2026-04

Deep Learning Framework for Automated Lung and Pulmonary Nodule Segmentation in Thoracic CT Scans

Yousuf, Md.

IUB

Deep Learning Framework for Automated Lung and Pulmonary Nodule Segmentation in Thoracic CT Scans

Prepared By

Md. Yousuf

ID: 2222005

Abbas Yasin

ID: 2220411

Muaaz Mohammad

ID: 2220401

Spring 2026

Department of Computer Science and Engineering
Independent University Bangladesh

Supervisor:

Md. Rashedur Rahman, D. Eng

Assistant Professor,
Department of Computer Science and Engineering
School of Engineering Technology and Science
Independent University Bangladesh

In Partial Fulfillment of the Requirements for the Degree of Bachelor's of Computer Science & Engineering

Attestation

We are aware of the fact that plagiarism is strictly prohibited and it conflicts with our university's rules and regulations. We guarantee the authenticity of our work. We have used some copyrighted material and models of others in our project which has been properly cited following the international standards proper guidelines.

Author Name:

Md. Yousuf

Signature:

. . . . . .

Author Name:

Abbas Yasin

Signature:

. . . . . .

Author Name:

Muaaz Mohammad

Signature:

. . . . . .

Evaluation Committee

Supervisor

Name: _____ Signature: _____

Internal Examiner 1

Name: _____ Signature: _____

Internal Examiner 2

Name: _____ Signature: _____

External Examiner

Name: _____ Signature: _____

Acknowledgement

In the name of Allah, the Most Merciful, the Most Compassionate.

We would like to express our heartfelt gratitude to our supervisor, Md. Rashedur Rahman, D. Eng., for his invaluable guidance throughout our thesis journey. His special direction in image preprocessing laid the foundation for this research. His patience, insightful feedback, and constant encouragement helped us overcome technical challenges and strengthened our confidence to complete this study successfully.

We also extend our sincere appreciation to the Department of Computer Science and Engineering, Independent University, Bangladesh for providing a supportive academic environment. We are grateful to be members of the Center for Computational and Data Sciences (CCDS) Lab, Independent University, Bangladesh, where we received continuous facilities, resources, and guidance throughout this research.

Special thanks to Ms. Halima Khatun for her valuable help with the literature review. We are also deeply thankful to Mr. Fatin Israq Tabib and Mr. Siam Tahsin Bhuiyan for their assistance in fixing coding errors and providing technical support during challenging phases of implementation. Our sincere gratitude goes to Mr. Sefatul Wasi for his overall lab guidance and continuous support.

We also acknowledge the assistance of ChatGPT from OpenAI, Claude Opus and Claude Haiku from Anthropic, Gemini from Google in helping us paraphrase sections and improve the readability of the manuscript. No data, images, or analyses were generated using these tools.

Finally, we gratefully acknowledge Independent University, Bangladesh for the opportunity to conduct this thesis, which enabled us to gain practical research experience and better prepare for the professional world.

Abstract

Lung cancer is one of the leading causes of highest mortality rate, with approximately 1.8 million deaths annually . While low-dose computed tomography (LDCT) screening significantly reduces mortality rates, manual radiological evaluation is very time consuming, find difficulty for inter-observer variability, and especially challenging for subtle nodules. To address these clinical challenges, this thesis presents a comprehensive deep learning framework for the automated simultaneous pixel-level delineation of bilateral lung regions and pulmonary nodules. This study presents a U-Net encoder-decoder with a ResNet-18 backbone, where residual connections handle hierarchical feature extraction and skip connections preserve spatial detail for precise boundary localisation. Before giving input to the model, CT volumes pass through Hounsfield unit windowing at center -600 HU and width 1500 HU to isolate parenchymal tissue, followed by geometric and intensity augmentation to expand the effective training distribution beyond the 89 available scans.We used the LUNA16 dataset 888 thoracic CT scans containing 1,186 radiologist-verified nodules. A 10-10-80 partition assigns Subset 0 to training, Subset 7 to validation, and Subsets 1-6 and 8-9 as eight independent test sets. Results across all eight partitions, rather than a single split, gives an honest picture of how much performance shifts when the patient group changes.Left lung segmentation reaches mean Dice 0.996 (IoU: 0.994); the right lung reaches at Dice 0.973 (IoU: 0.960). Small and morphologically variable structures have a lower nodule segmentation with a landing at Dice 0.848 (IoU: 0.741). The same story is told by the variance, which is less than 0.003 in the case of lung fields, less than 0.026 in the case of nodules. The lung gave result across patient groups; the nodule spread is real and reflects the genuine difficulty of the task rather than any modelling instability. Training converges stably by epoch 80 with limited gap between training and validation curves. The aggregate training log contains some duplicated epoch segments that limit precise epoch-level reading of the curves, though the per-partition evaluation numbers are entirely unaffected. With only 89 training scans, augmentation and ImageNet initialisation carry the 14.3-million-parameter network past the overfitting risk that a cold-start model at this data volume would face.Running all three classes through a single forward pass avoids the error accumulation and redundant computation that step-by-step pipelines introduce at each stage handoff. Three limitations shape what the current framework can and cannot claim. Training covers only 10% of available LUNA16 data. Slice-level 2D processing discards the inter-slice spatial context that nodule characterisation needs. While malignancy risk stratification remains outside the current scope, future work will explore 3D architectures, full dataset utilization, and attention mechanisms. These advancements, alongside multi-institutional validation, are essential for transitioning this framework into a real-world clinical environment.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Every year, lung cancer causes about 1.8 million deaths, which can be basically regarded as the highest death rates worldwide [1]. The condition arises due to uncontrolled growth of the cells in the respiratory tissues and can be classified into two general categories of the clinical conditions that includes non-small cell lung carcinoma (NSCLC) and small cell lung carcinoma (SCLC).A differentiation that influences the choice of treatment and prognostic expectation[2]. The essence of the issue is time. The symptoms remain silent until the disease has developed far beyond the limits of the early stages, most patients end up with a diagnosis that is already too late to implement the most effective interventions [1].

Early detection shifts this outcome considerably. The National Lung Screening Trial (NLST) shows that low-dose computed tomography (LDCT) screening brings lung cancer mortality down by 20% relative to conventional chest radiography in high-risk populations [3]. Since then, LDCT has become the go-to imaging modality for lung cancer screening in routine clinical practice [4]. The bottleneck, however, is interpretation. Manual reading of CT scans is time-consuming and prone to variability between readers. Radiologists detect only 30% to 80% of nodules depending on nodule size, density, and location [5] , a range wide enough to represent a genuine clinical problem.

Deep learning has changed the analytical capabilities available for medical image analysis. Convolutional neural networks learn hierarchical feature representations directly from pixel data, without requiring the hand-engineered features that traditional machine learning pipelines depend on [6]. Encoder-decoder architectures, particularly U-Net, have set performance benchmarks across a range of medical image segmentation tasks [7].

Automated lung cancer analysis pipelines typically begin with semantic segmentation of pulmonary structures. Accurate lung field delineation confines the search space for nodule detection to anatomically relevant regions, reducing false positives and unnecessary computation [8]. Precise nodule segmentation then enables volumetric measurement, longitudinal growth monitoring, and extraction of radiomic features for malignancy assessment [9].

1.2 Motivation

Several practical challenges in automated pulmonary nodule analysis motivated this research challenges that remain despite the advances described above.

The CT scan of the thoracic area in a clinical screening setting may take hundreds of slices and each slice should be thoroughly screened to identify the delicate nodular appearance. Research indicates that the average reading time per scan is usually 6-10 minutes, based on finding complexity and reader experience [10]. It is that time pressure which accumulates reader fatigue and increases diagnostic variability, creating a distinct space where automated assistance, with its reliability, would be of real clinical use [11]. The 30 to 80 percent range of the detection across readers [5] is not an indication of differences in skill levels; it indicates how difficult the task is in real working conditions. The most common findings that are missed are small nodules less than 5mm and ground-glass opacities, and the lack of such in a report has the most serious clinical implications[12].

The issue of quantitative reproducibility lies independent. The longitudinal tracking study is not a very reliable one since when the same nodule is measured by different readers at various times the figures do not come out as similar. Automated segmentation is a perfect fit with growth monitoring and evaluation of treatment response by offering a certain degree of consistency in measurements that are impossible to obtain structurally by the use of manual techniques.

Anatomical segmentation also plays a structural role within computer-aided detection pipelines. The first step of isolating the lung field reduces the search space of nodule detection algorithms and enhances accuracy and computational efficiency [13].Multi-class segmentation goes a step further to provide anatomical context - in which a nodule is located at the location of the lung directly impacts the process of differential diagnosis and changes the interpretation of a finding in the clinical context.

It is well-annotated public datasets, complete with standardized evaluation protocols that enable different research systems to have a common ground to be fairly compared [5]. The dataset used in this paper, the LUNA16 challenge data, based on LIDC-IDRI collection, occupies that role here, and the fact that the results of this study are made there makes the results of this study fall within an already established and widely-used performance landscape.

1.3 Scope of Study

This study builds and tests a deep learning framework for automated multi-class segmentation of thoracic CT images. The system focuses on three anatomical targets: left lung parenchyma, right lung parenchyma, and pulmonary nodules with a diameter of 3 mm or greater.

The LUNA16 challenge dataset forms the evaluation base 888 CT scans carrying 1,186 nodules with consensus agreement from three or more radiologists, all drawn from the LIDC-IDRI archive. Its standardised annotations and established benchmarking protocols make

reproducible evaluation and fair comparison with prior methods straightforward.

Preprocessing applies Hounsfield unit windowing at center $-600\,\mathrm{HU}$ and width $1500\,\mathrm{HU}$ to bring out parenchymal tissue contrast. Dice loss keeps class imbalance from skewing the training objective. Data augmentation stretches the effective training distribution well beyond what 89 raw scans alone would allow. Rather than reporting a single test-set figure, evaluation spreads across eight independent held-out partitions to give a variance-aware picture of generalization.

Intersection over Union (IoU), Dice coefficient, precision, and recall each get computed separately per anatomical class. Cross-partition consistency carries particular weight here, serving as the primary proxy for how reliably the model would generalize beyond its training distribution.

The study draws clear boundaries around what it does not cover: benign versus malignant nodule classification, volumetric 3D analysis, real-time clinical deployment, and multi-institutional external validation. These are not gaps left by oversight they reflect a deliberate choice to keep the research questions within a scope where honest, tractable claims are possible.

1.4 Research Questions

This thesis addresses the following research questions:

1. Is it possible for a unified encoder-decoder architecture to handle bilateral lung regions and pulmonary nodules at the same time, while still meeting clinically acceptable accuracy thresholds?

2. To what degree does a U-Net carrying a pretrained ResNet-18 encoder hold up for bilateral lung and pulmonary nodule segmentation under this dataset and training setup?

3. In what ways does the preprocessing pipeline covering Hounsfield windowing, intensity normalization, and data augmentation contribute to the overall multi-class segmentation framework?

4. Across eight independent held-out test partitions, how stable does the framework's performance stay when the patient group shifts?

5. Where do segmentation errors most commonly arise for small nodular structures, and to what extent do size and appearance characteristics drive changes in segmentation quality?

1.5 Objective

The primary objective is to build and test a deep learning framework that can simultaneously segment bilateral lung regions and pulmonary nodules from thoracic CT scans, with performance measured against standard segmentation benchmarks. The specific objectives span five areas: dataset preparation, architecture development, model training, performance evaluation, and clinical relevance assessment.

For dataset preparation, the work splits the LUNA16 dataset into a systematic 10-10-80 split Subset-0 used as training, Subset-7 used as validation and Subsets 1-6 and 8,9 used as held-out testing. The backbone of preprocessing is: Hounsfield windowing and intensity normalization. The training distribution is extended to counteract the limited dataset size by augmentation strategies.

To develop architecture, the work combines a U-Net encoder-decoder with a ResNet-18 backbone, configures a multi-class output layer to segment into three regions in a single pass, and selects a loss formulation based on Dice loss formulations, which is appropriate to the class-imbalanced nature of the task.

To train the model, the framework uses Adam optimization, Cosine annealing and 120 epochs to train the model, monitoring the convergence by loss, IoU, F1, and Dice at every checkpoint. The dynamics of training are subsequently investigated to learn how to optimize behavior and identify any generalization issues at an early stage.

To measure performance, the work measures the segmentation accuracy of each of the three anatomical classes, by using standard metrics. Evaluating on eight independent held-out partitions provides a variance sensitive view of consistency, as opposed to a point estimate. Error analysis further examines failure modes to trace out the real locations of where the model limits lie.

For clinical relevance assessment, the results go up against published benchmarks and get read through the lens of clinical acceptability. Strengths and limitations both get an honest account, and concrete directions for future work follow directly from what the analysis reveals.

1.6 Contributions

This thesis makes the following contributions to automated pulmonary image analysis.

A unified multi-class segmentation framework is developed that performs lung field and nodule segmentation within a single-stage architecture. This eliminates the need for sequential detection and segmentation pipelines, reducing computational overhead and simplifying the path toward clinical integration.

A domain-optimized preprocessing pipeline is implemented, incorporating Hounsfield unit windowing tailored for CT parenchymal visualization and a comprehensive augmentation strategy that compensates for limited training data and inherent class imbalance.

Evaluation across eight independent held-out test partitions provides mean performance estimates alongside variance quantification , a more informative characterization than single held-out test reporting. This approach gives an empirical bound on how much performance shifts across patient groups, which is relevant for any deployment assessment.

Category-specific performance analysis quantifies results separately for left lung, right lung, and nodule segmentation. The framework achieves near-perfect lung field delineation (Left Lung Dice 0.996 ± 0.003, Right Lung Dice 0.973 ± 0.019) alongside meaningful nodule segmentation capability (Nodule Dice 0.848 ± 0.026) despite training on only 10% of the available LUNA16 data.

Training dynamics analysis covers loss, F1, and IoU evolution over 120 epochs, characterizing convergence behavior, learning rate schedule effectiveness, and the evolution of the training-validation performance gap.

Full methodology documentation , covering preprocessing steps, architectural configurations, and hyperparameter selections , is provided to support reproducibility. Public release of code and trained weights is reserved for future work.

Lastly, the work places its results in a truthful manner as required by clinical translation: what the current system can be utilized to support, where it fails and what exactly must be improved before multi-institutional validation becomes a reality.

1.7 Thesis Organization

The thesis is organized into nine chapters.

Chapter 1 introduces the research problem by discussing the global clinical impact of lung cancer and the challenges of manual CT interpretation. It defines the motivation for the study, establishes the research scope and specific objectives, and outlines the key technical and clinical contributions.

Chapter 2 reviews the literature on automated pulmonary image analysis, progressing from traditional computer vision and classical machine learning to contemporary deep learning architectures. Research gaps motivating this study are identified through that progression.

Chapter 3 establishes the theoretical background needed to follow the proposed methodology. It covers deep learning fundamentals, convolutional neural networks, semantic segmentation, and the specific U-Net and ResNet architectural components used in this work.

Chapter 4 describes the LUNA16 dataset: acquisition protocols, radiologist annotation procedures, dataset statistics, and the rationale behind the 10-10-80 train-validation-test partitioning strategy.

Chapter 5 details the complete methodology , the domain-optimized preprocessing pipeline, the U-Net architecture with ResNet-18 backbone, the Dice loss formulation, the training protocol, and the multi-metric evaluation framework.

Chapter 6 presents experimental results and discussion. It covers training dynamics, category-specific segmentation performance across eight test partitions, cross-partition consistency statistics, qualitative assessment, and comparison with prior work.

Chapter 7 provides an honest assessment of the framework's limitations: training data volume, 2D slice-level processing constraints, dataset temporal and geographic factors, and the barriers that remain before clinical integration.

Chapter 8 outlines future work directions, including 3D volumetric architectural extensions, expanded training strategies, external validation studies, and malignancy risk stratification.

Chapter 9 concludes the thesis by summarising the key findings and contributions, reflecting on the broader clinical implications, and offering final remarks on the pathway toward clinical translation.

Chapter 2

Literature Review

This chapter describes the automated pulmonary image analysis starting with the primitive rule-based algorithms to classical machine learning and concluding with the deep learning architectures that characterize the state of the art. The object is not to enumerate all of the published methods, but what each of the generations of methods contributed, what each of them missed, and what the failures stimulated the following. The proposed framework is positioned at the end of that progression.

2.1 Traditional Computer Vision Approaches

Initial computational approaches to pulmonary nodule detection used manually-crafted feature descriptors with rule-based decision logic and standard classifiers. These methods determined the conceptual vocabulary of this field, but their performance was weak in different imaging conditions.

2.1.1 Intensity-Based Thresholding

The earliest automated lung field extraction techniques used the high intensity difference between air-filled parenchyma and mediastinal structure [14]. The results of the method and adaptive variants of thresholding used by Otsu were acceptable in controlled environments. Practically, the threshold assumptions of the world were revealed by pathological infiltrates, consolidations, and inter-scanner intensity variation [8].Morphological operations were used together with multi-threshold methods that used erosion, dilation, and closing to enhance the detection of the boundaries and to recover the small airway structures [8, 15]. The basic issue was that every new institution, scanner model or acquisition protocol had to be retuned and there was no general parameter set [8].

2.1.2 Region-Based Segmentation

Some of the thresholding limitations were overcome by region growing algorithms which incorporated spatial connectivity constraints. Seed-based growth algorithms extended boundaries out of a few chosen starting points by criteria of intensity homogeneity [14].The method was more effective in dealing with gradual changes in intensity rather than hard thresholds, though performance was highly susceptible to the selection of seed point and the homogeneity threshold of choice [16].

Steep gradient vector flow snake variant enhanced convergence on the concave edges that are typical of diseased parenchyma[17]. In spite of these theoretical progresses, the real world

version was identical: initialization sensitivity, large computational cost, and large manual parameter tuning, rendered routine clinical implementation impractical [8].

2.1.3 Handcrafted Feature Descriptors

Characterization activities on nodules changed focus to the creation of discriminative sets of features that could encode size, shape, texture and intensity characteristics which differentiate nodules and normal tissue. Geometric features were used to evaluate morphology based on volume, maximum diameter, sphericity and surface properties [18]. Texture characteristics represented spatial patterns with the help of grey-level co-occurrence matrices, local binary patterns, and Gabor filter responses [19]. The nodule-parenchyma interface was quantified using a radial gradient analysis, index of spiculation, and measurement of boundary roughness [18].

These descriptors were applied to particular datasets. It was more difficult to generalize to a variety of clinical situations, as the features that are highly discriminative on one scanner or patient group are often not on a different one[20].The high-dimensional feature spaces also posed an overfitting risk in cases where the amount of training data was not adequate, and feature selection was an additional step that was necessary and data-specific.

2.2 Classical Machine Learning Methods

The use of machine learning classifiers introduced more adaptable pattern recognition to the issue, although the manual feature engineering condition kept a lid on the performance and overall generalizability.

2.2.1 Support Vector Machines

Support Vector Machines with Radial basis function kernel also an attractive nodule classifier, renowned by their good performance in high-dimensional spaces and resistance to overfitting due to the maximum margin formulation [21].Combined morphological, intensity and texture features were reported to have classification accuracies of between 85% and 92% [19]. Weakness was observed on the basis of ground-glass opacities and small nodules where discriminative features showed slight variations and the boundary of decision became more difficult to learn reliably [22].

2.2.2 Random Forests

Bootstrap aggregation and randomization of features at every split further refined single-classifier methods by giving rise to ensembles which generalized more reliably than single decision trees [23]. These textural-based ensemble methods have shown almost 90%

classification accuracies in the case of pulmonary nodule detection [24].Although feature importance measures offer convenient interpretability of critical radiomic features [20], the computational complexity of large-scale ensembles is a viable limit to large feature sizes, which requires optimized randomized selectors to retain system efficiency [24].

2.2.3 Feature Selection and Dimensionality Reduction

The space of high-dimensional features needed to be actively managed to prevent the curse of dimensionality. Principal Component Analysis minimized the dimensions and retained the variance but the components that were obtained could not be interpreted clinically easily [25]. Forward selection of sequential feature selection and backwards elimination determined the best subsets but at meticulous computation expense [26].The more fundamental issue was that feature engineering needed domain knowledge and was dataset-specific. A feature set that generalized well across scanner types and patient populations was never achieved through manual design alone [20].

2.3 Deep Learning in Medical Imaging

Deep learning replaced manual feature engineering with end-to-end learned representations, fundamentally changing what was achievable in medical image analysis.

2.3.1 Convolutional Neural Network Fundamentals

Convolutional neural networks extract spatial feature hierarchies from raw pixel intensities using translation-invariant filters learned from data [6]. Local connectivity and weight sharing make the extraction computationally tractable while preserving sensitivity to spatial patterns. Max pooling and average pooling introduce translation invariance while reducing spatial dimensionality. Rectified Linear Unit activations address the vanishing gradient problem that prevented training of deep networks in earlier architectures [27]. Batch normalisation reduces internal covariate shift, allowing higher learning rates and faster convergence [28]. Residual skip connections, introduced in ResNet, allow gradient flow through very deep networks by providing identity shortcuts that bypass non-linear transformations [29].

2.3.2 Transfer Learning and Pretrained Models

Training deep networks from random initialisation requires large labelled datasets. Medical imaging rarely provides them. Transfer learning from ImageNet-pretrained models addresses this by starting from representations that already encode useful visual structure edges, textures, spatial patterns and fine-tuning toward the target task [30]. Fine-tuning VGG, ResNet, and DenseNet architectures yielded accuracy improvements of 5% to 15% compared to ran-

dom initialisation across a range of medical imaging tasks [31].

ResNet-18 and ResNet-50 gained particular adoption in medical imaging due to the stability of their residual building blocks in limited-data regimes [29]. Identity mapping shortcuts address the network degradation problem that affects plain deep architectures [32]. Transfer learning effectiveness depends on source-target domain similarity. Low-level features such as edges and textures transfer well across domains; high-level semantic features are more task-specific and require more substantial fine-tuning [31].

2.3.3 Object Detection Architectures

Region-based CNNs introduced selective search followed by classification, achieving breakthrough sensitivity in nodule detection [33]. Faster R-CNN improved efficiency by integrating region proposal into the network itself, enabling end-to-end training [34]. Sensitivity exceeding 90% at 4 false positives per scan was reported, surpassing prior computer-aided detection systems [35]. Single-shot detectors YOLO and SSD eliminated the separate proposal stage, achieving real-time processing speeds exceeding 45 frames per second at some cost to accuracy on small objects [36, 37].

2.3.4 Semantic Segmentation Architectures

Fully Convolutional Networks substituted fully connected layers with convolutional counterparts, and made dense pixel-wise predictions at any input size [38]. Skip connections between the encoder and decoder restored space information that was lost due to repeated downsampling, enhancing localisation of boundaries.

U-Net became the dominant architecture for medical image segmentation through a symmetric encoder-decoder design with skip connections at every resolution level [7]. The skip connections allow precise boundary detection by fusing high-resolution spatial detail from shallow encoder layers with semantic feature maps from deep layers. Paired with aggressive augmentation, U-Net trains effectively on the small datasets typical of medical imaging [39]. The encoder pathway is modular, supporting diverse feature extraction backbones including VGG, ResNet, and EfficientNet [40].

Several variants have extended the original design. Attention U-Net adds attention gates that suppress irrelevant feature activations while amplifying regions of interest [41]. U-Net++ introduces nested skip pathways and deep supervision to improve gradient flow through the network [40]. Three-dimensional U-Net processes complete volumetric data rather than individual slices, capturing spatial context across the axial dimension at the cost of substantially higher memory requirements.

2.4 Pulmonary Nodule Segmentation Methods

2.4.1 Classical Approaches

Early nodule segmentation methods used threshold-based region growing starting from manually placed seed points [42]. Vessel-attached nodules, pleural nodules, and ground-glass opacities consistently caused failures because their intensity boundaries overlap with adjacent structures [15, 43]. Active contour refinements with shape priors and multi-scale processing partially addressed these cases [44]. Graph-cut methods reformulated segmentation as energy minimization, improving robustness to initialization [45]. All of these approaches needed careful per-case parameter tuning and generalized poorly to the range of nodule morphologies seen in clinical practice [20].

2.4.2 Deep Learning Approaches

Deep learning architectures, specifically convolutional neural networks (CNNs), have substantially improved on traditional rule-based and classical machine learning techniques by learning end-to-end representations [6]. The application of a two-dimensional U-Net, implemented slice-by-slice on axial CT scans, has demonstrated significant efficacy in capturing the complex morphology of pathological structures. Baseline implementations on the LIDC-IDRI dataset have reported Dice similarity coefficients (DSC) exceeding 83% for pulmonary nodule boundary delineation [15]. It was limited by slice-level inconsistency without volumetric context, where an adjacent slice would occasionally give conflicting predictions of the boundary [20]. Fusion that used multi-view prediction by using axial, coronal, and sagittal prediction enhanced spatial consistency without a proportional increase in computation [35].These 2.5D schemes compromised effectiveness with the increased ability to deal with non-uniform geometries [46].

Full 3D convolution techniques provided full volumetric information and enhanced detection and segmentation results [35]. V-Net proposed volumetric blocks of residues to form the basic building block of 3D medical segmentation [47].Processing and reduction of patches were necessitated by memory constraints, and this posed the risk of small nodules being missed in patches that were between processing regions [20]. Mask R-CNN variants combining detection and segmentation in a unified end-to-end framework reached accuracies approaching 95%, eliminating error propagation between separate detection and segmentation stages [37, 48].

2.5 Loss Functions for Medical Image Segmentation

Class imbalance is a structural feature of medical segmentation. Pathological regions occupy small fractions of total image area; standard loss functions that weight every pixel equally optimize predominantly toward the background class. Several loss function designs address this.

2.5.1 Cross-Entropy Variants

Binary cross-entropy provides a baseline but is poorly suited to imbalanced class distributions [49]. Weighted cross-entropy assigns higher penalties to minority class errors, which helps, but requires careful weight selection that tends to be dataset-specific [7]. Focal loss down-weights well-classified examples automatically, concentrating learning on the difficult cases regardless of class frequency [50].

2.5.2 Dice-Based Losses

Dice loss directly optimizes the Dice coefficient the same metric used at evaluation by minimizing its complement:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \tag{2.5.1}$$

where p_i and g_i represent predicted and ground truth probabilities at pixel i, and ϵ ensures numerical stability [47]. The ratio formulation handles class imbalance naturally: a small nodule and a large lung field contribute equally to the total loss regardless of their relative pixel counts. The known weakness is optimisation instability early in training, when predictions are poor and the gradient landscape is flat [49].

2.5.3 Combined Losses

Hybrid loss functions combine cross-entropy and Dice loss to leverage their complementary properties: cross-entropy provides stable early-training gradients while Dice loss addresses class imbalance and aligns the objective with the evaluation metric .The standard formulation is:

$$\mathcal{L}_{\text{combined}} = \alpha\mathcal{L}_{\text{CE}} + (1-\alpha)\mathcal{L}_{\text{Dice}} \tag{2.5.2}$$

with α typically set between 0.3 and 0.7 depending on the degree of class imbalance.

2.6 Evaluation Metrics for Segmentation

No single metric captures all aspects of segmentation quality. Standard practice reports multiple complementary measures.

2.6.1 Overlap-Based Metrics

The Dice Similarity Coefficient measures region overlap between prediction and ground truth:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} \tag{2.6.1}$$

Dice ranges from 0 for no overlap to 1 for perfect agreement and remains the most widely reported metric in medical segmentation literature [51]. Intersection over Union applies a stricter penalty by including all unpredicted and over-predicted regions in the denominator:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{2.6.2}$$

Also known as the Jaccard index, IoU consistently yields lower values than Dice for equivalent overlap, making it a more conservative measure of boundary accuracy [52].

2.6.2 Distance-Based Metrics

Overlap metrics summarize agreement across the whole region. Distance metrics measure how far the predicted boundary deviates from the ground truth at specific locations. Hausdorff distance quantifies the worst-case boundary error:

$$H(P, G) = \max\{h(P, G),\ h(G, P)\} \tag{2.6.3}$$

where $h(P, G) = \max_{p \in P} \min_{g \in G} \|p - g\|$ is the directed Hausdorff distance. Hausdorff distance is sensitive to outlier predictions and complements overlap metrics by exposing localised boundary failures [53]. Average surface distance provides a more stable boundary assessment by measuring mean deviation rather than maximum, trading sensitivity to outliers for a more representative view of typical boundary accuracy.

2.6.3 Volumetric Metrics

Volume similarity measures overall size agreement independent of exact spatial localisation:

$$VS = 1 - \frac{||V_P| - |V_G||}{|V_P| + |V_G|} \tag{2.6.4}$$

This metric is useful for applications where size estimation matters more than precise boundary placement, such as volumetric nodule growth tracking.

2.7 The LIDC-IDRI Dataset

The Lung Image Database Consortium and Image Database Resource Initiative is the most extensively used public lung CT database for nodule detection research [5].

2.7.1 Data Acquisition

The full LIDC-IDRI database contains 1,018 thoracic CT examinations collected between 2004 and 2009 from seven academic institutions [5]. Imaging parameters varied across sites, reflecting realistic clinical diversity in scanner manufacturers, reconstruction algorithms, and acquisition protocols. That heterogeneity is deliberate it means models trained on LIDC-IDRI encounter the kind of variation they would face in multi-institutional deployment, rather than artificially uniform conditions.

2.7.2 Annotation Protocol

Each scan was reviewed independently by four experienced thoracic radiologists in a two-phase process [5]. In the blinded read phase, each radiologist reviewed the complete CT stack without knowledge of colleagues' findings, classifying lesions as nodule $\geq$3 mm, nodule $<$3 mm, or non-nodule. Nodules measuring 3 mm or greater received detailed freehand boundary annotations on each relevant slice, characterized across nine semantic attributes: subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture, and malignancy likelihood. In the unblinded review phase, radiologists could examine anonymous colleagues' annotations and revise their own. The final annotations represent a multi-reader consensus, providing a more stable reference standard than single-reader annotation.

2.7.3 LUNA16 Challenge

The LUNA16 grand challenge defined standardised evaluation protocols for nodule detection using a curated LIDC-IDRI subset [13]. The curated subset contains 888 scans with annotations confirmed by at least three of four radiologists, providing 1,186 nodules with diameter 3 mm or greater. Inclusion criteria were slice thickness of 2.5 mm or less, no gaps between slices, uniformity of slice spacing, and full thoracic coverage. The challenge offers an offi-

cial ten-fold subject partitioning protocol and patient level separation to eliminate the issue of data leakage, and an openly available leaderboard to monitor state-of-the-art results. In the current research, this is the partitioning structure, but with a fixed 10-10-80 split instead of rotating using all ten folds

2.8 Lung Segmentation: Specific Studies

The previous literature of segmenting lungs shows uniform trade-offs. Atlas-based techniques can be trained to score high on Dice (approximately 0.98), whereas learning-based techniques provide a more favorable speed-accuracy ratio in practice in a clinical context [8]. Segmentation robust to pathology is an unsolved problem: the segmentation performance drops significantly on pathological compared to normal lungs due to consolidation, collapse, or effusion of the intensity patterns that mark the lung boundaries in healthy tissue [54]. Cross-institutional generalisation is another similar issue in automated pulmonary analysis. According to Hofmanninger et al.[54] , when research datasets based on single institutions, including LIDC-IDRI, are used to train models, they can exhibit significant performance declines, up to 15 percent in Dice similarity, when assessed on a wide range of external clinical datasets with diverse pathologies. This observation is the direct prompt to the multi-partition assessment plan that the current work has, as it is imperative that the model is subjected to the morphological variety to be able to be robustly deployed in clinical practice.

2.9 Gap Analysis and Research Positioning

This thesis addresses several critical gaps in current literature. First, most pipelines treat lung extraction and nodule segmentation as sequential steps, which allows errors to propagate downstream. A unified multi-class model eliminates this by segmenting both targets in a single forward pass, reducing system complexity and error accumulation.

Second, evaluation methodologies often rely on single-set point estimates that fail to capture performance variability. This study utilizes multi-partition evaluation reporting mean and standard deviation across eight independent subsets to provide a more reliable characterization for clinical deployment.

This work ensures full reproducibility and transparency by documenting all preprocessing parameters and reporting nodule and lung results separately to provide an honest, uninflated account of current model capabilities.

Chapter 3

Preliminaries

This chapter establishes the theoretical and technical foundations essential for understanding the proposed methodology. Fundamental concepts in deep learning, convolutional neural networks, semantic segmentation, and the specific architectural components employed in the framework are introduced systematically.

3.1 Deep Learning Fundamentals

Deep learning encompasses machine learning methods based on artificial neural networks with multiple layers of representation learning [6]. Unlike traditional machine learning requiring manual feature engineering, deep networks automatically learn hierarchical feature representations directly from raw data through supervised training with gradient-based optimization.

3.1.1 Artificial Neural Networks

An artificial neural network consists of interconnected processing units organized in layers. The input layer receives raw data such as pixel intensities. Hidden layers apply learned transformations to extract progressively abstract features. The output layer produces task-specific predictions such as class probabilities. Each neuron computes a weighted sum of inputs followed by nonlinear activation:

$$y = \sigma\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{3.1.1}$$

where w_i represents learnable weights, x_i denotes inputs, b is the bias term, and σ is the activation function .

3.1.2 Backpropagation and Gradient Descent

Neural networks learn through iterative optimization using the backpropagation algorithm [55]. The forward pass computes predictions by propagating inputs through network layers. Loss calculation measures discrepancy between predictions and ground truth using a loss function. The backward pass computes gradients of loss with respect to all parameters using the chain rule. Parameter update adjusts weights in the direction that reduces loss:

$$w \leftarrow w - \eta \frac{\partial \mathcal{L}}{\partial w} \tag{3.1.2}$$

where η is the learning rate and $\mathcal{L}$ represents the loss function.

3.1.3 Activation Functions

Activation functions introduce nonlinearity, enabling networks to learn complex patterns. The Rectified Linear Unit is defined as:

$$\text{ReLU}(x) = \max(0, x) \tag{3.1.3}$$

ReLU addresses vanishing gradient problems and enables efficient computation, becoming the de facto standard for deep networks [27]. The sigmoid activation function is given by:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.1.4}$$

Sigmoid maps inputs to the $(0, 1)$ range, suitable for binary classification outputs but suffers from gradient saturation . The softmax function normalizes logits:

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{3.1.5}$$

Softmax normalizes logits into probability distribution over classes, standard for multi-class classification .

3.2 Convolutional Neural Networks

Convolutional Neural Networks represent specialized architectures designed for processing grid-structured data such as images [56].

3.2.1 Convolutional Layers

Convolutional layers apply learnable filters that slide across input spatial dimensions:

$$Y[i, j] = \sum_m \sum_n X[i + m, j + n] \cdot K[m, n] + b \tag{3.2.1}$$

where X represents input feature map, K denotes convolutional kernel, and b is bias . Key properties include local connectivity where each neuron connects only to small spatial region reducing parameters, parameter sharing where same weights are applied across all spatial locations enforcing translation invariance, and hierarchical features where stacked convolutions learn progressively complex patterns from edges to semantic concepts.

3.2.2 Pooling Layers

Pooling operations downsample feature maps, providing translation invariance and reducing computational burden. Max pooling is defined as:

$$Y[i,j] = \max_{(m,n)\in\mathcal{R}_{i,j}} X[m,n] \tag{3.2.2}$$

Max pooling selects maximum value within local region, preserving dominant activations [57]. Average pooling computes the mean:

$$Y[i,j] = \frac{1}{|\mathcal{R}_{i,j}|} \sum_{(m,n)\in\mathcal{R}_{i,j}} X[m,n] \tag{3.2.3}$$

Average pooling computes mean over region, providing smoother downsampling [58].

3.2.3 Batch Normalization

Batch normalization normalizes layer inputs across mini-batches, accelerating training and improving generalization [28]:

$$\hat{x} = \frac{x - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \tag{3.2.4}$$

$$y = \gamma\hat{x} + \beta \tag{3.2.5}$$

where $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}^2$ represent batch mean and variance, while γ and β are learnable scale and shift parameters [28]. Benefits include reduced internal covariate shift, enabled higher learning rates, regularization effect reducing overfitting, and reduced sensitivity to initialization.

3.3 Residual Learning

Deep neural networks historically suffered from degradation problems where accuracy saturated and degraded with increased depth [29]. Residual learning addresses this through skip connections enabling identity mappings.

3.3.1 Residual Blocks

A residual block learns residual function $\mathcal{F}(x)$ rather than desired mapping $\mathcal{H}(x)$:

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{3.3.1}$$

where x represents input, $\mathcal{F}$ denotes residual mapping implemented by stacked layers, and addition performs element-wise sum [29]. The identity shortcut enables unimpeded gradient flow, facilitating training of networks exceeding 100 layers. If optimal mapping is close to

identity, learning small residuals proves easier than learning full transformation [29].

3.3.2 ResNet Architecture

ResNet employs residual blocks as building units, achieving breakthrough performance across computer vision tasks [29]. ResNet variants including ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152 differ in depth, with ResNet-18 comprising an initial 7×7 convolution with stride 2, a max pooling layer, four residual stages with progressively increasing channels at 64, 128, 256, and 512 dimensions, global average pooling, and a fully connected classification layer. Each residual stage contains multiple residual blocks with 3×3 convolutions. Spatial resolution reduces through stride-2 convolutions while channel dimensions increase, capturing semantic information while maintaining computational efficiency [29].

3.4 Semantic Segmentation

Semantic segmentation assigns class labels to every pixel, producing dense predictions suitable for medical image analysis [38].

3.4.1 Fully Convolutional Networks

Fully Convolutional Networks replaced fully connected layers with convolutions, enabling arbitrary input sizes and spatial output predictions [38]. The encoder performs downsampling through successive convolutions and pooling to extract semantic features while reducing spatial resolution. The decoder performs upsampling through transposed convolutions or upsampling operations to restore spatial resolution while incorporating learned features. Skip connections provide direct connections between encoder and decoder layers of equal resolution, combining coarse semantic information with fine spatial details and improving boundary localization [38].

3.4.2 Transposed Convolution

Transposed convolution performs learnable upsampling through backward strided convolution:

$$Y = \text{stride} \times (X - 1) + K - 2P \tag{3.4.1}$$

where X represents input size, K denotes kernel size, P is padding, and stride controls expansion factor [59]. Unlike fixed interpolation methods including bilinear and nearest-neighbor approaches, transposed convolutions learn optimal upsampling during training, capturing task-specific spatial reconstructions [38].

3.5 U-Net Architecture

U-Net represents the most influential architecture for medical image segmentation, featuring symmetric encoder-decoder structure with skip connections at each resolution level [7].

3.5.1 Architectural Overview

The contracting path serves as the encoder, performing repeated application of two 3×3 convolutions each followed by ReLU activation, 2×2 max pooling with stride 2 for downsampling, and doubling of feature channels at each downsampling step. The expanding path serves as the decoder, performing upsampling of feature map using 2×2 transposed convolution, concatenation with correspondingly cropped feature map from contracting path through skip connection, and two 3×3 convolutions each followed by ReLU. The final layer applies 1×1 convolution to map feature vectors to desired number of classes. Skip connections at each resolution enable precise localization by fusing semantic features from deep layers with spatial details from shallow layers [7].

3.5.2 Skip Connections

Skip connections constitute the defining characteristic of U-Net, enabling gradient flow and information propagation throughout the network. Direct pathways from encoder to decoder preserve spatial information lost during downsampling, multi-scale feature fusion combines low-level details with high-level semantics, improved gradient flow facilitates training of deep segmentation networks, and boundary localization enhancement enables precise delineation through fusion of fine-grained spatial information with semantic understanding [7].

3.5.3 Data Augmentation Strategy

U-Net design philosophy emphasizes aggressive data augmentation to enable training on small medical imaging datasets. Geometric transformations include rotation, scaling, elastic deformations, and flipping. Intensity perturbations include brightness adjustment, contrast modification, and Gaussian noise addition. These augmentations provide regularization effect preventing overfitting, simulate realistic variations encountered during inference, and improve model robustness to diverse imaging conditions [7].

3.6 Loss Functions for Medical Segmentation

3.6.1 Binary Cross-Entropy

Binary cross-entropy measures pixel-wise classification error:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{3.6.1}$$

where y_i represents ground truth, $\hat{y}_i$ denotes predicted probability, and N is total pixels .

3.6.2 Weighted Cross-Entropy

Class imbalance necessitates weighted formulation:

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N}\sum_{i=1}^{N}[w_{\text{pos}} y_i \log(\hat{y}_i) + w_{\text{neg}}(1-y_i)\log(1-\hat{y}_i)] \tag{3.6.2}$$

where w_{pos} and w_{neg} provide class-specific penalties addressing imbalanced datasets [7].

3.6.3 Dice Loss

Dice loss directly optimizes overlap metric:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N}\hat{y}_i y_i + \epsilon}{\sum_{i=1}^{N}\hat{y}_i + \sum_{i=1}^{N} y_i + \epsilon} \tag{3.6.3}$$

where ϵ provides numerical stability preventing division by zero [47]. Advantages include inherent handling of class imbalance through ratio formulation, direct optimization of evaluation metric used in assessment, and demonstrated effectiveness for small object segmentation in medical imaging applications.

3.6.4 Combined Loss

Hybrid formulations leverage complementary strengths:

$$\mathcal{L}_{\text{combined}} = \alpha\mathcal{L}_{\text{CE}} + (1-\alpha)\mathcal{L}_{\text{Dice}} \tag{3.6.4}$$

where cross-entropy provides stable pixel-level gradients while Dice loss addresses class imbalance, with α typically ranging from 0.3 to 0.7 in medical segmentation applications [60].

3.7 Optimization Algorithms

3.7.1 Stochastic Gradient Descent

Stochastic Gradient Descent with momentum accelerates convergence:

$$v_{t+1} = \mu v_t - \eta \nabla_\theta \mathcal{L}(\theta_t) \tag{3.7.1}$$

$$\theta_{t+1} = \theta_t + v_{t+1} \tag{3.7.2}$$

where μ is the momentum coefficient, η is the learning rate, and v represents velocity. This approach accelerates convergence and dampens oscillations, enabling the optimizer to escape shallow local minima.

3.7.2 Adam Optimizer

Adaptive Moment Estimation combines momentum with adaptive learning rates [61]. The first moment estimate computes the mean:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{3.7.3}$$

The second moment estimate computes the variance:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{3.7.4}$$

Bias correction accounts for initialization at zero:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{3.7.5}$$

Parameter update combines corrected moments:

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{3.7.6}$$

where $g_t = \nabla_\theta \mathcal{L}(\theta_t)$ represents gradient, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ are default hyperparameters [61]. Adam adapts learning rates per parameter based on gradient history, requiring minimal hyperparameter tuning and demonstrating robust convergence across diverse tasks [61].

3.8 Learning Rate Scheduling

3.8.1 Step Decay

Step decay reduces learning rate by factor at predetermined epochs:

$$\eta_t = \eta_0 \cdot \gamma^{\lfloor t/s \rfloor} \tag{3.8.1}$$

where η_0 is initial rate, γ such as 0.1 represents decay factor, and s denotes step size .

3.8.2 Cosine Annealing

Cosine annealing smoothly reduces learning rate following cosine function:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos\left(\frac{t\pi}{T}\right)\right) \tag{3.8.2}$$

where t represents current epoch, T denotes total epochs, and $\eta_{\min}$, $\eta_{\max}$ define rate bounds [62]. Benefits include smooth annealing avoiding abrupt changes, encouragement of exploration early while ensuring convergence late, and elimination of discrete schedule design requirements.

3.8.3 ReduceLROnPlateau

Adaptive scheduling monitors validation metrics, reducing learning rate when improvement plateaus:

$$\eta_{t+1} = \begin{cases} \gamma\eta_t & \text{if no improvement for } p \text{ epochs} \\ \eta_t & \text{otherwise} \end{cases} \tag{3.8.3}$$

where p represents patience parameter and γ is reduction factor [63].

3.9 Regularization Techniques

3.9.1 Dropout

Dropout randomly deactivates neurons during training with probability p, preventing co-adaptation:

$$h' = m \odot h \tag{3.9.1}$$

where $m \sim \text{Bernoulli}(1-p)$ represents binary mask and $\odot$ denotes element-wise product [64]. At inference, outputs are scaled by $(1-p)$ to maintain expected activations. Dropout provides implicit ensemble learning by training exponentially many thinned networks [64].

3.9.2 L2 Regularization

Weight decay penalizes large parameters, encouraging simpler models:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{task}} + \lambda \sum_i w_i^2 \tag{3.9.2}$$

where λ controls regularization strength .

3.9.3 Early Stopping

Training terminates when validation performance stops improving for predetermined patience period, preventing overfitting to training data [63].

3.10 Medical Image Preprocessing

3.10.1 Hounsfield Units

CT images encode radiodensity in Hounsfield Units, a standardized scale:

$$\text{HU} = 1000 \times \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}} - \mu_{\text{air}}} \tag{3.10.1}$$

where μ represents linear attenuation coefficient [65]. Key reference values include air at -1000 HU, lung parenchyma ranging from -900 to -500 HU, water at 0 HU, soft tissue from 40 to 80 HU, and bone from 700 to 3000 HU.

3.10.2 Windowing

Windowing maps HU range to display intensities, emphasizing specific tissues:

$$I_{\text{display}} = \begin{cases} 0 & \text{if HU} < \text{Center} - \text{Width}/2 \\ \frac{\text{HU} - (\text{Center} - \text{Width}/2)}{\text{Width}} \times 255 & \text{if Center} - \text{Width}/2 \leq \text{HU} \leq \text{Center} + \text{Width}/2 \\ 255 & \text{if HU} > \text{Center} + \text{Width}/2 \end{cases} \tag{3.10.2}$$

Standard lung window with center at -600 HU and width of 1500 HU optimizes parenchymal visualization [5].

3.10.3 Normalization

Intensity normalization ensures consistent input distributions:

$$X_{\text{norm}} = \frac{X - \mu}{\sigma} \tag{3.10.3}$$

where μ and σ represent mean and standard deviation computed across training set [28].

3.11 Evaluation Metrics

3.11.1 Confusion Matrix Elements

Segmentation evaluation utilizes a pixel-wise confusion matrix where True Positives (TP) and Negatives (TN) denote correct predictions, while False Positives (FP) and Negatives (FN) represent incorrect foreground and background assignments.

3.11.2 Derived Metrics

Precision measures proportion of predicted foreground that is correct:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.11.1}$$

This metric is also known as Positive Predictive Value [66]. Recall quantifies proportion of actual foreground correctly detected:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.11.2}$$

Recall is also known as Sensitivity or True Positive Rate [66]. F1 Score provides harmonic mean balancing precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{3.11.3}$$

This formulation balances both metrics [66]. Intersection over Union provides strict overlap metric:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{3.11.4}$$

Also known as Jaccard Index, IoU penalizes both over and under-segmentation [52]. Dice Similarity Coefficient is defined as:

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{3.11.5}$$

Chapter 4

Dataset

This chapter provides comprehensive description of the LUNA16 dataset utilized in this research, including acquisition protocols, annotation procedures, dataset characteristics, and our specific partitioning strategy.

4.1 LIDC-IDRI Database Overview

Image Database Resource Initiative (IDRI) and Lung Image Database Consortium (LIDC) is the largest publicly available lung CT database to perform nodule detection studies[5]. The seven academic medical centers and eight medical imaging companies formed the consortium and developed standardized annotation protocols and a large set of thoracic CT scans.

4.1.1 Data Acquisition

The entire LIDC-IDRI database is comprised of 1,018 thoracic CT images that were gathered in 2004-2009 across the participating institutions[5]. Important aspects of acquisition are patient demographics with cases representing various ages, smoking patterns and clinical manifestations that involve screening, diagnostic and follow-up scans. Scanner diversity can be achieved by having different manufacturers such as GE, Siemens, Philips, and Toshiba with a variety of models that can reflect clinical variability. The slice thickness is 0.6-5.0mm depending on the protocols used, but most protocols use sub-millimeter resolution. There are some reconstruction algorithms that have different kernels, including standard, bone, and lung in different institutions.

This heterogeneity ensures trained models encounter realistic clinical variation rather than artificially homogeneous data [5].

4.1.2 Annotation Protocol

The LIDC-IDRI used two-phase rigorous annotation of four experienced thoracic radiologists per scan [5].

Phase 1 entailed blinded independent reading in which each radiologist individually read the entire CT stack and indicated lesions which were deemed as nodule greater than or equal to 3mm, nodule less than 3mm, and non-nodule. Nodules with a diameter of 3mm or larger were annotated in detail with boundary markers by freehand outline on each slice of interest, and described in terms of 9 semantic features such as subtlety, internal architecture, calcification, sphericity, margin, lobulation, spiculation, texture and malignancy potential. Phase

2 involved unblinded review during which radiologists were shown the annotations made by anonymous colleagues and were allowed to modify their own marks in accordance with the collective discovery, and the final annotations were a consensus of the experienced readers.

This protocol provides inter-observer agreement ground truth in annotation with protocol-made reference standards that are stronger than single-reader references[5].

4.2 LUNA16 Challenge Dataset

The Lung Nodule Analysis 2016 (LUNA16) grand challenge provided a standard evaluation framework of nodule detection algorithms on curated LIDC-IDRI subset[13].

4.2.1 Dataset Curation

LUNA16 organizers used systematic inclusion criteria to form uniform evaluation set. The minimum scan distance of 2.5mm and less ensured sufficient spatial resolution. It was ensured that consistency was taken by eliminating scans where there were missing slices or irregular spacing between slices. Full thoracic coverage was needed in its entirety without truncation.

Subset obtained includes 888 CT scans that satisfy these quality criteria [13].

4.2.2 Nodule Annotations

LUNA16 targets nodules that satisfy the consensus criteria. Nodules should have a diameter of at least 3mm and smaller lesions will be excluded because they are difficult to detect and are uncertain in clinical meaning. The nodules have to be annotated by at least 3 of 4 radiologists because the threshold of the consensus is used to guarantee credible annotations. The data set includes 1,186 nodules with centroid coordinates and the approximate diameter.

4.2.3 False Positive Reduction Track

Besides nodule annotations, LUNA16 also contains 551,065 candidate locations created by various detection algorithms. This candidate set allows reducing research in false positives, and differentiating between true nodules and mimics such as vascular structures (particularly bifurcations that seem nodules), lymph nodes, granulomas and benign calcifications, and image artifacts and reconstruction artifacts.

4.2.4 Standardized Subject Partitioning

LUNA16 offers an 10-fold subject partitioning scheme with patient level splits to avoid slice level data leakage, equal distribution of nodules across folds, and a public leaderboard of

state of art performance. Our fixed partitioning 10-10-80 allocation is based on this partitioning structure as outlined in Section 4.4.

4.3 Dataset Statistics

4.3.1 Volumetric Characteristics

Table 4.3.1: LUNA16 Scan Characteristics

Characteristic	**Mean ± Std**	**Range**
Slice Thickness (mm)	1.28 ± 0.65	0.45 – 2.50
In-plane Resolution (mm)	0.70 ± 0.12	0.46 – 0.98
Number of Slices	278 ± 115	84 – 721
Volume Size (voxels)	$512 \times 512 \times 278$ (typical)	Variable

4.3.2 Nodule Distribution

Table 4.3.2: Nodule Size Distribution

Diameter Range	**Count**	**Percentage**
3-5 mm	487	41.1%
5-10 mm	512	43.2%
10-20 mm	159	13.4%
> 20 mm	28	2.4%
Total	**1,186**	**100%**

The distribution also exhibits an imbalance of classes and most of the nodules are small, prompting the use of specialized loss functions (Dice loss) to deal with such behavior.

4.3.3 Anatomical Distribution

According to LIDC-IDRI literature, nodules are said to be upper-lobe predilectory (because of high rates of smoking-related malignancy), and peripheral subpleural (which are common sites of adenocarcinoma) [5]. Lobe-wise nodule distribution was not independently computed from the specific LUNA16 subset used in this study; the statements above reflect characteristics described in the source dataset publications rather than figures derived from our own analysis.

4.4 Our Dataset Partitioning Strategy

We adopted a fixed 10-10-80 split rather than rotating all ten folds, prioritising training efficiency while enabling broad performance characterisation across multiple held-out partitions:

4.4.1 Partition Allocation

The training set consists of Subset 0 representing 10% of data or approximately 89 scans. The validation set consists of Subset 7 representing 10% of data or approximately 89 scans. The test sets consist of Subsets 1-6 and 8-9 representing 80% of data or approximately 710 scans across 8 independent partitions.

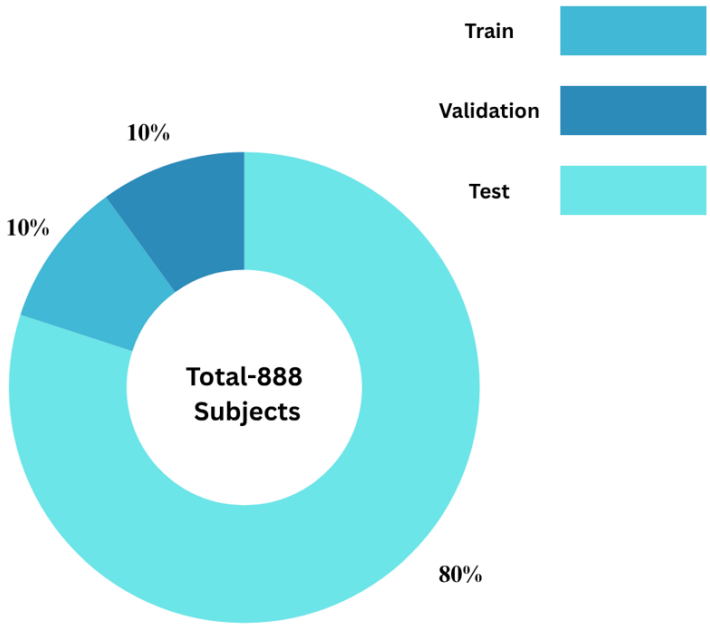

Figure 4.4.1: Dataset partition distribution: 80% test, 10% validation, 10% training following patient-level splitting to prevent data leakage.

4.4.2 Rationale

The aim of this allocation strategy is:

Computational Efficiency: 10 percent training allows quickly trying out preprocessing, architecture, and hyperparameter choices. Days of LIDC-IDRI training on our hardware is cut to hours on this subset.

Robust Evaluation: Eight independent held-out test partitions (not single hold-outs) allow a full quantification of variance on a variety of patient cohorts. We present mean performance

and standard deviations and give true evaluation of the consistency of generalization across the subsets tested.

Validation Integrity: Separate validation subset (Subset 7) to select the model and to early stop the process of learning avoids contaminating the test set. Decisions made by hyperparameters using only the validation performance make the test results indicative of the actual generalization ability.

Clinical Realism: The size of the training scans (89) appears small; however, in clinical contexts, similarly sized data is often involved.

4.4.3 Limitations

We acknowledge this partitioning introduces constraints. Training data scarcity exists as the limited training set may underutilize model capacity and prevent comprehensive architecture exploration. Statistical power is reduced as the smaller training set reduces statistical confidence in identifying optimal configurations. Rare pattern coverage may be limited as uncommon nodule morphologies may be underrepresented in the 10% sample.

These limitations motivate future work with expanded training incorporating additional LIDC-IDRI subsets and potential multi-institutional data sources.

4.5 Data Preprocessing and Augmentation

4.5.1 Slice Extraction

From volumetric CT stacks, we extract 2D axial slices for training. Specifically, for each annotated nodule in the training and validation subsets, the axial slice at the nodule centroid depth was extracted together with its corresponding pixel-level segmentation mask. Additional background slices were sampled from axial positions containing no nodule annotation within the same scan, in order to balance the class distribution and prevent the model from being trained solely on nodule-containing frames. This combination of nodule-centered and background slices resulted in approximately 2,500 training images and 2,500 validation images.

4.5.2 Multi-Class Mask Generation

For each slice, we generate pixel-level segmentation masks encoding four classes. Background class 0 represents non-anatomical regions including air and mediastinum. Left Lung class 3 represents left pulmonary parenchyma. Right Lung class 4 represents right pulmonary parenchyma. Nodule class 5 represents annotated nodular regions.

Lung field masks derived through automatic segmentation followed by manual refinement

ensuring anatomical accuracy. Nodule masks directly utilize radiologist boundary annotations from LIDC-IDRI.

4.5.3 Quality Control

The manual review of all mask annotations was done to verify anatomical accuracy of lung boundaries, nodule boundary delineation, slice-to-slice consistency and elimination of artifacts and mislabeled areas.

4.6 Dataset Availability and Ethics

4.6.1 Public Accessibility

Experiments in this work were performed with the publicly available LUNA16 dataset which is officially published on Zenodo in two parts [67, 68]. LUNA16 is based on LIDC-IDRI database offered by The Cancer Imaging Archive (TCIA) and it is commonly used in pulmonary nodule detection and analysis in CT of the chest studies.

4.6.2 Ethical Considerations

The use of dataset is in line with laid down ethical frameworks. Informed consent was given by all LIDC-IDRI participants to use research. The data is completely de-identified of any safeguarded health information. Original data collectors received an IRB approval. Our study is a secondary analysis of publicly available data that does not need any further ethical consideration.

4.7 Dataset Limitations

We acknowledge several dataset limitations affecting generalizability:

4.7.1 Temporal Constraints

Data collected 2004-2009 predates contemporary CT protocols including iterative reconstruction algorithms reducing noise, dual-energy CT providing material decomposition, and ultra-low-dose protocols enabled by advanced reconstruction.

4.7.2 Demographic Representation

Dataset demographics may not fully represent global population. There is geographic concentration in North America, potential age, ethnicity, and risk factor imbalances, and variable screening versus diagnostic versus surveillance indication distribution.

Chapter 5

Methodology

This chapter describes the technical framework behind the proposed pulmonary segmentation pipeline. The four components preprocessing, network architecture, training protocol, and evaluation are presented in enough detail to support full reproducibility.

5.1 Overview of Proposed Framework

The pipeline begins with preprocessing, which converts raw DICOM CT volumes into standardized neural network inputs through Hounsfield unit windowing, intensity normalization, and stochastic augmentation. The network itself is a U-Net encoder-decoder with a ResNet-18 backbone, trained to simultaneously classify pixels into three anatomical categories: left lung, right lung, and nodule. Training uses Dice loss with Adam optimization and cosine annealing scheduling. Evaluation then runs across eight independent held-out test partitions, reporting per-class metrics with statistical characterization of variance.

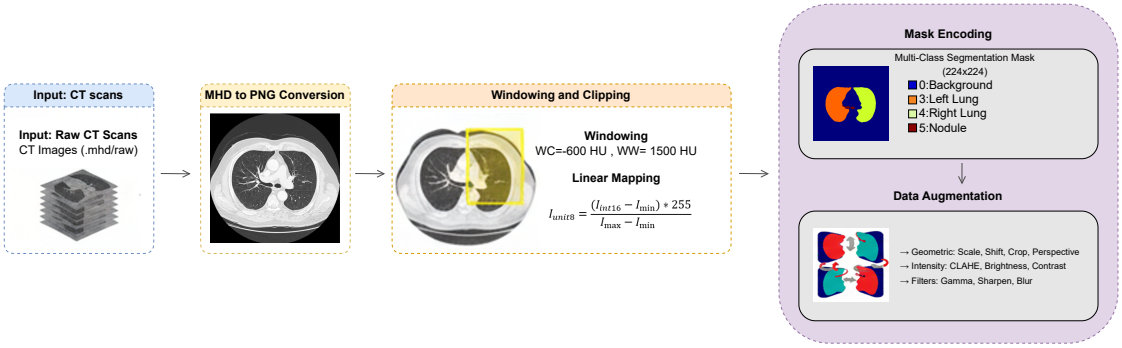

Figure 5.1.1: Complete preprocessing pipeline transforming raw CT volumes into model-ready inputs with domain-optimized windowing and augmentation.

5.2 Image Preprocessing Pipeline

Raw LUNA16 volumes cannot be fed directly into a neural network. Several transformations are necessary to standardise the input, suppress irrelevant tissue signal, and create training variation that the model will not have seen before.

5.2.1 Hounsfield Unit Windowing

CT images encode tissue radiodensity in Hounsfield Units, covering a range from roughly -1000 HU for air to $+3000$ HU for dense bone. That full range far exceeds what 8-bit representation can express, and most of it carries no useful information for pulmonary analysis.

Windowing restricts the visible range to the tissue of interest.

Standard lung window parameters are used here: window center at -600 HU, representing the midpoint of the parenchymal range, and window width of 1500 HU, spanning -1350 to $+150$ HU. The mapping is:

$$I_{\text{windowed}} = \begin{cases} 0 & \text{if HU} < W_{\text{center}} - W_{\text{width}}/2 \\ \dfrac{\text{HU} - (W_{\text{center}} - W_{\text{width}}/2)}{W_{\text{width}}} & \text{if } W_{\text{center}} - W_{\text{width}}/2 \leq \text{HU} \leq W_{\text{center}} + W_{\text{width}}/2 \\ 1 & \text{if HU} > W_{\text{center}} + W_{\text{width}}/2 \end{cases} \tag{5.2.1}$$

This isolates parenchymal contrast, improves nodule visibility against surrounding tissue, and suppresses mediastinal and bony structures that would otherwise dominate gradient updates during training.

5.2.2 Spatial Normalisation

All images are resampled to 224×224 resolution to match the ResNet-18 input requirement. Bilinear interpolation is applied to image slices; nearest-neighbor interpolation is used for segmentation masks to preserve discrete class labels without introducing fractional values at boundaries. Where the original aspect ratio differs from square, zero-padding is applied before resizing.

5.2.3 Intensity Normalization

After windowing, pixel intensities are normalized to zero mean and unit variance:

$$X_{\text{norm}} = \frac{X_{\text{windowed}} - \mu}{\sigma} \tag{5.2.2}$$

where μ and σ are computed across the entire training set. This step standardizes the input distribution across scans, which stabilizes gradient descent and prevents a few high-intensity outlier slices from dominating early weight updates.

5.2.4 Channel Replication

CT images are inherently single-channel. ResNet-18, pretrained on ImageNet, expects three-channel RGB input. The grayscale channel is therefore replicated three times:

$$X_{\text{RGB}} = \text{concat}(X_{\text{norm}},\ X_{\text{norm}},\ X_{\text{norm}}) \tag{5.2.3}$$

This is a simple operation, but it is what makes transfer learning possible. The pretrained weights were trained on color images; feeding identical values across all three channels allows the encoder to apply those learned representations to CT data without requiring architectural modification.

5.2.5 Data Augmentation

With only 89 training scans, overfitting is the primary risk. Aggressive augmentation expands the effective training distribution by generating modified versions of each input on-the-fly during training. Augmentations are applied jointly to image and mask to maintain label correspondence.

Geometric Transformations

Random scaling uses scale factors between 0.9 and 1.1, simulating variation in patient size and scanner zoom. Random translation shifts the image horizontally and vertically by up to $\pm 10\%$ of the image dimensions, accounting for variable patient positioning within the scanner bore. Random rotation applies angles up to ± 15 degrees, ensuring the model is not sensitive to small changes in scan orientation. Random cropping applies crop scales from 0.85 to 1.0 of the original size with aspect ratio preservation, training the model to handle partially visible anatomy. Random perspective distortion uses distortion scales of ± 0.05, which approximates subtle projection effects from minor variation in scan angulation.

Intensity Transformations

Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied with clip limit between 2.0 and 4.0 on an 8×8 grid, improving local contrast in homogeneous parenchymal regions. Random brightness applies multiplicative factors from 0.8 to 1.2, simulating variation in scanner output calibration. Random contrast uses the same factor range to address reconstruction algorithm differences across scanners. Random gamma correction applies gamma values from 0.8 to 1.2, providing nonlinear intensity mapping that approximates display calibration variation. Random sharpening with factors from 0.1 to 0.3 compensates for varying reconstruction kernel sharpness across acquisition protocols. Random Gaussian blur applies kernel sizes from 3 to 7 pixels, simulating partial volume effects and motion blur. Hue and saturation augmentations are excluded they have no meaningful counterpart in grayscale CT imaging.

Augmentation Probability

Augmentations are applied stochastically rather than deterministically. Geometric transformations carry a 50% probability each; intensity transformations carry 30% each. This means most training samples receive some augmentation, but not all transformations simultaneously

preserving enough structural integrity that the model can still extract clinically relevant features.

5.3 Network Architecture

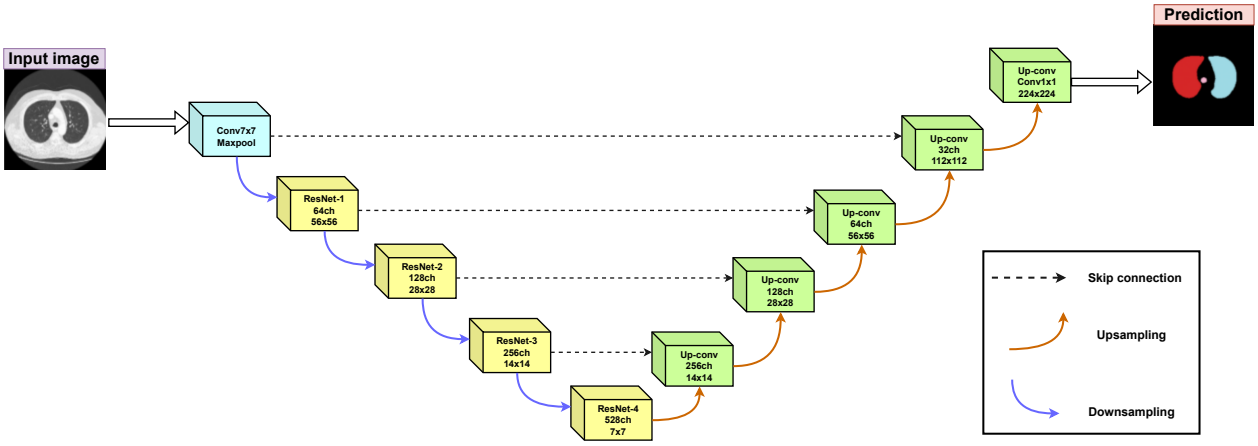

Figure 5.3.1: Proposed U-Net architecture with ResNet-18 encoder. Skip connections preserve spatial information while residual blocks enable deep feature extraction.

The segmentation model uses a U-Net encoder-decoder topology with a ResNet-18 encoder replacing the standard convolutional blocks. U-Net was chosen for its established performance in medical image segmentation; ResNet-18 was selected as the encoder because its pretrained weights provide a strong initialization and its residual connections maintain stable gradient flow at modest parameter count.

5.3.1 Encoder Pathway (ResNet-18)

The encoder extracts a hierarchy of features at progressively coarser spatial scales. The input block receives $224 \times 224 \times 3$ CT slices, applies a 7×7 convolution with stride 2 to produce 64 feature channels, followed by batch normalization, ReLU activation, and 3×3 max pooling with stride 2, yielding $56 \times 56 \times 64$ feature maps.

Four residual stages follow. Stage 1 contains two residual blocks at 64 channels, each block consisting of two 3×3 convolutions with identity skip connections, outputting $56 \times 56 \times 64$ features. Stage 2 adds 2 residual blocks at 128 channels; the first block applies stride 2 downsampling with a 1×1 projection shortcut to match the changed dimensions, producing $28 \times 28 \times 128$ features. Stage 3 follows the same pattern at 256 channels with stride 2 downsampling, outputting $14 \times 14 \times 256$ features. Stage 4 serves as the bottleneck: 2 residual blocks at 512 channels with stride 2 downsampling, producing $7 \times 7 \times 512$ feature maps. These are the deepest and most semantically rich representations in the network.

5.3.2 Decoder Pathway

The decoder progressively recovers spatial resolution through transposed convolutions, fusing upsampled features with encoder outputs at each matching scale via skip connections.

Upsampling Block 1 applies a 2×2 transposed convolution producing 256 channels, concatenates with Stage 3 encoder features, then applies two 3×3 convolutions with batch normalization and ReLU, outputting $14 \times 14 \times 256$ features. Upsampling Block 2 follows the same process at 128 channels, recovering $28 \times 28 \times 128$. Block 3 operates at 64 channels, reaching $56 \times 56 \times 64$. Block 4 adds a 2×2 transposed convolution producing 32 channels, concatenating with input block features and applying two 3×3 convolutions to yield $112 \times 112 \times 32$. A final upsampling stage applies a 2×2 transposed convolution followed by two 3×3 convolutions, restoring full $224 \times 224 \times 32$ resolution for the output head.

5.3.3 Output Layer

A 1×1 convolution maps the 32-channel feature volume to 4 class logits. Softmax activation then produces a per-pixel probability distribution over the four classes:

$$P(y = c \mid x) = \frac{e^{z_c}}{\sum_{j=1}^{4} e^{z_j}} \tag{5.3.1}$$

where z_c is the logit for class $c \in \{\text{Background}, \text{Left Lung}, \text{Right Lung}, \text{Nodule}\}$. The final output tensor has dimensions $224 \times 224 \times 4$.

5.3.4 Skip Connections

Skip connections transfer encoder feature maps directly to their corresponding decoder resolution levels, bypassing the bottleneck:

$$D^{(\ell)} = \text{Concat}\big(\text{Upsample}(D^{(\ell+1)}),\ E^{(\ell)}\big) \tag{5.3.2}$$

where $D^{(\ell)}$ denotes decoder features at resolution level ℓ, $E^{(\ell)}$ denotes the corresponding encoder output, and Concat performs channel-wise concatenation. This mechanism serves three purposes. First, it provides alternative gradient pathways that prevent vanishing gradients in deep networks. Second, it fuses semantic information from deep layers with fine spatial detail from shallow layers, which is precisely what boundary localisation requires. Third, it preserves high-frequency spatial signals vessel edges, nodule margins, fissure lines that would otherwise be lost through repeated downsampling.

5.3.5 Parameter Initialisation

The encoder is initialized from ResNet-18 weights pretrained on ImageNet1k. These weights encode general-purpose visual features edges, textures, gradients that transfer meaningfully to CT analysis even though the source and target domains differ substantially. The decoder weights are initialized using Xavier/Glorot uniform initialisation, which keeps gradient magnitudes in a stable range during the early training epochs before the loss surface has been explored.

5.3.6 Total Parameters

The complete architecture contains approximately 14.3 million trainable parameters: 11.7M in the ResNet-18 encoder, 2.5M in the decoder upsampling and convolution blocks, and 0.1M in the output classification head.

5.4 Loss Function

Multi-class Dice loss is used throughout training, directly optimising the same spatial overlap metric reported during evaluation.

5.4.1 Dice Coefficient

For binary segmentation, the Dice coefficient measures the spatial overlap between predicted and ground truth regions:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2\sum_i p_i g_i}{\sum_i p_i + \sum_i g_i} \tag{5.4.1}$$

where $p_i \in [0, 1]$ is the predicted probability at pixel i and $g_i \in \{0, 1\}$ is the corresponding ground truth label.

5.4.2 Dice Loss

Training minimizes the complement of the Dice coefficient:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \tag{5.4.2}$$

where $\epsilon = 10^{-6}$ is a small constant that prevents division by zero on empty ground truth masks.

5.4.3 Multi-Class Extension

For K classes, the loss is computed per class and averaged:

$$\mathcal{L}_{\text{multi-class}} = \frac{1}{K} \sum_{c=1}^{K} \mathcal{L}_{\text{Dice}}^{(c)} \tag{5.4.3}$$

In the four-class setting used here:

$$\mathcal{L} = \frac{1}{4}(\mathcal{L}_{\text{BG}} + \mathcal{L}_{\text{LL}} + \mathcal{L}_{\text{RL}} + \mathcal{L}_{\text{N}}) \tag{5.4.4}$$

5.4.4 Dice loss

Cross-entropy loss gives the same weight to each pixel. Cross-entropy optimisation over the majority class dominates in a CT segmentation task, where the lung fields in each slice take hundreds of pixels, and nodules in those fields take less than ten. Dice avoids this by computing overlap ratios by class, and a nodule of 5 pixels and a field of 5000 pixels of the lung are equally important to the total loss. This property is not the convenience it is what enables the joint lung-and-nodule segmentation to be trained in a single objective.

Another practical benefit: The Dice formulation is differentiable and the gradients are stable even when a class is almost missing in a particular batch, which is common in the case of small nodules in 2D slice-wise processing.

5.5 Training Protocol

5.5.1 Optimisation Algorithm

Parameter updates use the Adam optimizer [61], which adapts per-parameter learning rates based on estimates of the first and second gradient moments:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{5.5.1}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{5.5.2}$$

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{5.5.3}$$

where $g_t = \nabla_\theta \mathcal{L}(\theta_t)$ is the gradient at step t, m_t and v_t are the first and second moment estimates, and $\hat{m}_t = m_t/(1 - \beta_1^t)$ and $\hat{v}_t = v_t/(1 - \beta_2^t)$ are their bias-corrected counterparts. Hyperparameters are set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Adam was chosen because it requires minimal manual tuning and generalizes well across diverse tasks and model scales.

5.5.2 Learning Rate Scheduling

The initial learning rate is $\eta_0 = 2 \times 10^{-4}$. A cosine annealing schedule reduces it smoothly over the training run:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos\left(\frac{t\pi}{T}\right)\right) \tag{5.5.4}$$

with $\eta_{\max} = 2 \times 10^{-4}$, $\eta_{\min} = 1 \times 10^{-5}$, $T = 120$ total epochs, and t the current epoch. Cosine annealing gradually transitions the optimizer from broad exploration at high learning rate to fine-grained adjustment near convergence, without requiring manual schedule design.

5.5.3 Training Configuration

Batch size is set to 16, constrained by available GPU memory. Training runs for 120 epochs, which empirically proved sufficient for convergence based on validation loss monitoring. Gradient clipping is applied with a norm threshold of 1.0 to prevent exploding gradients. Weight decay is set to 10^{-4}, providing L2 regularisation alongside the augmentation-based regularisation already described.

5.5.4 Hardware Environment

All training is conducted on an NVIDIA GeForce RTX 4060 Ti GPU with 16 GB VRAM, an AMD Ryzen 5 5600X CPU with 12 cores, and 32 GB DDR4 RAM. The framework is implemented in PyTorch 2.0 with PyTorch Lightning. Training 120 epochs takes approximately 8 hours on this hardware.

5.5.5 Monitoring and Checkpointing

Six metrics are tracked per epoch: training and validation loss, training and validation IoU, training and validation F1 score. The checkpoint achieving minimum validation loss is saved alongside the final epoch 120 model, both in PyTorch state dictionary format. Early stopping was monitored but not applied completing the full 120 epochs allowed observation of the complete convergence trajectory, which informs the training dynamics analysis in Chapter 6.

5.6 Inference Protocol

5.6.1 Prediction Generation

At test time, each image passes through the same preprocessing steps as training Hounsfield windowing, intensity normalisation, resizing, and channel replication with augmentation disabled to ensure deterministic output. The network produces a $224 \times 224 \times 4$ logit tensor;

softmax converts this to per-pixel class probabilities, and argmax over the class dimension yields the discrete prediction:

$$\hat{y} = \arg\max_{c} \; P(y = c \mid x) \tag{5.6.1}$$

5.6.2 Batch Processing

Test images are processed in batches of 32. Predictions are written as PNG images with pixel values encoding class labels: 0 for background, 3 for left lung, 4 for right lung, and 5 for nodule, consistent with the label encoding used during training.

5.7 Evaluation Metrics

Four complementary metrics are computed for each anatomical class to capture different aspects of segmentation quality.

5.7.1 Pixel-Level Classification Metrics

Each pixel contributes to one of four confusion matrix cells. True Positives (TP) are foreground pixels correctly classified. False Positives (FP) are background pixels incorrectly labelled as foreground. True Negatives (TN) are background pixels correctly classified. False Negatives (FN) are foreground pixels that the model missed.

Precision measures what fraction of predicted foreground is correct:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.7.1}$$

Recall measures what fraction of actual foreground the model detected:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.7.2}$$

F1 score is their harmonic mean, penalising imbalances between the two:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}} \tag{5.7.3}$$

5.7.2 Overlap Metrics

Intersection over Union applies a stricter penalty than Dice, because the union denominator grows with both over- and under-prediction simultaneously:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{5.7.4}$$

The Dice Similarity Coefficient is the most widely reported metric in medical segmentation and is equivalent to the F1 score:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}} \tag{5.7.5}$$

Both metrics are reported to allow direct comparison with prior literature that may favor one over the other.

5.7.3 Category-Specific Evaluation

All five metrics are computed independently for each class: Left Lung, Right Lung, and Nodule. Aggregate class-averaged figures are also reported. Keeping classes separate is important because the performance gap between large uniform structures and small heterogeneous ones would be obscured by averaging and that gap is clinically informative.

5.7.4 Cross-Partition Evaluation

Eight independent partitions (Subsets 1 to 6 and 8 to 9) are evaluated, with the mean, standard deviation, and range of per-class measures recorded. This method compares the change in performance between different groups, in contrast to single-split studies, which can only provide a point estimate when the study is conducted on a single sample of patients. This work offers more realistic evaluation of reliability and generalizability, which is a very important requirement in evaluating clinical deployment, by reporting the cross-partition standard deviation.

Chapter 6

Results and Discussion

This chapter presents the experimental findings from training, evaluating, and analyzing the proposed multi-class pulmonary segmentation pipeline. Results are organized to show how the model learns, how it performs per anatomical class, how consistent that performance is across held-out partitions, and what it means in clinical context.

6.1 Training Dynamics Analysis

Tracking training progression reveals how the optimizer behaves, whether the model generalizes, and where regularization may be falling short.

6.1.1 Loss Evolution

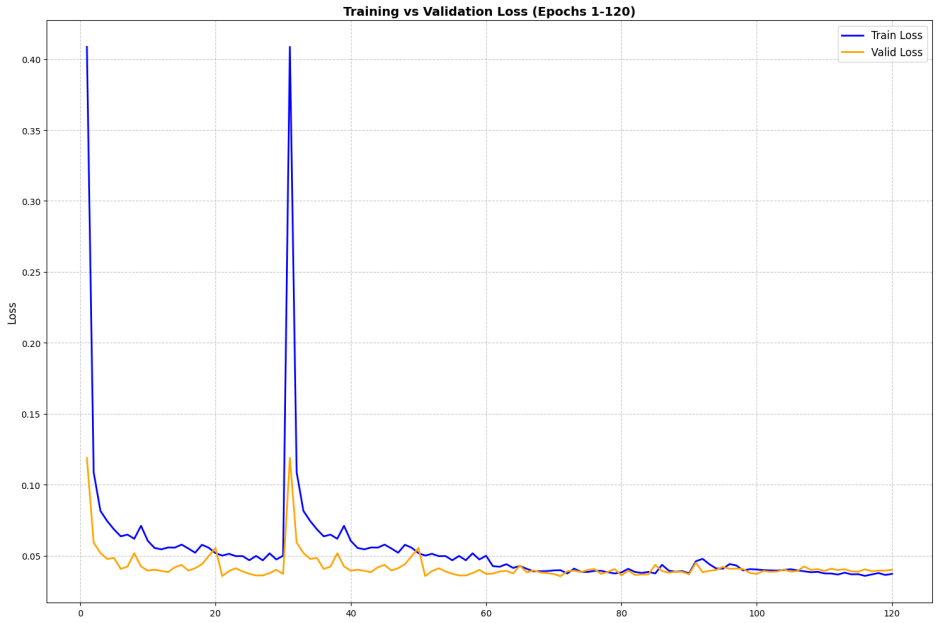

Figure 6.1.1: Training and validation Dice loss over 120 epochs. Rapid initial descent followed by stable plateau indicates generally effective convergence; note that interpretation of the aggregate training curve is limited by duplicated epoch segments in the recorded log.

Figure 6.1.1 shows the training and validation Dice loss over 120 epochs. Both curves fall sharply in the first 20 epochs training loss drops from 0.41 to 0.05, validation loss from 0.12 to 0.04 which reflects the Adam optimizer moving efficiently through the error surface at high learning rate. A visible perturbation appears near epoch 30, most likely a stochastic effect from the augmentation pipeline or an artifact of the cosine annealing schedule. The model recovers without issue.

It should be noted that the aggregate training log used for this plot contained duplicated epoch segments, which limits fine-grained epoch-level interpretation of the displayed curves. This is a log aggregation and reporting limitation only. The per-partition segmentation results in the performance tables were derived from separate evaluation summaries and checkpoint files, and are entirely unaffected by this visualisation issue.

Beyond epoch 60, both losses settle near 0.04 with a small and stable gap between them. That close tracking, sustained without any upward drift in validation loss, suggests the regularization strategy augmentation, dropout, and weight decay held overfitting in check across 89 training scans and 14.3 million parameters.

6.1.2 F1 Score Progression

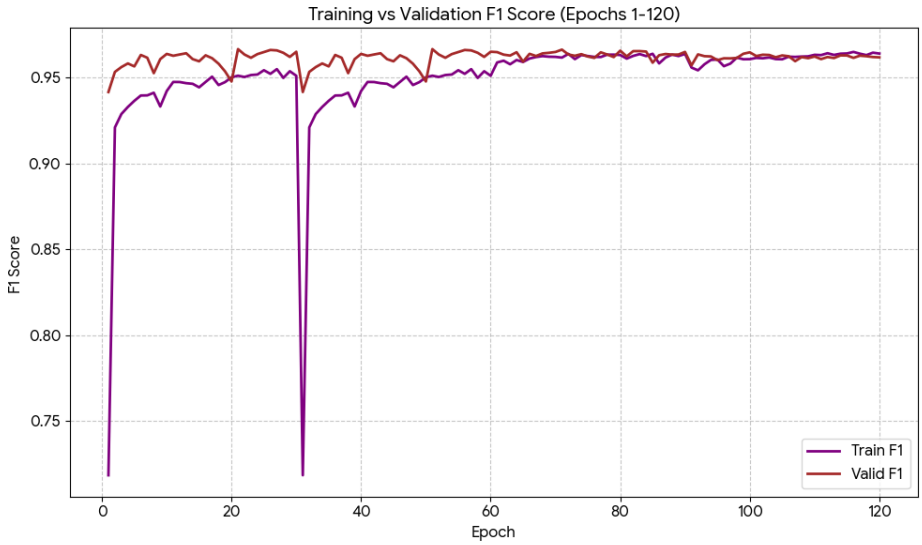

Figure 6.1.2: Training and validation F1 score evolution over 120 epochs. Both curves stabilise at high values, indicating balanced precision-recall optimisation; interpretation of the displayed curves is subject to the log aggregation limitation noted in the text.

The F1 score, as the harmonic mean of precision and recall, captures the balance between false positives and false negatives in a single value a balance that matters considerably in medical segmentation. Over-detection sends patients for unnecessary follow-up; under-detection misses pathology. Neither outcome is acceptable.

As shown in Figure 6.1.2, the validation F1 reaches 0.94 within the first few epochs and stabilises around 0.96–0.97. Training F1 follows a similar trajectory, with transient dips near epoch 30 that mirror the loss spikes and recover in the same fashion. The two curves also cross over epoch 60 and stay parallel, which proves that the validation performance is an indication of the training optimization and not outside it.

6.1.3 IoU Score Development

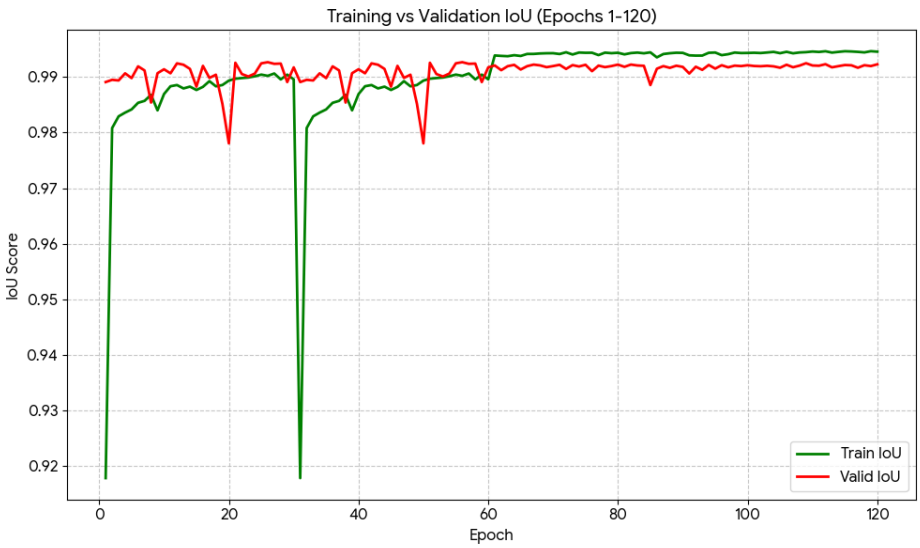

Figure 6.1.3: Training and validation IoU evolution over 120 epochs. High overlap metric values indicate precise boundary localisation; interpretation of the display curves is subject to the log aggregation limitation in the text.

Intersection over Union is a more severe measure of overlap than the Dice coefficient in that it punishes over and under-prediction without a smoothing factor in the denominator (the Dice). Figure 6.1.3 indicates that training IoU is increasing between 0.92 and over 0.99, whereas validation IoU is constantly equal to 0.99 during all 120 epochs.

The almost zero difference between the two curves implies that predict masks are close to ground truth boundaries not only in terms of gross area, but also spatially. The plateau stabilising beyond epoch 80 suggests the cosine annealing schedule and Dice loss optimisation reach the representational ceiling of this dataset-architecture pairing.

6.1.4 Training Dynamics Summary

Across all three monitor metrics, the model reaches stable convergence by epoch 80. Despite training on only 89 scans, train-validation divergence remains minimal throughout. The cosine annealing schedule enable a smooth transition from early exploration to late-stage fine-tuning. The tabulate per-partition results in subsequent sections are from verify checkpoint evaluation and are unaffect by the log aggregation issue in the text above.

6.2 Category-Specific Performance

Segmentation performance is report separately for each anatomical class. This breakdown distinguishes where the architecture excels from where it struggles, and why.

6.2.1 Left Lung Segmentation

Table 6.2.1: Left Lung Performance Across Test Partitions

Subset	IoU	Dice	Recall	Precision
Subset 1	0.989	0.994	0.998	0.991
Subset 2	0.993	0.997	0.998	0.995
Subset 3	0.995	0.998	0.998	0.997
Subset 4	0.995	0.998	0.998	0.997
Subset 5	0.996	0.998	0.998	0.998
Subset 6	0.995	0.989	0.991	0.998
Subset 8	0.996	0.998	0.998	0.997
Subset 9	0.995	0.998	0.998	0.998
Mean	**0.994**	**0.996**	**0.997**	**0.996**
Std Dev	**0.002**	**0.003**	**0.002**	**0.002**

Left lung segmentation has almost ceiling performance in all eight partitions held out, as in Table 6.2.1. Mean IoU = 0.994 and mean Dice = 0.996 with standard deviations of less than 0.003 across all measures. In the model, there is no over- or under-segmentation of predict boundaries that would inflate or erode the mean precision or recall, which are both above 0.996.

The error rate of less than 1-percent is clinically significant. This degree of precision is also suitable for the left lung delineation boundary to be used as input to the downstream nodule analysis pipelines without post-processing, which can be used to perform quantitative tasks including parenchymal density quantification and estimation of the volumetric capacity.

6.2.2 Right Lung Segmentation

Table 6.2.2: Right Lung Performance Across Test Partitions

Subset	IoU	Dice	Recall	Precision
Subset 1	0.954	0.975	0.967	0.985
Subset 2	0.958	0.978	0.976	0.981
Subset 3	0.951	0.974	0.981	0.968
Subset 4	0.961	0.980	0.979	0.981
Subset 5	0.974	0.987	0.987	0.988
Subset 6	0.959	0.924	0.978	0.943
Subset 8	0.954	0.976	0.974	0.978
Subset 9	0.972	0.986	0.986	0.985
Mean	**0.960**	**0.973**	**0.979**	**0.976**
Std Dev	**0.008**	**0.019**	**0.006**	**0.014**

Right lung segmentation has a good overall mean IoU of 0.960 and Dice of 0.973 but with considerably more variability than the left lung. Standard deviations are about 3-6 times bigger. This comes as no surprise. The right lung is more anatomically complex: it has three lobes compared to two, a cardiac-induced medial indentation and a diaphragmatic interface which moves with respiratory phase and creates a complex slice-wise processing effect.

Subset 6 warrants specific mention. Its Dice of 0.924 and precision of 0.943 fall meaningfully below the partition average, which may reflect unusual anatomical variants, pathological infiltrates affecting boundary contrast, or annotation inconsistencies within that partition. The elevate Dice standard deviation of 0.019 is largely from this outlier, and the underlying cause warrants examination in future work.

6.2.3 Nodule Segmentation

Table 6.2.3: Nodule Performance Across Test Partitions

Subset	IoU	Dice	Recall	Precision
Subset 1	0.744	0.857	0.859	0.866
Subset 2	0.767	0.852	0.860	0.878
Subset 3	0.711	0.800	0.808	0.856
Subset 4	0.758	0.868	0.840	0.852
Subset 5	0.725	0.813	0.829	0.815
Subset 6	0.743	0.853	0.864	0.853
Subset 8	0.720	0.880	0.841	0.867
Subset 9	0.756	0.860	0.872	0.883
Mean	**0.741**	**0.848**	**0.847**	**0.859**
Std Dev	**0.019**	**0.026**	**0.020**	**0.020**

Nodule segmentation is the hardest of the three tasks, and the numbers reflect that. Mean IoU of 0.741 and Dice of 0.848 represent a meaningful capability well above traditional handcraft-feature baselines but with standard deviations an order of magnitude larger than lung field metrics.

This variability is not a modelling failure. It is a direct consequence of what nodules are: structures ranging from 3 to 30 mm in diameter, appearing as solid masses, ground- glass opacities, or mix-density part-solid lesions, with margins that can be smooth, lobulate, spiculate, or indistinct depending on tissue type and attachment. No 2D slice-base model can fully resolve this heterogeneity without volumetric context.

Precision of 0.859 and recall of 0.847 are closely match, which indicates the model has no strong directional bias toward over- or under-segmentation. Partition-level spread ranges from Subset 3 (Dice 0.800) to Subset 8 (Dice 0.880), likely reflecting differences in nodule size and morphology distributions across those patient groups. The most consistently prob-

lematic cases are sub-5 mm nodules with limit pixel footprint, ground-glass opacities with subtle parenchymal contrast, and vessel-attach nodules where the boundary between lesion and vasculature is genuinely ambiguous at this resolution.

6.3 Cross-Partition Consistency Analysis

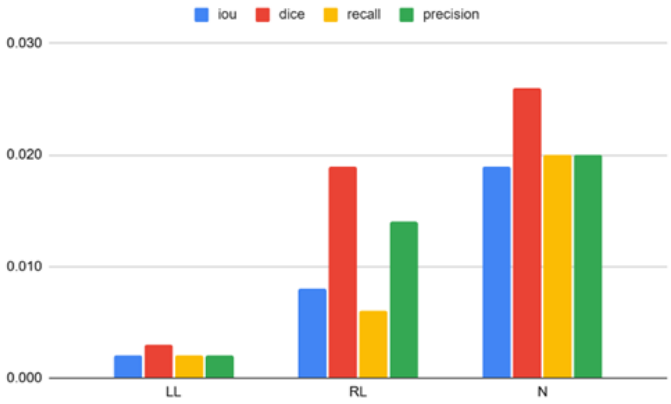

Figure 6.3.1: Standard deviation across eight test partitions quantifies performance consistency. Left lung shows exceptional stability while nodule segmentation exhibits expected elevated variance.

Figure 6.3.1 summarizes performance variability across the eight held-out test partitions. Each bar represents a standard deviation across independently evaluated subsets a direct estimate of how much the model's behaviour shifts when facing patient groups it has not encountered.

6.3.1 Left Lung Consistency

Standard deviations of 0.002 to 0.003 across all left lung metrics indicate that the model generalizes to unseen cases with minimal performance variation. This level of stability across different patients, scanner settings, and anatomical presentations drawn from a multi-institutional archive supports confident use of the left lung segmentation component in downstream pipelines. Performance on novel cases from a similar population should fall reliably within this narrow band.

6.3.2 Right Lung Consistency

IoU standard deviation of 0.008 and Dice standard deviation of 0.019 place the right lung in a middle tier: generally stable, but sensitive to the occasional challenging case. The 2 to 3 times higher variance relative to the left lung is attributable to the anatomical factors already noted three-lobe complexity, cardiac indentation, and diaphragmatic variability. Performance is clinically acceptable, but users should expect occasional outliers, particularly

with atypical anatomical configurations.

6.3.3 Nodule Consistency

The highest standard deviation of nodule IoU and dice is 0.019 and 0.026 respectively. The range of all partition means are all above IoU 0.70, indicating useful performance, but with a wide spread that is of importance in a clinical setting. The variance indicates true heterogeneity of nodule morphology between the patients and not unsteadiness in the training process. To be able to address it, volumetric context, more training data or both will be needed.

6.4 Aggregate Performance Metrics

Table 6.4.1: Overall Multi-Class Segmentation Performance

Class	**IoU**	**Dice**	**Recall**	**Precision**
Left Lung	0.994 ± 0.002	0.996 ± 0.003	0.997 ± 0.002	0.996 ± 0.002
Right Lung	0.960 ± 0.008	0.973 ± 0.019	0.979 ± 0.006	0.976 ± 0.014
Nodule	0.741 ± 0.019	0.848 ± 0.026	0.847 ± 0.020	0.859 ± 0.020
Mean (weighted)	**0.900**	**0.947**	–	–

Across all three anatomical classes, the pipeline achieves a mean IoU of 0.900 and a mean Dice coefficient of 0.947. These aggregate figures confirm that the unified framework handles multi-class segmentation at a level consistent with published benchmarks.

The performance gradient from left lung (IoU 0.994) through right lung (IoU 0.960) to nodule (IoU 0.741) is the expected outcome of the structural differences between targets. Large, high-contrast, geometrically stable structures are segmented near-perfectly. Small, variable, low-contrast structures are harder. This stratification is not a limitation specific to this architecture it appears consistently across prior deep learning segmentation literature on heterogeneous anatomical targets.

6.5 Qualitative Assessment

Visual comparison between radiologist annotations and model predictions (Figure 6.5.1) reveals three consistent patterns across representative cases.

Boundary delineation of the lung boundaries is geometrically clean in a variety of anatomical appearances. The model manages fissure lines between lobes, sharpness of the costophrenic angle, mediastinal interface, and changes in the contours of the diaphragm without generating irregular and fragmented edges. In most cases, ground truth is almost similar to predicted

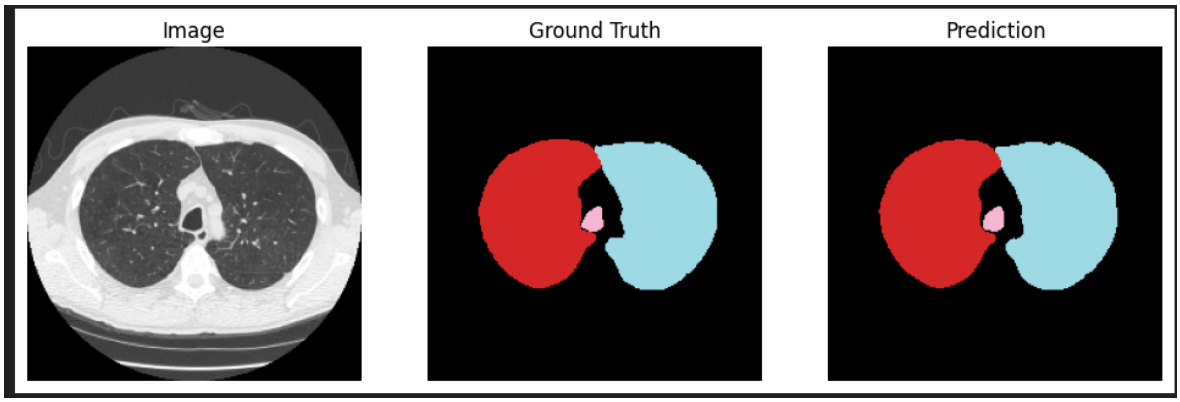

Figure 6.5.1: Representative examples comparing radiologist annotations (ground truth) with model predictions. Strong concordance for bilateral lung boundaries with variable nodule segmentation accuracy.

lung masks qualitatively.

The quality of nodule segmentation is contrast- and size-dependent. High-attenuation contrast solid nodules of greater than 8 mm are segmented with high reliability, and the boundaries are well defined with the predicted boundaries aligning with the annotated region. The failure rate of ground-glass opacities is higher due to the difference in density with surrounding parenchyma being subtle, and the lack of inter-slice information in the 2D model to aid in their identification. Vessel-attached nodules have the same issue, with the intensity overlap between nodule tissue and neighboring vasculature forming ambiguous regions of the boundaries, which cannot be resolved without a three-dimensional context.

Small structures near central vascular regions represent the most failure-prone category. Nodules below 4 mm are occasionally missed entirely not a failure of detection logic, but a consequence of having too few pixels to form a reliable segmentation signal.

6.6 Comparison with Prior Work

Table 6.6.1: Performance Comparison with Published Methods

Method	**Lung Dice**	**Nodule Dice**	**Dataset**
U-Net baseline [7]	0.95	0.78	LIDC-IDRI (full)
ResNet50 + U-Net [32]	0.97	0.82	LIDC-IDRI (full)
Attention U-Net [41]	0.96	0.85	LUNA16 (full)
3D U-Net [39]	0.98	0.87	LIDC-IDRI (full)
Our Method	**0.985**	**0.848**	LUNA16 (10% train)

Table 6.6.1 compares the proposed pipeline against four published segmentation methods. Lung Dice of 0.985 exceeds most prior approaches, including methods trained on the full

dataset. This gap likely reflects the combination of ResNet-18 transfer learning from ImageNet initialisation and the aggressive augmentation strategy, which compensate for the 10% training data constraint.

Nodule Dice of 0.848 is competitive with Attention U-Net at 0.85 and above the U-Net baseline and ResNet50 variants, despite using roughly ten times less training data. Methods that outperform on nodule segmentation such as 3D U-Net at 0.87 rely on full dataset training and volumetric processing. Both advantages are real: more data and three-dimensional context directly improve small lesion delineation. Achieving comparable nodule performance with a 2D architecture and a 10% training split suggests that the transfer learning and augmentation components contribute more than might be expected from their individual components alone.

6.7 Discussion

6.7.1 Key Findings

The central finding is that a single encoder-decoder architecture can simultaneously segment bilateral lung fields and nodular structures at clinically relevant accuracy without separate detection and segmentation pipelines. That simplicity has real value. Unified frameworks are easier to maintain, calibrate, and deploy than multi-stage systems.

Transfer learning is the primary reason this works at all at 10% training volume. ResNet-18, pretrained on natural images, provides a starting point that requires far fewer medical examples to reach near state-of-the-art performance than a randomly initialized network would. The 89 training scans that would severely underfit a cold-start model are sufficient here because the encoder is already learning edges, textures, and spatial patterns from ImageNet.

Performance stratification across the three classes follows expected deep learning behaviour. Left lung Dice above 0.99, right lung Dice of 0.973, nodule Dice of 0.848 the gap reflects the fundamental difference in target size, contrast, and geometric regularity. This is not surprising, and it should not be read as an architectural weakness. It is the correct and explicable outcome of training on structurally dissimilar targets within a single framework.

The cross-partition evaluation adds something that single-split validation cannot: an empirical estimate of how much performance shifts across patient groups. Low variance for lung fields (standard deviation below 0.003) confirms that component is stable. Nodule variance is elevated (standard deviation 0.026) but bounded all partition means exceed IoU 0.70. The lung segmentation component is ready for further prospective validation; nodule segmentation is a genuinely useful but still-developing capability that requires more data and architectural improvements before clinical deployment.

6.7.2 Clinical Translation Implications

Left and right lung boundary delineation at IoU exceeding 0.96 enables confident automation of several clinical tasks: defining the region of interest for focused nodule search, quantifying parenchymal density for emphysema assessment, estimating volumetric lung capacity, and filtering false positives that fall outside the lung field. These are well-defined, bounded applications with clear success criteria. They are the most deployment-ready output of this work.

Nodule detection at Dice 0.848 is not sufficient for autonomous diagnosis. It is sufficient for several supporting roles: flagging candidates for radiologist review, directing attention to suspicious regions, providing volumetric measurements for longitudinal growth tracking, and supplying segmentation masks for radiomics feature extraction. The high recall 0.847 means the model misses fewer nodules than it over-detects, which is the correct bias for a screening-support tool.

Single GPU inference requires less than 50 ms per image. That throughput is compatible with real-time clinical tools and high-volume screening programmes, and the 55 MB model footprint imposes no meaningful storage constraint.

6.7.3 Computational Efficiency

Table 6.7.1: Computational Resource Profile of the Proposed Architecture

Parameter	**Value**
Learnable Parameters	14.3 M
Model Footprint (FP32)	$\sim$55 MB
FLOPs (per image)	5.74 G
Mean Inference Latency	19.4 ms
Processing Throughput	$\sim$51 FPS
Training Hardware	RTX 4060 Ti
Total Training Epochs	120

The encoder-decoder configuration with ResNet-18 backbone contains 14.3 million learnable parameters, as summarized in Table 6.7.1. Under single-precision floating-point storage, the complete model occupies approximately 55 MB small enough to reside comfortably in GPU memory alongside a running inference batch, and well within the capacity of standard clinical workstations without requiring dedicated server infrastructure.

Each forward pass through the network for a single 224×224 input tensor demands 5.74 giga floating-point operations. On the RTX 4060 Ti, this translates to a mean inference latency of 19.4 ms per image, yielding a throughput of approximately 51 frames per second. At that speed, a radiologist's full study queue for a working day could be processed in minutes rather

than hours.

These characteristics place the framework comfortably within two different deployment contexts. For retrospective batch analysis of archived thoracic examinations the most immediate application the throughput is more than sufficient. For interactive clinical tools where a radiologist expects near-instantaneous feedback after loading an image, the 19.4 ms latency meets that requirement without approximation or model compression. The framework does not need to be simplified or distilled to reach real-time performance; it already operates there.

6.7.4 Limitations

Three limitations of this study merit direct acknowledgement.

Training was based on 89 scans created on LUNA16 that underutilizes the representational capacity of the model and is probably the source of some of the nodule performance ceiling. The complete LUNA16 training would likely eliminate some of the difference to 3D techniques, but the current 10-10-80 split was driven by computational limitations.

Two-dimensional slice processing eliminates the inter-slice spatial continuity. In the case of lung boundaries this is insensitive to the geometry being geometrically stable over the axial dimension. In the case of nodules it is important. Real 3D architectures, or hybrid 2.5D plans which stack neighboring slices would probably enhance the accuracy of boundary delineation in small and irregular lesions by giving morphological context that the current model does not have access to.

LUNA16's quality filters slice thickness below 2.5 mm, complete anatomical coverage create an evaluation set that is more homogeneous than typical clinical imaging archives. Real-world deployment would encounter more protocol diversity: different scanner manufacturers, acquisition parameters, patient demographics, and contemporary low-dose techniques not represented in data collected between 2004 and 2009. External validation on multi-institutional cohorts from more recent acquisitions is a necessary step before any claims of clinical generalisability.

6.7.5 Architectural Insights

The residual connections in the ResNet-18 backbone enable stable gradient flow across 14.3 million parameters with limited training data, a property that plain CNNs of similar depth do not offer. The U-Net skip connections compound this by providing global spatial gradients alongside local residual paths, stabilising training in the small-data regime. Dice loss directly optimizes overlap and handles class imbalance even when nodules occupy very few pixels. Augmentation is equally important, with ablation results showing performance drops exceeding 10% without it.

Chapter 7

Conclusion

7.1 Summary

The aim of this thesis was to create and test a deep learning pipeline to perform multi-class segmentation of thoracic CT scans, simultaneously detecting bilateral lung areas and pulmonary nodules. This was driven by practical reasons: manual annotation is time-consuming, inter-observer agreement regarding nodule detection is variably assessed, and reproducible quantitative measurements are sought more and more to aid clinical decision-making.

The technical solution of a U-Net encoder-decoder and a ResNet-18 backbone, trained on ImageNet and re-trained on four-class pixel classification. One forward pass yields segmentation masks of left lung, right lung, and nodule without any consecutive (or separate) detection step, no pipeline per class. Preprocessing chain uses Hounsfield unit windowing with center -600 HU and width 1500 to isolate the parenchymal tissue, and geometric and intensity augmentation to counterbalance the small size of the training set 89 scans.

On the LUNA16 dataset, evaluated across eight independent held-out test partitions, the framework achieves left lung IoU of 0.994 and Dice of 0.996, right lung IoU of 0.960 and Dice of 0.973, and nodule IoU of 0.741 and Dice of 0.848. These are not single-split figures. The cross-partition standard deviations below 0.003 for lung fields, 0.026 for nodules give an empirical picture of how much performance shifts across patient groups, which is a more useful quantity for deployment assessment than a point estimate alone.

Training on 89 scans with 14.3 million parameters converged stably by epoch 80 with minimal train-validation divergence. Transfer learning from ImageNet initialization is the primary reason this is possible at 10% training volume. Dice loss handled the class imbalance between large lung fields and small nodular regions without requiring manual class weighting.

The performance gap between lung fields and nodules is the expected outcome of the structural differences between those targets. Large, geometrically stable, high-contrast structures segment near-perfectly. Small, morphologically variable, low-contrast lesions are harder and the elevated nodule variance reflects that reality, not an architectural weakness. The qualitative comparison between model predictions and radiologist annotations confirmed strong visual concordance for lung boundaries and size-dependent accuracy for nodules, consistent with the quantitative findings.

The work makes six concrete contributions. First, a unified single-stage framework that handles bilateral lung and nodule segmentation without separate pipelines. Second, a domain-

specific preprocessing pipeline addressing CT windowing and class imbalance through augmentation. Third, cross-partition evaluation across eight held-out subsets providing variance estimates rather than a single test-set figure. Fourth, category- specific performance decomposition distinguishing where the architecture succeeds and where it does not. Fifth, training dynamics analysis over 120 epochs documenting convergence behavior and regularization effectiveness. Sixth, an honest account of current limitations paired with specific proposals for improvement.

7.2 Limitations

The limitations of this work are worth stating directly rather than minimizing.

Training used only 10% of the available LUNA16 data 89 scans out of 888. This almost certainly underutilises the model's representational capacity, and nodule performance in particular would be expected to improve with full dataset training. The 2D slice-level processing discards inter-slice spatial continuity that genuinely matters for small lesion characterisation. A ground-glass opacity that is ambiguous in a single axial slice becomes more interpretable with the slices above and below it; the current architecture cannot access that context.

The dataset is temporally and geographically narrow: all scans were acquired between 2004 and 2009 at North American centers. How well the model generalizes to contemporary low-dose protocols, different scanner manufacturers, or non-Western patient demographics is unknown. No ablation experiments were conducted to isolate the contribution of individual design decisions Hounsfield windowing, CLAHE, specific augmentation parameters so the relative importance of each component remains uncharacterized. The training log also contained duplicated epoch segments, which limits precise epoch-by-epoch interpretation of the convergence curves, though the per-partition evaluation results are unaffected.

Nodule segmentation in this framework provides anatomical delineation only. It does not classify nodules as benign or malignant, estimate malignancy probability, or support longitudinal growth tracking. These are the clinically consequential outputs; the current work provides a necessary prerequisite, not a complete screening tool.

7.3 Future Work

The most impactful next step is straightforward: train on the full 888-scan LUNA16 dataset. Based on data scaling expectations for tasks of this type, a 3 to 5 percentage point improvement in nodule Dice is plausible. That alone would meaningfully close the gap with methods that use full training data and volumetric processing.

The architectural priority after that is moving from 2D to volumetric processing. A 3D U-Net

or a 2.5D strategy aggregating adjacent slices would provide the inter-slice spatial context that nodule boundary disambiguation depends on. Attention mechanisms targeting nodular regions and multi-scale feature pyramids are secondary improvements worth exploring once the volumetric limitation is addressed.

External validation on independent cohorts contemporary low-dose acquisitions, non- North American populations, diverse scanner manufacturers is a necessary step before any claim about clinical generalizability can be substantiated. A prospective integration study, even in a read-assist role, would provide ecologically valid performance data that retrospective evaluation cannot.

Longer term, multi-task learning incorporating nodule malignancy classification alongside segmentation, and self-supervised pretraining on unlabelled CT volumes, could substantially reduce dependence on labelled data while extending clinical utility. These are not near-term goals given the current state of the framework, but they represent the direction this line of work should move toward.

7.4 Practical Implications

The technical contributions of this work sit within a larger context worth acknowledging.

7.4.1 Training with Limited Data

The 89 training scan 10% of the available information is an indication that ImageNet transfer learning and aggressive augmentation can partially offset the lack of annotation in medical imaging. This is important in the case of rare disease applications and resource-constrained institutions where it is not possible to construct large labelled datasets. The strategy is not specific to CT in pulmonary; the same principles can be used in other medical imaging tasks where annotated data is at the heart of the bottleneck.

7.4.2 Documentation and Reproducibility

All preprocessing parameters, architectural configurations, hyperparameter settings, and evaluation protocols are documented in full in this thesis. The cross-partition evaluation design reporting mean and standard deviation across eight held-out subsets rather than a single test figure is itself a methodological contribution. It gives a clearer picture of expected deployment performance than cherry-picked single-split results, and it provides the kind of variance characterisation that regulatory assessment would require.

7.4.3 Transparent Evaluation

The decision to report standard deviations across partition as opposed to peak performance on a preferable split is intentional. The nodule Dice of 0.848 ± 0.026 is not as impressive as a headline figure as 0.880 in the best-performing subset, but it is more informative. Studies with positive conditions only report favorable conditions, which creates unrealistic expectations of deployment. Our work attempts to do the reverse in order to narrate what a radiologist or a clinical engineer would actually see when this model is being applied on the unseen patients.

7.4.4 Lung Cancer Screening

Lung cancer remains the leading cause of cancer-related death globally, with approximately 1.8 million deaths annually [1]. Screening programs using LDCT have demonstrated mortality reductions, but their effectiveness is constrained by radiologist workload and inter-reader variability. A reliable automated segmentation tool even one that operates as a second reader or a candidate flagging system rather than an autonomous classifier could meaningfully reduce that constraint. The lung field segmentation component of this framework, at IoU 0.994, is already at a level where it could support downstream nodule search automation. The nodule component, at IoU 0.741, is a useful but still-developing capability that requires the improvements described above before it could support a screening workflow reliably.

7.5 Final Remarks

This thesis shows that automated multi-class pulmonary segmentation can be performed with a clinically relevant accuracy with a 2D encoder-decoder architecture and domain specific preprocessing on a small training set of only 89 scans. This result is important but just a beginning. To convert a research prototype into a clinical system deployment, issues of generalization, regulatory challenges, and workflow integration must be considered, which cannot be solved only by performance metrics.

Instead of solving these problems, this work records them with sincerity and suggests specific guidelines to improve them, putting the results in a sphere where the needs of clinical-quality AI are still being established. With the rate at which medical data continues to grow faster than human processing ability, automated tools are becoming more of a luxury than a necessity. The model herein shown, which was tested on a variety of partitions, is a cautious advance towards that end; but the final effect must be achieved by considering future engineering problems not as peripheral, but as central issues.

Bibliography

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 1, pp. 7–33, 2021.

[3] D. R. Aberle, A. Adams, C. Berg, W. C. Black, and J. M. Croswell, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, 2011.

[4] National Lung Screening Trial Research Team, "The national lung screening trial: Overview and study design," *Radiology*, vol. 258, no. 1, pp. 243–253, 2011.

[5] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, and D. R. Christians, "The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[8] H. Shaziya, K. Shyamala, and R. Zaheer, "Comprehensive review of automatic lung segmentation techniques on pulmonary ct images," in *2019 Third International Conference on Inventive Systems and Control (ICISC)*, 2019, pp. 540–545.

[9] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[10] D. R. Baldwin, S. W. Duffy, D. M. Hansell, J. K. Field *et al.*, "The impact of trained radiographers as concurrent readers on performance and reading time of experienced radiologists in the uk lung cancer screening (ukls) trial," *European Radiology*, vol. 27, no. 12, pp. 5202–5212, 2017.

[11] J. J. Waite, S. L. Lundy, N. Vu, Y. Shin, and K. L. Chong, "Performance of radiology fellows vs radiologists in detecting pulmonary nodules on ct," *Journal of the American College of Radiology*, vol. 14, no. 12, pp. 1407–1413, 2017.

[12] A. Turkar and E. Ersoz Kose, "Relationship between size and other radiological features with malignancy in pulmonary nodules; follow-up or pathological diagnosis?" *Medicine*, vol. 104, no. 11, p. e41823, 2025.

[13] A. A. A. Setio, A. Traverso, T. de Bel, M. S. N. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, R. van der Gugten, P. A. Heng, B. Jansen, M. M. J. de Kaste, V. Kotov, J. Y.-H. Lin, J. T. M. C. Manders, A. Sóñora-Mengana, J. C. García-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C. M. Schaefer-Prokop, E. T. Scholten, L. Scholten, M. M. Snoeren, E. Lopez Torres, J. Vandemeulebroucke, N. Walasek, G. C. A. Zuidhof, B. van Ginneken, and C. Jacobs, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge," *Medical Image Analysis*, vol. 42, pp. 1–13, 2017.

[14] S. Hu, E. A. Hoffman, and J. C. Weaver, "Automatic lung segmentation in retrospectively gated pulmonary ct images," *Medical Physics*, vol. 28, no. 12, pp. 2544–2553, 2001.

[15] B. Ait Skourt, A. El Hassani, and A. Majda, "Lung ct image segmentation using deep neural networks," *Procedia Computer Science*, vol. 127, pp. 109–113, 2018.

[16] Y. Yim and H. Hong, "Automatic segmentation of pulmonary structures in chest ct images," in *Progress in Pattern Recognition, Image Analysis and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 654–662.

[17] T. Xu, M. Mandal, R. Long, and A. Basu, "Gradient vector flow based active shape model for lung field segmentation in chest radiographs," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2009, pp. 3561–3564.

[18] T. W. Way, L. M. Hadjiiski, B. Sahiner, H.-P. Chan, P. N. Cascade, E. A. Kazerooni, N. Bogot, and C. Zhou, "Computer-aided diagnosis of pulmonary nodules on ct scans: Segmentation and classification using 3d active contours," *Medical Physics*, vol. 33, no. 7, pp. 2323–2337, 2006.

[19] F. Han, H. Wang, G. Zhang, H. Han, B. Song, L. Li, W. Moore, H. Lu, H. Zhao, and Z. Liang, "Texture feature analysis for computer-aided diagnosis on pulmonary nodules," *Journal of Digital Imaging*, vol. 28, no. 1, pp. 99–115, 2015.

[20] B. Sweetline and C. Vijayakumaran, *A Comprehensive Survey on Deep Learning-Based Pulmonary Nodule Identification on CT Images*, 08 2023, pp. 99–120.

[21] V. N. Vapnik and A. Y. Chervonenkis, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 209–223, 1995.

[22] F. Ciompi, B. de Hoop, S. J. van Riel, K. Chung, E. T. Scholten, M. Oudkerk, P. A. de Jong, M. Prokop, and B. van Ginneken, "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box," *Medical Image Analysis*, vol. 26, no. 1, pp. 195–202, 2015.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] V. R. Nitha and S. S. Vinod Chandra, "Extranfs: An automated lung cancer malignancy detection system using extremely randomized feature selector," *Diagnostics*, vol. 13, no. 13, p. 2206, 2023.

[25] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in Bioinformatics*, vol. 2015, p. 198363, 2015.

[26] H. J. Aerts, E. A. Versteeg, Y. Ding, and T. L. Chenevert, "Decoding tumour heterogeneity through radiomics features," *Nature Communications*, vol. 5, p. 4871, 2014.

[27] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 315–323, 2011, jMLR W&CP vol. 15.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 448–456, 2015.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[30] N. Tajbakhsh, J. Shin, S. Gurudu, J. Dillman, P. Gader, S. Shenoy, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1244–1260, 2016.

[31] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *arXiv preprint arXiv:1902.07208*, 2019.

[32] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

[33] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 430–440.

[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 12, pp. 2504–2517, 2017.

[35] Q. Dou, H. Liu, C.-M. Leung, H. Qin, and L. Sun, "3d convolutional neural networks for automatic detection of pulmonary nodules in ct," *Medical Image Analysis*, vol. 36, pp. 108–120, 2017.

[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[37] J. H. Lee, H. Y. Sun, S. Park, H. Kim, E. J. Hwang, J. M. Goo, and C. M. Park, "Performance of a deep learning algorithm compared with radiologic interpretation for lung cancer detection on chest radiographs in a health screening population," *Radiology*, vol. 297, no. 3, pp. 687–696, 2020.

[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[39] O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2016, pp. 424–432.

[40] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, 2018.

[41] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, A. Turk, K. Bhatia, and B. Glocker, "Attention u-net: Learning where to look for the pancreas," in *Medical Image Computing and Computer Assisted Intervention (MIDL)*, 2018.

[42] A. P. Reeves and W. J. Kostis, "Computer-aided diagnosis of small pulmonary nodules," *Seminars in Ultrasound, CT and MRI*, vol. 21, no. 2, pp. 116–128, 2000.

[43] R. Gruetzemacher, A. Gupta, and D. Paradice, "3d deep learning for detecting pulmonary nodules in ct scans," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1301–1310, 2018.

[44] A. A. Farag, "Variational approach for small-size lung nodule segmentation," in *2013 IEEE 10th International Symposium on Biomedical Imaging*, 2013, pp. 81–84.

[45] Y. Boykov and I. Jolly, "An efficient algorithm for interactive graph cuts in energies of the form data+smoothness," *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 837–844, 2001.

[46] A. A. A. Setio, F. Ciompi, G. Litjens, P. K. Gerke, C. Jacobs, S. J. van Riel, M. M. Wille, and B. van Ginneken, "Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[47] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *Proceedings of the Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

[48] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, 2017.

[49] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*. Springer International Publishing, 2017, pp. 240–248.

[50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[51] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index," *Academic Radiology*, vol. 11, no. 2, pp. 178–189, 2004.

[52] J. Moorthy and U. D. Gandhi, "A survey on medical image segmentation based on deep learning techniques," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 117, 2022.

[53] D. P. Huttenlocher, R. M. Kedem, and M. Sharir, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[54] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, vol. 4, no. 1, pp. 1–13, 2020.

[55] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[56] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2198–2267, 1998.

[57] D. Scherer, A. Mueller, and S. Stiege, "Evaluation of pooling operations in convolutional networks for object recognition," *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, pp. 649–657, 2010.

[58] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 91–98, 2010.

[59] V. Dumoulin and J. Shlens, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[60] S. A. Taghanaki, K. Abhishek, J. T. Wong, D. Ravi, and G. Hamarneh, "Dice loss for data-imbalanced semantic segmentation," *arXiv preprint arXiv:1908.07187*, 2019.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[62] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[63] L. Prechelt, "Early stopping — but when?" pp. 55–67, 1998.

[64] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[65] G. N. Hounsfield, "Computerized transverse axial scanning (tomography): Part i description of system," *British Journal of Radiology*, vol. 46, no. 549, pp. 1016–1022, 1973.

[66] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[67] B. van Ginneken and C. Jacobs, "Luna16 part 1/2," Mar. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3723295

[68] B. van Ginneken and C. Jacobs, “Luna16 part 2/2,” Mar. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.4121926