الجمهورية الشعبية الديمقراطية الجزائرية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
جامعة سعد دحلب البليدة -1
Saad Dahleb University
Blida - 1

Master's Dissertation

To obtain the diploma of **Master's Degree**

Field of Study: **Computer Science**

Specialization: **Intelligent Systems Engineering (AI)**

Specialization: **Computer Networks and Systems Engineering**

Theme

A 3D AI system for automated detection of pulmonary infected regions from HRCT scans

---

Presented by
Chouchaoui Mohamed El Bachir
Douadi Fatma Zohra

Defended on:
In front of the jury composed of

Oukid Saliha	President of the Jury
Ferdi Imene	Examiner
Benyahia Mohamed	Thesis Supervisor
Benbelkacem Samir	Co-Supervisor

Academic Year: 2024/2025

Acknowledgements

We would like to express our deepest gratitude to all those who supported and guided us throughout the journey of completing this thesis. Every contribution, whether academic, technical, or moral, has been instrumental to our progress and success.

First and foremost, we extend our sincere thanks to our supervisors for their invaluable guidance, constructive feedback, and unwavering support throughout this work. Their insightful direction helped shape the foundations of this research, and their belief in our potential was a continuous source of motivation.

We are also grateful to the members of the CDTA research center for their collaboration, technical assistance, and the resources made available to us during the course of this project. Their commitment to scientific rigor and innovation greatly enriched our research experience.

We further extend our thanks to the jury members for taking the time to evaluate our work and provide their critical insights. Their comments and observations have been both encouraging, and we are truly honored by their contribution to this academic milestone.

Our appreciation also goes to the staff and medical professionals at Mustapha Bacha Hospital. Their openness, collaboration, and valuable contributions played a key role in the practical aspects of this study.

To all who supported, challenged, and encouraged us directly or indirectly please accept our heartfelt thanks. This achievement reflects not only our efforts but also the collective support and generosity of many.

Dedication

This thesis is dedicated, first and foremost, to my parents.
Your unwavering love, quiet strength, and unshakable belief in my potential have been my greatest inspiration. This achievement stands as a testament to your countless sacrifices and the foundation of support you built beneath every step I took.

To my dearest architect sister,
for being my constant source of strength, joy,
and motivation throughout this challenging journey.

A special tribute is reserved for my beloved grandmother.
Upon my college acceptance, she gifted me this laptop and said, "I don't know if I will live long enough to see your success journey, but I am glad to help you take the first step. I know you will do big stuff with it." Her faith lit a fire I carry with me to this day.

To my brilliant colleagues,
thank you for the shared late nights, the collaborative breakthroughs, and the camaraderie that made the most difficult days meaningful. We built each other up, and I am grateful for every moment we shared on this path.

To my professors,
your patience, insight, and dedication helped illuminate even the most complex concepts. Your guidance was a compass throughout this academic voyage.

And to everyone who saw something in me
who offered a kind word, extended a hand, or simply believed—thank you. Your quiet faith became a hidden force behind every accomplishment.

To myself:
for surviving the storms, for pushing through the doubts, for showing up every single day, and for never giving up — this one's for you. This is not the end, but the beginning.

Finally, a nod to the universe's grand design:
I'm here, I believe, thanks to a healthy dose of God's good humor. It's funny how we meticulously plan, only for life to unfold with an even better, divinely orchestrated agenda.

Hani

Dedication

To my dear family,
who have been the foundation of my life.
Your endless love, sacrifices, and constant belief
in my potential have guided me through every obstacle.
Without your encouragement, patience, and prayers,
this journey would not have been possible.
You have given me strength in moments of doubt
and reminded me of who I am when I felt lost.

To my professors,
whose guidance, patience, and dedication have shaped
my academic path. Your encouragement and high expectations
pushed me to reach deeper and aim higher. Thank you for sharing your
knowledge, for your valuable feedback, and for inspiring excellence.

To my colleagues,
with whom I shared challenges, ideas, and growth.
Your collaboration, support, and shared passion made
this experience richer and more meaningful.
The discussions, the laughter, and the hard work
will always be part of this achievement.

And to myself,
for showing resilience when things got tough,
for embracing the long nights, the quiet
frustrations, and the invisible battles,
and for continuing to move forward even
when the finish line felt far away.
You've grown, endured, and made it through.
This is a celebration of all the versions of
you who never gave up.

Fatma-Zohra

Abstract

Interstitial lung diseases (ILDs) are a diverse group of pulmonary disorders that cause progressive scarring and inflammation of the lung tissue. Accurate diagnosis is challenging due to the heterogeneous radiological appearance of lesions and the high volume of HRCT data requiring expert interpretation. This complexity underscores the need for automated systems capable of reliable and efficient ILD analysis.

This Dissertation presents a modular 3D deep learning system for the automated detection and classification of ILD-related lesions from high-resolution computed tomography (HRCT) scans. The pipeline is composed of three main stages: initial lung segmentation using a 3D U-Net model, binary lesion detection using a patch-based 3D CNN classifier (Simple3DCNN), and a second classification stage that performs multi-class lesion categorization across the most common lesions using a fine-tuned version of the same CNN architecture. To improve interpretability and trustworthiness of the system's predictions, 3D Grad-CAM (Gradient-weighted Class Activation Mapping) was applied to highlight salient regions influencing the model's classification decisions.

The system was trained and validated on a carefully preprocessed dataset, with patch sampling strategies, and class balancing techniques. The lung segmentation model achieved excellent results (Dice coefficient: 0.99, Hausdorff distance: 3.17), while the binary lesion detector reached high sensitivity (0.993) and accuracy (0.994). The multi-class classification stage achieved an overall accuracy of 88.73% and macro F1-score of 88.7% across most common lesion types. To validate spatial reasoning, Grad-CAM heatmaps were overlaid on the original HRCT patches, confirming the network's attention to clinically relevant regions and structures.

These results demonstrate the pipeline's ability to accurately detect and differentiate ILD lesions in 3D space, while also providing visual interpretability through Grad-CAM that can enhance clinical confidence in automated decision-making. The system provides a solid foundation for future integration into diagnostic workflows and further extension toward real-time, explainable AI tools in pulmonary imaging.

Keywords: Interstitial Lung Diseases (ILD), 3D Medical Image Segmentation, Pulmonary Lesion Classification, Deep Learning in Thoracic Imaging

Résumé

Les maladies pulmonaires interstitielles (ILD) sont un groupe hétérogène de troubles respiratoires chroniques qui posent d'importants défis en matière de diagnostic en raison de similitudes radiologiques entre les différentes formes. Ce mémoire propose un système intelligent en 3D pour la détection et la classification automatiques des régions pulmonaires infectées à partir d'images TDM à haute résolution (HRCT).

Ce mémoire présente un système modulaire d'apprentissage profond 3D pour la détection et la classification automatisées des lésions liées aux maladies pulmonaires interstitielles (ILD) à partir de scanners thoraciques haute résolution (HRCT). Le pipeline se compose de trois étapes principales : une segmentation initiale des poumons à l'aide d'un modèle 3D U-Net, une détection binaire des lésions à l'aide d'un classificateur 3D CNN basé sur des patches (Simple3DCNN), et une deuxième étape de classification multi-classes identifiant les lésions les plus courantes à l'aide d'un modèle ayant subi un réglage fin de la même architecture. Pour renforcer l'interprétabilité et la fiabilité des prédictions, la méthode Grad-CAM 3D (Gradient-weighted Class Activation Mapping) a été utilisée pour mettre en évidence les régions influençant les décisions du modèle.

Le système a été développé et évalué sur une base des données MedGIFT ILD, avec un échantillonnage par patchs et des techniques d'équilibrage des classes. Le modèle de segmentation pulmonaire a atteint d'excellents résultats (coefficient de Dice : 0.99, distance de Hausdorff : 3.17), tandis que le détecteur binaire de lésions a atteint une sensibilité de 0.993 et une précision de 0.994. La classification multi-classes a atteint une précision globale de 88.73% et un score F1 macro de 88.7% pour les types de lésions les plus courants. Pour valider la cohérence spatiale, des cartes de chaleur Grad-CAM ont été superposées aux patches HRCT d'origine, confirmant l'attention du réseau sur des zones pertinentes cliniquement.

Ces résultats démontrent la capacité du pipeline à détecter et différencier avec précision les lésions ILD dans un espace tridimensionnel. L'ajout de Grad-CAM permet une interprétabilité visuelle qui renforcer la confiance clinique dans l'IA. Ce système constitue une base solide pour une future intégration dans les flux cliniques et pour le développement d'outils de support décisionnel explicables en temps réel.

Mots-clés : Maladies pulmonaires interstitielles, Segmentation d'images médicales 3D, Classification des lésions pulmonaires, Apprentissage profond en imagerie thoracique

ملخص

أمراض الرئة الخلالية (ILDs) هي مجموعة متنوعة من الاضطرابات التنفسية المزمنة التي تشكل تحديات كبيرة في التشخيص بسبب تشابه المظاهر الشعاعية بين الأنماط المختلفة. يقترح هذا العمل نظامًا ذكيًا ثلاثي الأبعاد للكشف التلقائي وتصنيف المناطق الرئوية المصابة اعتمادًا على صور التصوير المقطعي المحوسب عالي الدقة (HRCT).

تقدّم هذه الأطروحة نظاماً معيارياً للحفظ العميق ثلاثي الأبعاد للكشف التلقائي وتصنيف الآفات المرتبطة بأمراض الرئة الخلالية (ILDs) اعتماداً على صور الأشعة المقطعية عالية الدقة (HRCT). يتكوّن خط المعالجة من ثلاث مراحل رئيسية: التقسيم الأولي للرئتين باستخدام نموذج 3D U-Net، ثم تصنيف ثنائي للآفات باستخدام مصنف ثلاثي الأبعاد قائم على المقاطع (Simple3DCNN)، متبوعاً بمرحلة ثانية لتصنيف الآفات متعددة الفئات تشمل أكثر الآفات شيوعًا باستخدام نسخة محسّنة من نفس البنية. لتعزيز تفسيرية النموذج، تم دمج تقنية Grad-CAM ثلاثية الأبعاد لتوليد خرائط حرارية توضح المناطق المؤثرة في قرارات النموذج.

تم تطوير النظام وتقييمه باستخدام مجموعة بيانات مُعالجة مسبقاً من قاعدة البيانات ، حيث تم اعتماد استراتيجيات لاستخلاص المقاطع، وتقنيات موازنة لتوزيع الفئات. حقق نموذج التقسيم نتائج ممتازة (معامل Dice = 0,99، مسافة Hausdorff = 3,17)، بينما سجّل المصنف الثنائي حساسية عالية (0,993) ودقة بلغت (0,994). أما في مرحلة التصنيف متعددة الفئات، فقد بلغ متوسط الدقة 88,73% ومتوسط F1-score حوالي 88,7% أنواع الآفات الأكثر شيوعًا. كما ساهمت خرائط Grad-CAM في تأكيد تركيز النموذج على المناطق الهامة سريريا.

تُظهر هذه النتائج قدرة النظام على الكشف والتصنيف الدقيق للآفات الرئوية في الفضاء ثلاثي الأبعاد، مع توفير تفسير نظري داعم لقراراته، مما يجعله قاعدة متينة لتكامل مستقبلي في سير العمل السريري وتطوير أدوات دعم قرار ذكية وعملية.

الكلمات المفتاحية: أمراض الرئة الخلالية، تقسيم الصور الطبية ثلاثية الأبعاد، تصنيف الآفات الرئوية، الحفظ العميق في التصوير الصدري

Acronyms

Abbreviation	**Definition**
3D U-Net	Three-Dimensional U-Net Architecture
AUC	Area Under the Receiver Operating Characteristic Curve
CNN	Convolutional Neural Network
Dice	Sørensen–Dice Coefficient
Dicom	Digital Imaging and Communications in Medicine
FN	False Negative
FP	False Positive
F1	F1 Score (harmonic mean of precision and recall)
GGO	Ground Glass Opacity
Grad-CAM	Gradient-weighted Class Activation Mapping
HRCT	High-Resolution Computed Tomography
HU	Hounsfield Unit
ILD	Interstitial Lung Disease
ROC	Receiver Operating Characteristic
TN	True Negative
TP	True Positive
XAI	Explainable Artificial Intelligence

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Interstitial lung diseases (ILDs) are a diverse group of chronic lung disorders, characterized by inflammation and fibrosis of the lung interstitium. With absence of timely treatment, some subtypes of ILD can cause progressive and permanent damage, resulting in severe respiratory disfunction and quality of life [1]. Timely diagnosis is important, as some ILDs may be amenable to treatment in their early stages, while advanced fibrotic changes are generally irreversible [2].

For the latter indication HRCT has developed as an essential non-invasive method for ILD, allowing detailed insight into the architecture of the lung [3]. Yet, accurate analysis of HRCT scans remains a clinical dilemma. Many subtypes of ILD have overlapping radiological features; subtle imaging findings including ground glass opacities and early fibrotic change can be overlooked [4]. In addition, interpretation of hundreds of axial slices for volumetric scans is very labor-intensive and is subject to inter-observer variability, particularly in ambiguous or mix-pattern [5].

In such a situation, AI and deep learning techniques are promising resources for assisting clinical decision-making [6], [7]. Recent developments of medical imaging reveal the enormous potential of Convolutional Neural Networks (CNNs) to automatically recognize disease patterns with high accuracy. Specifically, 3D CNNs are appropriate to deal with volume data such as HRCT, as they can take advantage of spatial coherence across slices. Moreover, incorporating interpretability techniques like Grad-CAM can contribute to improve trust and transparency, by indicating which image regions drive the AI-derived result.

This increasing intersection of AI and radiology reveals new prospects for creating reliable, interpretable, and clinically relevant tools for early ILD detection and classification lesions.

Although several previous studies [7]–[9] have employed 2D CNNs based on individual slices, and despite their potential advantages, they may neglect the full anatomical and pathological context. ILD lesions were usually across some slices and showed the complicated 3D structure. A 3D CNN would thus be more appropriate to capture these 3D patterns and could potentially take advantage of enhancing the lesion localization, delineation, and the classification performance.

High-Resolution Computed Tomography (HRCT) offers critical visual insights for the

diagnosis of interstitial lung diseases (ILDs). However, its interpretation relies heavily on radiologist expertise and the manual review of hundreds of axial slices per patient. This diagnostic process is time-intensive, subject to inter-observer variability, and may yield inconsistent results particularly in early stage or mixed pattern ILDs where radiological features are often subtle or ambiguous.

The heterogeneity of ILD manifestations, their radiological overlap with other pulmonary conditions, and the absence of standardized diagnostic ground truths further complicate accurate assessment. These challenges highlight the urgent need for automated, interpretable, and scalable solutions capable of reducing diagnostic variability, improving workflow efficiency, and supporting the reliable detection and classification of ILD lesions from volumetric HRCT scans.

This dissertation seeks to build a 3D deep learning based pipeline for automatic detection and classification of lesions related to ILD from HRCT scans. The specific objectives of the work are:

- To develop a 3D U-Net model for precise segmentation of lung regions on volumetric CT data.
- To build a lightweight binary classifier module to segmentation of healthy and pathological regions.
- To design and evaluate a multi-class classification phase for detection of common ILD patterns (ground-glass opacities, fibrosis, and reticulation).
- To incorporate Grad-CAM-based interpretability methods into the classification pipeline to increase transparency and aid in clinical decision-making.

The remainder of this dissertation is structured as follows:

- **Chapter 1** introduces the clinical and technical context and defines the dissertation scope and objectives.
- **Chapter 2** reviews the state of the art in lung and lesion segmentation, ILD classification.
- **Chapter 3** describes the proposed methodology, including dataset handling, preprocessing, and model design.
- **Chapter 4** outlines the experimental design, data partitioning strategy, and evaluation setup.
- **Chapter 5** presents the results of segmentation, binary detection, multi-class classification, and Grad-CAM analysis.
- **Chapter 6** summarizes the key findings and suggests directions for future research and clinical integration.

CHAPTER 2

STATE OF THE ART

2.1 Introduction

In the past decade, the advancement of artificial intelligence (AI) has significantly changed the scene of the medical image, providing a variety of promising utilities for disease detection, segmentation and classification [10].

In the context of interstitial lung disease (ILD), a complex group of disorders with overlapping radiological and clinical manifestations, HRCT has become the gold standard for non invasive diagnosis, but its interpretation is still subjective and greatly benefits from multidisciplinary discussion (MDD) [11].

This chapter initially reviews the clinical and radiological background of ILD, and then provides an overview of the current state of deep learning driven methods in lung segmentation and disease detection and classification.

2.2 Clinical background of interstitial lung diseases

To understand the clinical implications of interstitial lung diseases it is important to clarify their definitions and the subtypes that fall under this broad category.

2.2.1 Definitions and types

More than 100 disorders with varying degrees of pulmonary interstitium inflammation and fibrosis are collectively referred to as interstitial lung diseases (ILDs). The lung architecture is disturbed by interstitial lung diseases, which also affects gas exchange and cause progressive respiratory failure. ILDs appear radiologically as pattern on HRCT scans and clinically with cough, breathlessness, and reduction in lung volumes [12].

According to the American Thoracic Society (ATS) and the European Respiratory Society (ERS) [13], ILDs can be divided into a number of general types based on their etiology and pathogen A brief categorization is shown in Table 2.1.

Category	Description	Key Examples
Idiopathic Interstitial Pneumonias (IIP)	ILDs of unknown cause, categorized by histopathology and clinical features.	Idiopathic Pulmonary Fibrosis (IPF), Non Specific Interstitial Pneumonia (NSIP), Acute Interstitial Pneumonia (AIP)
Autoimmune related ILD	Occurs in association with connective tissue diseases.	Systemic sclerosis associated ILD, Rheumatoid arthritis associated ILD
Exposure related ILD	Caused by environmental or occupational inhalants, or drug induced injury.	Hypersensitivity Pneumonitis, Asbestosis, Chemotherapy induced ILD
Cystic and Airspace Filling ILDs	Characterized by cyst formation or alveolar filling abnormalities.	Lymphangioleiomyomatosis (LAM), Pulmonary Langerhans Cell Histiocytosis
ILDs Related to Systemic Diseases	Secondary to systemic inflammatory or granulomatous diseases.	Sarcoidosis, Vasculitis
Other and Unclassifiable ILDs	Rare or overlapping ILDs that do not fit existing classification criteria.	Chronic eosinophilic pneumonia, ILD associated with malignancy

Table 2.1: Major categories of interstitial lung diseases

This classification framework is not only essential for differential diagnosis, but also for guiding treatment and prognosis [13]. For example, idiopathic pulmonary fibrosis (IPF) historically has a progressive and irreversible course and is unresponsive to immunosuppression, while ILDs in the context of autoimmune diseases may be responsive to corticosteroids or biologics [1].

Grouping ILDs into etiologically homogeneous categories, the classification makes it easier to develop targeted artificial intelligence (AI) models that can distinguish between disease entities based on clinical and radiological data.

The distribution of ILD subtypes varies significantly across regions, influenced by genetic, environmental, and clinical practice factors. A prospective cohort study conducted in Algeria by Abdelbassat Ketfi et al. [14] analyzed 455 newly diagnosed ILD patients between 2015 and 2019. The study revealed that connective tissue disease-associated ILD (CTD-ILD) was the most common subtype, accounting for 48.1% of cases, followed by idiopathic interstitial pneumonias (IIPs) at 23.5%, sarcoidosis at 16.9%, interstitial pneumonia with autoimmune features (IPAF) at 12.1%, and hypersensitivity pneumonitis (HP) at 2.4%. Notably, idiopathic pulmonary fibrosis (IPF)—a dominant subtype in Western datasets—represented only 8.6% of cases. These findings underline the epidemiological heterogeneity of ILDs and

support the need for region-specific diagnostic frameworks and datasets in both clinical and AI-based research.

2.2.2 Clinical and radiological characteristics of ILDs

Interstitial Lung Diseases (ILDs) are related with critical horribleness and mortality. Common indications incorporate progressive dyspnea (shortness of breath), chronic cough, and work out intolerance. As the disease progresses, pulmonary work ordinarily declines, driving to prohibitive patterns characterized by decreased lung volumes and diffusing capacity. Early detection and exact diagnosis are vital for compelling management and to moderate disease progression. Notably, ILDs can result in aggravation and scarring of the lung tissue, driving to disabled lung work and decreased quality of life [15].

High-Resolution Computed Tomography (HRCT) is considered as the best imaging modality for the diagnosis and follow-up of interstitial lung diseases (ILDs) as it depicts fine parenchymal details and specific lung patterns. Different from normal CT scanning, HRCT uses thin slice image (usually 1–1.5mm) with high resolution and low motion artifacts, which facilitates to capture the subtle lesion in lung interstitium, like ground-glass opacity, reticulation, and honeycombing [16]. These patterns have a tendency to identify specific ILD subtypes and limit differential diagnoses [11].

Pattern	**Description**	**Common Associations**
Ground Glass Opacities (GGO) [17]	Areas of hazy increased attenuation with preserved bronchial and vascular structures.	Non Specific Interstitial Pneumonia (NSIP), Acute Interstitial Pneumonia (AIP), viral infections
Reticular Pattern [16]	A network of intersecting linear opacities due to interstitial thickening.	Idiopathic Pulmonary Fibrosis (IPF), Connective Tissue Disease associated ILD
Honeycombing [18]	Subpleural clustered cystic airspaces with shared walls, typically in basal zones.	IPF with a Usual Interstitial Pneumonia (UIP) pattern
Fibrosis [16]	Irreversible scarring of the lung parenchyma characterized by architectural distortion, volume loss, and often associated with traction bronchiectasis.	UIP, NSIP, Chronic Hypersensitivity Pneumonitis, Connective Tissue Diseases
Consolidation [19]	Homogeneous increased density that obscures vessels and bronchi.	Organizing Pneumonia, Infection, Malignancy
Micronodules [20]	Tiny nodules (< 4 mm), either perilymphatic, centrilobular, or random in distribution.	Sarcoidosis, Hypersensitivity Pneumonitis

Table 2.2: Common HRCT Patterns Observed in ILDs

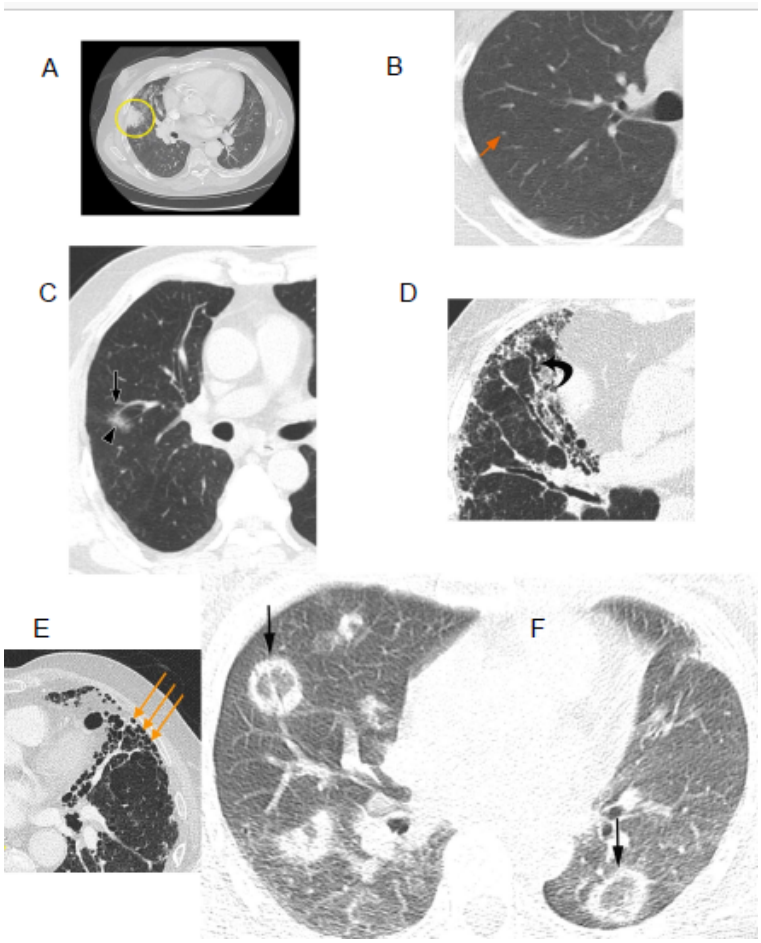

Figure 2.1: Representative HRCT patterns in ILD. A: Consolidation; B: Micronodules; C: Ground glass opacities; D: Reticular pattern; E: Honeycombing; F: Fibrosis.

Figure 2.1 illustrates representative HRCT patterns observed in ILD patients, including ground-glass opacities, reticulation, and fibrosis. These radiological patterns serve as key inputs for automated AI-based analysis, providing spatial and texture-based cues that aid in lesion classification and disease subtype differentiation.

2.2.3 Comparison with other pulmonary diseases

Accurate diagnosis of interstitial lung diseases (ILDs) is consistently difficult owing to the overlapping clinical and radiological features with other lung conditions. Some of the most commonly confused pathologies are COVID 19 pneumonia [21] and chronic obstructive pulmonary disease (COPD) [22], both of which may present with ground glass opacities, dyspnea, and impaired lung function. These conditions, however, significantly differ in etiology, radiological presentation, and disease course.

Table 2.3 compares ILDs, COVID 19 pneumonia, and COPD on various diagnostic features, including cause, clinical presentation, radiological findings on HRCT, progression, and treatment strategy.

Feature	ILDs	COVID 19 Pneumonia	COPD
Cause	Often idiopathic, or associated with autoimmune diseases or environmental exposure.	Viral infection caused by SARS CoV 2.	Chronic exposure to tobacco smoke or environmental pollutants.
Symptoms	Progressive dyspnea, dry cough, fatigue.	Fever, dry or productive cough, acute dyspnea, myalgia.	Chronic cough with sputum, exertional dyspnea, wheezing, recurrent lower respiratory infections (in late stages).
HRCT Findings	Ground glass opacities, reticulations, traction bronchiectasis, and honeycombing in subpleural and basal regions.	Bilateral, peripheral GGOs with or without consolidation; often lower lobe predominant.	Emphysema, bronchial wall thickening, air trapping; absence of fibrotic changes.
Progression	Chronic and usually irreversible; variable depending on subtype.	Acute to subacute course; may resolve or progress to post COVID fibrosis.	Chronic and progressive airflow limitation.
Treatment	Immunosuppressants, antifibrotics (e.g., pirfenidone), oxygen therapy, lung transplant in severe cases.	Supportive care, corticosteroids, antivirals, oxygen therapy (usually short term).	Bronchodilators, inhaled corticosteroids, pulmonary rehabilitation, long term oxygen therapy (LTOT) in severe cases.

Table 2.3: Comparative Clinical and HRCT Features of ILDs, COVID 19 Pneumonia, and COPD

It is of the utmost importance for both radiologists and diagnostic algorithms based on AI to be able to differentiate between these diseases. Misclassification, especially between ILDs and COVID 19 pneumonia or COPD, can lead to inappropriate treatment or overlooked diagnoses. As such, training datasets must be constructed to emphasize the subtle differences in both clinical and HRCT features in a bid to boost diagnostic accuracy and model generalizability.

2.3 ILD datasets

The development and evaluation of automated systems for interstitial lung disease (ILD) detection and classification require access to annotated datasets containing high-resolution computed tomography (HRCT) scans. While data availability remains a limiting factor in medical AI research, several publicly available datasets have been introduced in recent years to facilitate progress in this field.

One of the most widely referenced datasets is **the MedGIFT ILD Database**, developed by Depeursinge et al. It consists of HRCT scans from 113 patients diagnosed with various ILD subtypes. The dataset includes voxel-level lesion annotations grouped into six radiological patterns: consolidation, ground-glass opacity (GGO), reticulation, micronodules, and fibrosis. Each volume is accompanied by both lung segmentation masks and lesion-specific region of interest (ROI) masks, enabling training and evaluation of segmentation, detection, and classification models.

Another relevant resource is the **LTRC (Lung Tissue Research Consortium)** dataset, which provides HRCT scans from subjects with various lung diseases, including idiopathic pulmonary fibrosis (IPF). However, the LTRC dataset lacks detailed lesion annotations, which limits its use in supervised lesion-level classification tasks. Nonetheless, it can be employed for weakly supervised or volumetric-level prediction studies, especially those focused on fibrosis detection and lung function correlation.

Despite their importance, these datasets still present several limitations. Most suffer from limited diversity in disease subtypes, and variability in scan resolution. Moreover, publicly available ILD datasets often lack standardized diagnostic labels or ground truth established through multidisciplinary consensus, which can affect reproducibility and cross-study comparability.

Nevertheless, the availability of datasets like MedGIFT has catalyzed research in 3D segmentation and classification of ILDs.

2.4 Deep learning for ILD analysis

Deep learning has revolutionized the analysis of pulmonary images, especially for the detection and classification of Interstitial Lung Disease (ILD) patterns from High-Resolution Computed Tomography (HRCT) images. Unlike traditional image processing methods, deep learning models—particularly convolutional neural networks (CNNs)—provide a data-driven approach capable of capturing complex spatial and textural features relevant for lung disease assessment.

2.4.1 Lung segmentation studies

The development of CNN-based models has significantly improved the accuracy and robustness of lung segmentation in chest CT and HRCT. Encoder-decoder architectures with skip connections, such as U-Net and its variants, are commonly used to delineate lung boundaries, even in the presence of pathological artifacts.

- Alom et al. [23] introduced the R2U-Net model, where both recurrent and residual connections are integrated into the U-Net framework. The model enhances spatial feature learning and achieved a Dice coefficient of 0.981 on lung CT scans, outperforming standard U-Net variants.

- Jin et al. [24] proposed a 2.5D CNN approach that incorporates axial slices with minimal context from neighboring planes. Evaluated on the LIDC-IDRI dataset of over 1,000 patients, the model achieved a Dice score of 0.964, demonstrating the feasibility of hybrid 2.5D techniques for large-scale clinical segmentation tasks.

- Park et al. [25] trained their fully automated 3D U-Net on 196 volumetric chest CT scans from normal and mild-to-moderate COPD patients across three medical centers, and validated performance on an additional 40 external scans, achieving mean Dice scores of 0.97 ± 0.02 (internal) and 0.96 ± 0.02 (external), demonstrating robust performance and excellent generalization on both internal and external cohorts.

2.4.2 Binary ILD detection studies

Binary classification of HRCT images into healthy vs. pathological is a critical first step in automated ILD diagnosis. These methods are particularly useful for triage and screening in clinical workflows.

- Lu et al. [8] proposed a 2D patch-based CNN for ILD classification using a publicly available dataset. They achieved 85.5% accuracy in binary detection. However, their method lacked 3D context and used limited receptive fields, restricting its sensitivity to spatial dependencies.

- Silva de Araújo et al. [9] developed an ensemble model combining CNNs with radiomics-based multilayer perceptrons (MLPs) trained on in-house CT data. The CNN achieved 87.0% accuracy, while the ensemble yielded a slight improvement to 87.4%. Performance was affected by variability in acquisition protocols and disease manifestation.

2.4.3 Lesion segmentation studies

Lesion segmentation aims to isolate ILD-related patterns such as ground-glass opacities (GGOs), honeycombing, fibrosis, and reticulation. Precise lesion-level segmentation supports both quantitative analysis and targeted classification.

- Anthimopoulos et al. [7] used a 2D dilated CNN for patch-wise classification of ILD patterns. Their method achieved 85.5% accuracy in distinguishing between common ILD-associated features including GGOs and honeycombing.

- Wang et al. [26] proposed a cascaded dual U-Net framework, with the first network performing lung segmentation and the second responsible for lesion segmentation. Using the MedGIFT ILD dataset, their method achieved a Dice score of 0.78, illustrating the benefits of hierarchical localization.

- Zhang et al. [27] integrated attention gates into a 3D U-Net to enhance focus on lesion regions. On a dataset of 150 ILD cases, the model achieved a Dice coefficient of 0.82 for GGOs, outperforming baseline 3D models.
- Park et al. [28] combined 3D segmentation with clinical metrics to quantify fibrotic lesion volume. Their model achieved a Dice score of 0.89 across 40 patient scans, highlighting the utility of integrating visual outputs with quantitative analysis.

2.4.4 ILD classification studies

Subtyping ILDs is critical for treatment planning and prognosis. Deep learning models have increasingly been applied to multi-class classification of HRCT scans based on characteristic imaging features.

- Mei et al. [29] proposed a multimodal fusion framework that combines CNNs with Transformer-based models to classify ILD subtypes. Trained on a multicenter dataset of over 500 patients, their approach achieved an accuracy of 90.2%, showcasing the advantage of integrating local and global feature representations.
- Chassagnon et al. [10] trained a deep CNN on over 1,200 HRCT scans to classify ILD patterns into usual interstitial pneumonia (UIP) and non-UIP categories. Their system achieved 93% accuracy, performing comparably to expert radiologists.
- Walsh et al. [30] conducted one of the first large-scale comparisons between deep learning classifiers and radiologist agreement. The model yielded a Cohen's kappa of 0.74, comparable to the inter-observer variability among human experts, validating the potential of AI-assisted ILD diagnosis.

2.4.5 Summary of the studies

The following table 2.4 provides a comprehensive overview of the key studies reviewed in this chapter, summarizing their methodological approaches, datasets used, and performance metrics achieved. This comparison highlights the current state of research across different aspects of ILD analysis, from lung segmentation to ILD classification, and demonstrates the evolution of techniques in this field.

Study	Category	Method	Dataset	Performance
Alom et al. [23]	Lung Segmentation	R2U-Net (Recurrent + Residual U-Net)	Lung CT scans	Dice: 0.981
Jin et al. [24]	Lung Segmentation	2.5D CNN (axial slices with context)	LIDC-IDRI (1,000+ patients)	Dice: 0.964
Park et al. [25]	Lung Segmentation	3D U-NET	196 CT scans (internal), 40 external scans	Dice: 0.97 (internal), 0.96 (external)
Lu et al. [8]	Binary Detection	2D patch-based CNN	Public ILD dataset	Accuracy: 85.5%
Silva de Araújo et al. [9]	Binary Detection	CNN + Radiomics MLP ensemble	In-house CT data	Accuracy: 87.0%, Ensemble: 87.4%
Anthimopoulos et al. [7]	Pattern Classification	Dilated CNN (patch-wise)	ILD patterns dataset	Accuracy: 85.5%
Wang et al. [26]	Lesion Segmentation	Two-stage cascaded U-Net	MedGIFT ILD dataset	Dice: 0.78
Mei et al. [29]	Subtype Classification	Multimodal CNN + Transformer fusion	Multicenter (500+ patients)	Accuracy: 90.2%
Chassagnon et al. [10]	Subtype Classification	Deep CNN	1,200 HRCT images	Accuracy: 93%
Walsh et al. [30]	Subtype Classification	Deep CNN vs expert radiologists	Multicenter test dataset	Cohen's κ: 0.74

Table 2.4: Summary of studies in ILD Analysis

2.5 Methodological foundations in deep learning for medical image analysis

This section outlines the core methodological principles and architectural choices that underpin recent advancements in AI-based ILD analysis. The objective is to provide technical justification for the models and strategies adopted in this dissertation, particularly in relation to network dimensionality, segmentation paradigms, pipeline design, and explainability mechanisms.

2.5.1 Comparative analysis of 2D and 3D CNNs in thoracic CT imaging

Convolutional neural networks (CNNs) have been instrumental in medical image analysis, with 2D CNNs historically dominating early research[31]. These networks process single axial slices independently and offer lower computational overhead, making them suitable for classification tasks on slice-level annotations. However, 2D approaches suffer from key limitations in ILD diagnosis, notably the inability to leverage the spatial continuity between slices[32], [33]. Lesions such as ground-glass opacities or reticulations frequently span multiple adjacent planes, and processing them in isolation leads to inconsistent localization and decreased diagnostic reliability.

To address these shortcomings, 3D CNNs have emerged as a compelling alternative. These models operate on volumetric inputs, preserving contextual information along all three anatomical axes. As a result, 3D CNNs demonstrate superior performance in segmenting irregular lesions and capturing subtle inter-slice patterns. Studies such as Zhang et al. [27] and Park et al. [28] affirm that 3D models not only enhance segmentation accuracy but also support volumetric quantification vital to clinical workflows. Despite increased computational demand, their ability to model spatial dependencies makes them particularly suitable for ILD analysis.

Recent studies have demonstrated that 3D convolutional neural networks (CNNs) significantly outperform their 2D counterparts in volumetric chest imaging tasks by capturing richer spatial context across slices. For example, Alebiosu et al. [34] evaluated both 2D and 3D CNNs on the ImageCLEF 2021 dataset to classify high versus low tuberculosis severity on chest CT scans. The 3D CNN achieved an outstanding classification accuracy of 99.29% and an AUC of 0.9982, significantly surpassing the 2D model's performance. These findings underscore the advantage of 3D architectures in thoracic disease quantification and further justify their application in ILD analysis.

The following Figure 2.2 Adapted from [35] represents a schematic workflow of the 2D and 3D CNN models based on Xception. A: For 2D CNN, a single shoulder slide was the input, while 2D convolution layers were utilized to extract image features. Finally, 2048 features were extracted and fed into a classifier, from which the output was the probabilities of tear and normal. B: For the 3D CNN model, 3D shoulder image blocks were the input, and 3D convolution layers were utilized to extract image features. Finally, 2048 features were extracted and fed into a classifier, from which the output was the probabilities of tear and normal. CNN, convolutional neural network.

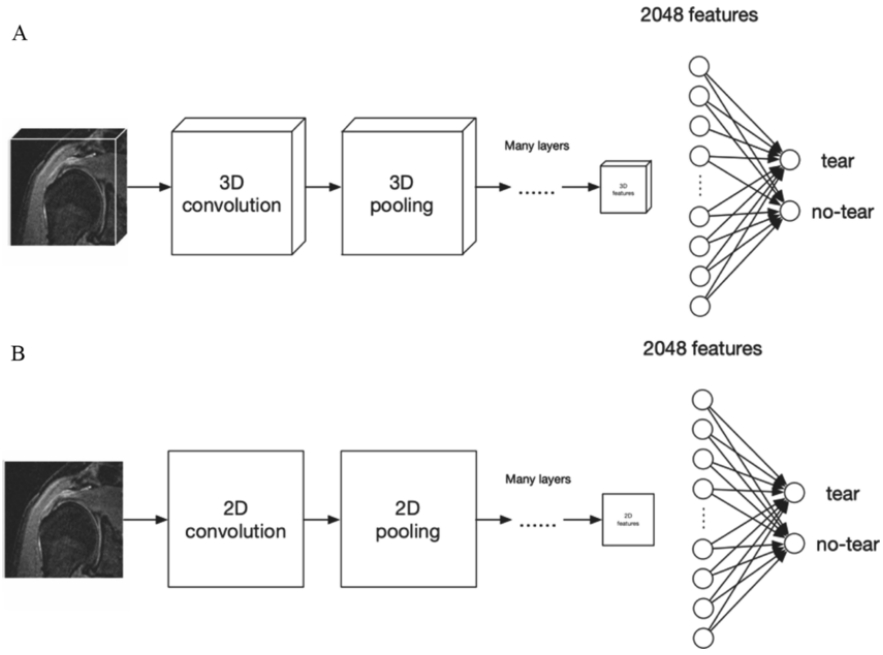

Figure 2.2: Comparison of 2D and 3D convolutional neural network pipelines for volumetric medical image segmentation.

2.5.2 U-Net and its variants

The U-Net architecture [36] has emerged as the prevailing approach in biomedical image segmentation due to its effectiveness in learning spatial features from limited annotated data. The architecture features a symmetric encoder-decoder structure, with skip connections that fuse coarse and fine feature maps, thereby enhancing localization precision.

Numerous adaptations of U-Net have been developed to improve its robustness and flexibility. R2U-Net [23], for example, integrates residual and recurrent layers to enhance the network's ability to capture spatial dependencies and reduce vanishing gradients. Attention U-Nets [37] incorporate gating mechanisms to focus on salient regions, thereby improving sensitivity to lesion boundaries. These variants have shown promising results in both lung and lesion segmentation tasks and are well-suited for complex, multi-label problems such as ILD assessment.

2.5.3 Multi-stage AI pipelines

A multi-stage pipeline decomposes a complex diagnostic task into sequential, functionally distinct modules commonly including lung segmentation, lesion detection, and multi-class classification. This modularity offers several advantages. First, it allows for stage-wise training and evaluation, facilitating targeted error analysis and fine-tuning. Second, it enhances the explainability and interpretability of model decisions by providing intermediate outputs that are clinically meaningful.

For ILD analysis, such multi-stage designs have proven effective in both research and clinical settings. Wang et al. [26] proposed a cascaded U-Net framework that sequentially segments lung fields and lesions, improving focus on disease-relevant regions. Similarly, our dissertation employs a modular architecture that first delineates anatomical structures

before proceeding to pathology-specific analysis. This ensures that computational resources and model attention are concentrated on actionable areas of the scan.

2.5.4 Explainable AI in medical imaging

The opacity of deep neural networks remains a barrier to clinical adoption, making explainability a critical requirement in medical imaging applications. Explainable AI (XAI) methods aim to elucidate the decision-making process of neural networks by highlighting input regions that influence specific outputs. Grad-CAM (Gradient-weighted Class Activation Mapping) [38] is one of the most widely adopted techniques for CNN visualization. It generates heatmaps that localize discriminative regions, offering intuitive insights into the model's reasoning.

In the context of ILD, XAI tools can help radiologists verify whether a model focuses on pathologically relevant structures, such as fibrotic streaks or GGOs [33]. While some studies question the alignment between saliency maps and expert judgment [39], Grad-CAM remains a practical and interpretable method for post-hoc analysis [38]. In this dissertation, we integrate Grad-CAM visualizations into the classification pipeline to enhance trust, auditability, and clinical usability.

Together, these methodological foundations inform the architectural choices and evaluation criteria in our proposed system for automated ILD detection and diagnosis.

2.6 Research gaps and our positioning

To motivate our methodology, we begin by outlining the main limitations in existing research on ILD segmentation and classification.

2.6.1 Limitations in the literature

Despite notable progress, the reviewed literature reveals several critical limitations that hinder the development of comprehensive ILD analysis systems:

- **Limited 3D volumetric context:** Many existing approaches rely on 2D slice-wise or patch-based analysis, which fails to capture the full volumetric continuity of lung structures and three-dimensional lesion patterns essential for accurate ILD characterization.
- **Lack of multi-class lesion classification:** Current research predominantly focuses on binary classification (healthy vs. pathological) or ILD subtype classification, with limited attention to detailed lesion-level pattern classification that distinguishes between different ILD manifestations such as ground-glass opacities, fibrosis, and reticulation.
- **Dataset and evaluation limitations:** Publicly available ILD datasets are relatively small, often class-imbalanced, and inconsistently labeled across studies, making it difficult to train robust, generalizable models and conduct meaningful performance comparisons.

- **Underrepresentation of modular pipelines:** There is a lack of integrated systems that combine lung segmentation, lesion detection, and classification in a unified framework, with most studies addressing individual components in isolation.
- **Limited clinical explainability:** Despite achieving high performance metrics, few systems integrate visual explainability mechanisms necessary for clinical trust and adoption in real-world medical settings.

2.6.2 Justification of our approach

To address these identified limitations, our work proposes a comprehensive modular 3D pipeline designed to reflect clinical reasoning and practical constraints:

- **Modular 3D architecture:** We implement a three-stage pipeline comprising lung segmentation, binary lesion detection, and multi-class lesion classification, enabling systematic processing of volumetric HRCT data while maintaining clinical workflow logic.
- **Volumetric processing with 3D CNNs:** Our approach utilizes 3D patch-based processing and U-Net architectures, enabling the model to leverage full volumetric spatial context while maintaining computational feasibility for clinical deployment.
- **Integrated explainability:** The system incorporates Grad-CAM-based visual interpretability throughout the classification pipeline, providing clinicians with transparent insights into model decision-making processes.
- **Comprehensive evaluation framework:** The system is evaluated on the standardized MedGIFT ILD dataset using robust clinical metrics, with realistic data partitioning strategies designed to simulate clinical inference scenarios and ensure reproducible results.
- **Clinical workflow consideration:** The modular design allows for flexible deployment, where individual components can be used independently or as part of the complete pipeline, depending on clinical requirements and computational constraints.

This approach integrates segmentation, detection, and classification into a cohesive framework that addresses multiple identified gaps while maintaining clinical relevance and practical applicability.

2.6.3 Contribution Outline

Based on the gaps identified and the proposed solutions, our main contributions are summarized as follows:

1. **3D lung segmentation module:** A robust 3D U-Net-based pipeline for accurate segmentation of lung regions from volumetric HRCT scans, providing the foundation for subsequent lesion analysis.

2. **Binary ILD detection system:** A lightweight binary classification module capable of distinguishing between healthy and pathological lung scans, enabling efficient clinical triage and pre-screening workflows.

3. **Multi-class lesion classification framework:** A comprehensive system for identifying and classifying specific ILD lesion patterns (ground-glass opacities, fibrosis, reticulation) within segmented lung regions, addressing the gap in detailed lesion-level analysis.

4. **Explainable AI integration:** Implementation of Grad-CAM-based visual interpretability mechanisms throughout the classification pipeline, enhancing clinical trust and providing actionable insights for radiological decision-making.

5. **Comprehensive evaluation and validation:** Systematic evaluation of each component using standardized metrics, providing benchmarks for future research and demonstrating clinical applicability.

2.7 Conclusion

The review indicates that there has been substantial progress in using AI for ILD detection and segmentation, but there are still some issues where this type of advancement has not been completed. The majority of the available works are 2D based and non-transferable across datasets or do not handle class imbalance well. Moreover, many pipelines do not make full use of the 3D spatial information contained in lung volumes. These shortcomings motivate the construction of a modular, 3D deep learning pipeline that is presented in this dissertation.

CHAPTER 3

MATERIALS AND METHODOLOGY

3.1 Introduction

This chapter introduces a brief methodology of the proposed system. It describes the pipeline developed for automated detection and classification of ILD based on HRCT scans. The strategy consists of pre-processing, data loading, lung segmentation with 3D U-Net, and two stage lesion analysis pipeline with 3D CNN and Grad-CAM [40], [41].

The models were designed for volumetric data and compensated for the class unbalance with advanced sampling and training procedures.

3.2 Dataset description

In this dissertation, the database used is the publicly available Interstitial Lung Disease (ILD) database that was put together by Professor Adrien Depeursinge and his team in HES SO, Valais, Switzerland. It's also known as **MedGIFT ILD dataset** [42], and is one of the largest public datasets for analyzing lung disease patterns on high resolution computed tomography (HRCT) images. The dataset consists of anonymized scans for a total of **113 patients**, each annotated by expert radiologists using both lung segmentation masks and voxel wise lesion labels.

The dataset was officially obtained on **May 6, 2025**, following permission for academic use and in accordance with the agreement established with the **CDTA research center**.

3.2.1 MedGIFT ILD dataset overview

Each patient scan is provided in the `DICOM` format. The imaging data was acquired using standardized HRCT protocols, with an in plane resolution of 512×512 pixels and variable numbers of axial, sagittal and coronal slices, depending on patient anatomy.

The following Figure 3.1 illustrates a representative HRCT scan from the ILD_DB dataset, highlighting the multiplanar views used for visual analysis and segmentation of interstitial lung disease patterns

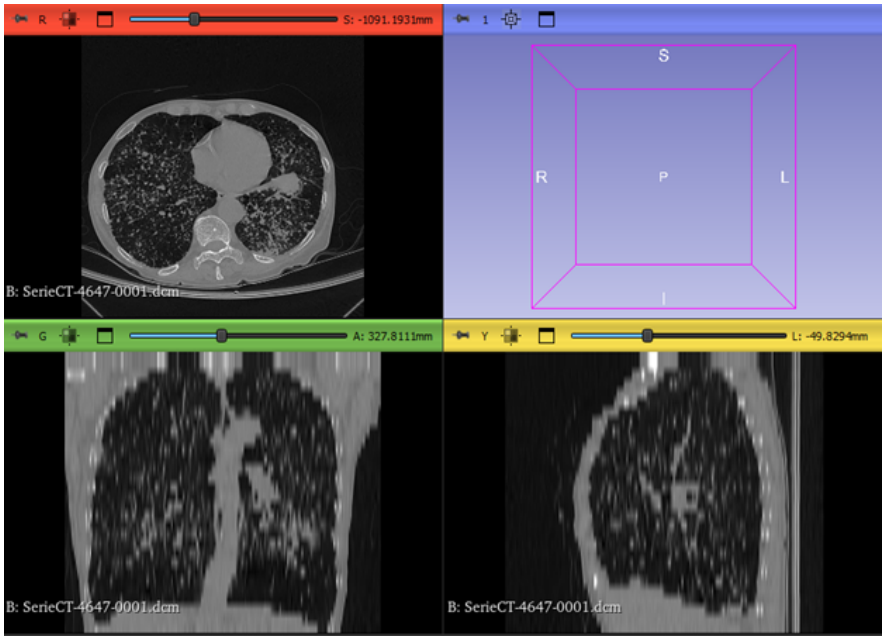

Figure 3.1: Example of an HRCT scan from the ILD DB dataset showing axial, sagittal, and coronal views with evident interstitial lung abnormalities, visualized using 3D Slicer.

3.2.2 Annotation types

There are three major classes of annotations in the MedGIFT dataset:

- **Lung masks**:

  Binary masks isolating the left and right lung fields were included in the *ILD_DB_lungMasks* directory. The masks are ground truth for the training of lung segmentation networks.

- **Lesion annotations**:

  Multiclass voxel level annotations are available in the *ILD_DB_volumeROIs* folder. Annotations for each annotated volume represent up to 17 pathological classes frequently found in ILD, including ground glass opacities, fibrosis, emphysema, and micronodules. The complete list of class indices and corresponding conditions is as follows:

  - 1 = Healthy
  - 2 = Emphysema
  - 3 = Ground glass
  - 4 = Fibrosis
  - 5 = Micronodules
  - 6 = Consolidation
  - 7 = Bronchial wall thickening
  - 8 = Reticulation
  - 9 = Macronodules
  - 10 = Cysts
  - 11 = Peripheral micronodules
  - 12 = Bronchiectasis
  - 13 = Air trapping
  - 14 = Early fibrosis
  - 15 = Increased attenuation
  - 16 = Tuberculosis
  - 17 = Pneumocystis pneumonia (PCP)

- **Text based ROI labels**:

  In the ILD_DB_txtROIs folder, the region of interest labels are available as label coordinate triplets. MATLAB and Java parsers come along with the dataset to facilitate easy integration into custom workflows.

3.2.3 Clinical metadata

Along with the segmentation and imaging data, the dataset also contains an Excel spreadsheet with relevant clinical information on all patients. These variables consist of age, sex, smoking history, existing comorbidities, and treatment plans. The availability of this extra information is useful when designing future studies involving multimodal AI systems or patient specific disease modeling.

3.2.4 Dataset summary analysis

To further understand the nature of the dataset, we conducted an in depth dataset analysis, as presented in Figure 3.2. This informed key pre-processing steps, model architecture, and evaluation strategies.

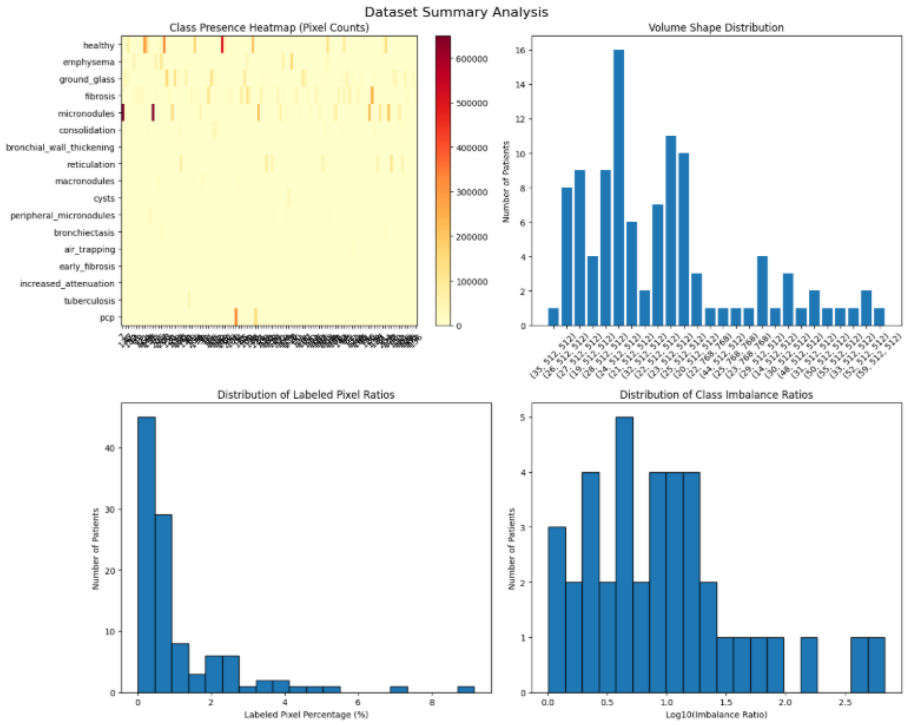

Figure 3.2: Dataset Summary Analysis: Top left: Class presence heatmap across patients; Top right: Distribution of volume shapes; Bottom left: Ratio of labeled pixels per patient; Bottom right: Log scaled class imbalance ratios.

Main findings:

1. **Class distribution:** The heatmap of class occurrence shows an imbalanced distribution where classes like *healthy*, *ground _glass*, *fibrosis*, and *micronodules* dominate. Rare classes like *tuberculosis* and *PCP* are either absent or represented showing an extreme class imbalance.

2. **Volume shape variability:** The variability in volume size calls for shape standardization as a pre-processing procedure. The majority of the volumes are concentrated around the $(51, 512, 512)$ shape with major outliers.

3. **Sparse lesion annotations:** As can be seen from the labeled pixel percentage histogram, lesions usually account for less than 1% of the lung region. This sparsity motivates the application of patch based and attention mechanisms for training models.

4. **Imbalance ratios:** The imbalance ratio (in log scale) ranges from about 3:1 to 300:1. This skew requires the use of robust methods such as weighted loss functions, oversampling, and data augmentation.

3.3 Understanding and pre-processing 3D medical image data

The application of deep learning to 3D medical imaging, and thoracic radiology in particular, requires from the input data not only quality but also a stable, well structured pre-processing pipeline. HRCT scans are, by definition, inhomogeneous in their spatial resolution, acquisition protocol, and format. For consistency and improved model performance, the Geneva HRCT dataset has undergone a series of pre-processing steps to unify its structure without compromising clinically relevant information [43]. This section outlines both the imaging modalities and the pre-processing pipeline used in this study.

3.3.1 Understanding 3D imaging modalities

DICOM (Digital Imaging and Communications in Medicine)[44] is the clinical radiology standard format. All CT scans are saved as a series of axial, sagittal and coronal slices with metadata that specify voxel spacing, orientation, slice thickness, and others. Although DICOM is suitable for clinical viewing and storage, it is not ideal for machine learning pipelines.

For this purpose, all DICOM series were transformed into the **NIfTI** format (`.nii.gz`)[45], a format that retains 3D or 4D volumes within one file and also keeps spatial metadata like affine transformations. Conversion was carried out with `SimpleITK`, with anatomical correctness in all three axes.

3.3.2 Pre-processing pipeline

The pre-processing pipeline involved several steps to normalize the spatial resolution, the intensity distributions, and to enable thedata to work with the 3D convolutional networks. These steps are elaborated upon below.

3.3.2.1 Pixel value transformation

An important pre-processing step involved transforming the pixel intensity values to Hounsfield Units (HU) a standardized scale for quantifying radiodensity in CT imaging.

This Figure 3.3 shows that the unit of measurement in CT imaging is the Hounsfield Unit (HU), which is a measure of radiodensity. The CT scanner is carefully calibrated to accurately measure this. From Wikipedia:

Substance	HU
Air	−1000
Lung	−500
Fat	−100 to −50
Water	0
CSF	15
Kidney	30
Blood	+30 to +45
Muscle	+10 to +40
Grey matter	+37 to +45
White matter	+20 to +30
Liver	+40 to +60
Soft Tissue, Contrast	+100 to +300
Bone	+700 (cancellous bone) to +3000 (cortical bone)

Figure 3.3: The unit of measurement in CT scans is the Hounsfield Unit (HU)

This figure 3.4 presents the plot on HU scale, and image data of that particular CT slice. The left panel shows a histogram of the voxel values in Hounsfield Units (HU), with annotated important tissue ranges such as air, soft tissue, and bone. Right: Preview of the axial CT slice, where normal thoracic anatomy is seen and lung parenchyma and surrounding structures have a good contrast.

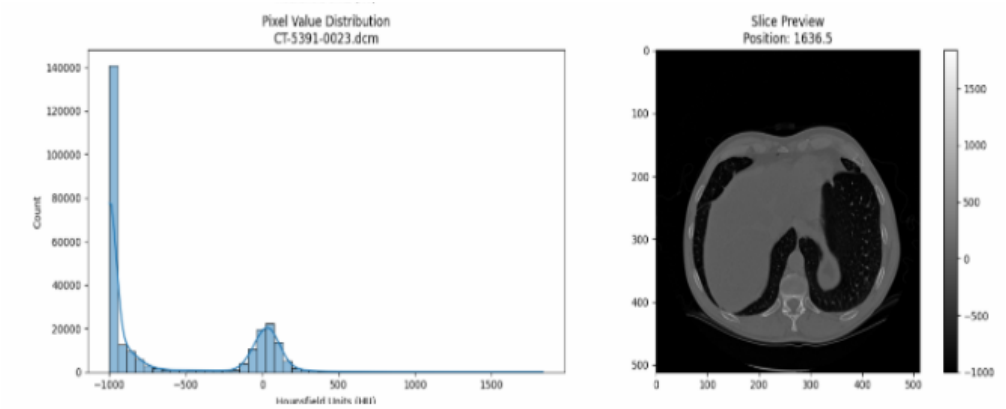

Figure 3.4: Pixel intensity distribution (left) and corresponding CT slice preview (right) from a DICOM scan. The histogram reveals typical lung and soft tissue contrast in Hounsfield Units (HU).

Raw pixel values were transformed using the DICOM metadata (Rescale Slope and Rescale Intercept), based on the formula:

$$HU = (\text{PixelValue} \times \text{RescaleSlope}) + \text{RescaleIntercept} \tag{3.1}$$

The conversion from raw pixel values to Hounsfield Units (HU) are computed from two DICOM metadata inputs as parameters: the **Rescale Slope** and the **Rescale Intercept**. These are used in the linear transformation in Equation (3.1).

- **PixelValue**: the gray level raw intensity of the CT scan, which usually lies between 0 and 4095 (12 bits).
- **Rescale Slope**: a scaling factor to rescale pixel value magnitude (typically 1)
- **Rescale Intercept**: a bias added to the scaled value, generally negative, in order to position the scaled value within the Hounsfield range.
- **HU**: the standardized value used for representing the density of tissue in CT slices.

After conversion, voxel intensities were windowed into [-1000, 400] HU in order to emphasize the lung regions and diminish the signal from unrelatively high density anatomies such as bones. To ensure numerical stability when training the neural network, the intensities were normalized to [0, 1].

3.3.2.2 Inconsistent pixel spacing and slice area

In the dataset, one of the main issues was that pixel spacing and slice area were not consistent between different CT images. 71 Spacing of voxels covered approximately 0.4 to 0.9 mm,and slice sizes varied between 0.17 mm^2 and 0.87 mm^2. Such inconsistency can lead to distortions in feature learning, as the same anatomical features may appear at different scales.

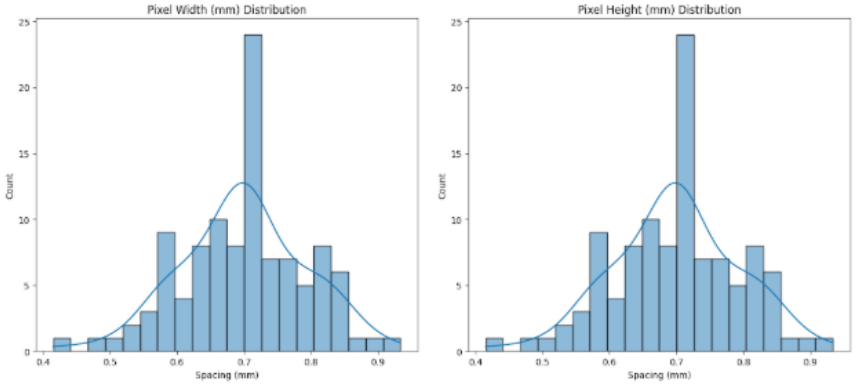

Figure 3.5: Distribution of in-plane resolution across the dataset.

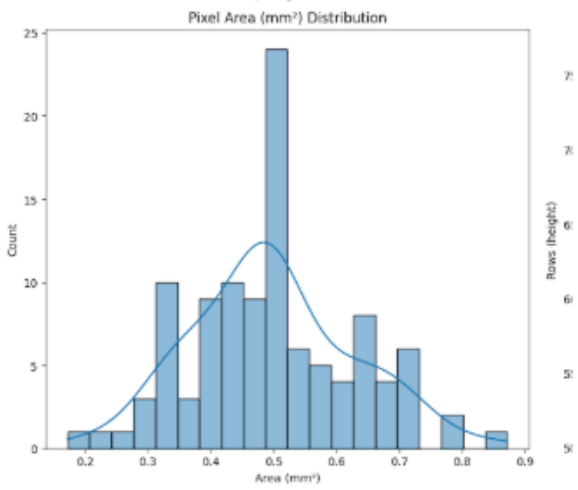

Figure 3.6: Distribution of pixel area (in mm^2) across the CT scans in the dataset.

In order to ensure this, all CT volumes and the respective paired lung and lesion masks were resampled to a common isotropic voxel size of 1.0 mm × 1.0 mm × 1.0 mm using trilinear interpolation on image data, and nearest neighbour interpolation on masks. This ensured evenly distributed spatial resolution, which is essential for effective and unbiased 3D CNN processing.

3.3.2.3 Varying in plane dimensions

Additionally, beside voxel spacing variation, the dataset was also variable in the in plane size of axial slices. The vast majority (approximately 97%)of scans were at 512×512 pixels and a minority (3%) were slightly larger at 768×768 pixels.

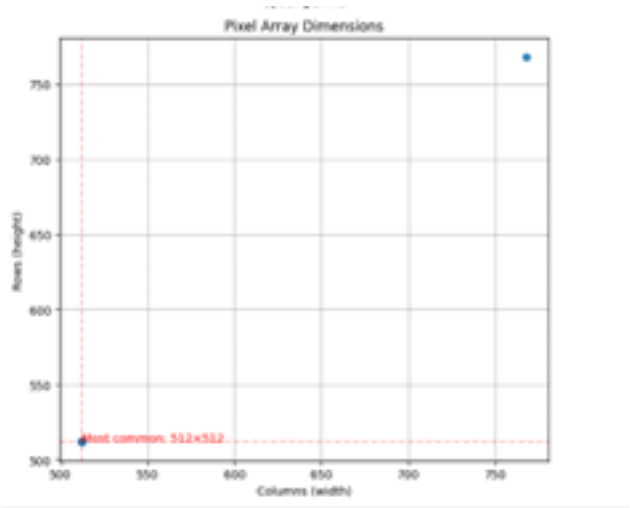

Figure 3.7: Standardizing in plane resolution.

All the 768×768 scans were down sampled to 512×512 using bilinear interpolation in order to ensure uniformity and to make them compatible with the input requirements of deep learning. It then preserved the anatomical integrity of the constructions and removed size based variability.

3.3.2.4 Variable scan depth (Z Dimension)

The depth of the Geneva HRCT scans were also inconsistent and the number of axial slices per scan varied significantly between patients. Scan depths per organ varied approximately between 5 mm and 25 mm such that emerged volumes showed strongly different Z dimensions.

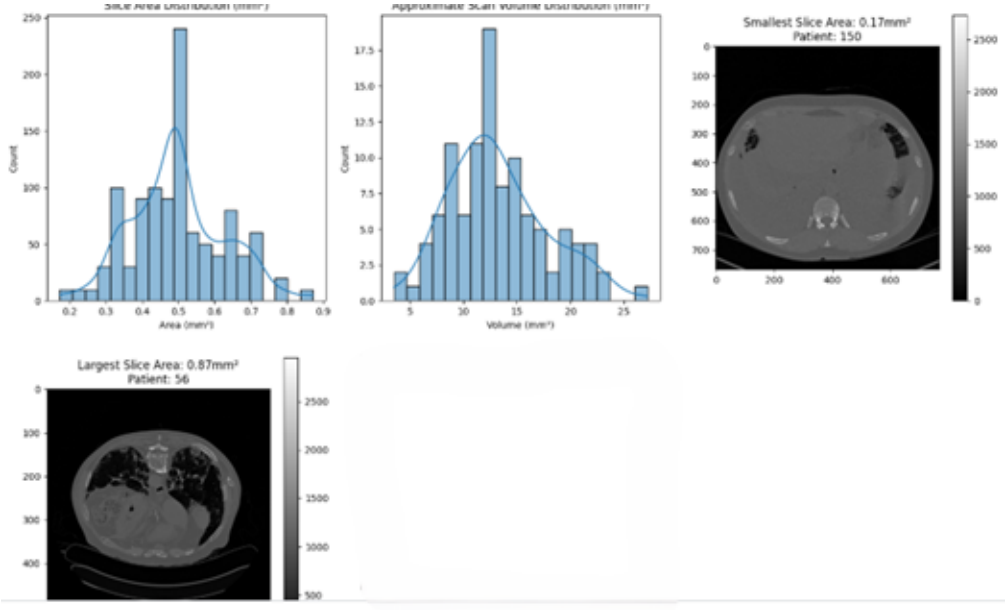

Figure 3.8: Patch based volume extraction strategy.

Patch based training was used instead of aggressive resizing of entire volumes (which might cause the loss of anatomical or pathological information). In particular, fixed size 32×320×20 voxel 3D patch were extracted from all the resampled volumes. This ensured you got both standardized input sizes and maintenance of clinically important spatial relationships.

3.3.2.5 Data augmentation

In order to add more strength and less overfits, data augmentation was performed during training by utilizing the `TorchIO` library. All transformations were applied in a spatially consistent fashion to both images and their masks. The augmentations were:

- Random flippings about sagittal and axial planes
- Minor rotations (±10 degrees)
- Elastic deformations
- Gaussian noise and intensity scaling

These augmentation mimicked scanner noise and anatomical variability, promoting model learning of more generalized features.

3.3.2.6 Export to NIfTI format

The pre-processed CT volumes and their corresponding lung and lesion masks were saved as the `. nii. gz` (compressed NIfTI) files. It is an appropriate format for volumetric medical images as it maintains important metadata (eg, voxel spacing, image orientation, affine transformations). Moreover, NIfTI files are well known format and they are largely supported by 3D medical imaging toolkits such as MONAI [46], TorchIO [47] and other deep learning frameworks, thus can be easily integrated into AI pipelines.

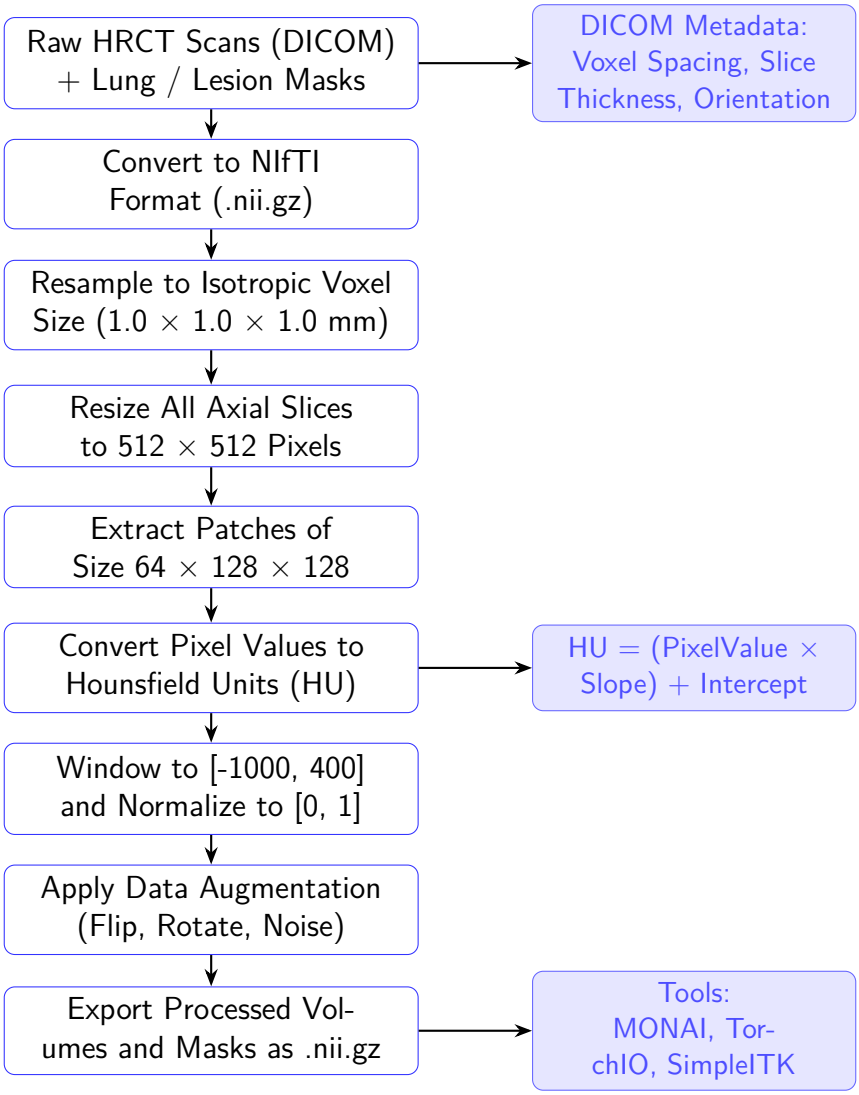

Figure 3.9: Overview of the pre processing pipeline applied to HRCT scans.

3.4 Model Architecture and Design

In this section, we describe the deep learning architectures used for lung segmentation, binary lesion detection, multi-class classifier and interpretability.

3.4.1 Lung segmentation model

Precise segmentation of the lung field is an important step in our pipeline. It guarantees that later processes, like lesion detection and classification, only operate within anatomically meaningful areas. This lowers the risk of extraneous false positives outside the lung regions and enhances overall model focus and efficiency.

- **Method overview**

  The 3D U-Net [32] is an encoder–decoder architecture where the encoder successively encodes the higher level features and the decoder is employed to regain the spatial resolution.

  Every encoding level features two $3 \times 3 \times 3$ convolutions with ReLU activation, and a $2 \times 2 \times 2$ max pooling.

  The decoder pathway is symmetrical with skip connections, transposed convolutions, and ReLU activations.

  Batch normalization is used prior to every activation function to enhance training stability.

  A final $1 \times 1 \times 1$ convolutional layer projects features onto the desired number of output channels.

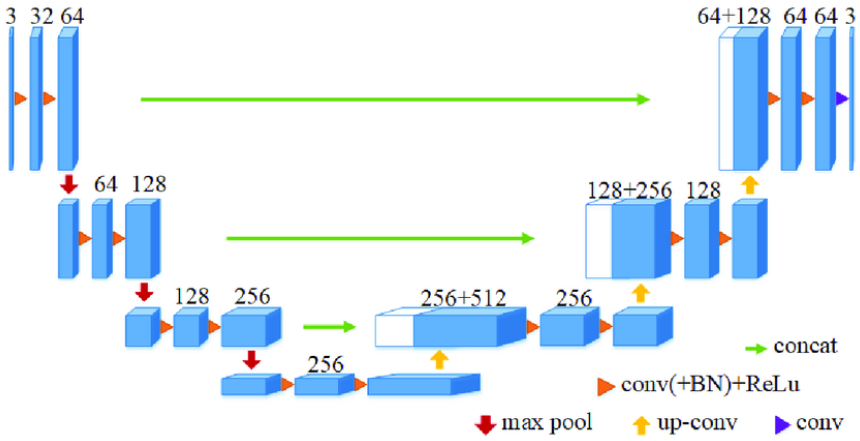

Figure 3.10: The 3D U Net Architecture. Adapted from [32].

This architecture allows for memory efficient compact voxel wise predictions, which is an extremely good fit for our 3D medical image problem.

- **Implementation in our pipeline**

  Our 3D U-Net is composed of:

  1. **Input:** pre-processed and masked HRCT lung volumes or patches ($128 \times 128 \times 128$).
  2. **Encoder:** 3D convolutional blocks with ReLU, followed by max pooling.
  3. **Bottleneck:** Deep convolutional layers extracting semantic abstraction.
  4. **Decoder:** Upsampling through transposed convolutions and skip connections to restore spatial details.
  5. **Output:** A $1 \times 1 \times 1$ convolution with softmax or sigmoid depending on the problem (multi class or binary).

The lung segmentation module follows the same standard 3D U-Net common architecture, ending with sigmoid activation to produce binary lung masks. The masks crop or mask the HRCT volumes and act as spatial filters for subsequent modules like lesion detection and classification.

3.4.2 Binary lesion detection model

The objective of the first stage in the pipeline is to determine whether an HRCT volume exhibits pathological signs related to interstitial lung disease (ILD). This task is framed as a binary classification problem, where each volume is classified as either `Healthy (0)` or `Pathological (1)`.

Input format:

The model processes 3D patches of size $32 \times 32 \times 32$ voxels, extracted from segmented lung regions. Patches are normalized to the range $[0, 1]$ and resampled to isotropic voxel spacing of 1 mm^3. All lesion types are grouped under a single "pathological" class for binary classification against the healthy class.

Architecture:

The feature extraction component of the `Simple3DCNN` [41] is composed of three hierarchical convolutional blocks.

The first block applies two 3D convolutional layers with channel dimensions increasing from $1 \rightarrow 32$, each followed by batch normalization and ReLU activation. A max pooling operation reduces the spatial dimensions, resulting in a feature map of shape $[32, 16, 16, 16]$.

The second block continues this progression, employing two 3D convolutions to expand the feature depth from $32 \rightarrow 64$, again followed by normalization, activation, and spatial down sampling, yielding an output of shape $[64, 8, 8, 8]$.

A third block increases the depth from $64 \rightarrow 128$, maintaining the same processing pattern, and produces a final feature representation of shape $[128, 4, 4, 4]$.

For classification, the network incorporates an adaptive average pooling layer that compresses the feature map to a compact $[128, 1, 1, 1]$ tensor. This tensor is flattened and passed through a fully connected layer of size $128 \rightarrow 64$ with ReLU activation, followed by a dropout layer (rate $= 0.5$) for regularization. A final dense layer maps the 64-dimensional latent space to 2 output neurons, corresponding to the binary prediction logits.

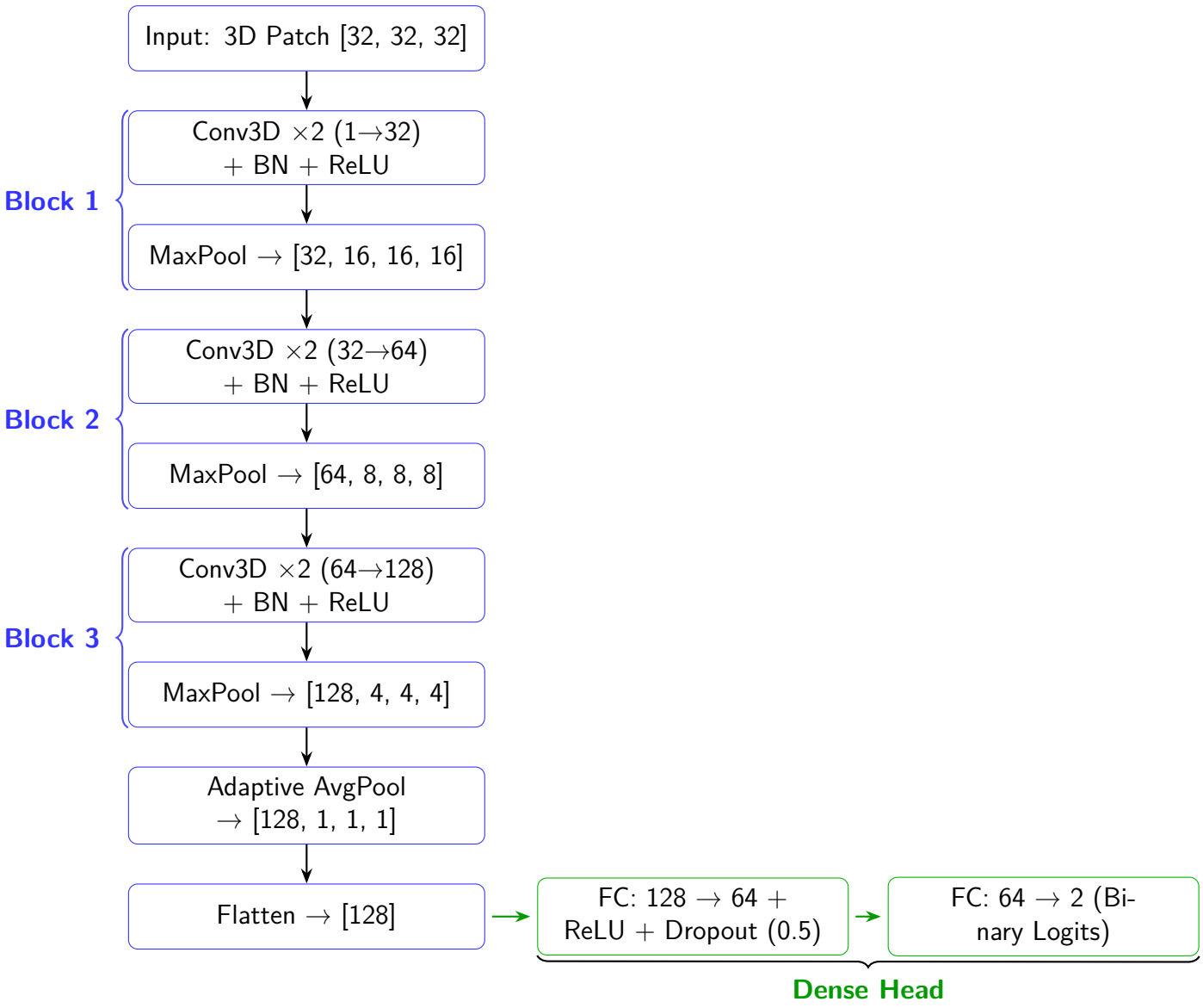

Figure 3.11: Structured view of the Simple3DCNN architecture.

The model contains approximately **867,682 trainable parameters** and is well-suited for patch-wise binary classification with limited GPU resources.

3.4.3 Multi-Class lesion classifier

This stage extends the pipeline to classify pathological lesion patches into three distinct ILD types: Ground Glass Opacity (GGO), Fibrosis, and Reticulation. It operates only on scans previously classified as pathological in Stage 1.

- **Dataset preparation:**
  From the full ILD dataset, patches with the following labels were extracted:

  – Label 3: Ground Glass Opacity (GGO) have 2030 patches

  – Label 4: Fibrosis have 7431 patches

  – Label 8: Reticulation have 5091 patches

  After filtering invalid patches, each class was balanced to 2030 samples, yielding a total of 6090 patches for training.

- **Model architecture:**
  The architecture is a modified version of the `Simple3DCNN` used in Stage 1. The final classification head was adapted for three class prediction by:

  – Changing the final fully connected layer to output 3 logits

  – Replacing sigmoid with softmax activation

  – Retaining transfer-learned weights from Stage 1

```python
# Load binary model
def create_multiclass_model_from_binary(binary_checkpoint_path: str):
    # Load binary model
    model = Simple3DCNN(in_channels=1, num_classes=2)
    checkpoint = torch.load(binary_checkpoint_path)
    model.load_state_dict(checkpoint['model_state_dict'])

    # Replace final layer for 3 classes
    model.classifier[-1] = nn.Linear(64, 3)

    # Freeze early layers ( helps retain learned features)
    for param in model.features.parameters():
        param.requires_grad = False

    return model
```

Figure 3.12: Modifying a pre trained binary classification model to support multi-class classification by updating the final layer and freezing early feature extraction layers for transfer learning.

- **Input format:**
  All inputs are 3D patches of size $32 \times 32 \times 32$, normalized to $[0, 1]$, and resampled to isotropic 1 mm^3 resolution.

- **Training strategy:**
  Training was conducted using the same infrastructure as Stage 1 with minor adjustments:

  – Loss: Multi-class cross-entropy

  – Balanced class sampling

  – Validation split to monitor performance

```python
# Key changes:
# 1. Class names
class_names = {0: 'GGO', 1: 'Fibrosis', 2: 'Reticulation'}

# 2. Weighted loss for class imbalance (if needed)
class_weights = torch.tensor([1.0, 1.2, 1.5])  # Adjust based on difficulty
criterion = nn.CrossEntropyLoss(weight=class_weights)

# 3. Per-class metrics
def compute_per_class_metrics(all_preds, all_labels):
    from sklearn.metrics import classification_report
    report = classification_report(
        all_labels, all_preds,
        target_names=['GGO', 'Fibrosis', 'Reticulation'],
        output_dict=True
    )
    return report
```

Figure 3.13: Implementing class-specific evaluation using weighted loss to address class imbalance and computing per-class performance metrics.

This stage provides detailed classification at the lesion level, enhancing clinical interpretability and supporting downstream visualization or triage modules.

3.4.4 Interpretability

To enhance interpretability of the ILD classification model and enable visual inspection of learned spatial features, we integrated a volumetric extension of Gradient-weighted Class Activation Mapping (Grad-CAM) into our pipeline. The 3D Grad-CAM [38] technique highlights class-discriminative regions within input CT volumes, allowing clinicians and researchers to understand model focus in the context of ILD subtype classification.

- **Method overview**

Grad-CAM works by computing the gradient of the class score with respect to feature maps in the last convolutional layer of a trained 3D CNN. For a given class c, the class activation map $L^c_{\text{Grad-CAM}}$ is computed as:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \tag{3.2}$$

where A^k is the k-th feature map of the final convolutional layer, and α^c_k is the global average of the gradients of the score for class c with respect to A^k:

$$\alpha^c_k = \frac{1}{Z}\sum_i \sum_j \sum_l \frac{\partial y^c}{\partial A^k_{i,j,l}} \tag{3.3}$$

Here, Z denotes the total number of voxels in the feature map. The resulting 3D activation map is upsampled to the input resolution and overlaid on the original CT patch for visualization.

- **Implementation details**

We implemented 3D Grad-CAM in PyTorch by:

- Registering a forward hook on the final convolutional layer to capture feature maps.
- Registering a backward hook to capture the gradients of the output class score with respect to these feature maps.
- Computing the voxel-wise attention maps as weighted combinations of channels, followed by a ReLU non-linearity.
- Upsampling the 3D Grad-CAM volume to the input size using trilinear interpolation for overlay visualization.

The internal flow of the 3D Grad-CAM module is illustrated in Figure 3.14. It shows how the trained 3D CNN processes an input CT volume to produce class scores, and how gradients are propagated back to compute the class-specific attention map. The resulting activation map is then overlaid on the CT volume across orthogonal planes to assist in visual interpretation of model focus during inference.

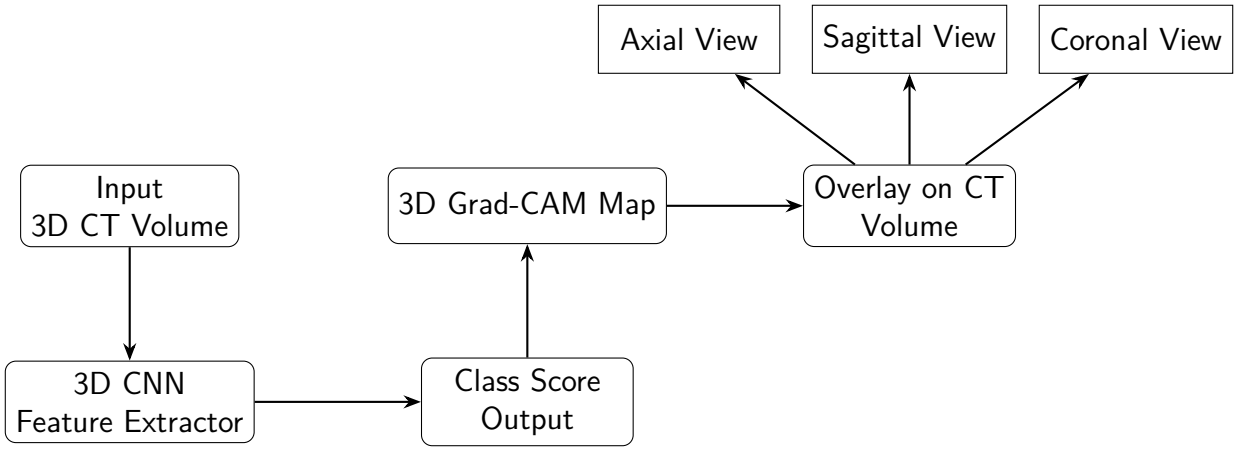

Figure 3.14: Pipeline of the 3D Grad CAM module.

To support spatial inspection, we extracted and visualized the Grad-CAM overlays in all three anatomical planes (axial, sagittal, and coronal). Additionally, for misclassified or ambiguous cases, we generated Grad-CAM montages to better understand the model's confusion and to identify potential lesion localization failures or data ambiguity.

This volumetric interpretability module is a critical component of our pipeline, offering transparency in model behavior and aiding in clinical validation of classification outcomes.

3.5 Conclusion

The architecture of the system is designed to be modular, efficient, and clinically meaningful. By segmenting lungs before lesion detection, the pipeline ensures that analysis is restricted to relevant anatomical regions, thereby reducing background noise. A two-stage classification strategy starting with binary lesion detection followed by multi-class lesion categorization allows computational resources to be focused on informative subregions. The use of 3D U-Net for segmentation and 3D CNNs for classification provides robust feature extraction and volumetric context modeling, which are critical in analyzing HRCT scans. To enhance interpretability, the system incorporates Gradient-weighted Class Activation Mapping (Grad-CAM), which generates 3D heatmaps that visualize the regions most influential in the model's decision-making process.

This methodological foundation supports both quantitative performance evaluation and qualitative clinical insights, establishing the framework for validating each component under realistic diagnostic scenarios.

CHAPTER 4

EXPERIMENTAL DESIGN

4.1 Introduction

This chapter describes the experimental setup and evaluation protocol for the training and testing of the proposed models. Its content includes hardware/software setup, data splitting details, and testing methodology for each stages lung segmentation, Stage 1 ILD binary Detection and Stage 2 multi-class classification. The hope is that reproducibility, fairness in scoring, and the robustness of the performance of subsequent assessments against multiple patient cohorts can be achieved.

The overarching objective of this experimental framework is to ensure a rigorous, unbiased quantitative assessment of the proposed models.

4.2 Data partitioning strategy

To ensure robust model evaluation and prevent data leakage, the dataset was carefully partitioned at the patient level, followed by patch-wise extraction and balancing techniques tailored to the classification and segmentation tasks.

4.2.1 Dataset split methodology

Acceptable model generalization will come only by imposing rigorous standards of evaluation. Dataset splits were performed on patient level stratified splits, where no slices of the same patient appeared in different dataset splits.

The cohort of 107 patients was split into three groups:

- 70% Training set: To determine the optimal model weights
- 15% Validation set: for estimation of early stopping and performance tuning.
- 15% Test set: Entirely held-out to assess generalization

The stratified random splitting ensures balanced representation of lesion types across all sets while maintaining the constraint that all patches from a single patient remain within the

same split. This patient-level separation is crucial for evaluating the model's generalization capability to unseen patients.

4.2.2 Patch extraction and balancing

To support both binary and multi-class lesion classification tasks, 3D image patches were extracted from the segmented lung volumes using a standardized patch-wise sampling strategy. All patches had a fixed spatial size of $32 \times 32 \times 32$ voxels and were generated using a sliding window approach with a stride of 16 voxels in all directions. This dense sampling ensured comprehensive coverage of the lung fields and minimized the likelihood of missing small or peripheral lesions.

Binary classification sampling:

For the binary lesion detection task, patches were labeled based on voxel-wise annotations within each 32^3 region. Two criteria were used for assigning class labels:

- **Healthy Patches:** Required at least 95% of voxels labeled as healthy lung parenchyma.
- **Lesion Patches:** Contained at least 50 voxels belonging to one or more lesion classes—Ground Glass Opacity (GGO), Fibrosis, or Reticulation.

To ensure the quality and relevance of patches, further filtering was applied:

- Minimum lung coverage: 70% of the patch must intersect the segmented lung mask.
- Minimum tissue heterogeneity: Standard deviation in Hounsfield Units (HU) ≥ 50.
- Intensity range normalization: HU values clipped to $[-1000, 1000]$ to remove outliers and harmonize inputs.

After filtering, the process yielded approximately 30,000 high-quality binary classification patches.

Multiclass classification balancing:

To mitigate class imbalance and improve classifier generalization, we restricted our multi-class patch-based classification task to the three most prevalent and well-separated ILD lesion types: Ground Glass Opacities (GGO), Fibrosis, and Reticulation. This decision was based on an analysis of lesion occurrence, patch availability, and HU characteristics, as summarized in Table 4.1.

Class	Patients	Avg Size (voxels)	HU Mean
3-GGO (Ground Glass Opacity)	31	33,257	∼404
4-Fibrosis	30	39,310	∼449
5-Micronodules	18	136,113	∼118
8-Reticulation	10	49,048	∼409
Others (Types 2, 6–17)	≤8	<10K or rare	mixed

Table 4.1: Class-wise summary of ILD lesions in the dataset.

Despite the large average size of micronodules, they exhibited considerable heterogeneity and poor HU separability from surrounding parenchyma. Additionally, they were often distributed diffusely across lung fields, making patch-level isolation challenging. Similarly, the "Others" category (including rare or ambiguous ILD patterns such as consolidation, honeycombing, and cysts) lacked sufficient annotated samples and were highly inconsistent in shape and texture.

For these reasons, both Micronodules and Others were excluded from training. The final patch-based classifier was trained on three well-defined classes: GGO, Fibrosis, and Reticulation.

After filtering, a balanced dataset was curated containing 2,030 patches per lesion class:

- Ground Glass Opacity (GGO) – Label 3
- Fibrosis – Label 4
- Reticulation – Label 8

This resulted in a final multiclass training set of 6,090 patches.

4.2.3 Data augmentation strategy

To improve model generalization and reduce overfitting due to the limited dataset size, online data augmentation techniques were applied during training. These augmentations introduce controlled variability while preserving anatomical and radiological consistency.

Spatial transformations:

- **Random 90° Rotations:** Performed independently along the axial, sagittal, and coronal planes with a probability of $p = 0.5$.
- **Random Flipping:** Random left-right, anterior-posterior, and superior-inferior flips applied with $p = 0.5$ per axis.

Intensity transformations:

- **Gaussian Noise Addition:** Zero-mean Gaussian noise with standard deviation $\sigma \in [0.01, 0.03]$ added with probability $p = 0.3$.
- **Intensity Shifting:** Global shift of voxel intensities by $\pm 5\%$ of the normalized range applied randomly ($p = 0.3$).

These augmentations were selected to simulate real-world imaging variability while preserving the diagnostic integrity of the HRCT volumes. They were implemented online during training using efficient augmentation pipelines integrated into the PyTorch data loader framework.

4.3 Hardware and software environment

Experimental evaluation took place in a hybrid computing scenario. Most training and evaluation were performed on **Kaggle Notebooks** [48] using GPU acceleration. pre-processing, patch extraction and rendering were partially performed on a local workstation. This had to be done to work around memory and session-time limits imposed by cloud platforms.

4.3.1 Computational infrastructure

- **Cloud platform:** Kaggle Notebooks
  - Equipped with dual NVIDIA Telsa T4 GPUs (16 GB VRAM)
  - CUDA version 11.7, cuDNN enabled
  - RAM: 30 GB, max run-time: 9 hours per session
- **Local workstation:**
  - Windows 64 bit, Intel Core i5, 16 GB RAM

4.3.2 Software stack

The use of these software tools ensures the systems modularity and reproducibility:

- **PyTorch 2.0:** Deep learning library for setting up, training, and evaluating models [49].
- **MONAI:** Medical AI library on the PyTorch framework for 3D segmentation, training loops, and metrics [46].
- **TorchIO:** Augmentation and patch based pre-processing for 3D volumes [47].
- **SimpleITK, Nibabel:** To load and save DICOM/NIfTI images along with spatial consistency [50], [51].

- **scikit learn, seaborn, matplotlib:** Used for computing metrics, and statistical evaluation, and for visual analytics [52]–[54] .

Figure 4.1 offers a visual overview of the software infrastructure employed.

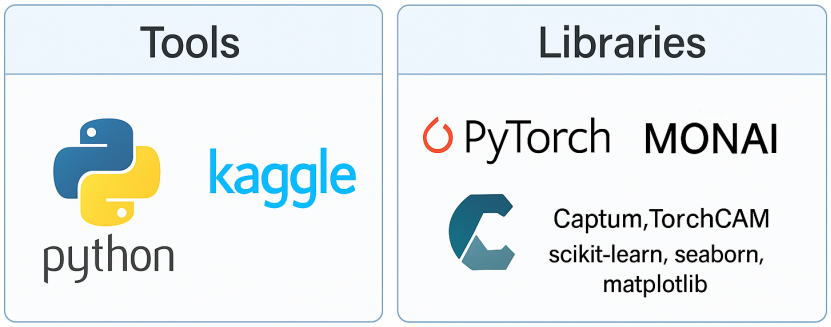

Figure 4.1: Overview of the software environment used in the study.

4.4 Performance evaluation metrics

Each metric employed in this work is selected to assess a different facet of performance ranging from segmentation accuracy and boundary precision to the effectiveness of classification [55].

4.4.1 Segmentation evaluation metrics

1. **Dice Similarity Coefficient (DSC):**

   Dice coefficient is a measure of overlap between the ground truth $\mathcal{G}$ and the predicted segmentation $\mathcal{P}$. It is particularly well suited to handle imbalanced datasets.

   $$\text{DSC} = \frac{2 \cdot |\mathcal{P} \cap \mathcal{G}|}{|\mathcal{P}| + |\mathcal{G}|}$$

   Where:

   - $\mathcal{P}$: Predicted segmentation voxels
   - $\mathcal{G}$: Ground truth segmentation voxels
   - $|\cdot|$: Cardinality (number of voxels)
   - $\cap$: Intersection of predicted and ground truth voxels

   DSC values range from 0 (no overlap) to 1 (perfect agreement).

2. **Jaccard Index (Intersection over Union, IoU):**

   The Jaccard Index calculates the intersecting over combined voxels ratio.

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

Where:

- $\cup$: Union of predicted and ground truth voxels

IoU is stricter than DSC and penalizes more over segmentation.

3. **Precision:**

   Precision calculates the ratio of true positive predictions to all predicted positives.

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

   Where:

   - TP: True Positives
   - FP: False Positives
   - TN: True Negatives
   - FN: False Negatives

4. **Recall (Sensitivity):**

   Recall calculates the model's ability in distinguishing true positives from all real positives.

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

5. **F1 Score:**

   The F1 score is the harmonic mean of precision and recall.

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

6. **Volumetric Overlap Error (VOE):**

   Relative error in the volume of predicted and ground truth segmentations.

4.4.2 Patch-Level classification metrics

Patch-level predictions were assessed on the held-out test set using standard classification metrics:

- **Accuracy:** Overall correctness of classification across all patches.

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

- **Confusion matrix:** Displays the distribution of predictions versus actual classes to identify patterns of misclassification.
- **Area Under the ROC Curve (AUC):** Evaluates the model's ability to distinguish between classes, especially for binary confidence-based classification.

4.4.3 Patient-Level evaluation metrics

To better reflect clinical applicability, patch-level predictions were aggregated to obtain patient-level outcomes.

Aggregation strategies:

- **Mean Probability:** Average softmax probability across all patches from a patient.
- **Majority Voting:** The class with the highest number of patch-level predictions is chosen.
- **Confidence-Weighted Aggregation:** Patches are weighted by prediction confidence or patch quality.

Single-Label metrics:

- Accuracy, precision, recall, and F1-score computed per patient.
- Patient-level confusion matrix.

Multi-Label metrics: (for overlapping or mixed lesions)

- **Multi-label Precision:**

$$\text{Precision}_{\text{multi}} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FP_i)}$$

- **Multi-label Recall:**

$$\text{Recall}_{\text{multi}} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FN_i)}$$

- **Hamming Loss:**

$$\text{Hamming Loss} = \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{j=1}^{L} \mathbb{1}[y_{ij} \neq \hat{y}_{ij}]$$

Where:

  - N: Number of patients
  - L: Number of possible labels
  - y_{ij}: True label indicator
  - $\hat{y}_{ij}$: Predicted label indicator

4.5 Experimental workflow overview

The experimental pipeline was aimed to investigate how the proposed multi stage system performs for detecting and classifying ILD using HRCT studies. The workflow is modular and corresponds to real world clinical diagnostic procedure: segment lung volume, detect abnormality, and categorize lesion type. Each step was implemented, validated and tested separately to ensure for robust performance and traceability.

4.5.1 Evaluation of lung segmentation

Before any lesion detection, the lung regions were segmented using a customized three-dimensional (3D) U-Net model. This and the following pre processing phase ensure that:

- Analysis on lung parenchymas only is considered
- Non-pulmonary structures, such as ribs or diaphragm, are not considered
- Patch acquisition and lesion categorization are spatially constrained

The output of this step is a set of lung-masked CT volumes, which are then used in Stage 1 for binary classification and Stage 2 for lesion-specific patch filtering and training.

4.5.2 Binary classification experiment

The first experiment in the evaluation pipeline aims to establish a baseline for binary classification, focusing on distinguishing between healthy lung tissue and regions exhibiting interstitial lung disease (ILD) pathology. This binary decision step represents a critical component of the diagnostic workflow, enabling the model to screen scans prior to multi-class lesion classification.

Experimental protocol:

- The Simple3DCNN architecture is trained to perform binary classification (healthy vs. pathological).
- The training procedure uses the Adam optimizer with a learning rate of 0.001 and a weight decay of 1×10^{-4} to prevent overfitting.
- Early stopping is implemented with a patience value of 10, based on the validation loss curve.
- The model is evaluated on the held-out test set using standard patch-level metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

This experiment provides insight into the model's sensitivity to early disease features and its robustness against false positives, serving as a foundation for subsequent multi-class classification tasks. Let me know if you'd like the LaTeX for a results table or confusion matrix associated with this experiment, or if you want to mention patient-level aggregation here as well.

This evaluation protocol enabled clinically interpretable patient-level decisions, balancing sensitivity to disease detection with robustness against false positives.

4.5.3 Multi-class classification experiment

This stage was evaluated using balanced data for the three lesion types. The testing strategy ensured that the model was validated in a clinically meaningful and fair setting.

- **Data handling:**
  The dataset was split into training and validation sets from the 6090 patches (2030 per class).
- **Patch-Based classification:**
  All predictions were made on 3D patches of size $32 \times 32 \times 32$, extracted from pathological lung volumes. Each patch was processed independently, and the model predicted one of the three lesion classes.
- **Model inference:**
  The model output a softmax probability distribution over the three classes. The predicted class was the one with maximum probability.
- **Training dynamics:**
  Training converged after approximately 7 epochs. No signs of overfitting were observed. Transfer learning enabled fast initial improvements.

This evaluation framework confirms the robustness of the multi-class classifier and its ability to distinguish between visually overlapping ILD lesion types using 3D information.

4.5.4 Model interpretability experiment

The final experiment focuses on interpreting the inner workings of the trained model and evaluating its decision-making transparency. Given the critical importance of explainability in clinical applications, this stage leverages Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the spatial regions that most strongly influence the network's predictions.

Experimental rotocol:

- Generate 3D Grad-CAM activation maps from the final convolutional layers of the trained model.
- Visualize activation regions for correctly classified samples from each lesion type (Ground Glass Opacity, Fibrosis, Reticulation).
- Aggregate individual Grad-CAM maps to construct average class-specific attention patterns.
- Examine misclassified patches to identify potential failure cases and confounding image characteristics.

This interpretability experiment enables qualitative assessment of model behavior and assists in identifying clinically relevant activation zones. Moreover, it provides valuable insights into the anatomical features contributing to each decision and highlights limitations in model generalization across ambiguous or overlapping lesion appearances.

4.5.5 Patient-Level inference experiment

In order to evaluate the clinical applicability of the proposed pipeline, this experiment investigates the diagnostic performance at the patient level by aggregating patch-wise predictions. While previous stages focused on localized classification of 3D patches, the goal here is to simulate a real-world scenario in which clinical decisions are made based on the analysis of entire CT scans.

Experimental protocol:

- Each patient scan in the test set is fully processed through the trained multiclass classifier to generate patch-wise softmax outputs.
- The entire CT volume is divided into overlapping 3D patches. Only patches passing quality and anatomical relevance thresholds are included.
- Three aggregation strategies are applied to produce a single prediction per patient:
    1. **Mean Probability Aggregation:** Computes the average softmax probability across all patches and selects the class with the highest average.
    2. **Maximum Confidence Aggregation:** Identifies the patch with the maximum class probability and uses its predicted label.
    3. **Majority Voting:** Determines the most frequently predicted class label among all patches.
- Final predictions are compared with ground truth patient-level labels derived from metadata.
- Special attention is given to ambiguous and mixed-pattern cases (e.g., co-occurrence of fibrosis and ground-glass opacity), which are qualitatively analyzed to assess robustness.

This experiment provides a realistic framework for translating patch-level outputs into clinically actionable patient-level predictions. By comparing aggregation strategies, the study aims to determine which method yields the most consistent and reliable performance in the presence of inter-lesion variability, potential class imbalance, and image heterogeneity.

4.5.6 Overview of the system workflow

Figure 4.2 shows a high level of the system's workflow, from CT data pre processing to final output.

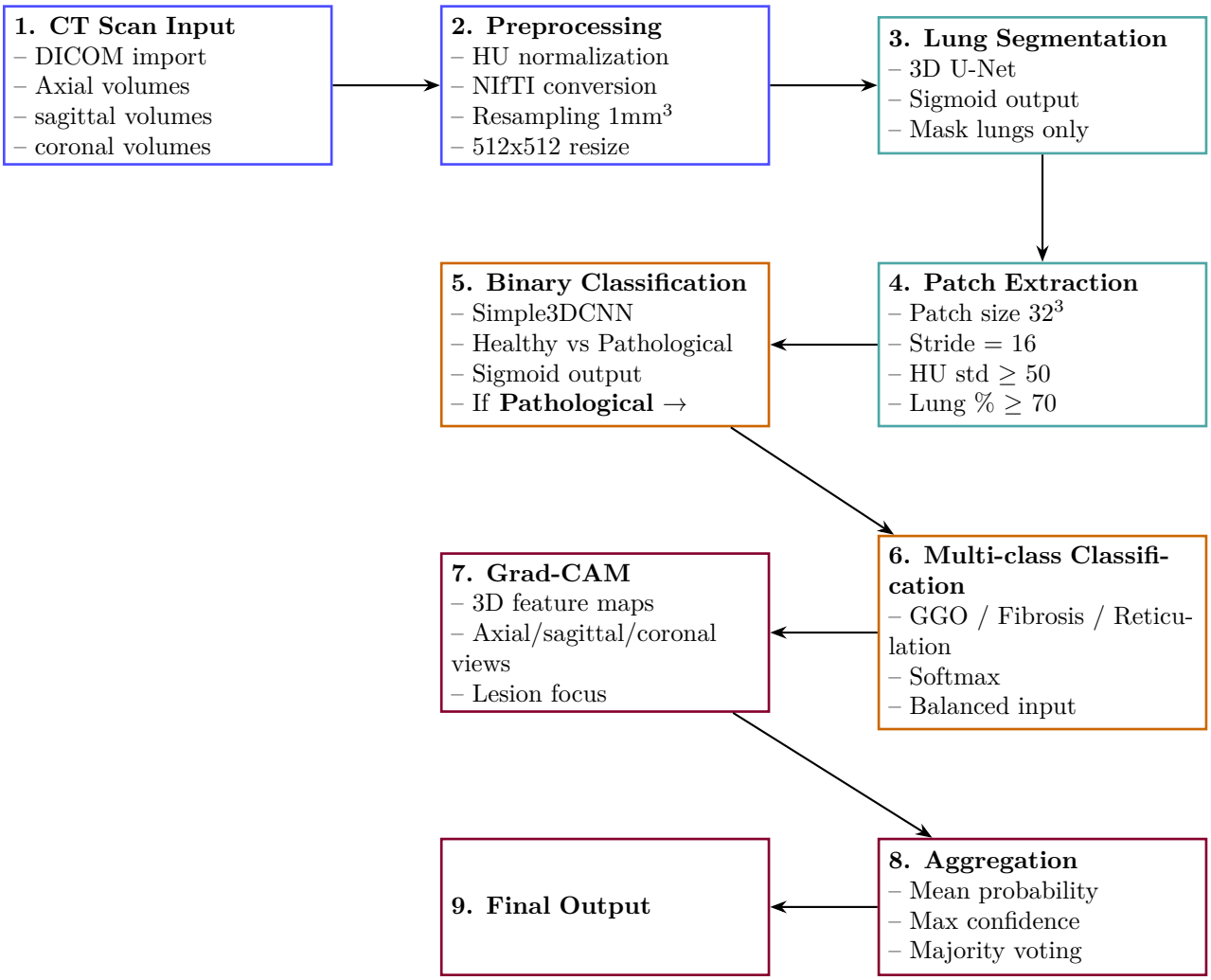

Figure 4.2: a complete pipeline diagram from input CT to final decision with Grad-CAM.

4.6 Implementation details

To ensure experimental rigor and reproducibility, this section outlines the specific training configuration used across all models, along with the measures taken to guarantee consistent and traceable results.

4.6.1 Training configuration

The experiments were conducted using a standardized training pipeline to ensure reproducibility and computational efficiency. The configuration is summarized as follows:

- **Batch size:** 32 (applied for both binary and multiclass classification tasks)
- **Number of epochs:** 50, with early stopping based on validation loss (patience = 10)
- **Learning rate strategy:** ReduceLROnPlateau scheduler (reduction factor = 0.5, patience = 5)
- **Optimizer:** Adam optimizer with default parameters

4.6.2 Reproducibility measures

To enhance reproducibility of experimental results, the following practices were adopted:

- **Random seed control:** Fixed seeds across all random number generators in NumPy, PyTorch, and CUDA
- **Deterministic behavior:** Enforced deterministic computation in PyTorch where feasible
- **Configuration logging:** Full experimental configuration saved alongside each model checkpoint
- **Environment management:** Experiments encapsulated in a Docker container replicating the exact runtime environment

4.7 Conclusion

This chapter has presented a comprehensive overview of the experimental design underpinning the ILD diagnosis pipeline. From standardized preprocessing and lung segmentation to the modular two-stage classification framework and Grad-CAM-based interpretability, each component was selected and configured to reflect both technical soundness and clinical applicability. Particular attention was given to balanced data handling, patient-level evaluation, and reproducibility to ensure robust performance under real-world conditions. The next chapter will detail the results obtained from this pipeline and analyze the effectiveness of each stage in achieving accurate and interpretable ILD classification.

CHAPTER 5

RESULTS AND ANALYSIS

5.1 Introduction

This chapter presents the experimental results obtained from the proposed 3D deep learning pipeline for ILD analysis. The evaluation focuses on three main components: lung segmentation, binary lesion detection (Stage 1), and multi-class ILD lesion classification (Stage 2). Each section provides quantitative performance metrics, training behavior analysis, and qualitative visualizations to illustrate the effectiveness of the system.

5.2 Lung segmentation results

In this section, we present the quantitative and qualitative results of the lung segmentation task.

5.2.1 Training and validation performance

In addition, Dice coefficient and segmentation loss were displayed over epochs to observe the learning process of the segmentation model.

- The right plot in Figure 5.1 shows the Dice coefficient curves for both training and validation sets. The metric steadily increases over epochs and plateaus near **0.99**, reflecting excellent spatial overlap with the ground truth masks. This plateau suggests that the model generalizes well without overfitting.

  These visual trends are supported by the quantitative metrics reported in Table 5.1, where the Dice coefficient reaches a mean of $\mathbf{0.9926 \pm 0.0054}$ across 17 test patients. The accompanying Hausdorff distance of $\mathbf{3.17 \pm 0.51}$ further confirms the spatial accuracy of the model, while high sensitivity (**0.9965**) and precision (**0.9956**) demonstrate its ability to segment lung structures completely and correctly.

- The intermediate and right plots in Figure 5.1 present the segmentation loss for training and validation. Both losses decrease smoothly and converge to low final values, with

no significant divergence between them. This convergence indicates training stability and validates that the model is not overfitting to the training data.

Together, these observations confirm that the segmentation model achieves high performance, robustness, and consistency, making it a reliable pre-processing component for the downstream classification pipeline.

Figure 5.1: Segmentation training and validation progress

Table 5.1 summarizes the evaluation metrics obtained for the lung segmentation model on the test cohort consisting of 17 patients. The results indicate high precision and reliability of the segmentation output, as detailed below:

Metric	**Mean**	**Standard Deviation**
Dice Coefficient	0.9926	± 0.0054
Hausdorff Distance (voxels)	3.17	± 0.51
Sensitivity	0.9965	–
Specificity	0.9994	–
Precision	0.9956	–

Table 5.1: Performance of the lung segmentation model on the test set (n = 17 patients).

Together, these metrics confirm that the 3D U-Net segmentation model performs with near-perfect accuracy, both in overlap and boundary precision, and is well suited as a pre-processing step for downstream lesion classification tasks.

Figure 5.2 shows a qualitative example from the test set, displaying the ground truth mask, the model's predicted segmentation, as well as a color coded error map showing true positives (green), false positives (red) and false negatives (yellow).

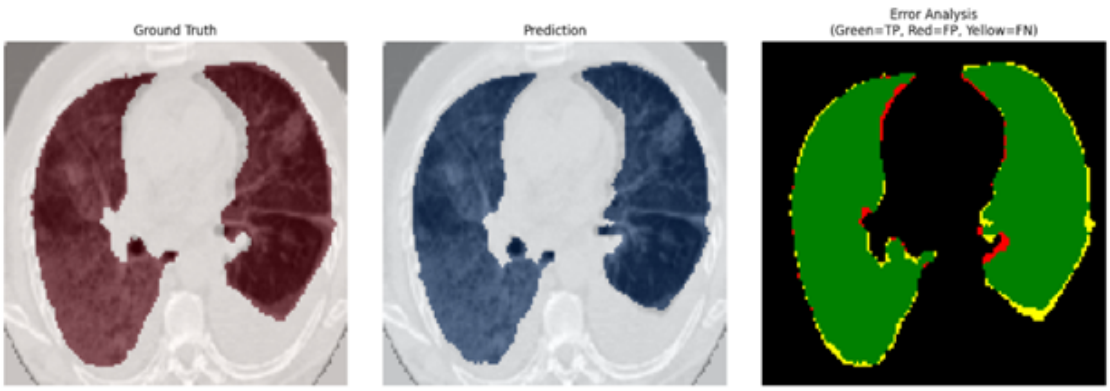

Figure 5.2: Lung segmentation visualization. Left: ground truth, Center: model prediction, Right: error map with TP (green), FP (red), FN (yellow).

We utilized a dedicated lung segmentation model to limit the analysis of lesions to the pulmonary area and to minimize the contamination of non lung tissue. The binary lung mask was then cropped and masked the HRCT volumes for further processing.

To illustrate the output of the lung segmentation model, Figure 5.3 presents an example case from the dataset. The left panel shows the original axial CT slice. The center panel displays the predicted lung mask overlaid on the same slice, highlighting the precise isolation of lung parenchyma. Finally, the right panel demonstrates the masked CT volume where only lung regions are preserved—this filtered volume serves as the input for subsequent lesion patch extraction and classification tasks. This visual confirms the anatomical relevance and spatial accuracy of the segmentation step.

Figure 5.3: Example of lung segmentation: (Left) original CT slice, (Center) segmented lung mask applied, (Right) masked lung volume ready for lesion detection.

5.3 Results and analysis: binary lesion detection

- Training and evaluation metrics

Figure 5.4 illustrates the training dynamics of the binary classifier over 50 epochs, highlighting both loss and accuracy for training and validation datasets.

The left plot demonstrates a smooth and consistent decline in both training and validation loss curves, converging to near-zero values, which indicates that the model successfully minimized the objective function without overfitting. Similarly, the right plot shows a progressive increase in accuracy, with training accuracy reaching approximately 98.9% and validation accuracy nearing 99.4%.

This convergence and minimal gap between training and validation curves suggest excellent generalization capability. Moreover, the absence of oscillations or divergence in the loss curves is indicative of stable learning behavior.

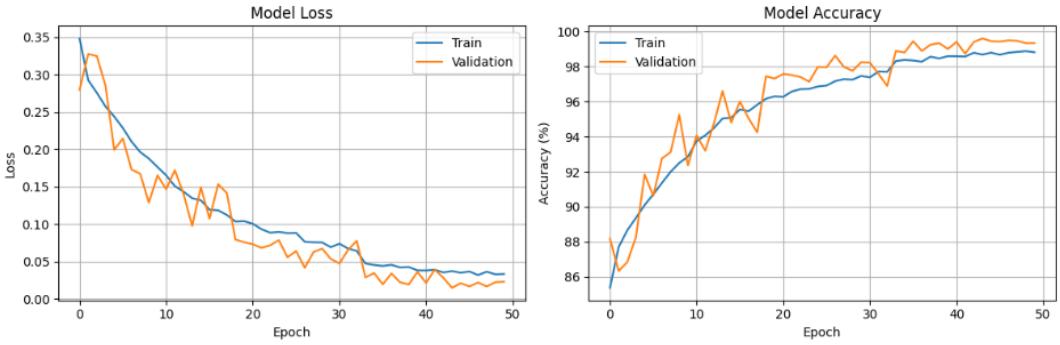

Figure 5.4: Collection of training and evaluation metrics for Stage 1.

Figure 5.5 presents the confusion matrix for the binary classification task distinguishing between healthy and lesion-containing samples. The classifier demonstrates exceptional performance, with 2,238 true negatives (correctly classified healthy samples) and 2,236 true positives (correctly classified lesion samples). Only 12 healthy samples were misclassified as lesions (false positives), and 14 lesion samples were incorrectly predicted as healthy (false negatives).

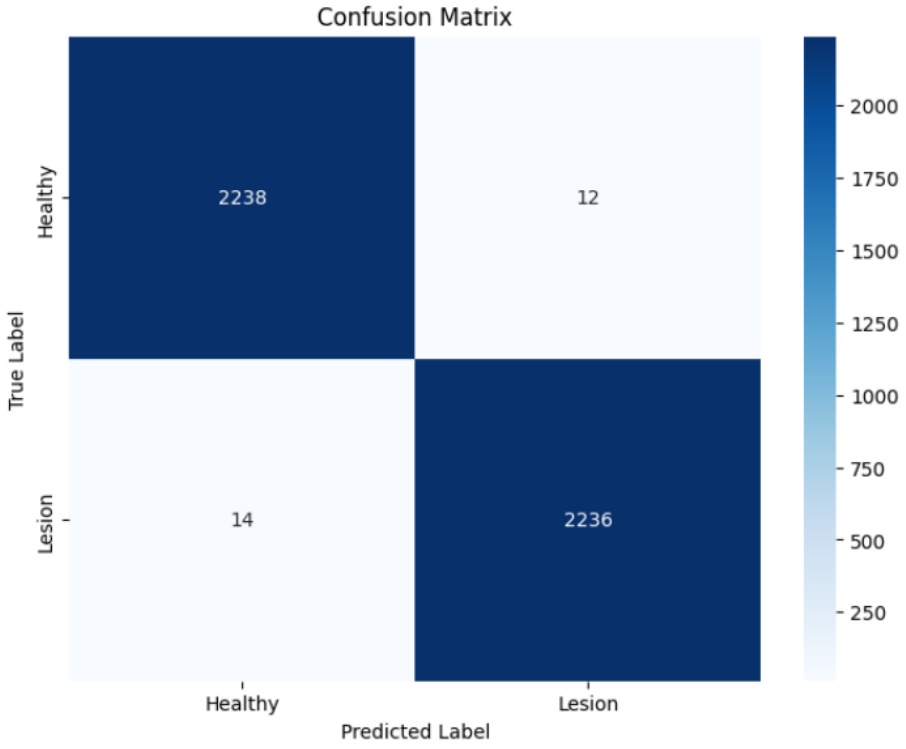

Figure 5.5: Confusion Matrix of Stage 1.

Table 5.2 summarizes the evaluation metrics obtained for the binary classification task. The results demonstrate high classification performance, highlighting the model's ability to effectively distinguish between healthy and pathological patches.

Metric	**Value**
Accuracy	0.9947
Balanced Accuracy	0.9944
Precision (Lesion)	0.9946
Recall / Sensitivity (Lesion)	0.9938
F1 Score (Lesion)	0.9942

Table 5.2: Evaluation metrics for binary lesion detection (patch-level inference).

These results show strong generalization and clinical usefulness in a high-recall context.

5.4 Results and analysis: Multi class lesion classification

After training and validation, the proposed multi-class ILD lesion classifier demonstrated strong generalization performance across all target classes.

- **Confusion matrix**

The confusion matrix (Figure 5.6) illustrates the model's performance across the three pathological lung patterns: Ground Glass Opacity (GGO), Fibrosis, and Reticulation.

High correct classification rates were achieved for all classes, with the most significant confusion observed between **GGO and Fibrosis** (45 cases), likely due to their overlapping radiological features and clinical progression in interstitial lung diseases.

A moderate level of confusion occurred between **Fibrosis and Reticulation** (32 cases), reflecting similarities in structural lung alterations. The **least confusion** was found between **GGO and Reticulation** (26 cases), suggesting these patterns are more visually distinct to the model.

The model exhibited **rapid convergence** by epoch 7, with training and validation curves closely aligned, indicating **no overfitting**. The application of transfer learning significantly accelerated training, improving accuracy from an initial 50% to a final test accuracy of **88.73%**.

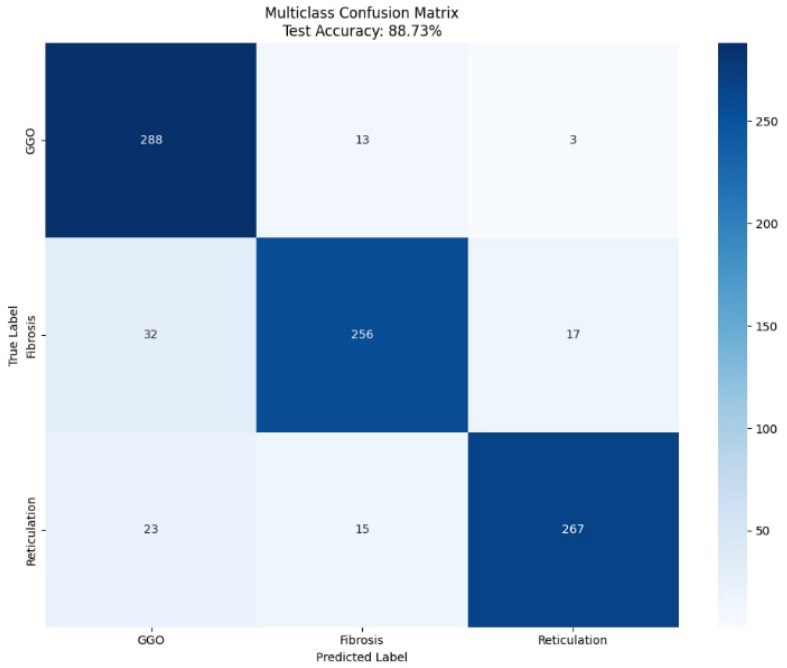

Figure 5.6: Confusion matrix for ILD lesion classification.

- **Training and evaluation metrics**

To evaluate the learning behavior and generalization capacity of the model, we examine the training and validation curves over 30 epochs for three core metrics: loss, accuracy, and F1 score (Figure 5.7).

- **Loss curve:**

  The loss curve exhibits a steady and consistent decrease in both training and validation loss across epochs. Importantly, the validation loss remains consistently below the training loss, suggesting good generalization and a lack of overfitting. The most significant reduction in loss occurs during the initial epochs (0–7), indicating rapid convergence. After this phase, the loss continues to decline gradually, reflecting a fine-tuning stage with diminishing updates.

- **Accuracy curve:**

  The accuracy curve demonstrates rapid improvement in both training and validation accuracy during the first few epochs, particularly from epoch 0 to 7. Training accuracy increases from approximately 50% to over 80%, while validation accuracy stabilizes around 88%. Beyond epoch 7, the improvements become marginal, and both curves exhibit a plateau. The alignment between the two curves suggests that the model is learning in a stable and generalizable manner.

- **F1 score curve:**

  The F1 score curve closely follows the trend observed in accuracy, with sharp early improvements and subsequent stabilization. The validation F1 score reaches approximately 0.88 and remains consistent thereafter, reflecting balanced performance in terms of both precision and recall. This is particularly important in imbalanced or multi-class classification tasks, where F1 score provides a more nuanced assessment than accuracy alone.

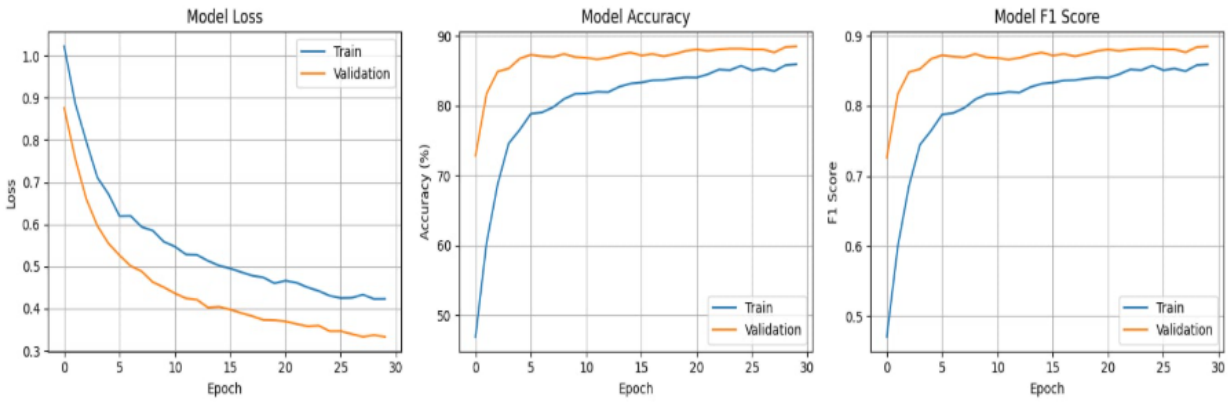

Figure 5.7: Training and validation curves for loss (left), accuracy (middle), and F1 score (right) over 30 epochs.

5.5 Grad-CAM visualizations and interpretation

This section presents qualitative results using Gradient-weighted Class Activation Mapping (Grad-CAM) to gain insight into the learned spatial attention of the multi-class lesion classifier. Grad-CAM overlays highlight which areas of the CT patch most influenced the model's decision.

5.5.1 Per-Class prediction and activation maps

Figures 5.8, 5.9, and 5.10 show representative Grad-CAM outputs for three common ILD lesion types: Reticulation, Fibrosis, and Ground Glass Opacity (GGO). Each figure displays a predicted patch alongside its heatmap overlay, with activation intensities indicating the most relevant regions for classification; areas highlighted in **red** correspond to the Regions of Interest (ROIs) where the model focuses its attention, while **blue regions** indicate low or no contribution to the prediction.

- **Reticulation (Figure 5.8):** The model focused on thin, linear structures consistent with fibrotic reticulation. The predicted probability exceeded 91%.

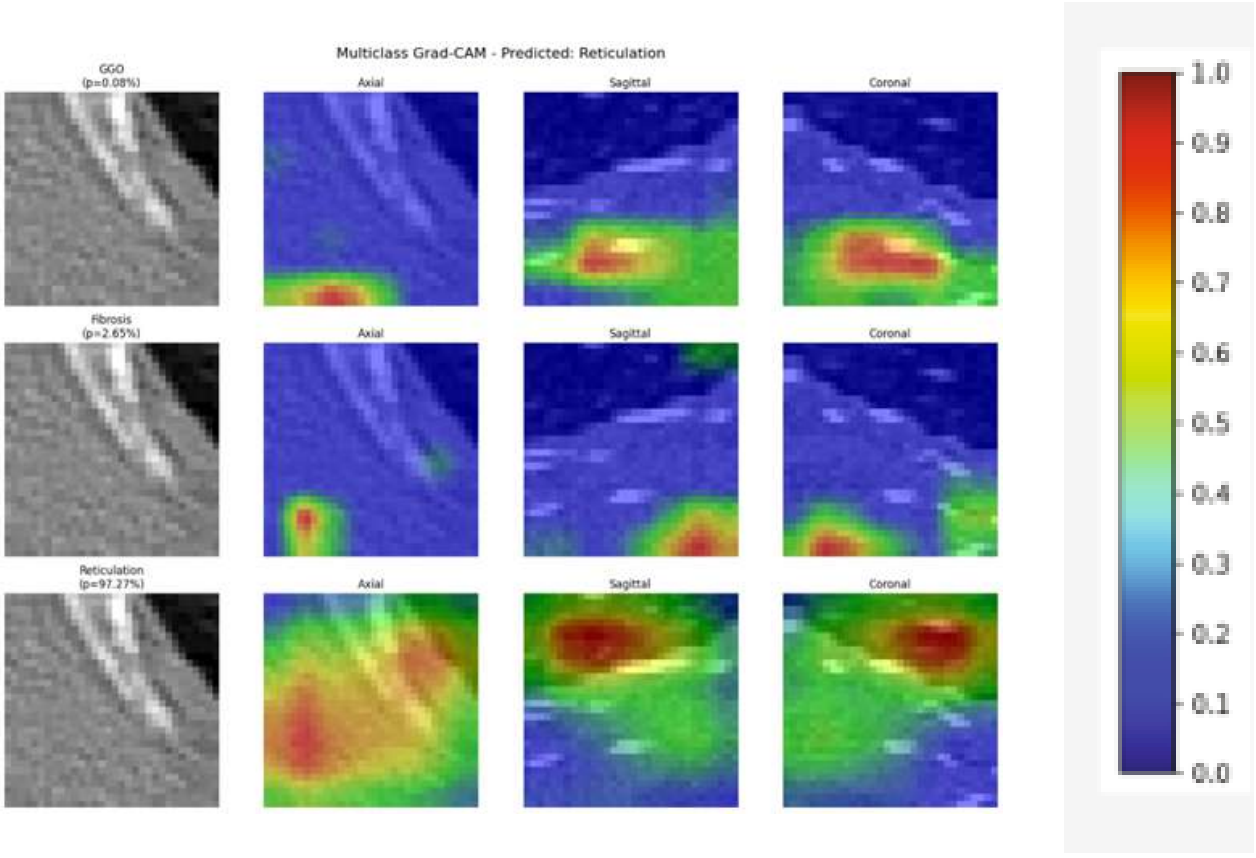

Figure 5.8: Grad-CAM overlay for a patch predicted as Reticulation. Confidence: 91

- **Fibrosis (Figure 5.9):** Activation was distributed over irregular, dense textures hallmarks of fibrosis. The model predicted this class with 89% confidence.

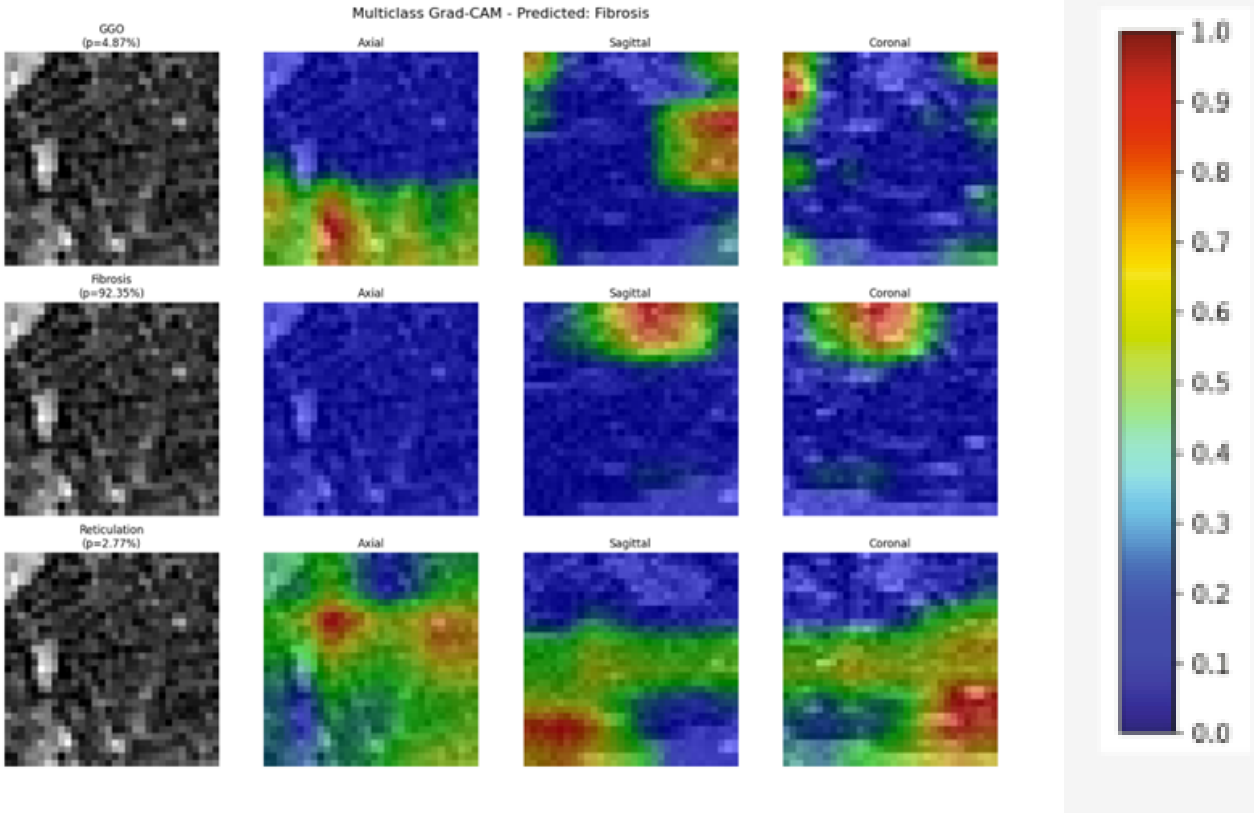

Figure 5.9: Grad-CAM overlay for a patch predicted as Fibrosis. Confidence: 89.

- **Ground Glass Opacity (GGO) (Figure 5.10):** The heatmap emphasized hazy areas of low intensity within the lungs, typical of GGO, with prediction confidence reaching 94%.

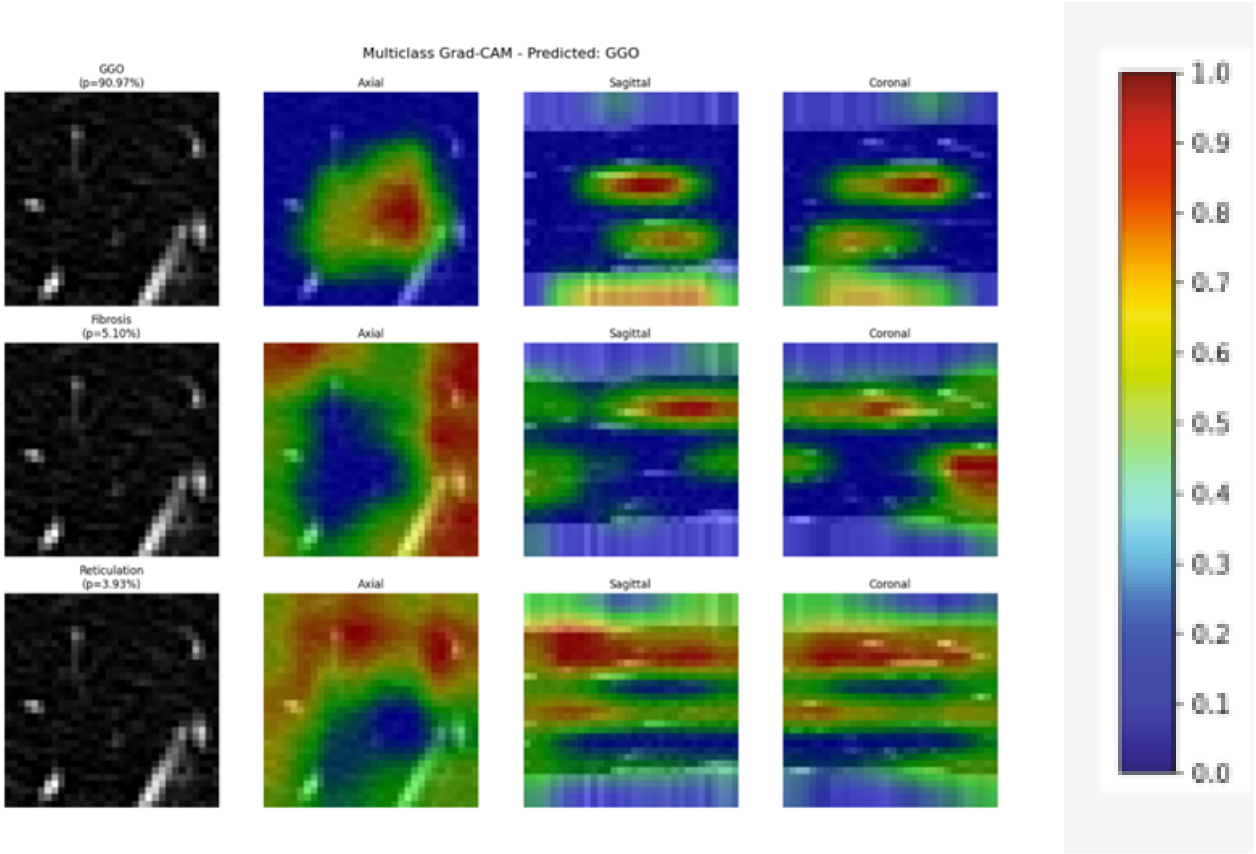

Figure 5.10: Grad-CAM overlay for a patch predicted as Ground Glass Opacity (GGO). Confidence: 94.

The table 5.3 summarizes key evaluation metrics.

Class	**Precision (%)**	**Recall (%)**	**F1-Score**
Ground Glass Opacity (GGO)	84.0	94.7	0.891
Fibrosis	90.1	83.9	0.869
Reticulation	93.0	87.5	0.902
Overall Accuracy	**88.73%**		
Weighted F1 Score	**0.887**		

Table 5.3: Per-Class Performance Metrics on the Test Set

5.5.2 Average class activation summary

To provide a global perspective of model attention, we aggregated Grad-CAM maps across multiple samples per class. The resulting heatmaps (Figure 5.11) illustrate common spatial activation patterns across correctly predicted samples for each ILD lesion.

- **Ground Glass Opacity (GGO):** Average accuracy of 85%. Activations were diffuse and located in mid-lung regions with hazy textures.
- **Reticulation:** Highest accuracy at 95%. Attention focused on structured, net-like fibrotic regions.
- **Fibrosis:** Lowest accuracy at 70%. Heatmaps were more dispersed, reflecting the variable and complex nature of fibrotic lesions.

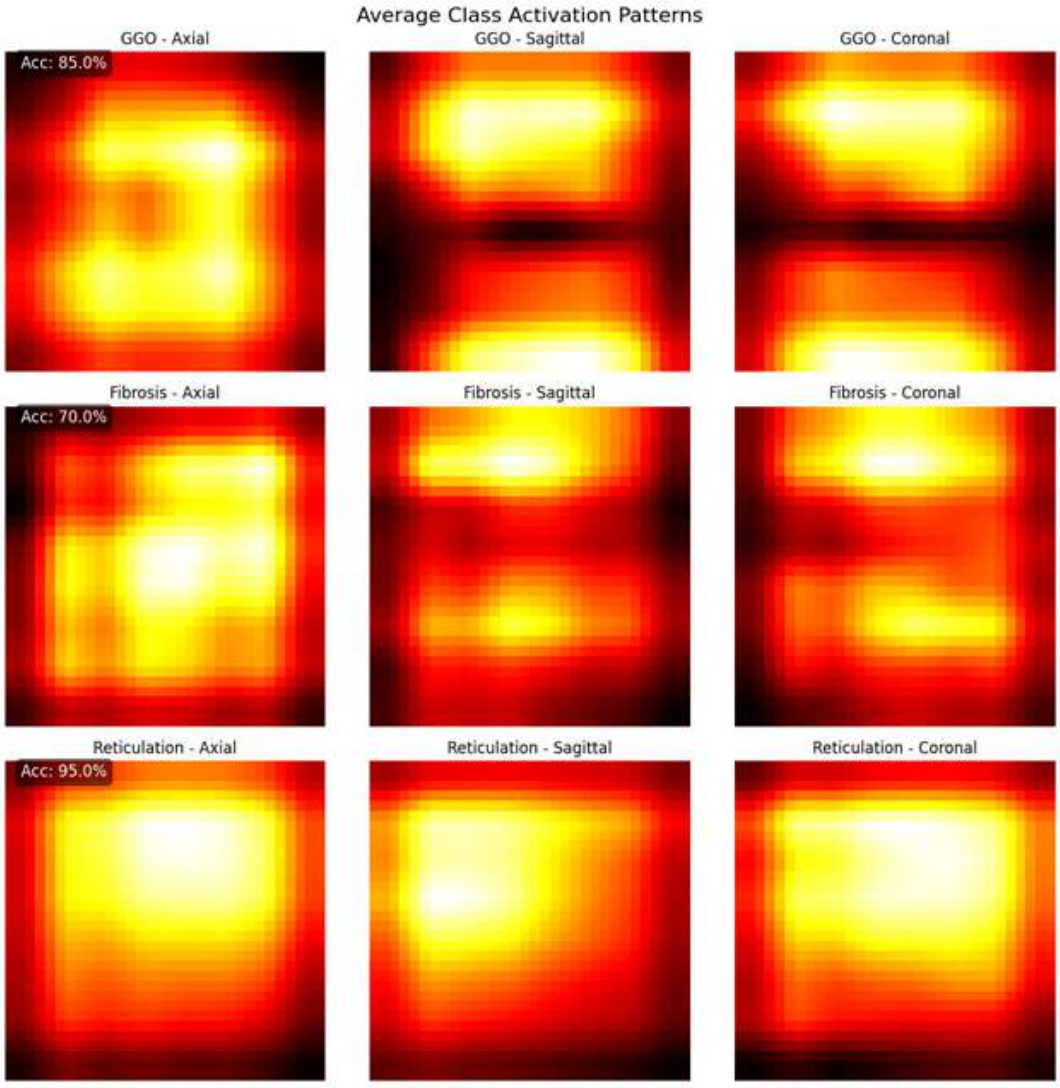

Figure 5.11: Average Grad-CAM maps across multiple correctly predicted samples for each ILD lesion type.

The Grad-CAM overlays confirm that the model learned clinically meaningful features and focused on relevant regions when predicting ILD subtypes. This interpretability supports model reliability and offers transparency that may improve trust in real-world deployment.

5.6 Patient-Level inference results

To assess the clinical viability of the proposed system, patch-level predictions were aggregated into patient-level diagnoses using confidence-based averaging. The patient-level evaluation was performed on a subset of ten test patients, covering all three ILD lesion types.

As shown in Figure 5.12, patients labeled with ground-glass opacity (GGO) tend to have sharply peaked softmax probabilities, reflecting consistent agreement across patch predictions. In contrast, patients with fibrosis exhibit a more diffuse confidence distribution, occasionally overlapping with GGO or reticulation, highlighting the inherent challenge in distinguishing early fibrotic changes.

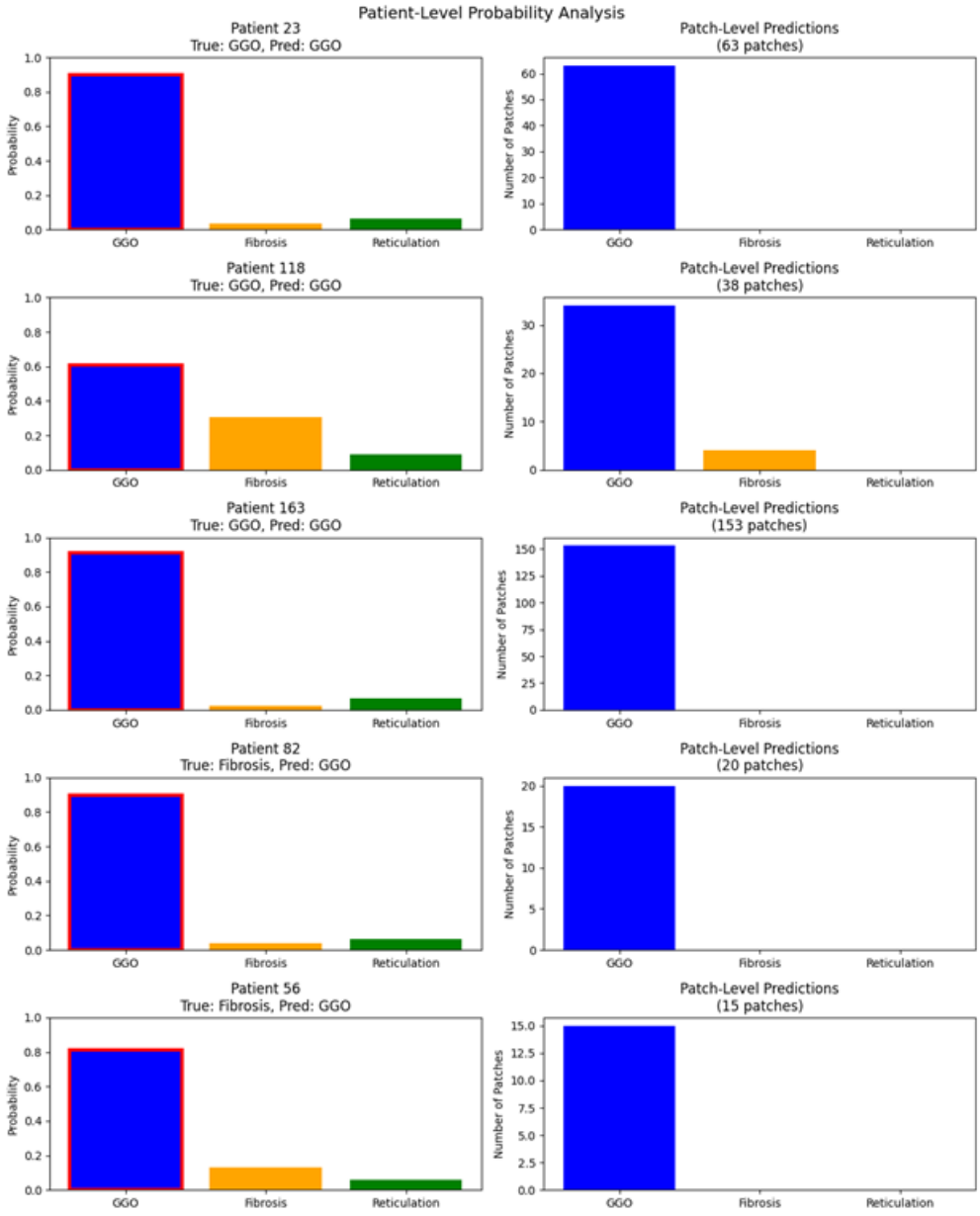

Figure 5.12: Patient-level class probability distributions across test patients.

Figure 5.13 illustrates the confusion matrix based on final patient-level predictions. GGO was correctly identified in most cases, achieving the highest classification accuracy. Reticulation was also consistently recognized with minimal confusion. However, fibrosis cases were frequently misclassified as GGO, underscoring the visual and radiological overlap between these two ILD patterns.

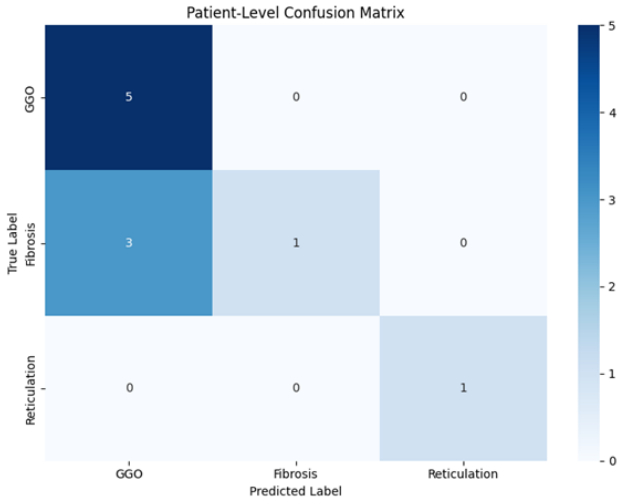

Figure 5.13: Patient-level confusion matrix using averaged softmax probabilities.

Class	Patients (Total)	Correctly Classified	Recall
Ground-Glass Opacity (GGO)	5	5	100%
Fibrosis	4	1	25%
Reticulation	1	1	100%
Overall	**10**	**7**	**70%**

Table 5.4: Patient-Level Classification Performance.

- **Overall accuracy:**

  The system achieved a patient-level classification accuracy of 70%, correctly identifying 7 out of 10 patients. This is notably lower than the patch-level accuracy (88.73%), highlighting the complexity introduced by patch aggregation and patient heterogeneity.

- **Class-Wise performance:**

  - **Ground-Glass Opacity (GGO):** Achieved 100% recall (5/5), consistently producing high-confidence predictions > 0.8.
  - **Fibrosis:** Demonstrated poor recall (25%), with only 1 out of 4 patients correctly classified. Most misclassifications were in favor of the GGO class.
  - **Reticulation:** Achieved 100% accuracy (1/1), with very high confidence 0.882 for Patient 128.

- **Misclassification patterns:**

  All three misclassified fibrosis cases (Patients 82, 56, 168) were incorrectly labeled as GGO. These errors were associated with:

  - High but ambiguous GGO probabilities 0.705–0.896
  - Limited patch count per patient (7–20 patches), leading to less stable aggregation

- **Noteworthy cases:**

  - **Patient 118 (GGO):** Despite having a minor mix of fibrosis patches (2/38), was correctly classified as GGO with moderate confidence (0.606).
  - **Patient 76 (Fibrosis):** Correctly predicted despite only three patches, suggesting robustness in certain low-data scenarios.
  - **Patient 128 (Reticulation):** Exceptionally reliable prediction supported by a large number of consistent patches (1,822).

These results emphasize both the promise and the limitations of patient-level diagnosis based on patch-wise CNN classification. They also reveal that class imbalance, limited patch coverage, and overlapping lesion appearances remain key challenges for real-world deployment.

5.7 Conclusion

The proposed pipeline showed strong performance in lung segmentation and binary lesion detection, achieving high Dice scores and sensitivity. While the multi-class classification results were promising, validation revealed challenges due to class imbalance and a potential risk of overfitting, indicating areas for further improvement in future work.

CHAPTER 6

CONCLUSION

This dissertation presented the design and development of a modular 3D deep-learning pipeline for automatic detection and classification interstitial lung diseases (ILD) lesions in high-resolution computed tomography (HRCT). In light of the diagnostic challenges and clinical burden of ILDs, we proposed a scalable approach, that combines anatomical lung segmentation, binary detection of pathology, and lesion level classification.

The pipeline follows a multi-stage structure. To begin, a 3D U-Net segments the lung focusing analysis on relevant anatomical structures. Then, a lightweight binary classifier (`Simple3DCNN`) performs patch-level classification to distinguish between healthy and pathological scans. For those scans predicted as pathological, a second stage performs multi-class classification of ILD lesions using a fine-tuned 3D CNN on balanced patch samples.

The testing was conducted using a preprocessed and filtered dataset of the MedGIFT ILD database. Strong performance was observed across all stages, including high Dice scores in segmentation, over 99% accuracy in binary detection, and approximately 88.7% macro F1 score for the 3-class lesion classification .

To enhance model interpretability, a 3D Grad-CAM method was integrated into the final classification stage. This visualization tool highlights the spatial regions that most influence the model's decision, providing initial steps toward clinical transparency and trust. These class activation maps were generated for selected patients and evaluated qualitatively to verify the model's attention on meaningful lesion regions.

In addition to patch-wise evaluation, patient-level classification was performed by aggregating predictions across volumetric patches. This strategy showed promising results, with high per-class precision and recall, demonstrating the model's ability to generalize diagnostic decisions at the patient scale.

Summary of Contributions

The main contributions of this dissertation are summarized as follows:

- **3D Lung Segmentation Module:** We developed a robust 3D U-Net-based pipeline for volumetric lung segmentation from HRCT scans. The system achieved a mean Dice coefficient of 0.9926, demonstrating high accuracy across patient scans.
- **Binary ILD Detection System:** A lightweight 3D classification module was implemented to distinguish healthy from pathological lungs, supporting automated triage and early screening.
- **Multi-Class Lesion Classification:** The system identifies and classifies lesion subtypes (ground-glass opacities, fibrosis, and reticulation) using a 3D CNN trained on segmented lesion patches, addressing the clinical need for pattern-specific ILD diagnosis.
- **Grad-CAM-Based Explainability:** Grad-CAM visualizations were integrated to highlight class-relevant lesion areas, improving model transparency and allowing radiologists to interpret decisions in clinically meaningful terms.
- **Lesion Volume Quantification:** Patient-level lesion volumes and lesion-to-lung ratios were computed to provide objective biomarkers for disease burden and potential severity scoring.
- **Modular AI Pipeline Design:** The architecture was structured into sequential modules, allowing stage-wise evaluation and adaptation for real-world clinical workflows.

Limitations

The limitations of this work are:

- **Limited Data and Class Imbalance:** The dataset is not only relatively small, but also unbalanced, with some ILD patterns underrepresented. This affects model generalization and robustness.
- **No End-to-End Volumetric Segmentation:** The system classifies individual patches but does not generate full-volume lesion maps, limiting lesion burden estimation and 3D visualization quality.
- **Absence of Clinical Metadata Integration:** The model is based purely on imaging without incorporating clinical characteristics such as age or pulmonary function tests (PFTs), which could improve diagnostic accuracy.
- **Limited External Validation:** All experiments were conducted using the MedGIFT dataset only. The generalizability of the model to other datasets, scanners, or clinical settings remains unverified.

Future Work:

We plan to follow up with several future directions to improve the system:

1. **Dataset Augmentation:** Expanding the dataset, especially for less common ILD patterns, to improve generalization and robustness.
2. **Advanced Learning Techniques:** Exploring semi-supervised learning, self-training, or transformer-based architectures to enhance classification performance in data-scarce conditions.
3. **Deployment Readiness:** Validating the system in real-world clinical settings through external testing and expert radiologist feedback.

In summary, this dissertation presents a novel 3D AI framework for ILD detection and pattern classification using HRCT data. While further development is required for clinical integration, the modular and interpretable system introduced in this work lays a solid foundation for future research into accurate and trustworthy computer-aided diagnosis of ILD.

BIBLIOGRAPHY

[1] G. Raghu, H. R. Collard, J. J. Egan, *et al.*, "An official ats/ers/jrs/alat statement: Idiopathic pulmonary fibrosis: Evidence-based guidelines for diagnosis and management", *American Journal of Respiratory and Critical Care Medicine*, vol. 183, no. 6, pp. 788–824, 2011.

[2] A. U. Wells, N. Hirani, *et al.*, "Interstitial lung disease: A clinical overview and general approach", *European Respiratory Review*, vol. 22, no. 128, pp. 102–115, 2013.

[3] D. A. Lynch, W. D. Travis, N. L. Müller, *et al.*, "High-resolution ct of idiopathic interstitial pneumonias", *Radiographics*, vol. 25, no. 4, pp. 777–794, 2005.

[4] S. K. Frankel, C. D. Cool, D. A. Lynch, and K. K. Brown, "High-resolution computed tomography findings in subacute hypersensitivity pneumonitis: Diagnostic considerations", *Journal of Thoracic Imaging*, vol. 19, no. 2, pp. 65–68, 2004.

[5] K. R. Flaherty, E. L. Thwaite, E. A. Kazerooni, *et al.*, "Radiological versus histological diagnosis in uip and nsip: Survival implications", *Thorax*, vol. 59, no. 2, pp. 143–149, 2004.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis", *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[7] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network", *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1–1, Feb. 2016. DOI: 10.1109/TMI.2016.2535865.

[8] L. Lu, Y. Zheng, G. Carneiro, and L. Yang, "Multilabel deep learning for fully automated multi-class labeling of chest ct images", *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[9] R. Silva de Araújo, J. Ferreira, G. Oliveira, and P. Azevedo-Marques, "Radiomics and deep learning fusion for interstitial lung disease classification on chest ct", *Computers in Biology and Medicine*, vol. 155, p. 106 617, 2023. DOI: 10.1016/j.compbiomed.2023.106617.

[10] G. Chassagnon, M. Vakalopoulou, N. Paragios, and M.-P. Revel, "Artificial intelligence applications for thoracic imaging", *European Journal of Radiology*, vol. 123, p. 108 774, 2020. [Online]. Available: https://doi.org/10.1016/j.ejrad.2019.108774.

[11] D. A. Lynch *et al.*, "Diagnostic criteria for idiopathic pulmonary fibrosis: A fleischner society white paper", *The Lancet Respiratory Medicine*, vol. 6, no. 2, pp. 138–153, 2018. DOI: 10.1016/S2213-2600(17)30433-2.

[12] P. Rivera-Ortega and M. Molina-Molina, "Interstitial lung diseases in developing countries", *Annals of Global Health*, 2019. DOI: 10.5334/aogh.2414.

[13] W. D. Travis, U. Costabel, D. M. Hansell, and et al., "An official american thoracic society/european respiratory society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias", *American Journal of Respiratory and Critical Care Medicine*, vol. 188, no. 6, pp. 733–748, 2013. DOI: 10.1164/rccm.201308-1483ST.

[14] A. Ketfi, F. Selatni, C. Djouadi, and R. Touahri, "Clinical and functional characteristics of interstitial lung disease in algeria: A single-center prospective study", *Journal of Respiration*, vol. 4, no. 1, pp. 12–25, 2024, ISSN: 2673-527X. DOI: 10.3390/jor4010002. [Online]. Available: https://www.mdpi.com/2673-527X/4/1/2.

[15] M. A. Althobiani, A. M. Russell, J. Jacob, *et al.*, "Interstitial lung disease: A review of classification, etiology, epidemiology, clinical diagnosis, pharmacological and non-pharmacological treatment", *Frontiers in Medicine*, vol. 11, p. 1 296 890, 2024. DOI: 10.3389/fmed.2024.1296890. [Online]. Available: https://doi.org/10.3389/fmed.2024.1296890.

[16] D. M. Hansell, A. A. Bankier, H. MacMahon, T. C. McLoud, N. L. Muller, and J. Remy, "Fleischner society: Glossary of terms for thoracic imaging", *Radiology*, vol. 246, no. 3, pp. 697–722, 2008.

[17] M. Remy-Jardin, F. Giraud, J. Remy, M. C. Copin, B. Gosselin, and A. Duhamel, "Importance of ground-glass attenuation in chronic diffuse infiltrative lung disease: Pathologic-ct correlation.", *Radiology*, vol. 189, no. 3, pp. 693–698, 1993, PMID: 8234692. DOI: 10.1148/radiology.189.3.8234692.

[18] T. Watadani, F. Sakai, T. Johkoh, *et al.*, "Interobserver variability in the ct assessment of honeycombing in the lungs", *Radiology*, vol. 266, no. 3, pp. 936–944, 2013, PMID: 23220902. DOI: 10.1148/radiol.12112516. [Online]. Available: https://doi.org/10.1148/radiol.12112516.

[19] E. Marchiori, T. Franquet, T. D. Gasparetto, L. P. Gonçalves, and D. L. Escuissato, "Consolidation with diffuse or focal high attenuation: Computed tomography findings", *Journal of thoracic imaging*, vol. 23, no. 4, pp. 298–304, 2008.

[20] J. Kim, B. Dabiri, and M. Hammer, "Micronodular lung disease on high-resolution ct: Patterns and differential diagnosis", *Clinical Radiology*, vol. 76, no. 6, pp. 399–406, 2021, ISSN: 0009-9260. DOI: https://doi.org/10.1016/j.crad.2020.12.025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0009926021000209.

[21] X. Li, X. Fang, Y. Bian, and J. Lu, "Comparison of chest ct findings between covid-19 pneumonia and other types of viral pneumonia: A two-center retrospective study", *European radiology*, vol. 30, pp. 5470–5478, 2020.

[22] Mayo Clinic Staff, *Copd – diagnosis and treatment*, Accessed on 27 May 2025, 2024. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/copd/diagnosis-treatment/drc-20353685.

[23] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, *Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation*, 2018. arXiv: 1802.06955 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1802.06955.

[24] Q. Jin, H. Cui, S. Sun, X. Wang, and L. Chen, "Lung segmentation using 2.5d convolutional neural networks on ct scans", *Journal of Healthcare Engineering*, vol. 2020, Article ID 8895181, 2020.

[25] J. Park, J. Yun, N. Kim, *et al.*, "Fully automated lung lobe segmentation in volumetric chest ct with 3d u-net: Validation with intra- and extra-datasets", *Journal of Digital Imaging*, vol. 33, no. 2, pp. 221–230, 2020. DOI: 10.1007/s10278-019-00223-1. [Online]. Available: https://doi.org/10.1007/s10278-019-00223-1.

[26] J. Wang, X. Liu, X. Chen, Y. Zhang, and Y. Liu, "Cascaded segmentation for interstitial lung disease based on deep convolutional neural networks", *Computers in Biology and Medicine*, vol. 123, p. 103 885, 2020.

[27] F. Zhang, M. Liu, R. Zhao, and X. Qiu, "Attention-gated 3d u-net for automatic segmentation of interstitial lung disease patterns", in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2021, pp. 2511–2518.

[28] J. H. Park, K. Y. Lee, S. H. Kim, and J. M. Choi, "Quantitative and visual analysis of interstitial lung disease based on 3d deep learning segmentation", *Medical Physics*, vol. 48, no. 8, pp. 4270–4281, 2021.

[29] J. Mei, X. Li, and L. Zhao, "Multimodal deep learning for ild classification using hrct and clinical data", *Computerized Medical Imaging and Graphics*, vol. 98, p. 102 153, 2023.

[30] S. L. Walsh, L. Calandriello, M. Silva, and N. Sverzellati, "Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: A case-cohort study", *The Lancet Respiratory Medicine*, vol. 6, no. 11, pp. 837–845, 2018.

[31] S. K. Zhou, H. Greenspan, and D. Shen, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises", *Proceedings of the IEEE*, vol. 108, no. 1, pp. 60–109, 2020.

[32] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, *3d u-net: Learning dense volumetric segmentation from sparse annotation*, 2016. arXiv: 1606.06650 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1606.06650.

[33] A. Shivdeo, A. Patil, *et al.*, "Comparison of 2d and 3d convolutional neural networks for detection of interstitial lung disease patterns in high-resolution ct scans", *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2021.

[34] D. O. Alebiosu, F. Hassan, A. Folayan, and M. Y. H. Yeow, "3d cnn outperforms 2d cnn for tuberculosis severity classification on chest ct: A comparative study", *Journal of Medical Imaging and Health Informatics*, 2025. [Online]. Available: https://www.ijcte.org/vol17/IJCTE-V17N1-1365.pdf.

[35] D. Guo, X. Liu, D. Wang, X. Tang, and Y. Qin, "Development and clinical validation of deep learning for auto-diagnosis of supraspinatus tears", *Journal of orthopaedic surgery and research*, vol. 18, p. 426, Jun. 2023. DOI: 10.1186/s13018-023-03909-z.

[36] 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*, Springer, 2015, pp. 234–241.

[37] O. Oktay, J. Schlemper, L. L. Folgoc, *et al.*, "Attention u-net: Learning where to look for the pancreas", in *arXiv preprint arXiv:1804.03999*, 2018.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization", in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[39] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps", in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[40] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "Automated pulmonary nodule detection via 3d convolutional neural networks", *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1188–1198, 2017.

[41] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013. DOI: 10.1109/TPAMI.2012.59.

[42] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P.-A. Poletti, and H. Muller, "Building a reference multimedia database for interstitial lung diseases", *Computerized Medical Imaging and Graphics*, vol. 36, pp. 227–238, 2012. DOI: 10.1016/j.compmedimag.2011.07.003.

[43] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation", *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[44] D. A. Clunie, "Dicom structured reporting and cancer clinical trials results", *Cancer Informatics*, vol. 2, pp. 33–56, 2006. DOI: 10.1177/117693510600200001.

[45] R. W. Cox, J. Ashburner, S. M. Smith, and et al., "A (sort of) new image data format standard: Nifti-1", *NeuroImage*, vol. 22, no. Supplement 2, e1440, 2004, Organization for Human Brain Mapping. [Online]. Available: https://nifti.nimh.nih.gov/.

[46] M. Cardoso, W. Li, T. Brown, *et al.*, "Monai: An open-source framework for deep learning in healthcare", *Computers in Biology and Medicine*, vol. 141, p. 105 111, 2022.

[47] F. Pérez-García, R. Sparks, and S. Ourselin, "Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning", *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106 236, 2021.

[48] Kaggle, *Kaggle: Your machine learning and data science community*, https://www.kaggle.com, 2024.

[49] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library", *Advances in neural information processing systems*, vol. 32, pp. 8024–8035, 2019.

[50] B. C. Lowekamp, D. T. Chen, L. Ibanez, and D. Blezek, "Design of a simplified interface for the insight toolkit", *Insight Journal*, vol. 2013, no. 1, pp. 1–18, 2013, https://simpleitk.readthedocs.io/.

[51] M. Brett, C. J. Markiewicz, *et al.*, *Nibabel: Access a cacophony of neuro-imaging file formats*, https://nipy.org/nibabel, Accessed: June 2025, 2020.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[53] M. L. Waskom, "Seaborn: Statistical data visualization", *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. DOI: 10.21105/joss.03021.

[54] J. D. Hunter, "Matplotlib: A 2d graphics environment", *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[55] D. Müller, I. Soto–Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation", *BMC Research Notes*, vol. 15, no. 1, p. 210, 2022. DOI: 10.1186/s13104-022-06096-y.