAP values for the Shape are strongly negative, indicating that when the shape feature influences the model towards the negative class, it does so with conviction. In contrast, the FOF and LBP exhibit much smaller contributions across all categories: FOF has a minimal impact, whereas LBP plays a slightly more significant role in negative-class predictions.

This analysis confirms that integrating radiomics features with deep learning is most effective when shape features are included, as they significantly influence model predictions. The SHAP analysis demonstrates that incorporating shape features enhances both interpretability and predictive power. In contrast, the contributions of FOF and LBP are modest, suggesting that while multi-feature fusion offers complementary insights, the primary benefit comes from shape information.

Chapter 7

Conclusions

7.1 Research Questions

This work set out to answer three principal research questions:

1. **Does fusing information from shallow and deep feature extractors improve classification or generalisation performance compared to using a deep extractor?**
   From all the experiments performed, we were able to identify robust evidence that fusing shallow radiomics features with deep neural representations leads to consistent improvements in both classification and generalisation performance when compared to conventional standalone deep learning models.

   The established base models - the families of ResNet and EfficientNet, along with ConvNeXt-Tiny, demonstrate an already competitive performance on the LIDC-IDRI, defining a target to be surpassed by fused models. The fact that lighter architectures match or even outperform deeper counterparts, such as EfficientNet-B0 and ResNet-18, respectively, suggests that merely increasing network depth or complexity would result in diminishing returns for this specific domain.

   For example, combining Shape, LBP, and FOF features at empirically optimal points in ResNet-18 led to an improvement in AUC from 0.82 to 0.86, along with a notable increase in accuracy, precision, and sensitivity compared to the baseline, as shown in Table 6.6.

   Notably, most effective gains were observed when complementary radiomics features were fused, indicating the existence of synergy in combining diverse shallow descriptors with hierarchical deep features. In particular, with the 2D Shape descriptors, this synergy consistently contributed to the most significant performance boost; however, combinations of two or three features performed better than any individual feature when fused.

   The benefits of fusion generalisation are evident in the stability of performance improvements, which remain consistent across variations in model hyperparameters and fused blocks (Tables 6.4–8.3). Most of the optimal fusion occurred at earlier ResNet blocks, suggesting

that shallow features can be most effective when complementing deep representations before excessive abstraction.

In summary, these results underscore that carefully designed fusion could lead to improvements in state-of-the-art shallow and deep-only approaches.

2. **How does the fusion approach behave under varying dataset conditions, such as different sample sizes, bounding-box definitions, and image representations?**
The ablation study conducted—detailed in Table 6.9—investigated how the Fused model and its corresponding baseline performed under various conditions. Across all scenarios, the fusion solution demonstrated superior performance, with the exception of sensitivity in the Central Scores case, where it matched. However, the magnitude of its improvements was influenced by the complexity and representation of the subsets presented.

   The results indicate that the auxiliary information derived from radiomics features is particularly valuable in challenging classification scenarios, where deep models alone may struggle to capture subtle and discriminative cues.

   Importantly, fusion retained its advantage even as the nature of the input data changed. This suggests that shallow features encode information orthogonal to that of deep network representations and that their integration is beneficial across diverse imaging resolutions and multi-view data.

   Collectively, these findings indicate that the fusion approach is more robust and adaptable than the non-fused approach, capable of better generalising across a range of clinical data conditions.

3. **In what ways does information fusion contribute to the explainability of lung nodule malignancy predictions?**
Using Grad-CAM (Figure 6.1), we exhibited that the fused model generates CAM that exhibit more anatomically relevant attention, centralising on nodule regions and boundaries. In contrast, the non-fused model tended to show a more diffuse or background-oriented focus. This suggests that the incorporation of shallow features assists in directing the model's attention toward appropriate clinical structures, proposing more trustworthy and easier decisions for clinicians to interpret.

   The SHAP analyses (Figure 6.2) offer additional insights by quantifying the influence of each radiomics feature vector on individual predictions. SHAP revealed that the 2D Shape feature positively affects true positives and false negatives, while negatively impacting true negatives and false positives, always leaning towards the correct decision. The other tested features, FOF and LBP, contribute in a complementary, but less pronounced manner. This direct correlation between radiomics descriptors and model outputs facilitates interpretations that coincide with medical reasoning, thereby contributing to the improvement in transparency.

The improvements in interpretability are not simply coincidental, since they are a direct outcome of the fusion process. The rationale behind the model becomes more transparent and easier to audit by systematically integrating features that have clear clinical relevance.

7.2 Hypothesis

Based on our research and previously analysed research questions, the fusion paradigm between complementary types of features yields higher predictive capability. This statement is supported by tangible improvements in AUC, accuracy, precision, and sensitivity across all tested architectures over each respective baseline. The most significant gains were achieved when multiple handcrafted features were injected, after an optimal block, supporting the hypothesis of enhanced model performance.

The higher adaptability was also observed across different sets of data. Since in all data cases the fused model surpassed or matched the non-fused one, with the most significant disparities in more ambiguous or complex data subsets. These exhibits indicate that the fusion approach could be better, more robust, and more adaptable, further validating its reliability for diverse clinical scenarios. Information fusion is essential for enhancing advanced deep classifiers when they either reach a performance plateau or show decreased accuracy as their depth or complexity increases.

Additionally, we not only observed increments in quantitative metrics but also in interpretable relationships, as the fused model presented a less diffuse visualisation. Additionally, the fused features present a leaning towards the right decision, mainly observed with the Shape descriptor. These findings address the need for more reliable and understandable diagnostic systems.

The evidence collected from this study strongly supports our hypothesis that the fusion of shallow and deep features provides a superior paradigm for the detection and diagnosis of lung nodules. Fusion methods deliver gains in predictive performance, generalizability, interpretability, and adaptability, which are crucial for lung cancer diagnostics and for supporting real-world practice.

7.3 Main Contributions

This dissertation makes significant strides in the realm of CAD for lung cancer, highlighting several pivotal advancements:

- **Development of a Fusion-Based Diagnostic Framework:** This research introduces a novel deep learning model that synergistically combines handcrafted radiomics features - such as Shape, LBP, and the FOF - with deep features derived from ResNet architectures. The integration of these diverse feature sets led to marked improvements in the model's classification accuracy compared to traditional baseline models, thereby enhancing diagnostic precision.

- **Systematic Evaluation of Fusion Strategies:** Through rigorous experimentation that spanned multiple stages of feature fusion and architectural variations, this work successfully identified the most effective fusion points, particularly at earlier layers of the network. These insights maximised the beneficial interactions between shallow and deep features, thus optimising overall model performance.
- **Enhanced Explainability in AI-Driven Diagnostics:** By employing advanced interpretability techniques such as Grad-CAM and SHAP, the decision-making process of the model was rendered more transparent. This enhancement not only validated the importance of shape features in the diagnostic process but also fostered greater trust and confidence in the AI 's predictive outputs among clinicians.
- **Empirical Validation Across Diverse Data Conditions:** The robustness of the proposed fusion approach was thoroughly validated under a variety of dataset conditions, encompassing different image representations, sizes, and bounding box strategies. This comprehensive examination underscored the framework's versatility and broad applicability in real-world scenarios.
- **Contribution to SDG 3:** This research aligns with and advances SDG 3, focusing on health and well-being. Creating non-invasive, interpretable, and accessible diagnostic tools contributes significantly to the early detection of lung cancer and promotes equitable access to healthcare resources.

7.4 Future Work

Building upon the foundation established for fusion-based lung nodule characterisation, numerous intriguing avenues for future research have been identified:

- **Application to 3D and Temporal Data:** Further development could involve transitioning from current 2D and 2.5D image representations to fully adapting 3D CT volumes and longitudinal patient scans. This change would facilitate improved nodule characterisation by leveraging the spatial nuances and temporal dynamics of lung nodules.
- **Automated Feature Selection and Optimisation :** The exploration of implementing attention mechanisms or strategies for feature selection holds the promise of dynamically adjusting the weighting of features. This could significantly enhance the performance and generalizability of the fusion models.
- **Hyperparameter Tuning of Feature Extractors:** A systematic approach to hyperparameter optimisation for each handcrafted feature extractor would likely yield improvements in the quality and discriminative power of the fused features, leading to better diagnostic outcomes.

- **Multi-Label Classification:** Investigating methods for multi-label classification could enable the model to predict multiple attributes or characteristics of a nodule simultaneously, encompassing features such as malignancy, calcification, and spiculation, thereby providing a more comprehensive diagnostic overview.
- **Clinical Validation and Deployment:** Engaging in collaboration with medical professionals to validate the model within real-world clinical environments is critical. This step would include its integration into existing radiology workflows to facilitate practical application and acceptance by the medical community.
- **Benchmarking Against Multicenter Datasets:** Conducting evaluations of the model across diverse datasets collected from multiple hospitals or geographic regions would serve to test its generalizability and fairness, ensuring that it works effectively across varied populations and clinical settings.

References

[1] Ali Abbasian Ardakani et al. “Interpretation of radiomics features–A pictorial review”. In: *Computer Methods and Programs in Biomedicine* 215 (Mar. 1, 2022), p. 106609. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2021.106609. URL: https://www.sciencedirect.com/science/article/pii/S0169260721006830.

[2] Imdad Ali et al. “Deep Feature Selection and Decision Level Fusion for Lungs Nodule Classification”. In: *IEEE Access* 9 (2021), pp. 18962–18973. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3054735. URL: https://ieeexplore.ieee.org/document/9335996/.

[3] Imdad Ali et al. “Efficient Lung Nodule Classification Using Transferable Texture Convolutional Neural Network”. In: *IEEE Access* 8 (2020), pp. 175859–175870. DOI: 10.1109/ACCESS.2020.3026080. URL: https://ieeexplore.ieee.org/document/9204580/.

[4] Sajid Ali et al. “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence”. In: *Information Fusion* 99 (Nov. 1, 2023), p. 101805. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.101805. URL: https://www.sciencedirect.com/science/article/pii/S1566253523001148.

[5] Ahmed Alksas et al. “A novel higher order appearance texture analysis to diagnose lung cancer based on a modified local ternary pattern”. In: *Computer Methods and Programs in Biomedicine* 240 (Oct. 2023), p. 107692. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2023.107692. URL: https://www.sciencedirect.com/science/article/pii/S0169260723003577.

[6] Samuel G. Armato III et al. *Data From LIDC-IDRI*. Published: Dataset. 2015. DOI: 10.7937/K9/TCIA.2015.LO9QL9SX. URL: https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX.

[7] Philip Baum et al. “Incidental Pulmonary Nodules: Differential Diagnosis and Clinical Management”. In: *Dtsch Arztebl Int* 121.25 (Dec. 13, 2024). Place: Germany, pp. 853–860. ISSN: 1866-0452. DOI: 10.3238/arztebl.m2024.0177. URL: http://dx.doi.org/10.3238/arztebl.m2024.0177.

[8] Thorsten M. Buzug. “Computed Tomography”. In: *Springer Handbook of Medical Technology*. Ed. by Rüdiger Kramme, Klaus-Peter Hoffmann, and Robert S. Pozos. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 311–342. ISBN: 978-3-540-74658-4. DOI: 10.1007/978-3-540-74658-4_16. URL: https://doi.org/10.1007/978-3-540-74658-4_16.

[9] Angela Cantatore and Pavel Müller. *Introduction to computed tomography*. Report. Publication Title: Introduction to computed tomography. Kgs.Lyngby: DTU Mechanical Engineering, 2011.

[10] Aditya Chattopadhay et al. “Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks”. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018, pp. 839–847. DOI: 10.1109/WACV.2018.00097.

[11] Dengyuhan Dai. “An Introduction of CNN: Models and Training on Neural Network Models”. In: *2021 International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR)*. 2021 International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR). Journal Abbreviation: 2021 International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR). Nov. 5, 2021, pp. 135–138. DOI: 10.1109/ICBAR55169.2021.00037.

[12] N. Dalal and B. Triggs. “Histograms of oriented gradients for human detection”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. ISSN: 1063-6919. June 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177. URL: https://ieeexplore.ieee.org/document/1467360/ (visited on 06/17/2025).

[13] Amal A Farag et al. “Feature fusion for lung nodule classification”. In: *Int J Comput Assist Radiol Surg* 12.10 (Oct. 2017). Place: Germany, pp. 1809–1818. ISSN: 1861-6410, 1861-6429. DOI: 10.1007/s11548-017-1626-1. URL: http://link.springer.com/10.1007/s11548-017-1626-1.

[14] Kunihiko Fukushima. “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”. In: *Biological Cybernetics* 36.4 (Apr. 1, 1980), pp. 193–202. ISSN: 1432-0770. DOI: 10.1007/BF00344251. URL: https://doi.org/10.1007/BF00344251.

[15] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. “Early vs Late Fusion in Multimodal Convolutional Neural Networks”. In: *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. 2020, pp. 1–6. DOI: 10.23919/FUSION45008.2020.9190246.

[16] Bram van Ginneken et al. “Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study”. In: *Medical Image Analysis* 14.6 (2010), pp. 707–722. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2010.05.005. URL: https://www.sciencedirect.com/science/article/pii/S1361841510000587.

[17] Yu Gu et al. “A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning”. In: *Computers in Biology and Medicine* 137 (Oct. 2021), p. 104806. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.104806. URL: https://www.sciencedirect.com/science/article/pii/S0010482521006004.

[18] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. “Textural Features for Image Classification”. In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3.6 (Nov. 1973), pp. 610–621. ISSN: 0018-9472, 2168-2909. DOI: 10.1109/TSMC.1973.4309314. URL: http://ieeexplore.ieee.org/document/4309314/ (visited on 06/16/2025).

[19] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385 (visited on 06/23/2025).

[20] Shih-Cheng Huang et al. “Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines”. In: *npj Digital Medicine* 3.1 (Oct. 16, 2020), p. 136. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00341-z. URL: https://doi.org/10.1038/s41746-020-00341-z.

[21] International Agency for Research on Cancer. *Trachea, Bronchus and Lung Cancer Fact Sheet*. 2024. URL: https://gco.iarc.who.int/media/globocan/factsheets/cancers/15-trachea-bronchus-and-lung-fact-sheet.pdf.

[22] Saeed Iqbal et al. “Fusion of Textural and Visual Information for Medical Image Modality Retrieval Using Deep Learning-Based Feature Engineering”. In: *IEEE Access* 11 (2023), pp. 93238–93253. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3310245. URL: https://ieeexplore.ieee.org/document/10234403/.

[23] Mustafa Mohammed Jassim and Mustafa Musa Jaber. “Systematic review for lung cancer detection and lung nodule classification: Taxonomy, challenges, and recommendation future works”. In: *Journal of Intelligent Systems*. Journal of Intelligent Systems 31.1 (2022), pp. 944–964. DOI: 10.1515/jisys-2022-0062. URL: https://doi.org/10.1515/jisys-2022-0062 (visited on 06/17/2025).

[24] Navneet Kaur, Nahida Nazir, and Manik. “A Review of Local Binary Pattern Based texture feature extraction”. In: *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). Sept. 2021, pp. 1–4. DOI: 10.1109/ICRITO51393.2021.9596485. URL: https://ieeexplore.ieee.org/document/9596485/ (visited on 06/16/2025).

[25] Justin S. Kirby et al. “LUNGx Challenge for computerized lung nodule classification”. In: *Journal of Medical Imaging* 3.4 (2016). Publisher: SPIE, p. 044506. DOI: 10.1117/1.JMI.3.4.044506. URL: https://doi.org/10.1117/1.JMI.3.4.044506.

[26] Anna Rita Larici et al. “Lung nodules: size still matters”. In: *European Respiratory Review* 26.146 (Dec. 20, 2017), p. 170025. DOI: 10.1183/16000617.0025-2017. URL: https://publications.ersnet.org//content/errev/26/146/170025.abstract.

[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature* 521.7553 (May 1, 2015), pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539. URL: https://doi.org/10.1038/nature14539.

[28] Xue Li, Lifeng Yang, and Xiong Jiao. “Comparison of Traditional Radiomics, Deep Learning Radiomics and Fusion Methods for Axillary Lymph Node Metastasis Prediction in Breast Cancer”. In: *Acad Radiol* 30.7 (Nov. 11, 2022). Place: United States, pp. 1281–1287. ISSN: 1878-4046. DOI: 10.1016/j.acra.2022.10.015. URL: http://dx.doi.org/10.1016/j.acra.2022.10.015.

[29] Yang Li et al. “Research on lung nodule recognition algorithm based on deep feature fusion and MKL-SVM-IPSO”. In: *Scientific Reports* 12.1 (Oct. 18, 2022), p. 17403. ISSN: 2045-2322. DOI: 10.1038/s41598-022-22442-3. URL: https://doi.org/10.1038/s41598-022-22442-3.

[30] Jonathan A. Liu, Issac Y. Yang, and Emily B. Tsai. “Artificial Intelligence (AI) for Lung Nodules, From the AJR Special Series on AI Applications”. In: *American Journal of Roentgenology* 219.5 (Nov. 2022), pp. 703–712. ISSN: 0361-803X, 1546-3141. DOI: 10.2214/AJR.22.27487. URL: https://doi.org/10.2214/AJR.22.27487.

[31] Shaojun Liu et al. “Classification of Benign and Malignant Pulmonary Nodules Based on Mixed Features”. In: *2023 42nd Chinese Control Conference (CCC)*. ISSN: 1934-1768. July 2023, pp. 8803–8808. DOI: 10.23919/CCC58697.2023.10240557.

[32] Xindong Liu, Mengnan Wang, and Rukhma Aftab. “Study on the Prediction Method of Long-term Benign and Malignant Pulmonary Lesions Based on LSTM”. In: *Frontiers in Bioengineering and Biotechnology* 10 (Mar. 2, 2022), p. 791424. ISSN: 2296-4185. DOI: 10.3389/fbioe.2022.791424. URL: https://www.frontiersin.org/articles/10.3389/fbioe.2022.791424/full.

[33] Zhuang Liu et al. *A ConvNet for the 2020s*. Mar. 2, 2022. DOI: 10.48550/arXiv.2201.03545. arXiv: 2201.03545[cs]. URL: http://arxiv.org/abs/2201.03545 (visited on 06/28/2025).

[34] Konstantinos Loverdos et al. “Lung nodules: A comprehensive review on current approach and management”. In: *Annals of Thoracic Medicine* 14.4 (2019). ISSN: 1817-1737. URL: https://journals.lww.com/aotm/fulltext/2019/14040/lung_nodules__a_comprehensive_review_on_current.2.aspx.

[35] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. Nov. 25, 2017. DOI: 10.48550/arXiv.1705.07874. arXiv: 1705.07874[cs]. URL: http://arxiv.org/abs/1705.07874 (visited on 06/29/2025).

[36] Ling Ma et al. “A novel fusion algorithm for benign-malignant lung nodule classification on CT images”. In: *BMC Pulmonary Medicine* 23.1 (Nov. 28, 2023), p. 474. ISSN: 1471-2466. DOI: 10.1186/s12890-023-02708-w. URL: https://doi.org/10.1186/s12890-023-02708-w.

[37] P. Mathumetha, R. Sivakumar, and Ananthakrishna Chintanpalli. *Feature Extraction Methods and Deep Learning Models for Detection of Cancerous Lung Nodules at an Early Stage – Recent Trends and Challenges*. Place: Vellore, India Publisher: Vellore Institute of Technology. 2024. DOI: 10.2139/ssrn.4756114.

[38] Michalis Mazonakis and John Damilakis. “Computed tomography: What and how does it measure?” In: *European Journal of Radiology* 85.8 (Aug. 1, 2016), pp. 1499–1504. ISSN: 0720-048X. DOI: 10.1016/j.ejrad.2016.03.002. URL: https://www.sciencedirect.com/science/article/pii/S0720048X16300754.

[39] John D Minna, Jack A Roth, and Adi F Gazdar. “Focus on lung cancer”. In: *Cancer Cell* 1.1 (Feb. 1, 2002). Publisher: Elsevier, pp. 49–52. ISSN: 1535-6108. DOI: 10.1016/S1535-6108(02)00027-2. URL: https://doi.org/10.1016/S1535-6108(02)00027-2 (visited on 06/14/2025).

[40] Elias Munoz et al. “3D-Morphomics, Morphological Features on CT Scans for Lung Nodule Malignancy Diagnosis”. In: *Cancer Prevention Through Early Detection*. Ed. by Sharib Ali et al. Section: 0. Cham: Springer Nature Switzerland, 2022, pp. 3–13. ISBN: 978-3-031-17979-2. DOI: 10.48550/arXiv.2207.13830. URL: http://arxiv.org/abs/2207.13830.

[41] Muhammad Muzammil et al. “Pulmonary Nodule Classification Using Feature and Ensemble Learning-Based Fusion Techniques”. In: *IEEE Access* 9 (2021), pp. 113415–113427. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3102707. URL: https://ieeexplore.ieee.org/document/9507437/.

[42] Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. "A Review of Evaluation Metrics in Machine Learning Algorithms". In: *Artificial Intelligence Application in Networks and Systems*. Ed. by Radek Silhavy and Petr Silhavy. Cham: Springer International Publishing, 2023, pp. 15–25. ISBN: 978-3-031-35314-7.

[43] Syed Muhammad Naqi, Muhammad Sharif, and Mussarat Yasmin. "Multistage segmentation model and SVM-ensemble for precise lung nodule detection". In: *International Journal of Computer Assisted Radiology and Surgery* 13.7 (July 1, 2018), pp. 1083–1095. ISSN: 1861-6429. DOI: 10.1007/s11548-018-1715-9. URL: https://doi.org/10.1007/s11548-018-1715-9.

[44] National Lung Screening Trial Research Team. *Data from the National Lung Screening Trial (NLST)*. Published: Dataset. 2013. DOI: 10.7937/TCIA.HMQ8-J677. URL: https://doi.org/10.7937/TCIA.HMQ8-J677.

[45] National Lung Screening Trial Research Team et al. "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening". In: *The New England journal of medicine* 365.5 (Aug. 4, 2011), pp. 395–409. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/nejmoa1102873. URL: https://europepmc.org/articles/PMC4356534.

[46] Jing Ning et al. "Early diagnosis of lung cancer: which is the optimal choice?" In: *Aging (Albany NY)* 13.4 (Feb. 11, 2021). Place: United States, pp. 6214–6227. ISSN: 1945-4589. DOI: 10.18632/aging.202504. URL: http://dx.doi.org/10.18632/aging.202504.

[47] Ana Oprisan and Sorinel Adrian Oprisan. "Bounds for Haralick features in synthetic images with sinusoidal gradients". In: *Frontiers in Signal Processing* 3 (Nov. 23, 2023). Publisher: Frontiers. ISSN: 2673-8198. DOI: 10.3389/frsip.2023.1271769. URL: https://www.frontiersin.org/journals/signal-processing/articles/10.3389/frsip.2023.1271769/full (visited on 06/17/2025).

[48] M. Pietikäinen. "Local Binary Patterns". In: *Scholarpedia* 5.3 (2010), p. 9775. DOI: 10.4249/scholarpedia.9775.

[49] Matti Pietikäinen. "Image Analysis with Local Binary Patterns". In: *Image Analysis*. Ed. by Heikki Kalviainen, Jussi Parkkinen, and Arto Kaarna. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 115–118. ISBN: 978-3-540-31566-7.

[50] Eduardo M. Rodrigues et al. "Efficient-Proto-Caps: A Parameter-Efficient and Interpretable Capsule Network for Lung Nodule Characterization". In: *IEEE Access* 13 (2025), pp. 56616–56630. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3555428.

[51] Tanzila Saba et al. "Lung Nodule Detection based on Ensemble of Hand Crafted and Deep Features". In: *Journal of Medical Systems* 43.12 (Nov. 8, 2019), p. 332. ISSN: 1573-689X. DOI: 10.1007/s10916-019-1455-6. URL: https://doi.org/10.1007/s10916-019-1455-6.

[52] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017 IEEE International Conference on Computer Vision (ICCV). Oct. 22, 2017, pp. 618–626. ISBN: 978-1-5386-1033-6. DOI: 10.1109/ICCV.2017.74.

[53] Arnaud Arindra Adiyoso Setio et al. *LUNA16: A Challenge for Automatic Nodule Detection in Low-Dose CT Scans*. 2016. URL: https://luna16.grand-challenge.org/.

[54] Ahmed Shaffie et al. "A Generalized Deep Learning-Based Diagnostic System for Early Diagnosis of Various Types of Pulmonary Nodules". In: *Technol. Cancer Res. Treat.* 17 (Jan. 1, 2018). Publisher: SAGE Publications, p. 1533033818798800. ISSN: 1533-0346, 1533-0338. DOI: 10.1177/1533033818798800. URL: https://journals.sagepub.com/doi/10.1177/1533033818798800 (visited on 02/04/2025).

[55] Ahmed Shaffie et al. "Computer-assisted image processing system for early assessment of lung nodule malignancy". In: *Cancers (Basel)* 14.5 (Feb. 22, 2022). Publisher: MDPI AG, p. 1117. ISSN: 2072-6694. DOI: 10.3390/cancers14051117. URL: https://www.mdpi.com/2072-6694/14/5/1117.

[56] Lamia H. Shehab et al. "An efficient brain tumor image segmentation based on deep residual networks (ResNets)". In: *Journal of King Saud University - Engineering Sciences* 33.6 (Sept. 1, 2021), pp. 404–412. ISSN: 1018-3639. DOI: 10.1016/j.jksues.2020.06.001. URL: https://www.sciencedirect.com/science/article/pii/S1018363920302506.

[57] Mingxing Tan and Quoc V. Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. Sept. 11, 2020. DOI: 10.48550/arXiv.1905.11946. arXiv: 1905.11946[cs]. URL: http://arxiv.org/abs/1905.11946 (visited on 06/23/2025).

[58] William D. Travis. "Pathology of Lung Cancer". In: *Clinics in Chest Medicine* 32.4 (Dec. 1, 2011). Publisher: Elsevier, pp. 669–692. ISSN: 0272-5231. DOI: 10.1016/j.ccm.2011.08.005. URL: https://doi.org/10.1016/j.ccm.2011.08.005 (visited on 06/14/2025).

[59] United Nations. *THE 17 GOALS | Sustainable Development*. 2015. URL: https://sdgs.un.org/goals (visited on 10/06/2024).

[60] Joost J.M. Van Griethuysen et al. "Computational Radiomics System to Decode the Radiographic Phenotype". In: *Cancer Research* 77.21 (Nov. 1, 2017), e104–e107. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-17-0339. URL: https://aacrjournals.org/cancerres/article/77/21/e104/662617/Computational-Radiomics-System-to-Decode-the (visited on 06/16/2025).

[61] Yoh Watanabe. "TNM Classification for Lung Cancer". In: *TNM Classification for Lung Cancer* 9.6 (2003).

[62] World Health Organization. *The Top 10 Causes of Death*. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (visited on 10/03/2024).

[63] Yun-Ju Wu and Fu-Zong Wu. "AI-Enhanced CAD in Low-Dose CT: Balancing Accuracy, Efficiency, and Overdiagnosis in Lung Cancer Screening". In: *Thoracic Cancer* 16.1 (Nov. 2024). Place: Singapore Publisher: Department of Radiology, Kaohsiung Veterans General Hospital, Kaohsiung, Taiwan.; Department of Software Engineering and Management, National Kaohsiung Normal University, Kaohsiung, Taiwan.; Department of Radiology, Kaohsiung Veterans General Hospital, Kaohsiung, Taiwan., e15499. ISSN: 1759-7714 (Electronic). DOI: 10.1111/1759-7714.15499. URL: https://pubmed.ncbi.nlm.nih.gov/39600243/.

[64] Yutong Xie et al. "Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT". In: *Information Fusion* 42 (July 2018), pp. 102–110. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2017.10.005. URL: https://www.sciencedirect.com/science/article/pii/S1566253516301063.

[65] Haiying Yuan, Yanrui Wu, and Mengfan Dai. "Multi-Modal Feature Fusion-Based Multi-Branch Classification Network for Pulmonary Nodule Malignancy Suspiciousness Diagnosis". In: *Journal of Digital Imaging* 36.2 (Apr. 1, 2023), pp. 617–626. ISSN: 1618-727X. DOI: 10.1007/s10278-022-00747-z. URL: https://doi.org/10.1007/s10278-022-00747-z.

[66] Jumin Zhao et al. "Combining multi-scale feature fusion with multi-attribute grading, a CNN model for benign and malignant classification of pulmonary nodules". In: *Journal of Digital Imaging* 33.4 (Aug. 1, 2020), pp. 869–878. ISSN: 1618-727X. DOI: 10.1007/s10278-020-00333-1. URL: https://doi.org/10.1007/s10278-020-00333-1.

[67] Yue Zhao et al. "Pulmonary Nodule Detection Based on Multiscale Feature Fusion". In: *Computational and Mathematical Methods in Medicine* 2022.1 (Dec. 21, 2022). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/8903037, pp. 1–13. ISSN: 1748-6718, 1748-670X. DOI: https://doi.org/10.1155/2022/8903037. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/8903037.

[68] Bolei Zhou et al. "Learning Deep Features for Discriminative Localization". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2921–2929. DOI: 10.1109/CVPR.2016.319.

[69] Nitish Zulpe and Vrushsen Pawar. "GLCM Textural Features for Brain Tumor Classification". In: *International Journal of Computer Science Issues* 9.3 (2012).

Chapter 8

Best AUC and Stage for Fusion

Table 8.1: Best AUC and respective Stage for HOG Feature Fusion with ResNet-18 Across Learning Rates and Batch Sizes

Batch Size	**Learning Rate**					
	$\mathbf{10^{-3}}$		$\mathbf{10^{-4}}$		$\mathbf{10^{-5}}$	
	AUC	**Stage**	**AUC**	**Stage**	**AUC**	**Stage**
32	0.81 ± 0.02	1	0.81 ± 0.01	4	0.81 ± 0.02	1
64	0.81 ± 0.01	4	0.80 ± 0.01	3	0.80 ± 0.02	1
128	0.82 ± 0.01	1	0.81 ± 0.01	1	0.79 ± 0.01	1

Table 8.2: Best AUC and respective Stage for Gabor Feature Fusion with ResNet-18 Across Learning Rates and Batch Sizes

Batch Size	**Learning Rate**					
	$\mathbf{10^{-3}}$		$\mathbf{10^{-4}}$		$\mathbf{10^{-5}}$	
	AUC	**Stage**	**AUC**	**Stage**	**AUC**	**Stage**
32	0.81 ± 0.01	4	0.81 ± 0.01	3	0.80 ± 0.01	2
64	0.82 ± 0.01	1	0.81 ± 0.01	2	0.81 ± 0.01	2
128	0.82 ± 0.02	4	0.81 ± 0.01	1	0.80 ± 0.01	2

Table 8.3: Best AUC and respective Stage for Shape Feature Fusion with ResNet-18 Across Learning Rates and Batch Sizes

Batch Size	**Learning Rate**					
	$\mathbf{10^{-3}}$		$\mathbf{10^{-4}}$		$\mathbf{10^{-5}}$	
	AUC	**Stage**	**AUC**	**Stage**	**AUC**	**Stage**
32	0.84 ± 0.02	1	0.83 ± 0.01	2	0.81 ± 0.01	2
64	0.83 ± 0.02	1	0.82 ± 0.01	1	0.81 ± 0.01	2
128	0.82 ± 0.01	1	0.82 ± 0.01	2	0.81 ± 0.00	2

Table 8.4: Best AUC and respective Stage for Haralick Feature Fusion with ResNet-18 Across Learning Rates and Batch Sizes

Batch Size	**Learning Rate**					
	$\mathbf{10^{-3}}$		$\mathbf{10^{-4}}$		$\mathbf{10^{-5}}$	
	AUC	**Stage**	**AUC**	**Stage**	**AUC**	**Stage**
32	0.82 ± 0.02	1	0.81 ± 0.02	2	0.80 ± 0.01	2
64	0.81 ± 0.02	2	0.81 ± 0.01	2	0.78 ± 0.01	1
128	0.81 ± 0.01	3	0.82 ± 0.01	2	0.79 ± 0.01	2