Politecnico di Torino

Politecnico di Torino

Computer Engineering

A.a. 2025/2026

Graduation Session March 2026

Imaging-Based Progression Prediction in Interstitial Lung Disease

Supervisors:

Tania Melo

Giuseppe Averta

Candidate:

Francesco Passiatore

Abstract

Idiopathic pulmonary fibrosis (IPF) is a chronic and progressive interstitial lung disease characterized by a highly heterogeneous clinical trajectory, which complicates prognostic assessment and therapeutic planning. In this thesis, we investigate a multimodal machine learning framework for predicting short-term functional progression in IPF, defined as a $\geq 10\%$ decline in forced vital capacity (FVC) at one-year follow-up.

Baseline chest computed tomography (CT) scans and structured patient data are integrated to assess whether imaging-derived representations provide additional prognostic value beyond clinical variables. From CT scans, convolutional neural networks (CNNs) are used as feature extractors to obtain high-dimensional slice-level embeddings, which are aggregated at the patient level through different pooling strategies. These imaging features are combined with handcrafted quantitative descriptors and demographic variables within a simplified neural architecture designed to reduce overfitting in small datasets.

A structured ablation study is conducted to compare clinical-only, imaging-only, and multimodal configurations. Model performance is evaluated using patient-wise cross-validation, with fold-wise statistical comparisons to assess the robustness of observed differences. In parallel, gradient boosting models are trained on pooled CNN embeddings and clinical features to examine whether deep imaging representations contribute complementary predictive signal.

Results show that handcrafted clinical features provide stable predictive performance, while CNN-based imaging representations exhibit higher variance across folds. The addition of imaging features to clinical variables yields limited and statistically non-significant improvements, suggesting that, in small cohorts, deep learning-based imaging embeddings may not consistently outperform structured clinical predictors.

Although constrained by sample size, this work provides a systematic evaluation of multimodal learning strategies for IPF progression prediction and highlights methodological considerations for applying deep learning in small medical datasets.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Interstitial lung diseases (ILDs) comprise a heterogeneous group of diffuse parenchymal lung disorders characterized by varying degrees of inflammation and fibrosis affecting the lung interstitium. These conditions often lead to impaired gas exchange, progressive respiratory symptoms, and structural abnormalities visible on high-resolution computed tomography (HRCT). Among the different ILDs, a subset of fibrosing diseases may develop a progressive phenotype characterized by worsening fibrosis, declining lung function, and increased mortality.

Idiopathic pulmonary fibrosis (IPF) represents the prototypical chronic progressive fibrosing interstitial lung disease. It is characterized by irreversible fibrotic remodeling of the lung parenchyma, leading to a gradual decline in respiratory function. Despite advances in diagnosis and therapy, IPF remains incurable and is associated with substantial morbidity and mortality. The disease exhibits considerable heterogeneity: while some patients experience a relatively slow decline, others worsen rapidly or develop acute exacerbations. This variability complicates prognosis, treatment planning, follow-up, and patient counseling [1].

High-resolution computed tomography (HRCT) is central to the diagnostic process, offering detailed images of structural abnormalities in the lungs. Current diagnostic pathways and definitions are outlined in the official ATS/ERS/JRS/ALAT guidelines for IPF diagnosis [2]. However, imaging alone is insufficient for accurately predicting the rate of functional deterioration. Consequently, clinicians often make decisions under uncertainty, balancing the early initiation or escalation of therapy with the potential for adverse effects and determining which patients require closer monitoring or referral.

Predictive modeling has proven to be valuable in medicine for stratifying patients based on risk and expected disease course. In IPF, a reliable prognostic model could facilitate early identification of high-risk patients, guide personalized care, and improve the design and interpretation of clinical trials by providing better estimates of disease trajectories.

This thesis investigates the use of machine learning to predict the evolution of lung function in patients with IPF, combining chest CT imaging with clinical and demographic information. The primary endpoint in this study is forced vital capacity (FVC), a key spirometric measure commonly used in IPF research to evaluate lung function and disease progression [3].

To address the task of predicting lung function decline, we propose a comprehensive machine learning pipeline that integrates multiple data sources. The core components of this pipeline are CT images, clinical data, and machine learning models, which together aim to provide a predictive model for disease progression in IPF.

The first step in the pipeline involves the processing of CT scans, the primary imaging modality for diagnosing IPF. These scans provide high-resolution images that allow clinicians to identify areas of fibrosis and other structural changes in the lungs. The pipeline extracts relevant features from specific CT slices, focusing on those that capture the most pertinent lung tissue characteristics.

In addition to these handcrafted features, such as lung volume estimates and average tissue density, which offer valuable insights into the degree of fibrosis, we utilize Convolutional Neural Networks (CNNs) to automatically learn complex patterns from raw image data. Unlike handcrafted features, CNNs can detect intricate, hierarchical features in the lung tissue, enabling the model to learn sophisticated representations that may not be easily captured by human-defined metrics. Pretrained CNN models, trained on large datasets, are employed to extract relevant features from the CT scans.

Clinical data, including patient-specific information such as age, sex, and smoking status, are integrated into the model to provide additional context. These variables play an important role in influencing disease progression, with factors like age and smoking history having significant effects on lung function and disease trajectory.

The ultimate goal of this study is to predict the future trajectory of FVC. By combining CT-derived features with clinical data, the model aims to estimate the decline in lung function over time. Such predictions could help clinicians identify high-risk patients at earlier stages, enabling more personalized care and improved treatment planning. To ensure the robustness of the model, cross-validation is used to assess its performance, reducing the risk of overfitting by training and testing the model on different subsets of the data.

While this study offers a promising approach, it has several limitations. The dataset includes a limited number of patients, and CT scans are available only at a single time point, lacking longitudinal imaging that could provide deeper insights into individual disease trajectories. Despite these constraints, combining CT-derived features with clinical data and machine learning techniques holds great potential for improving the prediction of IPF progression.

In summary, this thesis explores the relationship between CT-derived representations, clinical variables, and the progression of FVC in IPF. The proposed models aim to serve as research tools to support future work on multimodal predictors of disease progression and are not intended to replace clinical judgment.

Chapter 2

Background and State of the Art

2.1 Clinical Overview and Related Work

2.1.1 Interstitial Lung Diseases (ILDs)

Interstitial lung diseases (ILDs) comprise a broad and heterogeneous group of diffuse parenchymal lung disorders characterized by varying degrees of inflammation and/or fibrosis involving the interstitium, alveolar spaces, peripheral airways, and pulmonary vasculature.

This umbrella category includes ILDs of known cause (e.g., connective tissue disease-associated ILD, hypersensitivity pneumonitis, and occupational pneumoconioses) as well as idiopathic forms, among which idiopathic pulmonary fibrosis (IPF) is the most extensively studied chronic fibrosing phenotype.

Despite this etiological heterogeneity, many ILDs share common clinical and radiological manifestations, including exertional dyspnea, impaired gas exchange, restrictive ventilatory defects, and diffuse abnormalities on high-resolution computed tomography (HRCT). In fibrosing ILDs, disease behavior is particularly relevant, as a subset of patients develops a progressive fibrosing phenotype with worsening symptoms, increasing fibrotic abnormalities on imaging, and declining lung function over time [4].

Within this broader ILD spectrum, IPF represents the prototypical chronic progressive fibrosing interstitial pneumonia and constitutes the primary disease focus of this thesis.

2.1.2 Idiopathic Pulmonary Fibrosis (IPF)

Idiopathic pulmonary fibrosis (IPF) is a chronic, progressive fibrosing interstitial lung disease of unknown cause, characterized by irreversible scarring of the lung parenchyma and a steadily declining respiratory function [5, 1]. It predominantly affects older adults and is associated with poor prognosis [5].

Current evidence supports a pathobiological model in which repeated micro-injuries to the alveolar epithelium in genetically susceptible individuals lead to abnormal wound repair, excessive fibroblast activation, myofibroblast accumulation, and extracellular matrix deposition, ultimately causing architectural distortion of the lung [5]. Recognized risk factors include cigarette smoking and occupational/environmental inhalational exposures, although no single causative agent has been identified [5].

Clinically, patients typically present with progressive exertional dyspnea and chronic dry cough; physical examination often reveals bibasilar inspiratory "Velcro" crackles, and digital clubbing may be present [5]. Disease progression is highly variable: some patients experience slow functional decline, whereas others deteriorate rapidly or develop acute exacerbations with substantial impact on outcomes [1, 5].

Although IPF remains incurable, antifibrotic therapies (e.g., pirfenidone and nintedanib) can slow the rate of functional decline, while supportive care (pulmonary rehabilitation, oxygen therapy) and lung transplantation remain important components of patient management [6, 7, 5].

Diagnosis and HRCT Findings

The diagnosis of IPF requires the exclusion of other known causes of interstitial lung disease and relies on the integration of clinical data, high-resolution computed tomography (HRCT), and, when required, histopathology, ideally within a multidisciplinary discussion [2]. The 2018 ATS/ERS/JRS/ALAT guideline standardizes HRCT findings into four diagnostic categories (UIP, probable UIP, indeterminate for UIP, and alternative diagnosis), which guide the need for additional investigations such as bronchoalveolar lavage or lung biopsy and help reduce interobserver and inter-centre variability [2].

Usual interstitial pneumonia (UIP) refers to a specific histopathological and radiological pattern characterized by spatial and temporal heterogeneity of fibrosis. On HRCT, UIP is defined by a basal and subpleural predominance of reticular abnormalities, traction bronchiectasis, and honeycombing, with relative preservation of lung regions in earlier disease stages. Reticular abnormalities reflect a network of fine linear opacities caused by interstitial fibrosis, while traction bronchiectasis refers to irreversible bronchial dilatation secondary to fibrotic distortion of the surrounding lung parenchyma. Honeycombing consists of clustered, thick-walled

cystic air spaces, typically located in subpleural regions, and represents established end-stage fibrosis. Although UIP can be observed in other clinical contexts, including connective tissue disease–associated interstitial lung disease and chronic hypersensitivity pneumonitis, its presence in the appropriate clinical setting is highly suggestive of IPF. Figure 2.1 provides representative examples of these HRCT patterns, illustrating the spectrum of UIP, probable UIP, and alternative fibrotic interstitial lung disease appearances discussed above.

However, substantial overlap in imaging appearances with other fibrotic interstitial lung diseases may persist, leading to diagnostic uncertainty and potential misclassification despite guideline-based assessment [2, 5].

Prognosis and Functional Decline

The natural history of IPF is heterogeneous: patients may experience a gradual decline in lung function, periods of relative stability, or episodes of accelerated deterioration, including acute exacerbations [1]. Pulmonary function tests are central for monitoring disease evolution, and forced vital capacity (FVC) is widely used as a functional endpoint in both clinical practice and clinical trials [1]. Prognosis can be summarized using simple multivariable tools such as the Gender–Age–Physiology (GAP) index, which combines sex, age, and physiological impairment (FVC and DLCO) to stratify mortality risk [8].

Although current antifibrotic therapies can slow functional decline, they do not represent a cure, and guideline recommendations continue to evolve [6, 7]. Importantly, commonly used clinical prognostic tools do not explicitly incorporate imaging phenotypes, despite the marked heterogeneity observed on HRCT. This limitation motivates multimodal approaches that integrate HRCT-derived

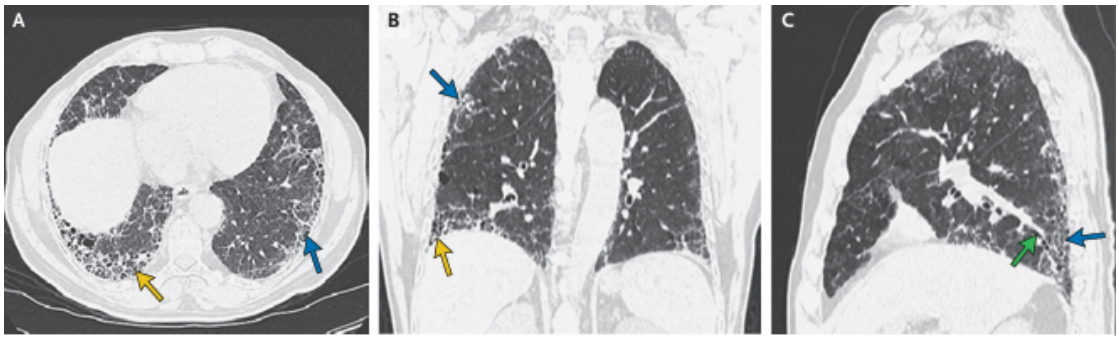

Figure 2.1: Representative high-resolution CT (HRCT) images illustrating common imaging patterns encountered in fibrotic interstitial lung diseases. Images are shown in axial, coronal, and sagittal reconstructions. Adapted from Lederer and Martinez.

information with clinical data to support more individualized risk stratification.

2.1.3 Forced Vital Capacity (FVC)

Forced vital capacity (FVC) is the maximum volume of air that can be forcibly exhaled after a full inspiration. In clinical practice, it is measured by spirometry and is commonly reported both as an absolute value (litres) and as a percentage of the predicted value for age, sex, height, and ethnicity. In IPF, percent-predicted FVC is widely used as a standardized measure of physiological impairment that enables comparison across patients and cohorts [3].

FVC represents a key functional endpoint in IPF because restrictive ventilatory impairment reflects the progressive loss of functional lung volume caused by fibrotic remodeling. It is also well suited for longitudinal monitoring, as it is reproducible and responsive to clinical change; validation studies have shown that even relatively small changes in FVC may be associated with clinically meaningful differences in outcome [3].

FVC is commonly used in conjunction with other clinical parameters, such as diffusing capacity for carbon monoxide (DLCO), to assess the overall pulmonary function and disease progression in IPF. However, FVC measurements can be influenced by factors such as patient effort, which may lead to variability, especially in longitudinal studies or clinical trials. Hence, careful standardization and consistent follow-up are crucial to ensure reliable data.

Given its responsiveness to clinical changes, FVC serves not only as a tool for monitoring but also as a key prognostic marker in IPF, where declines over time are linked to worsened outcomes.

FVC Decline as a Prognostic Marker

Changes in FVC over time are consistently associated with prognosis in IPF and are commonly used to define disease progression in clinical trials and observational studies. A decline of $\geq 10\%$ in percent-predicted FVC is widely accepted as a robust threshold for progression and has been linked to increased mortality risk [9]. Methodological analyses suggest that using a relative versus absolute change in FVC to define a $\geq 10\%$ decline can alter the classification of progressors, though both approaches maintain similar prognostic accuracy [10].

Importantly, even “marginal” declines (e.g., 5–10% over six months) have been associated with worse outcomes, indicating that smaller changes may still carry meaningful prognostic information [3, 11]. This highlights the importance of monitoring even subtle declines in FVC, particularly when evaluating individual patient trajectories or when making treatment decisions.

Imaging- and Data-driven Prediction of FVC and Disease Progression

A long-standing motivation for integrating imaging into prognostic modeling is that high-resolution computed tomography (HRCT) provides detailed information about the spatial extent and pattern of fibrosis that is not fully captured by spirometry. Early work in untreated IPF reported moderate correlations between the extent of disease on HRCT and both forced vital capacity (FVC) and diffusing capacity for carbon monoxide (DLCO) at diagnosis. Additionally, longitudinal changes in HRCT extent were associated with changes in both FVC and DLCO, further establishing the role of imaging in disease monitoring [12].

More recently, deep learning has been explored as a way to directly infer pulmonary function from CT scans. For example, the BeyondCT model predicts spirometric indices, including FVC, from chest CT, suggesting that imaging-derived representations might serve as surrogates for functional measurements when spirometry is unavailable or unreliable [13]. This development points toward an emerging paradigm where imaging-derived biomarkers could augment traditional clinical parameters, offering an alternative when functional data from spirometry is missing, noisy, or inconsistent.

Within IPF specifically, modeling the temporal evolution of FVC has also been approached with data-driven methods. Using longitudinal percent-predicted FVC (ppFVC) data from the PROFILE cohort, Fainberg et al. combined imputation of non-random missingness with unsupervised clustering (self-organizing maps) and identified four distinct FVC trajectory clusters associated with different baseline features and survival. This work highlighted the heterogeneity of progression beyond simple threshold definitions and emphasized the need for more refined models to track disease evolution [14].

Importantly, systematic evaluations in other chronic respiratory diseases have identified key methodological pitfalls when developing machine learning (ML) models for prognosis, including limited external validation and variable reporting quality. These challenges are critical when considering the clinical adoption of multimodal models, as robustness and reproducibility across different patient populations must be rigorously validated [15]. Transparent reporting and comprehensive external validation remain necessary steps before these models can be used reliably in clinical settings.

In parallel, several quantitative and deep learning approaches have demonstrated that automated imaging biomarkers extracted from baseline CT can predict clinically relevant outcomes in IPF and related fibrosing interstitial lung diseases (ILDs). One prominent quantitative CT framework, CALIPER, showed that vessel-related structures (VRS) were strong predictors of mortality and lung function decline, supporting the idea that imaging-derived features provide prognostic information

beyond standard physiological assessments [16]. More recently, deep learning-based segmentation techniques have been used to quantify lung volumes and fibrosis-related regions, generating imaging biomarkers that correlate with FVC and predict progression and mortality in IPF cohorts, including the PROFILE study [17]. Complementary work has explored deep learning classifiers for recognizing usual interstitial pneumonia (UIP) patterns on CT, reporting associations with outcomes such as mortality risk and annual FVC decline [18].

Since many pipelines ultimately convert CT volumes into compact representations (e.g., feature embeddings or quantitative summary measures) and then combine them with clinical covariates to create predictive models, studies that explicitly evaluate added value beyond visual assessment are particularly relevant. For instance, ML-derived measures of fibrosis extent have been shown to predict transplant-free survival and FVC decline in large fibrosing ILD registries, even after accounting for CT pattern categories [4]. Furthermore, multimodal models that combine automated CT staging with clinical function have also been proposed to predict IPF mortality, aligning with the general idea of integrating imaging and clinical data for improved risk stratification [19].

While these advancements illustrate the promising role of multimodal models, the challenge remains to translate these methods into robust, clinically applicable tools. This thesis aims to extend these efforts by focusing on a task-driven formulation, integrating baseline chest CT data with structured clinical variables to predict subsequent FVC decline in IPF. By bridging the gap between image-derived biomarkers and clinical measurements, this work aims to enhance individualized disease management and improve patient outcomes.

2.2 Technical Overview

Deep Learning: Definition and Motivation in the IPF Context

Deep learning (DL) is a family of machine learning methods that learn directly from data to recognize patterns useful for a given task (e.g., predicting a clinical outcome), without the need for these patterns to be manually defined. The term "deep" indicates that the model is composed of multiple layers: each layer transforms the input into a progressively more abstract representation, allowing it to move from local information (textures, edges, opacity) to more global concepts (e.g., extent of fibrosis, architectural distortions, patterns compatible with UIP).

Deep learning was chosen because the primary input consists of CT images, which are high-dimensional data with complex spatial relationships. In this scenario, the DL approach allows us to automatically extract relevant features (in the form

of feature embeddings) directly from the scans, reducing the reliance on manual feature engineering. These embeddings can then be combined with clinical variables (e.g., age, sex) and quantitative measures (e.g., lung volume) in a final predictor (e.g., an MLP), following a multimodal approach.

Deep learning is particularly suited for medical imaging because clinical images often present: (i) high variability between patients and across centers (different scanners, different protocols), (ii) pathological signals that are often diffuse and subtle compared to noise and artifacts, and (iii) significant morphological heterogeneity even with the same diagnosis. When properly trained and validated, DL models can learn robust representations that capture these variations, providing valuable insights for predictive tasks.

Compared to traditional methods (radiomics analysis based on predefined features, visual scoring, or statistical models on a few aggregated measures), DL differs mainly in its ability to automatically learn features from data, without imposing a priori which descriptors are most important. This is particularly relevant in IPF, where the HRCT pattern can be heterogeneous, and where small differences in the distribution/extent of fibrosis can lead to very different functional trajectories.

The connection to the clinical problem is direct: HRCT contains crucial information about the structural phenotype (e.g., extent and distribution of fibrosis, presence of honeycombing/traction bronchiectasis, lung volume) that affects pulmonary capacity and its evolution over time. A DL model that synthesizes these signals into embeddings can therefore support the prediction of outcomes such as FVC at 52 weeks and progression risk, especially when the goal is to integrate imaging information and clinical data into a single personalized estimate.

In summary, we expect deep learning to improve the prediction of FVC decline and progression because it allows us to: (i) capture complex patterns that cannot be easily reduced to a few measures, (ii) leverage the spatial information in CT scans efficiently, and (iii) produce compact representations (embeddings) that, when combined with clinical covariates, increase the model's ability to explain the prognostic heterogeneity observed in IPF patients.

2.2.1 CNNs and ResNet50 for Image Analysis

Convolutional Neural Networks (CNNs) are a class of deep learning models designed to analyze data with spatial structure, such as images. In simple terms, a CNN applies small filters (convolutions) that slide over the image, learning to recognize local patterns. By stacking multiple layers, the model builds progressively more complex representations, moving from basic features like edges and textures to larger structures and pathological patterns. In practice, a CNN can be used for both classification and as a feature extractor: the output of an intermediate layer (an embedding) synthesizes the image information into a compact vector, which

can then be used by downstream models.

In the context of medical imaging, particularly chest CT scans in IPF, CNNs are highly suitable because they leverage the hierarchical nature of radiological signals. Subtle alterations in the lung parenchyma (e.g., reticulation, honeycombing, traction bronchiectasis) and their spatial distribution are difficult to describe using a few manual measurements and exhibit high inter-patient and inter-center variability. A CNN-based approach allows for the automatic learning of robust and reproducible descriptors, reducing the reliance on hand-crafted features and explicit rules.

For feature extraction in this thesis, we utilized ResNet50, a deep CNN architecture based on skip connections (or residual connections) that facilitate training of very deep networks by stabilizing the gradient flow [21]. Compared to older architectures like AlexNet or VGG, ResNet50 generally offers a better trade-off between representational capacity and optimization, mitigating performance degradation issues as the network depth increases [22, 23].

From an application perspective, ResNet50 is an effective choice for our pipeline because it generates embeddings that capture complex patterns in CT images. These embeddings can then be combined with clinical variables and quantitative measures (e.g., lung volume) in an MLP to predict FVC at 52 weeks and progression risk. In other words, ResNet50 acts as the "visual front-end" that translates the CT scan into an informative numerical representation, making it easier for the final model to learn relationships between the radiological phenotype and clinical outcomes.

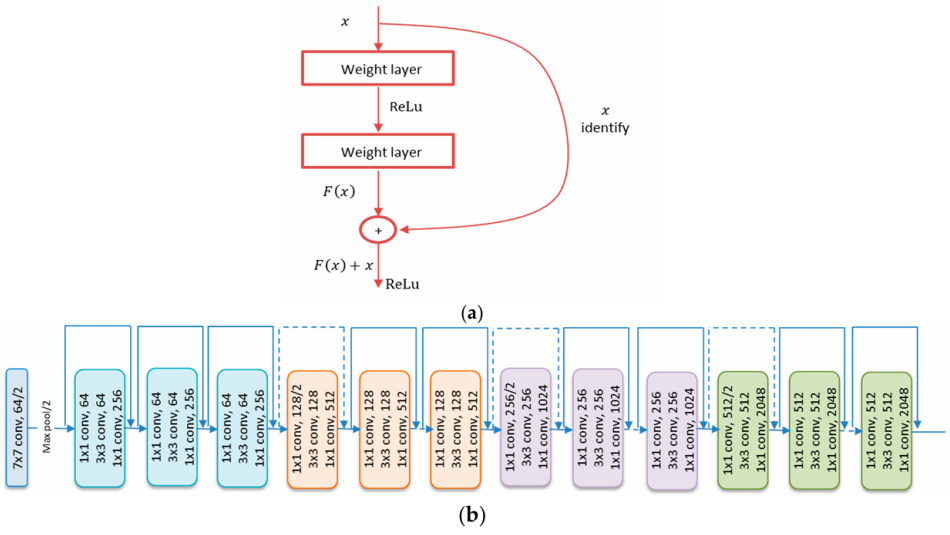

Figure 2.2: Residual block and overall ResNet50 architecture reproduced from [20].

2.2.2 Gradient Boosted Trees and LightGBM

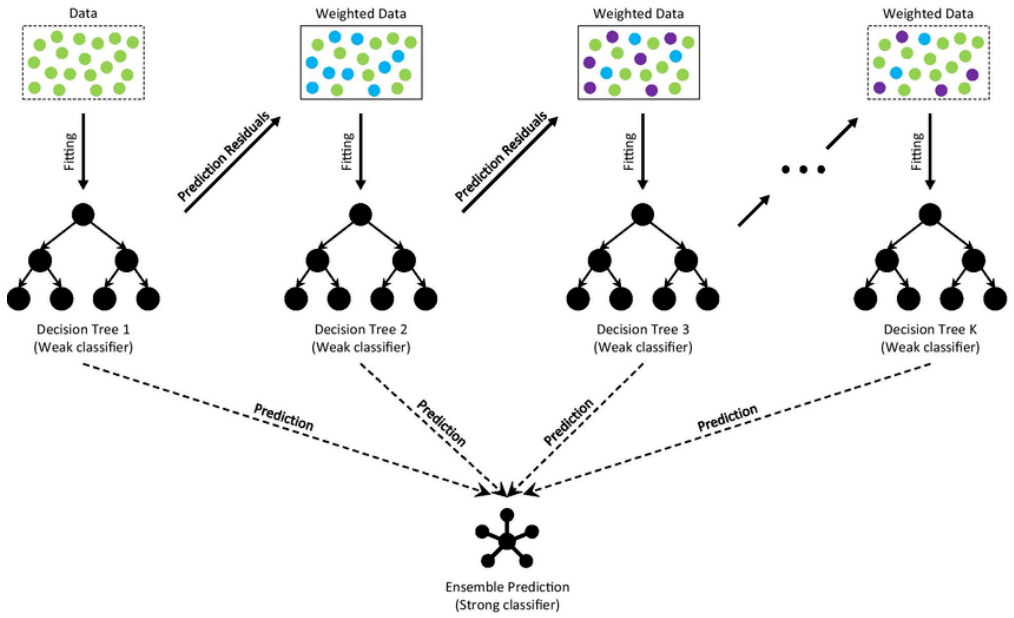

Figure 2.3: Schematic representation of gradient boosting using decision trees. Adapted from Deng et al. [24].

While deep learning models are particularly suited for extracting representations from high-dimensional imaging data, structured clinical variables are often better modeled using tree-based methods. Gradient boosting is an ensemble learning technique that builds a strong predictive model by sequentially combining multiple weak learners, typically shallow decision trees.

At each iteration, a new tree is trained to correct the errors of the previous ensemble by minimizing a differentiable loss function. Rather than fitting all trees independently, boosting follows a stage-wise additive strategy in which each learner focuses on the residual errors of the current model. The final prediction is obtained as the weighted sum of all trees.

LightGBM is an efficient implementation of gradient boosted decision trees designed to handle high-dimensional data with improved computational efficiency. It introduces optimizations such as histogram-based feature binning and a leaf-wise tree growth strategy, which allows faster convergence compared to traditional level-wise boosting algorithms.

In biomedical prediction tasks involving small cohorts and heterogeneous feature types (continuous, binary, categorical), gradient boosted trees often provide strong performance with relatively low risk of overfitting when properly regularized. Unlike neural networks, tree-based models do not require strict feature scaling,

naturally capture non-linear interactions, and are well suited for multimodal tabular representations.

In the context of this thesis, LightGBM complements the deep learning approach: while ResNet50 is used to extract compact representations from CT images, gradient boosted trees serve as a robust predictor operating on structured patient-level features. This combination enables a balanced evaluation between neural architectures and classical ensemble learning methods in the prediction of IPF progression.

Chapter 3

Materials and Methods

3.1 OSIC Dataset

The dataset used in this study was derived from the OSIC Pulmonary Fibrosis Challenge cohort [25] and consists of multimodal patient-level data, including longitudinal pulmonary function measurements and baseline thoracic CT scans. Each patient record includes a unique identifier (`PatientID`), a set of longitudinal FVC measurements acquired at different time points (`Weeks`), baseline demographic variables (age, sex, and smoking status), and a baseline thoracic CT scan provided in DICOM format.

3.1.1 Data Structure

Example of Patient-Level Data

To illustrate the patient-level structure of the dataset, an example of a single patient record is shown below. Each patient is associated with multiple longitudinal FVC measurements collected at irregular time intervals relative to baseline (Week 0), together with static demographic variables and a baseline CT scan The FVC measurements are irregularly spaced and include both negative and positive week indices relative to baseline. In contrast, demographic variables remain constant over time, reflecting the patient-level structure of the dataset. The baseline CT scan is a volumetric acquisition composed of multiple axial slices. The number of slices varies across patients, leading to heterogeneous scan depth. This variability, together with the irregular longitudinal sampling of FVC measurements, informed the preprocessing and aggregation strategies described in the following sections.

Table 3.1: Example of longitudinal FVC measurements for a representative patient (ID00007637202177411956430).

Weeks	FVC (mL)	Age	Sex	Smoking Status
-4	2315	79	Male	Ex-smoker
5	2214	79	Male	Ex-smoker
7	2061	79	Male	Ex-smoker
9	2144	79	Male	Ex-smoker
11	2069	79	Male	Ex-smoker
17	2101	79	Male	Ex-smoker
29	2000	79	Male	Ex-smoker
41	2064	79	Male	Ex-smoker
57	2057	79	Male	Ex-smoker

3.1.2 Dataset Characterization

Prior to cohort filtering, an exploratory analysis was conducted on the full dataset (N = 176 patients) to characterize demographic properties and longitudinal sampling patterns. This analysis was performed to identify potential sources of heterogeneity

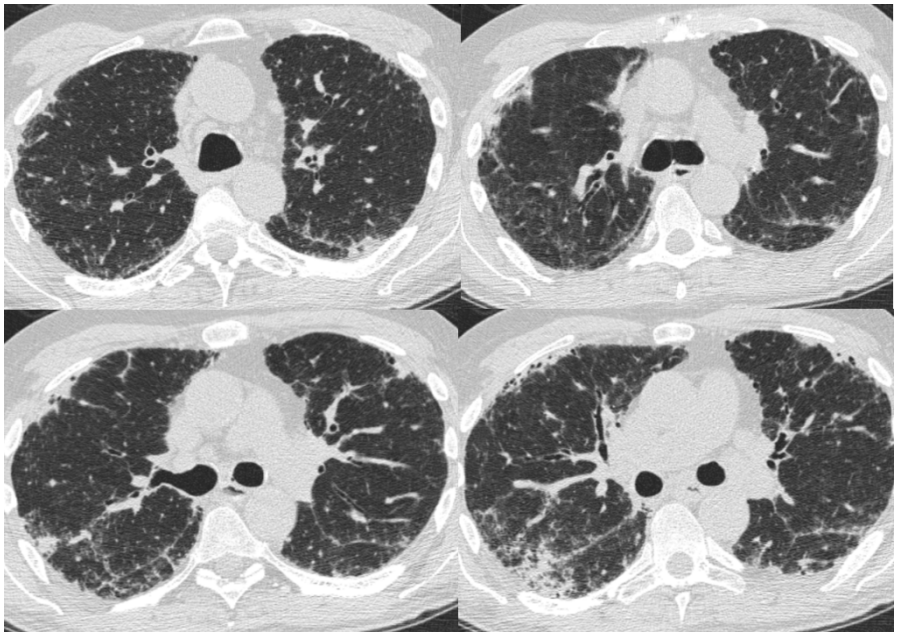

Figure 3.1: Representative axial slices from the baseline CT scan corresponding to the patient reported in Table 3.1.

Table 3.2: Cohort Summary Statistics

Statistic	**Value**
Number of Patients	176
Average FVC Measurements per Patient	8.8 ± 0.7
Average CT Slices per Scan	187.6 ± 179.4
Mean Age (years)	67.3 ± 7.1
Male (%)	79.0%
Female (%)	21.0%
Smoking Status (Never / Ex / Current)	49 / 118 / 9

that could affect subsequent modeling choices and progression label definition.

Demographic Distribution

The cohort shows moderate age variability (67.3 ± 7.1 years), with most patients in the seventh decade of life. The sex distribution is imbalanced, with 79% male and 21% female patients.

Although demographic variables were not used to define progression labels, their distribution provides context regarding cohort composition and potential subgroup imbalance that may affect model generalization.

Temporal Distribution of FVC Measurements

Longitudinal FVC measurements are irregularly sampled across patients, both in terms of the number of observations and their temporal spacing. Figure 3.2 illustrates the distribution of measurement time points across the cohort.

As shown in Figure 3.2, measurement density is higher during early follow-up and progressively decreases at later weeks. Importantly, observations are not consistently available at clinically meaningful reference points such as Week 0 and Week 52.

To assess the feasibility of using exact temporal alignment, patient-level availability at these precise time points was first evaluated. Only 18 patients (10.2%) have an FVC measurement exactly at Week 0, while only 10 patients (5.7%) have a measurement exactly at Week 52. The overlap between these two groups is minimal, indicating that strict reliance on exact time points would drastically reduce the effective study population.

To mitigate this issue while preserving temporal proximity to baseline and one-year follow-up, predefined temporal windows were considered. The availability of measurements within these windows is summarized in Table 3.3.

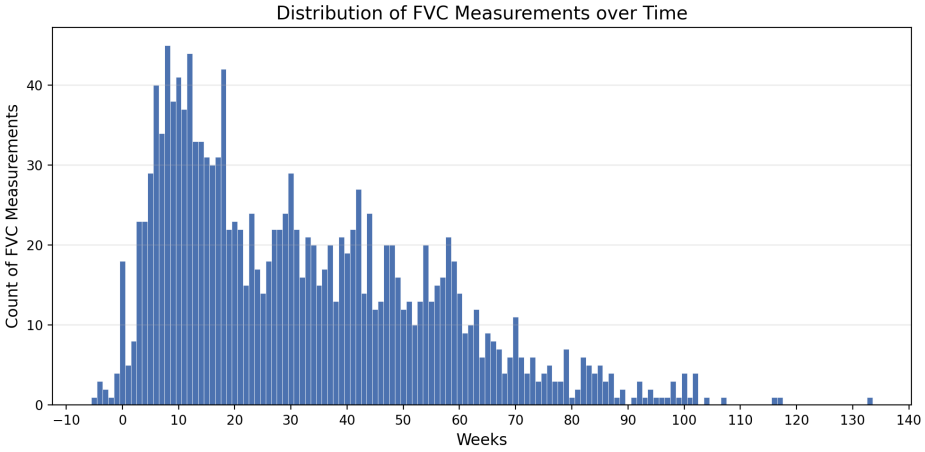

Figure 3.2: Distribution of FVC measurement time points across the full cohort.

The baseline window was defined asymmetrically to maximize patient inclusion while preserving proximity to the clinical baseline. While it substantially increases data availability, requiring valid measurements within both baseline and follow-up windows reduces the effective cohort to 47.7% of the original population.

These observations justify the use of temporal windows when defining progression labels.

Table 3.3: Patient-level measurement availability at exact time points and within predefined temporal windows.

Condition	Patients (N=176)
Exact Week 0	18 (10.2%)
Exact Week 52	10 (5.7%)
Baseline window [-5, 10]	92 (52.3%)
Week 52 window [42, 60]	166 (94.3%)
Both windows available	84 (47.7%)

3.1.3 Cohort Filtering and Ground Truth Definition

Based on the temporal analysis presented in the previous section, strict alignment at exact time points (Week 0 and Week 52) was not feasible due to the limited availability of measurements. Therefore, temporal windows were introduced to define baseline and one-year follow-up observations while preserving proximity to clinically relevant reference points.

Definition of Baseline and Follow-up Measurements

Baseline FVC (FVC_0) was defined as the measurement closest to Week 0 within the interval:

$$\text{Weeks} \in [-5, +10].$$

Similarly, follow-up FVC (FVC_{52}) was defined as the measurement closest to Week 52 within the interval:

$$\text{Weeks} \in [42, 60].$$

In cases where multiple measurements were available within a given window, the value with minimal absolute temporal distance from the reference week (0 or 52) was selected.

Progression Label Definition

Disease progression was defined based on the relative percentage decline in pulmonary function between baseline and one-year follow-up.

The percentage decline in FVC was computed as:

$$\text{FVC Decline}_{\%} = \frac{FVC_0 - FVC_{52}}{FVC_0} \times 100.$$

Patients were labeled as *progressors* if:

$$\text{FVC Decline}_{\%} \geq 10\%,$$

while the remaining patients were labeled as *non-progressors*.

This threshold reflects clinically meaningful functional decline and is consistent with commonly used criteria in IPF studies.

Final Study Cohort

Patients without valid FVC measurements within both temporal windows were excluded from further analysis. After applying these selection criteria, the effective study cohort was reduced from 176 to 84 patients.

The demographic and structural characteristics of the filtered cohort are summarized in Table 3.4.

Table 3.4: Cohort Summary Statistics of the Filtered Dataset (N = 84).

Statistic	**Value**
Number of Patients	84
Average FVC Measurements per Patient	8.9 ± 0.4
Average CT Slices per Scan	186.3 ± 188.6
Mean Age (years)	66.3 ± 6.2
Male (%)	84.5%
Female (%)	15.5%
Smoking (Never / Ex / Current)	26 / 54 / 4

Compared to the original cohort, the filtered population maintains similar age distribution and longitudinal measurement density. The average number of FVC measurements per patient and CT scan depth remained similar to those of the original cohort. Age distribution and longitudinal sampling density were also largely preserved. The proportion of male patients slightly increased (84.5% vs 79% in the full cohort), indicating a slight shift in sex distribution following temporal selection. However, no major demographic distortions are introduced by the filtering procedure.

Although this selection step reduces the overall sample size, it ensures temporal consistency in ground truth construction and prevents bias arising from poorly aligned longitudinal measurements. The resulting cohort was used for subsequent classification and regression experiments.

Progression Label Distribution

After applying the 10% decline criterion, 28 patients (33.3%) were labeled as *progressors*, while 56 patients (66.7%) were labeled as *non-progressors*, resulting in a moderate class imbalance within the filtered cohort (Table 3.5).

The mean FVC decline across the cohort was $7.20 \pm 9.60\%$. As expected, progressors exhibited a substantially higher decline ($17.77 \pm 6.41\%$) compared to non-progressors ($1.92 \pm 5.79\%$), indicating a clear difference between the two groups.

Table 3.5: Distribution of progression labels in the filtered cohort (N = 84).

Label	Patients
Progressors	28 (33.3%)
Non-progressors	56 (66.7%)

When stratified by sex, progression was observed in 24 out of 71 male patients (33.8%) and 4 out of 13 female patients (30.8%), indicating no substantial sex-specific imbalance in progression rates. Similarly, age distributions were nearly identical between progressors (66.3 ± 6.0 years) and non-progressors (66.3 ± 6.3 years), suggesting that age alone does not discriminate between progression groups within this cohort.

Regarding smoking status, progression occurred in 8 out of 26 never-smokers (30.8%), 20 out of 54 ex-smokers (37.0%), and none of the 4 current smokers. Given the limited number of current smokers, no definitive conclusions can be drawn for this subgroup.

Overall, progression labels appear to be more strongly associated with longitudinal functional decline than with demographic variables in this cohort.

Survival Ground Truth Definition (Time-to-Event)

In addition to fixed-horizon progression labeling, disease worsening was also formulated as a time-to-event outcome to enable survival analysis under right-censoring. The event of interest was defined as the first occurrence of clinically meaningful functional decline, operationalized as a relative FVC decrease of at least 10% compared to baseline.

Baseline reference. Baseline FVC (FVC_0) was defined using the same temporal window described previously (Weeks $\in [-5, +10]$) and served as the reference value for all subsequent decline computations.

Event definition. For each follow-up measurement (t, FVC_t) recorded after baseline, the relative decline with respect to baseline was computed as:

$$\Delta_{\%}(t) = \frac{FVC_0 - FVC_t}{FVC_0} \times 100.$$

An event was registered at the earliest follow-up time T such that:

$$\Delta_{\%}(T) \geq 10\%.$$

In this case, the survival outcome was encoded as $(T, \delta) = (T,1)$, where δ denotes the event indicator.

Right-censoring. If a patient never reached the 10% decline threshold during the observed follow-up period, the observation was treated as right-censored. Let T_{last} denote the time of the last available FVC measurement; the survival outcome was encoded as:

$$(T, \delta) = (T_{\text{last}},0).$$

This formulation preserves the available longitudinal information while explicitly accounting for incomplete observation of progression events.

Consistency with classification labeling. The survival endpoint is consistent with the classification criterion (10% relative decline), but differs in that it exploits the *full temporal trajectory* to identify the first occurrence of progression rather than enforcing a fixed one-year horizon.

Survival Endpoint Distribution

After constructing the time-to-event ground truth, 50 patients (59.5%) experienced a progression event during follow-up, while 34 patients (40.5%) were right-censored.

Descriptive statistics of the survival targets are summarized in Table 3.6.

Table 3.6: Summary statistics of the survival ground truth (N = 84).

	All	**Progressors**	**Censored**
Patients (n)	84	50	34
Time-to-event (weeks)	37.9 ± 20.7	25.9 ± 18.9	55.5 ± 2.7
Time range (weeks)	$[1, 63]$	$[1, 59]$	$[52, 63]$
Max FVC Drop (%)	12.5 ± 8.7	17.8 ± 7.2	4.7 ± 3.3
Baseline FVC (mL)	2828 ± 854	2773 ± 855	2910 ± 859

Progressors exhibited earlier functional decline (mean 25.9 weeks) compared with censored patients, whose follow-up times were concentrated near one year (mean 55.5 weeks). As expected, maximum FVC decline was substantially larger among patients experiencing an event. Baseline FVC values were comparable between groups, suggesting that progression timing is not trivially explained by baseline pulmonary function alone.

3.1.4 CT Image Preprocessing

Baseline thoracic CT scans were provided in DICOM format and consisted of volumetric acquisitions composed of multiple axial slices per patient. Given the heterogeneity in slice count, intensity scaling, and acquisition parameters, a standardized preprocessing pipeline was applied to ensure consistency across patients prior to feature extraction.

Conversion to Hounsfield Units

Each DICOM slice was first converted to Hounsfield Units (HU) using the rescale slope and intercept provided in the DICOM metadata:

$$HU = \text{pixel\_value} \times \text{RescaleSlope} + \text{RescaleIntercept}.$$

This step ensures physically meaningful intensity representation across scans acquired on different scanners.

Lung Region Segmentation

To focus the analysis on lung parenchyma while preserving anatomical context, an automatic lung segmentation procedure was applied to each slice. The segmentation approach was adapted from a publicly available implementation developed for the OSIC Pulmonary Fibrosis Challenge [26]. The segmentation pipeline consisted of:

- Intensity standardization of the slice using z-score normalization;
- Unsupervised threshold estimation via K-means clustering (2 clusters) applied to the central region of the image;
- Binary thresholding based on cluster centroids;
- Morphological erosion and dilation operations to remove noise and refine region boundaries;
- Connected component analysis to retain anatomically plausible lung regions;
- Final dilation to ensure full parenchymal coverage.

This approach provides a fully automated and reproducible lung mask without manual annotation.

Intensity Windowing and Soft Masking

To restrict the dynamic range to clinically relevant values, HU intensities were clipped to the interval:

$$[-1000, 400] \text{ HU}.$$

This range captures air (-1000 HU), lung tissue, and soft tissue structures, while excluding extreme bone intensities.

Instead of applying a hard mask, a soft masking strategy was adopted to preserve part of the surrounding thoracic context and reduce sharp boundary artifacts that could affect convolutional filters. Voxels inside the lung mask retained full intensity, whereas voxels outside the mask were attenuated to 10% of their original value.

$$I_{\text{soft}} = I \cdot M + 0.1 \cdot I \cdot (1 - M),$$

where M denotes the binary lung mask.

This strategy attenuates non-lung regions while preserving surrounding anatomical context.

Global Intensity Normalization

Following clipping and masking, intensities were linearly normalized to the range [0,1]:

$$I_{\text{norm}} = \frac{I_{\text{soft}} + 1000}{1400}.$$

This transformation maps:

- -1000 HU (air) $\rightarrow$ 0
- +400 HU (upper bound) $\rightarrow$ 1

The resulting normalized images are therefore intensity-consistent across patients.

Spatial Standardization

All slices were resized to a fixed spatial resolution of 224×224 pixels using area interpolation. This resolution was selected to match the input dimensionality expected by standard convolutional neural networks. Each preprocessed slice was saved as a NumPy array in float32 format. This reduces storage overhead compared to DICOM while preserving normalized intensity information and enabling efficient loading during model training.

The final output of the preprocessing pipeline consists of standardized, lung-focused axial slices stored as normalized NumPy arrays and organized at the patient level. The preprocessed CT slices were subsequently used for both handcrafted feature extraction and deep feature embedding, as described in the following sections.

3.1.5 Handcrafted CT-Derived Features

In addition to deep learning representations, we extracted a set of handcrafted CT-derived features designed to capture global morphological and attenuation-related characteristics of the lung parenchyma. These descriptors provide interpretable summaries of lung structure and tissue density patterns and are consistent with established radiomics approaches in quantitative imaging [27, 28].

Computed tomography attenuation values are expressed in Hounsfield Units (HU), where air typically corresponds to approximately -1000 HU and water to 0 HU. Because healthy lung tissue contains a large amount of air, normal parenchyma exhibits predominantly low attenuation values. In contrast, fibrotic remodeling increases tissue density, leading to higher attenuation values within affected regions. Consequently, statistical descriptors of the attenuation distribution can provide quantitative indicators of structural lung alterations associated with interstitial lung disease.

All handcrafted features were initially computed at the slice level using the automatically generated lung segmentation mask. Slice-level measurements were subsequently aggregated at the patient level using arithmetic averaging to obtain fixed-dimensional representations while accommodating CT scans with variable numbers of slices.

Morphological and Volumetric Descriptors

To approximate lung size and tissue burden, several morphological descriptors were derived from the binary lung segmentation mask.

For each CT slice, the number of lung pixels was obtained from the segmentation mask. An approximate lung volume proxy was then computed by summing lung pixels across slices and scaling by the in-plane pixel spacing and slice thickness obtained from the DICOM header. Because thoracic CT scans are typically acquired during maximal inspiration, this estimate provides a coarse proxy for total lung capacity and may correlate with pulmonary function measurements such as forced vital capacity (FVC).

To characterize parenchymal tissue burden, an inner lung mask was generated by applying a slight morphological erosion to the lung segmentation mask. This step reduces boundary artifacts caused by partial volume effects near pleural surfaces.

Within this inner lung region, several tissue-related quantities were computed, including:

- Number of tissue pixels within the inner lung mask,
- Tissue area estimates scaled by pixel spacing,
- Volume-adjusted tissue estimates incorporating slice thickness,
- Ratios between tissue pixels and total slice area,
- Ratios between tissue pixels and lung area.

These descriptors provide coarse but robust proxies of parenchymal structural burden, which may reflect fibrotic tissue involvement or increased parenchymal density in interstitial lung disease [16, 17].

Intensity Distribution Features

To characterize attenuation patterns within the lung parenchyma, first-order histogram statistics were extracted from lung-masked pixels following standard radiomics definitions [29]. The following descriptors were computed:

- **Mean intensity**, representing the average Hounsfield Unit (HU) value within the lung mask,
- **Skewness**, measuring asymmetry in the attenuation distribution,
- **Kurtosis**, quantifying the peakedness and tail heaviness of the distribution.

Mean intensity provides a global estimate of lung attenuation and tends to increase when fibrotic remodeling replaces normal air-filled parenchyma with denser tissue. Skewness captures the asymmetry of the attenuation distribution: normal lung parenchyma typically exhibits a left-skewed distribution dominated by low attenuation values, whereas fibrotic lungs show a shift toward higher attenuation values. Kurtosis reflects the concentration of values around the mean and the presence of extreme attenuation values, which may indicate heterogeneous mixtures of normal lung tissue and fibrotic regions.

Previous studies have shown that histogram-derived CT features correlate with pulmonary function measurements in idiopathic pulmonary fibrosis, with kurtosis demonstrating particularly strong associations with physiologic impairment [30].

Patient-Level Aggregation

For each patient, slice-level handcrafted features were aggregated across slices using arithmetic averaging to obtain a fixed-dimensional feature vector. This aggregation strategy ensures dimensional consistency across scans with heterogeneous slice counts while reducing sensitivity to slice-level noise and local segmentation artifacts.

Table 3.7: Handcrafted CT-derived features extracted at the patient level.

Feature	Description
Approximate Volume	Approximate lung volume proxy obtained by summing lung pixels across slices and scaling by pixel spacing and slice thickness.
Average Number of Tissue Pixels	Average number of tissue pixels per slice within the inner lung mask.
Average Tissue Area	Estimated tissue area obtained by scaling tissue pixels by pixel spacing.
Average Tissue Thickness	Volume-adjusted tissue estimate incorporating slice thickness.
Average Tissue By Total	Ratio between tissue pixels and total slice pixels.
Average Tissue By Lung	Ratio between tissue pixels and lung pixels.
Mean Intensity	Mean Hounsfield Unit value within the lung mask.
Skewness	Skewness of the lung attenuation distribution.
Kurtosis	Kurtosis of the lung attenuation distribution.

These handcrafted descriptors were subsequently combined with deep CNN-derived imaging features and clinical variables to construct multimodal patient-level representations used for downstream predictive modeling.

3.2 Methods

3.2.1 Experimental Design Overview

Problem Formulation

The objective of this study is to model one-year disease progression in idiopathic pulmonary fibrosis using baseline multimodal information. For each patient, baseline data include CT-derived imaging features, handcrafted radiomic descriptors, demographic variables (age, sex, smoking status), and baseline pulmonary function measurements.

Three complementary supervised learning tasks were investigated:

- **Binary classification:** prediction of clinically meaningful functional decline ($\geq 10\%$ relative FVC reduction within one year).
- **Regression:** prediction of absolute FVC at one-year follow-up (FVC_{52}).
- **Survival analysis:** time-to-event modeling of progression defined as the first occurrence of $\geq 10\%$ relative FVC decline.

All models operate strictly on baseline information, ensuring that predictions simulate a realistic clinical decision-making scenario.

Cross-Validation Protocol

Given the limited cohort size ($N = 84$), model performance was assessed using patient-level stratified 5-fold cross-validation.

Patients were treated as the atomic unit of splitting, ensuring that all CT slices and patient-level variables belonging to the same subject were assigned to a single split. This prevents information leakage between training, validation, and test partitions.

Classification and Regression. Binary classification and regression tasks shared identical fold assignments. Stratification was performed on the binary progression label (`has_progressed`) to maintain a balanced distribution of progressors and non-progressors across folds.

Across the five folds, each patient appeared exactly once in the test set and once in the validation set, while being included in the training set in the remaining folds.

Survival Analysis. Survival modeling required a distinct split definition due to the time-to-event outcome with right-censoring. Stratification was performed on the event indicator (`event`), and folds were generated such that each rotation used three partitions for training, one for validation, and one for testing.

Within-Fold Data Handling. All preprocessing steps involving data-dependent statistics (e.g., feature scaling or dimensionality reduction) were fitted exclusively on training patients within each fold and then applied unchanged to validation and test splits. This strict separation ensures unbiased performance estimation.

Feature Extraction and Patient-Level Representation

Baseline CT scans were processed slice-wise using a ResNet50 backbone pre-trained on ImageNet [31]. The final classification layer was removed and the 2048-dimensional embedding from the global average pooling layer was retained for each slice.

Since CT scans contain a variable number of slices per patient, slice-level embeddings were aggregated into fixed-dimensional patient-level representations using pooling strategies (mean pooling, max pooling, or concatenated mean+max pooling). Slice masks were used to exclude padded slices from the aggregation process.

In parallel, handcrafted CT-derived radiomic features and demographic variables were incorporated at patient level. Continuous variables were standardized within each fold using training statistics only. Categorical variables were encoded numerically and centered to improve numerical stability.

Depending on the experimental configuration, models operated on:

- deep imaging features only,
- handcrafted radiomic features only,
- clinical variables only,
- or multimodal concatenations.

Training and Optimization Strategy

All neural models were trained using mini-batch gradient-based optimization. Model selection was performed within each fold based on validation performance.

To mitigate overfitting given the limited cohort size, several regularization strategies were adopted, including early stopping, weight decay, and dropout within fully connected layers. Learning rate scheduling was applied when validation performance plateaued. Final reported metrics correspond to the model checkpoint achieving the best validation performance within each fold.

All experiments were initialized with fixed random seeds to ensure reproducibility.

Modeling Framework

While all tasks share a unified patient-level representation, the downstream modeling strategy was adapted to the statistical structure of each prediction problem.

Binary classification was addressed using discriminative neural models optimized for cross-entropy-based objectives and a LightGBM configuration. Regression models were trained using mean squared error to estimate continuous pulmonary function outcomes. Survival modeling employed a Cox proportional hazards formulation with regularization to account for limited event counts.

This modular yet unified design enables direct comparison across tasks while preserving methodological consistency.

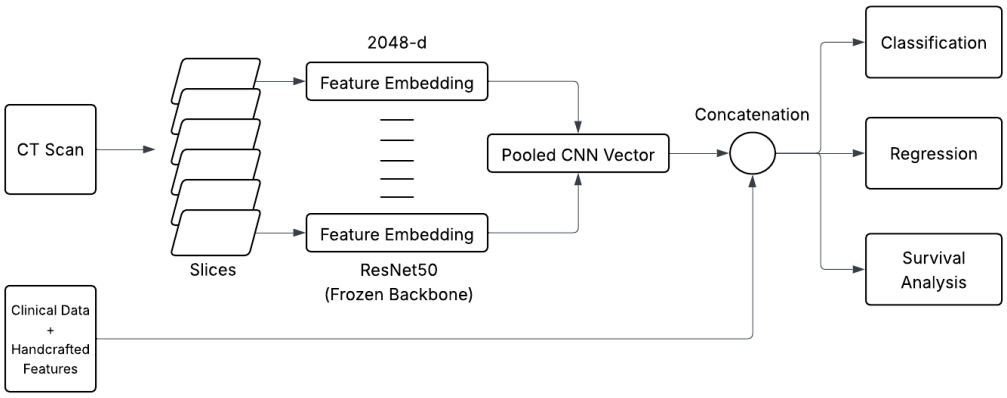

Figure 3.3: Overview of the proposed multimodal experimental pipeline for IPF disease progression modeling. Baseline CT scans are processed slice-wise through a ResNet50 backbone with frozen weights to extract 2048-dimensional feature embeddings. These embeddings are aggregated into a patient-level vector via pooling and concatenated with clinical variables and handcrafted radiomic features. The resulting multimodal representation serves as input for three distinct supervised learning tasks.

3.2.2 Binary Classification

Binary classification aims to estimate the probability that a patient will experience clinically meaningful progression within one year, defined as a relative FVC decline $\geq 10\%$. All experiments were evaluated using the 5-fold patient-level cross-validation protocol described in Section 3.2.1 and were organized as a structured ablation study comparing feature subsets and pooling strategies.

Experimental Design and Ablation Blocks

Classification experiments were organized into three experimental blocks: (i) a tabular baseline using handcrafted CT-derived descriptors alone and in combination with demographic variables; (ii) an imaging-only block comparing pooling strategies for deep CT embeddings (mean, max, and concatenated mean+max pooling); (iii) a multimodal block combining deep features with handcrafted descriptors and optionally demographic variables.

Neural Classifier (MLP)

The neural classifier is a lightweight multilayer perceptron designed to reduce overfitting risk in the small-cohort regime. Let $\mathbf{e}_i \in \mathbb{R}^d$ denote the slice-level embedding extracted from the frozen ResNet50 backbone for slice i, with $d = 2048$. Given a patient with n available slices, a fixed-dimensional patient representation is obtained via pooling over slice embeddings:

$$\mathbf{p} = \mathrm{Pool}\Big(\{\mathbf{e}_i\}_{i=1}^{n}\Big),$$

where $\mathrm{Pool}(\cdot)$ is either mean pooling, max pooling, or concatenated max+mean pooling. Variable-length scans were handled using slice masks to ensure that padded slices did not contribute to the aggregation. To control model capacity, the pooled CNN representation is projected to a 32-dimensional latent vector using a fully-connected reduction layer with ReLU activation and dropout:

$$\tilde{\mathbf{p}} = \mathrm{Dropout}(\mathrm{ReLU}(W\mathbf{p} + b)), \quad \tilde{\mathbf{p}} \in \mathbb{R}^{32}.$$

Handcrafted CT features and demographic variables (if included by the ablation setting) are concatenated to $\tilde{\mathbf{p}}$ to form the final patient-level vector $\mathbf{z}$. A shallow classification head then produces a single logit, subsequently mapped to a probability via sigmoid:

$$\hat{y} = \sigma(g_\theta(\mathbf{z})).$$

The head consists of a 32-unit hidden layer with ReLU and dropout, followed by a linear output layer.

Fold-wise Preprocessing and Missing Data Handling

Within each fold, missing values are handled prior to normalization using statistics computed on training patients only. Continuous variables (handcrafted features and age, when used) are imputed with the training-set median, whereas categorical variables are imputed with a constant value encoding an "unknown" category.

Handcrafted CT-derived features are standardized using a `StandardScaler` fitted on training patients only. Demographic variables are encoded in a numerically

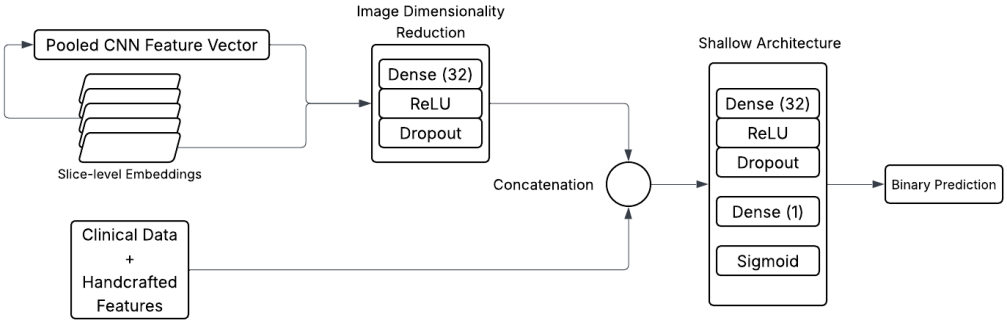

Figure 3.4: Architecture of the neural MLP classifier used for progression prediction. Patient-level representations derived from pooled CNN embeddings and clinical variables are processed through a lightweight multilayer perceptron to estimate the probability of disease progression.

stable representation form: (i) age is z-normalized on the training set; (ii) sex is mapped to $\{-1, +1\}$; (iii) smoking status is one-hot encoded and centered to $[-0.5, 0.5]$ to improve numerical stability. Original demographic columns are removed after encoding to avoid duplication.

Training Procedure

The model is trained using AdamW optimization with learning rate 3.9×10^{-5}, weight decay 3×10^{-3}, batch size 16, and a maximum of 60 epochs. To mitigate class imbalance, the binary cross-entropy loss is implemented as `BCEWithLogitsLoss` with a positive-class weight (`pos_weight`) computed from the training split of each fold. Gradient clipping with maximum norm 1.0 is applied at each update step.

A ReduceLROnPlateau scheduler monitors validation AUC and reduces the learning rate by a factor 0.5 after 5 stagnant epochs, with minimum learning rate 10^{-6}. Early stopping is applied with patience 15 based on validation AUC; the model parameters corresponding to the best validation AUC are restored before evaluation. To support reproducibility, each fold is trained under a deterministic seed initialisation scheme.

Threshold Selection and Evaluation

Model performance is primarily quantified using ROC AUC on validation and test sets. In addition, decision-threshold-dependent metrics are reported on the test split using: (i) the default threshold 0.5, and (ii) an *optimal* threshold selected on the validation split.

The optimal threshold is chosen by maximising Youden's J statistic on the validation ROC curve:

$$J(\tau) = \mathrm{TPR}(\tau) - \mathrm{FPR}(\tau),$$

and the corresponding threshold $\tau^\star$ is then applied to the test predictions. For the best threshold, the following metrics are reported: accuracy, precision, recall (sensitivity), F1-score, and specificity.

Gradient-Boosted Trees Classifier (LightGBM)

In addition to the neural classifier, we implemented a gradient-boosted decision tree (GBDT) model using LightGBM, which is well suited for small tabular datasets and can capture non-linear feature interactions. The LightGBM pipeline operates on patient-level feature vectors obtained by concatenating (depending on the ablation setting) pooled CNN embeddings, handcrafted CT-derived descriptors, demographic variables, and optionally baseline FVC.

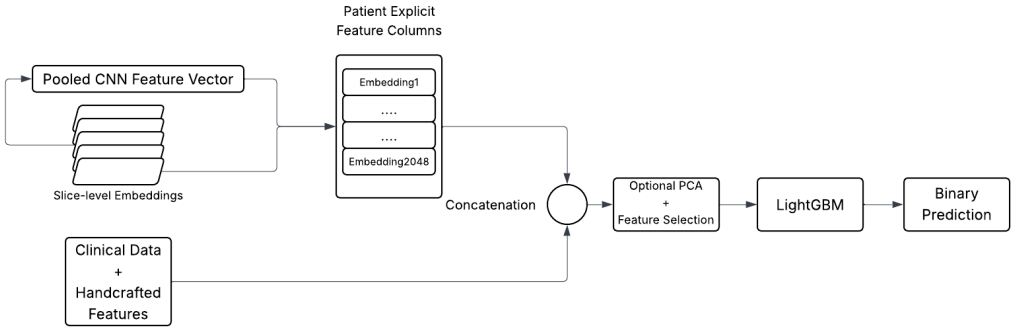

Figure 3.5: Architecture of the LightGBM classification pipeline. Slice-wise CNN embeddings are aggregated into a patient-level representation and expanded into explicit feature columns ($d = 2048$). These are concatenated with handcrafted radiomic descriptors and clinical variables. The resulting multimodal vector undergoes optional dimensionality reduction via PCA and standardization before being processed by a regularized ensemble of gradient-boosted decision trees to estimate the probability of disease progression.

Patient-level CNN aggregation. Slice-wise CNN embeddings (ResNet50, 2048-d per slice) are pooled to patient-level vectors using a deterministic aggregation function. In this study, the primary LightGBM configuration uses mean pooling across slices (selected as the best CNN setting in the ablation design), producing a 2048-dimensional patient embedding. The resulting vector is expanded into explicit

columns $\{\texttt{cnn\_emb\_i}\}_{i=1}^{d}$ before concatenation with the remaining patient-level variables.

Fold-wise preprocessing. All preprocessing is performed independently within each cross-validation fold, fitting transformations on the training split only and applying them unchanged to validation and test splits. Demographic variables are encoded in a numerically stable form: (i) age is z-normalized using a `StandardScaler`; (ii) sex is mapped to $\{-1, +1\}$; (iii) smoking status is one-hot encoded and centered to $\{-0.5, +0.5\}$. When baseline FVC is included, it is z-normalized on the training split and stored as `FVC_normalized`. Missing values are imputed using training statistics (median for continuous handcrafted/FVC variables; constant 0 for categorical variables).

After feature construction, all selected feature columns are standardized using a `StandardScaler` fitted on the training split. This global scaling step is applied to ensure comparable ranges across heterogeneous feature groups (CNN embeddings, handcrafted descriptors, and clinical variables).

Optional dimensionality reduction. Some ablation configurations enable PCA on the training feature matrix, with the number of components selected adaptively to explain at least 95% of the variance, subject to a maximum of 50 components. The fitted PCA transformation is then applied to validation and test sets. (The main experiments were conducted without PCA unless otherwise specified.)

Model configuration and training. LightGBM is trained with a strongly regularized configuration tailored to the small-cohort regime: shallow trees (maximum depth 2), a small number of leaves (3), learning rate 0.01, feature subsampling (0.7), bagging (0.7 with frequency 5), minimum 15 samples per leaf, and both L1/L2 regularization ($\lambda_1 = \lambda_2 = 0.5$), with a minimum gain threshold to split of 0.05. Training uses up to 1000 boosting iterations with early stopping (patience 100), monitoring validation AUC; the best iteration is retained and used for test inference. Deterministic execution is enforced through fixed seeds and column-wise training.

Threshold selection and evaluation. For each fold, an optimal decision threshold is determined on the validation split by maximizing Youden's J statistic on the ROC curve:

$$J(\tau) = \mathrm{TPR}(\tau) - \mathrm{FPR}(\tau).$$

The selected threshold $\tau^\star$ is then applied to the test predictions to compute accuracy, precision, recall (sensitivity), F1-score, and specificity, while AUC is reported using the continuous probabilities. Fold-level ROC and confusion matrix visualizations

are saved, together with aggregate ROC and aggregated confusion matrices across folds.

Model Explainability with SHAP

To support interpretability, SHAP (SHapley Additive exPlanations) analysis is performed for each LightGBM model. For every fold, SHAP values are computed on the test split using `TreeExplainer`, and then pooled across folds to obtain cross-validated explanations.

We report: (i) a cross-fold mean absolute SHAP importance ranking (top features), including standard deviation across folds; (ii) group-level importance by summing mean |SHAP| within feature groups (CNN embeddings, handcrafted descriptors, demographics, and baseline FVC when present).

3.2.3 FVC regression at 52 weeks

In addition to binary progression classification, a regression task was formulated to predict the Forced Vital Capacity (FVC) at 52 weeks using baseline information and CT-derived features. This formulation provides a continuous estimate of functional decline, complementing the binary progression outcome.

Target definition and fold-wise normalisation

The regression target is the absolute FVC value at 52 weeks (in mL). Since FVC is also used as an input feature (baseline FVC).To prevent data leakage, normalization parameters were estimated exclusively on the training patients within each fold. For each fold, we fit the normalisation parameters only on the training patients and then applied the resulting transform to validation and test sets. Concretely, we standardised jointly the pair {`baselinefvc`, `gt_fvc52`} using a `StandardScaler` fitted on the fold training set, and then transformed all samples in that fold. This guarantees that no statistics from validation/test patients influence training. Demographic variables were encoded using the same preprocessing strategy described in Section 3.2.2.

Input modalities and ablation configurations

The model operates at the patient level but receives slice-level CNN descriptors extracted from CT scans. We evaluated four feature configurations in an ablation study:

- **cnn_only**: CNN slice features + baseline FVC only;
- **cnn_hand**: CNN features + hand-crafted features + baseline FVC;

- **cnn_demo**: CNN features + demographics + baseline FVC;
- **full**: CNN + hand-crafted + demographics + baseline FVC.

Baseline FVC was included in all configurations due to its strong clinical association with future pulmonary function.

Model architecture

We implemented a multi-branch MLP combining (a) pooled CNN features from CT slices, (b) baseline FVC, and optionally (c) hand-crafted features and (d) demographics. Slice-level CNN embeddings are aggregated into a fixed-size patient representation using one of the pooling strategies: *max*, *mean* or *max+mean concatenation*.

The architecture is organised into dedicated branches:

- **FVC branch**: baseline FVC is processed through a small MLP to a higher-dimensional embedding (default: $1 \to 64 \to 128$), allowing the model to learn a dedicated representation of baseline pulmonary function;
- **CNN branch**: pooled CNN embedding is mapped to a compact representation (default: `cnn_dim` $\to 512 \to 256$) with Batch Normalisation and dropout;
- **Hand-crafted branch** (optional): hand-crafted radiomic features are projected to 64 dimensions;
- **Demographics branch** (optional): demographics are projected to 32 dimensions.

The concatenated representation is processed through a lightweight feature-attention gate that learns sigmoid weights for each modality, allowing the model to adaptively reweight the contribution of CNN, clinical, and handcrafted features.

Training protocol and evaluation

Training is performed within each fold using AdamW optimization (learning rate 5×10^{-4}, weight decay 0.05) and mean squared error (MSE) loss. A `ReduceLROnPlateau` scheduler monitors validation loss and reduces the learning rate when improvement stalls. Early stopping is applied with patience 20 epochs, restoring the best model checkpoint.

Model performance was evaluated using:

- Mean Absolute Error (MAE),
- Root Mean Squared Error (RMSE),

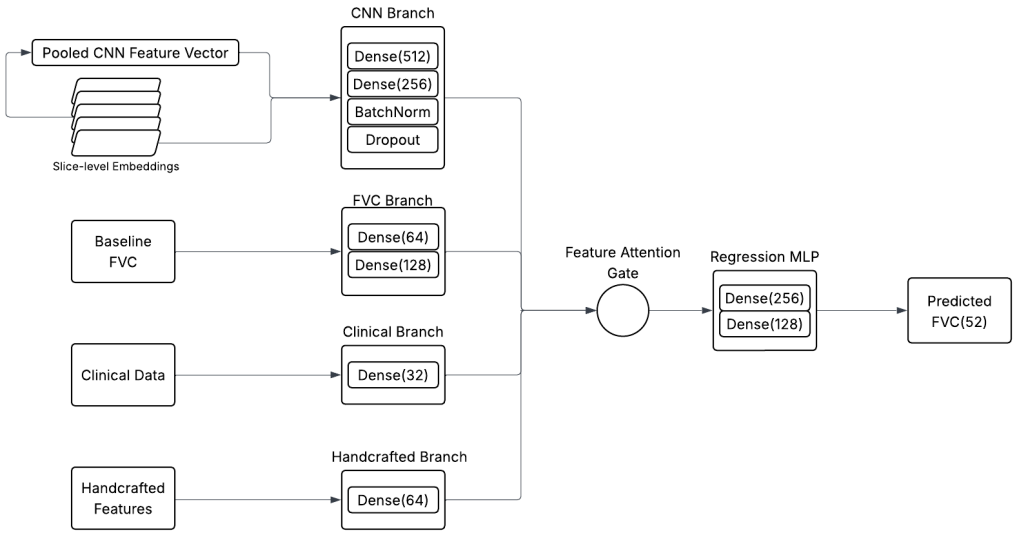

Figure 3.6: Architecture of the proposed multi-branch regression model for predicting FVC at 52 weeks. The network processes multimodal inputs through three dedicated pipelines: a CNN branch for pooled slice embeddings, an FVC branch to anchor the prediction using baseline values, and a clinical branch for handcrafted features. A feature-attention gate adaptively reweights these representations before a final regression MLP estimates the absolute FVC value at 52 weeks.

- Coefficient of determination (R^2).

Although training and validation operate on normalised FVC values, final test metrics are computed after *de-normalising* predictions back to mL using the fold-specific FVC scaler. In addition to scalar metrics, we generated diagnostic plots per fold: predicted vs true scatter with identity line, residual plots, residual distribution, and Bland–Altman analysis.

3.2.4 Survival Analysis

While binary classification models predicts a fixed one-year outcome and regression predicts continuous pulmonary function values, survival analysis explicitly models the temporal dynamics of disease progression while accounting for right-censored observations.

Outcome Definition

For each patient i, survival data consist of:

- a time variable T_i representing the time (in weeks) from baseline to either progression or last follow-up;
- an event indicator $\delta_i \in \{0,1\}$, where $\delta_i = 1$ indicates observed progression and $\delta_i = 0$ indicates right-censoring.

This formulation allows the model to incorporate both patients who experienced progression during follow-up and those whose event time is unknown due to censoring.

Model

To estimate progression risk over time, we employed a Cox Proportional Hazards (CoxPH) model. The Cox model defines the hazard function as:

$$h(t \mid \mathbf{x}) = h_0(t) \exp(\beta^\top \mathbf{x}),$$

where $h_0(t)$ is the baseline hazard function and $\mathbf{x}$ represents the patient-level feature vector. The exponential term models the multiplicative effect of covariates on the hazard of progression.

Given the relatively small cohort size and the potential correlation between features, the Cox model was regularized using a ridge penalty. Regularization stabilizes coefficient estimation and mitigates overfitting when the number of covariates approaches the number of observed events.

Model fitting was performed using the `lifelines` implementation of Cox proportional hazards regression with the following hyperparameters:

- penalization strength: `penalizer` $= 0.5$
- elastic-net mixing parameter: `l1_ratio` $= 0.0$ (pure ridge regularization)

Feature Selection and Dimensionality Control

Feature selection was performed to control model complexity and avoid overfitting given the limited cohort size. Candidate predictors included demographic variables and handcrafted radiomic descriptors derived from CT images.

First, univariate Cox models were evaluated to assess the stability of feature effects across cross-validation folds. Features exhibiting unstable coefficients or frequent sign reversals were excluded.

Second, redundancy among predictors was assessed through pairwise correlation analysis. When highly correlated variables were identified (Pearson $r > 0.85$), only one representative feature was retained.

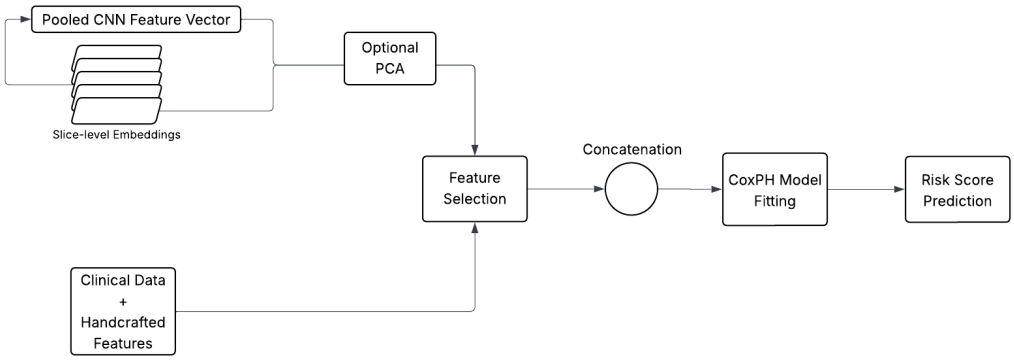

Figure 3.7: Survival analysis pipeline. Patient-level features derived from pooled CNN embeddings, handcrafted radiomic descriptors, and clinical variables are processed through feature selection and used to train a regularized Cox proportional hazards model, producing a risk score for progression prediction.

Finally, the number of predictors was restricted to maintain a favorable events-per-variable (EPV) ratio. Empirical analysis indicated that models including more than three handcrafted features exhibited increased validation variance and signs of overfitting.

Preprocessing

Continuous features were standardized within each cross-validation fold using statistics computed from the training patients only. The demographic variable Sex was encoded as a centered binary variable (Female = -1, Male = +1) to improve numerical stability. All preprocessing steps were performed independently within each fold to avoid information leakage.

CNN Feature Aggregation

CNN-derived imaging representations were incorporated into the survival model through statistical aggregation of slice-level embeddings. For each patient, the CNN encoder produces a high-dimensional feature vector extracted from the CT volume. To obtain a compact patient-level representation suitable for Cox regression, four summary statistics were computed from the embedding vector:

- **Mean**: average activation value across embedding dimensions;
- **Variance**: dispersion of feature activations;

- **L2 norm**: magnitude of the embedding vector, capturing overall feature intensity;
- **Entropy**: information entropy of normalized feature activations, reflecting the heterogeneity of the representation.

These aggregated descriptors provide complementary information about the distribution and structure of CNN activations while keeping the number of predictors small, which is essential for stable estimation in survival models with limited sample size. The resulting four statistics were used as imaging-derived covariates in the Cox proportional hazards model.

Model Evaluation

Model performance was evaluated using the concordance index (C-index), a standard metric in survival analysis that measures the ability of the model to correctly rank patients according to their predicted risk.

The C-index estimates the probability that, for a pair of comparable patients, the one predicted to have higher risk actually experiences the event earlier. Values range from 0.5 (random ordering) to 1.0 (perfect risk ranking).

For each cross-validation fold, the C-index was computed on training, validation, and test splits. The difference between training and validation performance was also monitored to assess potential overfitting.

Risk Stratification Analysis

To visualize the clinical interpretability of the predicted risk scores, patients in the validation set were stratified into two groups based on the median predicted risk:

- **High-risk group**: patients with predicted risk above the median;
- **Low-risk group**: patients with predicted risk below the median.

Kaplan–Meier survival curves were then estimated for the two groups to assess whether the model successfully separates patients with different progression trajectories. Additionally, hazard ratios derived from the fitted Cox model were reported to quantify the effect size of individual covariates.

Chapter 4

Results

This chapter presents the experimental results obtained for the three predictive tasks introduced in Section 3.2.1: binary classification of disease progression, regression of one-year pulmonary function, and survival analysis of time-to-event progression.

All experiments follow the patient-level 5-fold cross-validation protocol described in Section 3.2.1. Performance metrics are reported as the mean and standard deviation across folds to account for variability due to the limited cohort size.

4.1 Binary Classification Results

The binary classification task predicts clinically meaningful disease progression within one year, defined as a relative FVC decline of at least 10%. The evaluation follows the ablation design introduced in Section 3.2.2, comparing three groups of feature configurations: (i) tabular handcrafted baselines, (ii) CNN-based imaging representations with different pooling strategies, and (iii) multimodal models combining imaging and handcrafted descriptors.

Table 4.1 summarizes the cross-validated performance across all configurations.

Table 4.1: Cross-validated performance of classification models for progression prediction. Values are reported as mean ± standard deviation across folds.

Configuration	Accuracy	Precision	Recall	Specificity	F1	AUC
Handcrafted	0.57 ± 0.15	0.31 ± 0.25	0.44 ± 0.40	0.63 ± 0.34	0.34 ± 0.27	0.45 ± 0.22
Handcrafted + Demographics	0.43 ± 0.16	0.31 ± 0.10	0.54 ± 0.31	0.38 ± 0.31	0.37 ± 0.13	0.45 ± 0.15
CNN (max pooling)	0.60 ± 0.10	0.32 ± 0.22	0.36 ± 0.34	0.72 ± 0.24	0.31 ± 0.24	0.53 ± 0.16
CNN (max+mean pooling)	0.41 ± 0.13	0.32 ± 0.05	0.61 ± 0.16	0.31 ± 0.26	0.40 ± 0.02	0.49 ± 0.13
CNN (mean pooling)	0.57 ± 0.13	0.44 ± 0.15	0.47 ± 0.36	0.63 ± 0.36	0.38 ± 0.12	0.49 ± 0.13
CNN + Handcrafted	0.59 ± 0.20	0.43 ± 0.37	0.44 ± 0.34	0.67 ± 0.37	0.38 ± 0.28	**0.54 ± 0.25**
CNN + Handcrafted + Demo.	0.59 ± 0.05	0.32 ± 0.18	0.43 ± 0.33	0.67 ± 0.22	0.35 ± 0.22	0.46 ± 0.10

Overall Performance and Stability Analysis

Overall predictive performance remained modest across all configurations, with ROC AUC values ranging between approximately 0.45 and 0.54. The best discriminative performance was achieved by the multimodal configuration combining CNN embeddings with handcrafted CT descriptors (AUC = 0.54).

However, a critical observation arises from the analysis of performance stability. While the *CNN + Handcrafted* configuration yields the highest mean AUC, it is characterized by a very high standard deviation (±0.25). This suggests that the multimodal integration within a neural framework is highly sensitive to the specific patient distribution in each fold. In comparison, the standalone *CNN (max pooling)* model achieves a similar AUC (0.53) but with a much lower standard deviation (±0.16), indicating a more consistent predictive signal. The ROC curves aggregated across cross-validation folds for the best-performing configuration are shown in Figure 4.1.

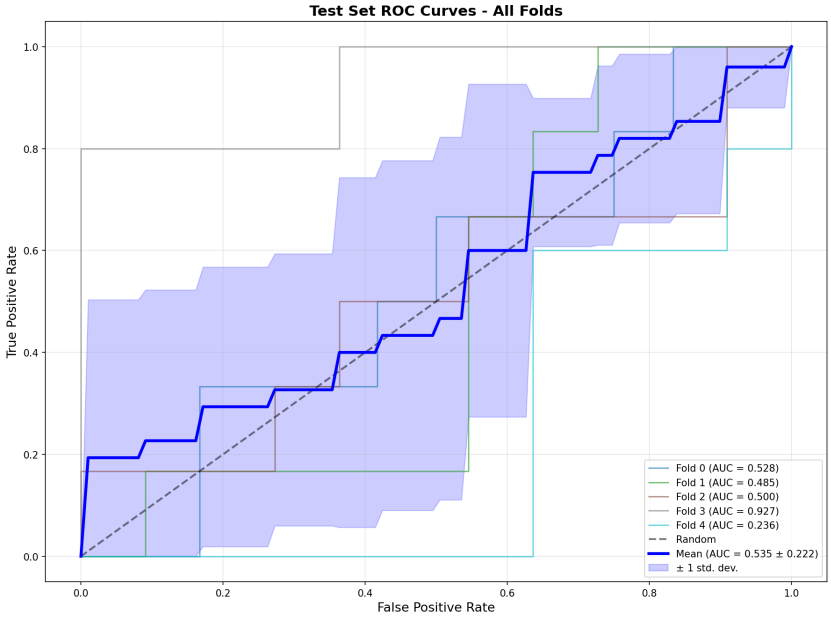

Figure 4.1: Aggregate ROC curve across the five cross-validation folds for the best-performing classification model (CNN + Handcrafted).

CNN Feature Aggregation Study

The second experimental block evaluated deep imaging representations using different pooling strategies. Among CNN-only models, max pooling produced the highest discriminative ability (AUC = 0.53), outperforming mean pooling (AUC = 0.49). This result suggests that localized high-activation patterns detected in specific slices may contain stronger predictive signals for progression than globally averaged representations.

Multimodal Fusion and the Regularizing Effect of Demographics

Combining CNN embeddings with handcrafted descriptors resulted in a marginal improvement compared to CNN features alone (AUC 0.54 vs 0.53). This suggests that handcrafted features provide limited complementary structural information beyond the deep features extracted by the CNN.

Interestingly, the inclusion of demographic variables (age, sex, and smoking status) revealed a Bias-Variance Tradeoff. When adding demographics to the multimodal representation, the mean AUC decreased to 0.46, but the standard deviation for accuracy dropped sharply from ± 0.20 to ± 0.05. This indicates that demographic features act as a "variance stabilizer" or a form of implicit regularization. By providing consistent clinical anchors, these variables prevent the model from over-fitting to noisy imaging patterns, though this stability comes at the cost of reduced discriminative power.

Threshold and Fold Evaluation

The validation-derived thresholds (Figure 4.2) were found to be located in a similar region of the ROC space, suggesting that the classifier output probabilities are relatively concentrated. However, as shown in the example fold evaluation (Figure 4.3), the predicted probability distributions exhibit significant overlap between progression and non-progression cases. This overlap visually confirms the model's high uncertainty and the intrinsic difficulty of the task in a small-cohort setting.

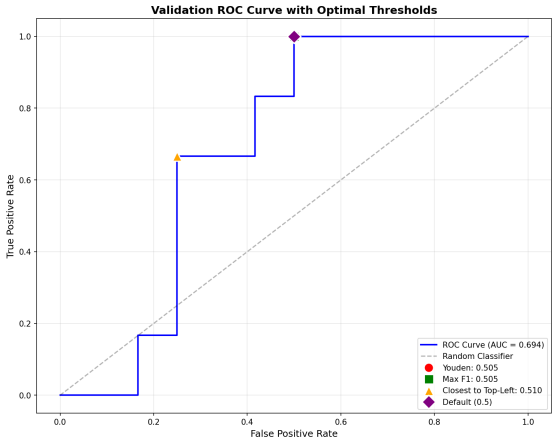

Figure 4.2: Validation ROC curve for a representative cross-validation fold. Markers indicate candidate decision thresholds obtained using different selection criteria, including Youden's statistic, maximum F1-score, and minimum distance to the top-left corner of the ROC space.

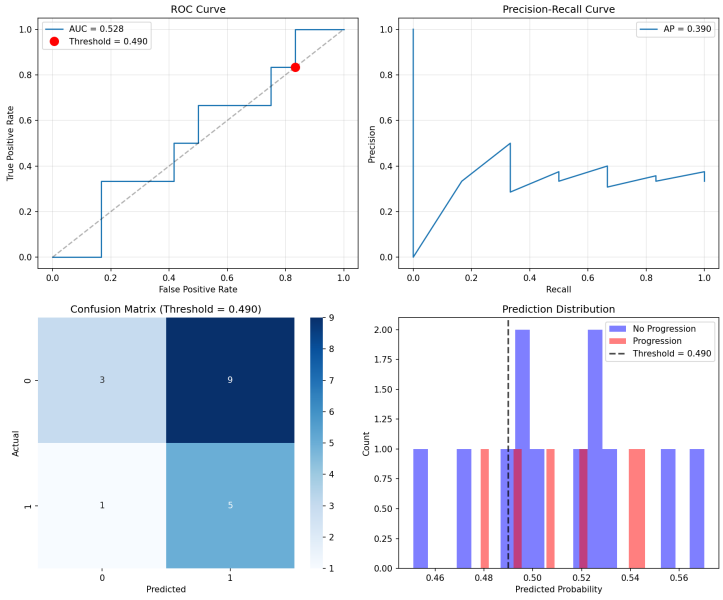

Figure 4.3: Example evaluation for a representative cross-validation fold. The probability distribution (bottom right) illustrates the considerable overlap between classes, reflecting the model's uncertainty.

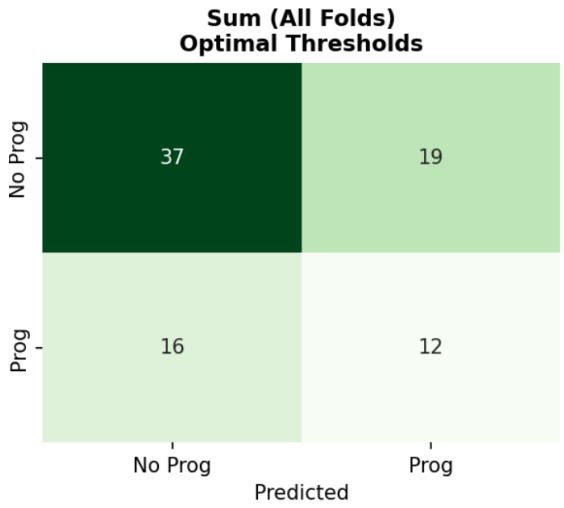

Figure 4.4: Confusion matrix for the test split with optimal threshold for best configuration (CNN + Handcrafted).

Transition to Gradient Boosting

In summary, the neural approach (MLP) demonstrated that imaging-derived features provide the most informative signal, yet the results remain hampered by high variance and limited AUC. These findings suggest that a standard neural architecture may not be optimal for integrating heterogeneous features with such a limited sample size. These limitations provide a direct motivation for the transition to the LightGBM framework in the following section, which is inherently more robust to feature scaling and noise in small-cohort tabular settings.

4.2 LightGBM Classification Results

In addition to the neural classifier described in the previous section, we evaluated a gradient-boosted decision tree model implemented using LightGBM. As described in Section 3.2.2, the model operates on patient-level tabular representations constructed by concatenating pooled CNN embeddings, handcrafted CT-derived descriptors, demographic variables, and optionally baseline pulmonary function (FVC at baseline).

All experiments follow the same patient-level 5-fold cross-validation protocol used for the neural classifier. Performance metrics are reported as the mean and standard deviation across folds.

Overall Performance

Table 4.2 summarizes the cross-validated performance across the evaluated feature configurations.

Table 4.2: Cross-validated classification performance across feature configurations. Values are reported as mean ± standard deviation across folds.

Configuration	Accuracy	Precision	Recall	Specificity	F1	AUC
Handcrafted	0.57 ± 0.09	0.37 ± 0.11	0.49 ± 0.19	0.61 ± 0.10	0.42 ± 0.14	0.60 ± 0.17
Handcrafted + Demographics	0.53 ± 0.09	0.27 ± 0.14	0.39 ± 0.25	0.60 ± 0.24	0.31 ± 0.17	0.55 ± 0.10
CNN (mean pooling)	0.58 ± 0.05	0.25 ± 0.20	0.42 ± 0.35	0.67 ± 0.22	0.31 ± 0.25	0.65 ± 0.08
CNN + Handcrafted + Demographics	0.60 ± 0.14	0.62 ± 0.31	0.43 ± 0.25	0.69 ± 0.31	0.39 ± 0.09	0.66 ± 0.10
CNN + Demographics	0.60 ± 0.11	0.44 ± 0.13	0.39 ± 0.20	0.70 ± 0.23	0.37 ± 0.10	0.55 ± 0.12
Handcrafted + FVC(0)	0.64 ± 0.17	0.54 ± 0.27	0.46 ± 0.17	0.73 ± 0.22	0.47 ± 0.18	0.62 ± 0.14
CNN + Handcrafted + FVC(0)	0.61 ± 0.12	0.34 ± 0.20	0.39 ± 0.25	0.72 ± 0.22	0.35 ± 0.21	0.64 ± 0.07
CNN + Handcrafted + Demo + FVC(0)	0.56 ± 0.11	0.39 ± 0.07	0.43 ± 0.25	0.63 ± 0.26	0.36 ± 0.11	**0.70 ± 0.05**

Overall, the LightGBM classifier achieved higher discriminative performance compared to the neural classifier presented earlier. The best-performing configuration combines CNN embeddings, handcrafted CT descriptors, demographic variables, and baseline pulmonary function, reaching a mean ROC AUC of approximately 0.70 across cross-validation folds.

Figure 4.5 shows the ROC curves obtained on the test split for each fold of the best-performing configuration.

These results indicate that gradient-boosted decision trees are able to effectively exploit heterogeneous tabular representations derived from imaging and clinical variables in this small-cohort setting. Interestingly, the inclusion of handcrafted radiomic descriptors provides only a limited improvement over CNN-based features alone. This suggests that the CNN embeddings already capture much of the structural information encoded by traditional handcrafted radiomic features. Such behaviour is consistent with previous findings in medical imaging, where deep

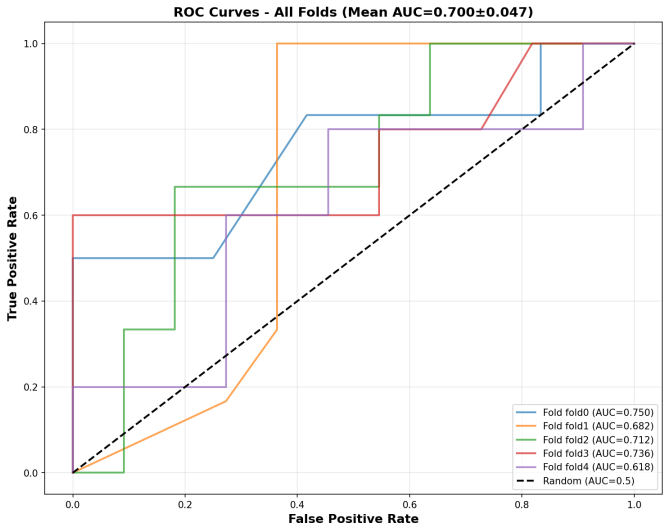

Figure 4.5: ROC curves across the five cross-validation folds for the best-performing LightGBM configuration.

convolutional representations often learn higher-level texture and morphology patterns that subsume classical radiomic descriptors.

Fold-wise Threshold Selection

During cross-validation, the decision threshold was not fixed a priori but selected independently for each fold using the validation split. Specifically, the optimal threshold was determined by maximizing Youden's J statistic on the validation ROC curve.

Figure 4.6 illustrates the threshold selection process for a representative cross-validation fold. The highlighted point corresponds to the threshold maximizing Youden's statistic on the validation ROC curve.

The fold-specific thresholds obtained from the validation sets were then applied to the corresponding test predictions to compute classification metrics.

Confusion Matrix Analysis

Figure 4.7 reports the confusion matrices obtained for each cross-validation fold together with the aggregated confusion matrix across folds.

The aggregated results show that the model correctly identifies a substantial portion of progression cases while maintaining moderate specificity. This reflects the inherent difficulty of the prediction task in a relatively small dataset.

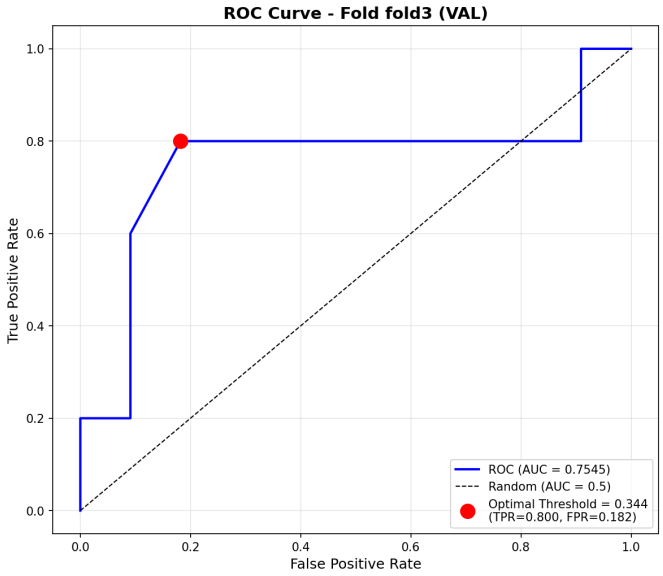

Figure 4.6: Validation ROC curve for a representative fold. The highlighted point indicates the threshold maximizing Youden's J statistic, which is subsequently applied to the test split of that fold.

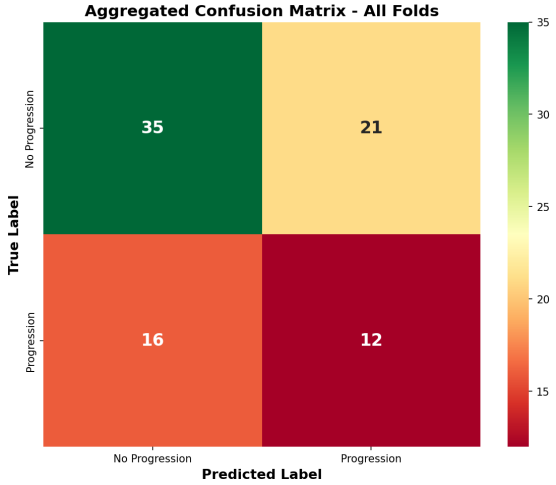

Figure 4.7: Confusion matrices for each cross-validation fold and the aggregated confusion matrix across all folds for the best-performing LightGBM configuration.

Threshold Sensitivity Analysis

In addition to the fold-wise threshold selection used during cross-validation, we performed a dedicated threshold sensitivity analysis for the best-performing LightGBM configuration (CNN embeddings + handcrafted descriptors + demographics

+ baseline FVC). This analysis was conducted only for this configuration because it achieved the highest cross-validated AUC and therefore represents the most relevant model for further inspection.

To characterize the behaviour of the classifier as a function of the decision threshold, we evaluated accuracy, precision, recall, specificity, and F1-score over a threshold grid ranging from 0.10 to 0.90. For each threshold value, predictions from all cross-validation folds were aggregated and the mean performance metrics were computed.

Figure 4.8 shows the evolution of the evaluation metrics as the classification threshold varies. The dashed vertical line indicates the threshold selected during validation ($\tau \approx 0.32$).

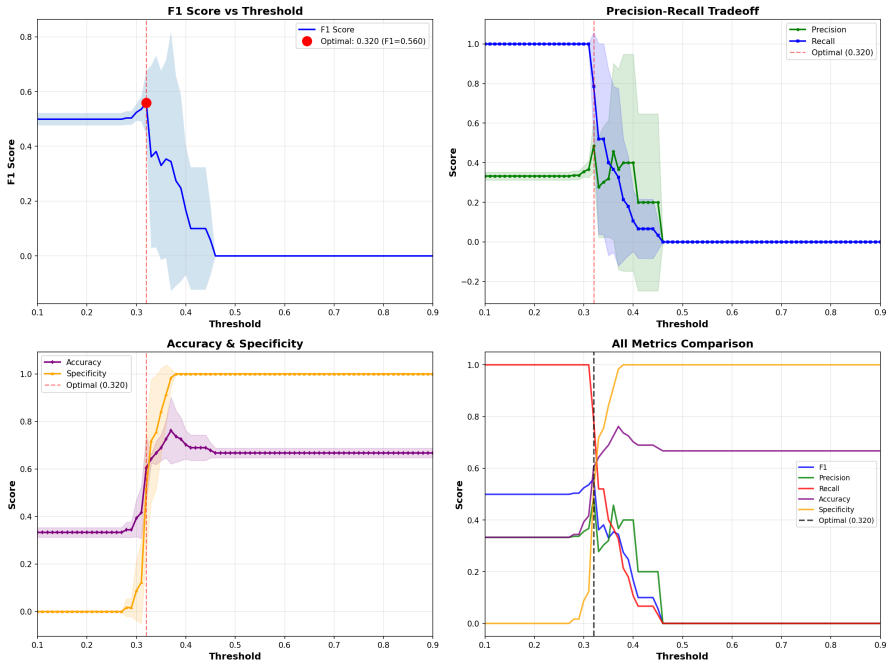

Figure 4.8: Performance metrics as a function of the decision threshold for the best-performing LightGBM configuration. The dashed vertical line indicates the operating threshold selected during validation ($\tau \approx 0.32$).

The threshold sweep reveals a clear trade-off between sensitivity and specificity. At very low thresholds, the classifier predicts progression for almost all patients, resulting in maximal recall but extremely low specificity. As the threshold increases, specificity improves while recall decreases progressively.

The selected operating point around $\tau \approx 0.32$ lies in the region that maximizes the mean F1-score while maintaining a balanced compromise between sensitivity and specificity. This threshold therefore represents a reasonable trade-off between

detecting progression events and limiting false positive predictions.

For completeness, Table 4.3 reports representative operating points illustrating how the classifier behaviour changes across different threshold regimes.

Table 4.3: Representative operating points from the threshold sensitivity analysis of the best-performing LightGBM configuration. Values correspond to cross-fold mean metrics.

Threshold	Accuracy	Precision	Recall	Specificity	F1
0.10	0.33	0.33	1.00	0.00	0.50
0.30	0.39	0.36	1.00	0.09	0.52
0.32	0.60	0.49	0.79	0.52	**0.56**
0.35	0.69	0.32	0.40	0.84	0.33
0.37	0.76	0.37	0.33	0.98	0.34
0.46	0.67	0.00	0.00	1.00	0.00

To further illustrate the practical impact of the threshold selection, Figure 4.9 reports the confusion matrices obtained across the cross-validation folds using the selected operating threshold ($\tau \approx 0.32$).

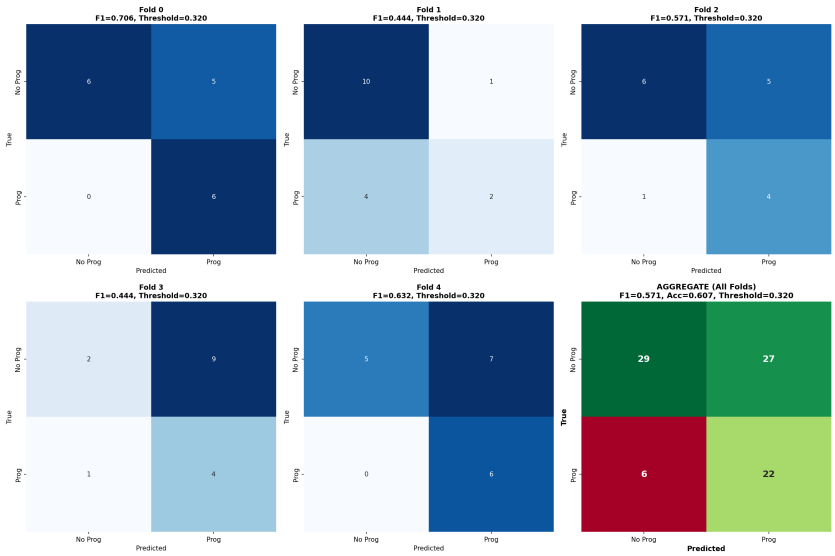

Figure 4.9: Confusion matrices across cross-validation folds for the best-performing LightGBM configuration using the validation-derived threshold ($\tau \approx 0.32$).

The confusion matrices confirm the behaviour observed in the threshold sweep analysis. Compared to very low thresholds that strongly favour sensitivity, the

selected threshold produces a more balanced distribution of predictions, reducing false positive predictions while still identifying a substantial proportion of progression cases. Although some variability across folds is expected due to the relatively small cohort size, the overall classification patterns remain consistent across cross-validation splits.

Model Explainability

To further investigate the contribution of different feature groups, SHAP analysis was performed as described in Section 3.2.2. SHAP values were computed on the test split of each fold and then aggregated across folds to obtain cross-validated explanations.

Figure 4.10 reports the mean absolute SHAP importance aggregated by feature group for the best-performing LightGBM configuration.

The results indicate that CNN-derived imaging embeddings contribute the largest share of the predictive signal. Baseline pulmonary function provides an additional but smaller contribution, while handcrafted radiomic descriptors and demographic variables have a more limited overall impact. This suggests that deep CNN embeddings capture structural patterns associated with disease progression, while clinical variables provide complementary information.

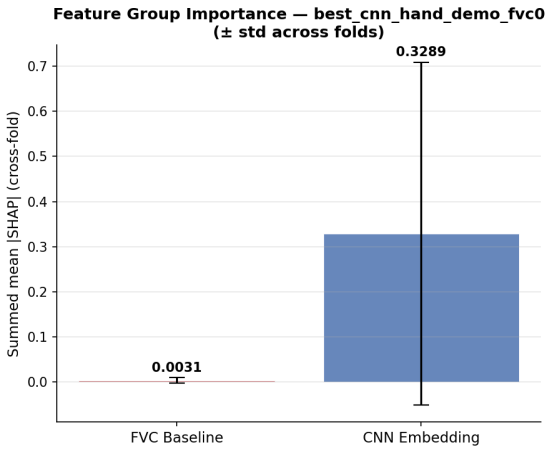

Figure 4.10: Cross-fold SHAP importance aggregated by feature group for the best-performing LightGBM configuration.

To better understand the role of handcrafted radiomic descriptors, we also analyzed SHAP explanations for the configuration using handcrafted features together with baseline pulmonary function.

Figure 4.11 shows the cross-validated SHAP importance of individual handcrafted features for this model.

The results highlight that baseline pulmonary function is the dominant predictor, providing the largest contribution to the model output. Among the handcrafted descriptors, only a small subset of features—such as the average number of tissue pixels and HU kurtosis—exhibit a noticeable impact on prediction. Most remaining radiomic descriptors show relatively small SHAP values, indicating limited predictive relevance.

Comparing these results with the feature-group SHAP analysis suggests that the information encoded by handcrafted radiomic descriptors is largely captured by the CNN embeddings. This may explain why adding handcrafted features to the multimodal model provides only modest improvements in predictive performance.

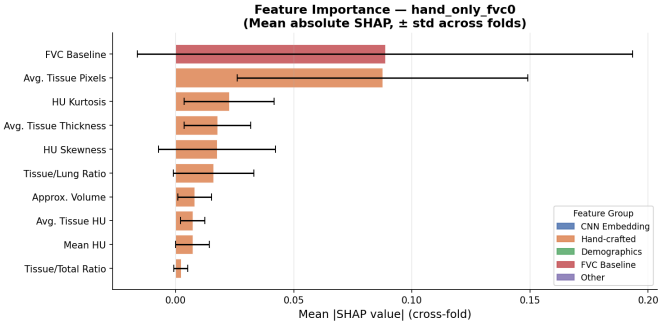

Figure 4.11: Cross-fold SHAP importance for the handcrafted + baseline FVC configuration. Error bars represent the standard deviation across folds. Baseline pulmonary function provides the strongest predictive signal, while only a subset of handcrafted radiomic descriptors contributes meaningfully to the prediction.

4.3 FVC Regression Results

In addition to binary progression prediction, we evaluated the ability of the proposed framework to estimate pulmonary function at one year through a regression task predicting FVC at 52 weeks. All experiments follow the patient-level 5-fold cross-validation protocol described in Section 3.2.3. Final metrics are computed on de-normalised predictions expressed in milliliters.

Overall Performance

Table 4.4 summarizes the predictive performance across the evaluated feature configurations.

Table 4.4: Cross-validated performance of regression models for predicting FVC at 52 weeks. Metrics are computed on de-normalised predictions (mL).

Configuration	MAE (mL)	RMSE (mL)	R^2
Handcrafted + FVC(0)	**254**	326	0.82
Handcrafted + Demographics + FVC(0)	260	319	**0.83**
CNN + Handcrafted + Demographics + FVC(0)	261	323	0.83
CNN (max+mean pooling) + FVC(0)	264	341	0.82
CNN + Handcrafted + FVC(0)	287	355	0.79
CNN (mean pooling) + FVC(0)	287	362	0.79
CNN (max pooling) + FVC(0)	295	362	0.78
Models without baseline FVC			
Handcrafted only	582	744	0.15
CNN + Handcrafted + Demographics (no FVC)	611	772	0.10
CNN only (no FVC)	619	781	0.05
CNN + Handcrafted (no FVC)	619	785	0.06

Overall, the regression models achieve strong predictive performance when baseline pulmonary function is included as an input feature. The best-performing configuration combines handcrafted CT-derived descriptors with baseline FVC, reaching a mean absolute error of approximately 254 mL and an R^2 of 0.82 across cross-validation folds.

Effect of Baseline FVC

A clear performance gap emerges between models that include baseline FVC and those that do not. When baseline FVC is removed from the feature set, prediction accuracy deteriorates substantially, with MAE values exceeding 580 mL and R^2 dropping below 0.15.

This result highlights the strong predictive value of baseline pulmonary function for estimating future FVC levels. Baseline FVC provides an anchor for the regression task, allowing the model to learn patterns of functional decline relative to each

patient's initial lung capacity. This behaviour is expected clinically, as lung function trajectories in fibrotic lung diseases tend to follow patient-specific decline patterns relative to baseline measurements.

Impact of Imaging and Multimodal Features

When baseline FVC is included, adding CNN-derived imaging features or handcrafted descriptors provides only moderate improvements over simpler models. The full multimodal configuration (CNN + handcrafted + demographics + baseline FVC) achieves performance comparable to the best tabular model, with an MAE of approximately 261 mL and R^2 around 0.83.

These findings suggest that while imaging-derived features contain relevant structural information about lung parenchyma, baseline pulmonary function already explains a large fraction of the variance in future FVC measurements. As a result, additional modalities provide only limited incremental predictive benefit in this dataset. In addition, structural imaging features may primarily reflect disease severity already captured by pulmonary function measurements, which could explain the limited incremental gain obtained when adding imaging-derived descriptors in this dataset.

Prediction Error Analysis

To further investigate the behaviour of the regression model, we report diagnostic plots for the best-performing configuration in Figure 4.12. These plots illustrate the agreement between predicted and observed FVC values and provide insight into the distribution of prediction errors.

The predicted-versus-true scatter plot shows a strong linear relationship between predicted and observed FVC values, with most samples lying close to the identity line. This indicates that the model is able to capture the general trend of pulmonary function trajectories.

The residual plot shows no clear systematic pattern across the range of FVC values, suggesting that prediction errors are relatively evenly distributed. The residual distribution is approximately centred around zero, indicating limited overall prediction bias.

Finally, the Bland–Altman analysis confirms good agreement between predicted and observed values, with most samples falling within the limits of agreement. Larger deviations are mainly observed for patients with more extreme FVC values, reflecting the inherent difficulty of modelling individual disease trajectories in a small cohort.

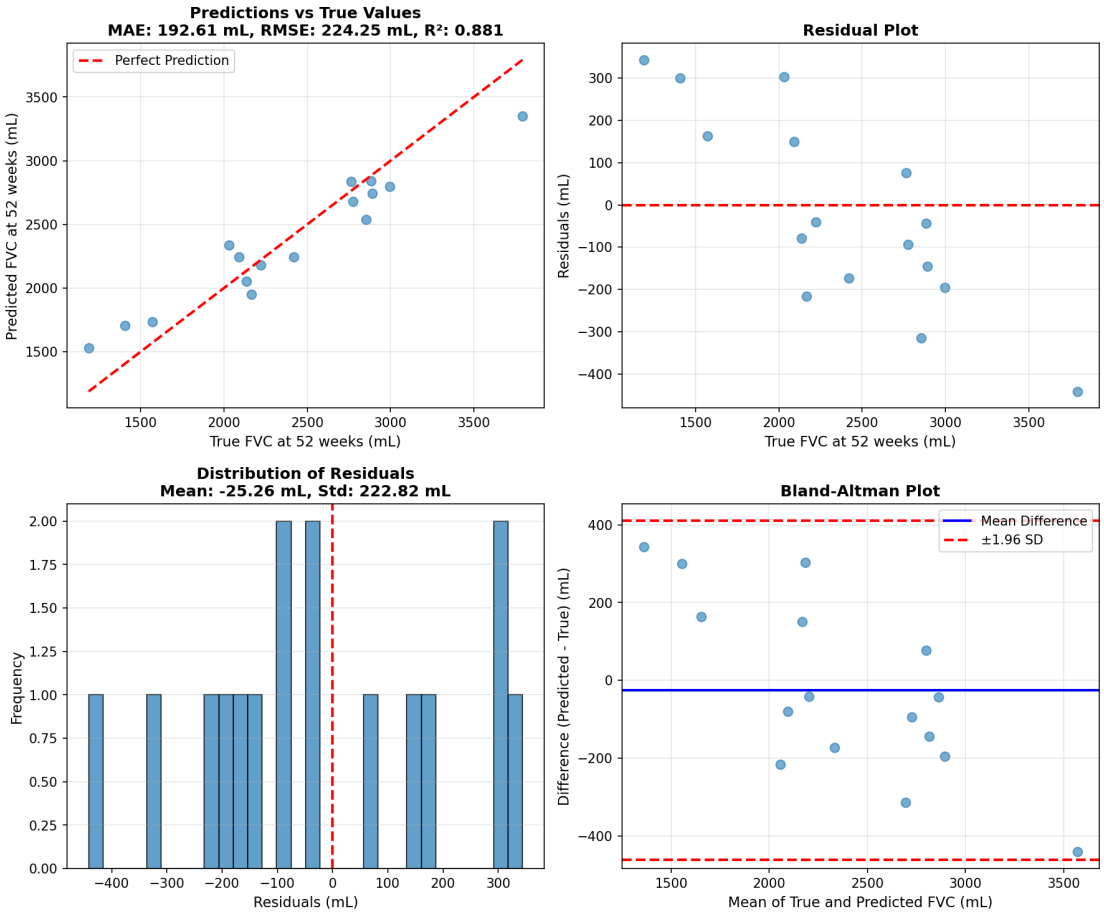

Figure 4.12: Diagnostic plots for the best-performing regression configuration. Top-left: predicted versus true FVC values at 52 weeks with the identity line. Top-right: residuals plotted against the true FVC values. Bottom-left: distribution of residuals. Bottom-right: Bland–Altman plot showing agreement between predicted and observed values.

4.4 Survival Analysis Results

We evaluated the Cox proportional hazards model using multiple feature configurations to assess the ability of CT-derived features, clinical variables, and their combinations to predict time to disease progression. Performance was measured using the concordance index (C-index), which quantifies the model's ability to correctly rank patients according to their progression risk.

Table 4.5 summarizes the cross-validated results for the most representative configurations from the ablation study.

Table 4.5: Cross-validated Cox proportional hazards performance across selected feature configurations. Values are reported as mean ± standard deviation across folds.

Configuration	Validation C-index	Test C-index
CNN (4-stat features)	0.628 ± 0.150	**0.652 ± 0.120**
CNN + Sex	0.622 ± 0.101	0.626 ± 0.129
Handcrafted (Kurtosis)	0.624 ± 0.063	0.624 ± 0.063
Handcrafted + Sex	0.617 ± 0.113	0.571 ± 0.070
CNN + Handcrafted + Sex	0.624 ± 0.089	0.623 ± 0.127
Demographics (Age only)	0.389 ± 0.090	0.477 ± 0.141

Overall, models incorporating CNN-derived radiomic descriptors achieved the highest concordance scores, with the configuration based on aggregated CNN statistics obtaining the best performance (test C-index = 0.652 ± 0.120). In contrast, demographic variables alone showed limited predictive power, with substantially lower concordance values.

Combining CNN features with handcrafted radiomic descriptors or demographic variables produced comparable but not substantially improved results, suggesting that the CNN-derived descriptors already capture much of the relevant structural information associated with disease progression risk.

Risk Stratification

To assess the clinical interpretability of the Cox model, patients were stratified into two groups according to the median predicted risk score. Kaplan–Meier survival curves were then estimated for the high-risk and low-risk groups.

Figure 4.13 shows the resulting survival curves for a representative cross-validation fold. The two groups exhibit clearly different progression trajectories, with the high-risk group experiencing earlier progression events and a steeper decline in progression-free probability over time. This separation indicates that the model is able to identify patients with different risk profiles.

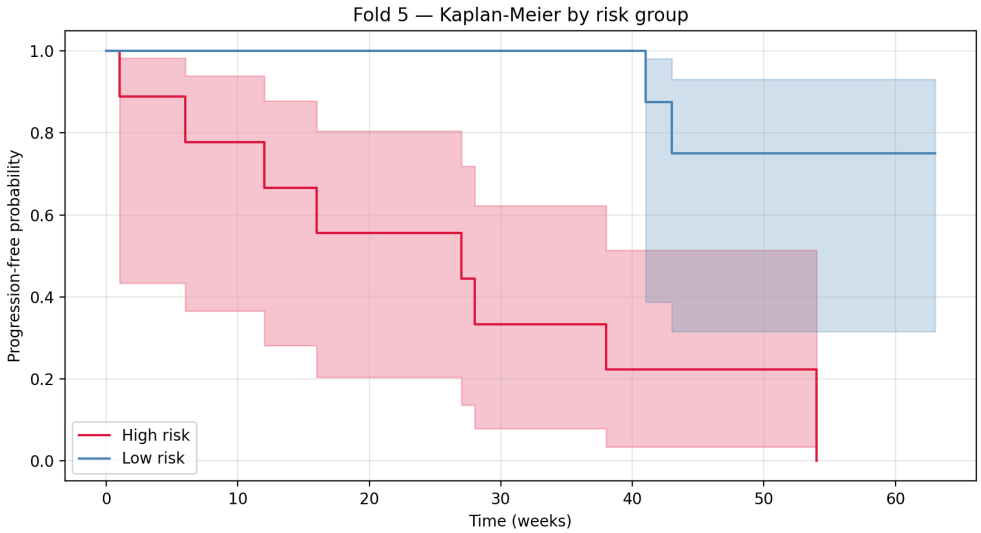

Figure 4.13: Kaplan–Meier survival curves obtained by stratifying patients into high- and low-risk groups based on the median predicted risk from the Cox model. The separation between the two curves indicates that the model captures meaningful differences in progression risk.

Model Interpretation

To further investigate which imaging features contribute to the estimated progression risk, we examined the hazard ratios derived from the fitted Cox model.

Figure 4.14 illustrates the hazard ratios for the CNN-derived radiomic descriptors in a representative fold. The hazard ratios are generally close to one, indicating that individual features exert relatively modest effects on the estimated risk. This behaviour is expected given the regularized Cox model and the limited sample size. Nevertheless, the combination of multiple radiomic descriptors enables the model to capture subtle variations in lung tissue characteristics that correlate with disease progression risk.

Overall, these results demonstrate that survival modeling can provide complementary insights beyond binary classification by capturing the temporal dynamics of disease progression and enabling clinically interpretable risk stratification.

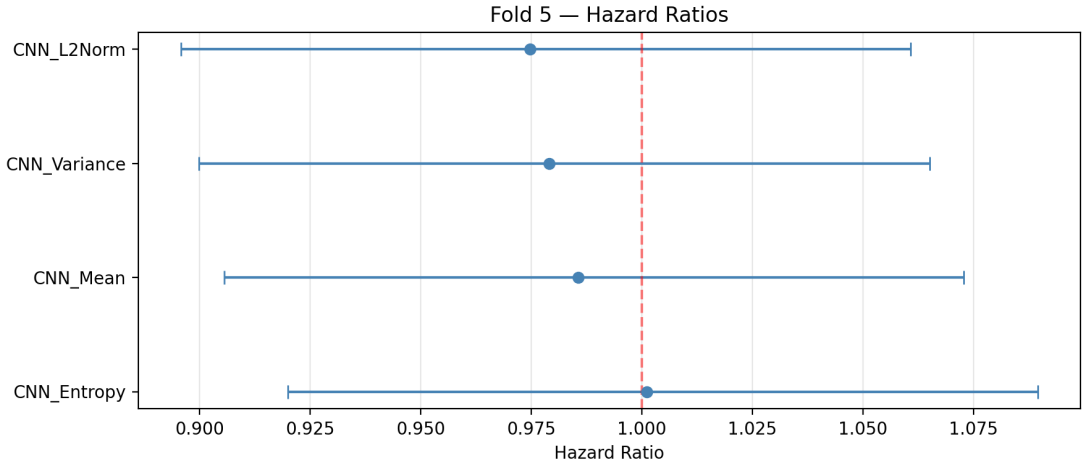

Figure 4.14: Estimated hazard ratios for CNN-derived radiomic features in a representative cross-validation fold. Error bars indicate confidence intervals. Values above 1 correspond to increased progression risk, while values below 1 indicate protective effects.

Chapter 5

Discussion

This work investigated the use of multimodal machine learning approaches to predict disease progression in idiopathic pulmonary fibrosis (IPF) by integrating imaging-derived features from chest CT scans with clinical and demographic information. Three complementary prediction tasks were explored: binary progression classification, continuous prediction of pulmonary function decline (FVC at 52 weeks), and survival analysis modeling time to progression.

Overall, the results indicate that imaging-derived representations extracted from CT scans provide meaningful predictive information and can capture structural patterns associated with disease progression.

Predictive Performance Across Tasks

Across the three predictive tasks, models incorporating CT-derived features consistently achieved the strongest performance. In the classification task, the best configuration obtained an average AUC of approximately 0.70, indicating a moderate ability to distinguish between patients who will experience progression within one year and those who will not. While this performance is not sufficient for direct clinical deployment, it demonstrates that meaningful progression signals can be extracted from baseline imaging data.

The regression analysis yielded particularly strong results. The best multimodal configuration achieved an R^2 value above 0.82 with a mean absolute error of approximately 260 mL when predicting FVC at 52 weeks. These results suggest that baseline imaging and clinical variables contain substantial information about future pulmonary function decline. Importantly, models that excluded baseline FVC exhibited a large performance degradation, highlighting the strong predictive value of the initial pulmonary function measurement.

In the survival analysis task, the Cox proportional hazards model achieved a concordance index of approximately 0.65. This result is consistent with previous

studies in small IPF cohorts, where survival modeling is particularly challenging due to limited sample size and censoring effects. Nevertheless, the Kaplan–Meier analysis demonstrated clear separation between predicted high-risk and low-risk groups, suggesting that the model captures clinically meaningful differences in progression trajectories.

Taken together, these results indicate that the different modeling approaches capture complementary aspects of disease progression. Regression models appear particularly effective at predicting continuous functional decline, while classification and survival analysis provide interpretable risk stratification.

Role of Imaging Features

An important observation across experiments is the strong contribution of imaging-derived features. In both the classification and survival tasks, configurations incorporating CNN-based radiomic descriptors achieved the highest performance. These features likely capture structural characteristics of fibrotic lung tissue that are not fully reflected in traditional clinical variables.

Interestingly, combining CNN features with handcrafted radiomic descriptors and demographic variables did not consistently produce large improvements. This suggests that the deep CNN embeddings already encode much of the relevant information contained in the handcrafted features, potentially reducing redundancy between feature groups.

The survival analysis results further support this observation. Models based on aggregated CNN statistics achieved the highest concordance scores, while demographic variables alone showed limited predictive power. This finding aligns with the clinical understanding that imaging biomarkers can provide direct information about the spatial distribution and severity of fibrotic changes.

Clinical Interpretation

From a clinical perspective, the ability to identify patients at higher risk of disease progression could support earlier treatment decisions and improved patient monitoring. The Kaplan–Meier risk stratification analysis demonstrated that the Cox model successfully separates patients into groups with distinct progression trajectories. Although the absolute predictive performance remains moderate, this type of risk stratification could still provide useful decision-support information in clinical settings.

Furthermore, the regression results suggest that predicting continuous pulmonary function decline may be a particularly promising direction. Since FVC decline is a key clinical endpoint in IPF trials, accurate prediction of future FVC values could potentially support individualized prognosis estimation.

Limitations

Several limitations should be considered when interpreting these results. First, the dataset size is relatively small, which limits the complexity of models that can be reliably trained and increases the risk of overfitting. This constraint also affects survival analysis, where the number of observed events determines the stability of hazard estimates.

Second, the study relies on a single dataset without external validation. Although cross-validation was used to estimate model performance, evaluating the proposed models on independent cohorts would be necessary to assess generalization.

Third, imaging features were extracted from pre-trained CNN models without task-specific fine-tuning on medical imaging datasets. While transfer learning enables efficient feature extraction, domain-specific pretraining or self-supervised learning approaches may improve the quality of learned representations.

Future Work

Future research could extend this work in several directions. First, larger multi-center datasets would allow training more expressive models and enable external validation of the proposed approaches. Second, incorporating longitudinal imaging information may provide a more complete picture of disease progression dynamics. Third, more advanced survival models, such as deep survival networks, could be explored to capture nonlinear relationships between imaging features and progression risk.

Additionally, integrating explainability techniques may help identify imaging patterns associated with rapid disease progression, potentially improving the clinical interpretability of the models.

Chapter 6

Conclusion

This thesis explored the use of multimodal machine learning approaches to model disease progression in idiopathic pulmonary fibrosis (IPF) by integrating imaging-derived information from chest CT scans with clinical and demographic variables. Three complementary predictive tasks were investigated: binary classification of one-year disease progression, regression of pulmonary function decline (FVC at 52 weeks), and survival analysis modeling time to progression.

The experimental results demonstrate that CT-derived imaging features contain meaningful predictive signals related to disease evolution. In particular, regression models achieved strong performance when predicting future pulmonary function, highlighting the strong predictive value of baseline clinical measurements. At the same time, classification and survival models showed that imaging-derived CNN features can capture structural patterns associated with disease progression and enable clinically interpretable risk stratification.

Across tasks, CNN-derived representations consistently provided the most informative imaging features, while handcrafted radiomic descriptors and demographic variables contributed only limited additional predictive value. These findings suggest that deep learning representations extracted from CT imaging can implicitly capture structural characteristics of fibrotic lung disease that are relevant for prognosis.

Although the predictive performance is constrained by the relatively small cohort size, the results confirm the feasibility of multimodal machine learning approaches for modeling IPF progression. Importantly, the experiments highlight the complementary roles of imaging-derived biomarkers and pulmonary function measurements in describing disease trajectories.

Future work should focus on evaluating these approaches on larger multi-center datasets, incorporating longitudinal imaging information, and exploring more advanced representation learning techniques tailored to medical imaging data. Such developments could improve the robustness and clinical applicability of predictive

models for IPF progression.

Overall, this work provides a systematic evaluation of multimodal machine learning strategies for IPF progression prediction and highlights the potential of combining imaging and clinical data to support more personalized prognosis and monitoring in interstitial lung diseases.

Appendix A

Additional Experimental Details

A.1 Model Hyperparameters

This section summarizes the main hyperparameters used for the machine learning models evaluated in this study. All models were trained using the same 5-fold cross-validation protocol described in the main text. Hyperparameters were selected empirically to ensure stable training in the small-cohort setting.

MLP Classification Model Architecture

The neural classification model is implemented as a lightweight multilayer perceptron (MLP) operating on pooled CNN embeddings extracted from a pre-trained ResNet50 backbone. Slice-level features are aggregated at the patient level using max pooling and combined with optional handcrafted and demographic variables.

The architecture is intentionally compact to reduce overfitting in the small-cohort regime. Table A.1 summarizes the main architectural components.

Table A.1: Architecture of the MLP classification model.

Component	**Configuration**
Backbone CNN	ResNet50 (pre-trained feature extractor)
Input image size	224×224
Pooling strategy	Max pooling across slices
CNN embedding reduction	$2048 \rightarrow 32$ (Linear + ReLU + Dropout 0.3)
Classifier architecture	$(32 + \text{handcrafted} + \text{demographics}) \rightarrow 32 \rightarrow 1$
Hidden layers	[32]
Dropout	0.3

Training Configuration

Model training was performed using the AdamW optimizer with early stopping based on validation performance. The main training parameters are summarized in Table A.2.

Table A.2: Training configuration of the MLP classifier.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	3.86×10^{-5}
Weight decay	0.003
Batch size	16
Maximum epochs	60
Early stopping patience	15
Gradient clipping	max_norm = 1.0
Label smoothing	0
Input normalization	Standard scaling
Random seed	42

Learning Rate Scheduler

Training uses a ReduceLROnPlateau scheduler monitoring validation AUC. The learning rate is reduced when validation performance stagnates.

Table A.3: Learning rate scheduler configuration.

Parameter	Value
Scheduler	ReduceLROnPlateau
Monitored metric	Validation AUC
Mode	max
Reduction factor	0.5
Patience	5 epochs
Minimum learning rate	1×10^{-6}

Loss Function and Threshold Selection

Binary classification was trained using the `BCEWithLogitsLoss`. To mitigate class imbalance, a positive class weight was computed from the training distribution within each fold.

During evaluation, the decision threshold was selected on the validation split by maximizing Youden's J statistic on the ROC curve. The selected threshold was then applied to the corresponding test predictions.

LightGBM Model Configuration

This section provides additional implementation details for the LightGBM classifier used in the classification experiments. The model operates on patient-level feature vectors obtained by concatenating pooled CNN embeddings, handcrafted CT-derived descriptors, demographic variables, and baseline pulmonary function (FVC at baseline).

The CNN embeddings were extracted from a pre-trained ResNet50 backbone and aggregated at the patient level using mean pooling across slices. All feature vectors were standardized using a `StandardScaler` fitted on the training split within each cross-validation fold to avoid information leakage.

Table A.4 summarizes the main configuration parameters used during training.

Table A.4: Fixed configuration of the LightGBM classifier.

Parameter	**Value**
Model	LightGBM (Gradient Boosted Decision Trees)
Objective	Binary classification (cross-entropy)
Evaluation metric	AUC
CNN backbone	ResNet50 (pre-trained)
CNN pooling	Mean pooling across slices
Learning rate	0.01
Class weighting	Balanced
Maximum boosting rounds	1000
Early stopping rounds	100
Normalization	StandardScaler (fit on training split)
PCA	Optional (95% explained variance, max 50 components)
Cross-validation	5-fold stratified
Threshold selection	Youden's J statistic (validation set)
Random seed	42

Hyperparameter Search Space

Hyperparameters were selected using a grid search performed independently within each cross-validation fold. The search space was designed to favor highly regularized tree structures in order to reduce overfitting in the small-cohort setting.

Table A.5 summarizes the explored hyperparameter ranges.

Best Hyperparameters per Fold

The grid search evaluated 1,458 configurations per fold. Table A.6 reports the best-performing hyperparameter configuration selected for each cross-validation fold in the best-performing experiment (*CNN + Handcrafted + Demographics + FVC(0)*).

Table A.5: Hyperparameter search space for the LightGBM model.

Parameter	**Search Range**
num_leaves	[3, 5, 7]
max_depth	[2, 3]
min_data_in_leaf	[10, 15, 20]
λ_1 (L1 regularization)	[0.0, 0.5, 1.0]
λ_2 (L2 regularization)	[0.0, 0.5, 1.0]
feature_fraction	[0.6, 0.7, 0.8]
bagging_fraction	[0.6, 0.7, 0.8]

Table A.6: Best LightGBM hyperparameters selected per cross-validation fold.

Fold	num_leaves	max_depth	min_data_leaf	λ_1	λ_2	feat_frac	bag_frac
0	3	2	15	0.0	0.0	0.6	0.7
1	3	2	10	0.0	0.0	0.6	0.6
2	3	2	10	0.0	0.0	0.8	0.8
3	3	2	10	0.0	0.5	0.6	0.6
4	3	2	10	1.0	0.5	0.8	0.7

Regression Model Configuration

This section provides additional implementation details for the regression model used to predict pulmonary function at 52 weeks. The model operates on patient-level representations obtained by combining pooled CNN embeddings extracted from CT scans, baseline pulmonary function, handcrafted CT-derived descriptors, and demographic variables.

The regression network is implemented as a multi-branch MLP designed to process heterogeneous feature groups separately before multimodal fusion. All input features were standardized within each cross-validation fold using statistics computed on the training split only in order to avoid information leakage.

The architecture was intentionally regularized through high dropout rates, batch normalization, and early stopping in order to reduce overfitting in the small-cohort setting.

Table A.7: Training configuration of the regression model.

Hyperparameter	**Value**
Backbone CNN	ResNet50 (pre-trained)
Input image size	224×224
Pooling strategy	Max + Mean concatenation
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	0.05
Loss function	MSELoss
Batch size	8
Maximum epochs	100
Early stopping patience	20
Gradient clipping	max_norm = 1.0
Random seed	42
Input normalization	StandardScaler (per fold)
Scheduler	ReduceLROnPlateau
Scheduler factor	0.5
Scheduler patience	5 epochs
Minimum learning rate	1×10^{-6}

Table A.8: Architecture of the multi-branch regression model for the best CNN (max+mean pooling).

Component	**Configuration**
FVC branch	Linear($1 \to 64$) + ReLU + Dropout(0.21) + Linear($64 \to 128$) + ReLU
CNN branch	Linear($4096 \to 512$) + BatchNorm + ReLU + Dropout(0.7) + Linear($512 \to 256$) + BatchNorm + ReLU
Handcrafted branch	Linear($9 \to 32$) + BatchNorm + ReLU + Dropout(0.35) + Linear($32 \to 64$) + ReLU
Demographics branch	Linear($3 \to 16$) + ReLU + Dropout(0.21) + Linear($16 \to 32$) + ReLU
Feature attention	Linear($fusion \to fusion/4$) + ReLU + Linear($fusion/4 \to fusion$) + Sigmoid
Fusion MLP	Linear($fusion \to 256$) + BatchNorm + ReLU + Dropout(0.7) + Linear($256 \to 128$) + BatchNorm + ReLU + Dropout(0.49) + Linear($128 \to 1$)

Cox Proportional Hazards Survival Model

This section summarizes the configuration of the Cox proportional hazards model used for the survival analysis experiments. The model was implemented using the `lifelines` Python library and trained to predict time to disease progression while accounting for right-censored observations.

Given the relatively small cohort size, the Cox model was regularized using ridge penalization to stabilize coefficient estimation and reduce overfitting. Imaging features were optionally included through pooled CNN representations summarized using statistical descriptors.

Table A.9: Configuration of the Cox proportional hazards survival model for the best configuration.

Parameter	**Value**
Model	Cox Proportional Hazards (lifelines CoxPHFitter)
Penalizer	0.5
L1 ratio	0.0 (pure ridge regularization)
CNN backbone	ResNet50 (pre-trained)
Image size	224×224
CNN pooling	Mean pooling across slices
CNN statistics	Mean, Variance, L2 norm, Entropy
CNN PCA	Disabled (statistics used instead)
Cross-validation	5-fold

A.2 Additional MLP Results

Validation ROC Curves

Figure A.1 shows the ROC curves obtained on the validation split for each cross-validation fold of the MLP classifier in the *CNN + Handcrafted features* configuration. The optimal decision threshold is selected by maximizing Youden's J statistic on the validation set.

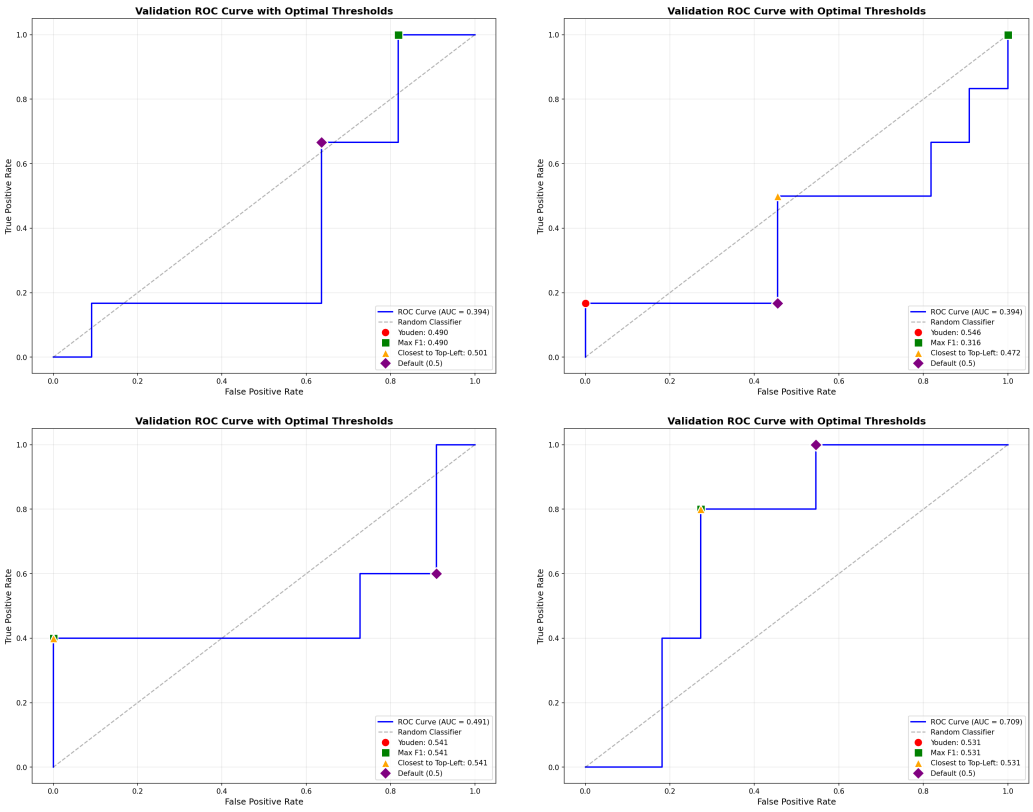

Figure A.1: Validation ROC curves for the four cross-validation folds of the MLP classifier in the *CNN + Handcrafted features* configuration. The optimal decision threshold is selected by maximizing Youden's J statistic on the validation set.

Test Evaluation per Fold

Figure A.2 and Figure A.3 report the evaluation results obtained on the test split for each cross-validation fold. The decision threshold selected on the validation set is applied to the corresponding test predictions to compute the final classification metrics.

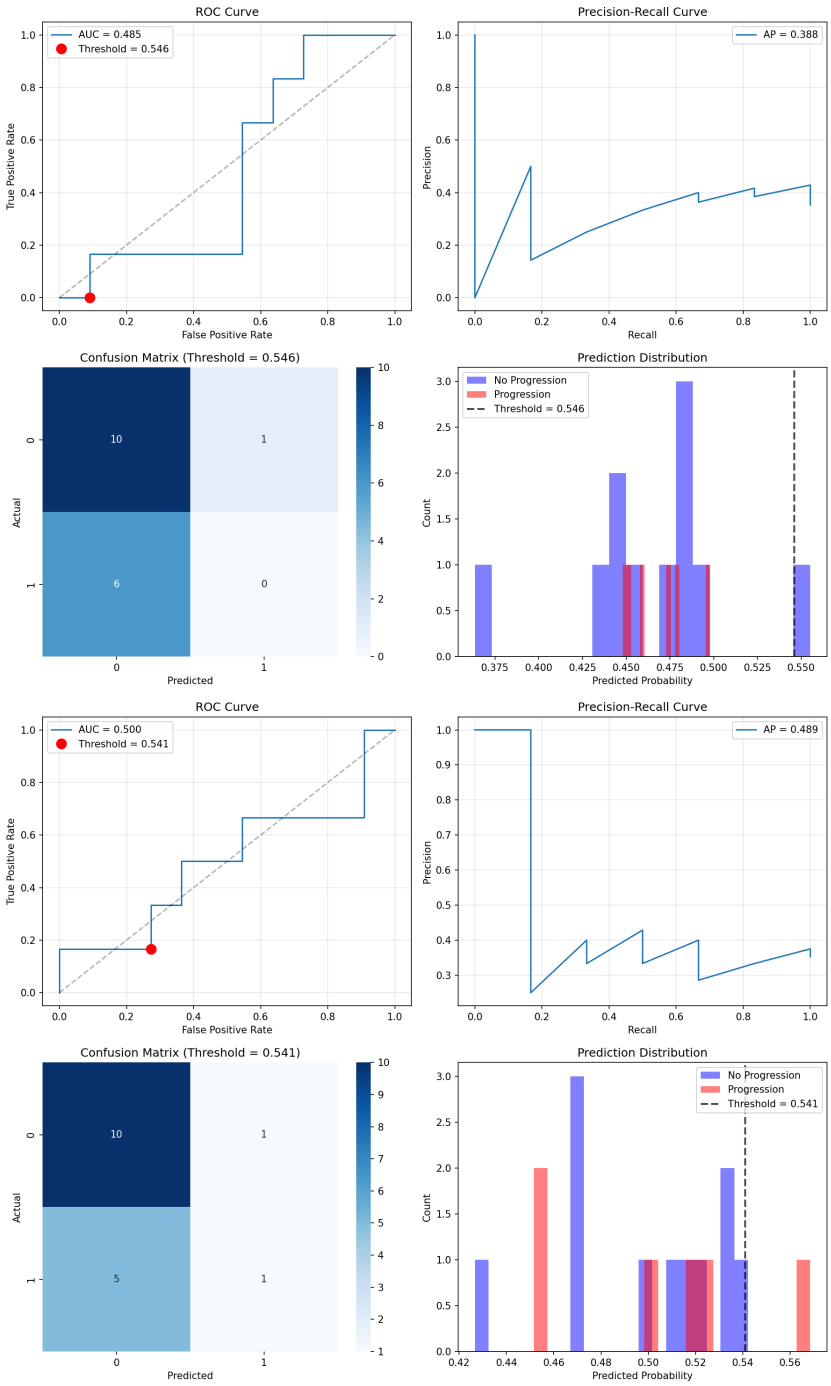

Figure A.2: Test evaluation plots for folds 1 and 2 of the MLP classifier in the *CNN + Handcrafted features* configuration.

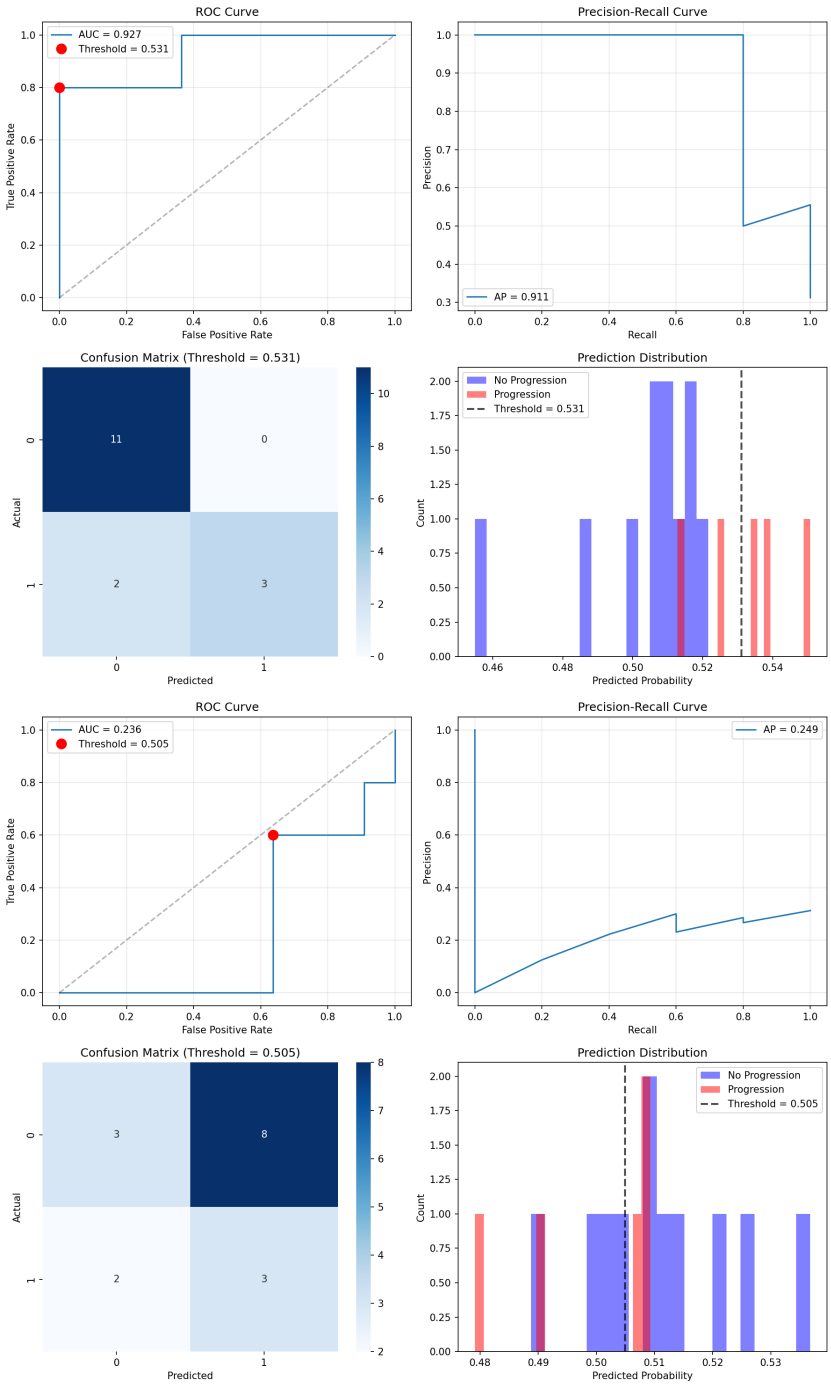

Figure A.3: Test evaluation plots for folds 3 and 4 of the MLP classifier in the *CNN + Handcrafted features* configuration.

A.3 Additional LightGBM Results

This section reports additional diagnostic plots for the LightGBM classifier. In particular, we show the ROC curves obtained on both the validation and test splits across the cross-validation folds.

Validation ROC Curves

Figure A.4 shows the ROC curves obtained on the validation split for each cross-validation fold of the LightGBM classifier.

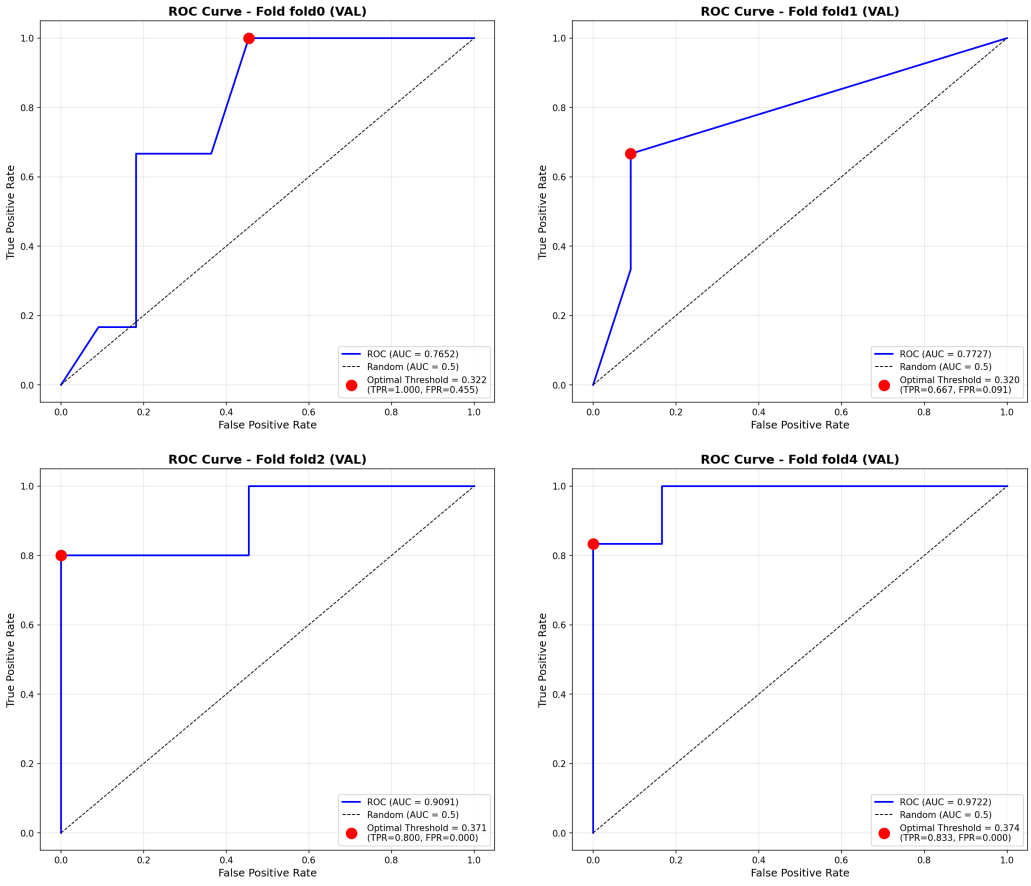

Figure A.4: Validation ROC curves for the LightGBM classifier across cross-validation folds.

Test ROC Curves

Figure A.5 reports the ROC curves obtained on the test split for each cross-validation fold.

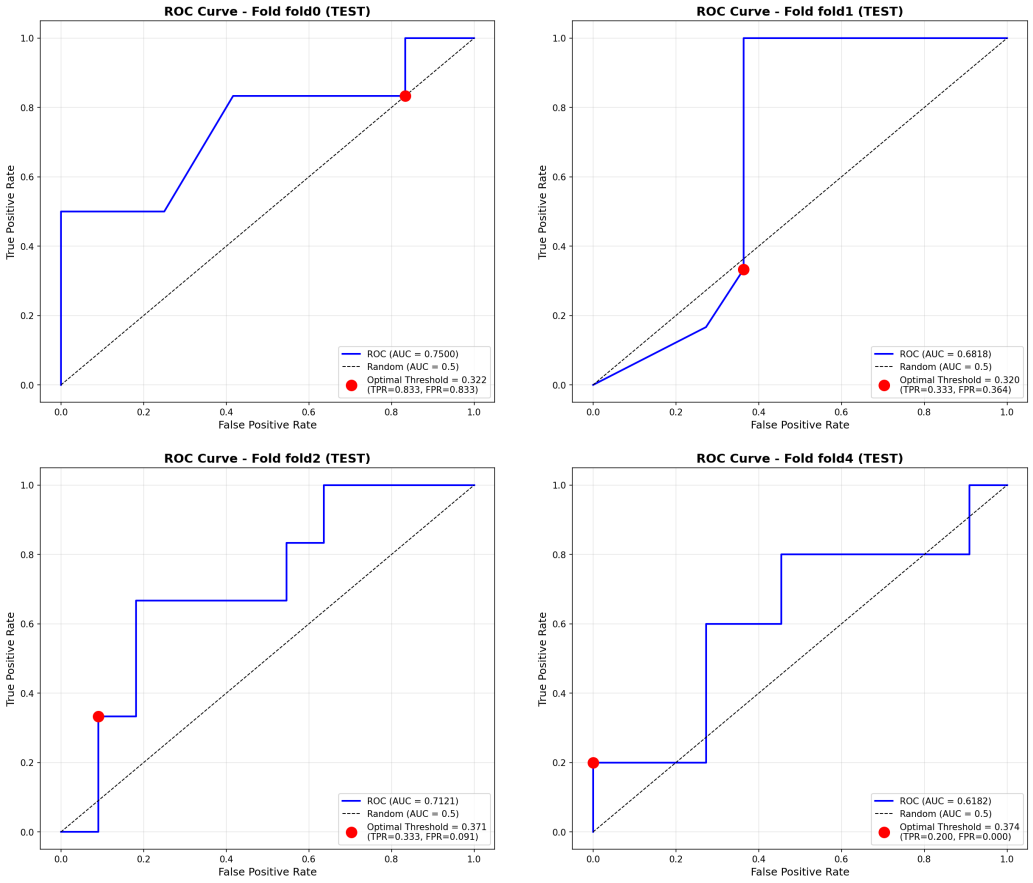

Figure A.5: Test ROC curves for the LightGBM classifier across cross-validation folds.

A.4 Additional Regression Results

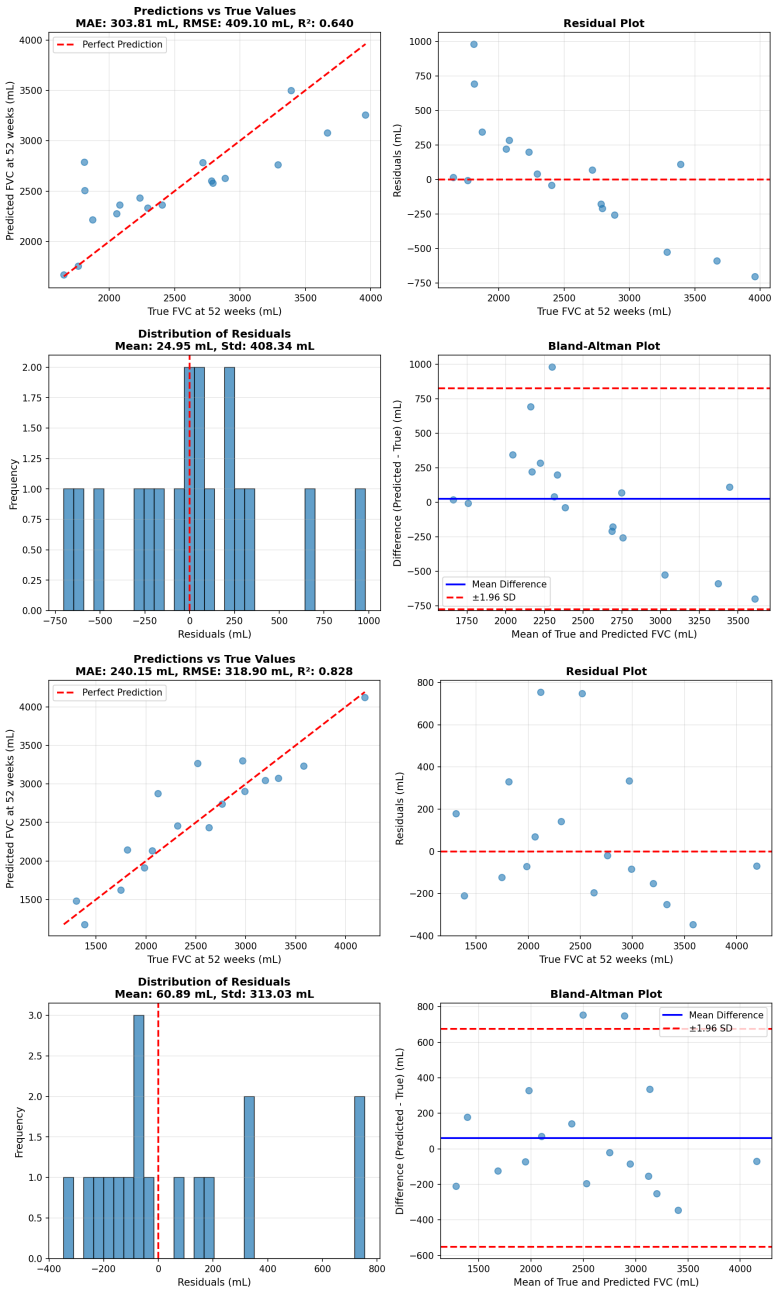

Figure A.6: Test evaluation plots for folds 1 and 2 of the regression model in the *Handcrafted features + FVC(0)* configuration.

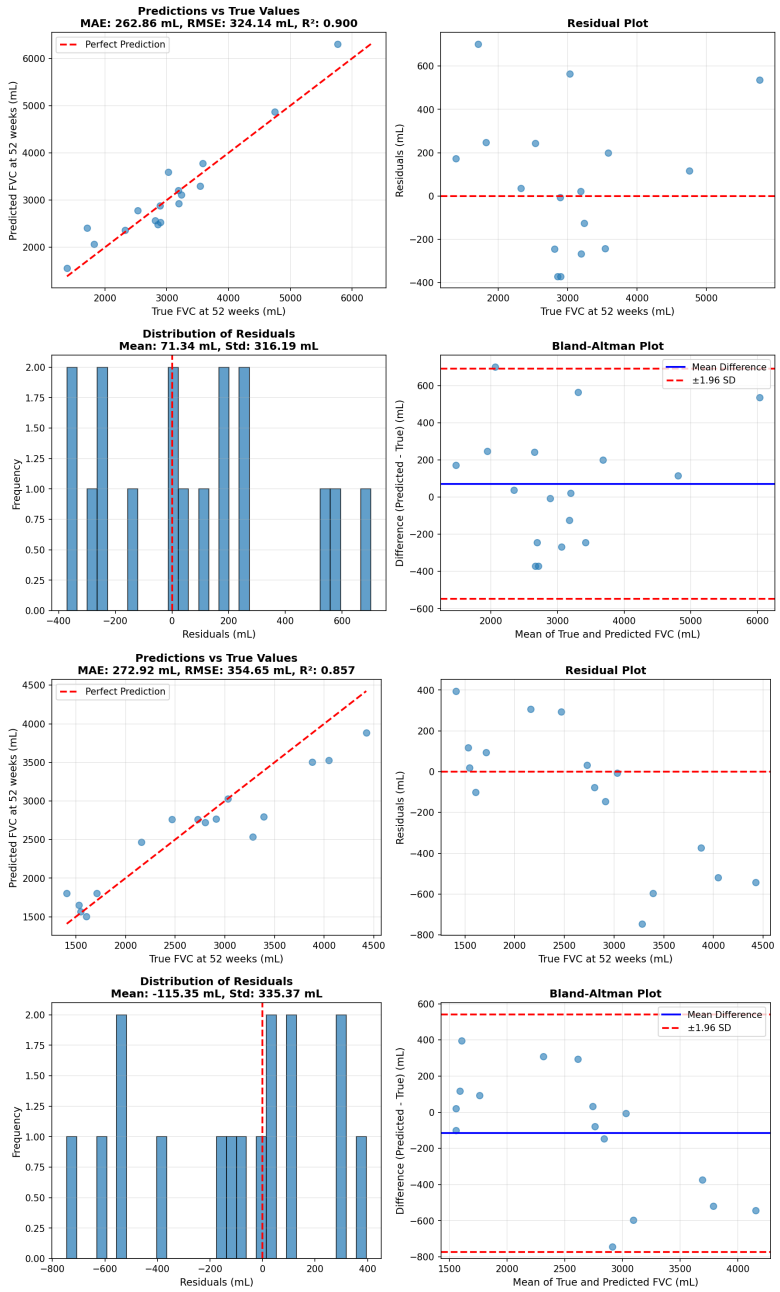

Figure A.7: Test evaluation plots for folds 3 and 4 of the regression model in the *Handcrafted features + FVC(0)* configuration.

A.5 Additional Survival Analysis Results

Table A.10 reports the complete set of survival model configurations evaluated during the ablation study. Each configuration corresponds to a different combination of imaging-derived descriptors, handcrafted CT features, and demographic variables. Performance is reported using the concordance index (C-index), averaged across the cross-validation folds.

Table A.10: Complete ablation study results for Cox survival modeling.

Configuration	Val C-index	Val Std	Test C-index	Test Std
cnn_4stat_only	0.628	0.150	0.652	0.120
cnn_4stat_sex	0.622	0.101	0.626	0.129
hand_kurtosis_only	0.624	0.063	0.624	0.063
cnn_4stat_kurtosis_sex	0.624	0.089	0.623	0.127
hand_kurtosis_approxvol	0.603	0.091	0.613	0.076
cnn_4stat_hand_all_sex	0.591	0.037	0.590	0.131
hand_approxvol_only	0.585	0.083	0.585	0.083
cnn_pca5_only	0.566	0.102	0.574	0.067
hand_kurtosis_sex	0.617	0.113	0.571	0.070
hand_skew_only	0.570	0.021	0.570	0.021
cnn_3stat_no_l2_hand_all_sex	0.574	0.036	0.566	0.140
hand_reduced_3	0.540	0.111	0.557	0.085
hand_top2_sex	0.597	0.144	0.556	0.062
hand_reduced_3_sex	0.566	0.143	0.542	0.080
cnn_pca5_hand_all_sex	0.533	0.093	0.542	0.090
hand_mean_only	0.457	0.056	0.526	0.066
cnn_pca7_hand_all_sex	0.533	0.101	0.521	0.081
cnn_pca10_hand_all_sex	0.540	0.094	0.514	0.060
cnn_maxpool_no_l2norm_only	0.482	0.140	0.510	0.142
cnn_maxpool_4stat	0.482	0.136	0.508	0.140
cnn_pca3_only	0.496	0.111	0.506	0.069
cnn_pca2_hand_all_sex	0.512	0.132	0.505	0.058
hand_lungfrac_only	0.464	0.075	0.502	0.083
cnn_pca3_hand_all_sex	0.510	0.136	0.493	0.046
demo_all_hand_all	0.521	0.120	0.481	0.081
hand_all_9	0.499	0.091	0.478	0.049
demo_age_only	0.389	0.090	0.477	0.141
demo_sex_only	0.511	0.070	0.474	0.066
cnn_pca3_sex	0.514	0.124	0.473	0.035
cnn_maxpool_4stat_sex	0.470	0.136	0.468	0.109
cnn_maxpool_4stat_hand_all_sex	0.470	0.123	0.462	0.090
hand_thickness_only	0.452	0.063	0.462	0.070
cnn_maxpool_3stat_no_l2_hand_all_sex	0.476	0.125	0.457	0.097
demo_smoking_only	0.479	0.087	0.456	0.069
hand_redundant_cluster	0.465	0.084	0.453	0.089
demo_sex_age	0.416	0.106	0.453	0.128
demo_all	0.418	0.178	0.441	0.062
demo_all_redundant_cluster	0.512	0.054	0.439	0.095
demo_all_thickness	0.476	0.075	0.437	0.068

To further assess model interpretability, we report the hazard ratio estimates and Kaplan–Meier survival curves obtained across the cross-validation folds.

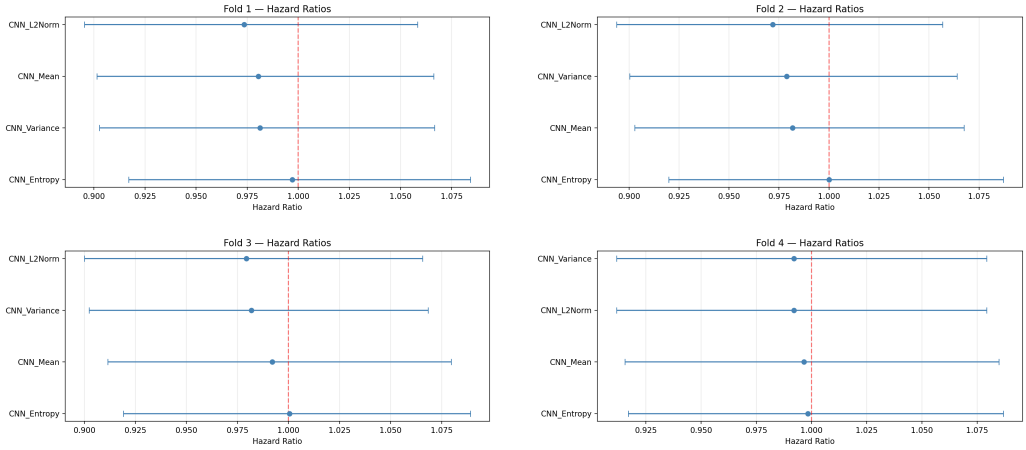

Figure A.8: Hazard ratio estimates for the Cox proportional hazards model across cross-validation folds. Error bars represent the 95% confidence intervals of the estimated coefficients.

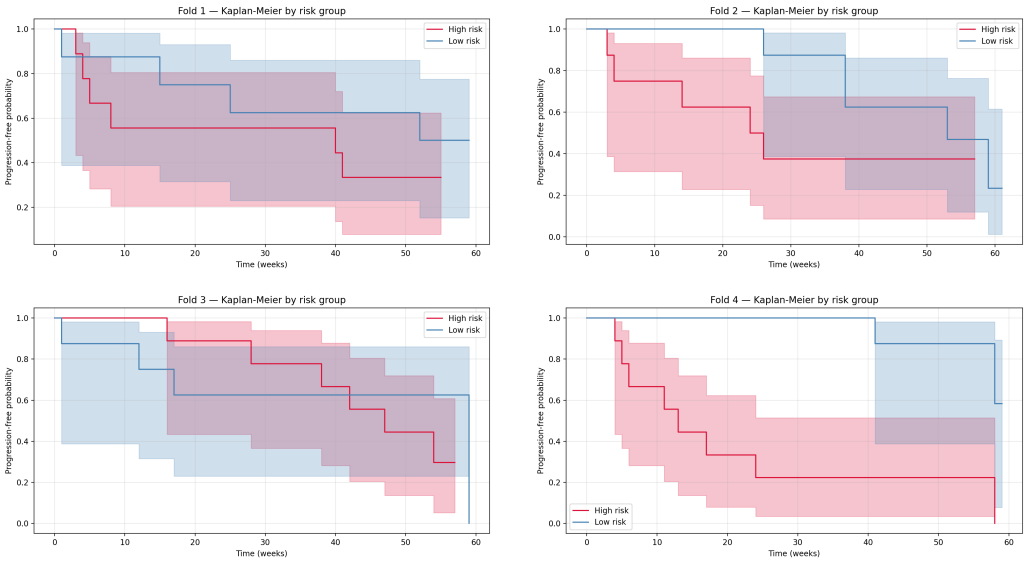

Figure A.9: Kaplan–Meier survival curves obtained by stratifying patients into high- and low-risk groups based on the predicted Cox model risk score across cross-validation folds.

Bibliography

[1] Hyun Joo Kim, David Perlman, and Rade Tomic. «Natural history of idiopathic pulmonary fibrosis». In: *Respiratory Medicine* 109.6 (2015), pp. 661–670. DOI: 10.1016/j.rmed.2015.02.002.

[2] Ganesh Raghu, Martine Remy-Jardin, Jeffrey L. Myers, et al. «Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline». In: *American Journal of Respiratory and Critical Care Medicine* 198.5 (2018), e44–e68. DOI: 10.1164/rccm.201807-1255ST.

[3] Roland M. du Bois et al. «Forced vital capacity in patients with idiopathic pulmonary fibrosis: test properties and minimal clinically important difference». In: *American Journal of Respiratory and Critical Care Medicine* 184.12 (2011), pp. 1382–1389. DOI: 10.1164/rccm.201105-0840OC.

[4] David A Lynch Andrea S Oh et al. «Deep Learning-based Fibrosis Extent on Computed Tomography Predicts Outcome of Fibrosing Interstitial Lung Disease Independent of Visually Assessed Computed Tomography Pattern». In: *Annals of the American Thoracic Society* 21.2 (2024), pp. 218–227. DOI: 10.1513/AnnalsATS.202301-084OC.

[5] David J. Lederer and Fernando J. Martinez. «Idiopathic Pulmonary Fibrosis». In: *New England Journal of Medicine* 378.19 (2018), pp. 1811–1823. DOI: 10.1056/NEJMra1705751.

[6] Ganesh Raghu, Bram Rochwerg, Yuan Zhang, et al. «An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline: Treatment of idiopathic pulmonary fibrosis. An update of the 2011 clinical practice guideline». In: *American Journal of Respiratory and Critical Care Medicine* 192.2 (2015), e3–e19. DOI: 10.1164/rccm.201506-1063ST.

[7] Ganesh Raghu, Martine Remy-Jardin, Luca Richeldi, et al. «Idiopathic Pulmonary Fibrosis (an Update) and Progressive Pulmonary Fibrosis in Adults: An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline». In: *American Journal of Respiratory and Critical Care Medicine* 205.9 (2022), e18–e47. DOI: 10.1164/rccm.202202-0399ST.

[8] Brett Ley et al. «A multidimensional index and staging system for idiopathic pulmonary fibrosis». In: *Annals of Internal Medicine* 156.10 (2012), pp. 684–695. DOI: 10.7326/0003-4819-156-10-201205150-00004.

[9] Talmadge E. King, Shira Safrin, Kathleen M. Starko, Kevin K. Brown, Paul W. Noble, Ganesh Raghu, and David A. Schwartz. «Analyses of efficacy end points in a controlled trial of interferon-gamma1b for idiopathic pulmonary fibrosis». In: *Chest* 127.1 (2005), pp. 171–177.

[10] Luca Richeldi et al. «Relative versus absolute change in forced vital capacity in idiopathic pulmonary fibrosis». In: *Thorax* 67.5 (2012), pp. 407–411. DOI: 10.1136/thoraxjnl-2011-201184.

[11] Christopher J. Zappala, Pinelopi I. Latsi, Andrew G. Nicholson, Thomas V. Colby, Douglas Cramer, Elisabetta A. Renzoni, David M. Hansell, Roland M. du Bois, and Athol U. Wells. «Marginal decline in forced vital capacity is associated with a poor outcome in idiopathic pulmonary fibrosis». In: *European Respiratory Journal* 35.4 (2010), pp. 830–835. DOI: 10.1183/09031936.00155108.

[12] A. Xaubet, C. Agustí, P. Luburich, J. Roca, C. Montón, M. C. Ayuso, J. A. Barberá, and R. Rodriguez-Roisin. «Pulmonary function tests and CT scan in the management of idiopathic pulmonary fibrosis». In: *American Journal of Respiratory and Critical Care Medicine* 158.2 (Aug. 1998), pp. 431–436. DOI: 10.1164/ajrccm.158.2.9709008.

[13] Kaiwen Geng, Zhiyi Shi, Xiaoyan Zhao, Alaa Ali, Jing Wang, Joseph Leader, and Jiantao Pu. *BeyondCT: A Deep Learning Model for Predicting Pulmonary Function from Chest CT Scans*. arXiv preprint. 2024. arXiv: 2408.05645 [eess.IV].

[14] Hernan P. Fainberg et al. «Forced vital capacity trajectories in patients with idiopathic pulmonary fibrosis: a secondary analysis of a multicentre, prospective, observational cohort». In: *The Lancet Digital Health* 4.12 (Dec. 2022), e862–e872. DOI: 10.1016/S2589-7500(22)00173-X.

[15] Luke A. Smith, Lauren Oakden-Rayner, Alix Bird, Minyan Zeng, Minh-Son To, Sutapa Mukherjee, and Lyle J. Palmer. «Machine learning and deep learning predictive models for long-term prognosis in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis». In: *The Lancet Digital Health* 5.12 (Dec. 2023), e872–e881. DOI: 10.1016/S2589-7500(23)00177-2.

[16] J. Jacob et al. «Predicting Outcomes in Idiopathic Pulmonary Fibrosis Using Automated Computed Tomographic Analysis». In: *American Journal of Respiratory and Critical Care Medicine* 198.6 (2018), pp. 767–776. DOI: 10.1164/rccm.201711-2174OC.

[17] 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.

[18] S. M. Humphries et al. «Deep Learning Classification of Usual Interstitial Pneumonia Predicts Outcomes». In: *American Journal of Respiratory and Critical Care Medicine* 209.9 (2024), pp. 1121–1131. DOI: 10.1164/rccm.202307-1191OC.

[19] Zhen Wu et al. «Idiopathic pulmonary fibrosis mortality risk prediction based on artificial intelligence: The CTPF model». In: *Frontiers in Pharmacology* 13 (2022), p. 878764. DOI: 10.3389/fphar.2022.878764.

[20] Mohamad M. AlRahhal, Yakoub Bazi, Haikel AlHichri, and Mansour Zuair. «Learning a Multi-Branch Neural Network from Multiple Sources for Knowledge Adaptation in Remote Sensing Imagery». In: *Remote Sensing* 10 (2018).

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. «Deep Residual Learning for Image Recognition». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. «ImageNet Classification with Deep Convolutional Neural Networks». In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2012.

[23] Karen Simonyan and Andrew Zisserman. «Very Deep Convolutional Networks for Large-Scale Image Recognition». In: *International Conference on Learning Representations (ICLR)*. 2015.

[24] Haowen Deng, Youyou Zhou, Lin Wang, and Cheng Zhang. «Ensemble learning for the early prediction of neonatal jaundice with genetic features». In: *BMC Medical Informatics and Decision Making* 21 (2021), p. 338. DOI: 10.1186/s12911-021-01701-9.

[25] Open Source Imaging Consortium (OSIC). *OSIC Pulmonary Fibrosis Progression*. https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression. Kaggle competition. 2020.

[26] Kaggle. *OSIC Pulmonary Fibrosis Challenge Solutions*. https://www.kaggle.com/competitions/osic-pulmonary-fibrosis-progression. Accessed: 2026. 2020.

[27] Hugo J. W. L. Aerts, Elisabeth R. Velazquez, Ralph T. H. Leijenaar, et al. «Radiomics: The bridge between medical imaging and personalized medicine». In: *Nature Communications* 5 (2014), p. 4006.

[28] Stephen S. F. Yip and Hugo J. W. L. Aerts. «Applications and limitations of radiomics». In: *Physics in Medicine and Biology* 61.13 (2016), R150–R166. DOI: 10.1088/0031-9155/61/13/R150.

[29] Alex Zwanenburg, Martin Vallieres, Maha A. Abdalah, et al. «The Image Biomarker Standardisation Initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping». In: *Radiology* 295.2 (2020), pp. 328–338. DOI: 10.1148/radiol.2020191145.

[30] Andrew C Best, David A Lynch, Christopher M Bozic, Robert R Miller, and Gary K Grunwald. «Quantitative CT indexes in idiopathic pulmonary fibrosis: relationship with physiologic impairment». In: *Radiology* 228.2 (2003), pp. 407–414.

[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. «ImageNet: A Large-Scale Hierarchical Image Database». In: *International Journal of Computer Vision* 115.3 (2009), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.