Politecnico di Torino

Politecnico di Torino

Computer Engineering

A.a. 2025/2026

Graduation Session March 2026

Quantitative CT Analysis of Airways and Parenchyma in Pulmonary Fibrosis

Characterization and FVC% Prediction

Supervisors:

Giuseppe Bruno Averta

João Pedrosa

Candidate:

Francesca Saglimbeni

Abstract

Segmenting the lungs and analyzing the results is fundamental for identifying possible pathological changes. Pulmonary fibrosis is one of the most common conditions characterized by specific alterations in the lung structure, involving both the vasculature and the airways.
The goal of this thesis is to analyze the airway tree, starting from its segmentation and continuing with the extraction of structural metrics. A primary focus is to quantify the variation of airway dimensions along their path from the carina and to map the branch distribution according to the Weibel generation model. Furthermore, the analysis incorporated parenchymal metrics that, according to the literature, are strongly correlated with fibrotic disease.
The proposed pipeline is developed and tested on CT scans from the CARVE14 and OSIC datasets. To validate the approach in the absence of a fully reliable structural ground truth, the correlation between the extracted metrics and the clinical parameter FVC% (available only for the OSIC data) measured at weeks 0 and 52 is evaluated. The most relevant metrics are subsequently used as input for various machine learning models to predict FVC% at 52 weeks, a time point of established clinical relevance, using only baseline CT scans (week 0). The results demonstrate the potential of quantitative airway and parenchymal analysis as a tool for disease characterization and functional prediction.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In recent years, the introduction of artificial intelligence in the medical field has significantly contributed to improving diagnostic analysis and support processes. The use of algorithms for the automatic analysis of images and clinical data has made it possible to reduce the complexity of many diagnostic tasks, assisting doctors in the evaluation of complex pathologies. The main goal of artificial intelligence systems in medicine is not to replace the clinician's role, but to provide reliable, consistent, and reproducible decision-support tools, capable of integrating the doctors' experience and specialization.

In the context of lung diseases, one of the most complex chronic conditions to diagnose and monitor is **pulmonary fibrosis**, it is a disease characterized by a progressive remodeling of lung tissue, leading to thickening of structures and a loss of lung elasticity, resulting in reduced respiratory capacity. Fibrosis affects not only the parenchyma but also vascular structures and airways, making its early identification particularly difficult through simple visual assessment of CT images. This thesis work fits into this context, proposing an approach for the analysis of pulmonary fibrosis through the study of the airway's structures. The study focuses on analyzing variations in airway structure during development, starting from the carina and moving through successive bronchial bifurcations. Abnormal structural alterations, particularly in airway diameters, can serve as a structural indicator of underlying fibrotic processes. Starting from the segmentation of the airways and their representation as a graph, several structural metrics were extracted and analyzed. These metrics provide a direct quantitative characterization of the airway tree, enabling an objective assessment of structural alterations that are indicative of the underlying disease level.

The analysis was conducted on CT scans from two distinct public competition datasets: **CARVE14** [1] and **OSIC Pulmonary Fibrosis Progression** [2]. The latter provides, in addition to images, associated FVC% values, making it particularly useful for validating the obtained results.

Given the lack of a fully reliable structural ground truth for airways, the validation strategy was based on a well-established clinical parameter: the percent forced vital capacity (FVC%). This index, which is available within the OSIC dataset, expresses the patient's FVC as a percentage of the predicted value for a healthy individual of comparable age, sex, and height.

A correlation analysis was first conducted to evaluate the relationship between the full set of extracted metrics (both airway and parenchymal) and the clinical parameter FVC%. The analysis was performed globally to identify which structural features showed the strongest association with the functional impairment measured by FVC%. Based on these initial results, the most relevant metrics were selected for a more granular, time-dependent analysis. This subsequent step assessed their correlation with FVC% separately at baseline (week 0) and at follow-up (week 52), as well as with the absolute decline in FVC% (the drop) between these time points.

The metrics demonstrating the strongest and most consistent associations across these analyses were then used as input features for several machine learning models, including Linear Regression, Random Forest, and a Multilayer Perceptron (MLP), with the aim of predicting the FVC% value at 52 weeks using only baseline CT scans (week 0).

1.1 Motivation and Clinical Context

The early and accurate diagnosis of pulmonary fibrosis represents one of the crucial challenges in modern medicine so identifying the disease at an early stage is indeed decisive for initiating antifibrotic therapies early, which can slow its course. The main difficulties encountered in diagnosis are clinical, symptomatic, and pathological. The first signs, exertional dyspnea and dry cough, are nonspecific and common to many other cardiopulmonary diseases (asthma, heart failure, COPD), often delay clinical suspicion.

Furthermore, fibrosis can present with different histological patterns (UIP, NSIP, etc.), and its distribution is typically heterogeneous and peripheral or sub-pleural, making a representative evaluation with targeted diagnostic methods problematic. This is compounded by the current lack of routine blood biomarkers capable of definitively diagnosing pulmonary fibrosis or reliably differentiating its various patterns.

In this context, Computed Tomography (CT) of the chest represents the cornerstone of diagnosis and follow-up because, in addition to characterizing the fibrotic picture with high precision, CT can identify the most representative and accessible areas for a potential transbronchial biopsy (cryobiopsy) or surgical biopsy, significantly increasing its diagnostic yield.

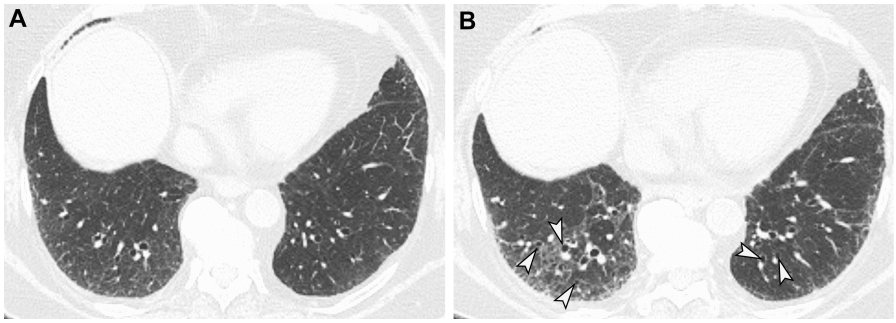


Figure 1.1: Progression of pulmonary fibrosis. (A) Axial CT image reveals mild interlobular septal thickening, subpleural irregularity, and fine reticular abnormality. (B) Follow-up axial CT image 2 years later shows increased reticulation, with new traction bronchiectasis or bronchiolectasis (arrowheads). The patient was diagnosed with IPF. Reprinted from [3].

1.2 Role of Artificial Intelligence in Medicine

Artificial intelligence is redefining the foundations of healthcare, introducing tools capable of analyzing large amounts of data with often superhuman speed and precision. In medicine, AI is not see as a substitute for the doctors but as a powerful ally, designed to reduce the clinician's cognitive load, enhance their decision-making capabilities, and, consequently, improve patient outcomes.

Despite this potential, the clinical implementation of AI is constrained by fundamental technological and ethical limitations, as already mentioned, the ultimate responsibility remains firmly in the hands of the doctors.

The main critical issues include:

1. the **blackbox** problem: advanced algorithms lack explanatory transparency. The inability to understand why an algorithm has formulated a specific recommendation raises issues of trust, accountability, and risks perpetuating potential biases hidden in the training data.

2. **generalizability and data bias**: an algorithm trained on a specific population may fail or be inaccurate when applied to different demographic contexts, thus risking amplifying existing health inequalities so to reduce this risk, continuous validation on external and representative datasets is necessary.

3. the doctors must possess and maintain the skills to verify, question, and, when necessary, override the algorithm's indications.

1.3 Pulmonary Fibrosis

When we talk about pulmonary fibrosis, we refer to a progressive and often irreversible process, in which healthy, elastic tissue is slowly replaced by rigid, non-functional scar tissue. Fibrosis is not just a microscopic alteration; it shapes the entire architecture of the organ, and Computed Tomography (CT) allows us to see what happens inside the lung without the need for invasive procedures.
One of the most characteristic effects is the traction-induced dilation and distortion of the airways caused by surrounding fibrotic tissue, along with a reduction in their terminal diameter. As a result, the bronchi and bronchioles appear widened, irregular, and constricted along their course, rather than following a regular path. This is a direct sign of the presence and evolution of the disease, which CT is able to document in detail, allowing not only diagnosis but also assessment of its severity and monitoring of its progression over time.

1.4 Airway Analysis as a Structural Biomarker

In a disease like pulmonary fibrosis, scar tissue generates continuous and disordered traction forces that visibly alter the airways, often even before the damage becomes clinically apparent. Unlike other more subjective radiological signs, the size and structure of the bronchial tree offer a quantitative and objective parameter, ideal for computational analysis becoming a sort of *structural signature* of the disease so a reliable process **biomarker**.
In a healthy lung, airway diameter progressively decreases from the root to the periphery, whereas in fibrosis, the contracting scar tissue exerts radial traction on the bronchial walls, causing abnormal and irregular dilation, thereby making the measurement of these variations a direct index of the fibrotic process's extent and distribution.
The disease also profoundly alters the architecture of the bronchial tree: bifurcations may appear abnormally angled or increased in number, branches may be deviated, and peripheral branching may be prematurely truncated. These changes not only reflect the severity of fibrosis but also reveal its spatial distribution, testifying to an active remodeling of the entire lung structure.
To enhance accuracy and clinical relevance, parenchymal metrics were strategically integrated into the analysis. This addition is motivated by evidence that these metrics exhibit a stronger correlation with fibrotic disease than airway-derived measurements alone, precisely because densification and textural alterations of the lung tissue are classic, well-documented hallmarks of fibrosis and directly reflect disease severity. The chosen metrics are Histogram Entropy and Mean Lung Density. When combined with airway geometry, they enable the model to capture a more

comprehensive and pathophysiologically coherent representation of the fibrotic process.

1.5 Thesis Objectives

This thesis aims to develop and validate a new imaging biomarker for the objective and personalized assessment of pulmonary fibrosis. The general objective of the work is to develop an automatic and quantitative methodology based on Chest Computed Tomography (CT) images to analyze airway morphology and establish to what extent their alterations are correlated with the presence, severity, and progression of the fibrotic disease. In summary, we want to answer the question: how much do the airways, in their shape and size, tell us about the disease state of the lung?

To translate this ambition into a concrete research path, the work will proceed through a series of sequential and interconnected steps. It will start with the segmentation of the entire bronchial tree from CT scans using a specific tool, TotalSegmentator[4]. Once the final segmentation is obtained, the focus will shift to the identification and calculation of the most significant quantitative descriptors, which capture the deformations induced by fibrosis.

The central objective is therefore to integrate these morphological measures into a score that synthesizes the patient's fibrotic burden. Ultimately, the clinical validation of this score will represent a critical step: its significance will be verified by comparison with established clinical parameters, specifically the previously discussed percent-predicted Forced Vital Capacity (FVC%).

1.6 Main Contributions

This work proposes a fully automated pipeline for quantitative airway morphological analysis from chest CT images. The pipeline automates all steps, from segmentation and bronchial graph extraction to detailed branch analysis and the calculation of a comprehensive set of structural and parenchymal parameters.

Given the lack of a complete and reliable structural ground truth, an innovative validation strategy was developed and implemented using the FVC% parameter as a form of weak supervision. Quantitative parameters derived from the pipeline are correlated with FVC% to assess their clinical relevance, and the most predictive ones are then used as input for machine learning models to predict future FVC% values from baseline scans.

This approach provides a robust method for validating CT-derived biomarkers.

1.7 Thesis Structure

The present thesis is structured to guide the reader through a logical progression, from the foundational clinical problem to the proposal, validation, and critical evaluation of the proposed methodological solution.
The journey begins with an introduction that establishes the clinical and motivational context for the research. This opening section outlines the primary objectives of this work, summarizes its main scientific contributions, and provides a roadmap for the entire document.
To ground the study in existing knowledge, a comprehensive State of the Art follows. Here, the relevant anatomical and pathophysiological foundations of pulmonary fibrosis are detailed. It also reviews essential medical imaging techniques, specifically computed tomography (CT), alongside established methods for lung and airway segmentation. Finally, this section critically surveys the existing literature on using airway morphology as a potential biomarker for disease.
The narrative then turns to the practical foundations of the study in a section dedicated to data and preprocessing where is described the specific datasets used in this research, detailing the preprocessing procedures applied to the CT images. Following this, the core methodology is laid out in detail. This chapter illustrates the entire proposed analysis pipeline step-by-step, from the initial airway segmentation and the subsequent extraction of a tree-like graph representation, to the branch-level analysis that computes the full set of structural and parenchymal metrics.
The subsequent section focuses on pipeline validation, including assessments of segmentation quality and the structural consistency of the extracted graphs. This section elaborates on the central validation strategy: the correlation of the extracted quantitative metrics with a clinical functional measure (Forced Vital Capacity percent), and their subsequent use for functional prediction.
The results of the experimental work are then systematically presented. Building on the results, a dedicated discussion section offers a deeper clinical and methodological interpretation, analyzing the robustness and limitations of the proposed approach. Finally, the thesis culminates with conclusions and future work.

Chapter 2

State of the Art

The automated analysis of pulmonary airways from CT imaging represents a rapidly evolving field that addresses a critical clinical need: the objective and reproducible quantification of airway alterations in lung diseases, particularly pulmonary fibrosis. Traditional radiological assessment relies on subjective visual evaluation, which suffers from inter-observer variability and limited reproducibility[5] so recent advances in deep learning and computational image analysis have enabled the development of automated pipelines capable of extracting quantitative biomarkers from routine chest CT scans.

This chapter reviews the main approaches and methodologies relevant to this thesis, focusing on four key areas: airway segmentation techniques, morphometric analysis frameworks, machine learning approaches for small datasets, and clinical validation strategies.

2.1 Airway Segmentation Approaches

Early approaches to airway segmentation relied on traditional image processing techniques such as region growing, thresholding, and morphological operations but these methods struggled with anatomical variability and pathological alterations characteristic of fibrotic disease.

The introduction of deep learning has transformed segmentation accuracy for example **TotalSegmentator** [6], the framework employed in this thesis, represents a significant advancement in multi-organ segmentation. Trained on a large and diverse dataset, it can automatically segment 104 anatomical structures and its generalpurpose design provides robust baseline segmentation across different patient populations and acquisition protocols.

A critical challenge in airway segmentation is maintaining topological continuity,

particularly for thin peripheral airways and in the presence of pathological alterations, in fact discontinuities in the segmented airway tree compromise subsequent analyses such as skeletonization and graph construction.
Recent work has addressed the challenge of topological continuity in airway segmentation through sophisticated post-processing and architecture-aware strategies. García-Uceda et al. [7] demonstrated that 3D U-Net architectures can effectively segment airways from CT scans when trained on sufficient annotated data, achieving good performance on central airways but struggling with peripheral branches. Their work highlighted the fundamental trade-off between sensitivity (capturing thin peripheral airways) and specificity (avoiding false positives in noise or vessels).

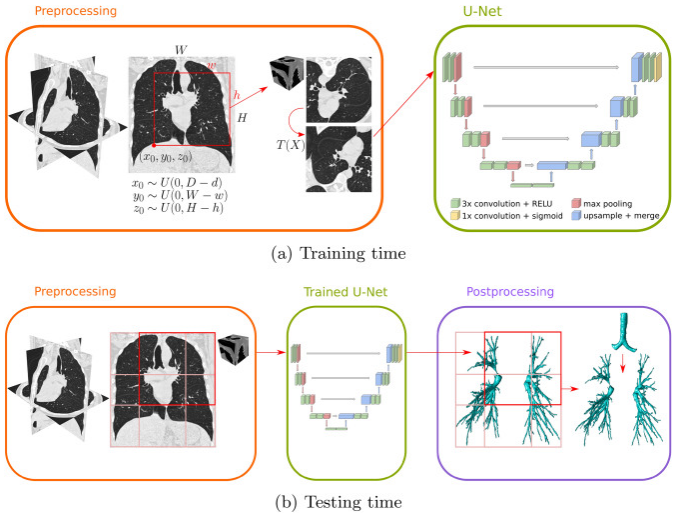


Figure 2.1: Schematics of the proposed airway segmentation method. (a) Overview at training time, showing the extraction of 3D random patches from the CT scan, to feed as input to the U-Net. (b) Overview at testing time, showing the extraction of 3D patches through sliding-window, and the generation of the airway tree segmentation from the patch-wise output of the U-Net. This image is taken from García-Uceda et al. [7] paper.

Cheung et al. [8] proposed a data-centric approach to airway segmentation that emphasizes training data quality and annotation consistency over pure architectural innovation. Their work demonstrated that careful curation of training examples, particularly for difficult cases with pathological alterations, can significantly improve model generalization without requiring complex multi-stage pipelines.
In terms of Topology-Preserving Method for Airway Segmentation, we have one important work in literature, Zhang et al. [9] introduced topology-aware loss functions and post-processing strategies specifically designed to maintain airway tree connectivity. Their AirMorph framework incorporates graph-based reasoning

to detect and correct topological errors, such as missing connections between parent and child branches, which are particularly common in diseased lungs where airways may be compressed or distorted by fibrotic tissue.
However, these advanced methods typically require either extensive training on annotated airway datasets or complex multi-stage pipelines with carefully tuned hyperparameters. For research focused primarily on morphometric analysis rather than segmentation methodology advancement, a robust pre-trained segmentation tool combined with targeted post-processing represents a pragmatic alternative.
Since the primary objective of this thesis is the development and validation of morphometric analysis rather than the advancement of segmentation algorithms, TotalSegmentator [10] was selected as the foundation. Its general-purpose design, trained on a large and diverse dataset, provides robust baseline segmentation across different patient populations and acquisition protocols. The strategy adopted here is to accept TotalSegmentator's output as a starting point and apply intelligent post-processing specifically designed to address the requirements of subsequent graph-based morphometric analysis—gap filling for topological continuity, artifact removal based on Hounsfield unit consistency, and conservative trachea removal to preserve the carina landmark.
This approach allows the pipeline to focus computational and development effort on the novel contributions of this work: robust carina detection, dual-mask morphometry, and correlation with functional outcomes, rather than replicating segmentation research already extensively covered in the literature.

2.2 Morphometric Analysis and Clinical Validation

The validation of automated segmentation pipelines requires a systematic approach that goes beyond simple visual evaluation in fact the validation of automated segmentation pipelines requires a systematic approach that extends beyond visual inspection.
Recent studies have proposed technical validation frameworks based on morphological and physical criteria [11], including verification of the physical plausibility of geometric parameters (e.g., tapering ratios, maximum reachable generations), checks for over-segmentation (total volumes, branch counts), and validation against known anatomical models such as Weibel's model [12].
Lo et al. [11] demonstrated that automated validation metrics can effectively identify pipeline failures before clinical application, reducing the need for manual quality control. Their framework evaluates segmentation reliability through multiple independent criteria: volume plausibility, generation depth consistency, peripheral-to-central ratio bounds, and tapering ratio constraints.

In the specific context of fibrotic lung disease, Pakzad et al. [13] developed and evaluated an automated computational pipeline, AirQuant, for airway quantification in CT. Their work demonstrated the clinical relevance of morphometric measurements, such as airway count and generation analysis, for assessing disease severity and progression in patients with interstitial lung diseases.

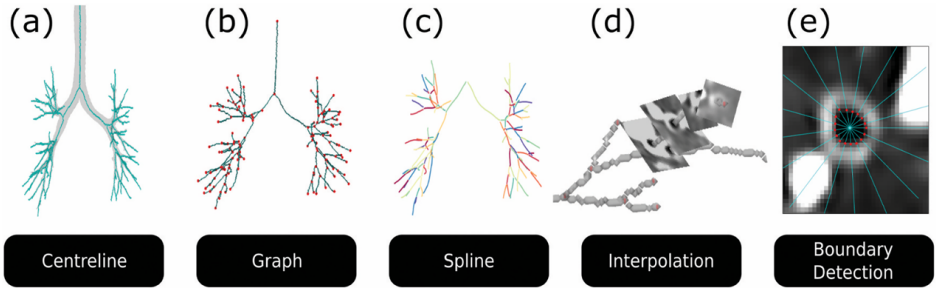


Figure 2.2: Demonstrating the AirQuant pipeline graphically from left to right for a given airway lumen segmentation through to the end where the lumen boundary is established taken from Pakzad et al. [13] paper.

AirQuant in fact represents the most comprehensive open-source framework for automated airway morphometry. Starting from deep learning-based segmentation, AirQuant performs lobe-based and generation-based subdivision of the bronchial tree, enabling anatomically stratified analysis.
The framework calculates three core metrics: **inter-branch tapering**, which assesses diameter changes between branches against the healthy standard of Weibel's law [12]; **tortuosity**, which quantifies the abnormal curvature of airways caused by fibrotic traction; and **generation-specific metrics**, which allow for localized comparison of structural changes against a theoretical anatomical model.
The analysis demonstrates that AirQuant provides a robust automated method for discriminating between healthy and IPF-affected lung morphology and its open-source framework has subsequently facilitated wider academic use.

2.2.1 Peripheral Airway Metrics

Peripheral airway analysis presents particular challenges in chest CT imaging due to the limited spatial resolution of routine CT and the small caliber of distal branches, so assessment often relies on indirect or aggregate descriptors, such as the peripheral-to-central distribution ratio.
In obstructive lung disease, CT-derived airway distribution metrics have been proposed to detect the relative loss of peripheral versus central airways; for example, Sasaki et al. [14] evaluated peripheral-to-central airway ratios, and Kirby et al. [15]

demonstrated that a CT-derived total airway count is associated with the risk of COPD progression, supporting the concept that airway visibility and the number of reconstructed branches contain clinically meaningful information beyond single-layer wall measurements.

For fibrosing interstitial lung disease, the interpretation of peripheral airway metrics is less straightforward because traction bronchiectasis and architectural distortion can increase the apparent visibility of the airways on CT. Consistent with this, micro-CT and stereology studies reported early small-airway loss and remodeling in IPF [16]. More recently, the study by Wang et al. [17] reported that functional impairment of the small airways and generation-specific airway changes on CT, particularly in the more distal generations, can also be detected in early-stage IPF.

In the context of this thesis, airway metrics such as distal branch counts/volumes, peripheral-to-central ratios, and distal diameter dispersion are therefore treated as composite biomarkers that capture both loss/remodeling and traction-related changes. These descriptors are aligned with existing automated airway quantification pipelines in fibrosing lung disease such as AirQuant [13].

A fundamental question for any imaging biomarker is its relationship to clinically meaningful outcomes. Beyond diameter and branching patterns, volumetric measurements provide an additional dimension of characterization and can capture both airway remodeling and lung parenchymal distortion.

Recent studies have further emphasized that airways do not exist in isolation and that predictive performance improves when multiple anatomical compartments are quantified jointly. Thillai et al. [18] demonstrated that deep learning-based segmentation, which allows visualization of lung, vascular, and fibrosis volumes on CT, can then provide prognostic information on progression and mortality in idiopathic pulmonary fibrosis (IPF).

This multicompartmental philosophy inspired the decision of this thesis to integrate parenchymal metrics such as mean lung density and histogram entropy with airway-derived features in the final predictive model since the parenchyma reflects the primary fibrotic process, while the airway reflects its mechanical consequences through traction and distortion [19].

2.3 Machine Learning with Limited Clinical Data

The application of machine learning techniques in clinical contexts is often constrained by the scarcity of annotated data. Large-scale datasets with comprehensive longitudinal follow-up and expert annotations are expensive to create and rarely available, particularly for rare diseases like IPF, so this data limitation poses significant challenges for model development and validation.

For datasets with fewer than 50 patients, approaches such as Ridge Regression

with L2 regularization and Random Forest with depth constraints have demonstrated greater robustness compared to deep neural networks [20]. Ridge regression addresses multicollinearity and overfitting by penalizing large coefficient values, allowing stable estimation even when the number of features approaches the number of observations.
Random Forest mitigates overfitting in small samples by constraining tree depth (typically 3–5 levels) and enforcing minimum sample sizes for splits. Its ensemble approach averaging predictions across hundreds of bootstrapped trees reduces variance while preserving model stability, making it well suited for noisy biomedical data [21].
Deep neural networks, despite their success in large-scale image classification, struggle with limited data due to their high parameter count relative to sample size. With fewer than 50 patients, even shallow networks tend to overfit training data and fail to generalize to unseen cases. Regularization techniques can mitigate overfitting, but fundamentally, deep learning requires larger datasets or transfer learning from related domains to achieve reliable performance.
Leave-One-Out Cross-Validation (LOOCV), while computationally expensive, provides the most reliable performance estimates when the sample size is limited [22]. In LOOCV, each patient serves once as a test case, with the model trained on all remaining patients, this maximizes training set size for each fold and provides an approximately unbiased estimate of generalization error. Because training sets overlap heavily, LOOCV estimates can have high variance and does not account for model selection bias when hyperparameters are tuned on the same data used for error estimation. To address this, nested cross-validation is recommended: an outer LOOCV loop for error estimation and an inner cross-validation loop for hyperparameter tuning [23].

2.4 Longitudinal Data and Temporal Interpolation

Pulmonary function tests (PFTs) are the gold standard for monitoring disease progression in IPF, but their timing is irregular and varies across patients due to clinical scheduling constraints, a lot of studies must often deal with incomplete or non-aligned time series, where measurements for different patients occur at different weeks post-diagnosis.
The interpolation of longitudinal spirometric measurements requires careful attention to data quality and temporal distribution. In this thesis, a pragmatic windowed approach was adopted, prioritizing measurements closest to the target timepoint ($\pm$5 weeks for baseline, 40–65 weeks for annual follow-up) to balance accuracy and dataset completeness. This approach recognizes that measurements very close to

the target date are nearly equivalent to the target value, while measurements far from the target introduce increasing uncertainty, a standard concern in longitudinal data analysis [24].
For missing timepoints with no measurements in the preferred window, linear regression can be applied when supported by ≥ 2 measurements with good temporal coverage and reasonable correlation ($r > 0.5$) [25] and the validity of linear interpolation rests on the assumption that FVC decline is approximately linear over short intervals (6–12 months), an assumption supported by longitudinal trajectory analyses in IPF [26].
Fitzmaurice et al. [25] emphasize that when interpolation is necessary, uncertainty should be propagated through the analysis. For instance, if FVC at week 52 is estimated from measurements at weeks 36 and 68, the standard error of the estimate should reflect both measurement variability and temporal extrapolation distance

Chapter 3

Datasets and Data

This study utilized two distinct datasets, both consisting of unenhanced thoracic computed tomography (CT) scans.
The **CARVE14**[1] dataset is comprised of 55 volumetric thoracic CT scans, acquired without the administration of contrast medium. These scans were obtained from a lung cancer screening cohort. The patients included in this dataset present with emphysema and interstitial lung diseases of varying severity, a characteristic that makes the dataset particularly challenging for the automatic analysis of vascular structures.
The **OSIC**[2] (Open Source Imaging Consortium) dataset was developed as part of a Kaggle competition focused on predicting the progression of pulmonary fibrosis. The primary objective is to predict the severity of pulmonary fibrosis based on thoracic CT scans. Lung function is assessed through the output of a spirometer, which measures **Forced Vital Capacity (FVC)**, defined as the volume of air expelled during a forced expiratory maneuver. FVC is a crucial indicator of respiratory function and its deterioration over time.

3.1 CT Preprocessing

All CT scans from both datasets were converted from the MHD+RAW (MetaImage) format to the NIfTI (Neuroimaging Informatics Technology Initiative) format, which is widely used in the neuroimaging and medical imaging community. This conversion was performed using the SimpleITK library, preserving all spatial metadata (spacing, origin, direction) and intensity information in Hounsfield Units (HU).
The preprocessing pipeline, initially developed and tested on the CARVE14 dataset, provided optimal results for the subsequent construction of the vascular graph.

However, applying the same pipeline to the OSIC dataset revealed several critical issues related to the CT acquisition characteristics of this dataset:

1. **Inadequate number of slices**: some acquisitions had an extremely low number of slices (e.g., only 30 slices per scan), insufficient for a reliable volumetric reconstruction of the pulmonary vascular structures.

2. **Variability in slice thickness**: high variability in slice thickness compromising the quality of 3D reconstruction.

3. **Non-uniformity in slice spacing**: In some acquisitions, the slices were not uniformly spaced along the z-axis, with variable gaps between consecutive slices introducing artifacts in the volumetric reconstruction.

To address these issues, a multi-criteria filtering pipeline based on rigorous technical parameters was implemented (Table 3.1).
The geometric characteristics of CARVE14, specifically in terms of slice thickness and axial resolution served as the baseline to establish filtering thresholds for the OSIC data set, this ensured that only acquisitions with physical properties compatible with the segmentation algorithms and subsequent graph construction were retained.

Criterion	**Min**	**Max**
Number of slices	300	1500
Slice thickness	0.5 mm	1.5 mm
XY resolution	0.3 mm	1.1 mm
Std slice spacing	–	1.0 mm
Maximum gap between slices	–	2 X avg spacing

Table 3.1: Filtering Criteria Applied to the OSIC Dataset

The application of these filtering criteria revealed significant quality heterogeneity within the original OSIC dataset (Figure 3.1), characterized by high variability in acquisition parameters. While XY resolution was mostly adequate, a substantial portion of the scans failed to meet requirements for slice thickness or total slice count. In particular, the insufficient number of slices was the primary cause of exclusion, as it prevented a complete and accurate capture of the pulmonary vascular tree.

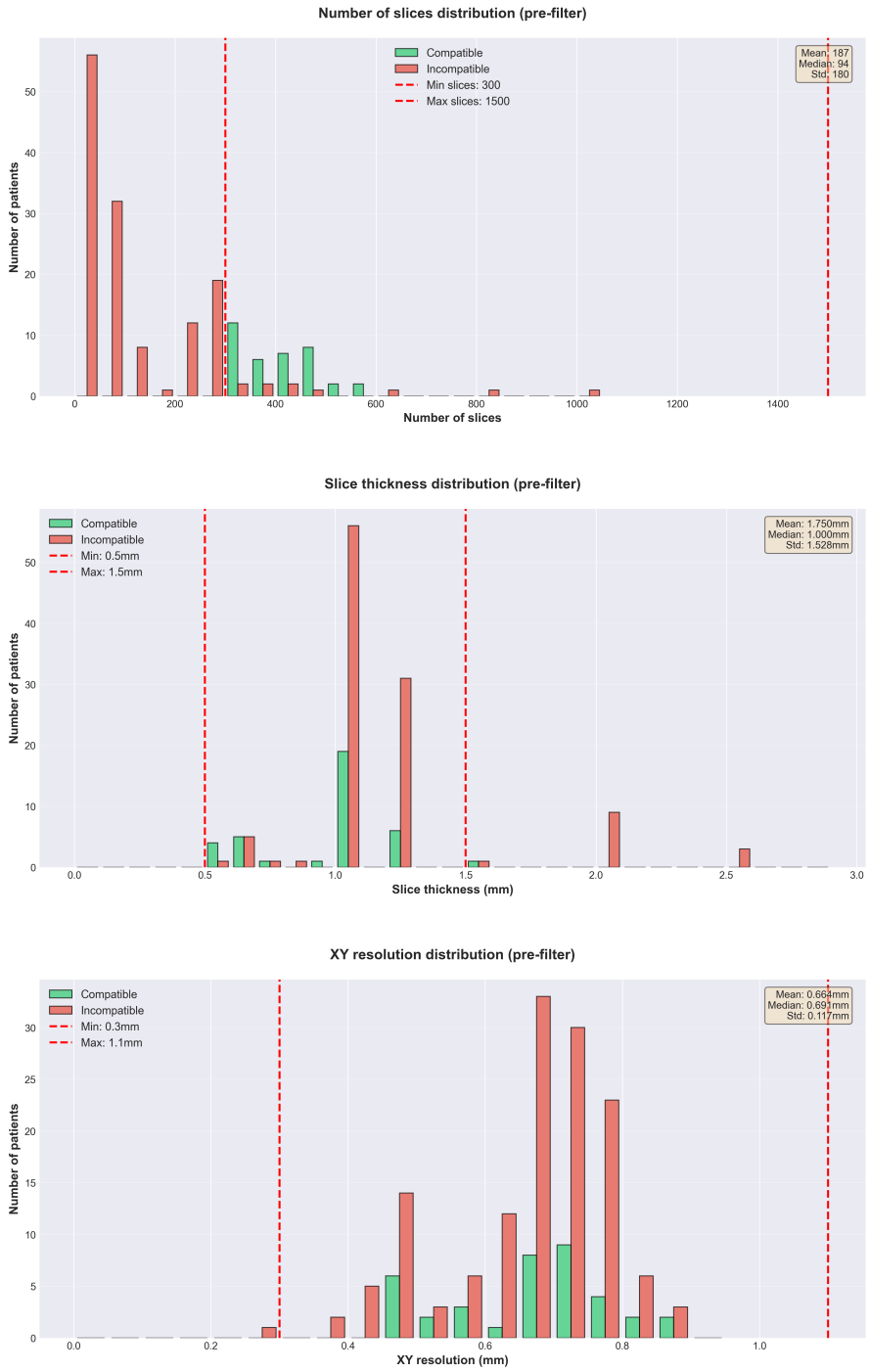


Figure 3.1: Pre-filtering distributions of the OSIC dataset metrics

All exclusion criteria discussed above are based solely on geometric and slice uniformity parameters. An additional source of heterogeneity was identified in the variability of the reconstruction kernels used. Some OSIC scans utilized high-frequency kernels (sharp kernels, such as Siemens' B60f, B70f, B80f or Toshiba's FC51, FC53) designed for high-resolution bone or lung imaging. These kernels, while providing greater anatomical detail, introduce high-frequency noise that can compromise the performance of automatic segmentation algorithms.
For these scans, a selective smoothing procedure was implemented by systematically evaluating four different 3D filtering methods: Median, Gaussian, Bilateral, and Curvature Flow filters.

1. **Median Filter**: it's a non-linear filter, it is highly effective at removing noise while preserving sharp edges, as it replaces voxel intensity with the median value of its neighborhood.

2. **Gaussian Filter**: it is a linear smoothing filter, performs uniform smoothing via convolution with a Gaussian kernel, efficiently suppressing high-frequency noise but potentially blurring fine edges.

3. **Bilateral Filter**: as a non-linear filter, it extends the Gaussian filter by introducing a second kernel in the intensity domain. It smooths homogeneous regions while inhibiting smoothing across strong intensity gradients, making it particularly suited for medical images where structure boundaries are critical.

4. **Curvature Flow Filter**: it is an example of anisotropic diffusion techniques, this filter iteratively smoothens the image based on local curvature, promoting intra-region smoothing while inhibiting inter-region smoothing across boundaries.

The performance (Table 3.2) of each filter was evaluated using a comprehensive set of quantitative metrics:

- **SNR (Signal-to-Noise Ratio)** in the Lung Region: measures how much the lung tissue is distinguished from the background, a higher value indicates a cleaner image, providing a better basis for segmentation.

- **Noise Reduction**: apositive percentage indicates effective noise suppression so a good primary efficacy, while a negative value suggests the introduction of smoothing artifacts.

- **Edge Preservation**: it measures the filter's ability to maintain the sharpness of anatomical boundaries, such as vessel walls.

- **Smoothness**: represents the degree of uniformity introduced the main problem is that a very high value may lead to blurred fine structures, while a very low value may indicate insufficient noise removal.

- **Histogram Similarity**: Evaluates how well the global intensity distribution is preserved, more it is close to 1.0 more the filter maintaining consistent segmentation thresholds.

- **Processing Time**: determines the feasibility of applying the filter to the entire dataset.

- **PSNR (Peak Signal-to-Noise Ratio)**: a traditional measure of reconstruction fidelity where an higher values indicate a filtered image that remains closer to the original.

Metric	Median	Gaussian	Bilateral	Curvature Flow
Lung SNR	8.34	8.20	8.06	8.26
Noise Reduction (%)	-0.7	-0.5	-0.7	-0.2
Edge Preservation (%)	68.1	78.2	93.6	70.4
Smoothness (%)	51.8	46.7	11.5	51.5
Histogram Similarity	0.957	0.956	0.997	0.951
Runtime (s)	12.48	1.67	9.11	11.45
PSNR (dB)	31.3	32.5	48.7	32.2

Table 3.2: Comparison of Smoothing Methods

The analysis of these metrics reveals that the **gaussian filter** presents the optimal compromise for our pipeline. Although the bilateral filter achieves superior edge preservation (93.6%) and histogram similarity, its prohibitive computational cost (more than 900 seconds per scan) makes it impractical for processing an entire dataset. The Gaussian filter, on the contrary, offers excellent edge preservation (78.2%), combined with exceptional computational efficiency (1.67 seconds), high SNR, and minimal noise amplification. This balance makes it the most suitable choice for ensuring both segmentation accuracy and pipeline scalability.

However, for the final filter selection, a manual, scan-by-scan analysis of the results within the pipeline was performed. This involved visualizing the final segmentation and graph construction output, ensuring the practical effectiveness of every component in the entire analysis pipeline.

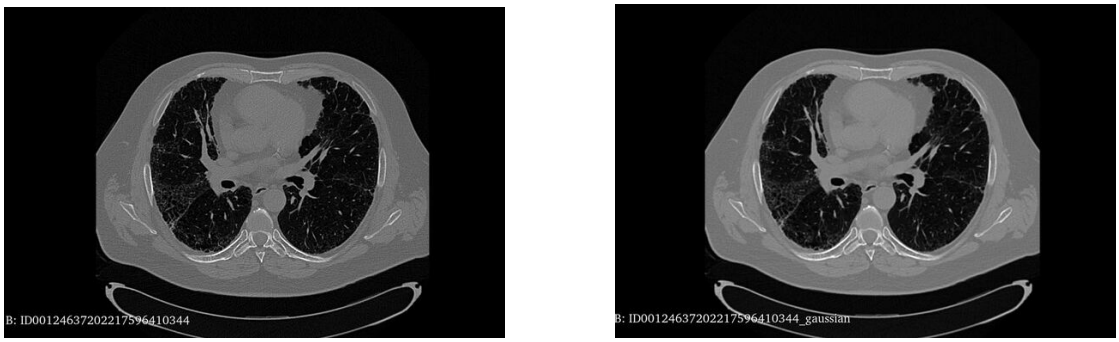


Figure 3.2: Comparison between the original CT slice and the result of the Gaussian filter

By sequentially applying geometric filtering criteria and selective smoothing, the OSIC dataset was reduced from approximately 200 initial cases to a validated subset of 45 CT scans. These selected acquisitions meet the technical requirements for automatic segmentation, maintain compatibility with the TotalSegmentator tool, and are associated with usable FVC measurements for downstream analyses.

3.2 Lack of Ground Truth and Related Challenges

A key methodological limitation of this study is the absence of a reference structural ground truth (GT) for validating the vascular graph. Although the OSIC dataset provides functional data in the form of Forced Vital Capacity (FVC), the derived FVC percent (FVC%) which represents the patient's FVC relative to the predicted value for a person of similar characteristics, presents challenges as a direct reference. Primarily, a significant temporal misalignment exists, as the FVC measurements often do not coincide with the baseline CT scan (recorded as Week 0). Consequently, FVC% cannot serve as a direct validator for the graph's structural features but serves as a clinical validator for the overall disease level, providing a functional benchmark against which the derived quantitative metrics are assessed.

Chapter 4

Initial Pipeline: Airway Segmentation and Feature Extraction

This thesis proposes an integrated pipeline for the quantitative assessment of pulmonary fibrosis from chest CT images, addresses the clinical need to obtain an objective and reproducible characterization of the morphological changes in the airways associated with fibrotic disease, overcoming the limits of qualitative visual analysis.

The overall architecture follows a precise logic: it begins with the segmentation of anatomical structures, proceeds through phases of cleaning and geometric reconstruction, and culminates in extracting the quantitative airway and parenchymal metrics essential for disease characterization.

The first module involves the segmentation of the airways using the **TotalSegmentator**[4] framework, followed by cleaning functions to preserve the distal branching while removing artifacts.

In the second module is implemented a conservative removal of the trachea based on the accurate detection of the carina, thus isolating the bronchial tree for subsequent analysis. Preprocessing and reconnection techniques are applied at every stage, each in a distinct way, but this approach ensures the topological continuity of the structure.

The next step involves creating a graphical representation of the bronchial tree using skeletonization and topological analysis algorithms. Following this, advanced airway and parenchymal metrics are calculated, such as volume, diameters, and others well described in the next chapters. These metrics are then used to perform a comprehensive quantitative structural and functional analysis of the disease, IPF.

An innovative aspect of architecture is the implementation of a **dual-mask strategy**, which maintains in parallel an original version of the segmentation for metric precision and a refined version for topological integrity, thus optimizing both measurement accuracy and structural analysis reliability.
The system's final output includes not only all computed metrics and a comprehensive summary of the quantitative analysis, but also a structured report that documents every intermediate phase, ensuring transparency and reproducibility. The pipeline is designed to operate on single exams or in batch mode, making it suitable for both clinical research and for potential applications in screening and monitoring programs.

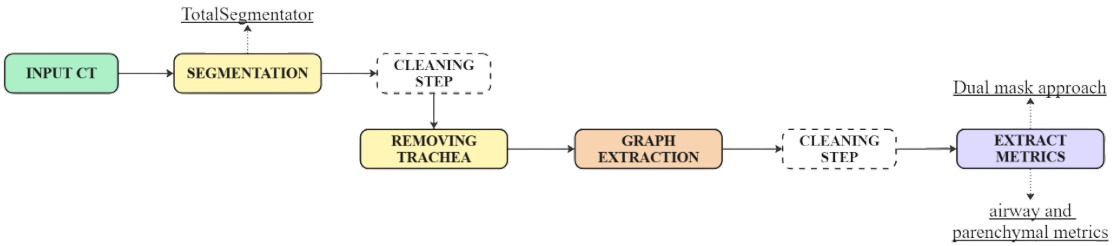


Figure 4.1: Block diagram of the pipeline

4.1 Airway Segmentation

The segmentation of the airways represents one of the most critical steps of the entire pipeline developed in this thesis, as the quality of this process directly affects all subsequent stages of the analysis, from skeletonization to bronchial graph construction, up to the extraction of metrics. For this reason, the segmentation step required an extensive phase of analysis, refinement, and validation.

The initial segmentation was performed using TotalSegmentator[4], an automatic deep learning based framework that provides a wide set of predefined anatomical labels for whole body CT images.

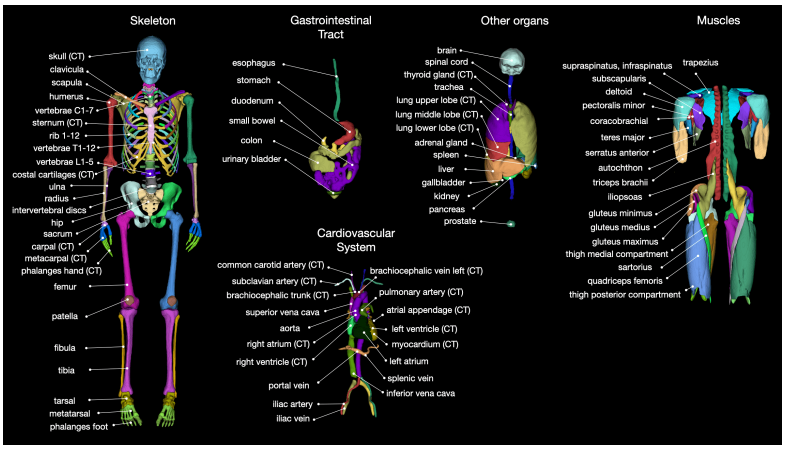


Figure 4.2: Main classes segmented by TotalSegmentator for CT and MR

As discussed in the chapter dedicated to the dataset, the CT images from the OSIC dataset are not immediately compatible with TotalSegmentator infact a preliminary preprocessing step was required to adapt the input data, which has been described in Chapter 3.1.

TotalSegmentator can be configured to perform specific segmentation tasks, each yielding a dedicated set of labels for particular anatomical regions, an overview of the full set of structures it can segment is provided in Figure 4.2.

In the context of this thesis, the *lung_vessels* task of TotalSegmentator was used even if the task name refers to pulmonary vessels, it also includes the **lung_trachea_bronchia label**, which provides the complete airway tree, including the trachea and bronchial branches, as a single connected structure. The output of this step (Figure 4.3) is therefore a binary mask containing the trachea and the complete bronchial tree.

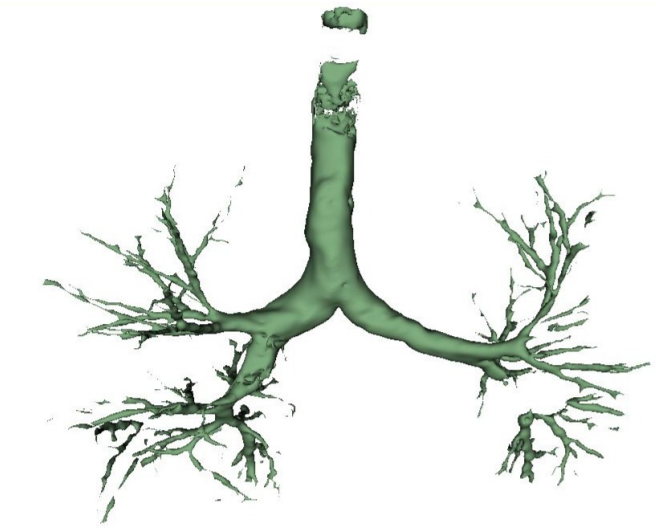


Figure 4.3: Original segmentation, the direct result of TotalSegmentator

A limitation of automatic segmentations is the presence of local disconnections or artificial interruptions of bronchial branches. To address this issue, an *intelligent gap-filling module* was developed that improves connectivity without introducing anatomically implausible structures, allowing the recovery of thin and interrupted branches.
The algorithm proceeds through a sequence of progressive steps; a connected component analysis identifies disconnected fragments. Small defects with volume < 100 mm^3 are filled only if the majority of voxels within them have Hounsfield Unit (HU) values consistent with air (HU $<$ -400), thus ensuring the anatomical plausibility of the intervention. Subsequently, isolated components within 10 mm of the main bronchial tree are reconnected using validated three-dimensional bridges, verifying that intermediate voxels have HU values consistent with air, preventing spurious connections through non-aerated tissue. Finally, a light morphological refinement smoothes the airway surface while preserving its tubular geometry.
After this initial segmentation and continuity reconstruction phase, an advanced refinement module is applied, designed to remove small, isolated artifacts while preserving fine tubular structures. Unlike traditional approaches that use fixed and very restrictive HU thresholds, the pipeline implements an adaptive strategy. This employs more permissive criteria in the gap-filling phases, where the goal is to recover anatomical continuity even in the presence of greater signal variability, and more selective criteria in the cleaning phases (HU $<$ -850), where the need to exclude soft tissue prevails.
The precise thresholds are automatically determined for each exam by analyzing the local histogram of HU values. In practice, the histogram inside the initial airway

mask is clustered to derive patient-specific class boundaries, and the resulting thresholds are then applied with a distance-based rule during region growing and gap filling.
This dynamic adaptation, applied during the critical phases of region growing and topological reconstruction, allows the pipeline to intelligently balance anatomical completeness with accuracy, adapting to the specifics of each individual scan. This flexibility is particularly advantageous in real-world clinical settings and multicenter studies, where the heterogeneity of acquisition protocols renders approaches based on rigid parameters ineffective.

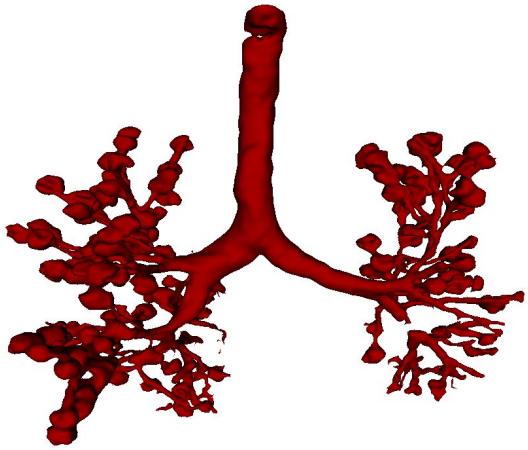


Figure 4.4: Refined segmentation after gap filling and artifact removal

4.1.1 Carina Detection and Trachea Removal

Although the refined segmentation is generally correct from an anatomical point of view, a detailed qualitative analysis revealed several critical issues. In particular, the upper portion of the trachea is often characterized by a highly irregular surface, which introduces significant noise in the subsequent steps of the pipeline, especially during skeletonization and graph construction. To mitigate this, a straight cut is applied slightly above the estimated carina position.
The **carina** is the main bifurcation point of the trachea into the right and left main bronchi and represents a fundamental anatomical landmark for the entire bronchial analysis, an incorrect identification of the carina would compromise both the skeleton construction and the assignment of bronchial generations.
For this reason, the pipeline implements a robust carina identification method that combines several complementary criteria. The analysis is conducted by examining the layer-by-layer segmentation to identify the axial level at which the single tracheal structure divides into two or more components of comparable size. Simultaneously,

a preliminary skeleton is computed to identify points with maximum local diameter, given that the carina typically represents the largest cross-section immediately before the first bifurcation. A further topological check builds a provisional graph and searches for nodes with high connectivity and high centrality, indicators of critical branching points. Finally, a layer-by-layer morphological analysis evaluates the compactness and number of objects present, detecting the transition from a single compact structure to multiple, separate components. This combination of information from geometry, topology, and morphology ensures accurate carina identification even in the presence of artifacts or anatomical variability.
The different candidates are then combined through a voting system. Each detection method produces a candidate position (z, y, x) and assigns it a method-specific weight:

- Connected components: weight = 3.0
- Diameter analysis: weight = 2.0
- Topological assessment: weight = 2.5
- Slice analysis: weight = 1.5

For each candidate, a consensus score is computed based on its proximity to other candidates, with closer candidates receiving higher scores. The final score for a candidate is the sum of its own weight and the one with the highest total score is selected as the carina.
The carina is identified *before* removing the trachea, to avoid the risk of cutting too deeply and losing the carina itself, which is required in subsequent steps.
Tracheal removal is performed extremely conservatively to preserve the integrity of critical anatomical structures. A preliminary upper cut is performed at a predetermined axial level to eliminate the highest and most irregular portion of the trachea, a potential source of noise. Finally, only a limited portion of this identified structure is removed, always maintaining a safety margin of at least 15 mm above the previously identified position of the carina.

This approach removes the noisiest tracheal regions while preserving the anatomical integrity of the bronchial tree and the main bifurcation point (Figure 4.5).

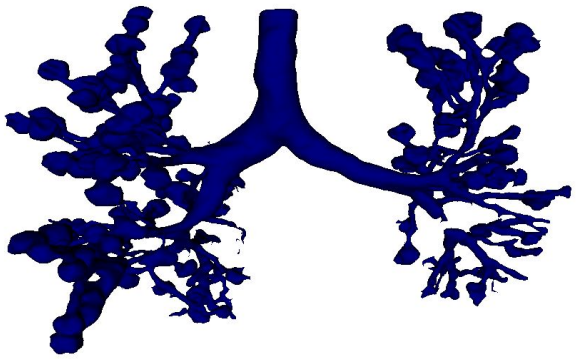


Figure 4.5: Refined segmentation after trachea removal

4.1.2 Preprocessing and Component Management

After tracheal removal, segmentation may still present disconnected components due to incomplete gap filling in severely diseased regions, anatomical variants in bronchial branching, or segmentation artifacts in the peripheral airways. A comprehensive preprocessing module addresses these issues through intelligent component analysis and possible reconnection.
Properties such as voxel number and volume, 3D centroid position, bounding box extension, and spatial location relative to anatomical landmarks are calculated for each component. Components are then classified by size, with the largest component assumed to represent the main bronchial tree. Smaller components are candidates for reconnection, if anatomically plausible, or removal, if likely artifacts.
Traditional reconnection algorithms calculate Euclidean (straight-line) distances between disconnected regions. However, for tubular anatomical structures such as airways, this approach is suboptimal because it ignores the actual anatomical connectivity path. The pipeline instead employs a path-based distance metric, which calculates distances along the airway skeleton graph rather than across 3D space. The algorithm follows a complex procedure: first, a temporary skeleton is computed from the core component to establish the connectivity network; this skeleton is then converted into a weighted graph in which nodes represent junctions and terminals, edges represent airway segments, and edge weights correspond to the physical path length in millimeters. Next, for each disconnected component, the minimum distance to the core tree is computed using Dijkstra's algorithm along the

physiological airway pathways, rather than along straight lines through the tissue. Components within a threshold distance (50.0 mm by default) are reconnected to the main tree using cylindrical connectors, the validity of which is confirmed by an analysis of the Hounsfield unit values of the affected voxels.

The final output is a complete bronchial graph, representing the airway tree starting from a point just above the carina, complete with a comprehensive set of morphometric properties for each branch.

After the construction of the bronchial graph, a branch-level analysis is performed in order to extract quantitative descriptors of the airway tree. This analysis is carried out on a per-branch basis and includes the computation of key metrics such as branch length, branch diameter, bifurcation properties, and airway generation assignment.

The calculation of quantitative metrics is based on a **double-mask strategy**, designed to ensure both topological accuracy and morphometric precision. The logic involves using both masks obtained in the initial phases in parallel: the original mask, produced by TotalSegmentator and subjected to minimal processing, and the refined mask, resulting from gap-filling and artifact removal operations.

The refined mask is optimized for topology and connectivity and is therefore used to calculate the skeleton, construct the graph, and assign generations according to the Weibel model. Its main advantage lies in improved connectivity, with fewer gaps and a more complete bronchial tree structure. Its limitation is that the gap-filling operations can alter the actual anatomical dimensions, for example by creating non-physiological terminal swellings.

The original mask, on the other hand, is dedicated to the calculation of quantitative metrics, such as diameters, volumes, and surface areas, as this mask preserves the true anatomical dimensions, without the distortions introduced by morphological corrections.

This separation ensures that the graph construction, which requires continuity, is not compromised by noise or fragmentation present in the original scan. At the same time, it prevents the morphological alterations introduced by gap correction (such as the terminal blob visible in Figure 4.4) from affecting the calculation of critical metrics such as diameter.

Basically, the skeleton is calculated from the *refined mask*, identifying the branch paths and their connectivity. For each point in the skeleton, the same point from the original mask is taken to obtain the value of the required metric. If a direct point-to-point correspondence is not possible, the value is extrapolated by calculating a weighted average from the closest available points in the original mask.

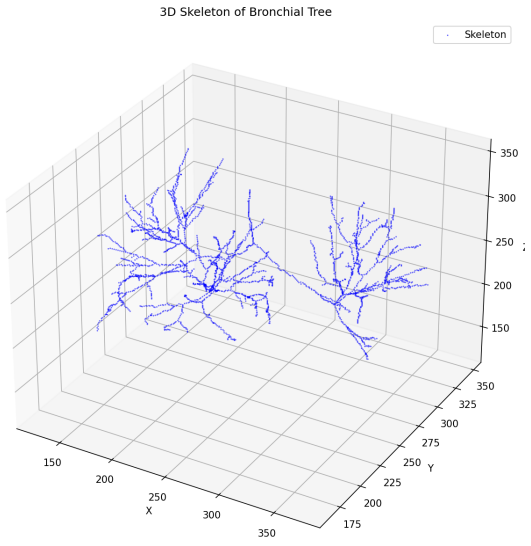


Figure 4.6: 3D visualization of resultant skeleton

4.1.3 Branch Length

The length of each bronchial branch is computed directly from the three-dimensional skeleton representation of the airways. For each branch identified by the skeleton analysis, the length is computed as *the cumulative sum of Euclidean distances between consecutive skeleton voxels along the branch path.*
Importantly, the branch length follows the actual skeleton path rather than the straight-line distance between the branch endpoints, this allows an accurate estimation of the true anatomical length of each airway segment, especially in the presence of curved or tortuous branches.
The computation explicitly accounts for the physical voxel spacing of the CT image, ensuring that branch lengths are expressed in millimeters and are comparable across subjects:

$$L_{\text{branch}} = \sum_{i=1}^{N-1} \sqrt{(\Delta x_i \cdot s_x)^2 + (\Delta y_i \cdot s_y)^2 + (\Delta z_i \cdot s_z)^2} \tag{4.1}$$

where N is the number of skeleton points along the branch, $(\Delta x_i, \Delta y_i, \Delta z_i)$ are the voxel coordinate differences, and (s_x, s_y, s_z) are the voxel spacings in mm.

4.1.4 Branch Diameter

It is important to underline that the airway diameter is not computed directly from the segmentation surface, but is estimated using the **distance transform**

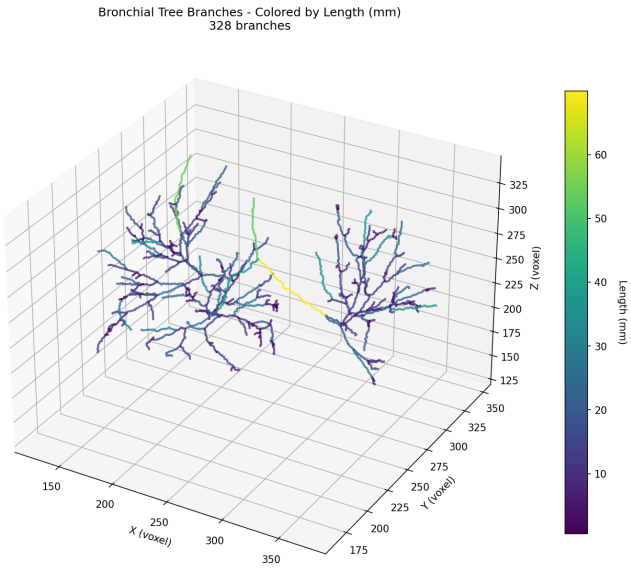


Figure 4.7: 3D visualization of airway branches with length annotation

applied to the binary airway mask. Specifically, given a binary mask in which airway voxels are set to 1 and background voxels to 0, the distance transform of the background is computed. For each voxel belonging to the airway skeleton, the distance transform represents *the minimum Euclidean distance from that skeleton voxel to the nearest background voxel.*

Standard mean diameter calculation along branches is susceptible to **terminal blob artifacts**, the small spherical dilations at airway endpoints caused by the refined mask's gap-filling operations.

Since this is a crucial aspect for the calculation of the final metrics, the pipeline implements a robust diameter estimation strategy. This consists of terminal exclusion, whereby the first and last 10% of voxels along each branch are excluded from the calculation to avoid artifacts at the extremities.

The measurement is then based on a percentile approach: instead of using the arithmetic mean, the 75th percentile of the distance transform values is used, make the measurement more resistant to outliers, while still capturing the typical diameter. However, as with all metrics, the measurement approach used is the double-mask measurement previously described well at chapter 4.1.2, where the skeleton is calculated from the refined mask, but the metrics values are taken from the original mask, thus avoiding distortions caused by gap-filling operations. Formal definitions of diameter metrics are provided in Appendix A, Table A.3.

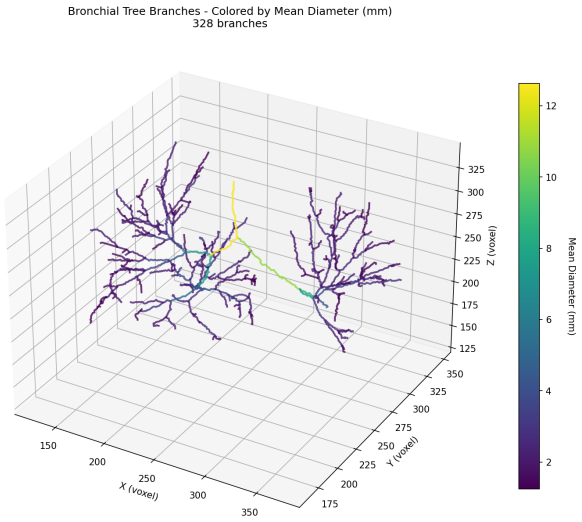


Figure 4.8: 3D visualization of airway branches with diameter annotation (using robust 75th percentile method)

4.1.5 Bifurcations and Generations of the Airway: The Weibel Model

Bronchial bifurcations are identified as the graph nodes with a degree greater than two representing the transition point from a parent bronchus to two or more child branches and plays a fundamental role in defining the tree's topology. Identifying bifurcations allows the tree to be separated into distinct branches and establishes the hierarchical organization necessary for assigning bronchial generations.
Generation assignment follows a hierarchical approach inspired by the classic **Weibel model** of the human bronchial tree [12]. In this model, the airway system is represented as a branched structure in which each bifurcation marks the transition from a parent bronchus to its child bronchi, and each level of branching corresponds to a subsequent generation.
According to the Weibel model, the tracheal carina represents the starting point and is defined as generation zero. Each subsequent bifurcation increases the generation index by one, proceeding distally from the central airways toward the lung periphery. This hierarchical framework provides a standardized reference for describing bronchial morphology and allows for consistent comparisons across subjects.
In addition to providing a topological nomenclature, the Weibel model offers a geometric-quantitative framework, characterizing each generation through key morphometric parameters. Key parameters proposed include the number of airways, mean bronchial diameter, and mean length of typical branches for each generation,

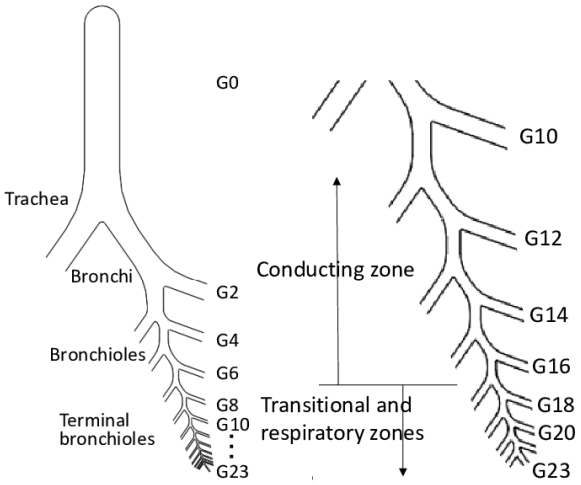


Figure 4.9: Schematic representation of the Weibel model of the bronchial tree, illustrating the hierarchical organization of airway generations starting from the carina (generation 0) and progressing distally through successive bifurcations, adapted from [27]

which follow predefined exponential relationships, providing a standard anatomical benchmark.

A key morphometric relationship described by the Weibel model is **bronchial tapering**: the progressive reduction in diameter observed from a parent branch to its offspring. Under normal anatomical conditions, this narrowing follows a consistent ratio approximated by Weibel's law, with a theoretical value of approximately 0.793, meaning that the diameter of a daughter branch is approximately 79.3% of that of the parent.

Analyzing the distribution of these ratios across the bronchial tree is particularly relevant in the context of pulmonary fibrosis because in this disease, pathogenic mechanisms create marked effects on bronchial structures, leading to progressive narrowing and architectural distortion. Translating this clinical and radiological evidence into robust and reproducible quantitative parameters, such as diameters and generation-specific thinning ratios, allows us to establish a solid morphometric basis for characterizing airway pathology. Detailed definitions of these metrics and their reference ranges are provided in Appendix A, Tables A.3 and A.2.

Figure 4.10 shows the result of this analysis: airway generations are color-coded across branches, and the carina is highlighted with a red marker (with its coordinates reported).

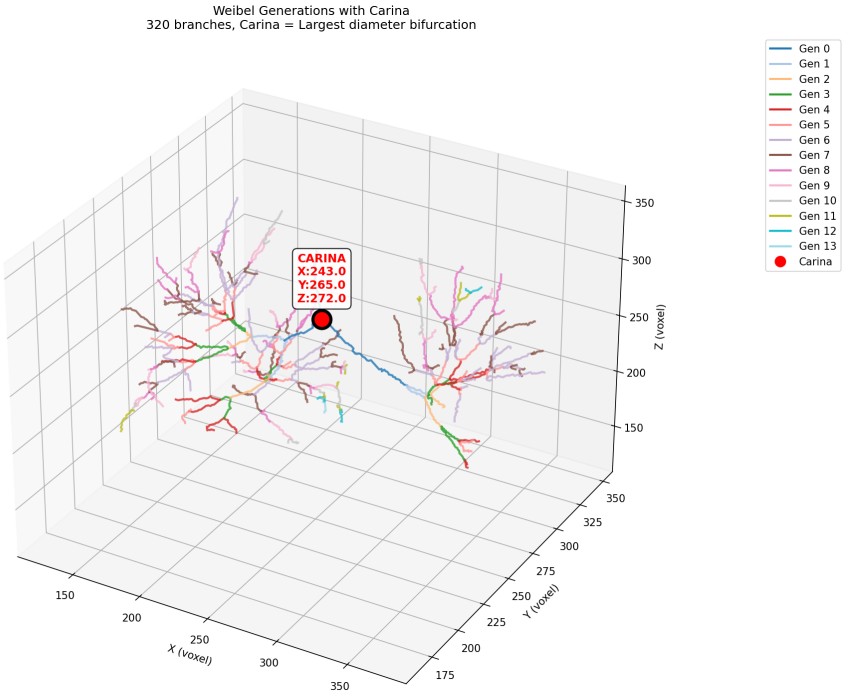


Figure 4.10: Weibel generations assigned to airway branches with carina highlighted.

Chapter 5

Validation Pipeline

The pipeline described in the previous chapter returns not only the airway analysis metrics, but also the related segmentations and graphs, also illustrated previously and after this process, the obtained results must be validated.
A crucial aspect, which consistently characterizes the work of this thesis, is the lack of a ground truth (GT) in all the steps addressed: from the segmentation itself to a reliable reference indicating whether the calculated metrics define a patient as sick or not.
Precisely for this reason, a dedicated validation pipeline was implemented (Figure 5.1), designed to evaluate each specific substep of the overall process. In the absence of a reference GT, the basis for validation was drawn from values consolidated in the scientific literature, consequently, the validation results are primarily theoretically grounded.

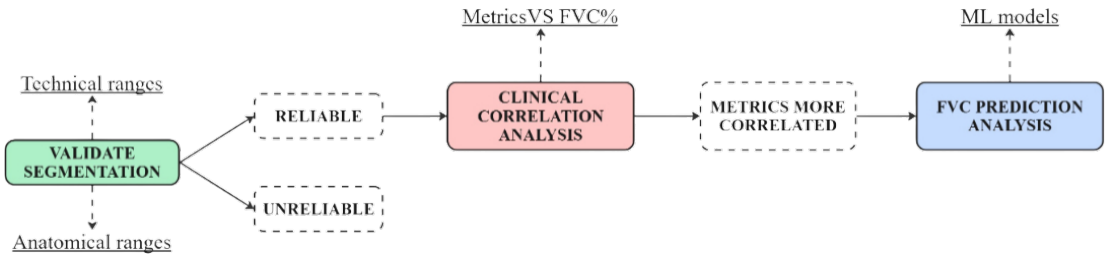


Figure 5.1: Flowchart of the developed validation pipeline

5.1 Segmentation quality assessment

This section describes the first level of validation for the airway analysis pipeline. In the absence of a clinically validated ground truth, validation focuses on the

technical reliability of the computational process through a two-step approach: first, disease-independent technical plausibility checks that detect pipeline errors regardless of patient condition, and only afterward, anatomical reference checks against literature-derived ranges. These, while based on healthy populations, serve as benchmarks for identifying gross computational errors rather than clinical anomalies.

So we have two distinct sets of criteria with different objectives (summarized in Appendix A, Tables A.1 and A.2).

The first criteria are based on thresholds that define physically plausible limits for algorithmic output and are designed to flag computational errors such as oversegmentation, the inclusion of non-airway tissue, or topological errors. These include airway volume, maximum generation count, peripheral-to-central ratio, overall branch count, tortuosity index, and taper ratio.

The second set of criteria consists of anatomical reference ranges. These are derived from normative data in healthy adult populations and include parameters such as total airway volume, surface area, branch count, mean branch length, tracheal diameter, fifth-generation diameter, bifurcation ratio, taper ratio, and maximum generation depth. A critical limitation is that pathological airways can legitimately deviate from these ranges.

The results are then classified using a three-level hierarchical scheme: **pass** if the value is within the range; **warning** if the value is outside the valid range but within two standard deviations of the reported mean and finally **fail**.

Recall that for technical controls, the classification is binary pass/fail because these thresholds define physically and computationally plausible limits, independent of the patient's disease state. For example, an airway volume greater than 600 ml always indicates algorithmic failure because it's the inclusion of non-respiratory tissue, regardless of whether the patient is healthy or fibrotic.

In contrast, anatomical controls use a three-tiered system because these reference ranges are derived from healthy populations, but the dataset includes diseased lungs. A patient with advanced pulmonary fibrosis may have a PC ratio of 0.15 not due to segmentation failure, but as a true pathological feature reflecting peripheral airway destruction. The *warning* level takes this physiological variability into account, distinguishing probable disease-related deviations from clear calculation errors.

Validation follows a modular pipeline that examines specific characteristics such as file presence, metadata completeness, and image dimensionality. Only then are the technical and anatomical checks discussed above performed.

For each case, a detailed JSON report is generated containing the patient ID, the results of the individual technical checks, the measured values with their reference ranges, the final classification, and explanatory notes. The results of all cases are finally aggregated into a summary CSV file, indicating that the final status for

each patient is either reliable or not reliable. A case is marked as reliable if it passes all technical checks, indicating that the calculation process is correct and the results are ready for clinical interpretation. It is also marked as unreliable if the case has one or more violations of the technical limits. Note that cases that pass the technical checks but fail the anatomical reference checks are still classified as technically valid.
At the dataset level, the aggregated CSV enables descriptive statistics, analysis of the distribution of quality levels (reliable/unreliable), and identification of potential bias patterns in the processing pipeline.

5.2 Clinical Correlation Analysis

After the technical validation of the segmentation pipeline (Section 5.1), this second level of validation assesses the clinical relevance of the extracted airway parameters by examining their correlation with forced vital capacity (FVC), a gold-standard functional outcome measure in pulmonary fibrosis. Unlike the technical validation, which assessed computational reliability, this analysis investigates whether the quantitative parameters derived from CT imaging, and therefore all extracted metrics, reflect disease severity and progression as measured by spirometry.
The analysis leverages the OSIC Pulmonary Fibrosis Progression dataset, which, as mentioned in Chapter 3, provides longitudinal FVC measurements coupled with baseline chest CT scans for patients with idiopathic pulmonary fibrosis (IPF). This creates a unique opportunity to perform cross-modal validation: correlating the airway structural features obtained from the pipeline with functional respiratory capacity to establish that the computational pipeline captures biologically meaningful information rather than simply producing technically valid but clinically irrelevant output.
The validation workflow integrates three different data sources: the first is the CSV file resulting from the previous validation, where cases are marked as reliable or not reliable, ensuring that only computationally valid segmentations are included. We then use the quantitative metrics extracted from the airway segmentation, including volumetric, topological, and morphological parameters, along with parenchymal metrics, both extracted from the main pipeline. Finally, we use the FVC% value corresponding to each CT scan in the OSIC dataset, aligned by patient ID and time point.

Methodology

Rather than analyzing the entire set of calculated metrics, the validation focuses on a targeted subset of features hypothesized to reflect disease mechanisms in IPF. The extracted metrics are organized into three main categories.

Key airway metrics include

- **airway volume** (ml), which represents the total volume of the segmented bronchial tree
- **mean tortuosity**, obtained as the average of the path tortuosity across all airway segments, an indicator of their distortion.

The peripheral airway metrics category includes:

- **peripheral diameter standard deviation**(mm), which measures the diameter variability in peripheral airways (beyond the fifth generation) and reflects the heterogeneity of peripheral remodeling;
- **central-to-peripheral diameter ratio** calculated as the ratio of the mean diameter of the central airways (generations ≤ 5) to that of the peripheral airways (generations > 5), which captures the gradient of airway narrowing;
- the **mean peripheral branch volume**(mm^3), which corresponds to the average volume of individual peripheral airway segments and is a parameter sensitive to airway loss in this region;
- **peripheral branch density**, defined as the number of distal branches, which reflects the integrity of the terminal conducting airway network.

Finally, lung parenchymal metrics are analyzed:

- **mean lung density** (HU), the average value in hounsfield units within the lung parenchyma, which reflects fibrotic burden;
- **histogram entropy**, calculated as *the shannon entropy* of the histogram of HU values, a parameter that quantifies textural heterogeneity associated with reticulation and honeycombing.

This selective approach avoids the multiple testing burden of an exhaustive metric screening, focusing instead on features with biological plausibility in the context of IPF pathophysiology.
The primary analysis assesses the cross-sectional relationship between baseline airway metrics and FVC% at all available time points. Although airway metrics are derived from baseline CT (week 0), FVC measurements span the entire follow-up period, resulting in multiple observations per patient. This design allows us to assess whether baseline airway structure predicts not only FVC% but also longitudinal functional trajectory.
Various metrics are used to assess the quality of prediction, including the **pearson correlation coefficient** (r), which measures the linear association between each

metric and FVC%. Another nonparametric metric is **the spearman correlation coefficient** (*rho*), which is used because it is robust to outliers and monotonic but nonlinear relationships. Finally, we perform a **significance test**, or *p-value analysis* for the pearson and spearman coefficients, defined as $p < 0.05$ to indicate a good correlation.

For each metric, scatterplots show the relationship between the metric (x-axis) and FVC% (y-axis), with individual data points colored by time (week) to visualize temporal heterogeneity.

To assess whether baseline airway metrics predict disease progression, patients are stratified into quartiles based on their baseline metric values (Q1 = lowest quartile, Q4 = highest quartile), and for each quartile, the mean rate of decline in FVC% is calculated via a patient-specific linear regression of FVC% versus time defined in 'weeks'. This identifies whether, for example, high baseline tortuosity or low peripheral airway volume at baseline predicts accelerated functional decline over follow-up.

It is important to remember that a subset of cases in the OSIC dataset was modified with a gaussian smoothing kernel applied during reconstruction, which may potentially impact parenchymal texture metrics.

To assess robustness, the correlation analyses are repeated first for all scans (original + smoothed), then for only the original scans, and finally for only the smoothed scans. The difference in correlation coefficients (Δr) quantifies the impact of including smoothed scans, for which a threshold of $\Delta r < 0.1$ is considered negligible, while $\Delta r > 0.2$ suggests significant confounding that warrants separate reporting.

The Δr value is computed as the absolute difference between the correlation coefficient obtained on the full set (original + smoothed) and the coefficient obtained on the original-only subset for each metric. This sensitivity analysis is used as a decision rule: when $\Delta r < 0.1$, smoothed scans are considered non-influential and are retained in the pooled analysis; when $\Delta r > 0.2$, the smoothed scans are treated as a confounding source and results are reported separately for original-only and smoothed-only subsets. Intermediate values are interpreted as potentially influential, prompting cautious interpretation but not an automatic exclusion. This logic informs which dataset variant is used in downstream analyses and which stratified results are emphasized in the final reporting.

The analysis specifically generates **correlation results** in a summary table of pearson and spearman correlations for each metric, with sample sizes and p-values, along with individual plots for each metric against FVC%, annotated with correlation statistics, and a bar chart comparing correlation levels across all metrics. Finally, plots of FVC evolution are also provided, as time curves of mean FVC% by quartile of the metric, illustrating differential rates of decline.

To interpret the significance of the observed correlation between radiomics metrics and FVC percentage, we adopted established criteria. A strong correlation is

considered when the correlation coefficient $|r| \geq 0.5$, indicating that the metric captures a substantial portion of functional variance. A moderate correlation, defined as values $0.3 \leq |r| < 0.5$, indicates a statistically detectable association, although other factors contribute more to FVC variability. Finally, a weak correlation, with $|r| < 0.3$, suggests that the metric has limited independent prognostic utility; however, it may contribute in multivariate models together with other predictors.
The direction of the correlation provides crucial clinical information; a positive correlation implies that higher values of the metric are associated with a higher FVC percentage and therefore with better lung function, as is the case with larger airway volume. Conversely, a negative correlation indicates that higher metric values correspond to a reduced FVC%, indicating functional deterioration; this is observed, for example, when increased lung density reflects extensive fibrosis.
It is important to emphasize that correlations do not establish causality. Observed associations may reflect, for example, direct mechanistic links, such as peripheral airway destruction causing ventilatory compromise, or a shared underlying pathology, such as fibrosis affecting both the airways and the parenchyma, or even confounding due to disease severity, whereby more severe disease simultaneously affects all metrics.
Nonetheless, significant correlations provide evidence that computational metrics encode clinically meaningful information, validating the biological relevance of the pipeline beyond technical accuracy.
Several limitations of this pipeline must be acknowledged. First, airway metrics are derived only from baseline CT, so *longitudinal imaging* that is, having scans from different time points of the same patient, perhaps from time points with available FVC% values would allow us to monitor structural changes parallel to functional decline, but this is not available in the OSIC dataset. This raises the problem of having FVC measurements that span multiple time points, while CT is performed at baseline, so correlations assume that baseline airway structure predicts functional trajectory. This is plausible given the irreversible nature of fibrotic remodeling, but remains an unverifiable hypothesis.
Another problem is that IPF is a heterogeneous disease with variable progression rates, so subgroup analyses by progression phenotype may reveal stronger correlations within homogeneous strata, but require larger sample sizes. Although metric selection is driven by hypotheses, multiple correlation analysis increases the risk of false positives so results with borderline significance ($p = 0.05$) should be interpreted with caution and validated in independent cohorts.
This clinical correlation analysis is conceptually and temporally distinct from the technical validation described in Section 5.1. Therefore, technical validation ensures that the pipeline produces computationally valid outputs, while clinical validation (the one just described) assesses whether the technically reliable outputs correlate with clinical outcomes (FVC), establishing biological relevance.

The two levels are obviously complementary: Level 1 filters out unusable data, while Level 2 confirms that usable data are clinically informative. A metric may be technically valid but show no correlation with FVC, indicating limited prognostic utility. Conversely, a metric showing a strong correlation with FVC in technically unreliable cases would be misleading, as the association could be driven by segmentation artifacts rather than true biology.

5.3 FVC Prediction Analysis

While the previous section established correlations between baseline airway metrics and longitudinal FVC% measurements, this third level of validation investigates the predictive capability of individual airway and parenchymal features. Specifically, the analysis aims to determine whether single morphological metrics measured at baseline can predict: FVC% at week 0, FVC% at week 52 (one-year follow-up), and the functional decline over one year.
This predictive validation addresses a fundamental clinical question: can quantitative features derived from baseline CT scans predict disease outcome?
Such predictive models, if validated, could support treatment throughout the disease's course.
The analysis is constrained by the fundamental challenge of temporal misalignment previously addressed, whereby CT scans are acquired at baseline (nominally week 0), while FVC measurements are performed at variable time points that can be within a year or even after or before baseline. This requires a robust interpolation framework to estimate FVC% values at standardized time points (weeks 0 and 52) from sparse and unevenly sampled spirometry data.
To enable prediction of FVC% at specific time points, the system must infer FVC% values at week 0 (baseline) and week 52 (one-year follow-up) from measurements that may occur before, after, or between these targets.
The main challenge is that FVC measurements are not uniformly distributed. Some patients have dense sampling (more than 10 measurements), while others have sparse data (2-3 measurements). The timing of measurements is heterogeneous: baseline assessments can be performed between weeks 5 and 10, and one-year follow-ups between weeks 40 and 65. This temporal variability requires a principled interpolation approach that balances data utilization with methodological rigor.

5.3.1 FVC Interpolation Strategy

The interpolation windows adopted in this work are summarized in Appendix A, Table A.5.
The interpolation methodology adopts a hierarchical approach that prioritizes measurement quality based on temporal proximity and data availability.

For week 0 (baseline) estimation, the strategy defines a preferred window between -5 and 10 weeks relative to the CT scan acquisition time, where the nearest measurement is selected and quality is assigned based on temporal distance (high quality if within 3 weeks, medium if within 8 weeks). When no measurements fall within this preferred window but data exist between 15 and 30 weeks, linear regression interpolation is employed using at least two available measurements, with quality marked as low. Beyond 30 weeks, the temporal gap is considered too large for reliable estimation, and no value is assigned.

For week 52 (one-year follow-up) estimation, a preferred window between 40 and 65 weeks is established, again using the nearest measurement with quality assignment based on distance from the target week (high if within 4 weeks, medium if within 8 weeks, low if within 13 weeks). In cases where measurements exist outside this window but regression-based estimation is feasible, linear interpolation or extrapolation is performed with appropriate quality downgrading. The interpolation quality is systematically tracked and documented for each estimated FVC% value, enabling subsequent stratification of the dataset by interpolation reliability.

A critical methodological consideration in this analysis concerns how to quantify the functional decline in FVC%.

Two distinct computational approaches were evaluated, each with different underlying assumptions and implications for prediction performance.

1. Traditional Drop Metric, difference between interpolated time points

The first approach, termed the **traditional drop metric**, calculates FVC decline as the arithmetic difference between interpolated values at two standardized time points:

$$\text{FVC drop}_{\text{traditional}} = \text{FVC\%}_{\text{week0}} - \text{FVC\%}_{\text{week52}} \tag{5.1}$$

This method directly depends on the quality and accuracy of the FVC% estimates at both week 0 and week 52, obtained using the interpolation framework described in the 5.3.1 section. The resulting decline value inherits uncertainty from both interpolation procedures, and its reliability therefore depends on the presence of high- or medium-quality estimates at both time points.

The main advantage of this approach is its **conceptual simplicity** and **straightforward interpretability**: it provides an estimate of a patient's lung function decline in the first year following baseline CT acquisition.

However, this method has several limitations, the most notable of which is that interpolation errors at both time points propagate into the decline calculation, potentially amplifying measurement uncertainty.

2. Direct Decline Metric

This approach differs fundamentally in that it uses the entire longitudinal trajectory of FVC% for each patient to calculate decline. Rather than relying on interpolated values at fixed time intervals, this method fits a patient-specific linear regression model to all available FVC% measurements during the follow-up period:

$$\text{FVC\%}(t) = \beta_0 + \beta_1 \cdot t + \epsilon \tag{5.2}$$

where t represents time in weeks, β_0 is the intercept, β_1 is the slope (weekly rate of decline), and ϵ is the residual error.
The direct decline metric is therefore defined as:

$$\text{Annual decline}_{\text{direct}} = \beta_1 \times 52 \tag{5.3}$$

This value represents the estimated change in FVC% per year, derived from the slope of the regression line. It is important to note that this metric is calculated only for patients with at least three FVC measurements spanning a minimum time interval, ensuring that the regression fit relies on sufficient data to capture the underlying trend rather than noise. Incorporating all available measurements maximizes data utilization, reduces dependence on a single time point, and provides a more reliable estimate of the true rate of decline by averaging measurement noise across multiple observations.
The direct decline metric is **independent** of the quality of the interpolation at weeks 0 and 52, as it derives the functional trajectory directly from observed data rather than from inferred values. This makes it particularly valuable for patients with dense longitudinal sampling, where regression-based estimation is likely more accurate than any interpolation-based calculation.
Limitations of this approach include the requirement for sufficient measurements, which excludes patients with sparse follow-up data. Furthermore, the linear assumption may not capture more complex and nonlinear progression patterns that can occur in IPF. The metric also assumes that decline is approximately constant over the observation period, which may not be valid for patients experiencing acute exacerbations or therapeutic interventions.

Comparative Analysis of the Two Metrics

The two metrics are fundamentally complementary rather than mutually exclusive. The traditional decline metric provides standardized, time-anchored estimates appropriate for patients with measurements close to target weeks, while the direct decline metric offers a longitudinally integrated assessment for patients with a richer temporal sample.

Fundamentally, the two metrics answer slightly different clinical questions. The traditional decline metric quantifies "How much has the FVC% changed between baseline and the one-year follow-up?", while the direct decline metric answers the question "What is the patient's average annual rate of decline in FVC% based on all available data?" For patients with stable linear progression, these should produce similar values, but may diverge substantially in cases of nonlinear decline or when the quality of the interpolation differs significantly between weeks 0 and 52.
In this analysis, **both metrics are calculated and evaluated independently** to assess whether morphological features derived from baseline CT predict functional outcomes differently depending on how decline is quantified.
For each patient, the prediction models were trained separately on four different objectives, corresponding to four different ways of quantifying pulmonary fibrosis progression.
The first objective involves estimating the FVC% value at week 0. Then, the FVC% value at week 52 is predicted. The third model focuses on predicting the absolute decline in FVC, defined according to the traditional approach as the difference between the value at week 0 and that at week 52. Finally, the fourth model aims to directly estimate the annual decline in FVC, expressed as the slope of the linear regression line calculated on all available values over time.
This dual-metric approach allows us to assess whether the predictive power of airway and parenchymal metrics is sensitive to the definition of functional decline. If a feature demonstrates strong predictive performance for the direct decline metric but weak performance for the traditional decline metric, this suggests that the feature captures the long-term disease trajectory better than short-term variation.

5.3.2 Prediction Methodology: Leave-One-Out Cross-Validation

The prediction analysis employs **leave-one-out cross-validation (LOOCV)** to assess the forecasting accuracy of individual morphological features. This method is particularly appropriate for the moderate sample size of the OSIC dataset and provides an unbiased estimate of generalization performance by maximizing the use of available data for both training and testing.
For each feature-target combination (e.g., mean peripheral branch volume predicting week 52 FVC%), the LOOCV procedure iteratively trains a univariate linear regression model on all patients except one, then predicts the held-out patient's target value. This process repeats until each patient has been predicted exactly once, yielding a vector of predicted values that can be compared against actual (interpolated or regression-derived) values.
The choice of univariate models (single-feature predictions) rather than multivariate

models is deliberate and serves multiple purposes. First, it isolates the independent predictive power of each feature, establishing which individual metrics encode the most prognostic information. Second, it avoids overfitting risks associated with high-dimensional models in small datasets. Third, it enables direct comparison of feature importance, informing subsequent development of multivariate models by identifying which features contribute most substantially to predictions.

5.3.3 Evaluation Metrics

Prediction performance is quantified through several complementary metrics that capture different aspects of model accuracy:

Coefficient of Determination (R^2): Represents the proportion of variance in the target variable explained by the feature. Values range from 0 (no predictive power) to 1 (perfect prediction), with negative values possible if the model performs worse than a horizontal line. Strong features typically achieve $R^2 > 0.3$.

Mean Absolute Error (MAE): The average absolute difference between predicted and actual values, expressed in the same units as the target (FVC% points or %/year for decline). MAE provides an interpretable measure of typical prediction error magnitude.

Root Mean Squared Error (RMSE): Similar to MAE but with quadratic weighting that penalizes large errors more heavily. Useful for identifying features that produce occasional large prediction errors even if average performance is acceptable.

Pearson Correlation (r): Measures the linear relationship between predicted and actual values. Unlike R^2, Pearson r is sensitive to both correlation strength and calibration. Strong predictors show $r > 0.5$ with $p < 0.05$.

5.3.4 Visualization Outputs

For each feature analyzed, the prediction results are visualized through a comprehensive set of plots that enable assessment of both overall performance and detailed diagnostic evaluation. Given the dual-metric approach described in Section 5.3.1, visualization outputs are generated for all four prediction targets: FVC% at week 0, FVC% at week 52, traditional drop (week0 - week52), and direct annual decline.

Correlation Plots

The scatterplots show the predicted versus actual FVC% for each of the four forecasting tasks.
The key elements of the figure include, first, the perfect forecast line (or identity line), represented by the diagonal of the equation y=x. This line is the reference for an ideal calibration of the model, any systematic deviation of the points from it indicates the presence of a bias in the forecasts, which can translate into a tendency to overestimate (if the points fall above the diagonal) or underestimate (if they fall below) the actual values.
The regression line is then shown, fitted to the pairs of predicted and actual values, the interpretation of its slope is informative: a slope less than 1 suggests a compression of the forecast range, with a typical underestimation of the highest actual values and an overestimation of the lowest ones. Conversely, a slope greater than 1 indicates an expansion of the forecasts relative to reality.
Finally, the graph is accompanied by performance annotations, which summarize the main model evaluation metrics. These typically include the coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), and Pearson correlation coefficient (r), along with the corresponding p-value to assess statistical significance.
Strong predictors show tight clustering around the identity line with minimal dispersion. Systematic deviations from the identity line suggest a model calibration error that could be corrected through recalibration techniques in future work.
Importantly, when comparing correlation plots across the four targets, differences in scatter patterns and R^2 values reveal whether a given morphological feature predicts static FVC% values (at weeks 0 or 52) more accurately than dynamic change metrics (drop or decline). Such differential performance can provide insight into the biological mechanisms by which airway or parenchymal structure relates to functional outcomes.

Bland-Altman Plots

To assess the agreement between predicted and actual values, Bland-Altman plots are generated for each prediction task.
These plots show, on the x-axis, the mean between the predicted and actual FVC values for each patient, calculated as $(\hat{y}_i + y_i)/2$. The y-axis, on the other hand, represents the difference between the predicted and the actual value, i.e., $\hat{y}_i - y_i$.
Two key reference elements are plotted on the graph. The first is a horizontal line indicating the average difference (bias) of the entire set of forecasts. The position of this line reveals any systematic tendency in the model: a bias other than zero indicates a consistent over- or under-prediction.

The second element is the dashed lines of the limits of agreement, placed at a distance of ±1.96 times the standard deviation of the errors from the bias line. These intervals define the area within which approximately 95% of the differences between predicted and actual values are expected to fall, assuming that these errors follow a normal distribution.
Bland-Altman analysis identifies four main phenomena in forecast data. It detects the presence of a constant bias, represented by a fixed offset that systematically shifts all forecasts relative to the actual values. It can also highlight a proportional bias, in which the magnitude of forecast errors increases or decreases in a manner correlated with the level of the measure of interest (in this case, FVC%.)
Furthermore, the analysis can diagnose heteroskedasticity, i.e., the circumstance in which the variance (or dispersion) of errors is not constant but changes depending on the value of FVC%.
Finally, the plot allows for the identification of outliers, i.e., individual cases with exceptionally large forecast errors that deviate significantly from the general pattern of the other data.
Features with minimal bias and narrow limits of agreement are preferred for clinical applications, as they provide consistent forecast errors across the entire FVC% range.
Comparing Bland-Altman plots between the traditional drop and direct decline targets is particularly informative. If a feature shows minimal bias when predicting direct decline but exhibits systematic over- or under-prediction for traditional drop, this suggests that interpolation errors at weeks 0 or 52 may be confounding the traditional metric. Conversely, if both metrics show similar bias patterns, this strengthens confidence that the observed relationships reflect true biological associations rather than methodological artifacts.

5.3.5 Dataset Generation and Model Selection Workflow

The validation analyses described in the preceding sections have revealed significant limitations in the available data, including variability in interpolation quality, measurement sparsity, and uncertainty in target variable definitions. Given these constraints and the absence of a single, unambiguously optimal data configuration, a systematic dataset generation strategy was implemented to explore different methodological approaches for subsequent machine learning prediction tasks.
A unified python pipeline was developed to generate a comprehensive family of datasets that simultaneously addresses both quality stratification and dual-target evaluation challenges. This unified approach represents a methodological advancement over separate pipeline strategies, as it ensures consistent patient populations and feature sets across different quality levels while integrating both traditional drop and direct decline metrics within the same analytical framework.

The pipeline initially generates two datasets containing exclusively the target variables, regardless of airway metrics.
The first dataset collects interpolated FVC% values at week zero and week fifty-two for all patients, calculated using a balanced interpolation logic that returns not only the estimated values, but also the quality rating, the method used, the actual weeks of measurement, the distance from the target weeks, and the number of points used, thus ensuring full transparency into the reliability of the interpolations for each subsequent filtering.
The second dataset contains annual FVC decline rates derived from linear regression on all available longitudinal measurements, limited to patients with at least three observations. For each patient, the slope, the correlation coefficient (r), the number of measurements, the time interval covered, and a multifactorial quality rating are reported. Specifically, the quality score is based on the number of available time points, the duration of follow-up, and the magnitude of (r).
From these two datasets, airway metrics, parenchymal features, and the two families of target variables are integrated into a unified dataset, applying different filtering strategies to assess the tradeoff between data reliability and sample size.
The integration step is performed only for cases marked as `RELIABLE` by the segmentation validation pipeline.
Patients are retained if at least one of the two targets (traditional decline or directed decline) is available after applying the selected quality rules, then a final completeness filter removes subjects with missing values in any of the selected CT-derived features.
In the most stringent configuration (`dataset_strict.csv`), only patients with high or medium-quality interpolations at both time points and high or medium-quality regression for directed decline are included, prioritizing measurement certainty over sample size. . The balanced configuration (`dataset_balanced.csv`), which is the recommended choice, extends inclusion for interpolations to the low-quality class while maintaining the high or medium-quality requirement for regression. This achieves a balance between acceptable precision and statistical power, recovering patients with less precise estimates that are still usable for decline from weeks 0 to 52, without compromising the reliability of long-term progression estimates.
Finally, the full configuration (`dataset_all.csv`) removes all quality filters regardless of the reliability of the interpolations or regressions. This serves as a basis for verifying whether filtering actually improves predictive performance or whether the pipeline is robust to measurement uncertainty.
Each unified dataset contains, for each patient, the complete set of features, including airway metrics, parenchymal characteristics, demographics, and interpolated FVC values at week 0 and week 52, along with their associated quality labels. The two calculated progression targets are then saved: the traditional FVC decline $\Delta\text{FVC} = \text{FVC}_0 - \text{FVC}_{52}$ expressed in percentage points and derived from the

interpolated values, and the direct annual decline measured as a percentage per year, obtained via linear regression on all available FVC measurements for that patient.
Quality metadata for both targets, such as high, medium, or low labels assigned to the interpolated values and the decline estimate, enable stratified analyses. Patients are included in a dataset if they have at least one valid target after applying the chosen quality filter; As a result, some individuals contribute both targets while others contribute only one. This design preserves the maximum sample size while allowing for a flexible, quality-conscious subset for subsequent modeling.
A critical issue is that not all patients simultaneously have valid interpolations and valid direct decline measures, which is why the merged datasets have a variable availability of patients to analyze.
The first is restricted to patients with valid interpolated FVC decline values, i.e., both interpolations of acceptable quality (`dataset_traditional_only.csv`). The second includes exclusively patients with valid direct decline measures, characterized by sufficient longitudinal follow-up and adequate regression quality, making it ideal for models oriented towards long-term progression (`dataset_decline_only.csv`). The third one collects only patients who have both valid measures at the same time (`dataset_both_targets.csv`).

5.3.6 Interpretation of Prediction Results

Predictive analysis provides information on which airway and parenchymal characteristics encode functional and prognostic information.
Based on the theory of idiopathic pulmonary fibrosis, it is hypothesized that certain characteristics exhibit strong predictive power:

Peripheral airway metrics: a lower peripheral bronchial branch volume should predict both a lower FVC% at baseline and follow-up, as well as a more rapid decline because a reduced volume reflects distal airway destruction and the presence of small airway disease. A low peripheral branch density, meaning a lower number of distal branches, also correlates with impaired respiratory function and accelerated decline, indicative of loss of terminal conducting airways. Finally, a high central-to-peripheral diameter ratio, which indicates relatively larger central airways compared to peripheral ones, should predict worse outcomes. Such a ratio is a sign of remodeling and loss of the peripheral compartment of the bronchial tree.

Parenchymal metrics: a higher mean lung density, expressed in more positive Hounsfield units (HU), is a direct indicator of fibrotic infiltration of lung tissue. This parameter is therefore expected to predict a lower FVC% value and a greater decline

over time, reflecting the replacement of healthy parenchyma with nonfunctional tissue. A higher entropy value in the density histogram reflects greater heterogeneity in lung texture; this inhomogeneity is typically associated with pathological patterns characteristic of advanced fibrosis. Consequently, high entropy could predict worse respiratory function and more rapid disease progression.

Differential Predictions for Decline Metrics: Given the methodological differences between traditional drop and direct decline metrics outlined in Section 5.3.1, interpretation of prediction results must consider which metric is being forecast:

- **Features predicting both metrics equivalently:** When a morphological feature demonstrates comparable predictive performance for both traditional drop and direct decline, this provides strong evidence of a genuine biological relationship between airway or parenchymal structure and functional trajectory. Such features are robust to the choice of decline quantification method.
- **Features preferentially predicting direct decline:** Superior performance for regression-derived decline compared to interpolated drop may indicate that the feature captures long-term disease progression patterns rather than short-term fluctuations. These features may be particularly valuable for identifying patients with consistent, linear decline trajectories. The independence from interpolation artifacts further validates these associations.
- **Features preferentially predicting traditional drop:** Stronger performance for week 0-to-52 drop might suggest sensitivity to early disease changes occurring specifically within the first year. Alternatively, this pattern could arise if the feature correlates with measurement timing patterns that affect interpolation quality, warranting cautious interpretation.

Factors Limiting Predictive Accuracy: as we know, airway metrics are static (baseline CT only), while FVC% is dynamic. The prediction assumes that baseline structure determines the future trajectory, which may not be true if disease progression involves unpredictable remodeling events.
For the traditional drop metric, the predicted targets ($\text{FVC\%}_{\text{week0}}$, $\text{FVC\%}_{\text{week52}}$) are themselves estimates, not directly measured values. Therefore, interpolation errors propagate through the prediction models, amplifying the apparent prediction error. The quality stratification framework (high/medium/low interpolation quality) provides some control over this limitation by enabling sensitivity analyses restricted to cases with high-quality interpolated values.
For the direct decline metric, the requirement of at least three measurements restricts the analysis to patients with sufficient follow-up data. This may introduce

selection bias if patients with more frequent monitoring differ systematically from those with sparse data (e.g., healthier patients may have less frequent spirometry). Additionally, the linear regression assumption may oversimplify complex, non-linear progression patterns.
IPF presents variable progression phenotypes (rapid vs. slow progressors), so a single linear model may not accommodate this heterogeneity. Future analyses incorporating non-linear models or trajectory clustering methods could address this limitation.
Finally, we must take into account the limited number of samples which increases the variance of the LOOCV and limits the detection of weak but real effects. The split of the cohort into traditional drop and direct decline subsets, while methodologically necessary, further reduces effective sample sizes for each analysis.

5.4 Predictive Model Validation

Following the construction of unified datasets with different quality stratifications (Section 5.2), the final validation step consists of evaluating the predictive capacity of the extracted airway and parenchymal metrics through machine learning models. This section describes a three-phase validation approach: first, a comparative assessment of multiple datasets using baseline model configurations to identify the optimal quality-size tradeoff; second, an extended assessment incorporating baseline functional status as a predictive feature; and third, a systematic hyperparameter optimization on the best-performing dataset to maximize predictive accuracy.

5.4.1 Phase 1: Baseline Model Testing Across Datasets

The first phase aims to determine which of the unified datasets provides the best foundation for predictive modeling. This evaluation uses a standardized Leave-One-Out Cross-Validation (LOOCV) framework with fixed baseline hyperparameters, ensuring that performance differences reflect dataset characteristics rather than model tuning artifacts.

Model Architecture and Configuration

The baseline validation employs a multi-model ensemble approach, combining neural network, regularized regression, and tree-based methods to capture different aspects of the feature-target relationships.

Multi-Layer Perceptron (MLP): a feedforward neural network with two hidden layers (16 and 8 neurons respectively), ReLU activation, and 0.2 dropout rate. The network is trained using Adam optimizer with learning rate 1×10^{-3}

and weight decay 1×10^{-4}, with early stopping based on validation loss (patience = 100 epochs, maximum 500 epochs). Thanks to the *inner split logic*, during each LOOCV fold, 20% of the training data is reserved for validation through 10-fold inner cross-validation to optimize stopping time.

Ridge Regression: L2-regularized linear regression with regularization parameter $\alpha = 5.0$, providing *a robust baseline for capturing linear relationships* while mitigating multicollinearity.

Lasso Regression: L1-regularized linear regression with $\alpha = 0.5$, performing implicit feature selection by shrinking less important coefficients toward zero.

Random Forest: ensemble of 100 decision trees with maximum depth = 2, minimum samples for split = 5, and minimum samples per leaf = 2. These conservative parameters *prevent overfitting on small sample sizes*.

XGBoost: gradient-boosted decision tree model included to capture non-linear interactions while maintaining strong regularization. In our setting, XGBoost is configured with 100 estimators, maximum depth 2, learning rate 0.1, and L1/L2 regularization parameters `reg_alpha` = 1.0 and `reg_lambda` = 1.0 to reduce overfitting risk in small cohorts.

Weighted Ensemble: linear combination of Ridge and Random Forest predictions with weights 0.7 and 0.3 respectively, combining the complementary strengths of linear and non-linear models.

Feature Set and Target Variable

The prediction task focuses on predict $\text{FVC\%}_{\text{week52}}$, the forced vital capacity percentage at 52 weeks, using six of the baseline CT-derived features:

1. Mean peripheral branch volume (mm^3)
2. Peripheral branch density (branches per unit volume)
3. Mean peripheral diameter (mm)
4. Central-to-peripheral diameter ratio
5. Mean lung density (HU)
6. Histogram entropy (parenchymal heterogeneity)

These features were selected because they emerged as the most strongly correlated with the target variable in the correlation analysis (Section 5.2).
In addition, $FVC\%_{week0}$ was tested as an input feature in a separate phase to assess how baseline functional capacity influences the results.

Model General Evaluation

Model performance is assessed using three complementary metrics:

- **R^2 (Coefficient of Determination)**: measures the proportion of variance in $FVC\%_{week52}$ explained by the model. Values closer to 1 indicate better predictive power, while negative values indicate performance worse than a constant predictor.
- **MAE (Mean Absolute Error)**: average absolute difference between predicted and actual FVC% values, expressed in percentage points. This metric is robust to outliers and directly interpretable in clinical terms.
- **RMSE (Root Mean Squared Error)**: square root of the mean squared prediction error, penalizing larger deviations more heavily than MAE. The MAE/RMSE ratio provides insight into prediction consistency (values near 1 indicate uniform error distribution).

Given the limited sample sizes, **LOOCV** is employed to maximize training data utilization while providing nearly unbiased performance estimates.
For each fold, one patient is excluded as a test case, while the remaining n-1 constitute the training set. In the specific case of MLP, as already mentioned, the *inner split* technique is used, whereby 20% of the training set is further divided into a validation set through a 10-fold internal cross-validation, in order to prevent overfitting and determine the early stopping criterion.
All models are trained on the training set and evaluated on the excluded patient. Feature importance is calculated via permutation importance, measuring the increase in Mean Absolute Error (MAE) after shuffling each feature's values. For each feature, its values are randomly permuted 100 times while keeping other features unchanged, and the resulting MAE is compared to the baseline. This ΔMAE (change in MAE) quantifies how much the model's predictive performance deteriorates without that feature, thereby revealing its importance.
This nested cross-validation structure ensures that no test data leaks into model selection or hyperparameter tuning.
The best-performing dataset is used as the baseline for the subsequent phases, to assess how the predictive metrics can be further improved through the addition of baseline functional status and hyperparameter optimization.

5.4.2 Phase 2: Assessment of Baseline Functional Status as a Predictive Feature

Phase 2 extends the analysis by incorporating **FVC% at week 0** (baseline functional status) as an additional predictive feature so the extended feature set therefore contains seven features.
This second phase directly addresses a fundamental question: to what extent does baseline functional state itself encode prognostic information about future function? In many chronic diseases, baseline severity is a strong predictor of future trajectoryIncluding baseline FVC% as a feature tests whether knowledge of current functional status substantially enhances predictive capacity beyond structural imaging alone.
The identical five dataset configurations and six-model ensemble are re-evaluated using the extended feature set with the same LOOCV protocol. By direct comparison of Phase 1 and Phase 2 results, we quantify the **marginal improvement** from including baseline functional status:

$$\Delta R^2 = R^2_{\text{extended}} - R^2_{\text{baseline}} \tag{5.4}$$

a large positive ΔR^2 indicates that baseline function dominates predictions, potentially reducing the clinical utility of imaging metrics while a small or negligible ΔR^2 suggests that imaging metrics capture independent information beyond baseline functional status.

5.4.3 Phase 3: Hyperparameter Optimization via Compact Grid Search

The third phase performs systematic hyperparameter tuning on the best-performing dataset identified in Phase 1. Given the small sample size, the search space is carefully constrained to configurations with strong theoretical justification for small-data regimes, avoiding exhaustive searches that would risk overfitting to dataset idiosyncrasies.
Traditionally, grid search with wide hyperparameter ranges has been poorly suited for small medical datasets for several reasons. First, with validation procedures like leave-one-out, even genuine performance differences can be obscured by random variation, making it difficult to distinguish truly superior configurations from those that simply benefited by a lucky combination. Furthermore, the computational cost is significant: each configuration requires numerous model training runs, and since these are neural networks, each training run requires multiple epochs. Testing hundreds of configurations therefore becomes prohibitive in terms of time and resources. Finally, a multiple comparisons problem arises: evaluating a large

number of configurations increases the likelihood of obtaining spurious "better" results, which in reality do not generalize beyond the analyzed sample.
The compact search strategy addresses these challenges by testing only carefully selected configurations that prioritize strong regularization and conservative architectures to prevent overfitting on small samples, reducing the search space from potentially thousands to dozens of combinations.
After identifying the best-performing dataset and model in Phase 1, a focused hyperparameter tuning was carried out only on that specific model, using the selected dataset as the baseline. The specific model selected for fine-tuning therefore depends on the Phase 1 outcome (and can differ across dataset configurations), and the concrete choices and results are reported in the next chapter. The goal of this step is to assess whether the predictive metrics can be further improved, while maintaining strong regularization to avoid overfitting in a small-sample setting. The tested hyperparameter values and the corresponding search space are reported in Table A.7 in the appendix.
Each candidate configuration was evaluated with the same Leave-One-Out Cross-Validation (LOOCV) protocol used in Phase 1, and performance was measured with R^2, MAE, and RMSE. The final tuned model is therefore the best variant within the hyperparameter space of the top-performing model, rather than a new cross-model comparison.
A **parameter stability analysis** examines the distribution of hyperparameter values across top-performing configurations. If multiple different parameter combinations achieve nearly identical performance, the optimization is likely robust and generalizable. Conversely, if performance depends on narrow parameter ranges, overfitting to the specific LOOCV structure is more likely.
The optimized configuration is compared to the baseline configuration (fixed parameters from Phases 1–2), quantifying the marginal gains achievable through hyperparameter tuning. This comparison provides insight into whether the identified best configuration represents a local optimum or whether substantial unexploited performance gains remain.

Model Selection Rationale

The final model selection follows a multi-level decision-making process, balancing statistical performance with the practical requirements of clinical interpretability. The primary criterion is the coefficient of determination (R^2), which measures the proportion of variance in the target variable explained by the model. The model with the highest R^2 is generally preferred, as it indicates superior predictive power. In the event that two or more models achieve equivalent R^2 (within a predefined tolerance), secondary criteria are considered. The first is the Mean Absolute Error (MAE), where a lower value is favored due to its direct clinical relevance—it

represents the average magnitude of prediction error in the original percentage units. The second is the MAE/RMSE, a ratio close to one suggests that prediction errors are consistently small and uniformly distributed, without the influence of large outliers that would disproportionately increase the RMSE.
Finally, when models show substantially equivalent performance across these statistical metrics, a principle of parsimony is applied to favor clinical interpretability. Simpler models are preferred over more complex ones, as they offer greater transparency and ease of explanation to clinicians. Specifically, linear models (e.g., Ridge Regression) are chosen over tree-based ensembles (Random Forest, XGBoost), which in turn are preferred over neural networks (MLP). This hierarchy ensures that the selected model is not only accurate but also understandable in a medical setting, where insight into the decision-making process is as important as the prediction itself.
This selection strategy ensures that the final model is both statistically robust and practically useful for deployment, balancing raw accuracy with the interpretability required for clinical decision-making.

5.4.4 Validation of Validation step

Starting from the Leave-One-Out Cross-Validation technique, although it represents a methodological choice that maximizes the use of available data, it is important to recognize some intrinsic limitations of this approach for a correct interpretation of the results.
First, the performance estimate obtained through LOOCV is characterized by an intrinsically high variance. With few folds corresponding to the patients in the sample, the confidence intervals of the predictive metrics are necessarily wide, and random variability can sometimes obscure substantial differences between different models.
A second aspect concerns the dependence between the folds. In LOOCV, each validation fold shares $n - 2$ training samples with all the other folds, generating a structural correlation that can lead to an optimistic performance estimate compared to what would be observed with a validation on an independent, unseen sample.
Finally, there is an inherent risk of overfitting in the model selection phase. Choosing the optimal configuration based on the average R^2 obtained from LOOCV could reward a model that, despite never having seen a single fold during training, has nevertheless adapted to the overall structure and peculiarities of the dataset.
In other words, model selection still takes place using information from the entire sample, and this can introduce an optimistic bias in the final performance estimate.
A first measure involves the conservative selection of hyperparameters. The search grid was deliberately limited, excluding overly flexible configurations such as deep neural networks or trees with great depth, which could have exploited accidental

peculiarities of the sample to the detriment of generalizability.
At the same time, particular emphasis was placed on regularization techniques. All neural networks implement dropout and weight decay mechanisms, while decision tree-based models use limited depth and strong regularization penalties. These measures prevent memorization phenomena, instead favoring the learning of generalizable patterns.
The adoption of multiple metrics provides an additional level of control. Simultaneously evaluating R^2, MAE, and RMSE reduces the risk of optimizing a single indicator, which could, by chance, reward overfitting models. Therefore, the convergence of positive results across multiple metrics offers a greater guarantee of the model's quality.
A particularly important aspect is the feature stability analysis. The variable importance rankings, obtained via permutation importance, are systematically examined across different models and configurations. When a stable hierarchy emerges in the features, it is a strong indicator that the model is capturing a genuine biological signal rather than artifacts specific to the dataset.
The validation results must be interpreted with awareness of their specific epistemic context, carefully distinguishing what they actually demonstrate from what they cannot prove.
The developed models demonstrate the ability to explain a portion of the variance in FVC% at 52 weeks starting from the baseline CT characteristics. This result establishes a non-trivial fact: the metrics extracted from the airways and parenchyma actually encode prognostic information relevant to the evolution of the disease.
In other words, there is a predictive signal in the images that the models were able to capture.
However, it is equally important to clarify what these results do not demonstrate. They do not establish causal relationships: the predictive capacity does not imply that the identified characteristics are causal mechanisms of functional decline, and they do not guarantee generalizability to other populations, as the validation is limited to patients with idiopathic pulmonary fibrosis from the OSIC cohort. All the results are described in a specific way in the next chapter.

Chapter 6

Validation Results

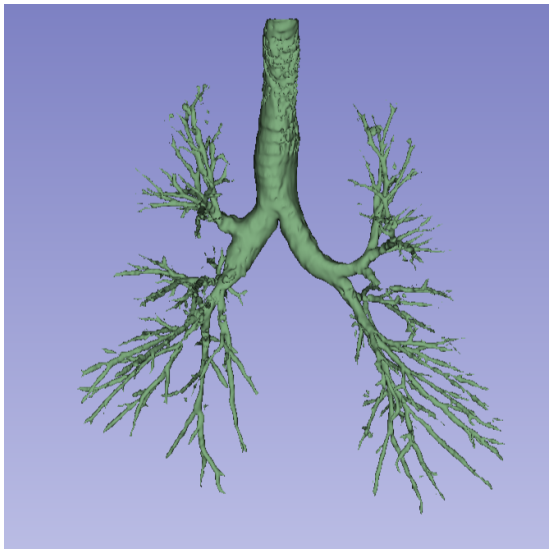
This chapter presents and discusses the results obtained in the validation phase described in Chapter 5. The structure mirrors the steps introduced in the previous chapter, in order to ensure continuity between methodology and results.

6.1 Segmentation validation: qualitative examples

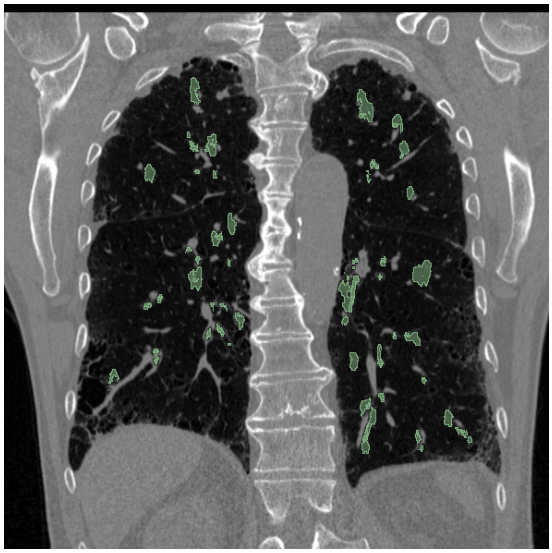
In Chapter 5.1 we introduced the technical validation protocol adopted to assess the anatomical plausibility of the extracted airway trees, complementing aggregated metrics with a set of rule-based checks derived from known morphological constraints (e.g., limits on maximum generation depth and related indicators; see Appendix A, Table A.1). In this section, we report qualitative examples to illustrate how those criteria manifest in practice on real cases: first, a representative failure mode flagged as *unreliable*, and then a *reliable* case that provides a visual reference for technically plausible output.

Example of an unreliable segmentation - excessive generation depth Beyond the aggregated statistics, qualitative examples help clarify how the technical checks translate into concrete failure modes. Figure 6.1 shows a representative case flagged as *unreliable* because the extracted airway tree reaches a maximum generation depth of 40, exceeding the physically plausible upper bound of 35 generations defined in the technical limits (Appendix A, Table A.1).

In this case, both the 3D rendering of the segmentation and the CT slices with the overlay clearly show that the mask extends unrealistically far into the lung periphery, suggesting oversegmentation.



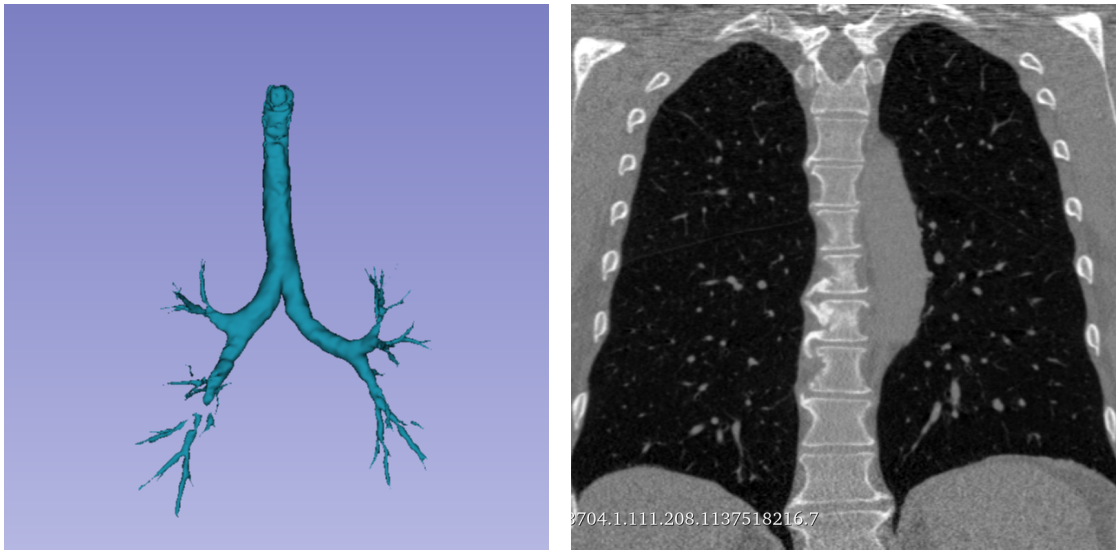
(a) 3D rendering of the airway segmentation.



(b) CT slices with segmentation overlay highlighting distal leakage.

Figure 6.1: Qualitative example of an *unreliable* segmentation flagged by an excessive maximum generation depth (40 vs. the technical upper bound of 35).

Example of a reliable segmentation Including a *reliable* example alongside the failure case is useful to provide a visual reference of what the pipeline considers technically plausible output. Figure 6.2 shows a representative case that passes all technical checks; the 3D airway tree appears coherent and the overlay remains confined to anatomically plausible airway lumen regions without distal leakage. While the approach does not replace clinical validation by experienced radiologists, nor does it account for subtle artifacts detectable only by visual inspection (e.g., motion or metal artifacts), it does provide a level of rigorous and reproducible quality control that strengthens the methodological foundation for downstream quantitative analysis in heterogeneous disease populations.



(a) 3D rendering of the airway segmentation (reliable case).

(b) CT slices with segmentation overlay (reliable case).

Figure 6.2: Qualitative example of a *reliable* segmentation passing the technical plausibility checks.

6.2 Segmentation Validation: quantitative Results

The technical validation process, described in detail in Section 5.1, was applied to 45 cases that successfully passed the initial filtering criteria defined in Chapter 3 (see Table 3.1). Of these, 39 cases (86.7%) were classified as *reliable*, while 6 cases (13.3%) were considered *unreliable*. Therefore, only technically reliable cases were considered for downstream clinical correlation and predictive analyses, ensuring that subsequent results were not influenced by segmentation artifacts.
All 45 cases passed the volume plausibility check, which highlights how automatic segmentation consistently produced anatomically plausible volume estimates without preventing errors in other technical parameters, particularly in the unreliable cases (Figure 6.3).

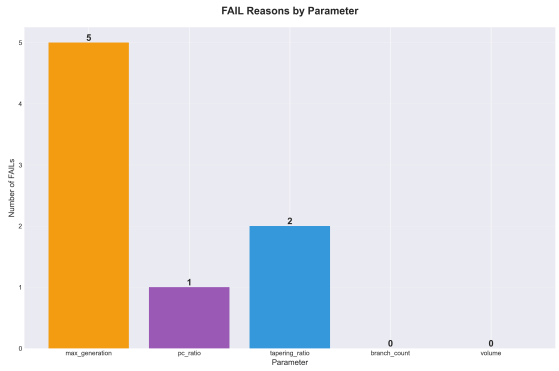


Figure 6.3: Distribution of technical failures by parameter. Maximum generation depth was the most common cause of failure, occurring in 4 out of 6 unreliable cases.

The most common failure was caused by maximum generation depth, which exceeded acceptable limits in 4 of the 6 unreliable cases. Additional failures were observed for the tapering ratio (2 cases) and the peripheral-to-central ratio (1 case). This pattern suggests that excessive branch depth is the primary cause of segmentation unreliability in our cohort.

The relationship between volume and branch count (Figure 6.4) reveals that unreliable cases (red dots) tend to occupy a distinct region of the feature space, characterized by a high number of branches disproportionate to the corresponding airway volume. This pattern is consistent with oversegmentation artifacts that generate an excessive number of terminal branches without proportionally increasing the overall airway volume.

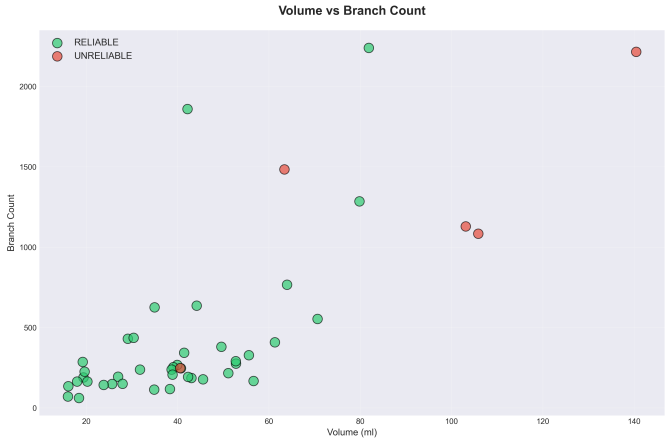


Figure 6.4: Relationship between airway volume and branch count.

Therefore, it is safe to say that the technical validation successfully identified 6 cases (13.3% of the cohort) with segmentation artifacts that could potentially confound clinical interpretation. The main source of error was excessive maximum generation depth, often accompanied by high branch counts and abnormal peripheral-to-central ratios. These patterns are consistent with known difficulties in airway segmentation, particularly in low-contrast image regions, where algorithms may erroneously extend branches beyond anatomically plausible limits.
Excluding these 6 unreliable cases from subsequent clinical analyses ensures that the remaining 39 cases represent a technically robust dataset.
This filtering step, as mentioned above, is essential to maintain the integrity of downstream statistical analyses and clinical correlations, as segmentation artifacts could otherwise introduce systematic biases or spurious associations.

6.3 Clinical Correlation Analysis

This section presents the results of the correlation between baseline metrics (both airway andparenchymal) and FVC%, as described in Section 5.2.
The primary objective of this analysis was to determine whether metrics extracted from baseline CT scans correlated with clinical progression of lung disease, as measured by percent predicted Forced Vital Capacity (FVC%).
We investigated two main aspects of the clinical correlation. First, we performed a cross-temporal correlation analysis: for each patient, we took the baseline imaging metric (measured at baseline CT) and correlated it with all available FVC% measurements from that patient across all follow-up timepoints (weeks 0, 10, 26, 52, etc.). This approach allows us to assess whether baseline morphometric and densitometric phenotypes are associated with the patient's overall lung function trajectory.
Second, we stratified patients based on their baseline metric values into quartiles and analyzed how different levels of these imaging biomarkers are associated with the subsequent evolution of FVC% over time, including calculating individual patient rates of decline.

6.3.1 Cross-Temporal Correlation: Baseline Metrics with Overall FVC% Trajectory

The linear relationship between each imaging metric and FVC% was assessed using the **pearson correlation coefficient**, computed across all patient-timepoint pairs. The full set of detailed results is visually summarized in Figure 6.5.

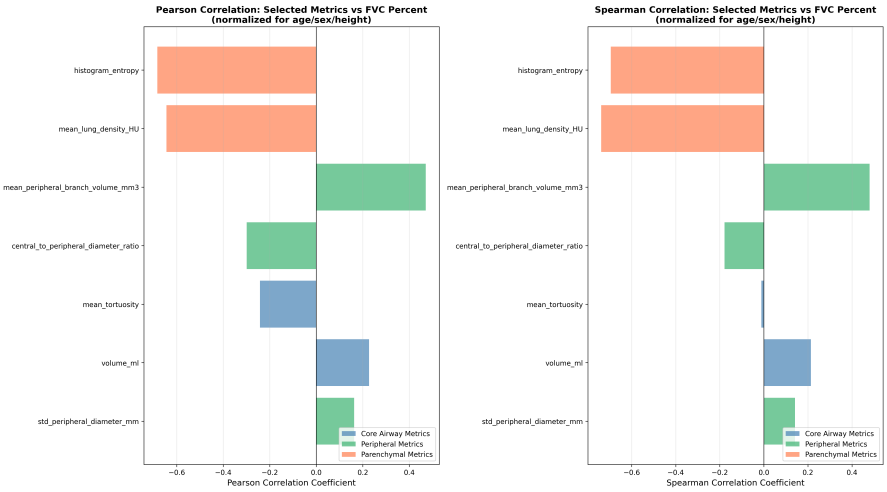
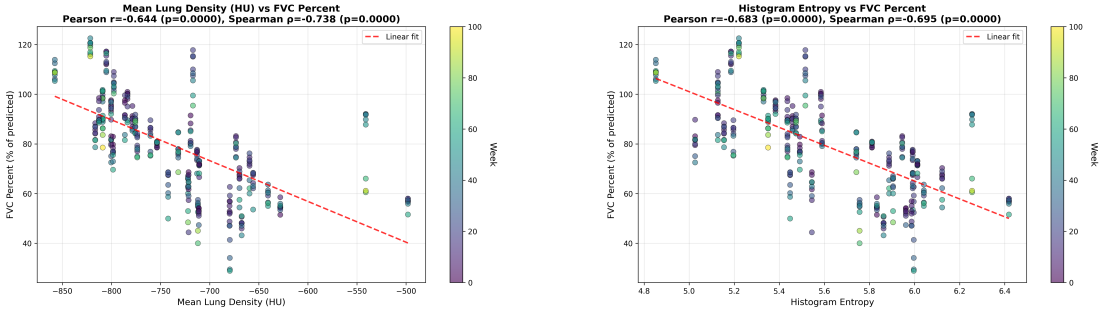


Figure 6.5: Summary of Pearson and Spearman correlation coefficients between baseline imaging metrics and FVC%.

The analysis revealed a spectrum of correlation strengths across different types of metrics. The most striking results were observed for parenchymal metrics, which demonstrated strong and statistically significant negative correlations with FVC%. Specifically, both mean lung density (HU) (pearson's r = -0.64, $p < 0.001$) and histogram entropy (pearson's r = -0.68, $p < 0.001$) were strongly associated with lower FVC%.
This indicates that patients with denser lungs and more complex and heterogeneous parenchymal structure, as quantified by entropy, are more likely to have worse lung function across their disease trajectory.

This can be seen in the scatterplots for these two metrics, shown in Figures 6.6a and 6.6b.



(a) Mean Lung Density **(b)** Histogram Entropy

Figure 6.6: Scatter plots of selected metrics vs. FVC%, with week as the color scale. The red dashed line represents a linear fit. (a, b) Parenchymal metrics show strong negative correlations.

Among the airway metrics, several showed statistically significant, though generally weaker, correlations. In particular, `mean_peripheral_branch_volume_mm3` exhibited a moderate positive correlation with FVC% ($r = 0.47$, $p < 0.001$), indicating that larger peripheral airway volumes are associated with better respiratory function. This trend is clearly visible in the scatter plot in Figure 6.7.

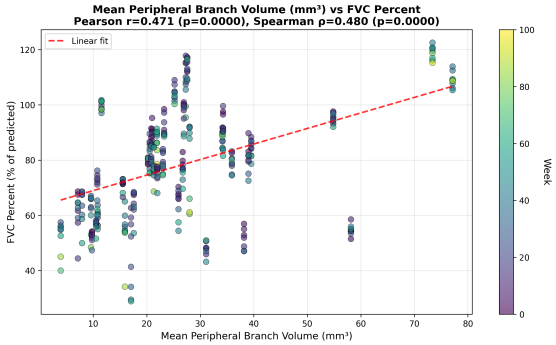


Figure 6.7: Relationship between `mean_peripheral_branch_volume_mm3` and FVC%.

The remaining airway metrics showed only weak associations with FVC%. Specifically, `central_to_peripheral_diameter_ratio` was weakly negative ($r = -0.30$, $p < 0.001$), while `volume_ml`, `mean_tortuosity`, and `std_peripheral_diameter_mm` exhibited small-magnitude correlations ($|r| = 0.16$–0.24).
The additional scatter plots for these other metrics are reported in the dedicated Appendix section (Appendix B.1, Chapter B), specifically in Figure B.1.

A sensitivity analysis was conducted to assess the impact of including scans that underwent a smoothing kernel correction. The results were consistent across the original, smoothed, and combined datasets, indicating that smoothing does not materially affect the parenchymal correlations. This supports analyzing the full dataset without introducing relevant bias.

6.3.2 Association of Baseline Metrics with FVC% Evolution

To determine whether CT-derived baseline metrics could predict disease progression, patients were divided into quartiles based on their baseline metric values, and their FVC% trajectories were analyzed over time. For each patient, a subject-by-subject linear regression of FVC% over the weeks of follow-up was used to estimate an individual rate (%/week), which was then averaged within each quartile group.
Of all the metrics examined, **mean lung density** and **mean peripheral branch volume** emerged as the most important information for capturing differences in longitudinal disease behavior.
Both metrics revealed a consistent pattern: patients at the extremes of the distribution, that is, with very low or very high baseline values, tended to decline more slowly than those in the middle quartiles.
Regarding peripheral branch volume, Q3 patients (mean 29.83 mm^3) showed the fastest decline (-0.193 ± 0.168 %/week), while Q1 and Q4 were comparably slower (-0.106 and -0.096 %/week, respectively), as shown in Table 6.1 and Figure 6.8.

Quartile	Baseline (mm^3)	Decline (%/week)	SD
Q1	13.16	−0.106	0.102
Q2	22.58	−0.122	0.084
Q3	29.83	−0.193	0.168
Q4	45.22	−0.096	0.133

Table 6.1: FVC% decline rates stratified by baseline quartiles of mean peripheral branch volume (in mm^3).

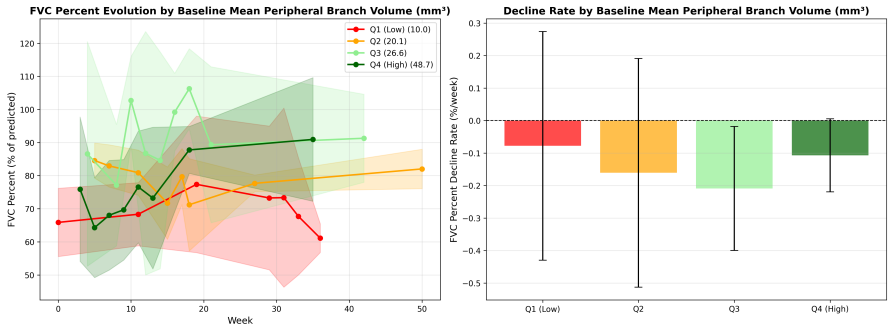


Figure 6.8: FVC% trajectories by baseline quartiles of mean peripheral branch volume. Q3 shows the steepest decline, while Q1 and Q4 decline more slowly, consistent with a non-linear trend.

A similar picture emerged for lung density: Q2 patients (mean −770.46 HU) showed the steepest decline (-0.205 ± 0.162 %/week), while those with less dense lungs (Q4, mean −627.35 HU) showed the most stable trajectory (−0.092 %/week), as reported in Table 6.2 and illustrated in Figure 6.9.

Quartile	Baseline (HU)	Decline (%/week)	SD
Q1	−806.31	−0.132	0.124
Q2	−770.46	−0.205	0.162
Q3	−705.56	−0.180	0.193
Q4	−627.35	−0.092	0.125

Table 6.2: FVC% decline rates stratified by baseline quartiles of mean lung density (in HU).

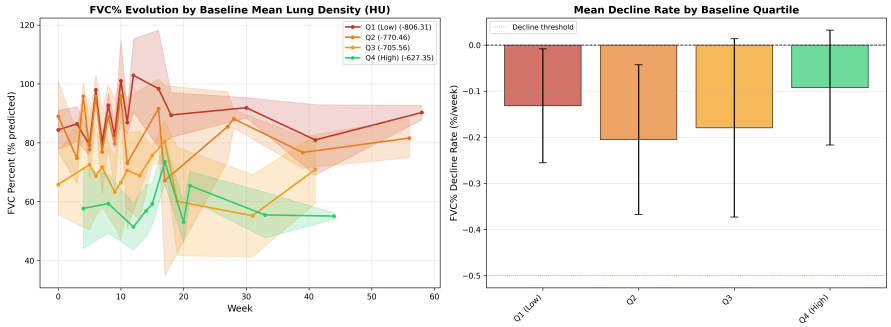


Figure 6.9: FVC% trajectories by baseline quartiles of mean lung density. Q2 and Q3 show the fastest decline, while Q4 declines more slowly, consistent with a non-linear trend.

Overall, these results suggest that the relationship between baseline imaging biomarkers and subsequent decline in FVC% is inherently nonlinearand that no single metric captures the full complexity of IPF progression. Nonetheless, peripheral airway volume and parenchymal density stand out as the most promising candidates for patient stratification, given their ability to distinguish groups with significantly different decline trajectories.
Additional decline plots for the other metrics are reported in the dedicated Appendix section (Appendix B.1, Chapter B).

6.4 Results of FVC Prediction Analysis

In this section, we report the predictive performance of six baseline imaging metrics across four distinct FVC-related targets: FVC% at week 0, FVC% at week 52, traditional FVC% decline (week 0 value - week 52 value), and direct annual FVC decline derived from longitudinal regression.
For readability, the main text reports only representative feature-specific prediction plots, the remaining plots from are provided in Appendix B.3 (Chapter B).

6.4.1 Data Composition and Sample Availability

Four predictive endpoints were defined, encompassing two conceptual approaches to quantifying disease progression. The first approach captures absolute lung function at fixed clinical intervals: predicted FVC% at baseline (week 0) and at one-year follow-up (week 52). The second approach captures functional decline either as traditional decline (week 0-week 52, a simple difference between the two interval values) or as direct annual decline (estimated by linear regression on all longitudinal FVC measurements).
A critical distinction emerges in sample availability based on the prediction target. Traditional endpoints require both a baseline measurement and a one-year measurement of adequate quality from the same patient whose quality has been assessed per measurement, as described in detail in 5.3.1.
Looking at the data in Table 6.3, a clear picture emerges regarding the four predictive targets considered in this study.
All target values are first estimated at the patient level and then summarized at the cohort level using mean, standard deviation, minimum, and maximum.
Specifically, *FVC% at Week 0*, *FVC% at Week 52*, and *Decline (Week 0-Week 52)* are calculated on the subset of 32 patients (`dataset_both_targets.csv`), i.e., patients for whom both time-point endpoints are available with acceptable quality. For each patient, the decline is calculated as:

$$\text{FVC\_drop\_percent} = \text{FVC\_percent\_week0} - \text{FVC\_percent\_week52}.$$

Annual decline (%/year) is instead calculated on the subset of 39 patients, where each patient has at least three longitudinal FVC measurements and the decline is estimated as the slope of a linear regression over time.
With this in mind, the FVC% values at week 0 and week 52 (approximately 82 and 72, respectively) indicate that the cohort includes patients with significantly different baseline conditions: some with relatively preserved lung function and others with already more compromised lung function. This heterogeneity is further supported by the high standard deviations (approximately 19-20 points), which

show that the values are widely dispersed rather than concentrated around a single profile.
The same trend is observed for progression metrics. For endpoint decline, the standard deviation is close to the mean, indicating strong interindividual variability in disease progression.

Target Variable	**N**	**Mean**	**SD**	**Min**	**Max**
FVC% at Week 0	32	81.9	19.4	48.1	129.8
FVC% at Week 52	32	71.9	20.2	28.9	109.5
Decline (Week 0 – Week 52)	32	10.0	12.7	−0.5	65.0
Annual Decline (%/year)	39	9.3	9.8	−1.0	46.0

Table 6.3: Summary statistics for the four prediction targets.

In practical terms, **idiopathic pulmonary fibrosis does not follow a single, uniform trajectory**: some patients deteriorate rapidly, while others remain relatively stable throughout the observation period.

6.4.2 Visual Summary of Feature Performance

A comprehensive overview of the predictive performance across all feature-target combinations is provided in Figure 6.10. Color separation in the heatmap immediately reveals the previously discussed hierarchy of predictive value, whereby parenchymal metrics significantly outperform airway features. This image illustrates the key point of the analysis that in this patient group, FVC depends on disease burden and tissue remodeling, not just on how open the airways are.

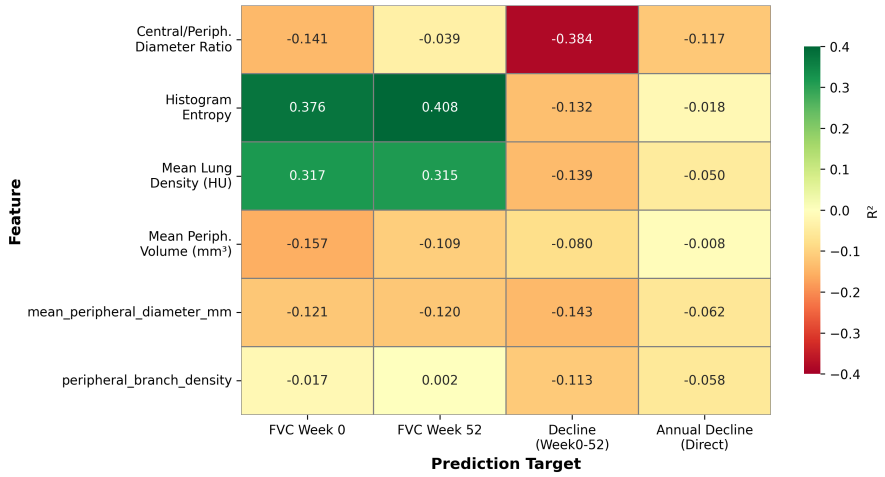


Figure 6.10: R^2 performance heatmap across all features and targets.

In the Table 6.4 are presented all six features ranked by their R^2 for the primary clinical endpoint (FVC% at week 52). The ranking reveals a pronounced separation: the top two features (histogram entropy and mean lung density) achieve $R^2 > 0.3$, while the remaining four airway metrics cluster at $R^2 < 0.02$ demonstrating the fundamental mismatch between airway morphometry and FVC status.

Feature	R^2	MAE	Pearson r
`histogram_entropy`	0.408	13.35	0.64
`mean_lung_density_HU`	0.315	11.74	0.59
`periphery_branching_density`	0.002	15.18	0.17
`central_to_peripheral_diameter_ratio`	-0.039	16.48	0.09
`mean_peripheral_branch_volume_mm3`	-0.109	16.51	-0.10
`peripheral_mean_diameter_mm`	-0.120	16.41	-0.03

Table 6.4: Feature ranking by R^2 for the primary target FVC% at week 52.

The superiority of parenchymal metrics is visually evident in Figure 6.11, where `mean_lung_density_HU` shows a clear linear trend with preserved scatter structure across the prediction range. The corresponding Bland-Altman plot reveals a symmetric distribution of residuals around the mean difference, suggesting minimal systematic bias in predictions at the extremes of the functional spectrum.

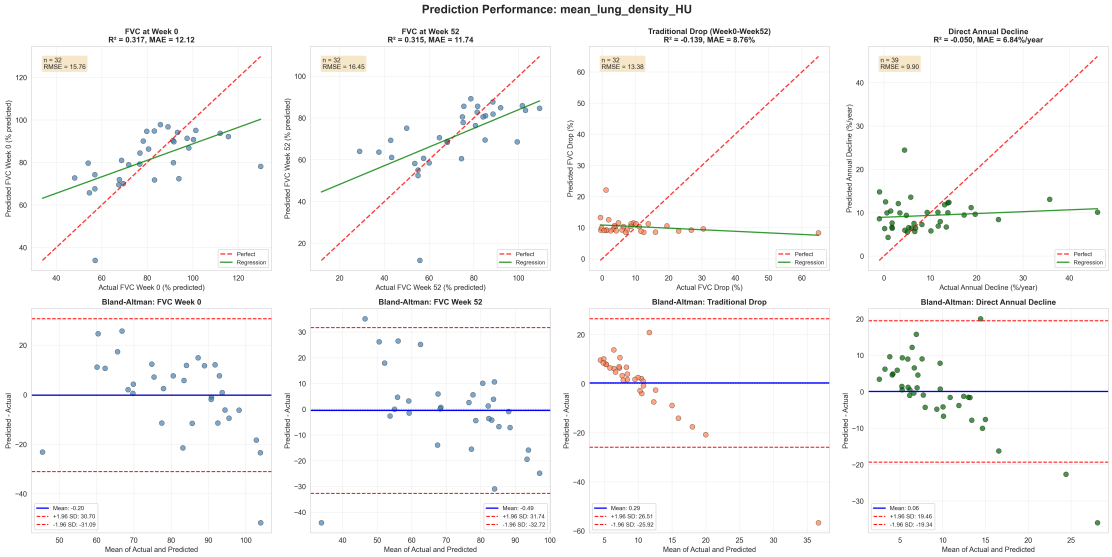


Figure 6.11: Prediction performance for `mean_lung_density_HU`. Scatter plots (top) and Bland–Altman plots (bottom) are shown for week 0, week 52, traditional decline, and annual decline. The metric predicts cross-sectional FVC% well ($R^2 > 0.31$) but not decline ($R^2 < 0$), with minimal bias at week 52.

A quantitative comparison of feature categories is presented in Figure 6.12, FVC week 52 has the mean R^2 of parenchymal features ($R^2 = 0.36$) exceeds that of airway features ($R^2 = -0.07$) by more than a factor of five. This gap persists across all cross-sectional targets but collapses for decline prediction, where both categories show poor performance (mean $R^2 < 0$ for both).

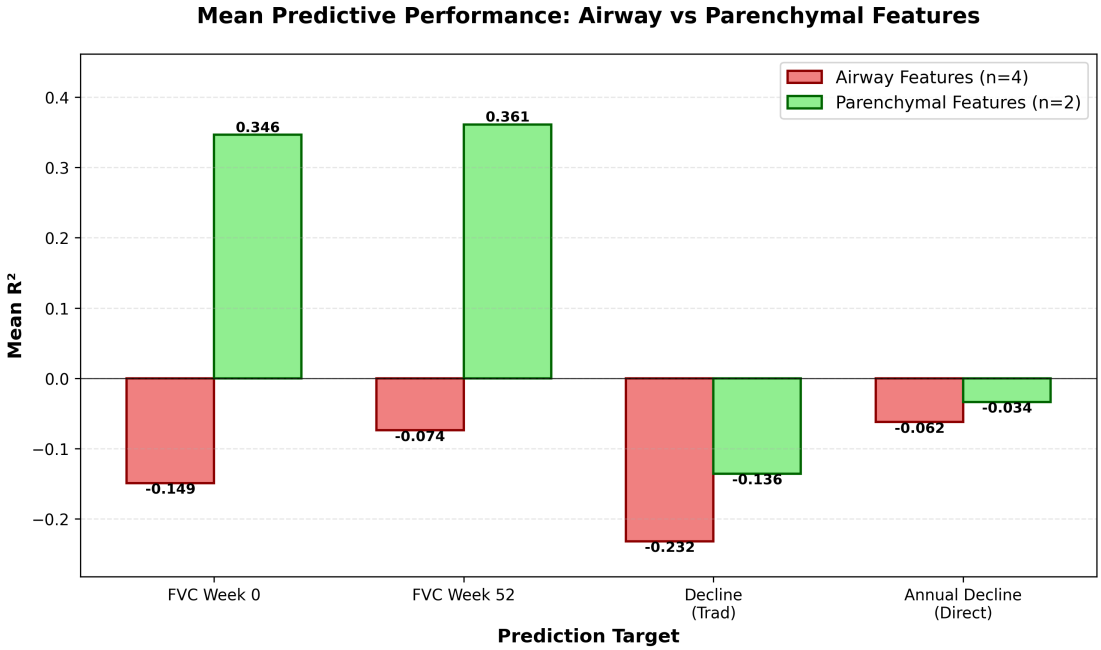


Figure 6.12: Mean R^2 by feature category across the four prediction targets: airway (red, n=4) versus parenchymal (green, n=2). Parenchymal features perform better for FVC% at week 0 and week 52, while both categories show similarly low performance for decline targets.

6.4.3 Airway Metrics Lack Predictive Signal

Unlike parenchymal metrics, airway-derived characteristics have proven to be of little use in predicting FVC status. Their weakness is clear from the initial analysis: parameters such as mean peripheral branch volume or peripheral diameter show negligible values, with inconsistent correlations and no statistical significance. The same pattern is repeated for other metrics such as the central-to-peripheral diameter ratio, confirming that, regardless of the metric considered, airways do not offer reliable predictive signals. In practice, what happens at the airway level appears to matter little in determining lung function in this cohort.

Figure 6.13 confirms this pattern: scatter plots show no clear trend and Bland–Altman plots show wide variability. Even for annual decline, the best airway feature remained near zero-skill ($R^2 = -0.008$).

Additional prediction plots are available in Appendix B.3. A full numerical summary of all feature-target combinations is reported in Appendix Table B.1.

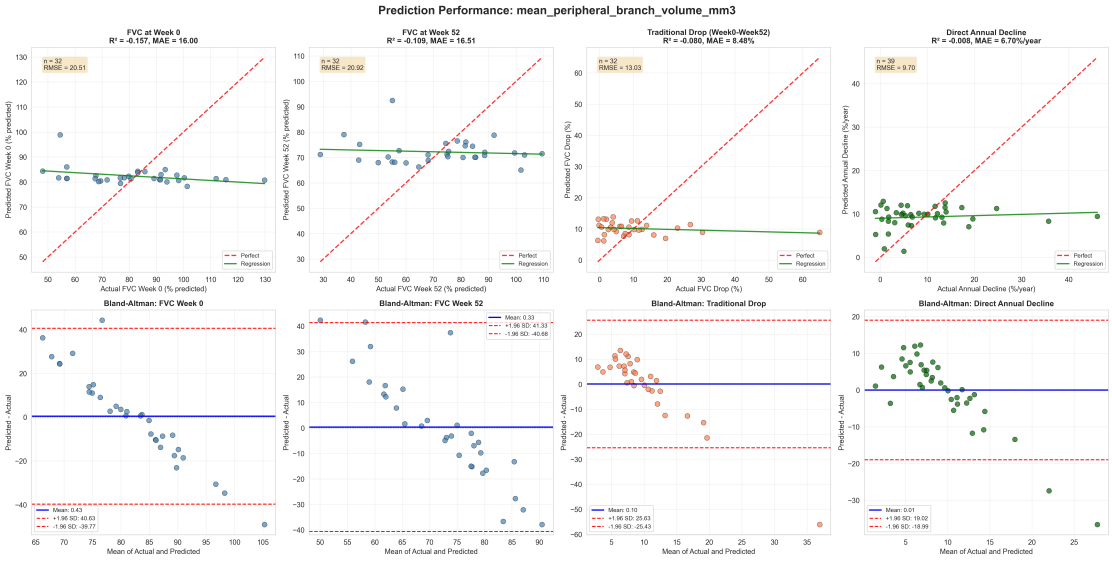


Figure 6.13: Prediction results for `mean_peripheral_branch_volume_mm3`. Top: scatter plots; bottom: Bland–Altman plots for week 0, week 52, traditional decline, and annual decline. The metric shows weak performance in all tasks ($R^2 < 0$), with high variability and limited clinical usefulness.

The lack of predictive power in airway metrics is noteworthy given their mechanistic relevance to airflow obstruction. This disconnect may reflect several possibilities: FVC is a global measure of lung function integrating multiple physiological compartments (airway, parenchyma, pleura, chest wall) not equally represented by local airway metrics; simple morphometric measures (diameter, volume) may not

adequately capture dynamic airway behavior or compliance; or the relationship between airway remodeling and functional impairment may be highly nonlinear and patient-specific, requiring more sophisticated modeling approaches.

6.4.4 Decline Prediction Remains Poor Across All Features

Both traditional and direct annual decline were difficult to predict from baseline characteristics, regardless of the metric used. For traditional decline, no model was able to provide reliable predictions: in practice, all performed worse than a simple mean-based prediction, producing errors so large that they overlapped with normal physiological patient variability. The only partial exception was histogram entropy, which showed a modest negative correlation, but still not enough to build a useful predictive model.

The picture is no better for direct annual decline. Even the best-performing metric proved essentially useless, with predictive power virtually indistinguishable from a chance prediction.

Figure 6.14 shows a clear pattern: for cross-sectional measures, percentage errors are higher but still small relative to the overall value range. For decline measures, percentage errors are lower in absolute terms, yet large compared with the very small average decline.

This highlights a fundamental difficulty: predicting individual evolution of lung function would require capturing highly personal factors that are simply not reflected in the lung anatomy measured at baseline.

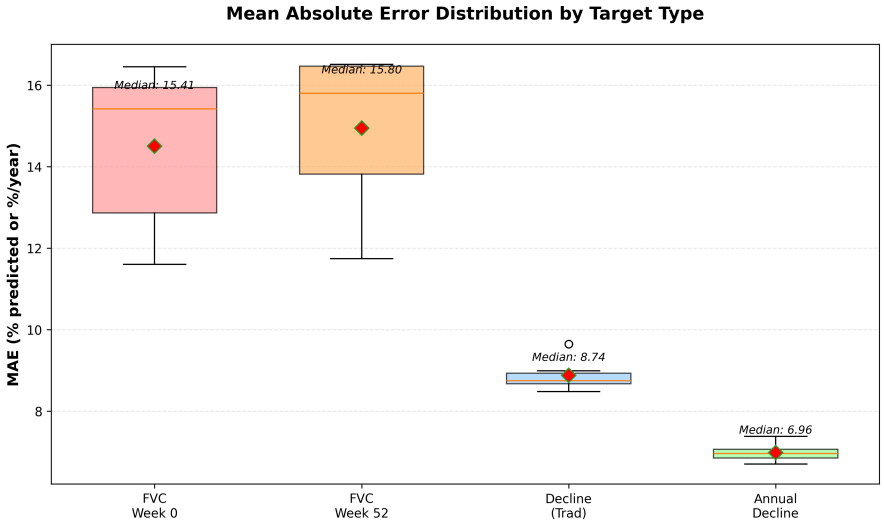


Figure 6.14: Distribution of mean absolute error (MAE) across features and targets.

This finding suggests a fundamental limit: **baseline imaging features, whether airway or parenchymal, contain insufficient information to predict individual patient decline trajectories**. The weak predictive signal for decline contrasts sharply with the moderate signal present for cross-sectional FVC status, indicating that the mechanisms driving longitudinal decline differ substantially from those associated with absolute functional level at a given timepoint.

6.5 Baseline-Method Analysis for FVC% Prediction at Week 52

In light of the results presented so far and their clinical relevance, FVC% at week 52 was selected as the primary prediction target.
This endpoint provides a clinically meaningful one-year functional outcome and allows us to evaluate whether baseline imaging phenotypes capture prognostic information beyond cross-sectional associations.
This section presents the baseline prediction analysis for FVC% at week 52 using just the six features discussed until now.
The baseline analysis encompassed five datasets with sample sizes ranging from 22 to 34 patients as descrier befor in Chapter 5.3.5. The `all` dataset, comprising all available patient records (N=34), demonstrated the highest predictive performance with R^2=0.4162. Notably, the `strict` dataset, despite containing the smallest sample (N=22), achieved a competitive MAE of 11.68%, suggesting robust prediction accuracy despite limited data.
Leave-One-Out Cross-Validation (LOOCV) was implemented to assess model generalization performance across all datasets and algorithms and are evaluated eight machine learning models during the baseline analysis.
A complete comparison of all evaluated models across all dataset configurations is provided in Appendix C, Table C.1.

Dataset	Model	N	R^2	MAE	RMSE
all	Ensemble (Ridge+RF)	34	0.4162	11.72	14.90
strict	Random Forest	22	0.3568	11.68	14.23
balanced	Ensemble (Ridge+RF)	32	0.4075	11.94	15.29

Table 6.5: Best performing model for each dataset.

Although five dataset configurations were created in the previous step, only three are reported here all, `strict`, and balanced because he traditional_only and both_targets configurations were not analyzed separately in this section because they contain the same patient sample as balanced, and therefore do not provide an additional independent comparison.
The ensemble method combining Ridge regression with Random Forest emerged as the optimal model for the larger datasets, achieving R^2 values while for the strict dataset, Random Forest slightly outperformed the ensemble approach (R^2=0.3568 vs. 0.3017).
The analysis raises some critical observations. First, deep learning models, particularly MLP networks, showed disappointing performance across all datasets, with

negative R^2 values indicating severe overfitting, likely due to the small sample size. Linear regression based on the single best feature also proved inadequate, with R^2 consistently below zero, indicating ineffective variable selection. Regularized linear models (Ridge and Lasso) achieved moderate results, with R^2 ranging from 0.06 to 0.36; in particular, Ridge generally outperformed Lasso in baseline performance. Finally, ensemble methods based on decisive trees stood out for their robustness, proving to be the most reliable overall.
The performance variation across datasets suggests that dataset composition and patient cohort characteristics significantly impact model predictability.
Model performance varies significantly based on the dataset used, highlighting how sample composition and the specific characteristics of the patient cohort significantly impact predictive power. The dataset called **all**, which with its 34 patients represents the largest and most diverse sample, achieved the highest R^2 value overall, equal to 0.4162. This suggests that a larger sample size and variety favor better regularization and generalization of the model. The ensemble combining Ridge and Random Forest achieved a mean absolute error (MAE) of 11.72%, meaning that predictions deviated on average by approximately ±12% from the actual FVC values.
If we consider the **strict** dataset, consisting of only 22 patients, despite its small size, competitive predictive accuracy is still recorded, with an MAE of 11.68%, only marginally higher than that obtained on larger samples. However, the R^2 value of the best model is significantly lower (0.3568), indicating greater variance in the quality of the predictions. The good performance of Random Forest on this specific dataset suggests that the restrictive selection criteria isolated patients with specific clinical characteristics, for which nonlinear interactions between variables play an important role.

The strongest predictive result in the baseline analysis is obtained on the *all* dataset (N=34): the Ensemble model (Ridge+RF) achieved the highest performance, with $R^2 = 0.4162$ and MAE = 11.72%.

6.6 Extended FVC% at Week 52 Prediction Analysis Including Baseline FVC% at Week 0

The incorporation of baseline FVC measurement (week 0) as an additional feature represents a critical methodological refinement, as baseline lung function is a well-established predictor of disease progression in respiratory conditions. This section documents the impact of this feature augmentation on predictive performance.
Baseline FVC is highly predictive of subsequent absolute FVC values due to strong inter-individual heterogeneity in baseline lung function. Including this

feature reduces prediction variance attributable to patient-level baseline differences, allowing models to better capture disease-specific progression patterns. However, this improvement comes at the cost of practical interpretability: the model becomes more of a baseline-adjustment model than a purely prognostic model based on structural features alone.

The integration of baseline FVC produced dramatic improvements across all datasets.

Dataset	**With FVC Week 0**			ΔR^2
	Model	**R^2**	**MAE**	
strict	Lasso	0.7769	6.87	+0.4201
balanced	Ensemble	0.5354	9.37	+0.1279
all	XGBoost	0.5024	10.31	+0.0862

Table 6.6: Best Models with FVC Week 0 and Relative Improvement (ΔR^2).

The addition of FVC week 0 produced substantial R^2 improvements ranging from +0.0862 to +0.4201. The *strict* dataset exhibited the largest absolute gain, with R^2 improving from 0.3568 to 0.7769 (ΔR^2=+0.4201), representing a 118% relative improvement. Mean Absolute Error decreased from 11.68% to 6.87%, an improvement of 41%.

The analysis of the performance increase, measured as a change in R^2, reveals a differentiated trend depending on the composition of the sample considered. It is clear that the incremental contribution of the model is more pronounced the smaller and more clinically homogeneous the dataset, as in the case of the strict dataset.

In the most selected sample, defined as strict and composed of only 22 subjects, the introduction of the predictor resulted in a significant leap, with a ΔR^2 of +0.4201. This data suggests that in a context where inter-patient variability is low, the baseline forced vital capacity (FVC) acquires proportionally greater weight, becoming a decisive discriminating factor.

Expanding our focus to a larger, more balanced sample, such as the balanced sample that includes 32 observations with traditional characteristics, the absolute gain stands at a ΔR^2 of +0.1279; the contribution, while remaining solid, begins to decline.

This phenomenon is further accentuated in the dataset that includes all observations (all, N=34), where the ΔR^2 drops to +0.0862, a relative increase of 21%.

This decreasing pattern can be interpreted in light of the increasing heterogeneity of larger samples. In more diverse datasets, individual variability is already partially absorbed both by the greater variety of available features and by the "implicit

ensemble" effect provided by the average over a larger number of cases. Consequently, the additional predictive power of the single FVC variable tends to decline, offering a marginally positive but less dramatic contribution compared to more uniform clinical contexts.

6.7 Fine-Tuning the Best Model

Based on the findings in Section 6.6, the *strict* dataset combined with the extended feature set (including FVC% week 0) was identified as the optimal configuration.
In this chapter is described the systematic hyperparameter optimization of the best model (Lasso with L1 regularization) using grid search methodology.
To systematically explore the effect of regularization, the α parameter was varied across a grid of ten discrete values, covering a range from 0.01 to 2.0.
The choice of this specific range responds to the need to test the model under increasingly restrictive conditions. Starting with a very weak regularization, gradually increasing the regularization, one encounters moderately strong values, such as those between 0.3 and 0.7, which represent a typical compromise between reducing variance and containing bias. Finally, at the upper end of the grid, α values equal to or greater than 1.0 require strong regularization, pushing the model toward sparse solutions and effectively favoring more aggressive feature selection.
Leave-One-Out Cross-Validation (LOOCV) was applied to each configuration to ensure reliable performance estimates on the small sample (N=22). Each of the 10 configurations underwent complete LOOCV, generating prediction residuals for all 22 patients and the results are reported in Table 6.7.

α	$\mathbf{R^2}$	**MAE (%)**	**RMSE (%)**
2.0	0.8104	6.40	7.72
1.5	0.8087	6.40	7.76
1.0	0.7978	6.56	7.98
0.7	0.7823	6.82	8.28
0.5	0.7769	6.87	8.38
0.3	0.7729	6.88	8.45
0.2	0.7599	7.13	8.69
0.1	0.7432	7.43	8.99
0.05	0.7369	7.46	9.10
0.01	0.7325	7.49	9.18

Table 6.7: Lasso grid gearch results on the strict data with FVC Week 0.

Weak Regularization ($\alpha \leq 0.1$): Models with $\alpha \in [0.01, 0.1]$ demonstrated the poorest performance ($R^2 \in [0.7325, 0.7432]$, MAE $\geq 7.43\%$), consistent with insufficient feature selection and potential overfitting to LOOCV training folds. The increasing error metrics in this range suggest that without adequate regularization pressure, Lasso fails to achieve appropriate sparsity in the 7-dimensional feature space.

Moderate Regularization ($\alpha \in [0.2, 1.0]$): In this range, performance improved consistently with increasing α, reflecting better balance between model complexity and bias. The transition from $\alpha = 0.5$ to $\alpha = 1.0$ yielded $\Delta R^2 = +0.0209$, demonstrating continued gains from stronger regularization.

Strong Regularization ($\alpha \geq 1.0$): Performance peaked at $\alpha = 2.0$ ($R^2 = 0.8104$) and declined slightly for $\alpha = 1.5$ ($R^2 = 0.8087$), with the difference ($\Delta R^2 = +0.02017$) likely within noise margins of LOOCV estimation. The very close performance between $\alpha = 1.5$ and $\alpha = 2.0$ suggests that the true optimum lies in this range, with potential overfitting risk for $\alpha > 2.0$ not yet evident in the tested range.

The optimal configuration is identified at $\alpha = 2.0$=2.0 , which achieves the highest R^2 and the lowest prediction errors, striking the ideal balance between sparsity and bias.

Chapter 7

Conclusions and Future Work

This thesis has developed **a comprehensive, integrated pipeline for quantitative assessment of pulmonary fibrosis from chest CT images**, addressing critical limitations of visual radiological analysis through automated, reproducible computational methods.

7.1 Conclusions

We have implemented an innovative **dual-mask strategy** that maintains segmentation precision while ensuring topological integrity. The pipeline integrates multiple stages—airway segmentation via deep learning (TotalSegmentator), intelligent gap-filling with adaptive Hounsfield thresholds, carina detection and trachea removal, topological graph construction, and quantitative analysis grounded in the Weibel model.

A **dedicated validation pipeline** was developed to operate without reference ground truth—a realistic constraint in clinical research. Of 45 cases processed, 39 (86.7%) passed technical validation.

The systematic hyperparameter optimization identified that stronger regularization (Lasso with $\alpha = 2.0$) yielded optimal predictive performance while maintaining interpretability—underscoring the importance of balancing empirical performance with clinical usability in machine learning for healthcare.

However, the work explicitly acknowledges **critical limitations**. The validation pipeline does not replace manual radiological review—technical checks identify computational errors, and anatomical reference checks assess plausibility relative to healthy populations, but cannot certify clinical accuracy in diseased lungs.

All results derive from the OSIC cohort, generalization to other fibrotic lung

diseases or different populations is not established. Although FVC measurements are longitudinal, baseline airway metrics were extracted from single time points, establishing **correlation rather than causation**. The final analysis cohort of 22 patients, while maximizing Leave-One-Out Cross-Validation utility, produces wide confidence intervals where random variability may obscure performance differences.

7.2 Future Work

The most significant barriers to advancement are twofold. First, the absence of clinically validated reference segmentations: expert manual annotation of airway trees on 30–50 cases spanning disease severity would enable computation of inter-rater agreement and validation of TotalSegmentator performance. Once reference segmentations exist, preprocessing parameters can be optimized through supervised learning. Second, and perhaps more fundamentally, the lack of a generic ground truth regarding fibrosis itself, there is no definitive clinical or imaging-based reference standard defining what constitutes IPF severity or progression in objective, measurable terms. Current classification relies on visual radiological assessment (UIP pattern vs. probable UIP) and functional measures (FVC decline), but these lack the quantitative, morphological rigor necessary to validate whether the computed airway metrics truly capture disease mechanisms or merely correlate with existing functional surrogates. Resolving this requires close collaboration with pulmonologists and pathologists to establish consensus definitions of fibrosis severity linked to histopathological findings and long-term clinical outcomes.

The current cohort of 45 cases is insufficient for robust machine learning; extended dataset acquisition targeting N=200–300 cases from multiple centers would support algorithm refinement and cross-population validation. Longitudinal CT acquisitions at 6, 12, and 24 months would also enable direct measurement of airway structural changes over time, providing temporal ground truth independent of spirometry and strengthening causal inference.

Methodologically, fine-tuning U-Net or Vision Transformer architectures on IPF-specific manually annotated data should outperform generic whole-body networks. Ensemble approaches combining multiple segmentation models would improve robustness at tissue boundaries, while post-processing refinement via graph-cut or conditional random field optimization could enforce anatomical priors and reduce oversegmentation artifacts.

Prospective clinical validation is essential before deployment: a blinded clinical trial enrolling prospectively recruited IPF patients with baseline CT and quarterly FVC measurements over 2 years would validate whether pipeline-derived metrics predict longitudinal FVC decline and transplant-free survival. Independent thoracic radiologists should assess segmentation quality on stratified pipeline outputs to

identify systematic failure modes. Extension to other fibrotic lung diseases would establish generalizability beyond IPF. Once larger validated datasets become available, temporal modeling via recurrent neural networks (LSTM, GRU) or transformers could capture longitudinal FVC trajectories. Explainability methods (GradCAM, SHAP) could identify which CT regions and metrics most strongly influence predictions, providing actionable clinical insights. Functional imaging integration could enhance predictive capability by linking structural metrics with functional gas exchange.

This work establishes a comprehensive computational framework for quantitative airway assessment in pulmonary fibrosis, demonstrating technical feasibility and clinical relevance. However **it represents a foundation, not a conclusion**. The pathway from research algorithm to clinical utility requires sustained effort across engineering, clinical science, and translational research. The convergence of segmentation accuracy, validation rigor, and clinical correlation positions this work to contribute meaningfully to precision medicine in pulmonary fibrosis. With continued refinement and expanded validation, automated airway quantification could become a standard radiological tool, improving diagnostic accuracy, prognostication, and clinical decision-making for patients with this serious, progressive disease.

Appendix A

Validation Parameters

A.1 Validation Parameters and Ranges

A.1.1 Technical Plausibility Limits

Table A.1: Technical limits for pipeline reliability assessment (disease-agnostic)

Parameter	**Min**	**Max**	**Interpretation of Violation**
Airway volume	5 ml	600 ml	Volume $>$ 600 ml suggests inclusion of non-airway tissues (parenchyma, vessels)
Maximum generation	5	35	Beyond 35 indicates noise-driven over-segmentation or artifacts
PC ratio	0.0	5.0	Values $>$ 5.0 indicate calculation errors or invalid region definitions
Tapering ratio	0.5	1.0	Outside range violates physical plausibility of diameter reduction
Branch count	50	5000	$<$ 50: insufficient segmentation; $>$ 5000: computational artifacts
Tortuosity index	1.0	3.0	Beyond 3.0 likely reflects segmentation artifacts

A.1.2 Anatomical Reference Ranges from Literature

Table A.2: Literature-based reference ranges for anatomical plausibility (healthy populations)

Parameter	Mean	SD	Valid Range	Unit	Source
Segmentation Metrics					
Airway volume	180	50	80–350	ml	Montaudon et al. 2007
Surface area	200	80	80–400	cm^2	CT morphometry studies
Graph/Tree Topology					
Branch count	1500	500	500–3000	–	Weibel 1963
Bifurcation ratio	0.15	0.05	0.05–0.30	–	Horsfield & Cumming 1968
Avg. branch length	12.0	5.0	5.0–25.0	mm	Weibel & Horsfield models
Weibel Model Parameters					
Max generation	18	3	12–23	–	Weibel 1963
Tapering ratio	0.793	0.05	0.70–0.88	–	Weibel ($2^{-1/3}$ ideal)
Trachea diameter	18.0	2.0	14.0–22.0	mm	CT normative data
Gen-5 diameter	3.5	1.0	2.0–6.0	mm	Weibel tables
Morphological Indices					
Tortuosity	1.25	0.15	1.0–1.6	–	CT morphometry studies

Classification logic for anatomical checks:

- **PASS**: Value within valid range
- **WARNING**: Outside valid range but within $\pm 2\sigma$ of mean (potential biological variation or mild pathology)
- **FAIL**: Beyond 2σ from mean (requires manual review to distinguish pathology from segmentation failure)

A.1.3 Metric Definitions and Computational Methods

Table A.3: Definitions of validated metrics

Metric	Definition
Airway volume	Total segmented airway lumen volume (ml).
Surface area	Airway lumen-wall interface area (cm^2).
Branch count	Number of airway segments in centerline graph.
Bifurcation ratio	Ratio of bifurcation nodes to total nodes.
Avg. branch length	Mean Euclidean length of airway segments (mm).
Maximum generation	Maximum tree depth from trachea (gen 0).
Tapering ratio	Child/parent diameter ratio at bifurcations (ideal: $2^{-1/3} \approx 0.793$).
Trachea diameter	Hydraulic diameter of generation-0 airway (mm).
Gen-5 diameter	Mean hydraulic diameter of generation-5 airways (mm).
PC ratio	Peripheral/central volume ratio (gen > 5 / gen ≤ 5).
Tortuosity	Centerline path length / straight-line distance (1.0 = straight).

A.2 Selected Metrics for Clinical Correlation

Table A.4: Airway and parenchymal metrics selected for FVC correlation analysis

Metric	Category	Unit	Biological Rationale
Core Airway Metrics			
Airway volume	Airway	ml	Total conducting airway capacity; reduced in advanced disease due to peripheral loss
Mean tortuosity	Airway	–	Airway distortion index; elevated in fibrotic remodeling and traction bronchiectasis
Peripheral Airway Metrics			
Std peripheral diameter	Peripheral	mm	Heterogeneity of peripheral airway caliber; increased in patchy fibrosis
Central/peripheral diameter ratio	Peripheral	–	Gradient of airway tapering; disrupted in peripheral-predominant disease
Mean peripheral branch volume	Peripheral	mm^3	Average volume of small airways (gen > 5); directly reflects peripheral loss
Parenchymal Metrics			
Mean lung density	Parenchymal	HU	Average CT attenuation; increases with fibrotic consolidation
Histogram entropy	Parenchymal	–	Texture heterogeneity linked to reticulation and honeycombing

A.3 Temporal Interpolation Windows

Table A.5: Hierarchical windowing strategy for FVC% interpolation

Target	Window Type	Time Range	Strategy
Week 0 (Baseline)	Preferred	[-5, 10] weeks	Use nearest measurement
	Extended	[10, 30] weeks	Linear regression ($\geq$ 2 points)
	Last resort	All data if max week ≤ 40	Regression on all available points
Week 52 (1-year follow-up)	Preferred	[40, 65] weeks	Use nearest measurement
	Regression	All data if max week ≥ 20 OR ≥ 3 points	Linear regression with extrapolation

Rationale: The week 0 strategy prioritizes temporal proximity to enrollment, as baseline FVC% should reflect function at the time of CT acquisition. The week 52 strategy accommodates greater extrapolation, recognizing that many patients lack precise one-year measurements but have sufficient early-to-mid follow-up data to estimate trajectory.

A.4 Performance Metrics

Table A.6: Metrics for evaluating single-feature prediction performance

Metric	Definition
MAE	Mean absolute error: $\frac{1}{N}\sum_{i=1}^{N}\lvert\hat{y}_i - y_i\rvert$ (percentage points). Thresholds: $< 5\%$ (excellent), $< 10\%$ (good), $> 15\%$ (poor).
RMSE	Root mean squared error: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$ (percentage points). Penalizes large errors.
$\mathbf{R}^2$	Proportion of FVC% variance explained: $1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2}$. Thresholds: > 0.5 (strong), 0.3–0.5 (moderate), < 0.3 (weak).
Pearson r	Linear correlation between predictions and actual values. Range: -1 to 1.
Spearman ρ	Rank-based correlation, robust to outliers.

A.5 Hyperparameter Tuning for Prediction Models

To ensure robust model evaluation and prevent overfitting given the limited sample size, a systematic grid search was conducted over a predefined set of hyperparameters. Table A.7 details the ranges explored for each model. All 108 unique combinations were evaluated using the same LOOCV framework described in the main text.

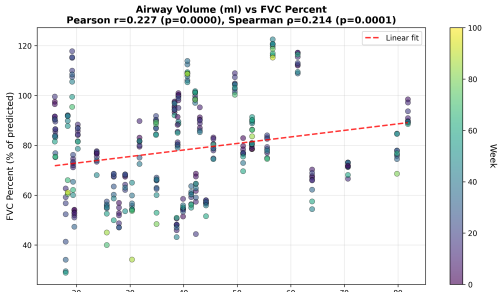
Table A.7: Hyperparameter search space for machine learning models

Model	Hyperparameter	Values Tested
MLP	Architecture	16-8, 20-10
	Dropout rate	0.2, 0.3
	Learning rate	1×10^{-3}, 5×10^{-4}
	Weight decay	1×10^{-4}, 1×10^{-3}
Ridge	Penalty (α)	1.0, 2.0, 5.0
Lasso	Penalty (α)	0.2, 0.5
Random Forest	Number of trees	100 (fixed)
	Max depth	2, 3, 4
	Min. samples split/leaf	5 / 2 (fixed)
Ensemble	Ridge / RF weight	0.6/0.4, 0.7/0.3

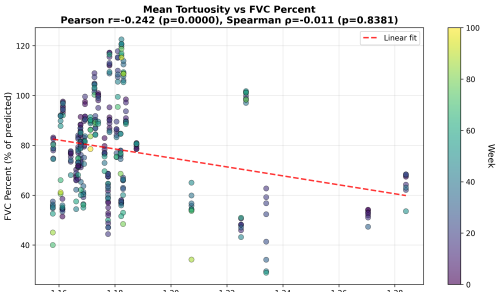
Appendix B

Validation Results

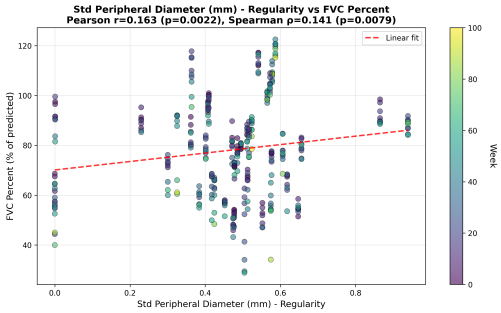
B.1 Supplementary Correlation Plots



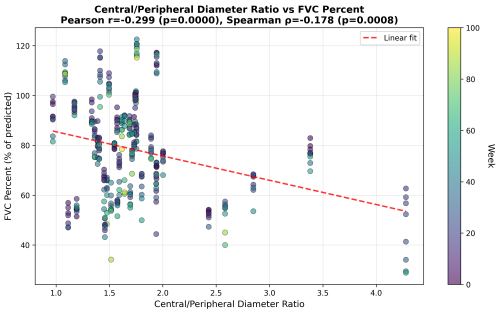
(a) Airway Volume



(b) Mean Tortuosity



(c) Std Peripheral Diameter



(d) Central/Peripheral Ratio

Figure B.1: Scatter plots for additional airway metrics vs. FVC%, with week as the color scale. The red dashed line represents a linear fit.

B.2 Supplementary Table of Feature-Wise Prediction Performance

Feature	Target	N	R^2	MAE	Pearson r (p)
`mean_peripheral_branch_volume_mm3`	FVC week 0	32	-0.157	16.00	-0.35 (0.047)
	FVC week 52	32	-0.109	16.51	-0.10 (0.599)
	Decline (traditional)	32	-0.080	8.48	-0.17 (0.353)
	Annual decline	39	-0.008	6.70	0.11 (0.489)
`periphery_branching_density`	FVC week 0	32	-0.017	15.08	0.13 (0.476)
	FVC week 52	32	0.002	15.18	0.17 (0.339)
	Decline (traditional)	32	-0.113	8.65	-0.48 (0.006)
	Annual decline	39	-0.058	7.06	-0.05 (0.748)
`peripheral_mean_diameter_mm`	FVC week 0	32	-0.121	15.74	-0.19 (0.302)
	FVC week 52	32	-0.120	16.41	-0.03 (0.862)
	Decline (traditional)	32	-0.143	8.98	-0.25 (0.171)
	Annual decline	39	-0.062	7.05	0.04 (0.818)
`central_to_peripheral_diameter_ratio`	FVC week 0	32	-0.141	16.45	-0.38 (0.031)
	FVC week 52	32	-0.039	16.48	0.09 (0.624)
	Decline (traditional)	32	-0.384	9.64	-0.20 (0.260)
	Annual decline	39	-0.117	7.37	0.04 (0.824)
`mean_lung_density_HU`	FVC week 0	32	0.317	12.12	0.57 (0.001)
	FVC week 52	32	0.315	11.74	0.59 (0.0004)
	Decline (traditional)	32	-0.139	8.76	-0.26 (0.151)
	Annual decline	39	-0.050	6.84	0.12 (0.481)
`histogram_entropy`	FVC week 0	32	0.376	11.60	0.62 (0.0002)
	FVC week 52	32	0.408	13.35	0.64 (7.5e-5)
	Decline (traditional)	32	-0.132	8.72	-0.49 (0.004)
	Annual decline	39	-0.018	6.86	0.13 (0.426)

Table B.1: Prediction performance of baseline imaging metrics for FVC targets. Results are obtained via leave-one-out cross-validation (LOOCV) with linear regression. FVC week 52 targets are highlighted in light yellow. MAE is expressed in % predicted for week 0 and week 52, and as %/year for annual decline metrics. All models are univariate linear regressions.

B.3 Supplementary Prediction Plots (Additional Features)

This section reports additional feature-specific prediction plots from `images/cap6/plots/`, presented separately for easier interpretation.

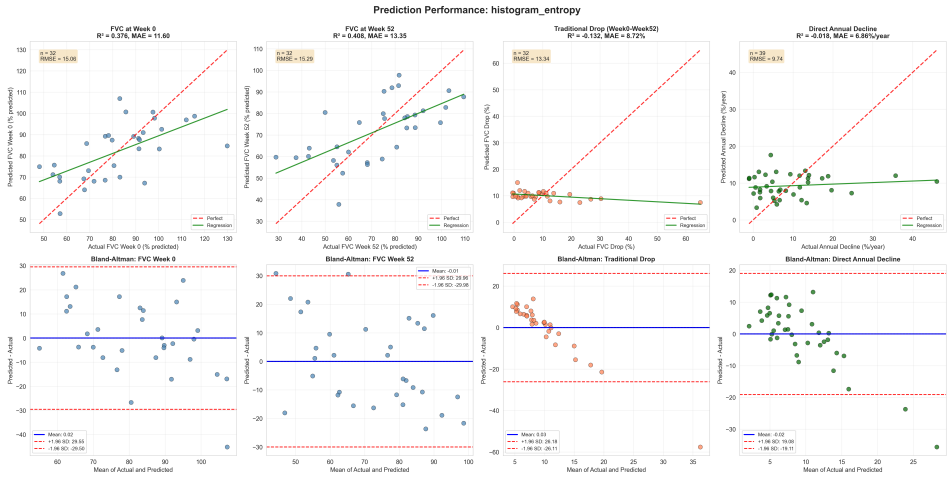


Figure B.2: Prediction plots for `histogram_entropy`. This parenchymal feature generally shows clearer structure in cross-sectional targets than in decline targets, where dispersion remains high.

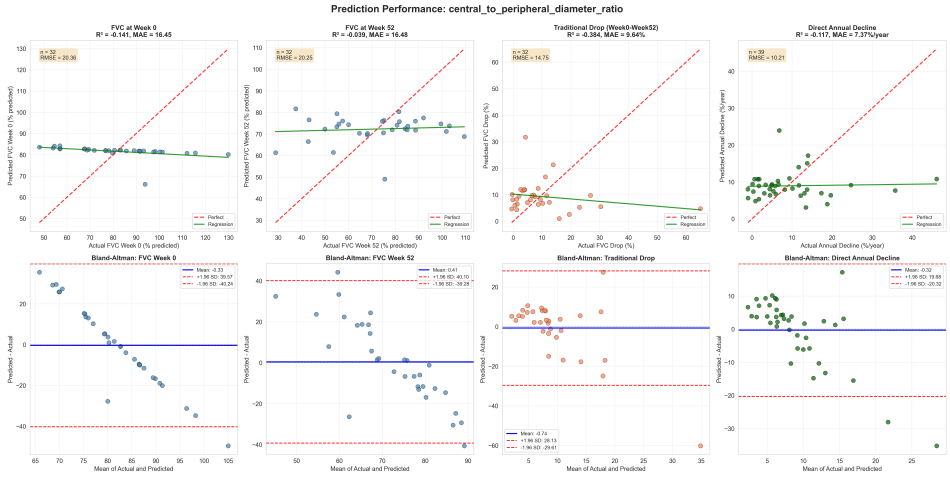


Figure B.3: Prediction plots for `central_to_peripheral_diameter_ratio`. The weak trend and broad Bland–Altman limits indicate limited predictive value for both status and decline outcomes.

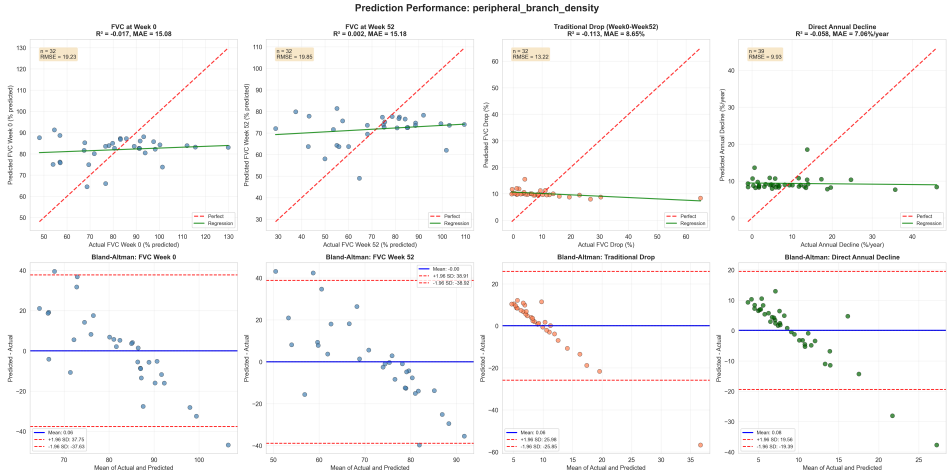


Figure B.4: Prediction plots for `periphery_branching_density`. Predictions remain widely scattered, suggesting that this airway feature alone does not robustly explain FVC variability.

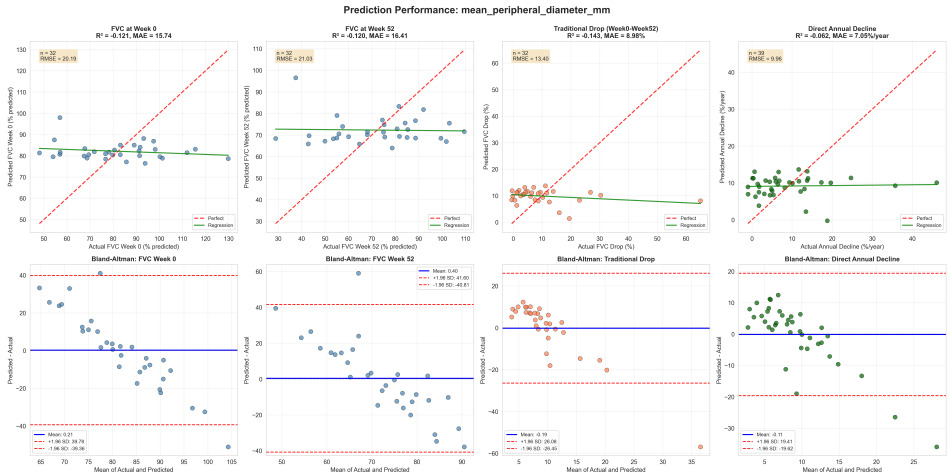


Figure B.5: Prediction plots for `peripheral_mean_diameter_mm`. Error spread is substantial across all tasks, confirming weak standalone performance for clinical prediction.

Appendix C

FVC% Week 52 Model Benchmark

C.1 Comprehensive Model Results Across Datasets

Dataset	Model	N	R^2	MAE	RMSE
all	Ensemble (Ridge+RF)	34	0.4162	11.72	14.90
	Ridge (L2 reg, $\alpha = 5.0$)	34	0.3631	12.11	15.57
	Random Forest	34	0.4017	12.24	15.09
	XGBoost	34	0.2772	13.98	16.58
	Lasso (L1 reg, $\alpha = 0.5$)	34	0.3206	12.46	16.08
	LR (multi-feature)	34	0.2851	13.12	16.49
strict	Ensemble (Ridge+RF)	22	0.3017	12.12	14.82
	Ridge (L2 reg, $\alpha = 5.0$)	22	0.1976	12.59	15.89
	Random Forest	22	0.3568	11.68	14.23
	XGBoost	22	0.3072	12.15	14.77
	Lasso (L1 reg, $\alpha = 0.5$)	22	0.0567	13.71	17.23
balanced	Ensemble (Ridge+RF)	32	0.4075	11.94	15.29
	Ridge (L2 reg, $\alpha = 5.0$)	32	0.3591	12.38	15.91
	Random Forest	32	0.3947	12.23	15.46
	XGBoost	32	0.2617	14.56	17.07
	Lasso (L1 reg, $\alpha = 0.5$)	32	0.3118	12.77	16.48
	LR (multi-feature)	32	0.2712	13.59	16.96

Table C.1: Performance comparison of models with non-negative performance across all datasets.

Bibliography

[1] J. P. Charbonnier, M. Brink, F. Ciompi, E. T. Scholten, C. M. Schaefer-Prokop, and E. M. van Rikxoort. «Automatic Separation and Classification of Pulmonary Arteries and Veins in CT Scans». In: *IEEE Trans. Med. Imaging* 35.3 (2016), pp. 882–892 (cit. on pp. 1, 14).

[2] Kaggle. *OSIC Pulmonary Fibrosis Progression Dataset*. Kaggle Competition Dataset. 2020 (cit. on pp. 1, 14).

[3] Kum Ju Chae, Hye Jeon Hwang, Rosane Duarte Achcar, Joseph C. Cooley, Stephen M. Humphries, Seth Kligerman, and David A. Lynch. «Central Role of CT in Management of Pulmonary Fibrosis». In: *RadioGraphics* 44.6 (2024). DOI: 10.1148/rg.230165. URL: https://doi.org/10.1148/rg.230165 (cit. on p. 3).

[4] Jakob Wasserthal et al. «TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images». In: *Radiology: Artificial Intelligence* 5.5 (2023), e230024. DOI: 10.1148/ryai.230024. eprint: https://doi.org/10.1148/ryai.230024. URL: https://doi.org/10.1148/ryai.230024 (cit. on pp. 5, 20, 22).

[5] Simon L F Walsh, Lucio Calandriello, Nicola Sverzellati, Athol U Wells, and David M Hansell. «Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT». In: *Thorax* 71.1 (2016). Ed. by on behalf of The UIP Observer Consort et al., pp. 45–51. ISSN: 0040-6376. DOI: 10.1136/thoraxjnl-2015-207252. eprint: https://thorax.bmj.com/content/71/1/45.full.pdf. URL: https://thorax.bmj.com/content/71/1/45 (cit. on p. 7).

[6] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, et al. «TotalSegmentator: Reliable Segmentation of Body Parts in CT». In: *Radiology: Artificial Intelligence* 5.5 (2023), e230024 (cit. on p. 7).

[7] Ana García-Uceda, Raghavendra Selvan, Zaigham Saghir, Harm AWM Tiddens, and Marleen de Bruijne. «Automatic Airway Segmentation from Computed Tomography using Robust and Efficient 3-D Convolutional Neural Networks». In: *Scientific Reports* 11.1 (2021), p. 16001. DOI: 10.1038/s41598-021-95364-1 (cit. on p. 8).

[8] Wing Keung Cheung et al. «Interpolation-split: a data-centric deep learning approach with big interpolated data to boost airway segmentation performance». In: *Journal of Big Data* 11.1 (Aug. 2024), p. 104. ISSN: 2196-1115. DOI: 10.1186/s40537-024-00974-x. URL: https://doi.org/10.1186/s40537-024-00974-x (cit. on p. 8).

[9] Minghui Zhang et al. *AirMorph: Topology-Preserving Deep Learning for Pulmonary Airway Analysis*. 2025. arXiv: 2412.11039 [eess.IV]. URL: https://arxiv.org/abs/2412.11039 (cit. on p. 8).

[10] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. «TotalSegmentator: Robust Segmentation of 104 Anatomical Structures in CT Images». In: *Radiology: Artificial Intelligence* 5.5 (2023), e230024. DOI: 10.1148/ryai.230024 (cit. on p. 9).

[11] Pechin Lo et al. «Extraction of Airways From CT (EXACT'09)». In: *IEEE Transactions on Medical Imaging* 31.11 (2012), pp. 2093–2107. DOI: 10.1109/TMI.2012.2209674 (cit. on p. 9).

[12] Ewald R. Weibel. *Morphometry of the Human Lung*. Berlin, Heidelberg: Springer, 1963 (cit. on pp. 9, 10, 30).

[13] Amir Pakzad, Martin J Willemink, Hamid Dehghani, Vivek S Iyer, John Pickering, Margaret L Salisbury, David A Lynch, Stephen M Humphries, and Samir Peters. «Evaluation of Automated Airway Morphological Quantification for Assessing Fibrosing Lung Disease». In: *Scientific Reports* 11.1 (2021), p. 23937. DOI: 10.1038/s41598-021-03419-0 (cit. on pp. 10, 11).

[14] Tomoaki Sasaki, Koji Takahashi, Nobuhisa Takada, and Yoshinobu Ohsaki. «Ratios of Peripheral-to-Central Airway Lumen Area and Percentage Wall Area as Predictors of Severity of Chronic Obstructive Pulmonary Disease». In: *American Journal of Roentgenology* 203.1 (2014), pp. 78–84. DOI: 10.2214/AJR.13.11748 (cit. on p. 10).

[15] Miranda Kirby et al. «Total Airway Count on Computed Tomography and the Risk of Chronic Obstructive Pulmonary Disease Progression: Findings from a Population-based Study». In: *American Journal of Respiratory and Critical Care Medicine* 197.1 (2018), pp. 56–65. DOI: 10.1164/rccm.201704-0692OC (cit. on p. 10).

[16] Kohei Ikezoe et al. «Small Airway Reduction and Fibrosis Is an Early Pathologic Feature of Idiopathic Pulmonary Fibrosis». In: *American Journal of Respiratory and Critical Care Medicine* 204.9 (2021), pp. 1048–1059. DOI: 10.1164/rccm.202103-0585OC (cit. on p. 11).

[17] Xiaoyan Wang, Ling Zhao, Dingyun Song, Xinran Zhang, Min Liu, and Huaping Dai. «Small airway lesions appear with the course of IPF and relate to the severity of pulmonary fibrosis progression». In: *BMC Pulmonary Medicine* 25 (2025), p. 465. DOI: 10.1186/s12890-025-03939-9 (cit. on p. 11).

[18] Muhunthan Thillai et al. «Deep Learning-based Segmentation of Computed Tomography Scans Predicts Disease Progression and Mortality in Idiopathic Pulmonary Fibrosis». In: *American Journal of Respiratory and Critical Care Medicine* 210.4 (2024), pp. 465–472. DOI: 10.1164/rccm.202311-2185OC (cit. on p. 11).

[19] Ganesh Raghu, Martine Remy-Jardin, Jeffrey L. Myers, et al. «Clinical Guidelines for Diagnosing Idiopathic Pulmonary Fibrosis». In: *American Journal of Respiratory and Critical Care Medicine* 198.5 (2018), e44–e68 (cit. on p. 11).

[20] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009 (cit. on p. 12).

[21] Leo Breiman. «Random Forests». In: *Machine Learning* 45.1 (2001), pp. 5–32 (cit. on p. 12).

[22] Annette M Molinaro, Richard Simon, and Ruth M Pfeiffer. «Comparing Methods for Estimating Prediction Errors». In: *Bioinformatics* 21.15 (2005), pp. 3301–3307 (cit. on p. 12).

[23] Sudhir Varma and Richard Simon. «Bias in Cross-Validation Error Estimation». In: *BMC Bioinformatics* 7 (2006), p. 91 (cit. on p. 12).

[24] Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott L Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 2002 (cit. on p. 13).

[25] Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied Longitudinal Analysis*. Wiley, 2011 (cit. on p. 13).

[26] Megan L. Neely et al. «Lung function trajectories in patients with idiopathic pulmonary fibrosis». In: *Respiratory Research* 24.1 (2023), p. 209. DOI: 10.1186/s12931-023-02503-5 (cit. on p. 13).

[27] Akira Kurozawa, Kazuya Tatsumi, and Koichi Nakabe. «Effect of Gas Oscillation-Induced Irreversible Flow in Transitional Bronchioles of Human Lung». In: *Journal of Flow Control, Measurement & Visualization* 4.4 (2016). Licensed under CC BY 4.0, pp. 171–193. DOI: 10.4236/jfcmv.2016.44015 (cit. on p. 31).