MULTIMODAL MACHINE LEARNING FOR PROGNOSTIC MODELLING IN IDIOPATHIC PULMONARY FIBROSIS

Ahmed H. Shahin

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

April 21, 2025

I, Ahmed H. Shahin, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Idiopathic Pulmonary Fibrosis (IPF) is a severe lung disease characterized by rapid progression and high mortality, with a highly variable prognosis between patients. This thesis leverages machine learning to enhance prognosis prediction in IPF by analysing clinical data and volumetric imaging. We first address the challenge of missing data in patient records by applying latent variable models to accurately impute missing attributes based on the available information in each record. Next, we use the Cox proportional hazards model to predict mortality risk from patient data. As a ranking objective, the Cox model requires many samples per training iteration, which is computationally expensive and often infeasible for volumetric data. We introduce a scalable memory bank-based training approach for efficient model training with volumetric data. Recognizing the inherent constraints of the Cox model, we also propose a new method, CenTime, which better utilizes censored data and directly predicts the time-to-mortality. CenTime relaxes the assumptions of the Cox model, provides a more precise estimation of patient outcomes, and leverages right-censored data more effectively. Our methods are validated on a comprehensive dataset of IPF patients, demonstrating significant improvements in prediction accuracy over existing approaches. This work can advance personalized prognosis in IPF, aiding clinicians in developing tailored treatment strategies.

Impact Statement

This thesis advances prognosis prediction in Idiopathic Pulmonary Fibrosis (IPF) using machine learning techniques. IPF is a severe lung disease with a median survival of 2–3 years post-diagnosis and a highly variable prognosis among patients. This work tackles key challenges in IPF prognosis, including missing data imputation, computationally efficient training with high-resolution volumetric imaging, and precise time-to-death prediction. By improving mortality risk assessment, these models enable clinicians to identify high-risk patients and develop personalized treatment strategies. Additionally, this research facilitates the discovery of novel imaging biomarkers, paving the way for improved disease understanding and targeted therapies.

Beyond IPF, the presented methods extend to broader prognosis prediction tasks, including interstitial lung diseases, cancer, and chronic conditions. The CenTime model, in particular, enhances survival prediction by effectively leveraging censored data and providing more accurate time-to-event estimations. This work contributes to the broader field of machine learning for healthcare, advancing personalized medicine and data-driven clinical decision-making.

Acknowledgements

﴿ الْحَمْدُ لِلَّهِ الَّذِى هَدَانَا لِهَذَا وَمَا كُنَّا لِنَهْتَدِىَ لَوْلَا أَنْ هَدَانَا اللَّهُ ﴾

القرآن الكريم – الأعراف ﴿٤٣﴾

"Praise to Allah, who has guided us to this; and we would never have been guided if Allah had not guided us"

Quran, Al-A'raf 43

I would like to sincerely thank my supervisor, David Barber, for his invaluable guidance, unwavering support, and continuous encouragement throughout my PhD journey. His technical expertise, mentorship, and belief in my abilities have been instrumental in shaping this thesis. Despite his demanding schedule, David always made time for discussions, constructive criticism, and motivational advice. His mentorship has fundamentally changed my approach to research, and I am truly fortunate to have had him as my supervisor.

I am deeply grateful to my secondary supervisors, Daniel C. Alexander and Joseph Jacob, whose valuable feedback and clinical insights significantly shaped the direction of my research. Their guidance has greatly improved the quality and relevance of my work.

My sincere thanks go to my viva examiners, Paul Taylor and Guang Yang, for their detailed comments and thoughtful suggestions, which substantially strengthened this thesis. I deeply appreciate their time and constructive criticism.

This research would not have been possible without the generous support and funding from the Open Source Imaging Consortium (OSIC). Special thanks go to Carmela Wegworth, Elizabeth Estes, Justin Zita, and the entire OSIC team for

their collaborative spirit, friendship, and dedication to advancing pulmonary fibrosis research. I am grateful for the resources, data, and opportunities provided by OSIC, which have been instrumental in this work.

I am also thankful to Noha El-Zehiry and Siemens Healthineers for offering me an internship opportunity during my PhD. The resources, collaboration, and mentorship provided during my internship at Siemens significantly enriched my research experience.

I extend my gratitude to friends and colleagues at the UCL AI Centre and the Centre for Medical Image Computing (CMIC) for their continuous support, engaging discussions, and fruitful collaborations. Special thanks go to An Zhao, Shahab Aslani, and Moucheng Xu for their friendship, technical advice, and shared experiences. I also warmly thank my friends at Nansen Village, London–Abdelrahman Abdeldayem, Mohamed Nabil, Osama Nabil, Mohamed Rabie, Abdullah Shoair, and Mohamed Wahba–whose companionship greatly enriched this journey, making it both enjoyable and rewarding.

A heartfelt thank you to my parents, Hassaan and Mona, and my siblings, Akram and Aya, for their unwavering love, support, and encouragement throughout my life. Studying abroad meant enduring the pain of separation, and I am deeply grateful for their sacrifices and belief in me, which have been a source of strength and motivation. No matter how old I grow, making my parents proud and happy remains my ultimate goal and greatest motivation.

My deepest gratitude goes to my wife, Salma, who joined me halfway through this journey and has been my rock ever since. Her sacrifice, patience, and understanding have been among my greatest blessings. Salma has been my pillar of strength and my confidante. Her love and encouragement have been my driving force, and I am truly blessed and grateful to have her by my side.

Last but not least, I thank my lovely daughter, Layla—only five months old at the time of writing these words—for making my life meaningful and showing me the true meaning of love and happiness. The stresses of the PhD journey fade away with every smile of hers. I hope this small token makes her proud someday.

UCL Research Paper Declaration Form A

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?**

   Survival Analysis for Idiopathic Pulmonary Fibrosis using CT Images and Incomplete Clinical Data

   (b) **Please include a link to or doi for the work:**

   https://proceedings.mlr.press/v172/shahin22a.html

   (c) **Where was the work published?**

   International Conference on Medical Imaging with Deep Learning

   (d) **Who published the work?**

   Proceedings of Machine Learning Research (PMLR)

   (e) **When was the work published?**

   2022

   (f) **List the manuscript's authors in the order they appear on the publication:**

   Ahmed H. Shahin, Joseph Jacob, Daniel C. Alexander, David Barber

   (g) **Was the work peer reviewd?**

   Yes

   (h) **Have you retained the copyright?**

   No

(i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

Yes. https://arxiv.org/abs/2203.11391

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**

   If 'Yes', please please give a link or doi:

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**

   AHS and DB designed the study. AHS implemented the methods, conducted the experiments, and wrote the first draft of the manuscript. DB and DCA contributed to the refinement of the research idea. JJ provided clinical insights and guidance on the medical aspects of the work. DB reviewed the manuscript and provided critical feedback. All authors contributed to the final version of the manuscript.

4. **In which chapter(s) of your thesis can this material be found?**

   Chapter 4 and Chapter 5

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

Date:

February 9, 2025

Supervisor/Senior Author signature (where appropriate)**:**

Date: 10th February 2025

UCL Research Paper Declaration Form B

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?**

   CenTime: Event-conditional modelling of censoring in survival analysis

   (b) **Please include a link to or doi for the work:**

   https://www.sciencedirect.com/science/article/pii/S1361841523002761

   (c) **Where was the work published?**

   Medical Image Analysis

   (d) **Who published the work?**

   Elsevier

   (e) **When was the work published?**

   2024

   (f) **List the manuscript's authors in the order they appear on the publication:**

   Ahmed H. Shahin, An Zhao, Alexander C. Whitehead, Daniel C. Alexander, Joseph Jacob, David Barber

   (g) **Was the work peer reviewd?**

   Yes

   (h) **Have you retained the copyright?**

   No

(i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

Yes. https://arxiv.org/abs/2309.03851

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**

   If 'Yes', please please give a link or doi:

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**

   DB and AHS designed the study. AHS implemented the methods and wrote the first draft of the manuscript. AHS and AZ conducted the experiments. JJ provided clinical insights and guidance on the medical aspects of the work. DB reviewed the manuscript and provided critical feedback. AZ and AHS refined the manuscript. All authors contributed to the final version of the manuscript.

4. **In which chapter(s) of your thesis can this material be found?**

   Chapter 6

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

Date:

February 9, 2025

Supervisor/Senior Author signature (where appropriate)**:**

Date: 10th February 2025

Contents

List of Figures

List of Tables

List of Abbreviations

6MWT 6-Minute Walk Test

AI Artificial Intelligence

C-Index Concordance Index

CNN Convolutional Neural Network

COPD Chronic Obstructive Pulmonary Disease

CoxMB Cox Proportional Hazards with Memory Banks

CoxPH Cox Proportional Hazards

CPI Composite Physiologic Index

DL_{CO} Diffusing Capacity of the Lung for Carbon Monoxide

EM Expectation-Maximization

FDA US Food and Drug Administration

FEV1 Forced Expiratory Volume in 1 second

FVC Forced Vital Capacity

GAP Gender Age Physiology

GBMs Gradient Boosting Machines

GGO Ground-Glass Opacity

GPU Graphics Processing Unit

HRCT High-Resolution Computed Tomography

HU Hounsfield Units

IIP Idiopathic Interstitial Pneumonia

ILDs Interstitial Lung Diseases

IPF Idiopathic Pulmonary Fibrosis

KM Kaplan-Meier

LLM Large Language Model

LSTM Long Short-Term Memory

LVM Latent Variable Model

MAE Mean Absolute Error

MAR Missing at Random

MCAR Missing Completely at Random

MICE Multiple Imputation by Chained Equations

MLP Multi-Layer Perceptron

MNAR Missing Not at Random

NRMSE Normalized Root Mean Squared Error

NTP Next-Token Prediction

OSIC Open Source Imaging Consortium

PDF Probability Density Function

RAE Relative Absolute Error

RSFs Random Survival Forests

SDTs Survival Decision Trees

SLB Surgical Lung Biopsy

SVMs Support Vector Machines

SVR Support Vector Regression

UIP Usual Interstitial Pneumonia

UK United Kingdom

VLMs Vision Language Models

Chapter 1

Introduction

1.1 Idiopathic Pulmonary Fibrosis

Idiopathic Pulmonary Fibrosis (IPF) is a fibrotic lung disease that belongs to the group of Interstitial Lung Diseases (ILDs) and is characterized by stiffening and scarring (fibrosis) of the lung tissue. This leads to shortness of breath, progressive decline in lung function, and ultimately respiratory failure and death [1, 2, 3, 4]. IPF is the most common and severe fibrotic lung disease, with a median survival rate ranging from two to three years, worse than many cancers [5, 6]. The incidence and prevalence of IPF are rising; in the United Kingdom (UK), rates increased by 78% from 2000 to 2012 [7]. As an idiopathic disease, IPF has no known cause [1]. For treatment, there is no definitive cure for the disease, and the available drugs aim to slow down the progression of the disease and manage the symptoms [8]. Diagnosing IPF is also challenging and requires a multidisciplinary approach that includes clinical, radiological, and histopathological assessments [6].

A significant challenge in the management of IPF is the heterogeneous and highly unpredictable disease progression, making it difficult to predict the prognosis and response to treatment for individual patients. Although numerous computer-based methods have been applied to improve disease prognosis predictions, accurate and widely accepted models for predicting disease progression and outcomes in

clinical practice for IPF patients are still lacking [9].

These challenges underscore the need for reliable prognostic models to predict IPF progression, guiding clinical decisions and identifying high-risk patients for early intervention. Furthermore, probing these models can help identify potential prognostic markers and risk factors associated with the disease, guiding future research and clinical practice and ultimately improving the understanding and management of IPF.

Several studies have shown the potential of machine learning methods in predicting the future progression of IPF using clinical data, such as demographic information, pulmonary function tests, and blood tests [10, 11]. However, these models often rely on clinical data alone, which may not capture the full complexity of the disease or provide accurate predictions. High-Resolution Computed Tomography (HRCT) offers precise anatomical insights into the characteristics and progression of lung disease, serving as an essential tool in supporting clinicians with IPF diagnosis, prognosis, and monitoring. Furthermore, HRCT has demonstrated higher sensitivity than pulmonary function measurements in some instances, particularly for asymptomatic IPF patients [12, 13]. This enhanced sensitivity enables the potential for earlier and more reliable prognoses in IPF.

Given the complexity and variability of IPF progression, conventional methods have struggled to provide reliable prognostic models. Machine learning, with its ability to process vast amounts of data and uncover complex patterns, presents a promising approach to improving IPF prognosis. In particular, multimodal machine learning — which integrates diverse data sources like clinical records and imaging — can capture the disease's full complexity and enhance prediction accuracy.

1.2 Multimodal Machine Learning

Machine learning is a subfield of Artificial Intelligence (AI) which focuses on developing algorithms that can learn from vast amounts of data, without being explicitly

programmed, to make predictions or decisions [14]. It has shown outstanding performance and capabilities in various domains, such as computer vision [15], natural language processing [16], and healthcare [17]. Multimodal machine learning pertains to developing machine learning algorithms that can learn from multiple data sources, such as images, text, speech, and other modalities, to make decisions [18].

Integrating imaging and clinical data through multimodal machine learning has shown promise in enhancing prognostic models for various chronic diseases, including cancer [19], cardiovascular [20, 21], and respiratory diseases [22, 23].

In the context of IPF, multimodal machine learning can integrate and learn from multiple data sources, such as clinical and imaging data, to predict the future progression of the disease accurately and reliably. Clinical data can include demographic information, pulmonary function tests, blood tests, and other clinical assessments, while imaging data includes HRCT images. By integrating and learning from multiple data sources, multimodal machine learning models can capture the full complexity of the disease, provide accurate predictions, and improve the understanding and management of IPF. This thesis proposes a multimodal approach to address critical gaps in IPF prognosis models.

1.2.1 Survival Analysis

While machine learning offers a broad framework for predictive modelling, IPF prognosis requires a method that can specifically account for time-dependent events, such as disease progression or mortality. Survival analysis is well-suited to this task, providing tools for time-to-event prediction and allowing clinicians to anticipate critical outcomes.

Survival analysis [24] is a statistical technique commonly used in medical research to predict the time until an event of interest, often called 'time-to-event' analysis. This method is useful for estimating either the time until an event (*e.g.*, disease progression, cancer recurrence, or death) or the risk of an event occurring

within a specific time frame [25, 26]. Typical events in survival analysis include death [25, 27], cancer recurrence [28], or even nonmedical outcomes like machine failure [29].

For IPF, progression is often evaluated based on the time to critical events, such as death, respiratory failure, or lung transplantation, making it particularly suited to survival analysis modelling [30]. Survival analysis models offer valuable insights into the progression of IPF, helping to identify high-risk patients and supporting clinical decision-making.

Despite the utility of survival analysis in modelling disease progression, existing models face limitations, particularly in handling the high rate of censored data. To address these challenges, this thesis proposes a novel framework combining advanced imputation techniques, multimodal integration, and survival analysis enhancements to produce more accurate and reliable prognostic predictions for IPF.

1.3 Research Problem

Building on this foundation, the primary research problem addressed in this thesis is the development of accurate and reliable prognostic models for IPF that can predict the future progression of the disease using clinical and imaging data. Despite the advances in machine learning and survival analysis, predicting the future progression of IPF remains relatively challenging and unexplored. Existing prognostic models for IPF depend heavily on clinical data alone, often excluding valuable imaging data that could reveal structural changes associated with disease progression [10, 11]. Furthermore, clinical data often suffer from missing values with which current methods struggle, impairing the performance of machine learning models and introducing bias into predictions. These methods often impute missing values in ways that do not account for the dependencies among clinical variables, thereby reducing the reliability of predictions [31, 32].

On the other hand, existing survival analysis models suffer from several lim-

itations. For instance, the Cox Proportional Hazards (CoxPH) model [26] has the following limitations

- **Proportional Hazards Assumption:** The CoxPH model assumes that the hazard ratio between any two individuals is constant over time, which may not hold in practice, especially for complex diseases like IPF.
- **Requirement for Large Batch Sizes:** Training CoxPH models requires ranking samples by risk scores, demanding large batch sizes, which is often infeasible with high-resolution data like HRCT due to memory constraints. Using small batch sizes for training can impair the model's performance and restrict its ability to learn effectively from the data, particularly for censored samples.
- **Limited Utilization of Censored Samples:** The CoxPH model often underutilizes censored samples in the dataset, the samples that have not died by the end of the study. These samples dominate survival analysis datasets and contain valuable information that can improve the model's performance and reliability.
- **Outputting Risk Scores:** The CoxPH model outputs a risk score for each patient, which may not be directly interpretable or valuable for clinicians. Clinicians often require a rough estimate of each patient's expected time to death, which the CoxPH model does not provide.

By addressing these limitations, this thesis aims to set a new standard for IPF prognosis models, ultimately supporting clinicians in making more personalized and timely decisions that enhance patient care.

1.4 Contributions

In this thesis, we aim to develop multimodal machine learning models that integrate and learn from multiple data sources, such as clinical and imaging data, to predict

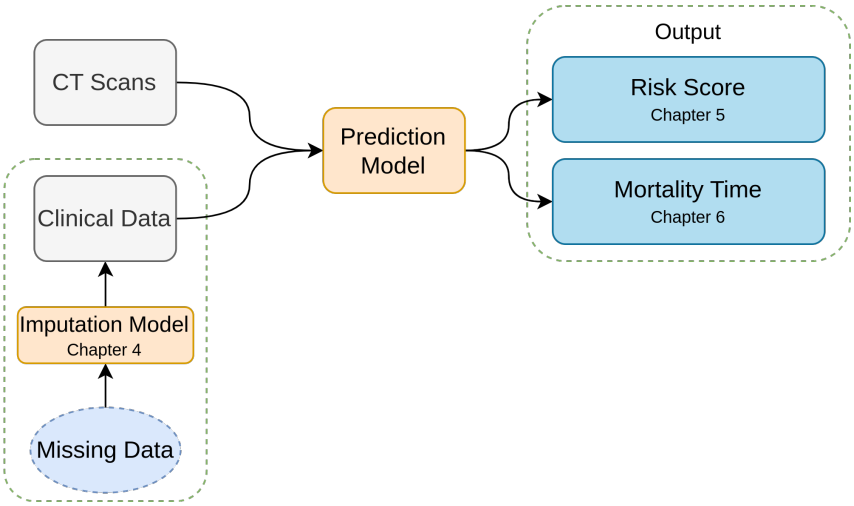


Figure 1.1: Overview of the thesis structure and contributions.

the future progression of IPF, see Figure 1.1. In real-world datasets, quality control and data preprocessing are essential to ensure the reliability and validity of the data, including developing methods to handle missing data without impairing or biasing the developed models. We aim to develop machine learning methods to predict IPF progression, providing a mortality risk score or estimated time to death. Further, we aim to develop novel survival analysis models that can better handle the censoring process, leverage the censored samples in the dataset, and relax the assumptions of existing models to improve prediction accuracy and reliability. The developed models are evaluated on a large and diverse dataset of IPF patients and compared to existing models to demonstrate their effectiveness and reliability. The contributions of this thesis can be summarized as follows.

1.4.1 Handling Missing Data in IPF Clinical Records

Real-world clinical data often contain missing values, impairing machine learning model performance and introducing bias into predictions. Missing data is especially common in IPF due to the nature of the disease and the variability in clinical assessments. For instance, some hospitals may lack equipment for specific tests, or patients may be unable to complete assessments due to their health condition. Standard

imputation methods generally assume that clinical attributes are independent, which may not hold in practice for IPF data.

In this thesis, we hypothesize that missing values for a patient can be accurately estimated using the observed clinical values. To address this, we propose a method based on Latent Variable Models (LVMs) and the Expectation-Maximization (EM) algorithm to handle missing data in IPF clinical records. This imputation method is applied to complete clinical datasets in our survival analysis experiments, aiming to improve prediction accuracy by providing a more comprehensive dataset for model training and reducing the bias introduced by missing values.

Furthermore, effective handling of missing data will facilitate a robust integration of clinical and imaging data, enhancing the reliability of our multimodal prognosis model for IPF. This contribution addresses the problem of missing data in clinical records, ensuring the reliability and validity of the developed models.

1.4.2 Improving the Cox Proportional Hazards Model with Memory Banks

The CoxPH model is widely used in survival analysis due to its simplicity and interpretability. However, training the CoxPH model on high-resolution imaging data (such as HRCT images) is computationally and memory-intensive, often requiring a reduced batch size to fit the model within the Graphics Processing Unit (GPU) memory. This limitation can impair model performance and restrict its ability to learn effectively from data, particularly for censored samples.

Inspired by the contrastive learning literature, this thesis proposes a novel approach to enhance the CoxPH model's performance by integrating memory banks during training [33]. Memory banks refer to a technique in which predictions from previous training iterations are stored temporarily and reused in later stages. This approach enables the model to leverage previously seen information, effectively addressing computational constraints.

By incorporating memory banks, we show that the CoxPH model achieves improved performance, scalability, and a better utilization of censored samples. We evaluate this method on a large and diverse dataset of IPF patients, comparing it to the standard CoxPH model to demonstrate its effectiveness and reliability. This chapter addresses the limitations of the CoxPH model when training on high-resolution imaging data and provides a novel approach to improve its performance and scalability.

1.4.3 Event-Conditional Modelling of Censoring in Survival Analysis

In this contribution, we address the foundational assumptions of traditional survival analysis models, such as the CoxPH model [26] and DeepHit [25], and introduce CenTime, a novel event-conditional model designed to handle the censoring process in survival analysis better. CenTime leverages censored samples in the dataset more effectively through a novel objective function for training survival analysis models.

Our approach relaxes the proportional hazards assumption, which may not hold in complex diseases like IPF, and outputs individualized survival time estimates. Through extensive evaluation, we demonstrate that CenTime improves the prediction accuracy, outperforming existing methods, including the CoxPH model and DeepHit. Although CenTime is a general framework applicable to a wide range of survival analysis tasks and datasets, we evaluate it specifically on our IPF dataset to highlight its effectiveness in this context. This chapter addresses the limitations of existing survival analysis models being less effective in handling the censoring process. It introduces a novel event-conditional model that can better leverage censored samples in the dataset and improve prediction accuracy and reliability while providing direct estimates of each patient's expected time to death.

1.5 Thesis Overview

This thesis is organized as follows

- **Chapter 2** provides a comprehensive background on IPF, including the lung anatomy, physiology of the disease, as well as the clinical assessment, diagnosis, prognosis, and management of IPF.
- **Chapter 3** provides a technical background on machine learning methods that are relevant to the thesis, including missing data imputation, survival analysis, and multimodal machine learning.
- **Chapter 4** presents our proposed method for handling missing data in IPF clinical records, which is essential for the reliability and validity of the developed models. This method is used in the subsequent chapters to train survival analysis models.
- **Chapter 5** discusses the limitation of the CoxPH model when limited by the GPU memory to small batch sizes and proposes the integration of memory banks to alleviate this limitation and improve the model's performance.
- **Chapter 6** discusses the limitation of existing survival analysis models and proposes a novel event-conditional model, CenTime, that can model the censoring process in survival analysis and better leverage the censored samples in the dataset.
- **Chapter 7** concludes the thesis and discusses future work.

1.6 Publications

The content of this thesis is based on the following publications

- **Shahin A. H.**, Jacob J., Alexander D. C., and Barber D., "*Survival Analysis for Idiopathic Pulmonary Fibrosis using CT Images and Incomplete Clinical*

Data", **Oral** presentation at the International Conference on Medical Imaging with Deep Learning (MIDL), 2022 [34].

- **Shahin A. H.**, Zhao A., Whitehead A. C., Alexander D. C., Jacob J., and Barber D., "*CenTime: Event-conditional modelling of censoring in survival analysis*", Medical Image Analysis, 2024 [35].

In addition, the following publications were conducted during the PhD but are not included in the thesis:

- Zhao A., **Shahin A. H.**, Zhou Y., Gudmundsson E., Szmul A., Mogulkoc N., Van Beek F., Brereton C. J., Van Es H. W., Pontoppidan K., Savas R., Wallis T., Unat O., Veltkamp M., Jones M. G., Van Moorsel C. H. M., Barber D., Jacob J., and Alexander D. C., "*Prognostic Imaging Biomarker Discovery in Survival Analysis for Idiopathic Pulmonary Fibrosis*", International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2022 [27].

- Lu Y., Aslani S., Zhao A., **Shahin A. H.**, Barber D., Emberton M., Alexander D. C., and Jacob J., "*A hybrid CNN-RNN approach for survival analysis in a Lung Cancer Screening study*", Heliyon, 2023 [36].

- **Shahin A. H.**, Zhuang Y., and El-Zehiry N., "*From Sparse to Precise: A Practical Editing Approach for Intracardiac Echocardiography Segmentation*", International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2023 [37].

- Whitehead A. C., **Shahin A. H.**, Zhao A., Alexander D. C., Jacob J., and Barber D., "*Neural Network Based Methods for the Survival Analysis of Idiopathic Pulmonary Fibrosis Patients from a Baseline CT Acquisition*", Medical Imaging Conference and International Symposium on Room-Temperature Semiconductor Detectors (NSS MIC RTSD), 2023 [38].

Chapter 2

Clinical Background

We give a clinical background on the lung, the disease of IPF, its diagnosis, prognosis, treatment, and the type of data primarily used in the clinical setting for diagnosis, prognosis, and monitoring of IPF.

2.1 Lung Anatomy

The lung is the primary organ of the respiratory system, whose main function is to provide oxygen to the blood. Generally, the respiratory system is divided into airways and lung parenchyma. The airways consist of the bronchus, which split off from the trachea and divide into bronchioles and alveoli. The lung parenchyma is responsible for gas exchange and includes alveoli, alveolar ducts, and bronchioles. Anatomically, lungs have an apex, three borders, and three surfaces. They are also subdivided into lobes and segments [39, 40, 41].

The lung apex lies above the first rib. The three lung borders are the anterior, posterior, and inferior borders. The anterior border has a cardiac notch in the left lung to accommodate the heart, while the posterior border extends from the seventh cervical vertebra (C7) to the tenth thoracic vertebra (T10). The inferior border is a thin border that separates the lung base from the costal surface.

The three lung surfaces are the costal, medial, and diaphragmatic surfaces. The costal surface is covered by the costal pleura and faces the sternum and ribs. The

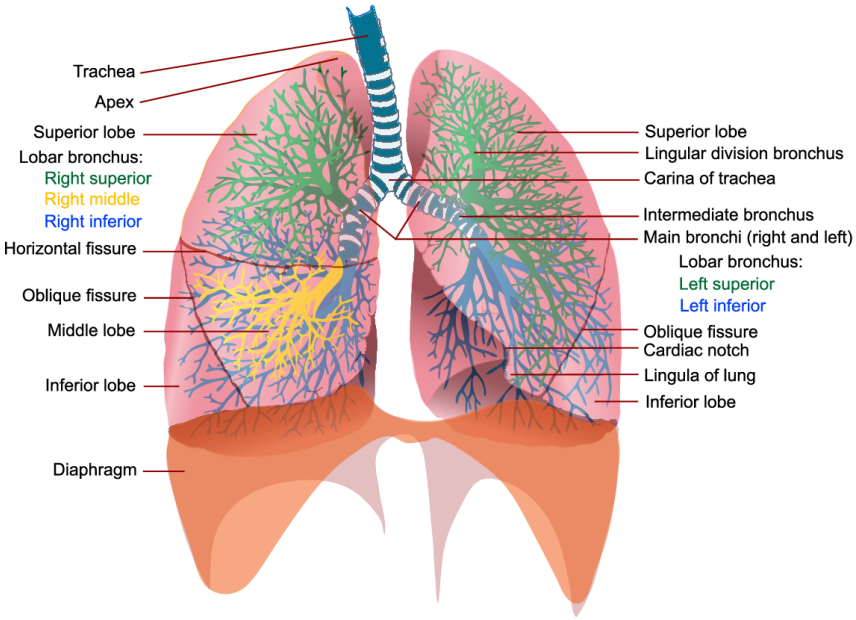


Figure 2.1: Anatomy of the lungs showing the pulmonary lobes, the bronchi, and other pulmonary structures. Taken from the public domain.

medial surface has two parts: anterior, which is related to the sternum, and posterior, which is related to the vertebra. The diaphragmatic surface is concave and lies on top of the diaphragm, with its right part higher than the left because of the existence of the liver.

The two lungs are similar but asymmetric (see Figure 2.1). Each lung comprises smaller units called lobes, which ultimately subdivide into millions of alveoli. The alveoli are the primary site of gas exchange in the lungs. The right lung has three lobes, separated by horizontal and oblique fissures, while the left has two lobes divided by an oblique fissure.

The hilum is at the centre of the medial surface and lies anterior to the T5 to T7 vertebra. It is the entry and exit point of various structures within the lung. The hilum contains mainly bronchi and pulmonary vasculature. In the right hilum, there are two bronchi, the eparterial and hyparterial bronchi, while in the left hilum, there is only one bronchus, the principal bronchus [40].

2.2 Lung Volumes and Capacities

Lung volume is a key metric for assessing lung function, essential for diagnosing and monitoring pulmonary diseases like IPF. Common lung volumes and capacities include (see Figure 2.2) [39, 42]

- Tidal Volume (TV): The volume of air inhaled or exhaled during normal breathing.
- Inspiratory Reserve Volume (IRV): The volume of air a patient can forcefully inhale after a normal inspiration.
- Expiratory Reserve Volume (ERV): The volume of air a patient can forcefully exhale after a normal expiration.
- Residual Volume (RV): The volume of air that remains in the lung after maximal exhalation.
- Inspiratory Capacity (IC): The maximum volume of air a patient can inhale after a normal expiration.
- Functional Residual Capacity (FRC): The volume of air that remains in the lung after normal expiration.
- Vital Capacity (VC): The maximum volume of air a patient can exhale after a normal inspiration.
- Total Lung Capacity (TLC): The volume of air that remains in the lung after maximal inspiration.
- Forced Vital Capacity (FVC): The maximum volume of air a patient can exhale after a maximal inspiration.
- Forced Expiratory Volume in 1 second (FEV1): The volume of air a patient can exhale in the first second of a forced expiration.

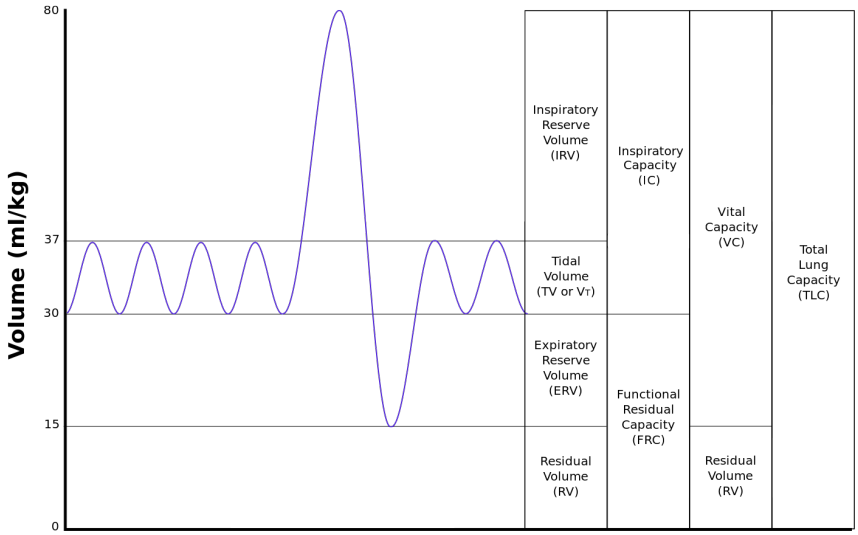


Figure 2.2: The lung volumes and capacities. Adapted from [43].

2.3 Major Types of Lung Diseases

Many diseases can affect the lungs. Lung diseases can be classified into two main categories: obstructive and restrictive. Distinguishing between obstructive and restrictive lung diseases provides context for developing machine learning models specific to IPF, a restrictive disease. By training models on data specific to IPF rather than a broad category of lung diseases, it is possible to achieve greater accuracy in disease-specific predictions, especially when distinguishing IPF from other ILDs.

2.3.1 Obstructive Lung Diseases

Obstructive lung diseases, such as Chronic Obstructive Pulmonary Disease (COPD), impair expiration, leading to air trapping and decreased FVC, FEV1, and FEV1/FVC ratios.

2.3.2 Restrictive Lung Diseases

These are diseases where specific abnormalities (*e.g.*, fibrosis or scarring) restrict lung expansion. This restriction leads to decreased lung volumes. Both FVC and FEV1 are decreased in restrictive lung diseases, but the FVC is decreased more than the FEV1, which leads to an increased FEV1/FVC ratio. An example of a restrictive

lung disease is IPF, which is the focus of this thesis.

2.4 Idiopathic Pulmonary Fibrosis

IPF is part of the broader ILDs family, which are lung disorders that cause inflammation or fibrosis in several lung areas and share similar clinical, physiologic or pathological features [44]. There are some ILDs that have known causes as well as disorders that happen due to unknown aetiology. Intuitively, the diagnosis of the latter category of diseases is more challenging. IIPs is a group of pulmonary disorders that belong to that category and have distinct histologic patterns.

IPF, the most common IIP, is a chronic and progressive disease with a median survival of two to three years from the time of diagnosis. The incidence of IPF is increasing and is more common in older adults [5, 6, 30]. It is characterised by the pattern of Usual Interstitial Pneumonia (UIP) on the chest HRCT scan, see Figure 2.4. UIP is a specific fibrosis pattern seen in the lung tissue. Lung abnormalities, like reticular opacities, traction bronchiectasis, honeycombing, and ground-glass opacities, are common in UIP [45].

2.4.1 Epidemiology

The epidemiology of IPF varies widely across different regions because of blurred diagnostic criteria and changes in the official diagnostic guidelines [46, 47]. Generally, IPF incidence has been increasing in the past few decades. However, it is unclear whether this increase is due to an actual increase in the disease incidence or better recognition and diagnosis of the disease [4].

In the UK, over 5000 IPF cases are diagnosed annually, with primary care cases rising 35% from 2000 to 2008. IPF mortality rates continue to increase; see Figure 2.3 [48]. In Europe and North America, IPF incidence is reported to be between 2.8 to 19 cases per 100,000 people [49, 50]. In comparison, it is estimated to be less than 4 cases per 100,000 persons in East Asia and South America [49].

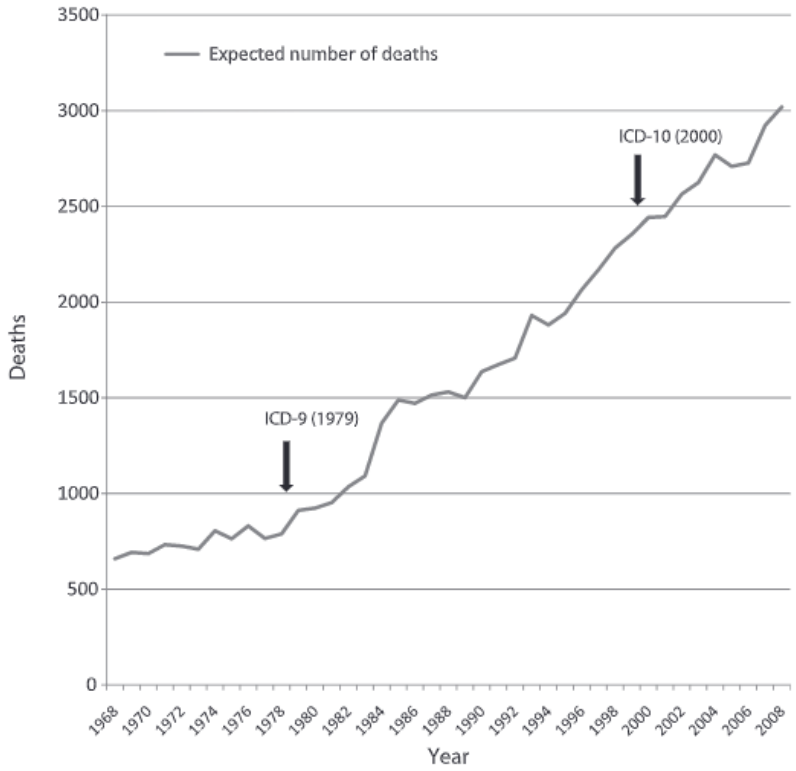


Figure 2.3: Estimated number of deaths from IPF in the UK. Age standardised to the 2008 population of England and Wales. ICD: International Classification of Diseases. Adapted from [48].

Raghu *et al.* reported that IPF prevalence in people aged 65 years and older in the US increased from 202.2 cases per 100,000 in 2001 to 494.5 in 2011 [51]. Several studies have reported a consistent increase in IPF incidence and prevalence rates among older males, with the majority being over 50 years old [52]. IPF mortality is also increasing worldwide but might be underestimated due to the lack of recognition and diagnosis difficulties associated with IPF [53].

Machine learning models can analyse epidemiological data to identify risk factors, predict disease incidence trends, and assess outcomes in IPF populations. For instance, models trained on demographic and clinical data can help stratify patient risk levels based on age, gender, or other factors [54, 55].

Change in Dyspnea Grade	Change in FVC (L)	Change in FVC (%)
Much better	2.3 ± 7.3	5.1 ± 7.3
Somewhat better	-2.1 ± 6.4	0.7 ± 6.4
Same	-2.8 ± 5.8	0.0 ± 5.8
Somewhat worse	-6.5 ± 6.6	-3.7 ± 6.6
Much worse	-6.1 ± 9.5	-3.3 ± 9.5

Table 2.1: Change in FVC with change in dyspnea grade. Adapted from [58].

2.4.2 Clinical Presentation

2.4.2.1 Symptoms and Physical Examination

Common symptoms of IPF include dyspnea, cough, and fatigue. Dyspnea (shortness of breath) is the most common symptom and is usually progressive. It limits the patient's daily activities and is often the reason for seeking medical attention as it impairs the quality of life [56, 57]. Some studies have shown that the patient-reported sensation of change of dyspnea grade is associated with changes in FVC [58], see Table 2.1. Cough is another common symptom which is more likely to happen in patients with a history of smoking [56]. Late stage IPF patients sometimes report general fatigue [59].

Physical examination of IPF patients may reveal clubbed fingers and *velcro crackles*. Finger clubbing refers to a deformity of the nail base, characterised by a swollen, spongy, and convex shape of the distal phalanx, accompanied by a reduction in the nail-fold angle [60]. It is thought to affect around 50% of IPF patients and has shown to be associated with poor prognosis [61, 62]. *Velcro crackles* are a distinctive sound heard on auscultation of the chest, which is thought to be due to the opening of small airways that are closed by fibrosis [62]. *Velcro crackles* are reported by many IPF patients, and clinical guidelines recommend their presence as a diagnostic criterion for IPF [6].

2.4.2.2 Physiological Measurements

Physiological measurements are essential for diagnosing and monitoring IPF. The most common measurements used are FVC, FEV1, and Diffusing Capacity of the Lung for Carbon Monoxide (DL_{CO}), with FVC being the most important. The majority of IPF patients suffer from a decreased FVC and DL_{CO} [6]. However, the physiological measurements can be normal in the early stages of the disease, and the decline in FVC is inconsistent across patients [4].

Physiological measurements, such as FVC and DL_{CO}, are critical features for machine learning models in IPF. Predictive algorithms can utilise these measurements to assess disease progression, with FVC serving as a valuable indicator of patient health that models can use to predict mortality [63, 64, 65, 66].

2.4.2.3 HRCT Findings

The clinical guidelines recommend the use of HRCT imaging in the diagnosis of IPF [30]. Radiographic changes in IPF patients can be observed before the onset of symptoms. Dong Soon *et al.* showed that asymptomatic IPF patients developed symptoms after more than 2 years of the diagnosis based on the HRCT findings [67].

IPF is defined by a histopathologic and/or radiologic pattern known as UIP (Figure 2.4), which involves paraseptal fibrosis and architectural distortion [30]. The most common features of UIP on HRCT include honeycombing, traction bronchiectasis, and traction bronchiolectasis, often accompanied by ground-glass opacification and fine reticulation. These terms are defined as follows

- Honeycombing, a hallmark of UIP, represents advanced pulmonary fibrosis and is identified by clusters of cystic air spaces, typically subpleural, with thick, well-defined walls [68, 69, 45] (Figure 2.4).

- Traction bronchiectasis/bronchiolectasis refers to the dilation of bronchi or bronchioles due to the retraction caused by surrounding fibrosis [68] (Figure 2.4).

- Reticulation generally signifies lung fibrosis [11], characterised by numerous short linear opacities that form net-like patterns on HRCT scans [68] (Figure 2.5).

- Ground-Glass Opacity (GGO) is identified as areas of hazy increased lung density where the margins of bronchi and vessels remain visible [68]. While pure GGO is not typically associated with UIP, it is common in patients with fibrotic lung diseases to observe GGO mixed with reticular abnormalities, traction bronchiectasis, or both [45].

The guidelines outline four patterns based on HRCT features to assist in diagnosing IPF: the UIP pattern (Figure 2.4), probable UIP pattern (Figure 2.5), indeterminate for UIP pattern, and patterns suggestive of an alternative diagnosis [30]. The UIP pattern is the most critical for IPF diagnosis, characterised by bilateral reticulation and honeycombing with subpleural and basal predominance [4, 30]. On HRCT, the UIP pattern strongly predicts the presence of histopathologic UIP, often making Surgical Lung Biopsy (SLB) unnecessary for diagnosis in patients with a typical UIP pattern and no other clear cause.

A probable UIP pattern, which includes bilateral reticulations that are subpleural and basal with peripheral traction bronchiectasis or bronchiolectasis but without honeycombing, is also strongly indicative of histopathologic UIP. Some patients with this probable UIP pattern on HRCT can be diagnosed without SLB, while others may require additional clinical evaluations, such as SLB, to confirm the diagnosis [30, 45].

HRCT scans are vital for diagnosing IPF. They can be used as input data for machine learning models designed to detect IPF patterns, such as honeycombing and ground-glass opacities [70, 71, 72, 73, 74]. By training models on HRCT images, it is possible to develop automated methods for recognising UIP patterns, supporting radiologists in diagnosis and monitoring [75]. In addition, deep learning models can be trained end-to-end to predict patient outcomes based on HRCT images without

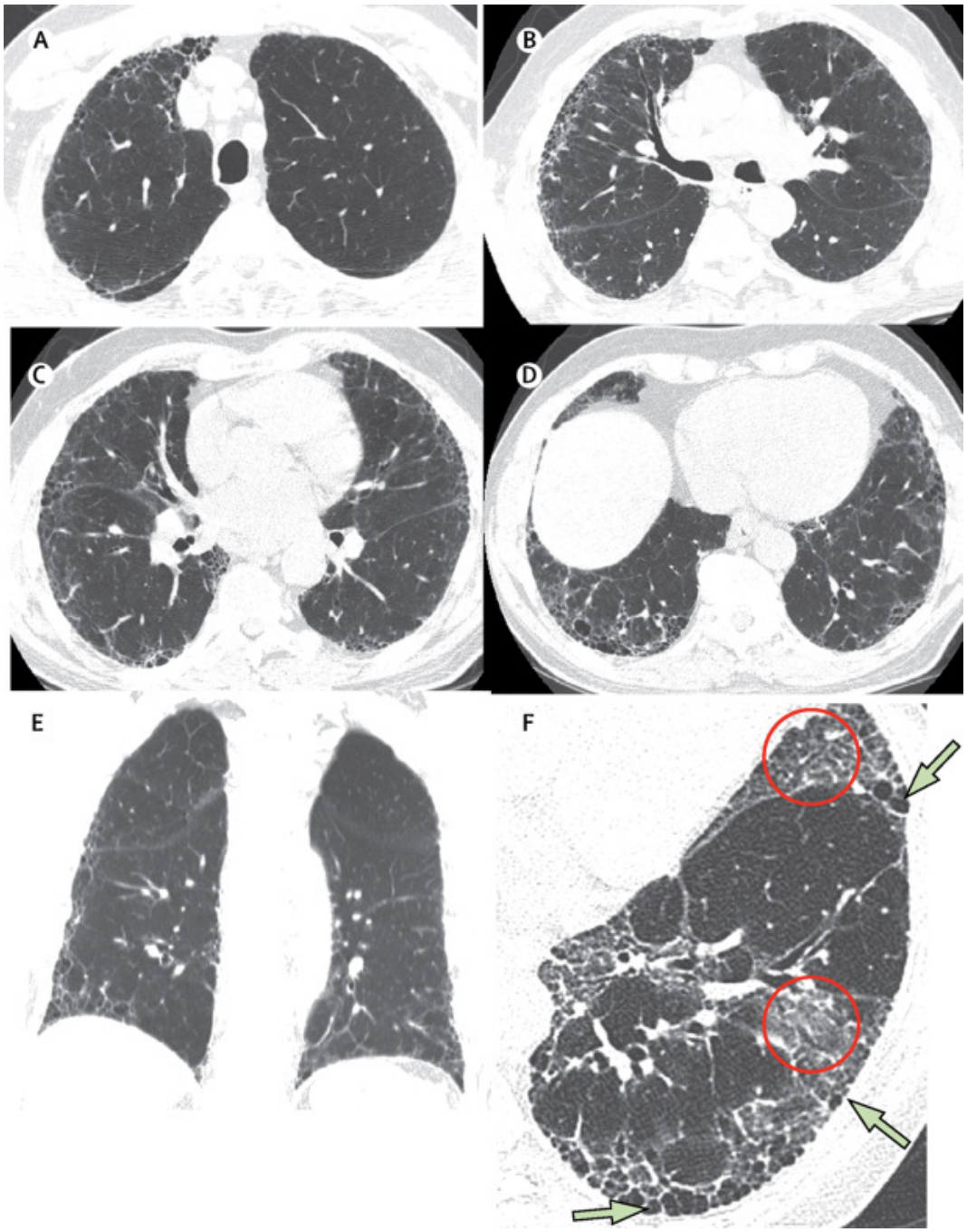


Figure 2.4: Typical UIP patterns in HRCT. (A–F) Axial and coronal HRCT scans from a patient with UIP display a subpleural predominant reticular abnormality, along with traction bronchiectasis and honeycombing on the coronal images (E). (F) A magnified view from another patient reveals honeycombing areas occurring in single and multiple layers (indicated by arrows). Additionally, two regions of apparent GGO (circled) contain dilated bronchi (traction bronchiectasis), suggesting these areas likely represent fibrosis. Adapted from [45].

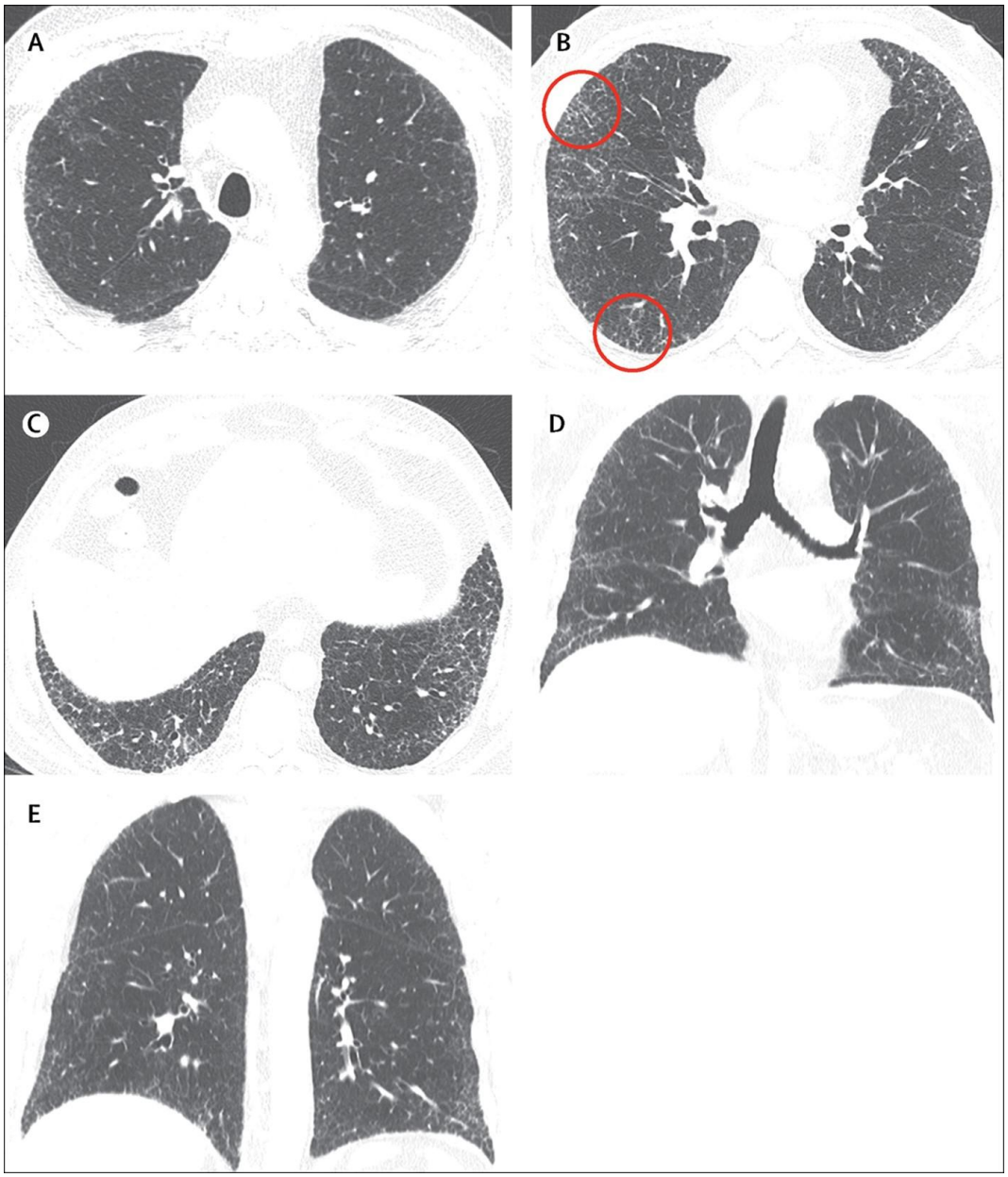


Figure 2.5: Probable UIP HRCT pattern. (A–E) The HRCT images reveal a basal-predominant and subpleural-predominant reticular abnormality, with peripheral traction bronchiectasis (circled in B) but without honeycombing. In this case, the diagnosis of UIP was confirmed through histological analysis. Adapted from [45].

requiring manual feature extraction. Probing the extracted features can also provide insights into the disease mechanisms and progression.

2.4.3 Diagnosis

Pathologically, IPF is characterised by stiffening and scarring (fibrosis) of the lung tissue with unknown causes. This leads to shortness of breath and progressive reductions in lung volume. Diagnosis of IPF requires the existence of UIP pattern and SLB [30, 76]. In patients not subjected to biopsy, the diagnosis is made when a definite UIP is present on the HRCT and other known causes of ILDs are excluded, see Figure 2.6.

IPF diagnosis is challenging due to several reasons. The clinical presentation of IPF, such as dyspnea and cough, overlaps with symptoms of other respiratory diseases like COPD or non-IPF ILDs, leading to potential misdiagnoses [77]. Additionally, even experienced radiologists may struggle to distinguish between UIP patterns and non-UIP patterns on HRCT scans due to subtle differences, increasing the risk of diagnostic errors [47, 78, 79, 80].

The reliance on HRCT and SLB for definitive diagnosis poses challenges [81]. While HRCT can identify UIP patterns, these patterns may not always be distinct, and SLB is invasive and not feasible for all patients, especially those with advanced disease or comorbidities. This diagnostic uncertainty often delays accurate diagnosis and timely treatment initiation.

Machine learning offers opportunities to assist in diagnosing IPF by automating the detection of UIP patterns in HRCT scans and combining imaging data with clinical records [82, 83].

2.4.4 Prognosis

Besides the lack of confident diagnostic criteria of IPF, another challenge is the highly variable and unpredictable progression of IPF across individuals. Disease progression in IPF is assessed by monitoring respiratory symptoms, progressive fibrosis on the HRCT, pulmonary tests (*e.g.*, FVC), or mortality. We explore FVC decline and mortality as established methods for tracking IPF progression.

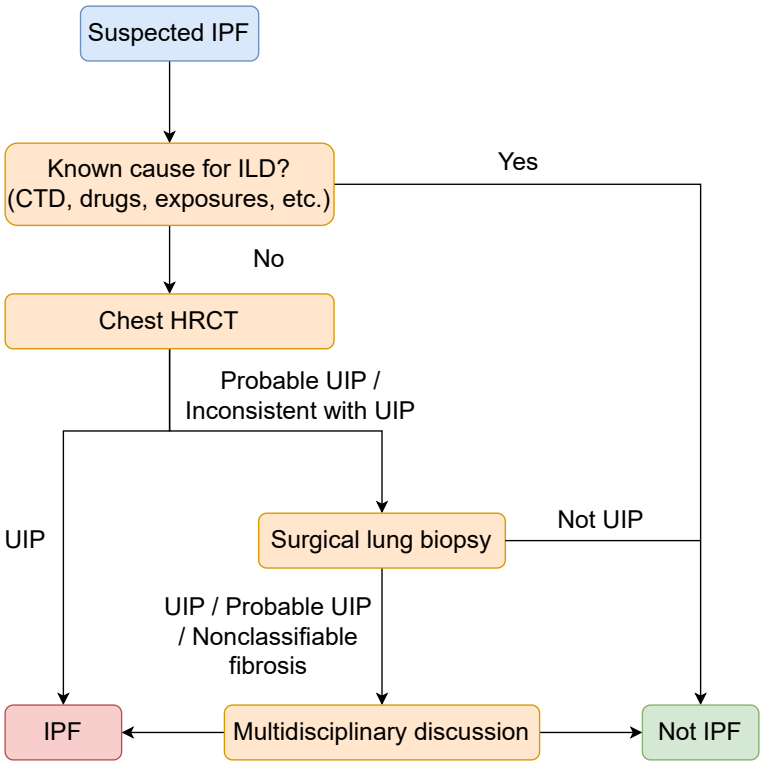


Figure 2.6: Diagnostic criteria for IPF [6]. ILD: Interstitial Lung Disease, CTD: Connective Tissue Disease, HRCT: High-Resolution Computed Tomography, UIP: Usual Interstitial Pneumonia.

2.4.4.1 Forced Vital Capacity Decline

FVC is one of the vital lung function tests used to track the progression of IPF. Importantly, FVC decline is shown to be correlated with patient mortality [84]. Pulmonary function interpretations involve comparing with FVC reference values obtained from a healthy population. Several factors contribute to the calculation of these predicted values, such as height (which reflects chest size), age (reflects maturity), sex, and, ideally, ethnicity [85]. The obtained typical values are then used to calculate the FVC per cent predicted, which is the ratio of the measured FVC to the predicted FVC.

Paterniti *et al.* showed that a decline in FVC of $\geq 10\%$ is associated with mortality [86]. FVC was the basis of the US Food and Drug Administration (FDA) approval

of two major treatments (antifibrotic agents) for IPF: pirfenidone and nintedanib [87]. While FVC decline is consistent in IPF patients, it varies significantly between individuals and over time, and prior declines are not a reliable predictor for future ones [6, 88]. However, a decline of $\geq 10\%$ in the first 24 weeks was shown to predict mortality in the following 24 weeks [89].

Despite the importance of FVC decline in IPF prognosis, it has significant limitations [90, 91]. FVC results can be within the normal range during the early stages of the disease. FVC may also be artificially elevated when emphysema is present [4]. Moreover, FVC depends on the patient's effort and cooperation during the test, leading to variability in the results. Finally, there is an inherent noise in the spirometer measurements, estimated to be around 140 mL in the case of FVC [91].

2.4.4.2 Mortality

Mortality is considered the most reliable endpoint in IPF. It can be interpreted in any of the following forms: all-cause mortality, respiratory-related mortality, or IPF-related mortality. Further, it can be recognised as the time-to-death endpoint or an endpoint at a fixed time (*e.g.*, one year). The most clinically relevant type mentioned is all-cause mortality, which is reliable and easy to define and measure [92]. One concern of mortality studies is that they may be impractical for IPF, as it will require large sample sizes and long duration to reach the endpoint, leading to the need for more resources than non-survival-based studies. However, King *et al.* showed that it is possible to conduct a successful clinical trial using all-cause mortality as the primary endpoint in IPF [93].

We have limited information about mortality predictors in IPF patients. Despite the poor survival rate (two to three years), some patients survive for much longer, and the clinical course varies from slow progression to acute failure and death. It will be clinically beneficial to have prediction models that can yield individual mortality risk [94]. This thesis focuses on all-cause mortality as the primary endpoint for

prognosis prediction in IPF patients.

2.4.4.3 Prognosis Challenges

The course of IPF progression is unpredictable; some patients may experience a rapid decline, while others show a more stable disease course for years. This variability complicates prognosis and makes it challenging to predict patient outcomes based solely on clinical and physiological measures [95]. The lack of reliable prognostic markers hinders the ability to stratify patients based on risk and progression rates.

Traditional prognostic measures, such as FVC decline, are inconsistent predictors of individual outcomes. FVC can remain normal in early-stage IPF, and its decline does not always correlate with disease progression [96, 97]. Additionally, FVC measurements depend on patient effort and can be influenced by comorbid conditions, leading to variability and potential inaccuracies in assessment.

2.4.4.4 Unmet Need for Predictive Models

Given the complexities in diagnosis, the unpredictable nature of disease progression, and the limitations of existing treatment options, there is a pressing need for advanced predictive models to better assist clinicians in managing IPF. The following section explores how machine learning approaches can be leveraged to address these unmet needs.

The diagnostic ambiguity, unpredictable progression, and limited treatment options in IPF highlight an urgent need for reliable predictive models. Traditional clinical approaches and prognostic measures have shown limited effectiveness in accurately predicting patient outcomes, leading to suboptimal treatment strategies. There is a lack of widely accepted prognostic tools that can integrate the diverse and complex data sources available, such as clinical records and HRCT scans.

Machine learning offers a promising solution by leveraging multimodal data to uncover patterns and relationships that may not be apparent through conventional analysis. By integrating clinical and imaging data, predictive models can provide

more accurate diagnoses, assess disease progression, and offer personalised treatment recommendations. These advancements can potentially transform the clinical management of IPF, leading to improved patient outcomes and quality of life.

2.4.5 Treatment

There is currently no cure for IPF. Treatment aims to slow down progression, manage symptoms, and improve the quality of life. Antifibrotic drugs, such as pirfenidone and nintedanib, aim to slow the rate of decline in FVC [98, 99]. However, their effectiveness varies significantly among patients, and side effects can limit their use [100]. Additionally, there is no reliable method to predict which patients will respond favourably to these treatments, complicating clinical decision-making.

On the other hand, lung transplantation is the only reported treatment method to improve both symptoms and survival likelihood [101, 102]. A lung transplant may be either unilateral or bilateral. Outcomes from bilateral transplants are often better regarding survival rate and lung function, but unilateral transplants benefit from shorter wait times and less complicated procedures. The overall five-year survival rate post-transplant is around 50% [101]. However, lung transplantation is only an option for a small subset of patients. The procedure has high risks, limited availability due to donor shortages, and variable outcomes, making it a challenging choice even for eligible patients.

Machine learning models could be used to evaluate treatment effectiveness by analysing longitudinal data from patients receiving antifibrotic treatments. Predictive models could also aid in identifying patients who may benefit most from lung transplantation based on their disease progression and risk factors [103, 104, 105].

2.5 Idiopathic Pulmonary Fibrosis Data

IPF is a complex disease requiring a multidisciplinary diagnosis and management approach. The data used in the clinical setting for IPF diagnosis, prognosis, and

monitoring are primarily clinical data, imaging data, and sometimes genetic data.

Data heterogeneity and missing values are significant obstacles in IPF research. Clinical data often contain gaps due to incomplete patient records, while imaging data can be inconsistent in quality and resolution. The small sample sizes in IPF studies, combined with the variability in patient demographics and disease stages, make it challenging to develop generalisable models.

Clinical and imaging data form the basis for the machine learning models developed in this thesis. By integrating these multimodal data sources, machine learning algorithms can be trained to predict disease progression, assess mortality risk, and support clinical decision-making in IPF.

2.5.1 Clinical Data

Clinical data includes but is not limited to patient demographics, lung function measurements, treatments, and symptoms information.

2.5.1.1 Patient Demographics

Patient demographics include information like patient age, gender, smoking history, exposures, and comorbidities. IPF is more common in older males, and smoking is a significant risk factor for the disease [6]. Environmental exposures, such as silica and wood dust, are also crucial for the disease assessment. Comorbidities, such as emphysema, pulmonary hypertension, and lung cancer, are common in IPF patients and can affect the prognosis [6].

2.5.1.2 Lung Function Measurements

This includes measurements like FVC, FEV1, DL_{CO}, and 6-Minute Walk Test (6MWT). FVC and DL_{CO} are the most common measurements used in IPF diagnosis and prognosis. DL_{CO} measures the lung's ability to transfer gases between the air sacs and the blood. Specifically, it measures how well the Carbon Monoxide (CO), a surrogate for oxygen, is transferred from the lungs to the blood. The 6MWT test

assesses the patient's functional exercise capacity and endurance. It measures how far a patient can walk on a flat surface in six minutes.

2.5.1.3 Treatments

Treatments include whether the patient is under any antifibrotic treatment and which drug. In addition, it should include information about the lung transplantation, if performed.

2.5.1.4 Symptoms

This section includes any symptoms reported by the patient at any point in time. Typical symptoms in IPF include shortness of breath, dyspnea, cough, and clubbed fingers [106].

2.5.2 Imaging Data

Imaging data is essential for IPF diagnosis and prognosis. The primary imaging modality used for IPF is the HRCT. It is a non-invasive imaging technique that provides detailed images of the lung tissue. Non-contrast HRCT with thin slices (≤ 3 mm) is used to detect fibrosis patterns, honeycombing, ground-glass opacities, and other abnormalities in the lung tissue [45]. Figure 2.4 shows examples of HRCT images with different lung patterns.

2.5.3 The Open Source Imaging Consortium Data

The OSIC dataset is the world's largest and most diverse dataset for IPF and ILDs in general. It comprises HRCT and contemporaneous clinical data from multiple sites worldwide. In addition, the dataset includes mortality information for the patients. This thesis uses the OSIC data for developing and evaluating the proposed models.

The OSIC dataset provides a rich, multimodal source of clinical and imaging data suitable for developing robust machine learning models. By using this dataset, machine learning models can learn from real-world data to improve the accuracy and reliability of IPF prognosis predictions.

Criteria	Number of Scans	Number of Patients
All samples	2603	1639
IPF diagnosis	1158	860
Slice thickness < 3 mm	834	621
Volumetric scan	832	619
Non-contrast scan	766	589
Exclude noisy scans	728	555

Table 2.2: Inclusion criteria for the OSIC data used in this thesis.

It contains data from 1639 patients with ILDs[1]. The patient data included in this thesis is collected from six different sites worldwide. Some patients have multiple follow-up scans, while others only have one baseline scan. In this thesis, we filter the data to include only patients with an IPF diagnosis with non-contrast volumetric HRCT scans and a slice thickness of less than 3 mm. For patients with multiple physiological measurements (*i.e.*, FVC, FEV1, and DL_{CO}), we use the average of measurements taken within a 90-day window of the scan date. This procedure ensures that the physiological measurements are as temporally aligned as possible with the scan and reduces the effect of variability in the measurements. If no physiological measurements are available within this window, they are treated as missing. Table 2.2 details the inclusion criteria for the OSIC data used in this thesis. Figure 2.7 shows the distributions of age, gender, smoking status, FVC, and survival status in the OSIC data.

[1] https://www.osicild.org

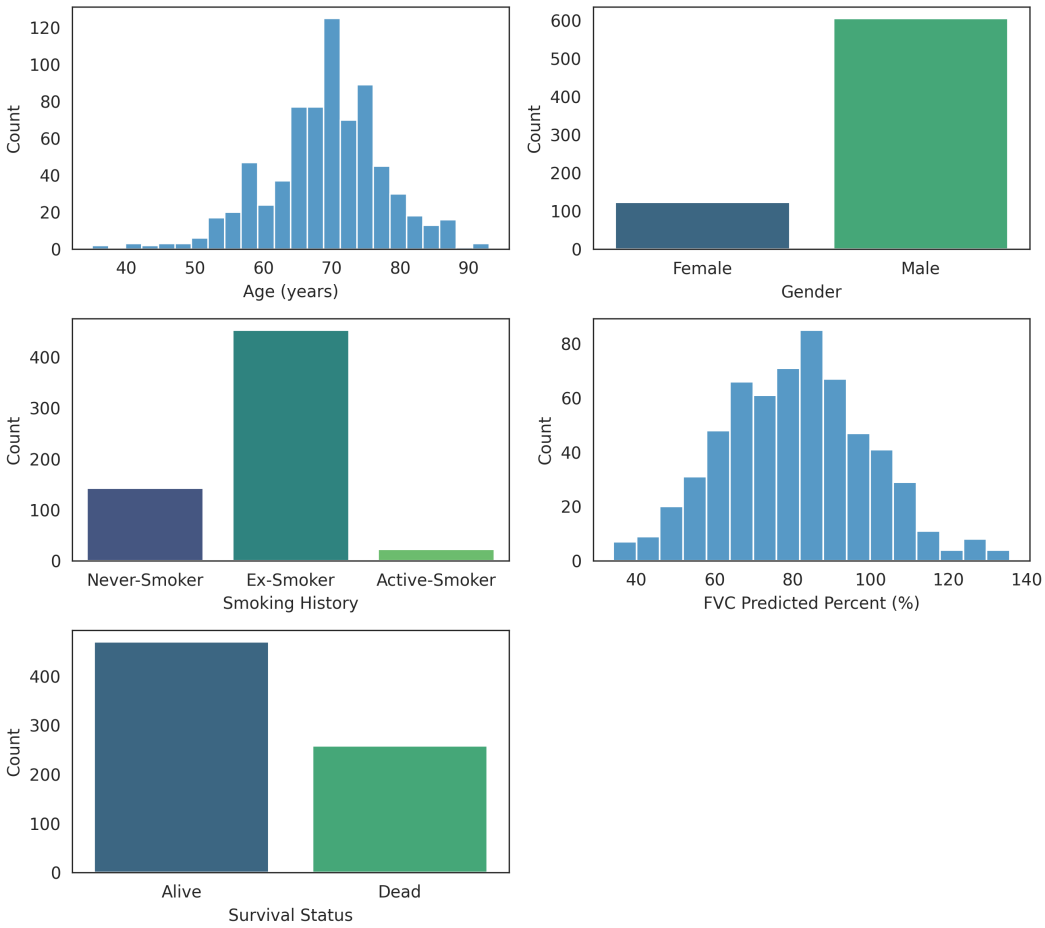


Figure 2.7: Distributions of age, gender, smoking status, FVC, and survival status in the OSIC data.

Chapter 3

Technical Background

In this chapter, we provide a technical overview of machine learning and some relevant methods and models used in the field, focusing on missing data imputation, survival analysis, and multimodal learning.

Machine learning is a subfield of AI that pertains to developing algorithms to learn from data without explicit instructions. The goal is to make predictions or decisions based on the patterns learned from the data. Machine learning has made outstanding progress in various domains, such as computer vision [107, 108, 109, 110], natural language processing [111, 112, 113], and healthcare [114, 115]. We provide an overview of machine learning methods and models relevant to this thesis and refer the reader to [14, 116, 117, 118, 119] for a comprehensive introduction to machine learning.

3.1 Notation

We use bold lowercase letters such as $\mathbf{x}$ for vectors. For matrices, we use bold upper case letters such as $\mathbf{X}$. Scalars are non-bold lowercase, *e.g.*, x. We denote the i-th element of a vector $\mathbf{x}$ as x_i. The i,j-th element of a 2D matrix $\mathbf{X}$ is denoted as x_{ij}. $p(X = x)$ is a probability distribution of the random variable X taking the value x. $p(X = x|Y = y)$ is the conditional probability of X taking the value x given that Y takes the value y. For presentation clarity, we omit the random variables in the

notation when it is clear from the context and use $p(x)$ and $p(x|y)$ instead. $p_\theta(\cdot)$ is a probability distribution parameterized by θ.

3.2 Missing Data Imputation

Missing data is a common problem in real-world datasets. It can occur for various reasons, such as data entry errors, equipment failure, patients not showing up for appointments, or patients not being able to do some tests due to their condition. Missing data can lead to biased conclusions and reduce the statistical power of the models. Therefore, it is essential to handle missing data appropriately.

Missing data imputation is estimating the missing values in the dataset. There are several methods for missing data imputation; see, for example, [116, 120, 121]. However, incautious handling can bias the model adversely. We discuss missing data mechanisms and then present some popular imputation methods.

3.2.1 Missing Data Mechanisms

Understanding the missing data mechanism is crucial for selecting the appropriate imputation method. There are three main missing data mechanisms [122, 123, 124]; we briefly describe them below.

3.2.1.1 Missing Completely at Random (MCAR)

In the Missing Completely at Random (MCAR) mechanism, the probability of a data point being missing is independent of both the observed and unobserved data. The missing data process is entirely random. For example, when a patient misses an appointment due to a random event like a traffic jam, the data is missing completely at random. In the MCAR mechanism, the missing values do not introduce bias into the model.

3.2.1.2 Missing at Random (MAR)

The probability of a data point being missed depends on the observed data but not the missing data themselves. For example, a DL_{CO} test might be more likely to be missing for patients with lower observed FVC values or older patients due to the difficulty of performing the test. In the Missing at Random (MAR) mechanism, the missing data process can introduce bias into the model, but it can be handled by conditioning on the observed data.

3.2.1.3 Missing Not at Random (MNAR)

The probability of data being missed depends on the missing data themselves. The missing values are directly related to the reason they are missing. For example, patients with severe lung function impairment might be more likely to miss lung function tests. In the Missing Not at Random (MNAR) mechanism, the missing data process introduces bias into the model, and it is challenging to handle because the missing data process cannot be modelled by conditioning on the observed data alone.

In this thesis, we assume that the missing data mechanism is MAR, as we condition on the observed data when imputing missing values. This assumption aligns with the proposed method for imputing missing data in the OSIC dataset, as described in Chapter 4.

3.2.2 Imputation Methods

3.2.2.1 Zero Imputation

Zero imputation is the simplest method for handling missing data. It replaces the missing values with zeros. In neural networks, zero imputation sounds reasonable as it prevents the weights associated with the missing nodes from being updated. However, several studies have reported that zero imputation harms the model performance [125, 126, 127]. In addition, zero imputation can introduce bias in the model. For example, it might correlate a missing lung function value with a poor prognosis

due to the inability of patients in late disease stages to perform lung function tests, which is not always true [128]. Zero imputation could be appropriate when the missing data are MCAR. However, it is not recommended for MAR or MNAR missing data mechanisms.

3.2.2.2 Mean Imputation

Mean imputation replaces the missing values in a feature with the mean (the mode in the case of categorical features) of the observed values in that feature [120]. Mean imputation is simple and easy to implement. However, it assumes all data attributes are independent, which is an invalid assumption in the case of IPF clinical records. Mean imputation is suitable for MCAR missing data but can introduce bias in the model when the missing data are MAR or MNAR.

3.2.2.3 Multiple Imputation

Considering dependency between attributes, Multiple Imputation by Chained Equations (MICE) iteratively performs supervised regression to model missing data conditioned on observed data [129]. MICE starts by using a simple imputation method like mean imputation to fill in the missing values. Then, missing values in each feature are regressed given the other features. The process is repeated multiple times to generate multiple imputed datasets. The final imputed dataset is obtained by averaging the multiple imputed datasets. MICE is a popular method for imputing missing data in clinical datasets [121]. It is suitable for MAR missing data but may not be appropriate for MNAR missing data.

3.3 Survival Analysis

Survival analysis is a valuable tool for estimating the time until specific events, such as death or cancer recurrence, based on baseline observations. This is particularly useful in healthcare to prognostically predict clinically important events based on patient data. However, existing approaches often have limitations; some focus only

on ranking patients by survivability, neglecting to estimate the actual event time, while others treat the problem as a classification task, ignoring the inherent time-ordered structure of the events. Furthermore, effectively utilising censored samples, training data points where the exact event time is unknown, is essential for improving the model's predictive accuracy.

This thesis uses survival analysis (or time-to-event prediction) as a proxy for disease prognosis in IPF patients. Given a dataset of patients with observed event times and covariates, the goal is to predict the time until an event of interest occurs for a new patient. The following subsections provide an overview of survival analysis and some popular models used in the field.

3.3.1 Survival Function

The time to event of interest is represented by the random variable T. The survival distribution is typically characterised by three functions: the Probability Density Function (PDF), the survival function, and the hazard function. The PDF $f(t)$ is defined as

$$f(t) = \frac{dF(t)}{dt} \tag{3.1}$$

where $F(t)$ is the cumulative distribution function of the event time $F(t) = p(T \leq t)$.

The survival and hazard functions are equivalent because if one is known, the other can be derived [130]. The survival function $S(t)$ is defined as

$$S(t) = p(T > t) = 1 - F(t) \tag{3.2}$$

It represents the probability that the event of interest has not occurred by time t. The hazard function $h(t)$ is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{p(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)} \tag{3.3}$$

where $p(t \leq T < t + \Delta t | T \geq t)$ is the conditional probability that the event occurs in the interval $[t, t + \Delta t)$ given that the event has not occurred before time t. The hazard function $h(t)$ represents the instantaneous risk of the event occurring at time t given that the event has not occurred before time t. The event in our case is mortality; however, it can be any event of interest, such as cancer recurrence, exacerbation, or machine failure [131, 132, 133, 134].

$f(t)$ can be expressed as

$$f(t) = \frac{dF(t)}{dt} = \frac{d}{dt}(1 - S(t)) = -\frac{dS(t)}{dt} \tag{3.4}$$

then, the hazard function can be expressed as

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}S(t) \cdot \frac{1}{S(t)} = -\frac{d}{dt}\log S(t)^{1} \tag{3.5}$$

Thus, the survival function is related to the hazard function as

$$S(t) = \exp\left(-\int_0^t h(u)du\right) = \exp\left(-H(t)\right) \tag{3.6}$$

where $H(t) = \int_0^t h(u)du$ is the cumulative hazard function. Additionally, the PDF $f(t)$ can be expressed as

$$f(t) = h(t) \cdot S(t) = h(t) \cdot \exp\left(-H(t)\right) \tag{3.7}$$

3.3.2 Censoring

In survival analysis, the task is to predict the time until an event of interest (mortality in our case) occurs from the time of covariates observation. However, in practice, collecting this data for training is only sometimes possible. For example, a patient may drop out of the study, stop visiting the hospital, or the study may end before

[1] Unless otherwise stated, log denotes the natural logarithm $\log(.) = \ln(.)$

the event occurs. In such cases, the event time is unknown and is partially observed in the sense that we know the event has not occurred up to a specific time. This process is called censoring, and samples with this property are called censored samples [131, 135, 136]. Censoring is a common issue in survival analysis, and it is essential to handle it appropriately to avoid biasing the model.

There are three types of censoring, as shown in Figure 3.1:

- Left censoring: The event of interest occurred *before* the study started, and the exact event time is unknown. For example, if we study the time until a patient dies, and the patient dies before the study starts, the event time is left censored. Left censoring is usually unobserved in practice because we usually include only patients alive at the start of the study.

- Interval censoring: The event of interest occurred *between* two time points, and the exact event time is unknown. For example, if we study the time until a patient dies, and the patient dies between two visits to the hospital, and we do not know the exact time of death, the event time is interval-censored.

- Right censoring: The event of interest has not occurred *by* a specific time, and the exact event time is unknown. For example, if we study the time until a patient dies and the patient stops visiting the hospital before death occurs, the event time is right-censored, and the censoring time is the last visit to the hospital. Right censoring is the most common type of censoring in survival analysis.

3.3.3 Data Representation in Survival Analysis

Let $\mathcal{D}$ be the entire dataset, and let $\mathcal{N}$ be the index set of all observations in the dataset, such that each sample is indexed by $n \in \mathcal{N}$. We further define $\mathcal{N}_{\text{uncens}}$ as the subset of indices corresponding to uncensored samples $(\delta_n = 1)$, and $\mathcal{N}_{\text{cens}}$ as the subset of indices corresponding to censored samples $(\delta_n = 0)$. In this thesis,

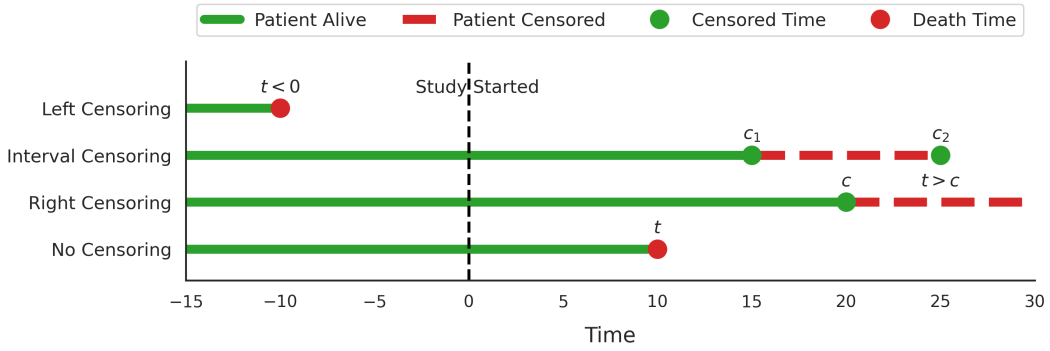


Figure 3.1: Examples of left-censored, interval-censored, right-censored, and uncensored samples.

we focus on right-censored data, as the other forms of censoring are not observed in the OSIC data or most of the other clinical datasets. Our training data $\mathcal{D}$ is a collection of uncensored and right-censored observations. For each sample $n \in \mathcal{N}$, the observation for an uncensored sample is represented as $(\delta_n = 1, \mathbf{x}_n, t_n)$, where $\delta_n = 1$ indicates that the death time t_n is known. For a right-censored sample $n \in \mathcal{N}_{\text{cens}}$, the observation is represented as $(\delta_n = 0, \mathbf{x}_n, c_n)$, where $\delta_n = 0$ indicates that the death time t_n is unknown, and only the censoring time $c_n < t_n$ is known. The covariates $\mathbf{x}_n$ are the patient features (HRCT scans, clinical data, ...), and the time t_n is the time until death. The censoring time c_n is the last visit to the hospital.

3.3.4 Popular Survival Analysis Models

Several models for survival analysis exist, each with strengths and weaknesses.

3.3.4.1 Kaplan-Meier Estimator

The Kaplan-Meier (KM) estimator is the most widely used non-parametric survival analysis model [137]. It estimates the survival function by computing the proportions of individuals surviving over time

$$\hat{S}_{\text{KM}}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{3.8}$$

where d_i is the number of deaths at time t_i and n_i is the number of individuals at risk at time t_i. The KM estimator is a step function that decreases at each event time. It is a non-parametric model that does not assume any specific form for the survival function. The KM estimator is useful for visualising the survival function and comparing survival functions between different groups. However, it is not suitable for predicting the survival time for new patients as it cannot account for the effects of covariates.

3.3.4.2 Cox Proportional Hazards Model

The most widely used model to learn from censored survival data is the CoxPH model [26]. CoxPH models the conditional hazard function $h(t|\mathbf{x})$ given the covariates $\mathbf{x}$ as

$$h(t|\mathbf{x}) = h_0(t)\exp\left(\beta^T\mathbf{x}\right) \tag{3.9}$$

where $h_0(t)$ is the baseline hazard function, and β are the model parameters. Given two patients with covariates $\mathbf{x}_i$ and $\mathbf{x}_j$, the hazard ratio is

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_j)} = \frac{h_0(t)\exp(\beta^T\mathbf{x}_i)}{h_0(t)\exp(\beta^T\mathbf{x}_j)} = \exp(\beta^T(\mathbf{x}_i - \mathbf{x}_j)) \tag{3.10}$$

The proportional hazards assumption assumes that the hazard ratio between two patients is constant over time [138].

The model is semi-parametric, as it does not assume a specific form for the baseline hazard function $h_0(t)$ (which does not depend on the covariates) but assumes a linear relationship between the covariates and the log hazard.

The model parameters β are learned by maximising the partial log-likelihood function [26]. To do this, for each patient n, we define the risk set $\mathcal{R}_n$ as all those patients that have not died before patient n and define the relative death risk as

$$p(D_n = t_n|\mathcal{R}_n) = \frac{h(t_n|\mathbf{x}_n)}{\sum_{m\in\mathcal{R}_n} h(t_m|\mathbf{x}_m)} = \frac{\exp(\beta^T\mathbf{x}_n)}{\sum_{m\in\mathcal{R}_n}\exp(\beta^T\mathbf{x}_m)} \tag{3.11}$$

The partial log-likelihood function is then defined as the sum of $\log P(D_n = t_n|\mathcal{R}_n)$ for all patients who died $n \in \mathcal{N}_{\text{uncens}}$

$$\mathbb{L}(\beta) = \sum_{n \in \mathcal{N}_{\text{uncens}}} \log p(D_n = t_n|\mathcal{R}_n) \tag{3.12}$$

$$= \sum_{n \in \mathcal{N}_{\text{uncens}}} \left(\beta^T \mathbf{x}_n - \log \sum_{m \in \mathcal{R}_n} \exp(\beta^T \mathbf{x}_m) \right) \tag{3.13}$$

The CoxPH model has been widely used in survival analysis due to its simplicity and interpretability. However, it has several limitations. The main limitation is its assumption of a linear relationship between the covariates and the log hazard, which may not hold in practice. In addition, the proportional hazards assumption is a strong assumption that may not hold in some cases, especially in a disease like IPF where the risk of death may change over time and the progression of the disease is heterogeneous and highly unpredictable [139, 30].

Further, the CoxPH model estimates the relative risk of death between patients rather than predicting the actual survival time, which is more useful and easier to interpret. This is because the CoxPH model is a semi-parametric model that does not estimate the baseline hazard function $h_0(t)$, which is required to predict the survival time. Non-parametric methods like the Breslow estimator [140] are often used to estimate the baseline hazard function and then compute the hazard function and survival function. However, the performance of this method is unsatisfactory in practice [131, 141].

3.3.4.3 Random Survival Forests

Survival Decision Trees (SDTs) are developed by modifying standard decision trees with specialised splitting rules to accommodate right-censored survival data. Numerous splitting rules have been suggested [142]. They can be generally classified into two categories: splitting rules that maximise heterogeneity between nodes [143, 144, 145] and those that minimise homogeneity within nodes [146, 147, 148]. To

address the instability of individual SDTs, Random Survival Forests (RSFs) [149] were introduced, which ensemble multiple random SDTs based on Breiman's random forest algorithm [150].

3.3.4.4 Gradient Boosting Machines

Boosting is a widely used ensemble learning technique that combines the predictions of multiple weak models (often called base learners) to create a more robust model. Gradient Boosting Machines (GBMs) [151] are among the most popular boosting methods and can be adapted for survival analysis by incorporating base survival models. For instance, CoxBoost was developed to estimate the coefficients of a Cox model using GBMs [152].

3.3.4.5 Support Vector Machines

Support Vector Machines (SVMs) [153] are commonly used in classification tasks, aiming to identify the optimal hyperplane separating different classes. SVMs can also be adapted for regression tasks [154], seeking a hyperplane that best fits the data while minimising error, a variant known as Support Vector Regression (SVR). SVR can predict survival time, rank scores, or both, though it does not model the survival distribution. Shivaswamy *et al.* [155] were the first to modify SVR for survival analysis, proposing a support vector approach for regression with censored targets, specifically to predict survival times within a target interval. Later, [156] reformulated survival analysis as a ranking problem with penalties for discordant pairs, and [157] introduced an SVR-based hybrid model that approaches survival prediction as both a regression and ranking problem.

3.3.4.6 Bayesian Survival Analysis

Bayesian methods offer a framework for inference and prediction by connecting posterior and prior probabilities. In survival analysis, Bayesian approaches are frequently used to predict the survival distribution. For example, Fernández *et*

al. [158] proposed a semi-parametric Bayesian model for survival analysis, designed to avoid strong constraints. Their approach modelled the hazard function as a product of a parametric baseline hazard and a non-parametric component, which employed a Gaussian process to capture the combined effects of time and covariates.

3.3.4.7 DeepSurv

One limitation of the CoxPH model is that it assumes a linear relationship between the covariates and the log-hazard. DeepSurv relaxes this assumption by using a neural network to model the hazard function [159]

$$h(t|\mathbf{x}) = h_0(t)\exp\left(f_\theta(\mathbf{x})\right) \tag{3.14}$$

where $f_\theta(\mathbf{x})$ is a neural network parameterized by θ. The model parameters θ are learned by minimising the negative log-partial likelihood function, similar to Equation 3.13. As a deep learning model, DeepSurv automatically learns the relevant features from the data and does not require any manual feature engineering [159].

3.3.4.8 Classical Censoring Model

In contrast to approaches that model the hazard function, other methods model the death distribution directly. The model is trained to maximise the likelihood of the observed death and censoring times. However, as we are interested in predicting the death times, one needs to assume the data generation process to model the censoring times [131].

One common approach in the literature is to assume that censoring times follow a distribution $p(C = c|x)$ and death times follow a distribution $p_\theta(D = t|x)$. These times are independently sampled and then compared: if the censoring time is less than the death time, the observation is the censoring time; otherwise, it is the death time [25, 160]. This approach is called the classical censoring model and leads to

the following model

$$p_\theta(\delta, c, t|x) = p_\theta(t|x)p(c|x)p(\delta|c,t) \tag{3.15}$$

where $p(\delta = 1|c,t) = 1$ if $c \geq t$ and $p(\delta = 0|c,t) = 1$ if $c < t$. For a uniform censoring distribution $p(C = c|x) = \frac{1}{T_{\max}}$ a right-censored observation then has the following likelihood[2]

$$p_\theta(\delta = 0, C = c|x) = \frac{1}{T_{\max}} \sum_{t=c+1}^{T_{\max}} p_\theta(D = t|x) \tag{3.16}$$

and the likelihood of an uncensored observation is given by

$$p(\delta = 1, D = t|x) = \frac{T_{\max} - t + 1}{T_{\max}} p_\theta(D = t|x) \tag{3.17}$$

Omitting additive constants, the objective then is to maximise

$$\mathbb{L}(\theta) \equiv \sum_{n \in \mathcal{N}_{\text{uncens}}} \log p_\theta(D = t_n|x_n) + \sum_{i \in \mathcal{N}_{\text{cens}}} \log \sum_{t=c_i+1}^{T_{\max}} p_\theta(D = t|x_i) \tag{3.18}$$

The model parameters θ are learned by maximising the likelihood function in Equation 3.18.

3.3.4.9 DeepHit

Lee *et al.* approach survival analysis as a classification task with $T_{\max}$ categories [25]. Specifically, a neural network predicts a vector of $T_{\max}$ values, which a softmax function then transforms into a death distribution, $p_\theta(D = t|x)$. DeepHit combines a classical censoring term (Equation 3.18) and a ranking objective to leverage the

[2]Any other censoring distribution can be used here, and it can also be learned from the data. However, for simplicity, we use a uniform distribution.

censored data. Specifically, the objective function is composed of two terms

$$\mathbb{L}_{\text{DeepHit}} = \mathbb{L}^{c}_{\text{lik.}} + \mathbb{L}_{\text{rank.}} \tag{3.19}$$

where $\mathbb{L}^{c}_{\text{lik.}}$ represents the classical likelihood (Equation 3.18) with a softmax function to model the death time distribution, and $\mathbb{L}_{\text{rank.}}$ is a ranking term that penalises the model for inaccuracies in predicting the ranking of patients' survival times, mirroring the Cox objective

$$\mathbb{L}_{\text{rank.}} = \eta(F_\theta(t_i|x_i), F_\theta(t_i|x_j)) \quad \forall i,j \in \mathcal{N} \quad \text{s.t.} \quad t_i < t_j \tag{3.20}$$

where $\eta(x,y) = \exp\left(\frac{-(x-y)}{s}\right)$, $F_\theta(t|x)$ represents the cumulative distribution function of the predicted distribution $p_\theta(t|x)$. The model parameters θ are learned by maximising the likelihood function $\mathbb{L}_{\text{DeepHit}}$.

3.3.5 Evaluation Metrics

3.3.5.1 Mean Absolute Error

The Mean Absolute Error (MAE) assesses the difference between death times predicted by the model and the true death times and is computed for uncensored samples

$$\text{MAE} = \frac{1}{|\mathcal{N}_{\text{uncens}}|} \sum_{i \in \mathcal{N}_{\text{uncens}}} |\hat{t}_i - t_i| \tag{3.21}$$

where $\hat{t}_i$ is the predicted death time for patient i.

3.3.5.2 Relative Absolute Error

Similarly, the Relative Absolute Error (RAE), which quantifies the relative deviation of the predicted time from the true death time

$$\text{RAE} = \frac{1}{|\mathcal{N}_{\text{uncens}}|} \sum_{i \in \mathcal{N}_{\text{uncens}}} \frac{|\hat{t}_i - t_i|}{t_i} \tag{3.22}$$

3.3.5.3 Concordance Index

The Concordance Index (C-Index) estimates the probability that the predicted risks or survival times of a randomly chosen pair of patients will have the same ordering as their actual survival times [161]. C-Index is a rank-correlation metric that assesses the model's ability to accurately rank individuals according to their survival times. It measures how effectively the model differentiates between high-risk and low-risk individuals. For a pair of patients i and j, whose true survival times are t_i and t_j, and the predicted survival times are $\hat{t}_i$ and $\hat{t}_j$, the concordance probability is

$$\mathrm{C} = p(\hat{t}_j > \hat{t}_i | t_j > t_i) \tag{3.23}$$

The C-Index is then defined as the fraction of concordant pairs to all pairs

$$\text{C-Index} = \frac{\sum_{i \neq j} \mathbb{I}(\hat{t}_j > \hat{t}_i)\mathbb{I}(t_j > t_i)\delta_i}{\sum_{i \neq j} \mathbb{I}(t_j > t_i)\delta_i} \tag{3.24}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The formula can be written in simpler terms, such as

$$\text{C-Index} = \frac{\#\text{concordant pairs}}{\#\text{concordant pairs} + \#\text{discordant pairs}} \tag{3.25}$$

A pair is considered concordant if the ranking predicted by the model matches the true ranking and discordant if it does not. A perfect model will have a C-Index=1.0. It is worth noting that the C-Index is a ranking metric, which only assesses the order in which the predicted values should be ranked compared to the true ranking. It does not evaluate the accuracy of the predicted values themselves. Therefore, the CoxPH model, which only estimates the relative risk of death, can achieve a high C-Index even if the predicted values of the death times are inaccurate.

Despite being widely used in survival analysis, the C-Index has notable limitations. First, it solely evaluates the ranking of predicted survival times and does not account for the magnitude of errors in absolute time-to-event predictions. This means

that a model with highly inaccurate survival time estimates can still achieve a high C-Index as long as the ranking is preserved. Second, C-Index does not fully account for censored samples, which can introduce bias, particularly when the censoring rate is high [162]. This sensitivity to censored data limits its effectiveness in real-world clinical applications, where censored data is common.

3.4 Whole HRCT Scans for Prognostic Modelling in IPF

Due to the high memory requirements for processing full 3D images, several studies used features extracted from the HRCT scans by expert radiologists [83, 163] or quantitative analysis tools [164, 165]. These extracted features are human-defined (*e.g.* honeycombing, reticulation, GGO) and provide a structured representation of imaging findings while reducing the computational demands of model training.

However, using extracted features from the HRCT scans has several limitations. First, it relies on features manually defined by radiologists, which introduces biases and inter-observer variabilities. Second, predefined features may not fully capture the complexity of lung abnormalities in IPF, potentially missing critical prognostic patterns. Third, this approach is constrained by current medical knowledge, limiting the discovery of novel imaging biomarkers. Finally, manual feature extraction is time-consuming and does not scale efficiently for large datasets.

In this thesis, we adopt an end-to-end learning approach using full HRCT scans as model inputs. This allows the model to learn directly from the imaging data, capturing subtle spatial and textural patterns that may be missed in manual feature selection. By eliminating the need for handcrafted features, we reduce inter-observer variability and provide a more objective, reproducible method for prognosis prediction. Additionally, interpretability techniques can reveal new imaging biomarkers that contribute to disease progression, as discussed in Subsection 7.2.5 [166], by

visualising the regions of the HRCT scans that the model uses for prediction. This approach also removes the need for time-consuming manual feature extraction, making it feasible for large-scale datasets.

Despite these advantages, processing full 3D HRCT scans presents computational challenges, particularly regarding GPU memory limitations and batch size constraints. To address this, we leverage memory banks to optimise training efficiency, as detailed in Chapter 5.

3.5 Multimodal Learning

In the context of IPF prognosis, combining HRCT scans and clinical data is particularly advantageous because these modalities offer complementary information. HRCT scans provide detailed insights into the structural abnormalities and progression of lung fibrosis, capturing visual patterns indicative of disease severity. On the other hand, clinical data (*e.g.*, pulmonary function tests, demographic information) offer a comprehensive view of the patient's systemic health and underlying risk factors. By integrating both sources, the model can leverage structural and systemic indicators, improving predictive accuracy and robustness compared to using either modality alone.

Multimodal machine learning aims to build models to learn and make predictions from multiple data modalities (*e.g.*, images, text, audio, ...) [18]. Learning from multiple modalities is vital in many applications, such as healthcare, where patient data is collected in several forms, such as medical images, clinical data, and genetic data. There are several ways to combine multiple modalities, such as early and late fusion [18].

3.5.1 Early Fusion

Early fusion is one of the most straightforward approaches to multimodal learning. In early fusion, inputs are combined (*e.g.*, by concatenation) before being fed into

the model. The model then learns the relationship between the combined inputs and the target output. Early fusion is simple and easy to implement but may not capture the complex relationships between the modalities, potentially leading to suboptimal performance.

3.5.2 Late Fusion

In contrast, late fusion keeps the modalities separate and learns separate representations for each modality. The model then combines the learned modality representations to make the final prediction. Late fusion is more flexible than early fusion, as it allows the model to learn separate representations for each modality.

One should make sensible choices for model branches that learn the representations of each modality. This thesis uses late fusion to combine the HRCT scans and clinical data. We use a Convolutional Neural Network (CNN) to learn the image representations and a feedforward neural network to learn the clinical data representations. We then combine the learned representations using a Multi-Layer Perceptron (MLP) to predict mortality in IPF patients.

Chapter 4

Latent Variable Models for Missing Data Imputation

4.1 Introduction

Patient clinical data is vital for the diagnosis and prognosis of IPF. These include patient demographics, physiological measurements, and treatment history. One of the main challenges in using clinical data for prognostic modelling is the presence of missing data. Most clinical records in the OSIC dataset contain at least one missing value. Consequently, training models on only complete samples would drastically reduce the amount of training data and negatively impact the subsequent model performance. In addition, this would also introduce bias in the model and the drawn conclusions.

This chapter proposes a novel approach to imputing missing clinical data in IPF patients using LVMs. We first describe the clinical data used in the thesis and the missing data patterns. We then present the LVM and the proposed imputation method. Finally, we evaluate the performance of the proposed method on the OSIC dataset. We use the method explained in this chapter to impute missing data in the OSIC dataset used in the experiments in Chapter 5 and Chapter 6.

4.2 Clinical Data Used in the Thesis

As explained in Section 2.5.1, clinical data is integral to the diagnosis and prognosis of IPF. Following the clinical guidelines [6, 167], we use the following clinical variables in the thesis

- **Demographics**:
  - Patient's age at the time of the HRCT scan.
  - Patient sex.
  - Patient smoking status (current, former, or never-smoker).
- **Physiological Measurements**:
  - Percent predicted FVC within 3 months of the HRCT scan.
  - DL_{CO} within three months of the HRCT scan (not corrected for haemoglobin).
- **Treatment History**:
  - Treatment with antifibrotic drugs (pirfenidone or nintedanib) before or at the time of the HRCT scan.

4.3 Missing Data Patterns

The clinical data used in this thesis is often incomplete due to various reasons such as missing measurements, data entry errors, or patient inability to perform specific tests. For example, IPF patients in the late stages of the disease may not be able to perform the DL_{CO} test due to their poor health condition. Table 4.1 shows the percentage of missing values for each clinical variable in the OSIC dataset. While the age and sex of the patients are complete, the DL_{CO} and antifibrotic treatment variables have a high percentage of missing values. The FVC predicted variable also has many missing values.

Clinical Variable	Percentage of Missing Values
Age	0.0%
Sex	0.0%
Smoking History	10.8%
FVC Predicted	32.4%
DL_{CO}	74.2%
Antifibrotic	85.9%

Table 4.1: Percentage of missing values for each clinical variable in the OSIC dataset.

Standard imputation methods, such as mean, median, or zero imputation (see Section 3.2), are not suitable for imputing missing data in clinical IPF records. These methods do not consider the relationships between the variables and assume that they are independent. However, there is a dependency between the clinical variables. For example, the percent predicted FVC and DL_{CO} measurements are correlated [168, 169, 170]. In addition, the computation of the predicted FVC values depends on the patient's age and gender, so there is a relationship between these variables [59, 171]. It is also sensible to assume that the antifibrotic treatment is related to the FVC and DL_{CO} measurements, as clinicians prescribe these treatments based on the patient's lung function [168, 169, 171].

In this chapter, we assume an MAR mechanism, where the likelihood of missing values depends on the observed data but not directly on the missing values themselves. This assumption is reasonable in our context, as the clinical variables used for imputation (*e.g.*, FVC, DL_{CO}, age, and sex) are correlated and can provide informative cues for estimating the missing values. While this approach may not fully capture scenarios where missing data depends on unobserved factors (*i.e.*, Missing Not at Random), using latent variables helps mitigate some of these effects by capturing underlying patterns in the observed data.

To this end, we propose a novel approach to impute missing clinical data using LVMs. We use LVMs to model the relationships between the clinical variables and impute the missing values based on the learned relationships. The following section

describes the LVM used in this thesis and the proposed imputation method.

4.4 Latent Variable Model

We introduce a LVM to model the relationships between the clinical variables and impute the missing values. To impute missing values, we assume the clinical features $\mathbf{x}$ are modelled by independent categorical distributions when conditioned on a hidden state h, see Figure 4.1. For patient $n \in \{1,\ldots,N\}$, the probability of clinical record $\mathbf{x}^n$ under the model is therefore given by

$$p(\mathbf{x}^n) = \sum_{h=1}^{H} p(h) \prod_{k=1}^{K} p(x_k^n|h) \tag{4.1}$$

where $p(h)$ denotes a categorical distribution with state $h \in \{1,\ldots,H\}$; K is the number of clinical features, and $p(x_k^n|h)$ is a categorical distribution. Writing each record in terms of observed and missing elements, $\mathbf{x} = (\mathbf{x}_o, \mathbf{x}_m)$, the likelihood of record $\mathbf{x}^n$ is given by

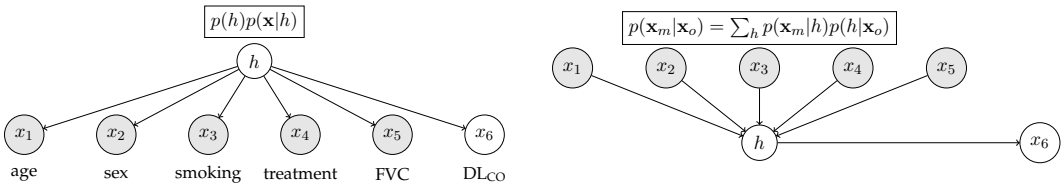
$$p(\mathbf{x}^n) = \sum_{h} p(\mathbf{x}_o^n|h) p(\mathbf{x}_m^n|h) p(h) \tag{4.2}$$

where

$$p(\mathbf{x}_o^n|h) = \prod_{i \in \mathbf{x}_o^n} p(x_i|h) \tag{4.3}$$

$$p(\mathbf{x}_m^n|h) = \prod_{i \in \mathbf{x}_m^n} p(x_i|h) \tag{4.4}$$

To model continuous features, we convert them into discrete variables by equal-frequency binning and model them as categorical variables. Equal-frequency binning ensures that all bins have the same number of samples, which helps to capture the underlying distribution of the continuous variables. We train the model using the EM algorithm [172] to learn the hidden distribution $p(h)$ and the categorical distributions $p(x_i|h)$. The model can then impute the missing values in the clinical records.



(a) Generative model. **(b)** Imputation of missing values.

Figure 4.1: Latent variable model for imputing missing clinical data in IPF records.

4.4.1 Training the Latent Variable Model

The model has two sets of parameters, the hidden distribution $p(h)$ and the categorical distributions $p(x_i|h)$. The EM algorithm [172] is a convenient choice to learn these distributions. Note that the EM algorithm can make use of all training data, even records that contain missing data.

The EM algorithm maximises the energy term (see [116]), given a posterior $q(h|\mathbf{x})$

$$\sum_{n,h} \mathbb{E}_{q(h|\mathbf{x}^n)} \log p(\mathbf{x}^n, h) = \sum_{n,h} \sum_{i \in \mathbf{x}^n} \mathbb{E}_{q(h|\mathbf{x}^n)} \log p(x_i|h) + \sum_{n,h} \mathbb{E}_{q(h|\mathbf{x}^n)} \log p(h) \tag{4.5}$$

where $q(h|\mathbf{x}^n)$ is given by the E-step

$$q(h|\mathbf{x}^n) \propto p(h)p(\mathbf{x}^n|h) \propto p(h) \prod_{i \in \mathbf{x}_o^n} p(x_i|h) \tag{4.6}$$

The E-step computes the posterior distribution of the hidden states h given the observed data x^n. The M-step maximises the energy term in Equation 4.5 with respect to the model parameters by updating the hidden distribution $p(h)$ and the categorical distributions $p(x_i|h)$, as follows

$$p(h) \propto \sum_n q(h|\mathbf{x}^n) \tag{4.7}$$

$$p(x_i = k|h) \propto \sum_n \mathbb{I}(x_i^n = k) q(h|\mathbf{x}^n) \tag{4.8}$$

The E-step and M-step are iterated until convergence. We can then use the learned model to compute the distribution of the missing values given the observed data

$$p(\mathbf{x}_m^n|\mathbf{x}_o^n) = \sum_h p(\mathbf{x}_m^n|h)p(h|\mathbf{x}_o^n) \propto \sum_h p(\mathbf{x}_m^n|h)p(h) \prod_{i \in \mathbf{x}_o^n} p(x_i^n|h) \tag{4.9}$$

Calculating missing data statistics or drawing samples as required is then straightforward. During the training of subsequent models that use the imputed clinical data, we sample the missing values from the distribution in Equation 4.9 to account for the uncertainty in the imputed values. During inference, we use the expectation of $p(\mathbf{x}_m^n|\mathbf{x}_o^n)$ to impute the missing values.

4.5 Experiments

4.5.1 Data

We evaluate the performance of the proposed method on the OSIC dataset. We include all IPF records in the dataset and use the clinical variables described in Section 4.2. This results in a dataset of 1853 records from 1484 patients. We divide the available records into training and test sets. We selected records with at most one missing value and split them on a patient level into training (80%) and test (20%) sets. The remaining records with more than one missing value are added to the training set. To evaluate the model performance, we drop one of the clinical variables from each record in the test set and impute the missing values using the different imputation methods described in Section 3.2. We assume that age and sex are always observed as they are complete in the dataset and are usually available in clinical records. Therefore, we only impute the missing values for the smoking history, FVC predicted, $\mathrm{DL_{CO}}$, and antifibrotic treatment variables.

4.5.2 Evaluation Metrics

Missing values include features from different types, such as categorical (smoking history and antifibrotic treatment) and continuous (predicted FVC and DL_{CO}). We use suitable evaluation metrics for each feature type and describe them below.

4.5.2.1 Accuracy

Binary accuracy is the simplest and most common evaluation metric for categorical features. It is defined as the proportion of correctly imputed values over the total number of imputed values. Given the imputed values $\hat{x}_i$ and the true values x_i, the binary accuracy is given by

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{x}_i = x_i) \tag{4.10}$$

Where N is the number of imputed values. Binary accuracy ranges from 0 to 1, where a higher accuracy indicates better imputation performance. In imbalanced datasets, accuracy can be misleading as it does not account for the class distribution. Therefore, we also report the F1-score.

4.5.2.2 F1-score

The F1-score is the harmonic mean of precision and recall. It is a suitable metric for imbalanced datasets because it considers both false positives and false negatives and consequently provides a balanced evaluation. Similar to accuracy, the F1-score ranges from 0 to 1, where a higher score indicates better imputation performance. The F1-score is defined as

$$\text{F1-score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{4.11}$$

TP, FP, and FN are the true positives, false positives, and false negatives, respectively. We calculate the F1-score for each class and report the average F1-score across all

classes in a given categorical feature.

4.5.2.3 Mean Absolute Error

We use the MAE as an evaluation metric for continuous features. The MAE measures the average absolute difference between the imputed values and the true values

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{x}_i - x_i| \tag{4.12}$$

The MAE has values from 0 to ∞, where a lower MAE indicates better imputation performance.

4.5.2.4 Normalised Root Mean Squared Error

The Normalized Root Mean Squared Error (NRMSE) is another evaluation metric for continuous features. The NRMSE is the square root of the mean squared error divided by the range of the true values

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2}}{\max(\mathbf{x}) - \min(\mathbf{x})} \tag{4.13}$$

The NRMSE has values from 0 to ∞, where a lower NRMSE indicates better imputation performance. Compared to the MAE, the NRMSE penalises significant errors more heavily.

4.5.3 Implementation Details

To evaluate the imputation methods, we simulate missing values by dropping a feature from each validation sample and imputing it using mean imputation (Section 3.2.2.2), MICE (Section 3.2.2.3), and our proposed latent variable model (Section 4.4).

For the LVM, the prediction is the expectation of the posterior distribution of the missing values given the observed data $p(\mathbf{x}_m|\mathbf{x}_o)$ (see Equation 4.9). We set the number of hidden states $H = 10$ and train the model until the improvement in the training log-likelihood is less than 10^{-8}. Using equal-frequency binning, we

Table 4.2: Imputation results for categorical features. The table shows the imputation performance where we drop one of the categorical features from each record in the validation set and impute it using the different imputation methods. We report the results as the mean and standard deviation over five folds. The best results are highlighted in bold. The higher, the better.

Method	Smoking		Antifibrotic	
	F1-score	Accuracy	F1-score	Accuracy
MICE	0.3776 ± 0.0474	0.5446 ± 0.0294	0.5581 ± 0.0245	0.5780 ± 0.0191
Mean	0.2587 ± 0.0084	0.6346 ± 0.0338	0.3750 ± 0.0331	0.6034 ± 0.0815
LVM (ours)	$\mathbf{0.4026 \pm 0.0174}$	$\mathbf{0.6820 \pm 0.0101}$	$\mathbf{0.6005 \pm 0.0427}$	$\mathbf{0.6515 \pm 0.0633}$

Table 4.3: Imputation results for continuous features. The table shows the imputation performance where we drop one of the continuous features from each record in the validation set and impute it using the different imputation methods. We report the results as the mean and standard deviation over five folds. The best results are highlighted in bold. The lower, the better.

Method	FVC		DL_{CO}	
	MAE	NRMSE	MAE	NRMSE
MICE	20.5322 ± 1.1288	0.2653 ± 0.0282	1.4751 ± 0.1382	0.2622 ± 0.0279
Mean	15.4792 ± 0.5193	0.1930 ± 0.0186	1.1434 ± 0.0588	0.2026 ± 0.0181
LVM (ours)	$\mathbf{14.5654 \pm 0.6270}$	$\mathbf{0.1819 \pm 0.0169}$	$\mathbf{1.0379 \pm 0.0959}$	$\mathbf{0.1906 \pm 0.0184}$

discretise the age values into 6 bins, the FVC into 8 bins, and the DL_{CO} into 6 bins. These hyperparameters were selected based on the model's performance on one fold of the validation set. We implemented the LVM in Python using NumPy library [173].

For the MICE method, we use the implementation in the statsmodels library [174]. We set the number of imputations to 10 and the number of iterations to 10. We use the default linear regression model for imputation.

4.5.4 Results

In Table 4.2 and Table 4.3, we report the imputation performance for the categorical and continuous features, respectively. These results represent the imputation performance where we drop one of the features from each record in the validation set and impute it using the different imputation methods. We report the results as the mean

and standard deviation over five folds to account for the data variability.

The results show that the proposed LVM method outperforms the mean and MICE imputation methods for both categorical and continuous features. For the categorical features, the LVM method achieves the highest F1-score and accuracy for both smoking history and antifibrotic treatment. The LVM method achieves an F1-score of 0.4026 and accuracy of 0.682 for smoking history and an F1-score of 0.6005 and accuracy of 0.6515 for antifibrotic treatment. The mean imputation method achieves the lowest F1-score for both features because it assigns the same value to all missing values, the mode of the feature values in the training set. The MICE method performs better than the mean imputation method in terms of F1-score but is outperformed by the LVM method.

For the continuous features, the LVM method achieves the lowest MAE and NRMSE for both FVC and DL_{CO}. The LVM method achieves an MAE of 14.5654 and NRMSE of 0.1819 for FVC and an MAE of 1.0379 and NRMSE of 0.1906 for DL_{CO}. The mean imputation method achieves the second-best performance for both FVC and DL_{CO}, while the MICE method achieves the worst performance. The MICE method uses linear regression to impute the missing values, which may not capture the complex relationships between the clinical variables.

These results demonstrate the superior performance of the proposed model compared to the standard imputation methods. The LVM method captures the relationships between the clinical variables and imputes the missing values based on the learned relationships.

In addition to the superior performance, the LVM imputation method outputs a distribution of the missing values given the observed data rather than point estimates, as in the mean and MICE imputation methods. This probabilistic approach enhances the robustness to noisy clinical data, as it prevents over-reliance on fixed, potentially erroneous imputed values. Further, it provides a measure of uncertainty in the imputed values, which can be used to make more informed predictions. During our

experiments in Chapter 5 and Chapter 6, we sample the missing values from the distribution output by the LVM method to account for the uncertainty in the imputed values. This variability ensures that minor distortions in the data do not lead to overconfident predictions, a feature not available in mean and MICE imputation.

4.6 Conclusion

In this chapter, we proposed a novel approach to impute missing clinical data in IPF records using LVMs. We introduced an LVM to model the relationships between the clinical variables and impute the missing values based on the learned relationships. The proposed method is beneficial for imputing missing data in clinical IPF records, where the variables are correlated, and the relationships between the variables are essential for the diagnosis and prognosis of the disease. Our experiments on the OSIC dataset demonstrated the superior performance of the proposed method compared to the standard imputation methods. The LVM is used in the subsequent chapters to impute missing data in the OSIC dataset and train prognostic models for IPF using both clinical and imaging data. In Chapter 5 and Chapter 6, we sample the missing values from the distribution output by the LVM method to account for the uncertainty in the imputed values and make more informed predictions about the disease progression and patient survival.

Chapter 5

Improving Cox proportional hazards Model with Memory Banks

5.1 Introduction

CoxPH model is the most popular model for survival analysis. However, one limitation of the CoxPH model is that it assumes a linear relationship between the covariates and the hazard function. This assumption may not hold in practice, especially when the covariates are high-dimensional. To address this limitation, previous studies have proposed to use deep learning models to learn the non-linear relationship between the covariates and the hazard function [159]

$$h_\theta(t|\mathbf{x}) = h_0(t)\exp\left(f_\theta(\mathbf{x})\right) \tag{5.1}$$

In this case, the objective is the partial log-likelihood function of the CoxPH model (See Subsection 3.3.4.2 for the detailed explanation of the CoxPH model.)

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{N}_{\text{uncens}}|}\sum_{n\in\mathcal{N}_{\text{uncens}}}\left(f_\theta(\mathbf{x}_n) - \log\sum_{m\in\mathcal{R}_n}\exp(f_\theta(\mathbf{x}_m))\right) \tag{5.2}$$

where $f_\theta(\mathbf{x})$ is the output of the deep learning model whose parameters θ. However, minimising $\mathcal{L}$ with respect to the θ using standard stochastic gradient descent based

on selecting batches of patients [159] is problematic since:

- Equation 5.2 represents a ranking loss that compares between patients that died in the batch according to their predicted mortality risk. This requires large batch sizes for robust training; however, for high-resolution inputs (3D scans), we are limited by GPU memory to small batch sizes.

- With small batch sizes and a high censoring percentage, there will often be batches containing only censored patients. In this case, the loss cannot be calculated, and these batches will be ignored.

In this chapter, we propose a novel approach to address these limitations and allow for stable training of deep learning models for survival analysis with limited GPU memory. We propose to use memory banks to store the model predictions for later iterations. This allows us to use small batch sizes in alignment with the GPU memory constraints while still having a stable training process. We use the proposed methods to predict the survival of IPF patients using their 3D HRCT scans and clinical data.

5.2 Memory Banks for Improving Cox Proportional Hazards Model

To overcome the limitations of the standard training procedure of the CoxPH model, we propose to use memory banks to store model predictions for later iterations [175, 33]. The memory bank, represented as $\mathcal{MB}$, is a queue of size $\lfloor K \times |\mathcal{N}| \rfloor$ with K representing the fraction of the training dataset stored, $|\mathcal{N}|$ representing the size of the training dataset, and $\lfloor . \rfloor$ representing the floor function. The function of the memory bank is to store the model predictions of training samples, along with their event indicators and death times. In the later iterations, all the samples in the memory bank are used to calculate the CoxPH loss. This allows us to approximate the CoxPH

loss on a larger sample size than the current batch size, which is limited by the GPU memory.

A K value of 1 corresponds to the storage of predictions of the entire training set in the memory bank, while a $K = 0$ means that no samples are stored, equivalent to the standard CoxPH objective. For every training iteration i, we calculate predictions $f_{\theta^i}(x^i)$ for the minibatch m^i and store them in $\mathcal{MB}$, along with the corresponding event indicators δ^i and death times t^i (or censoring times c^i for censored samples). $m^i = \{x_j, \delta_j, t_j\}$ represents the minibatch at iteration i with j representing the sample index $j \in \{1, 2, \ldots, |m^i|\}$. The memory bank $\mathcal{MB}$ is updated as[1]

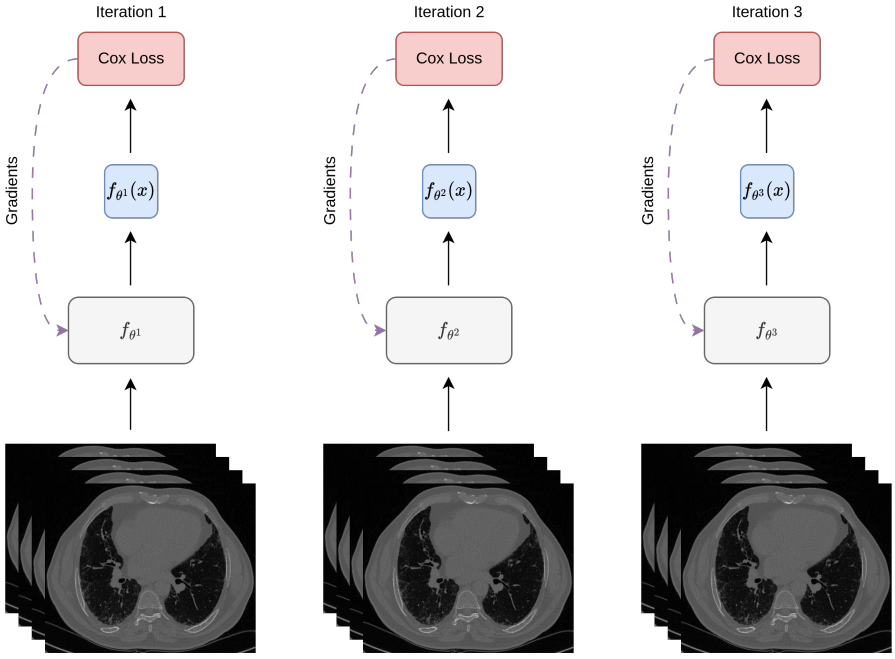
$$\mathcal{MB} \leftarrow \mathcal{MB} \,\|\, \{f_{\theta^i}(x^i), \delta^i, t^i, c^i\} \tag{5.3}$$

where $\|$ denotes concatenation. If the memory bank is full (*i.e.*, $|\mathcal{MB}| = \lfloor K \times |\mathcal{N}| \rfloor$), the oldest samples are removed, and new samples are added. After I iterations, $\mathcal{MB}$ will contain the tuples $\{f_{\theta^i}(x^i), \delta^i, t^i, c^i\}_{i=1}^{I}$. At each iteration i, we calculate the risk set $\mathcal{R}_n^i$ for each uncensored patient n in $\mathcal{MB}$ using the stored event indicators and times. The CoxPH loss for samples in $\mathcal{MB}$ is then calculated using the risk set $\mathcal{R}_n^i$ and the available predictions in the memory bank as

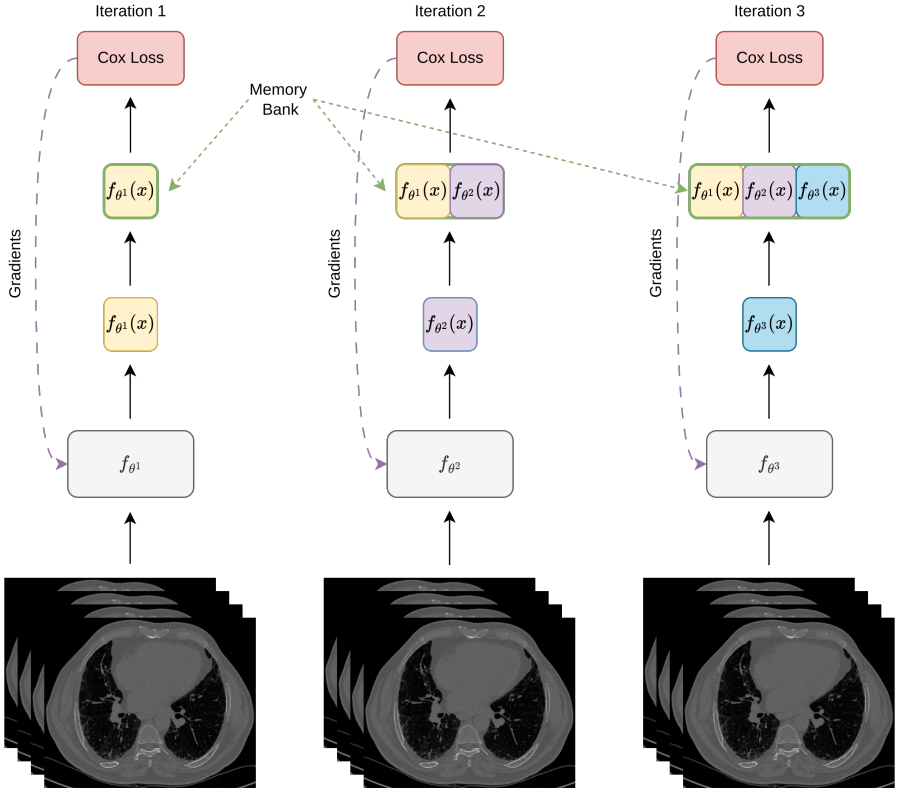
$$\mathcal{L}(\theta^i) \equiv \frac{1}{\mathcal{N}^i_{\text{uncensMB}}} \sum_{n \in \mathcal{N}^i_{\text{uncensMB}}} \left(f_{\theta^{\leq i}}(\mathcal{MB}_n) - \log \sum_{m \in \mathcal{R}_n^i} \exp(f_{\theta^{\leq i}}(\mathcal{MB}_m)) \right) \tag{5.4}$$

where $\mathcal{N}^i_{\text{uncensMB}}$ is the set of uncensored samples in $\mathcal{MB}$ at iteration i, and $f_{\theta^{\leq i}}(\mathcal{MB}_n)$ and $f_{\theta^{\leq i}}(\mathcal{MB}_m)$ are the predictions for patients n and m in $\mathcal{MB}$, respectively, and are functions of the model parameters at iteration i or any previous iteration $< i$. The loss is used to update the current parameters of the model θ^i. By updating $\mathcal{MB}$ at each iteration and using it to calculate the loss, we can effectively

[1]We use the superscript i to denote the batch number and the subscript j to denote the sample index within the batch.



(a) Standard training of CoxPH model. The model is trained on a batch of patients, and the loss is calculated based on the ranking of the patients in the batch.



(b) Proposed CoxMB training. The model predictions are accumulated in a memory bank, and the loss is calculated based on the stored predictions from the current and previous iterations. This allows for computing the loss on a larger sample size than the current batch size.

Figure 5.1: Comparison between the standard CoxPH training and the proposed Cox Proportional Hazards with Memory Banks (CoxMB) training.

Algorithm 1: Pseudocode of CoxMB training in a PyTorch-like style

```
# delta: indicator function (1 if experienced the event and 0 if censored)
# times: event or censoring time
# K: fraction of the training dataset to store in the memory bank
# initialize memory bank with maximum size K * len(data_loader.dataset)
mbank_preds = deque(maxlen=K * len(data_loader.dataset))
mbank_delta = deque(maxlen=K * len(data_loader.dataset))
mbank_times = deque(maxlen=K * len(data_loader.dataset))
for img in loader: # load a minibatch with n samples
 pred = model(img) # get predictions
 mbank_preds.append(pred) # store current predictions in the memory bank
 mbank_delta.append(delta) # store current delta in the memory bank
 mbank_times.append(times) # store current times in the memory bank
 # calculate loss using data in the memory bank
 loss = CoxLoss(mbank_preds, mbank_delta, mbank_times)
 loss.backward() # calculate gradients
 update(model.params) # update model parameters
```

approximate the Cox loss on a sample size larger than allowed by the standard Cox objective, which is limited by the batch size. This allows us to use small batch sizes in alignment with the GPU memory constraints while still having a stable training process. The proposed method is illustrated in Figure 5.1. We refer to this method as CoxMB and compare its performance to the standard Cox objective in our experiments.

In addition, we include a PyTorch-like pseudocode of the CoxMB training in Algorithm 1. The pseudocode shows the training loop of the CoxMB model, where the model predictions are stored in the memory bank, and the loss is calculated using the stored predictions from the current and previous iterations.

5.3 Experiments

We evaluate the proposed CoxMB method and the standard training method of the CoxPH model on the task of predicting survival in IPF patients using their 3D HRCT scans alone and in combination with clinical data.

5.3.1 Data

We use the OSIC dataset described in Section 2.5.3. We select cases with a confirmed diagnosis of IPF and an HRCT with a slice thickness of less than 3.0 mm. The clinical data includes age, sex, smoking history, FVC predicted percent, DL_{CO}, and

antifibrotic treatment; see Section 4.2 for more details. We examine the performance of different methods using exclusively HRCT images or a combination of HRCT images and clinical data. The dataset consists of 728 samples, which we randomly divided into training (70%), validation (15%), and test (15%) sets. The mean and standard deviation of the metrics are reported over five runs with different random splits. Approximately 65% (470 samples) of the dataset are right-censored.

We evaluate the performance in terms of the C-Index and the MAE and RAE of the predicted survival times, see Section 3.3.5.

5.3.2 Preprocessing

5.3.2.1 HRCT Preprocessing

All scans are cropped to the lung area using the lung segmentation model trained by [176]. These scans are then resampled to achieve an isotropic pixel spacing of $1 \times 1 \times 1$ mm^3 via linear interpolation. Following this, the scans are resized to dimensions of $256 \times 256 \times 256$ voxels using bicubic interpolation. Next, we apply histogram matching and a windowing operation within the range [-1024, 150] Hounsfield Units (HU) to remove irrelevant information. Finally, we normalise the scans to have zero mean and unit variance based on the statistics drawn from the training set.

We apply data augmentation techniques during training to mitigate overfitting and improve generalisation. Specifically, random rotation (up to 15 degrees) and translation (up to 20 pixels) are used to introduce variability while preserving anatomical structures. Additionally, scans that fail to meet quality standards due to severe motion artefacts, incomplete lung coverage, or significant noise are excluded from the dataset. Details on the inclusion criteria are provided in Subsection 2.5.3.

5.3.2.2 Clinical Data Preprocessing

For the clinical data, we normalise the continuous features (age, FVC, and DL_{CO}) to have zero mean and unit variance. We encode the categorical features (sex, smoking history, and antifibrotic treatment) using one-hot encoding. We then concatenate the normalised continuous features with the one-hot encoded categorical features to form the clinical data input to the models.

5.3.3 Implementation Details

In our experimental setup, we use a deep learning model to learn the non-linear relationship between the covariates (HRCT and optionally clinical data) and the hazard function; we detail the model architecture in Section 5.3.3.1 and the hyperparameters in Section 5.3.3.2.

5.3.3.1 Model Architecture

To process HRCT scans, we use a 3D CNN, as illustrated in Figure 5.2 (left). The network initiates with a 3D convolutional layer, followed by an instance normalisation layer and a leaky ReLU activation function. We then stack four residual blocks, each comprising three 3D convolutional layers [177]. After each convolutional layer, we use instance normalisation [178] and leaky ReLU [179] layers. We utilised $1 \times 1 \times 1$ kernels for the first and last convolutional layers, while the middle layer used a $3 \times 3 \times 3$ kernel. In a parallel branch, we use a single convolutional layer, and the outputs of the two branches are concatenated. The output of this series of layers is then passed through another convolutional layer, designed with a stride of 2, to halve the spatial dimension. Finally, we use a convolutional layer with 16 filters and a $1 \times 1 \times 1$ kernel to produce a compact feature representation. We flatten this representation and input it into the final fully connected layer. In designing this network, we were aware that the progression of IPF manifests itself in fine pulmonary patterns, such as honeycombing, reticulation, and ground glass opacities. To capture these nuances, we opt for small kernels and deliberately avoid pooling

layers, which could result in the loss of fine image details.

When we incorporate clinical data, we use a MLP that consists of two fully connected layers with 32 neurones each, each followed by batch normalisation [180] and leaky ReLU activation [179], as detailed in Figure 5.2 (right). The MLP output is concatenated with the CNN output. The CNN output, which represents imaging data, is projected to a 32-element vector to balance the contributions from imaging and clinical data. The combined output is subsequently propagated through a final fully connected layer.

5.3.3.2 Hyperparameters

We use AdamW optimiser [181] with a learning rate of 5×10^{-4} and weight decay of 1×10^{-2} for optimisation. Using weight decay is crucial to mitigate overfitting, especially when training deep learning models on small and potentially noisy datasets. The optimal learning rate value was tuned via a random search based on the performance on the validation set. Additionally, we apply a cosine annealing learning rate scheduler and gradient clipping. Due to the high resolution of the imaging data ($256 \times 256 \times 256$), we use a batch size of 2. We train the models for an initial 300 epochs. However, training is halted if there is no improvement in validation performance for 50 consecutive epochs. In CoxMB, we use a K value of 1.0. The models are implemented using PyTorch and trained on a single NVIDIA A6000 GPU.

5.3.4 Results

To evaluate the performance of the proposed CoxMB method, we compare it to the standard CoxPH model on the OSIC dataset. In Table 5.1, we report the test performance of the two approaches on the OSIC dataset. We notice that the introduction of memory banks during training (CoxMB) leads to a significant performance improvement compared to the DeepSurv model, which employs the standard CoxPH objective function [26, 159]. This improvement can be seen through the increase in

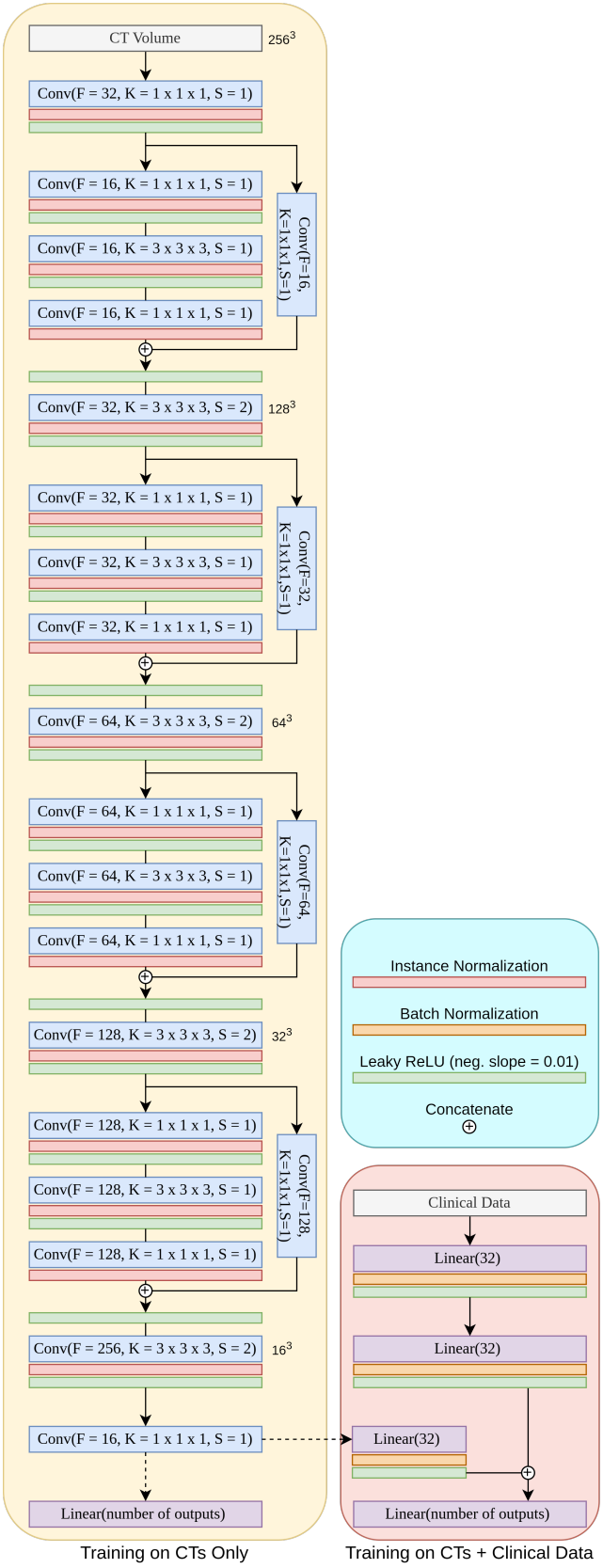


Figure 5.2: Deep learning model architecture. Left: 3D CNN for processing HRCT scans. Right: MLP to process clinical data. F: Number of filters, K: kernel size, S: stride. In the case of using HRCTs only, the architecture on the left is used. In the case of using HRCT and clinical data, the outputs of CNN and MLP are concatenated.

Table 5.1: Comparison of the test performance of CoxPH and CoxMB on OSIC dataset when trained on imaging data only, as well as combined imaging and clinical data. The mean and standard deviation are reported over five runs with different random train/val/test splits. The best results are highlighted in bold.

Data	Method	C-Index ↑	MAE ↓	RAE ↓
Imaging	DeepSurv (Cox)	67.441 ± 4.572	44.898 ± 19.505	2.286 ± 1.414
	CoxMB	**71.067 ± 5.572**	**28.887 ± 2.315**	**1.762 ± 0.807**
Imaging + Clinical	DeepSurv (Cox)	**72.1 ± 2.186**	27.603 ± 3.345	**1.718 ± 0.742**
	CoxMB	68.877 ± 2.413	**24.413 ± 2.548**	1.892 ± 0.868

the C-Index by 3.63, a reduction of the MAE by 16 months, and a decrease in the RAE by 0.046.

Upon inclusion of clinical data, CoxMB upholds superior performance on MAE compared to DeepSurv, whereas DeepSurv excels in ranking performance. This performance divergence, particularly with respect to the decline of the C-Index in the CoxMB case, can likely be attributed to the high noise level and the presence of missing values in clinical data, see Section 4.3 and Section 2.4.4.1. DeepSurv seems to benefit more from including clinical data than CoxMB, where the improvements are marginal. CoxMB already performs reasonably well on the imaging data, and the clinical data do not provide much additional information.

5.3.4.1 Effect of Memory Bank Size

We examine the effect of the size of the memory bank in the CoxMB model, trained on imaging data. K is the fraction of training samples stored in the memory bank during training. We train the CoxMB model with different values of K and report the results in Table 5.2. We observe that the performance of the CoxMB model improves as the memory bank size increases. This is expected, as a larger memory bank allows the model to store more information about the ranking of patients' survival times, which is then used to penalise the model for inaccuracies in predicting the ranking. We anticipate that this depends on the size of the training set and thus requires tuning for each dataset.

5.3.4.2 Performance Under Limited Uncensored Training Data

One notable limitation of the standard CoxPH model is that the loss cannot be calculated if the minibatch contains only censored samples, $\mathcal{N}_{\text{uncens}} = \emptyset$. As a result, these training batches are ignored during training. This is a common issue in survival datasets with high censoring rates and becomes more pronounced when using small batch sizes.

In contrast, the CoxMB training procedure alleviates this issue because the loss is calculated using the samples stored in the memory bank, which are accumulated over multiple iterations. This leads to fewer batches being ignored during training and, subsequently, more efficient use of the training data and more stable training.

We evaluate the performance of the two methods when trained on training sets with varying fractions of uncensored samples. Specifically, we train the models on training sets with $20\%, 40\%, \ldots, 100\%$ of the uncensored samples. The randomly sampled fraction of uncensored cases is added to the censored samples to form the training set in each experiment. It is worth mentioning that cox-based models are untrainable when the training set contains only censored samples; this is a limitation we address in the Chapter 6. In Figure 5.3, we report the performance in terms of the C-Index and MAE when training the models on training sets according to the fractions mentioned above and when using imaging data only. We report the mean and standard deviation over five runs with different random train/val/test splits.

As expected, the performance improves as the fraction of uncensored samples

Table 5.2: Effect of memory bank size on the performance of CoxMB model.

K	C-Index
0.0	67.441 ± 4.572
0.2	67.968 ± 2.712
0.4	70.884 ± 3.844
0.6	70.154 ± 0.975
0.8	$\mathbf{73.294 \pm 4.056}$
1.0	71.067 ± 5.572

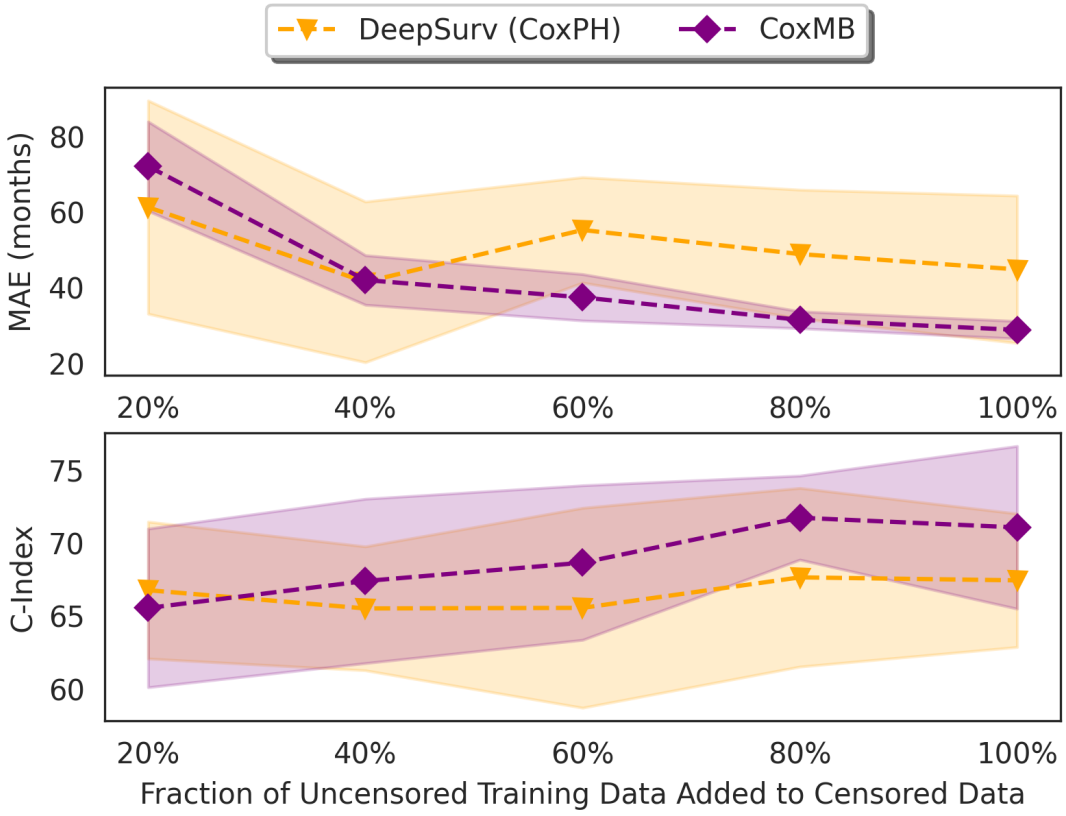


Figure 5.3: Performance of CoxPH and CoxMB models under limited uncensored training data. The models are trained on training sets with varying fractions of uncensored samples. The mean and standard deviation are reported over five runs with different random train/val/test splits.

in the training set increases. Furthermore, we observe that the performance of the CoxMB model, when trained with a limited amount of uncensored data (20%), is comparable to that of the CoxPH model. This can be attributed to the lessened effectiveness of the memory bank when the amount of uncensored data is limited. However, as the amount of uncensored data increases, the memory bank efficacy improves, and the performance of CoxMB consistently surpasses that of the CoxPH model. This is evident in both the C-Index and MAE metrics.

5.4 Related Work

Several studies have used the CoxPH model to predict mortality in IPF patients. Gonzalez *et al.* [10] used the CoxPH model [26] to predict mortality from the Gender

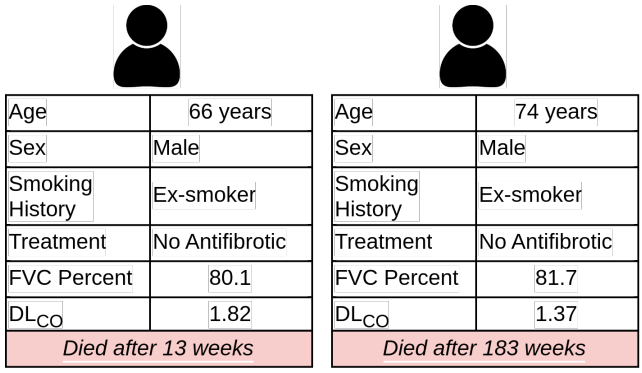
Age Physiology (GAP) index and Composite Physiologic Index (CPI). The GAP index is a clinical scoring system used to predict mortality in patients with IPF. It incorporates gender, age, FVC, and DL_{CO}. The CPI, on the other hand, combines FVC, DL_{CO}, and the FEV1 to estimate the extent of fibrosis and predict disease progression [182, 183, 184].

Collard *et al.* [185] adopted a similar approach and concluded that six-month changes in pulmonary function tests were predictive of mortality risk. However, HRCT scans of the lungs constitute an important part of the clinical assessment of IPF patients and contain pertinent information related to disease progression. It can also be shown that patients with similar clinical information may have different prognoses, see Figure 5.4. Therefore, we investigate the performance of survival models that use both imaging and clinical data.

Other studies have used extracted features from HRCT to predict mortality. Jacob *et al.* [164] compared between mortality prediction using features extracted by an expert radiologist (visual scoring) and features automatically extracted by CALIPER software (Computer-Aided Lung Informatics for Pathology Evaluation and Ratings) [186, 187]. CALIPER quantifies the extent of specified radiological patterns of lung damage[2] seen on the HRCT scan. However, both the visual scoring and CALIPER approaches are unsupervised feature extraction methods in the sense that they are not designed to be maximally predictive of mortality. Visual scoring is also a time-consuming approach that requires clinical expertise and is prone to inter-observer variability.

We are therefore interested in estimating a patient's mortality risk based on their clinical and imaging data. We train an end-to-end neural network to extract imaging features that are maximally predictive of mortality.

[2]Ground glass opacity, reticulation, honeycombing, emphysema, pulmonary vessels volume, and others.



Age	66 years
Sex	Male
Smoking History	Ex-smoker
Treatment	No Antifibrotic
FVC Percent	80.1
DL_{CO}	1.82
Died after 13 weeks	

Age	74 years
Sex	Male
Smoking History	Ex-smoker
Treatment	No Antifibrotic
FVC Percent	81.7
DL_{CO}	1.37
Died after 183 weeks	

Figure 5.4: An example from the OSIC dataset of two patients with very similar clinical features and different survival outcomes. This illustrates the limitations that exist when only using clinical data to predict disease progression in IPF. Our study examined the additional value that might be gained by using imaging data to predict disease progression. Time of death is reported relative to the time of lung function tests.

5.5 Conclusion and Limitations

In this chapter, we proposed a novel approach to address the limitations of the standard training procedure of the CoxPH model for survival analysis. We introduced memory banks to store model predictions for later iterations, allowing for stable training of deep learning models for survival analysis with limited GPU memory. We evaluated the proposed CoxMB method on the task of predicting survival in IPF patients using their 3D CT scans and clinical data. Our results show that the CoxMB method outperforms the standard CoxPH model, achieving a significant improvement in the C-Index, MAE, and RAE when trained on imaging data alone. The CoxMB model offers a more robust training strategy by employing memory banks, which is especially beneficial when training on high-resolution imaging data. The performance of the CoxMB model improves as the memory bank size increases.

However, the proposed method and Cox-based methods generally have some limitations. For example, the proportional hazards assumption in the CoxPH objective is a strong assumption that may not hold in practice. In addition, the CoxPH objective can only be computed if the minibatch contains at least one event, which can be challenging when the censoring rate is high, as is the case in many survival

datasets. Finally, the objective is a ranking objective. It does not optimise for the actual survival times, leading to suboptimal performance in terms of metrics such as MAE and RAE. We address some of these limitations in the next chapter by proposing a novel objective function for survival analysis that directly optimises for the survival times and does not require the proportional hazards assumption.

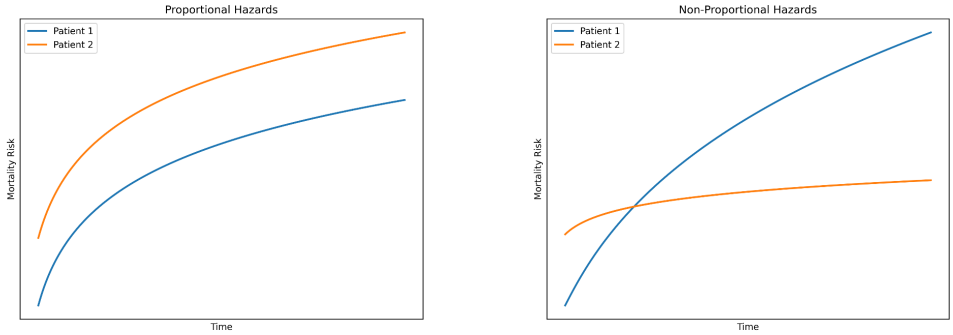
Chapter 6

CenTime: Event-Conditional Modelling of Censoring in Survival Analysis

6.1 Introduction

Despite the wide adoption of the CoxPH model and its variants in survival analysis [26, 159], these models have several limitations. Having discussed memory bank techniques to address the training stability issues in Chapter 5, we now focus on the limitations of the CoxPH model itself. Other methods have been proposed as a remedy to these limitations, such as DeepHit [25], but they have their own drawbacks as well.

In this chapter, we discuss the limitations of the standard methods in survival analysis and introduce a novel event-conditional objective function, CenTime, for training survival models. CenTime leverages censored data more effectively, relaxes restrictive assumptions compared to the CoxPH model, and directly estimates the time-to-event, providing valuable prognostic insights. We evaluate the proposed method on the OSIC dataset and show that it outperforms the state-of-the-art techniques in survival analysis.



(a) Proportional hazards. **(b)** Non-proportional hazards.

Figure 6.1: Proportional hazards assumption. (a) The hazard ratio between two samples is constant over time. (b) Violation of the proportional hazards assumption.

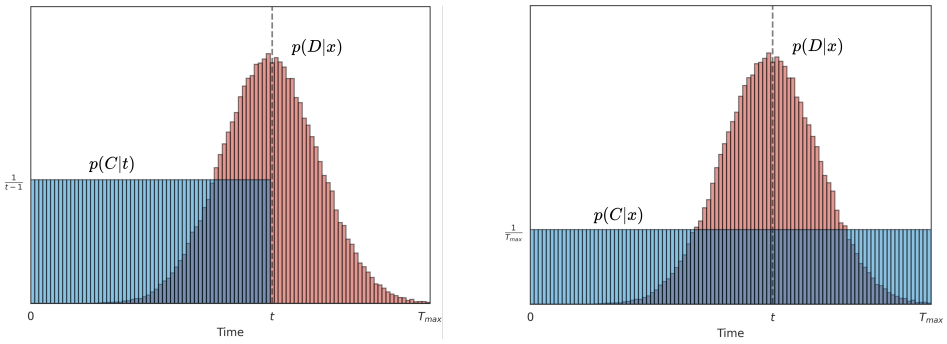
6.2 Limitations of existing survival analysis models

6.2.1 Cox Proportional Hazards Model

The CoxPH model is widely used due to its simplicity and interpretability. However, it has several limitations. Firstly, the proportional hazards assumption (see Figure 6.1), which states that the hazard ratio between two samples is constant over time, is often violated in practice, especially in heterogeneous diseases such as IPF. Secondly, the CoxPH model does not provide a direct estimate of the time-to-event, which is a crucial piece of information for prognosis. Thirdly, as can be seen from the partial log-likelihood function (detailed in Section 3.3.4.2)

$$\mathbb{L}(\beta) = \frac{1}{|\mathcal{N}_{\text{uncens}}|} \sum_{n \in \mathcal{N}_{\text{uncens}}} \left(\beta^T \mathbf{x}_n - \log \sum_{m \in \mathbb{R}_n} \exp(\beta^T \mathbf{x}_m) \right) \tag{6.1}$$

the model does not explicitly model the censored samples. The model only considers the uncensored samples in the likelihood function, which can lead to suboptimal performance, especially in datasets with a high proportion of censored samples. Finally, the CoxPH objective is a ranking objective, thus requiring a large batch size to ensure that the model is trained effectively, which is often computationally expensive and limited by the GPU memory.



(a) The CenTime data generation mechanism. **(b)** The classical data generation mechanism.

Figure 6.2: Survival analysis data generation mechanisms. (a) In the proposed event-conditional censoring model (CenTime), t is drawn from the death time distribution and c is uniformly sampled up to t. (b) In the classical model, t and c represent randomly drawn death and censoring times from the corresponding distributions. If $c < t$, the patient is censored; the observation is the censoring time. Otherwise, the patient is uncensored, and observation is the death time.

6.2.2 DeepHit

DeepHit [25], previously discussed in Section 3.3.4.9, has a few challenges as well. Firstly, the model does not capture the ordinal nature of the time-to-event data. DeepHit uses a softmax function to predict the time-to-event, which treats different death times as separate classes. This can lead to suboptimal performance, especially when the number of classes (T_{max}) is large. Secondly, the model requires a large number of parameters, especially when the maximum time-to-event is large, which can lead to overfitting. Finally, some death times might not be represented in the training data, which could reduce the softmax probabilities to zero, yielding no gradient and impeding the learning process for these times.

6.2.3 Classical Censoring Model

The classical censoring model assumes that censoring times are independent of the event times, see Subsection 3.3.4.8. We propose a novel alternative to this model, CenTime. CenTime is an event-conditional model, and we explain its formulation below.

6.3 CenTime: Event-Conditional Modelling of Censoring

We introduce CenTime, which enables the direct learning of a death time distribution $p_\theta(D = t|x)$ from either censored or uncensored data. Our objective is to maximise the log-likelihood of the data, which includes both censored and uncensored samples

$$\mathbb{L}(\theta) \equiv \sum_{n \in \mathcal{N}_{\text{uncens}}} \log p_\theta(D = t_n|x_n) + \sum_{i \in \mathcal{N}_{\text{cens}}} \log p_\theta(C = c_n|x_n) \tag{6.2}$$

where $p_\theta(C = c|x)$ is the censoring distribution. CenTime uses a novel censoring mechanism that we believe is more representative of censoring in some clinical situations. Here, we concentrate on right censoring while the method is generally applicable to other forms of censoring; see Appendix A. Specifically, we first sample the death time and then generate a censoring time from a distribution up to the death time. This results in the censored time model

$$p_\theta(C = c|x) = \sum_{t=1}^{T_{\max}} p(C = c|D = t, x) p_\theta(D = t|x) \tag{6.3}$$

The objective in Equation 6.2 is the likelihood of a mixture model containing contributions from the uncensored data and censored data, with each term being a consistent objective for the event model parameters θ (*i.e.*, estimators based on either contribution converge to the true parameters as the number of samples increases). This implies that even in the scenario where we only have censored training data, the model can learn the underlying event model.

The model also has the advantage that, if needed, we can easily sample data from this model given the proportion of censored to uncensored data. If a proportion of censored to uncensored data $p_c : p_n$ is required, for a chosen N one can simply sample Np_c censored datapoints from $p_\theta(C = c_n|x_n)$ and Np_n uncensored datapoints

from $p_\theta(D = t_n|x_n)$. This feature is absent in classical censoring models, in which it is not possible to sample data with a required proportion of censored to uncensored data.

We still need to make two assumptions – the censoring distribution $p(C|D,x)$ and the event distribution $p_\theta(D = t|x)$. We define the event distribution $p_\theta(D = t|x)$ below in Section 6.3.1, and here we define the censoring distribution $p(C|D,x)$. In principle, this can also be learned from the data, but for simplicity, we assume a uniform censoring distribution $p(C = c|D = t,x) = const$ for $c < t$ and 0 elsewhere (see Figure 6.2a), giving

$$p_\theta(C = c|x) = \sum_{t=c+1}^{T_{\max}} \frac{1}{t-1} p_\theta(D = t|x) \tag{6.4}$$

For any event distribution model $p_\theta(D = t|x)$ the likelihood objective to maximise is

$$\mathbb{L}(\theta) \equiv \sum_{n \in \mathcal{N}_{\text{uncens}}} \log p_\theta(D = t_n|x_n) + \sum_{i \in \mathcal{N}_{\text{cens}}} \log \sum_{t=c_i+1}^{T_{\max}} \frac{1}{t-1} p_\theta(D = t|x_i) \tag{6.5}$$

6.3.1 Event Time Distribution

We need to make an appropriate choice for the event time distribution $p_\theta(D = t|x)$. We employ a discretised form of the Gaussian distribution

$$p_\theta(D = t|x) = \frac{1}{Z} \exp\left(\frac{-(t - \mu_\theta(x))^2}{2\sigma_\theta^2(x)}\right) \tag{6.6}$$

In this formulation, $\mu_\theta(x)$ and $\sigma_\theta(x)$ are parameters of the distribution that are predicted by the model (a neural network parameterised by θ), and Z is a normalisation factor, defined as

$$Z = \sum_{t=1}^{T_{\max}} \exp\left(\frac{-(t - \mu_\theta(x))^2}{2\sigma_\theta^2(x)}\right) \tag{6.7}$$

This formulation has the following advantages

- The term $(t - \mu_\theta(x))^2$ ensures a heavier penalty for predictions that deviate sig-

nificantly from the true death time, promoting closer predictions. This stands in contrast to approaches that treat death times as independent categories [25], which do not fully capture this relationship.

- The model only outputs two quantities $(\mu_\theta(x), \sigma_\theta(x))$. This keeps the number of parameters low, reducing the risks of overfitting compared to treating this as a $T_{\max}$ classification task, with a category for each timepoint [25].

In principle, the form of the distribution $p_\theta(D = t|x)$ is also learnable, but we found that the discrete Gaussian performed well in our experiments.

6.4 Experiments

6.4.1 Data and Preprocessing

We evaluate the proposed method on the OSIC dataset. We use the same dataset, preprocessing, and splits as in Chapter 5.

6.4.2 Baselines

We compare the proposed method with the following baselines

- **DeepSurv** [159]: the standard CoxPH model with a deep neural network as the base model.
- **CoxMB**: the CoxPH model with the memory bank technique proposed in Chapter 5.
- **DeepHit** [25]: a state-of-the-art survival model that uses a deep neural network to predict the time-to-event, see Section 3.3.4.9.
- **DeepHit ($\mathbb{L}^C_{lik.}$) only**: a variant of DeepHit that uses only the likelihood term, $\mathbb{L}^C_{lik.}$, in the objective function, without the ranking term $\mathbb{L}_{rank.}$. This is to evaluate the performance of DeepHit when the ranking term is removed and relying only on the likelihood term, similar to CenTime.

- **Classical Censoring Model**: an alternative approach to model the censoring distribution, see Section 3.3.4.8. In contrast to DeepHit, we propose to use a discrete Gaussian distribution to model the event time distribution, see Section 6.3.1.

DeepSurv and CoxMB are ranking-based models, while the others are distribution-based models in the sense that they directly model the time-to-event.

6.4.3 Implementation Details

The event distribution-based models parameterise the distribution $p_\theta(t|x)$ using μ_θ and σ_θ. A deep learning model parameterised by θ is used to learn μ_θ, while σ is fixed at 12 months. This helps to stabilise the training process and mitigate overfitting (see [188] for a similar observation). For DeepHit, the model's output is a vector of size $T_{\max}$, representing the logits of the 1-of-$T_{\max}$ classification labels. For the DeepSurv and CoxMB models, the output is a single scalar representing the risk score, as explained in Chapter 5. We use AdamW optimiser [181] with a learning rate of 10^{-4} for the classical and event-conditional censoring models and 5×10^{-4} for DeepHit, DeepSurv, and CoxMB. Unless otherwise stated, we use the same architecture, hyperparameters, and training setup as in Chapter 5.

6.4.4 Results

The evaluation of survival analysis performance depends on the particular clinical objective. For instance, if the aim is to stratify patients into high- and low-risk groups, the C-Index is a suitable metric, whereas if the objective is to precisely predict each patient's time-to-death, metrics such as MAE and RAE are more appropriate. Since CenTime directly predicts the mortality time, MAE and RAE are the most relevant metrics for assessing its performance. However, we also report the C-Index for completeness and to compare CenTime's ranking performance with other methods. See Subsection 3.3.5 for a detailed explanation of survival analysis metrics.

Table 6.1: Comparison of the test performance of the different methods on OSIC dataset when trained on imaging data only, as well as combined imaging and clinical data. The mean and standard deviation are reported over five runs with different random train/val/test splits. The best results are highlighted in bold.

Data	Method	C-Index ↑	MAE ↓	RAE ↓
Imaging	DeepSurv (Cox)	67.441 ± 4.572	44.898 ± 19.505	2.286 ± 1.414
	CoxMB	**71.067 ± 5.572**	28.887 ± 2.315	1.762 ± 0.807
	DeepHit	53.165 ± 8.313	31.074 ± 7.765	1.830 ± 0.522
	DeepHit ($\mathbb{L}^c_{\text{lik.}}$ only)	57.607 ± 4.813	29.862 ± 3.742	1.926 ± 0.869
	Classical Censoring	68.844 ± 5.313	20.448 ± 4.787	1.407 ± 0.853
	CenTime	69.273 ± 0.946	**19.319 ± 1.613**	**1.338 ± 0.665**
Imaging + Clinical	DeepSurv (Cox)	**72.1 ± 2.186**	27.603 ± 3.345	1.718 ± 0.742
	CoxMB	68.877 ± 2.413	24.413 ± 2.548	1.892 ± 0.868
	DeepHit	54.980 ± 3.490	31.246 ± 4.599	2.240 ± 0.862
	DeepHit ($\mathbb{L}^c_{\text{lik.}}$ only)	52.882 ± 3.843	28.718 ± 2.077	2.059 ± 0.722
	Classical Censoring	70.35 ± 2.947	20.476 ± 1.85	1.546 ± 0.611
	CenTime	70.957 ± 3.048	**19.178 ± 0.795**	**1.48 ± 0.671**

In Table 6.1, we report the test performance of the different methods on the OSIC dataset. For distribution-based methods (DeepHit, Classical Censoring, and CenTime), CenTime outperforms all other distribution-based baselines in C-Index, MAE, and RAE metrics, whether trained solely on imaging data or a combination of imaging and clinical data. The superiority of our method is particularly noticeable in the hybrid case, where the MAE decreases by 9.92 and 1.3 months compared to the DeepHit and the classical censoring models, respectively. Similarly, the C-Index improves by 12.22 and 0.61 compared to these models. Compared to DeepSurv and CoxMB, CenTime offers a remarkable improvement in MAE (8.43 and 5.23 months, respectively) and a comparable ranking performance. This demonstrates the effectiveness of CenTime in efficiently capturing the censoring process. Interestingly, CenTime significantly outperforms DeepHit. In addition to the different modelling of the censoring process, this can be attributed to how each model handles the event distribution. CenTime applies a discretised version of the Gaussian distribution (as per Equation 6.6), whereas DeepHit considers it as a classification problem

comprising $T_{\max}$ classes, executed using a fully-connected layer followed by a softmax function. By disregarding the ordinal nature of the time variable and facing the potentially large class number, $T_{\max}$, DeepHit is more susceptible to overfitting.

CenTime outperforms all the baselines in predicting the time of death for IPF patients, whether trained solely on imaging data or a combination of imaging and clinical data. Additionally, it delivers competitive C-Index performance despite not incorporating a ranking objective. This makes it a more appropriate choice for clinical scenarios where the precise prediction of the time of death takes precedence over the ranking of patients' survival times.

Given the potential imbalance in the dataset (*e.g.*, long-term survivors or rapid-progressor patients), we experimented with oversampling techniques to mitigate class imbalance and improve performance on underrepresented subgroups. However, these techniques did not yield significant improvements in predictive performance, likely due to the increased variance and noise introduced by duplicating minority cases. As a result, we did not report these results. This suggests that alternative approaches, such as data augmentation using generative models, may be more effective in handling imbalance in future work.

6.4.4.1 Performance Under Limited Uncensored Training Data

The amount of uncensored data available for training survival models is typically limited. Therefore, learning algorithms must use the available censored data effectively to improve performance. In this subsection, we examine the performance of the different methods when trained on a limited amount of uncensored data in addition to the censored data (imaging only). We randomly sample 0% (purely censored), 20%, 40%, 60%, 80%, and 100% of the uncensored data. In each scenario, all the censored data is added to compose the training set. The results are presented in Figure 6.3.

The initial observation is that Cox-based models (DeepSurv and CoxMB) are

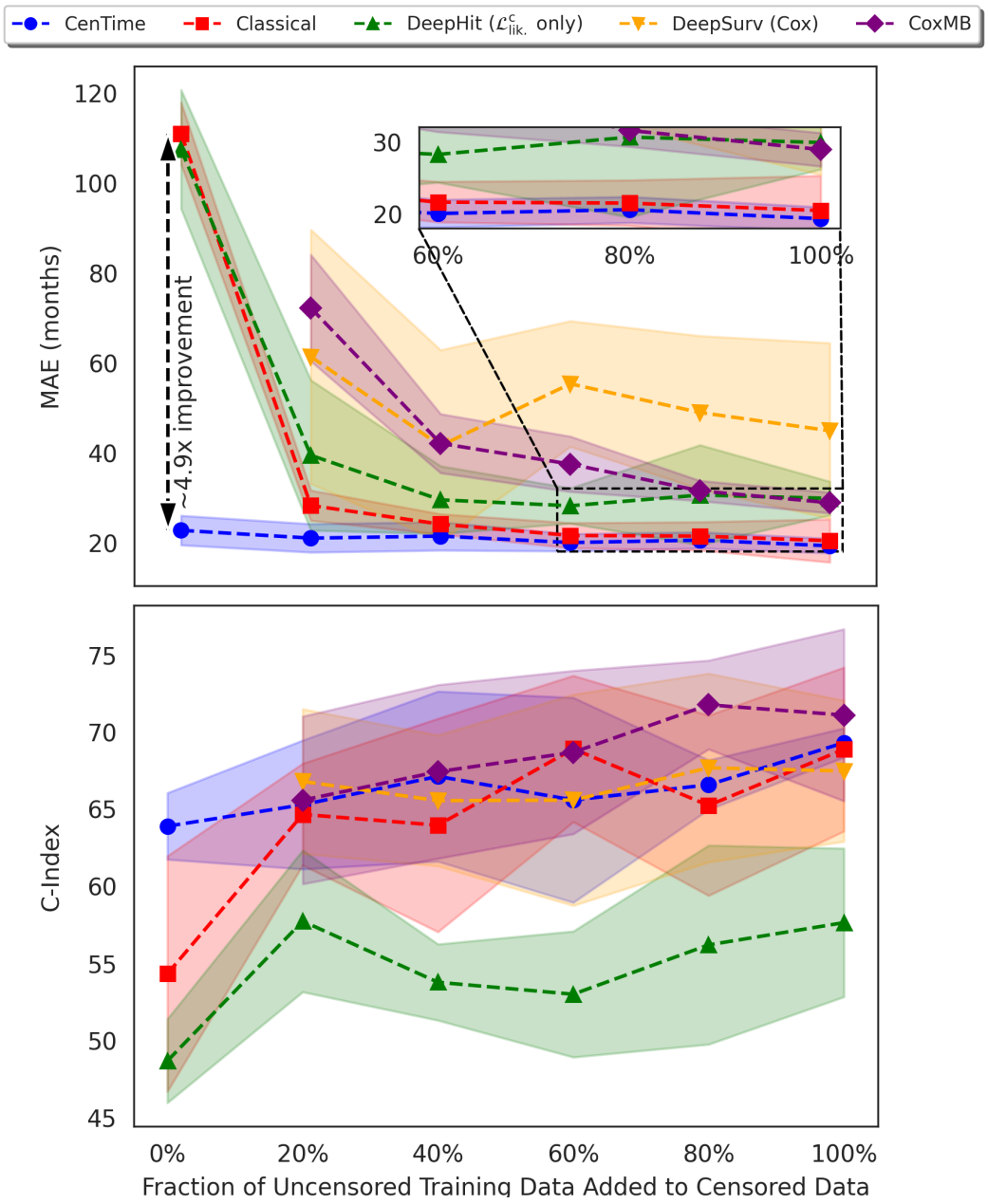


Figure 6.3: Performance of the different methods when trained on gradually increasing percentages of uncensored data added to the censored data. 0% corresponds to training on purely censored data, while 100% corresponds to training on the full training set. The mean and standard deviation are reported over five runs with different random train/val/test splits.

only trainable when uncensored examples are available during training. This is because the objective function is defined solely for uncensored examples (see Equation 5.2). Second, when utilising purely censored data, CenTime shows a significant

improvement ($\approx$ 4.9x in terms of MAE) over the classical and DeepHit models. This is because CenTime forms a consistent estimator of the model parameter θ even with purely censored data, a feature not shared by the classical and DeepHit models. As the amount of uncensored data included in the training data increases, we generally observe an improvement in the performance of all models, and the differences between the various methods diminish. However, CenTime continues to outperform the other methods in terms of MAE and offers competitive performance in terms of the C-Index. These findings underscore the effectiveness of our proposed approach in modelling the censoring process and utilising it efficiently.

Intriguingly, the C-Index performance of CenTime is comparable to that of DeepSurv, even though it does not use a ranking objective. This further underlines the robustness and versatility of our proposed event-conditional censoring model.

6.4.4.2 Effect of Lung Segmentation

Idiopathic Pulmonary Fibrosis predominantly affects the lungs, making this area the most relevant in CT scans. However, some evidence suggests that the disease can also affect other organs, such as the heart [189]. Therefore, we examine the effect of lung segmentation on the performance of CenTime when trained on imaging data. We train the model with and without lung segmentation (using [176]) and report the results in Figure 6.4. We do not observe a significant difference in the performance, suggesting that the model can learn the relevant features from the lung area without explicit segmentation. This also allows the model to benefit from information in the non-lung area (*e.g.*, heart) if it is relevant to the survival prediction task.

6.5 Conclusions

Our work demonstrates the limitations of existing survival methods and addresses them. Traditional Cox-based methods **(i)** assume the strong proportional hazards assumption, which is not always true, **(ii)** estimate the relative hazard rather than

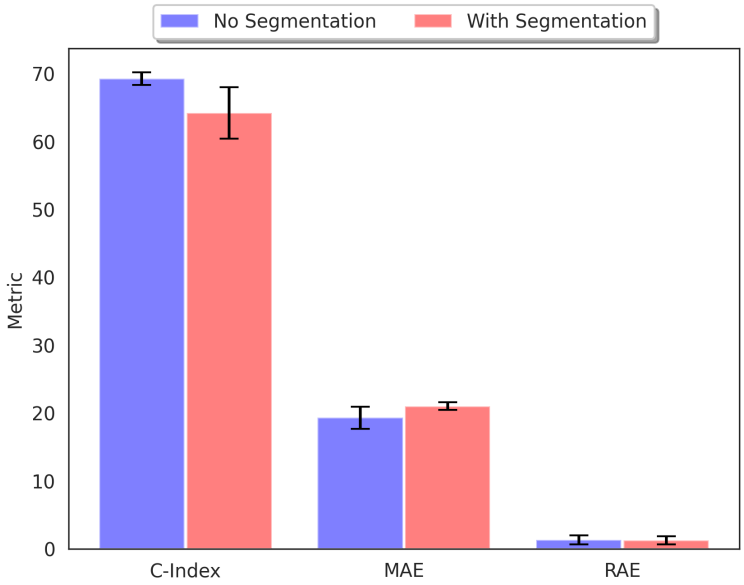


Figure 6.4: Effect of lung segmentation on the performance of CenTime.

the actual death time, which is often more helpful and easier to interpret, and **(iii)** represent a ranking method and, therefore, require a large batch size, which is not always feasible. DeepHit **(iv)** does not encode the ordinal nature of the target survival time variable, **(v)** approaches the problem as a classification task, which becomes prone to overfitting with too many classes. Our CenTime model addresses all these limitations. By modelling the death and censoring likelihoods, it circumvents the hazards proportionality assumption **(i)**, directly estimates the death time **(ii)**, and imposes no batch size restrictions **(iii)**. Furthermore, because of the adoption of the discretised Gaussian distribution, our model naturally encodes the ordinal nature of the target survival time variable **(iv)** and, by outputting only the discretised Gaussian distribution parameters, is less susceptible to overfitting **(v)**. Finally, compared to the classical censoring mechanism, CenTime offers a convenient alternative to the classical censoring model by providing a consistent estimator even with purely censored data alone and should be particularly useful in situations with only minimal uncensored entries.

Our results underscore the effectiveness of CenTime in predicting the time of death while offering competitive ranking performance, even without a ranking

objective. This makes CenTime a compelling choice for clinical scenarios where accurate prediction of the time of death takes precedence over ranking patients' survival times, particularly when dealing with limited observed death time data.

Chapter 7

Conclusions, Limitations and Future Work

In this thesis, we presented a comprehensive framework for modelling disease prognosis in terms of mortality prediction (*i.e.*, survival analysis). Although we focused on IPF as a challenging and heterogeneous disease, the presented methods generally apply to other diseases. In addition, the proposed framework is not limited to a specific modality or type of data and integrates both clinical and imaging data.

7.1 Summary of Contributions

We first addressed the problem of missing values in clinical records. In Chapter 4, we relied on the assumption that there is a relationship or dependency between different features in a patient record. Consequently, we can fit a model to predict the missing values based on the observed ones. Therefore, we proposed a novel method for imputing missing values in clinical records based on a LVMs. We showed that this method outperforms other state-of-the-art methods in terms of the imputation accuracy of both continuous and categorical features. This method was then used to impute missing values in OSIC dataset and used in the subsequent chapters.

In Chapter 5 and Chapter 6, we moved to study the problem of survival analysis, the limitations of the current methods, and how to improve and adapt them to our

problem. Survival analysis methods can be broadly categorized into two main categories: I) Ranking-based models (*e.g.*, CoxPH [26]), and II) Distribution-based models (*e.g.*, DeepHit [25]. In the former category, we aim to train the model to correctly rank patients in a dataset according to their mortality risk without estimating the exact death time. In the latter category, we train the models to output an accurate probability distribution of the time of death for each patient. It is worth mentioning that the two families of models have valid clinical use cases, and the choice between them depends on the clinical question. Therefore, we proposed contributions in the two categories.

In Chapter 5, we highlight the limitations of the CoxPH model, the most widely used model in survival analysis. The linearity assumption can be easily alleviated using non-linear transformations of the features (*e.g.*, a deep neural network). However, the ranking nature of the objective requires a large batch size to be trained effectively, which is not always feasible due to memory constraints, especially when using high-resolution imaging data. To address this issue, we proposed a novel method for training the CoxPH model using memory banks to accumulate model predictions over the training set. This allows us to compute the CoxPH loss over a much larger set of samples. We showed that this method outperforms the standard CoxPH model in terms of concordance index while being more memory efficient. This allows the application of the CoxPH model to high-resolution imaging data while maintaining high performance and a stable training process, in contrast to the standard CoxPH model, which does not scale well to high-resolution imaging data due to its memory requirements.

In Chapter 6, we shifted our focus to the second category of survival analysis, distribution-based models. Ranking-based models have limitations, such as the inability to estimate the exact time of death and the strong assumption of proportional hazards. Distribution-based models can overcome these limitations by estimating the full distribution of the time of death for each patient. We proposed a novel

objective function for training survival analysis models. CenTime, our proposed method, is a maximum likelihood-based method that proposes an alternative data generation mechanism for the censoring process. We showed that CenTime has several advantages over the state-of-the-art methods, such as CoxPH [26, 159], DeepHit, and the classical censoring model [25]. It performs better in accurately predicting the time of death while maintaining comparable ranking performance to the ranking-based models. In addition, CenTime excels in the presence of a limited amount of uncensored data due to its ability to model the censoring process more effectively.

We believe that these methods address practical problems in healthcare and medical imaging (*e.g.*, missing data, limited memory resources, and the abundance of censored data) and can be used to improve the prognosis of patients with IPF and other diseases. However, we discuss some limitations and future work in the following section.

7.2 Limitations and Future Work

7.2.1 Imputation of Missing Data

The proposed method for imputing missing values in Chapter 4 relies on the observed features to predict the missing ones in a patient record. While this method outperforms other imputation methods, it does not consider other sources of information that might give more information about the missing values. For example, the HRCT images can be used with the observed clinical features to predict the missing values. In addition, the proposed method does not consider the data's temporal nature and the features' previous values. Further, the EM algorithm used to train the LVM model is sensitive to the initialization of the parameters, does not scale well to high dimensional data, and is prone to local minima. Future work can address these limitations by incorporating the HRCT images, the temporal nature of the

data and the previous records of the patients (*e.g.*, using Long Short-Term Memory (LSTM) models [190], or Transformers [191]), and using more robust optimization methods [116].

Additionally, a more rigorous evaluation of the imputation method is necessary to assess its robustness to noise and missing data patterns. This is particularly critical in clinical settings, where data quality is often compromised due to human errors, inconsistent reporting, or variations in measurement protocols. Evaluating the method's stability under these real-world conditions will ensure its reliability in practical applications.

7.2.2 Cox Proportional Hazards with Memory Banks

One limitation with the proposed method in Chapter 5 is that after some training iterations $i >>> 1$, some information from the early iterations will be irrelevant to the current model parameters θ_i and might hurt the performance. There are several ways to address this issue; one notable way is to use a momentum factor, which gives more weight to the recent updates of the model parameters. Specifically, we could have two versions of the model parameters θ_i and ϕ_i, where θ_i is the current model parameters, and ϕ_i is an exponentially moving average of the previous model parameters

$$\phi_i = \beta\phi_{i-1} + (1-\beta)\theta_{i-1} \tag{7.1}$$

where β is a momentum factor $0 < \beta < 1$, where $\beta = 0$ corresponds to the standard CoxPH model, and $\beta = 1$ corresponds to the proposed method in Chapter 5, the higher the value of β, the more weight is given to the recent updates of the model parameters. θ_i is normally updated using the gradients of the loss function with respect to the model parameters, while no gradients are used to update ϕ_i. The model parameters used to compute the loss function are then ϕ_i instead of θ_i. This allows the model to have a more stable training process and avoid irrelevant information in the memory bank. This is similar to the momentum idea used in the MoCo method

for training contrastive learning models [33].

7.2.3 CenTime

In Chapter 6, we assumed that the censoring distribution $p(c|x)$ follows a uniform distribution from 0 to the observed time of death t. However, this could be a strong assumption, and future work should explore more flexible distributions for the censoring process. Additionally, we assumed that the censoring process is independent of the covariates x, which might not hold in some cases. Relaxing this assumption by incorporating a censoring model that explicitly depends on covariates could improve the robustness of the model.

For the death distribution $p(t|x)$, we modelled it as a discrete Gaussian distribution. Future work could investigate other distributions, such as the Weibull distribution, to assess their suitability for capturing survival times more effectively.

While CenTime demonstrates strong performance across multiple evaluation metrics, dataset imbalance remains challenging, particularly for long-term survivors and rapid-progressing patients. As discussed in Section 6.4.4, we experimented with oversampling techniques to mitigate this issue but did not observe significant improvements in predictive performance. Future work should explore alternative strategies, such as focal loss, reweighting, or generative data augmentation, to improve model robustness for minority subgroups.

7.2.4 Selection Bias and Generalizability

While the OSIC dataset is sourced from six sites worldwide, potential selection biases may still affect the generalizability of the models developed. Despite its global nature, the dataset primarily consists of data from specialized centres, which may not fully represent the broader IPF population. Patients from underrepresented geographic regions, ethnic groups, or community-based hospitals may be missing or underrepresented, potentially limiting the model's ability to generalize across diverse clinical settings.

To address this, future work should evaluate model performance on external datasets from more diverse clinical environments. Furthermore, stratified analysis of model predictions across demographic subgroups could provide insights into any disparities in performance and help identify potential biases.

7.2.5 Clinical Interpretability

One limitation of the proposed methods is that they need to be interpretable to clinicians. While the proposed methods achieve state-of-the-art performance in terms of prediction accuracy, they do not provide insights into the underlying mechanisms of the disease. Interpretability is crucial for clinical acceptance, especially in a disease like IPF, where the underlying mechanisms are not well understood. We explored the use of GradCAM [166] to generate saliency maps to provide insights into the model predictions and showed that the model highlights areas of fibrosis in the HRCT images, see Figure 7.1. However, the model also highlights other areas, such as bones, whose relevance to the prediction is unclear. In addition, GradCAM generates local explanations but does not provide a global view of the model predictions. Future work can explore other methods for model interpretability, such as SHAP [192], LIME [193], DeepLIFT [194], and counterfactual explanations [195]. These methods can provide more insights into the model predictions and help clinicians understand the underlying mechanisms of the disease.

7.2.6 Clinical Implementation Considerations

While the proposed methods demonstrate strong predictive capabilities, their effective integration into clinical workflows requires careful consideration. Extensive validation studies are essential to assess model generalizability across diverse patient populations and clinical settings, ideally involving external validation on multi-centre datasets.

The presented models can be integrated into clinical workflows to assist healthcare providers in patient prognosis and treatment planning. For example, the models

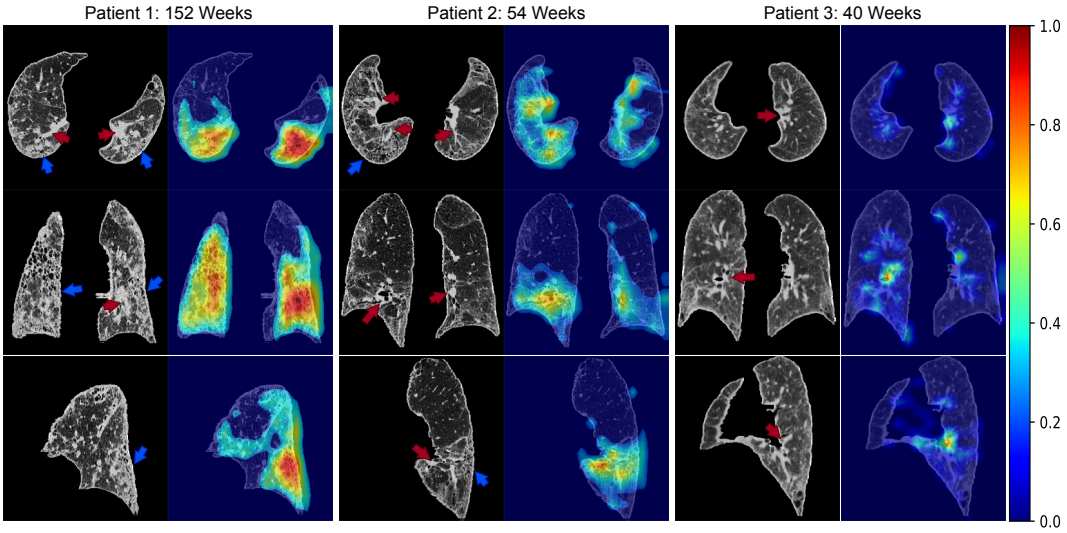


Figure 7.1: Saliency maps for the CoxMB model using the Grad-CAM method with the reported time of death. The model highlights areas of fibrosis (blue arrows) but also pulmonary vessels (red arrows).

can identify high-risk patients who may benefit from early interventions or more aggressive treatment strategies. Additionally, leveraging HRCT imaging in a screening setting could help flag patients with early-stage disease before respiratory symptoms manifest, potentially informing timely clinical decisions. Finally, equipped with interpretability tools (see Subsection 7.2.5), clinicians can use the models to gain insights into the underlying disease mechanisms, especially in diseases like IPF where we have limited understanding of disease progression and response to treatment.

Despite their promise, these models face several barriers to adoption. First, the accessibility of HRCT imaging may be limited, particularly in resource-constrained healthcare settings [196]. Second, the computational demands for model inference and deployment—such as GPU and memory requirements—may limit feasibility in standard clinical environments. Future work should explore resource-efficient architectures and hardware optimization strategies to facilitate real-world adoption.

Beyond technical challenges, the current "grading" of IPF remains insufficiently defined, complicating patient stratification. Despite the existence of clinical guidelines for diagnosing and managing IPF [6, 76], refining disease staging criteria could

improve IPF data labelling and consequently enhance prognostic models performance. Future clinical research should prioritize precise phenotyping and robust labelling of IPF progression, ensuring that prognostic models can deliver more reliable and clinically relevant predictions.

7.2.7 Vision Language Models

The lack of training data is a common problem in medical imaging, especially in diseases like IPF, where the number of patients is limited. Pretrained models can alleviate this issue by learning from large datasets. Large Language Models (LLMs) have shown outstanding performance in downstream tasks in natural language processing after being trained on large corpora of text data [197, 198, 199].

In addition, not only in the case of IPF, but in general, clinical diagnosis, treatment, and prognosis are often based on the interpretation of medical images and reports. Machine learning models that can understand and generate text and interpret medical images have the potential to assist clinicians in their decision-making and improve patient outcomes. Vision Language Models (VLMs) have shown impressive performance on various tasks and benchmarks by jointly learning from the visual and textual modalities [200, 201, 202, 203, 204]. The general framework for these models is to finetune a pretrained LLM on a specific task and dataset. Thanks to the large number of parameters in these models and the large amounts of data used for pretraining, they have shown excellent capabilities in generating high-quality text and generalization to new tasks and datasets [200, 201, 202].

7.2.7.1 Limitations of the Next-Token Prediction Objective

LLMs/VLMs are trained autoregressively using a Next-Token Prediction (NTP) objective function, where the next token in a sequence is classified into one of the tokens in the vocabulary based on the preceding tokens.

$$\max_{\theta} \sum_{t=1}^{T} \log p_{\theta}(x_t | x_{<t}) \tag{7.2}$$

where x_t is the token at time t, $x_{<t}$ is the sequence of tokens before time t, and θ are the model parameters. The distribution $p_\theta(x_t|x_{<t})$ is usually parameterized by a neural network, such as a Transformer [191]. The model outputs a softmax distribution over the vocabulary

$$p_\theta(x_t|x_{<t}) = \frac{\exp(f_\theta(x_t, x_{<t}))}{\sum_{x\in\mathcal{V}} \exp(f_\theta(x, x_{<t}))} \tag{7.3}$$

where $f_\theta(x_t, x_{<t})$ is the output of the neural network for token x_t given the sequence of tokens $x_{<t}$, and $\mathcal{V}$ is the vocabulary. The model is trained to maximize the log-likelihood of the observed tokens in the training set, as shown in Equation 7.2.

While this approach has shown an impressive performance in language generation [197, 198, 199], it has limitations when predicting numerical quantities of high importance in the medical domain (*e.g.*, age or clinical measurements). First, as a classification objective, NTP does not encode the ordinal nature of these variables. Second, models trained using NTP cannot generalize to numbers not in the training set.

7.2.7.2 Possible Remedies

A natural direction for future work is to extend the VLMs to accurately predict numerical quantities using a regression objective function instead of a classification one. A possible approach is to augment the NTP objective function with a regression objective to predict numerical quantities. We could add a regression head to the model, which takes the transformer's output and predicts the exact value of the numerical quantity. The model is then trained to minimize the mean squared error (or another regression loss) between the predicted value and the ground truth.

More formally, the model will have two heads, one for the standard classification task $c_\theta(x_t|x_{<t})$ and one for the regression task $r_\theta(x_t|x_{<t})$, the model final output will

depend on the type of the token x_t as follows

$$p_\theta(x_t|x_{<t}) = \begin{cases} \frac{\exp(f_\theta(x_t, x_{<t}))}{\sum_{x \in \mathcal{V}} \exp(f_\theta(x, x_{<t}))} & \text{if } x_t \text{ is a token} \\ r_\theta(x_t|x_{<t}) & \text{if } x_t \text{ is a numerical quantity} \end{cases} \tag{7.4}$$

The model is then trained to minimize the combined loss function

$$\max_\theta \sum_{t=1}^{T} (\mathbb{I}(x_t \text{ is a token}) \log p_\theta(x_t|x_{<t}) \tag{7.5}$$

$$-\mathbb{I}(x_t \text{ is a numerical quantity}) \mathbb{L}_{\text{MSE}}(r_\theta(x_t|x_{<t}), x_t)) \tag{7.6}$$

where $\mathbb{I}$ is the indicator function, and $\mathbb{L}_{\text{MSE}}$ is the mean squared error loss.

The question of how to identify numerical quantities in the text is an interesting research question. One possible approach is to use a named entity recognition model to identify numerical quantities in the text and then use the regression head to predict the exact value of the quantity [205]. Another more straightforward approach is to use two special tokens added to the vocabulary, one for the start of the numerical quantity and one for the end of the numerical quantity. The model is then trained to predict the start and end tokens, and the regression head is used to predict the exact value of the numerical quantity between the two tokens. A rough sketch of the latter method is shown in Figure 7.2.

7.2.7.3 Application to IPF

The suggested method can be used to predict the time of death in survival analysis and other numerical quantities, such as the FVC and the DL_{CO} in IPF, while leveraging the pretrained VLMs on large corpora of text and image data. The model can be finetuned on the OSIC dataset to predict the time of death, the FVC, and the DL_{CO} of patients with IPF. The model can then generate reports for patients with IPF and assist clinicians in decision-making. Further, the model can be used to generate

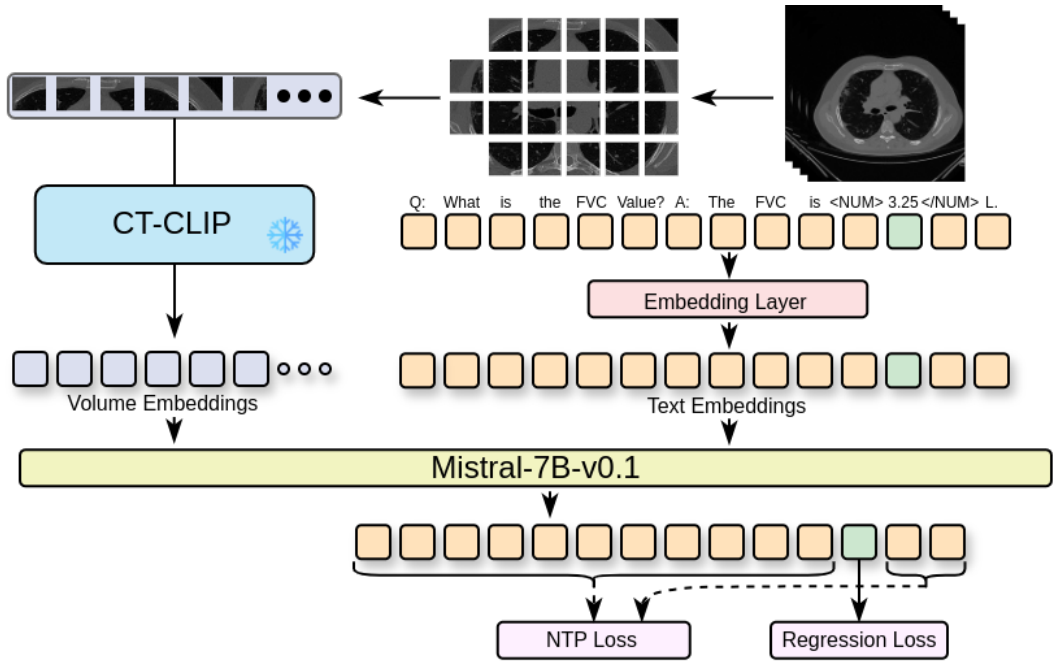


Figure 7.2: A rough sketch of the suggested method for extending Vision Language Models to predict numerical quantities. The model is trained to predict the next token in a sequence based on the preceding tokens. If the next token is a numerical quantity, the model is trained to predict the exact quantity value using a regression objective function. Otherwise, the model is trained to predict the next token in the vocabulary using the standard NTP objective function.

explanations for the model predictions, provide insights into the underlying mechanisms of the disease, and answer clinical questions (*i.e.*, medical visual question answering).

7.3 Outlook

This thesis advanced the field of survival analysis by addressing critical challenges related to missing data, computational constraints, and the complexities of censored data. Our contributions—ranging from innovative imputation techniques to novel survival models—highlight the potential of leveraging deep learning and probabilistic modelling to enhance patient prognosis. As we look to the future, several promising directions could build on this foundation.

7.3.1 Impact on Treatment and Drug Discovery

While this thesis primarily focuses on prognosis, its findings have potential implications for treatment planning and drug discovery. Future research should leverage the prognostic modelling methods presented in the thesis in combination with machine learning interpretability techniques (see Subsection 7.2.5) to identify novel IPF biomarkers. These biomarkers could provide deeper insights into disease mechanisms, facilitating the development of more targeted treatment strategies.

Furthermore, integrating predictive modelling with biomarker discovery could aid in designing personalized treatment plans and optimizing patient stratification for clinical trials. Identifying high-risk patients earlier may enable timely interventions, while biomarker-driven stratification could improve the efficiency of drug trials by selecting patients more likely to respond to specific therapies. These advancements would be central to improving IPF management and accelerating the development of novel therapeutics.

7.3.2 Broader Integration of Multimodal Data

Future research should explore the collection and integration of additional data modalities to improve the accuracy and robustness of survival models. For example, genetic, proteomic, and other omics data could provide valuable insights into the underlying mechanisms of diseases like IPF. In addition, integrating electronic health records, patient-reported outcomes, and other clinical data could enhance the predictive power of survival models. Researchers can develop more comprehensive and personalized prognostic models by combining diverse data sources.

7.3.3 Expanding to Rare Diseases

The methods developed in this thesis could be applied to a wide range of rare diseases, where limited data and high variability present significant challenges for prognosis. By adapting and extending the proposed techniques, researchers can develop tailored survival models for rare diseases, improving patient outcomes and advancing our

understanding of these conditions.

7.3.4 Leveraging Advances in Foundation Models

The recent advances in foundation models, such as GPT-3 [206], CLIP [207], and DALL-E [208], offer exciting opportunities for survival analysis. By leveraging these models, researchers can develop more powerful and flexible survival models that can learn from large-scale text and image data. These foundation models could be finetuned on medical datasets to improve the accuracy and generalization of survival models. An example is the VLMs discussed in the previous section.

Appendix A

CenTime for Interval Censoring

In the main text, we focused on right-censoring, which is the most common form of censoring in survival analysis. Nevertheless, the versatility of CenTime enables its application to interval censoring as well. In this section, we delineate how CenTime can be naturally adapted to handle interval censoring.

A.1 Interval Censoring

Interval censoring arises when the event is known to have occurred *within* a specific time interval $\{c_1, \ldots, c_2\}$. For instance, a patient is reported to be alive at time c_1 and subsequently reported dead at time c_2. Although the exact time of death is unknown, we know that it occurred within $\{c_1, \ldots, c_2\}$. According to our conditional censoring model, we will first sample a death time t from the distribution $p_\theta(t|x)$, then sample a lower censoring time c_1 from a distribution whose support is $\{1, \ldots, t-1\}$ and an upper censoring time c_2 from a distribution whose support is $\{t+1, \ldots, T_{\max}\}$. Similar to the right-censoring case, we assume a uniform censoring distribution for the states $c < t$ and $c > t$ for the two censoring distributions, respectively. The likelihood for an interval-censored observation is then

$$p(C_1 = c_1, C_2 = c_2|x) = \sum_{t=c_1+1}^{c_2-1} \frac{1}{t\,(T_{\max} - t)} p_\theta(D = t|x) \tag{A.1}$$

The objective function is then

$$\mathbb{L}(\theta) = \sum_{i \in \mathcal{N}_{\text{uncens}}} \log p_\theta(D = t_i | x_i) \tag{A.2}$$

$$+ \sum_{i \in \mathcal{N}_{\text{interval-cens}}} \log \sum_{t=c_1+1}^{c_2-1} \frac{1}{t\,(T_{\max} - t)} p_\theta(D = t | x) \tag{A.3}$$

where $\mathcal{N}_{\text{interval-cens}}$ is the set of interval-censored observations in the dataset.

Bibliography

[1] Ganesh Raghu, Martine Remy-Jardin, Luca Richeldi, Carey C Thomson, Yoshikazu Inoue, Takeshi Johkoh, Michael Kreuter, David A Lynch, Toby M Maher, Fernando J Martinez, et al. Idiopathic Pulmonary Fibrosis (an Update) and Progressive Pulmonary Fibrosis in Adults: an Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *American Journal of Respiratory and Critical Care Medicine*, 205(9):e18–e47, 2022. 24

[2] David A Zisman, Michael P Keane, John A Belperio, Robert M Strieter, and Joseph P Lynch. Pulmonary Fibrosis. *Fibrosis Research: Methods and Protocols*, pages 3–44, 2005. 24

[3] Martin Kolb and Martina Vašáková. The Natural History of Progressive Fibrosing Interstitial Lung Diseases. *Respiratory Research*, 20(1):57, 2019. 24

[4] David J. Lederer and Fernando J. Martinez. Idiopathic Pulmonary Fibrosis. *New England Journal of Medicine*, 378(19):1811–1823, 2018. 24, 38, 41, 42, 47

[5] Shaney L Barratt, Andrew Creamer, Conal Hayton, and Nazia Chaudhuri. Idiopathic Pulmonary Fibrosis (IPF): an Overview. *Journal of Clinical Medicine*, 7(8):201, 2018. 24, 38

[6] Ganesh Raghu, Harold R Collard, Jim J Egan, Fernando J Martinez, Juergen

Behr, Kevin K Brown, Thomas V Colby, Jean-François Cordier, Kevin R Flaherty, Joseph A Lasky, et al. An Official ATS/ERS/JRS/ALAT Statement: Idiopathic Pulmonary Fibrosis: Evidence-Based Guidelines for Diagnosis and Management. *American Journal of Respiratory and Critical Care Medicine*, 183(6):788–824, 2011. 24, 38, 40, 41, 46, 47, 50, 73, 117

[7] Helen Strongman, Imran Kausar, and Toby M Maher. Incidence, Prevalence, and Survival of Patients with Idiopathic Pulmonary Fibrosis in the UK. *Advances in Therapy*, 35:724–736, 2018. 24

[8] Roy Pleasants and Robert M Tighe. Management of Idiopathic Pulmonary Fibrosis. *Annals of Pharmacotherapy*, 53(12):1238–1248, 2019. 24

[9] Brett Ley, Williamson Z Bradford, Eric Vittinghoff, Derek Weycker, Roland M du Bois, and Harold R Collard. Predictors of Mortality Poorly Predict Common Measures of Disease Progression in Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 194, 2016. 25

[10] Amy Taylor Gonzalez and Toby Maher. Predicting Mortality in Idiopathic Pulmonary Fibrosis. Which Parameters Should Be Used to Determine Eligibility for Treatment? Analysis of a UK Prospective Cohort. *European Respiratory Journal*, 48(suppl 60), 2016. 25, 27, 94

[11] Harold R Collard, Talmadge E King Jr, Becki Bucher Bartelson, Jason S Vourlekis, Marvin I Schwarz, and Kevin K Brown. Changes in Clinical and Physiologic Variables Predict Survival in Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 168(5):538–542, 2003. 25, 27

[12] Ivan O Rosas, Ping Ren, Nilo A Avila, Catherine K Chow, Teri J Franks, William D Travis, J Philip McCoy Jr, Rose M May, Hai-Ping Wu, Dao M

Nguyen, et al. Early Interstitial Lung Disease in Familial Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 176(7):698–705, 2007. 25

[13] Brett M Elicker, Kimberly G Kallianos, and Travis S Henry. The Role of High-Resolution Computed Tomography in the Follow-Up of Diffuse Lung Disease. *European Respiratory Review*, 26(144), 2017. 25

[14] Christopher M Bishop and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006. 26, 54

[15] Nicu Sebe. *Machine Learning in Computer Vision*, volume 29. Springer Science & Business Media, 2005. 26

[16] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624, 2020. 26

[17] Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. Secure and Robust Machine Learning for Healthcare: A Survey. *IEEE Reviews in Biomedical Engineering*, 14:156–180, 2020. 26

[18] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018. 26, 70

[19] Barbara Lobato-Delgado, Blanca Priego-Torres, and Daniel Sanchez-Morillo. Combining Molecular, Imaging, and Clinical Data Analysis for Predicting Cancer Prognosis. *Cancers*, 14(13), 2022. 26

[20] Mohammad Moshawrab, Mehdi Adda, Abdenour Bouzouane, Hussein Ibrahim, and Ali Raad. Reviewing Multimodal Machine Learning and Its Use in Cardiovascular Diseases Detection. *Electronics*, 12(7), 2023. 26

[21] Saeed Amal, Lida Safarnejad, Jesutofunmi A. Omiye, Ilies Ghanzouri, John Hanson Cabot, and Elsie Gyang Ross. Use of Multi-Modal Data and Machine Learning to Improve Cardiovascular Disease Care. *Frontiers in Cardiovascular Medicine*, 9, 2022. 26

[22] Sachin Kumar and Shivani Sharma. An Improved Deep Learning Framework for Multimodal Medical Data Analysis. *Big Data and Cognitive Computing*, 8(10), 2024. 26

[23] Sachin Kumar, Olga Ivanova, Artyom Melyokhin, and Prayag Tiwari. Deep-Learning-Enabled Multimodal Data Fusion for Lung Disease Classification. *Informatics in Medicine Unlocked*, 42:101367, 2023. 26

[24] Frank Emmert-Streib and Matthias Dehmer. Introduction to Survival Analysis in Practice. *Machine Learning and Knowledge Extraction*, 1(3):1013–1038, 2019. 26

[25] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A Deep Learning Approach to Survival Analysis with Competing Risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 27, 31, 65, 66, 98, 100, 103, 112, 113

[26] David R Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. 27, 28, 31, 62, 90, 94, 98, 112, 113

[27] An Zhao, Ahmed H Shahin, Yukun Zhou, Eyjolfur Gudmundsson, Adam Szmul, Nesrin Mogulkoc, Frouke Van Beek, Christopher J Brereton, Hendrik W Van Es, Katarina Pontoppidan, et al. Prognostic Imaging Biomarker Discovery in Survival Analysis for Idiopathic Pulmonary Fibrosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 223–233. Springer, 2022. 27, 33

[28] Parviz Shahmirzalou, Majid Jafari Khaledi, Maryam Khayamzadeh, and Aliakbar Rasekhi. Survival Analysis of Recurrent Breast Cancer Patients Using Mix Bayesian Network. *Heliyon*, 9(10), 2023. 27

[29] Frederic Richardeau and Thi Thuy Linh Pham. Reliability Calculation of Multilevel Converters: Theory and Applications. *IEEE Transactions on Industrial Electronics*, 60(10):4225–4233, 2012. 27

[30] Ganesh Raghu, Martine Remy-Jardin, Jeffrey L Myers, Luca Richeldi, Christopher J Ryerson, David J Lederer, Juergen Behr, Vincent Cottin, Sonye K Danoff, Ferran Morell, et al. Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *American Journal of Respiratory and Critical Care Medicine*, 198(5):e44–e68, 2018. 27, 38, 41, 42, 45, 63

[31] Constanza L. Andaur Navarro, J. Damen, T. Takada, Steven W. J. Nijman, P. Dhiman, Jie Ma, G. Collins, R. Bajpai, R. Riley, K. Moons, and L. Hooft. Risk of Bias in Studies on Prediction Models Developed Using Supervised Machine Learning Techniques: Systematic Review. *The BMJ*, 375, 2021. 27

[32] Zhenkun Shi, Sen Wang, Lin Yue, Lixin Pang, Xianglin Zuo, Wanli Zuo, and Xue Li. Deep Dynamic Imputation of Clinical Time Series for Mortality Prediction. *Inf. Sci.*, 579:607–622, 2021. 27

[33] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 30, 84, 115

[34] Ahmed Shahin, Joseph Jacob, Daniel Alexander, and David Barber. Survival Analysis for Idiopathic Pulmonary Fibrosis Using CT Images and Incomplete

Clinical Data. In *International Conference on Medical Imaging with Deep Learning*, pages 1057–1074. PMLR, 2022. 33

[35] Ahmed H Shahin, An Zhao, Alexander C Whitehead, Daniel C Alexander, Joseph Jacob, and David Barber. Centime: Event-Conditional Modelling of Censoring in Survival Analysis. *Medical Image Analysis*, 91:103016, 2024. 33

[36] Yaozhi Lu, Shahab Aslani, An Zhao, Ahmed Shahin, David Barber, Mark Emberton, Daniel C Alexander, and Joseph Jacob. A Hybrid CNN-RNN Approach for Survival Analysis in a Lung Cancer Screening Study. *Heliyon*, 9(8), 2023. 33

[37] Ahmed H Shahin, Yan Zhuang, and Noha El-Zehiry. From Sparse to Precise: A Practical Editing Approach for Intracardiac Echocardiography Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 766–775. Springer, 2023. 33

[38] Alexander C Whitehead, Ahmed H Shahin, An Zhao, Daniel C Alexander, Joseph Jacob, and David Barber. Neural Network Based Methods for the Survival Analysis of Idiopathic Pulmonary Fibrosis Patients From a Baseline CT Acquisition. In *2023 IEEE Nuclear Science Symposium, Medical Imaging Conference and International Symposium on Room-Temperature Semiconductor Detectors (NSS MIC RTSD)*. IEEE, 2023. 33

[39] Michael G Levitzky. *Pulmonary Physiology*. McGraw-Hill Education: New York, NY, USA, 2007. 34, 36

[40] Susan Standring, Harold Ellis, and Caroline Wigley. *Gray's Anatomy: the Anatomical Basis of Clinical Practice*. Elsevier Churchill Livingstone, 39 edition, 2005. 34, 35

[41] Raheel Chaudhry and Bruno Bordoni. Anatomy, Thorax, Lungs. *Statpearls*, 2023. 34

[42] Moshe Haddad and Sandeep Sharma. Physiology, Lung. *Statpearls*, 2019. 36

[43] Kapwatt. Lung Volumes. https://commons.wikimedia.org/wiki/File:Illu_lung.jpg, 2023. 37

[44] Talmadge E. King. Clinical Advances in the Diagnosis and Therapy of the Interstitial Lung Diseases. *American Journal of Respiratory and Critical Care Medicine*, 172:268–279, 2005. 38

[45] David A Lynch, Nicola Sverzellati, William D Travis, Kevin K Brown, Thomas V Colby, Jeffrey R Galvin, Jonathan G Goldin, David M Hansell, Yoshikazu Inoue, Takeshi Johkoh, et al. Diagnostic Criteria for Idiopathic Pulmonary Fibrosis: a Fleischner Society White Paper. *The Lancet Respiratory Medicine*, 6(2):138–153, 2018. 38, 41, 42, 43, 44, 51

[46] Jaume Sauleda, Belén Núñez, Ernest Sala, and Joan B Soriano. Idiopathic Pulmonary Fibrosis: Epidemiology, Natural History, Phenotypes. *Medical Sciences*, 6(4):110, 2018. 38

[47] Giacomo Sgalla, Alice Biffi, and Luca Richeldi. Idiopathic Pulmonary Fibrosis: Diagnosis, Epidemiology and Natural History. *Respirology*, 21(3):427–437, 2016. 38, 45

[48] Vidya Navaratnam, KM Fleming, J West, CJP Smith, RG Jenkins, A Fogarty, and RB Hubbard. The Rising Incidence of Idiopathic Pulmonary Fibrosis in the UK. *Thorax*, 66(6):462–467, 2011. 38, 39

[49] John Hutchinson, Andrew Fogarty, Richard Hubbard, and Tricia McKeever. Global Incidence and Mortality of Idiopathic Pulmonary Fibrosis: a Systematic Review. *European Respiratory Journal*, 46(3):795–806, 2015. 38

[50] Giovanni Ferrara, Lisen Arnheim-Dahlström, Karen Bartley, Christer Janson, Klaus-Uwe Kirchgässler, Aaron Levine, and C Magnus Sköld. Epidemiology of Pulmonary Fibrosis: a Cohort Study Using Healthcare Data in Sweden. *Pulmonary Therapy*, 5:55–68, 2019. 38

[51] Ganesh Raghu, Shih-Yin Chen, Wei-Shi Yeh, Brad Maroni, Qian Li, Yuan-Chi Lee, and Harold R Collard. Idiopathic Pulmonary Fibrosis in US Medicare Beneficiaries Aged 65 Years and Older: Incidence, Prevalence, and Survival, 2001–11. *The Lancet Respiratory Medicine*, 2(7):566–572, 2014. 39

[52] Amy L Olson, Alex H Gifford, Naohiko Inase, Evans R Fernández Pérez, and Takafumi Suda. The Epidemiology of Idiopathic Pulmonary Fibrosis and Interstitial Lung Diseases at Risk of a Progressive-Fibrosing Phenotype. *European Respiratory Review*, 27(150), 2018. 39

[53] John P Hutchinson, Tricia M McKeever, Andrew W Fogarty, Vidya Navaratnam, and Richard B Hubbard. Increasing Global Mortality From Idiopathic Pulmonary Fibrosis in the Twenty-First Century. *Annals of the American Thoracic Society*, 11(8):1176–1185, 2014. 39

[54] J. Morgenstern, Emmalin Buajitti, Meghan O'Neill, Thomas Piggott, V. Goel, Daniel Fridman, K. Kornas, and L. Rosella. Predicting Population Health with Machine Learning: a Scoping Review. *BMJ Open*, 10, 2020. 39

[55] T. Wiemken and R. Kelley. Machine Learning in Epidemiology and Health Outcomes Research. *Annual Review of Public Health*, 2020. 39

[56] Jeffrey J Swigris, Michael K Gould, and Sandra R Wilson. Health-Related Quality of Life Among Patients with Idiopathic Pulmonary Fibrosis. *Chest*, 127(1):284–294, 2005. 40

[57] Jeffrey J Swigris, David L Streiner, Kevin K Brown, Amanda Belkin, Kathy E Green, Frederick S Wamboldt, IPFnet Investigators, et al. Assessing Exertional Dyspnea in Patients with Idiopathic Pulmonary Fibrosis. *Respiratory Medicine*, 108(1):181–188, 2014. 40

[58] Roland M Du Bois, Derek Weycker, Carlo Albera, Williamson Z Bradford, Ulrich Costabel, Alex Kartashov, Talmadge E King Jr, Lisa Lancaster, Paul W Noble, Steven A Sahn, et al. Forced Vital Capacity in Patients with Idiopathic Pulmonary Fibrosis: Test Properties and Minimal Clinically Important Difference. *American Journal of Respiratory and Critical Care Medicine*, 184(12):1382–1389, 2011. 40

[59] Andreas Guenther, Ekaterina Krauss, Silke Tello, Jasmin Wagner, Bettina Paul, Stefan Kuhn, Olga Maurer, Sabine Heinemann, Ulrich Costabel, María Asunción Nieto Barbero, et al. The European IPF Registry (Euripfreg): Baseline Characteristics and Survival of Patients with Idiopathic Pulmonary Fibrosis. *Respiratory Research*, 19:1–10, 2018. 40, 74

[60] Bernard Karnath. Digital Clubbing: A Sign of Underlying Disease. *Hospital Physician*, 39(9):25–27, 2003. 40

[61] Miaotian Cai, Min Zhu, Chengjun Ban, Jin Su, Qiao Ye, Yan Liu, Wen Zhao, Chen Wang, and Huaping Dai. Clinical Features and Outcomes of 210 Patients with Idiopathic Pulmonary Fibrosis. *Chinese Medical Journal*, 127(10):1868–1873, 2014. 40

[62] Talmadge E King Jr, Janet A Tooze, Marvin I Schwarz, Kevin R Brown, and Reuben M Cherniack. Predicting Survival in Idiopathic Pulmonary Fibrosis: Scoring System and Survival Model. *American Journal of Respiratory and Critical Care Medicine*, 164(7):1171–1181, 2001. 40

[63] An Zhao, Eyjolfur Gudmundsson, Nesrin Mogulkoc, Coline van Moorsel, Tamera J Corte, Pardeep Vasudev, Chiara Romei, Robert Chapman, Tim J M Wallis, Emma Denneny, et al. Mortality Surrogates in Combined Pulmonary Fibrosis and Emphysema. *The European Respiratory Journal*, 2023. 41

[64] Fasihul A Khan, Iain Stewart, Samuel Moss, Laura Fabbri, Karen A Robinson, Simon Johnson, and R Gisli Jenkins. Three month fvc change: a trial endpoint for ipf based on individual participant data meta-analysis. *medRxiv*, pages 2021–09, 2021. 41

[65] Julie Morisset, Eric Vittinghoff, Bo Young Lee, Roberto Tonelli, Xiaowen Hu, Brett M Elicker, Jay H Ryu, Kirk D Jones, Stefania Cerri, Andreina Manfredi, Marco Sebastiani, Andrew J Gross, Brett Ley, Paul J Wolters, Talmadge E King Jr, Dong Soon Kim, Harold R Collard, and Joyce S Lee. The Performance of the GAP Model in Patients with Rheumatoid Arthritis Associated Interstitial Lung Disease. *Respiratory Medicine*, 127:51–56, 2017. 41

[66] Xuening Wu, Chengsheng Yin, Xianqiu Chen, Yuan Zhang, Yiliang Su, Jingyun Shi, Dong Weng, Xing Jiang, Aihong Zhang, Wenqiang Zhang, and Huiping Li. Idiopathic Pulmonary Fibrosis Mortality Risk Prediction Based on Artificial Intelligence: the CTPF Model. *Frontiers in Pharmacology*, 13, 2022. 41

[67] Dong Soon Kim, Harold R Collard, and Talmadge E King Jr. Classification and Natural History of the Idiopathic Interstitial Pneumonias. *Proceedings of the American Thoracic Society*, 3(4):285–292, 2006. 41

[68] David M Hansell, Alexander A Bankier, Heber MacMahon, Theresa C McLoud, Nestor L Muller, and Jacques Remy. Fleischner Society: Glos-

sary of Terms for Thoracic Imaging. *Radiology*, 246(3):697–722, 2008. 41, 42

[69] Bruno Hochhegger, Edson Marchiori, Matheus Zanon, Adalberto Sperb Rubin, Renata Fragomeni, Stephan Altmayer, Carlos Roberto Ribeiro Carvalho, and Bruno Guedes Baldi. Imaging in Idiopathic Pulmonary Fibrosis: Diagnosis and Mimics. *Clinics*, 74, 2019. 41

[70] Emre Egriboz, Furkan Kaynar, Songül Varlı Albayrak, Benan Müsellim, and Tuba Selçuk. Finding and Following of Honeycombing Regions in Computed Tomography Lung Images By Deep Learning. *arXiv*, abs/1811.02651, 2018. 42

[71] Minna E Mononen, Hannu-Pekka Kettunen, Sanna-Katja Suoranta, Miia S Kärkkäinen, Tuomas A Selander, Minna K Purokivi, and Riitta L Kaarteenaho. Several Specific High-Resolution Computed Tomography Patterns Correlate with Survival in Patients with Idiopathic Pulmonary Fibrosis. *Journal of Thoracic Disease*, 13:2319 – 2330, 2021. 42

[72] David A Lynch, J David Godwin, Sharon Safrin, Karen M Starko, Phil Hormel, Kevin K Brown, Ganesh Raghu, Talmadge E King Jr, Williamson Z Bradford, David A Schwartz, and W Richard Webb. High-Resolution Computed Tomography in Idiopathic Pulmonary Fibrosis: Diagnosis and Prognosis. *American Journal of Respiratory and Critical Care Medicine*, 172 4:488–93, 2005. 42

[73] Nicola Sverzellati. Highlights of HRCT Imaging in IPF. *Respiratory Research*, 14, 2013. 42

[74] Avand Devaraj. Imaging: How to Recognise Idiopathic Pulmonary Fibrosis. *European Respiratory Review*, 23:215 – 219, 2014. 42

[75] Simon L F Walsh, John A Mackintosh, Lucio Calandriello, Mario Silva, Nicola Sverzellati, Anna Rita Larici, Stephen M Humphries, David A Lynch, Helen E Jo, Ian Glaspole, Christopher Grainge, Nicole Goh, Peter M A Hopkins, Yuben Moodley, Paul N Reynolds, Christopher Zappala, Gregory Keir, Wendy A Cooper, Annabelle M Mahar, Samantha Ellis, Athol U Wells, and Tamera J Corte. Deep Learning-Based Outcome Prediction in Progressive Fibrotic Lung Disease Using High-Resolution Computed Tomography. *American Journal of Respiratory and Critical Care Medicine*, 2022. 42

[76] Vincent Cottin, Bruno Crestani, Dominique Valeyre, Benoit Wallaert, Jacques Cadranel, Jean-Charles Dalphin, Philippe Delaval, Dominique Israel-Biet, Romain Kessler, Martine Reynaud-Gaubert, et al. Diagnosis and Management of Idiopathic Pulmonary Fibrosis: French Practical Guidelines. *European Respiratory Review*, 23(132):193–214, 2014. 45, 117

[77] Maria D. Martin, Jonathan H. Chung, and J. Kanne. Idiopathic Pulmonary Fibrosis. *Journal of Thoracic Imaging*, 31:127–139, 2016. 45

[78] Athol U Wells. The Revised ATS/ERS/JRS/ALAT Diagnostic Criteria for Idiopathic Pulmonary Fibrosis (IPF) - Practical Implications. *Respiratory Research*, 14, 2013. 45

[79] Yasuhiro Kondoh, Hiroyuki Taniguchi, Kensuke Kataoka, Taiki Furukawa, Ayumi Shintani, Tomoyuki Fujisawa, Takafumi Suda, Machiko Arita, Tomohisa Baba, Kazuya Ichikado, Yoshikazu Inoue, Kazuma Kishi, Tomoo Kishaba, Osamu Nishiyama, Takashi Ogura, Keisuke Tomii, and Sakae Homma. Clinical Spectrum and Prognostic Factors of Possible UIP Pattern on High-Resolution CT in Patients Who Underwent Surgical Lung Biopsy. *PLoS One*, 13, 2018. 45

[80] Gabriella Pezzuto, Giulia Claroni, Ermanno Puxeddu, Armando Fusco,

Francesco Cavalli, Simone Altobelli, Silvia Portalone, Maurizio Zompatori, Giovanni Simonetti, Cesare Saltini, and Gianluigi Sergiacomi. Structured Multidisciplinary Discussion of HRCT Scans for IPF/UIP Diagnosis May Result in Indefinite Outcomes. *Sarcoidosis, Vasculitis, and Diffuse Lung Diseases: Official Journal of WASOG*, 32 1:32–6, 2015. 45

[81] Laura M Glenn and Tamera J Corte. Diagnosing Idiopathic Pulmonary Fibrosis: Has the Time for Surgical Lung Biopsy Passed? *Respirology*, 25:1112 – 1113, 2020. 45

[82] Manoj V. Maddali, A. Kalra, Michael Muelly, and Joshua J. Reicher. Development and Validation of a CT-Based Deep Learning Algorithm to Augment Non-Invasive Diagnosis of Idiopathic Pulmonary Fibrosis. *Respiratory Medicine*, 2023. 45

[83] Turkey Refaee, Zohaib Salahuddin, Anne-Noelle Frix, Chenggong Yan, Guangyao Wu, Henry C Woodruff, Hester Gietema, Paul Meunier, Renaud Louis, Julien Guiot, and Philippe Lambin. Diagnosis of Idiopathic Pulmonary Fibrosis in High-Resolution Computed Tomography Scans Using a Combination of Handcrafted Radiomics and Deep Learning. *Frontiers in Medicine*, 9, 2022. 45, 69

[84] Harold R Collard, Talmadge E King, Becki Bucher Bartelson, Jason S Vourlekis, Marvin I Schwarz, and Kevin K Brown. Changes in Clinical and Physiologic Variables Predict Survival in Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 168(5):538–542, 2003. 46

[85] Sanja Stanojevic, Angie Wade, and Janet Stocks. Reference Values for Lung Function: Past, Present and Future. *European Respiratory Journal*, 36(1):12–19, 2010. 46

[86] Miya O. Paterniti, Youwei Bi, Dinko Rekić, Yaning Wang, Banu A. Karimi-Shah, and Badrul A. Chowdhury. Acute Exacerbation and Decline in Forced Vital Capacity Are Associated with Increased Mortality in Idiopathic Pulmonary Fibrosis. *Annals of the American Thoracic Society*, 14(9):1395–1402, 2017. 46

[87] Banu A. Karimi-Shah and Badrul A. Chowdhury. Forced Vital Capacity in Idiopathic Pulmonary Fibrosis — FDA Review of Pirfenidone and Nintedanib. *New England Journal of Medicine*, 372(13):1189–1191, 2015. 47

[88] Shelley L. Schmidt, Nabihah Tayob, Meilan K. Han, Christopher Zappala, Dolly Kervitsky, Susan Murray, Athol U. Wells, Kevin K. Brown, Fernando J. Martinez, and Kevin R. Flaherty. Predicting Pulmonary Fibrosis Disease Course From Past Trends in Pulmonary Function. *Chest*, 145(3):579–585, 2014. 47

[89] Luca Richeldi, Bruno Crestani, Arata Azuma, Martin Kolb, Moisés Selman, Wibke Stansen, Manuel Quaresma, Susanne Stowasser, and Vincent Cottin. Outcomes Following Decline in Forced Vital Capacity in Patients with Idiopathic Pulmonary Fibrosis Results From the INPULSIS and INPULSIS-ON Trials of Nintedanib. *Respiratory Medicine*, 156:20–25, 2019. 47

[90] Amirala Alavi Foumani, Seyyed Ali Alavi Foumani, Mirsaeed Attarchi, Alireza Etemadi Deilami, Behzad Majlesi, Shima Ildari, and Habib Eslami-Kenarsari. Quality of Spirometry Tests in the Field of Occupational Health. *BMC Research Notes*, 17(1):11, 2024. 47

[91] Matthew J Hegewald, Heather M Gallo, and Emily L Wilson. Accuracy and Quality of Spirometry in Primary Care Offices. *Annals of the American Thoracic Society*, 13(12):2119–2124, 2016. 47

[92] Ganesh Raghu, Harold R Collard, Kevin J Anstrom, Kevin R Flaherty, Thomas R Fleming, Talmadge E King, Fernando J Martinez, and Kevin K Brown. Idiopathic Pulmonary Fibrosis: Clinically Meaningful Primary Endpoints in Phase 3 Clinical Trials. *American Journal of Respiratory and Critical Care Medicine*, 185(10):1044–1048, 2012. 47

[93] Talmadge E King, Carlo Albera, Williamson Z. Bradford, Ulrich Costabel, Phil Hormel, Lisa Lancaster, Paul W. Noble, Steven A. Sahn, Javier Szwarcberg, Michiel Thomeer, Dominique Valeyre, and Roland M. du Bois. Effect of Interferon Gamma-1b on Survival in Patients with Idiopathic Pulmonary Fibrosis (INSPIRE): Multicentre, Randomised, Placebo-Controlled Trial. *The Lancet*, 374(9685):222–228, 2009. 47

[94] Brett Ley, Harold R Collard, and Talmadge E King. Clinical Course and Prediction of Survival in Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 183(4):431–440, 2011. 47

[95] Jessica Adkins and Harold R Collard. Idiopathic Pulmonary Fibrosis. *Seminars in Respiratory and Critical Care Medicine*, 33 5:433–9, 2019. 48

[96] Steven Nathan, Ming Yang, Elizabeth Morgenthien, and John Stauffer. FVC Variability in Patients with Idiopathic Pulmonary Fibrosis and Role of 6-Min Walk Test to Predict Further Change. *The European Respiratory Journal*, 55, 2020. 48

[97] Helen E Jo, Ian Glaspole, Yuben Moodley, Sally Chapman, Samantha Ellis, Nicole Goh, Peter Hopkins, Greg Keir, Annabelle Mahar, Wendy Cooper, et al. Disease Progression in Idiopathic Pulmonary Fibrosis with Mild Physiological Impairment: Analysis From the Australian IPF Registry. *BMC Pulmonary Medicine*, 18, 2018. 48

[98] Daniel S Glass, David Grossfeld, Heather A Renna, Priya Agarwala, Peter Spiegler, Joshua DeLeon, and Allison B Reiss. Idiopathic Pulmonary Fibrosis: Current and Future Treatment. *The Clinical Respiratory Journal*, 16(2):84–96, 2022. 49

[99] T. King, W. Bradford, S. Castro-Bernardini, E. Fagan, I. Glaspole, M. Glassberg, E. Gorina, P. Hopkins, D. Kardatzke, L. Lancaster, D. Lederer, S. Nathan, C. Pereira, S. Sahn, R. Sussman, J. Swigris, and P. Noble. A Phase 3 Trial of Pirfenidone in Patients with Idiopathic Pulmonary Fibrosis. *The New England Journal of Medicine*, 370 22:2083–92, 2014. 49

[100] Antje Moeller, Kjetil Ask, David Warburton, Jack Gauldie, and Martin Kolb. The Bleomycin Animal Model: a Useful Tool to Investigate Treatment Options for Idiopathic Pulmonary Fibrosis? *The International Journal of Biochemistry & Cell Biology*, 40 3:362–82, 2008. 49

[101] Vivien Somogyi, Nazia Chaudhuri, Sebastiano Emanuele Torrisi, Nicolas Kahn, Veronika Müller, and Michael Kreuter. The Therapy of Idiopathic Pulmonary Fibrosis: What Is Next? *European Respiratory Review*, 28(153), 2019. 49

[102] David Weill, Christian Benden, Paul A Corris, John H Dark, R Duane Davis, Shaf Keshavjee, David J Lederer, Michael J Mulligan, G Alexander Patterson, Lianne G Singer, et al. A Consensus Document for the Selection of Lung Transplant Candidates: 2014—an Update From the Pulmonary Transplantation Council of the International Society for Heart and Lung Transplantation. *The Journal of Heart and Lung Transplantation*, 34(1):1–15, 2015. 49

[103] Yiwen Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R. Mak, and H. Aerts. Deep Learning Predicts Lung Cancer Treatment Response From Serial Medical Imaging. *Clinical Cancer Research*, 25:3266 – 3275, 2019. 49

[104] Oh-Beom Kwon, Solji Han, Hwawook Lee, H. Kang, Sung Kyoung Kim, J. S. Kim, Chankwon Park, S. H. Lee, Seungjoon Kim, Jin Woo Kim, and C. Yeo. Prediction of Postoperative Lung Function in Lung Cancer Patients Using Machine Learning Models. *Tuberculosis and Respiratory Diseases*, 86:203 – 215, 2023. 49

[105] Sébastien Benzekry, Mathieu Grangeon, Mélanie Karlsen, Maria Alexa, Isabella Bicalho-Frazeto, Solène Chaléat, Pascale Tomasini, Dominique Barbolosi, Fabrice Barlesi, and Laurent Greillier. Machine Learning for Prediction of Immunotherapy Efficacy in Non-Small Cell Lung Cancer From Simple Clinical and Biological Data. *Cancers*, 13, 2021. 49

[106] Yutaro Nakamura and Takafumi Suda. Idiopathic Pulmonary Fibrosis: Diagnosis and Clinical Manifestations. *Clinical Medicine Insights: Circulatory, Respiratory and Pulmonary Medicine*, 9, 2015. 51

[107] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object Detection in 20 Years: A Survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 54

[108] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020. 54

[109] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A Survey on Instance Segmentation: State of the Art. *International Journal of Multimedia Information Retrieval*, 9(3):171–189, 2020. 54

[110] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in Vision: A Survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–41, 2022. 54

[111] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018. 54

[112] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep Learning–Based Text Classification: a Comprehensive Review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021. 54

[113] Cristina Garbacea and Qiaozhu Mei. Neural Language Generation: Formulation, Methods, and Evaluation. *arXiv Preprint arXiv:2007.15780*, 2020. 54

[114] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. Significance of Machine Learning in Healthcare: Features, Pillars and Applications. *International Journal of Intelligent Networks*, 3:58–73, 2022. 54

[115] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 54

[116] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012. 54, 55, 76, 114

[117] Kevin P Murphy. *Machine Learning: a Probabilistic Perspective*. MIT press, 2012. 54

[118] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016. 54

[119] Andriy Burkov. *The Hundred-Page Machine Learning Book*, volume 1. Andriy Burkov Quebec City, QC, Canada, 2019. 54

[120] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. A Gentle Introduction to Imputation of Missing Values. *Journal of Clinical Epidemiology*, 59(10):1087–1091, 2006. 55, 57

[121] Marianne Riksheim Stavseth, Thomas Clausen, and Jo Røislien. How Handling Missing Data May Impact Conclusions: A Comparison of Six Different Imputation Methods for Categorical Questionnaire Data. *SAGE Open Medicine*, 7:2050312118822912, 2019. 55, 57

[122] Donald B Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976. 55

[123] Frederico Z. Poleto, J. Singer, and C. Paulino. Missing Data Mechanisms and Their Implications on the Analysis of Categorical Data. *Statistics and Computing*, 21:31–43, 2011. 55

[124] Mortaza Jamshidian and Matthew Mata. Postmodeling Sensitivity Analysis to Detect the Effect of Missing Data Mechanisms. *Multivariate Behavioral Research*, 43:432 – 452, 2008. 55

[125] Elad Hazan, Roi Livni, and Yishay Mansour. Classification with Low Rank and Missing Data. In *International Conference on Machine Learning*, pages 257–266. PMLR, 2015. 56

[126] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. Multivariate Time Series Imputation with Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 31, 2018. 56

[127] Marek Śmieja, Łukasz Struski, Jacek Tabor, Bartosz Zieliński, and Przemysław Spurek. Processing of Missing Data By Neural Networks. *Advances in Neural Information Processing Systems*, 31, 2018. 56

[128] Joonyoung Yi, Juhyuk Lee, Kwang Joon Kim, Sung Ju Hwang, and Eunho Yang. Why Not to Use Zero Imputation? Correcting Sparsity Bias in Training Neural Networks. In *International Conference on Learning Representations*, 2019. 57

[129] Stef Van Buuren. Multiple Imputation of Discrete and Continuous Data By Fully Conditional Specification. *Statistical Methods in Medical Research*, 16(3):219–242, 2007. 57

[130] Elisa T Lee and John Wang. *Statistical Methods for Survival Data Analysis*, volume 476. John Wiley & Sons, 2003. 58

[131] Ping Wang, Yan Li, and Chandan K Reddy. Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019. 59, 60, 63, 65

[132] Olive Jean Dunn and Virginia A Clark. *Basic Statistics: a Primer for the Biomedical Sciences*. John Wiley & Sons, 2009. 59

[133] John Simes and Marvin Zelen. Exploratory Data Analysis and the Use of the Hazard Function for Interpreting Survival Data: an Investigator's Primer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, pages 1418–1431, 1985. 59

[134] Kenneth Hess and Victor A. Levin. Getting More Out of Survival Data By Using the Hazard Function. *Clinical Cancer Research*, 20:1404 – 1409, 2014. 59

[135] D. Watt, T. Aitchison, R. MacKie, and J. Sirel. Survival Analysis: the Importance of Censored Observations. *Melanoma Research*, 6:379–385, 1996. 60

[136] Stephen W Lagakos. General Right Censoring and Its Impact on the Analysis of Survival Data. *Biometrics*, 35:139, 1979. 60

[137] Edward L Kaplan and Paul Meier. Nonparametric Estimation From Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. 61

[138] Madison Jane Mathiason. *Survival Analysis of Treatment Effect For Brain Cancer Based on the Surveillance, Epidemiology, and End Results Database*. PhD thesis, North Dakota State University, 2020. 62

[139] Elisabetta Balestro, Fiorella Calabrese, Graziella Turato, Francesca Lunardi, Erica Bazzan, Giuseppe Marulli, Davide Biondini, Emanuela Rossi, Alessandro Sanduzzi, Federico Rea, et al. Immune Inflammation and Disease Progression in Idiopathic Pulmonary Fibrosis. *PLoS One*, 11(5), 2016. 63

[140] Norman Breslow. Covariance Analysis of Censored Survival Data. *Biometrics*, pages 89–99, 1974. 63

[141] Eu-Tteum Baek, Hyung Jeong Yang, Soo Hyung Kim, Guee Sang Lee, In-Jae Oh, Sae-Ryung Kang, and Jung-Joon Min. Survival Time Prediction By Integrating Cox Proportional Hazards Network and Distribution Function Network. *BMC Bioinformatics*, 22:1–15, 2021. 63

[142] Imad Bou-Hamad, Denis Larocque, and Hatem Ben-Ameur. A Review of Survival Trees. *Statistics Surveys*, 5:44 – 71, 2011. 63

[143] Louis Gordon and Richard A Olshen. Tree-Structured Survival Analysis. *Cancer Treatment Reports*, 69(10):1065–1069, 1985. 63

[144] Roger B Davis and James R Anderson. Exponential Survival Trees. *Statistics in Medicine*, 8(8):947–961, 1989. 63

[145] Michael LeBlanc and John Crowley. Relative Risk Trees for Censored Survival Data. *Biometrics*, pages 411–425, 1992. 63

[146] Antonio Ciampi, Johanne Thiffault, Jean-Pierre Nakache, and Bernard Asselain. Stratification By Stepwise Regression, Correspondence Analysis and Recursive Partition: a Comparison of Three Methods of Analysis for Survival Data with Covariates. *Computational Statistics & Data Analysis*, 4(3):185–204, 1986. 63

[147] Antonio Ciampi, Ching-Haur Chang, Sheilah Hogg, and Steve McKinney. Recursive Partition: A Versatile Method for Exploratory-Data Analysis in Biostatistics. *Biostatistics: Advances in Statistical Sciences Festschrift in Honor of Professor VM Joshi's 70th Birthday Volume V*, pages 23–50, 1987. 63

[148] Mark Robert Segal. Regression Trees for Censored Data. *Biometrics*, pages 35–47, 1988. 63

[149] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random Survival Forests. *The Annals of Applied Statistics*, pages 841–860, 2008. 64

[150] Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001. 64

[151] Jerome H Friedman. Greedy Function Approximation: a Gradient Boosting Machine. *Annals of Statistics*, pages 1189–1232, 2001. 64

[152] Harald Binder and Martin Schumacher. Allowing for Mandatory Covariates in Boosting Estimation of Sparse High-Dimensional Survival Models. *BMC Bioinformatics*, 9:1–10, 2008. 64

[153] Corinna Cortes. Support-Vector Networks. *Machine Learning*, 1995. 64

[154] Alex J Smola and Bernhard Schölkopf. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14:199–222, 2004. 64

[155] Pannagadatta K Shivaswamy, Wei Chu, and Martin Jansche. A Support Vector Approach to Censored Targets. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 655–660. IEEE, 2007. 64

[156] Vanya Van Belle, Kristiaan Pelckmans, Johan AK Suykens, and Sabine Van Huffel. Support Vector Machines for Survival Analysis. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (Cimed2007)*, pages 1–8, 2007. 64

[157] Vanya Van Belle, Kristiaan Pelckmans, Sabine Van Huffel, and Johan AK Suykens. Support Vector Methods for Survival Analysis: a Comparison Between Ranking and Regression Approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, 2011. 64

[158] Tamara Fernández, Nicolás Rivera, and Yee Whye Teh. Gaussian Processes for Survival Analysis. *Advances in Neural Information Processing Systems*, 29, 2016. 65

[159] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology*, 18(1):1–12, 2018. 65, 83, 84, 90, 98, 103, 113

[160] John P Klein and Melvin L Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*, volume 1230. Springer, 2003. 65

[161] Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine*, 15(4):361–387, 1996. 68

[162] Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011. 69

[163] Yu Shi, Weng Kee Wong, Jonathan G. Goldin, Matthew S. Brown, and Grace Hyun J. Kim. Prediction of progression in idiopathic pulmonary fibrosis using CT scans at baseline: A quantum particle swarm optimization - Random forest approach. *Artificial intelligence in medicine*, 100, 2019. 69

[164] Joseph Jacob, Brian J. Bartholmai, Srinivasan Rajagopalan, Maria Kokosi, Arjun Nair, Ronald Karwoski, Simon L.F. Walsh, Athol U. Wells, and David M. Hansell. Mortality Prediction in Idiopathic Pulmonary Fibrosis: Evaluation of Computer-Based CT Analysis with Conventional Severity Measures. *European Respiratory Journal*, 49(1):1601011, 2017. 69, 95

[165] Tomohiro Handa, Kiminobu Tanizawa, Tsuyoshi Oguma, Ryuji Uozumi, Kizuku Watanabe, Naoya Tanabe, Takafumi Niwamoto, Hiroshi Shima, Ryobu Mori, Tomomi W. Nobashi, Ryo Sakamoto, Takeshi Kubo, Atsuko Kurosaki, Kazuma Kishi, Yuji Nakamoto, and Toyohiro Hirai. Novel Artificial Intelligence-based Technology for Chest Computed Tomography Analysis of Idiopathic Pulmonary Fibrosis. *Annals of the American Thoracic Society*, 2021. 69

[166] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations

From Deep Networks Via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, volume 128, pages 618–626, 2017. 69, 116

[167] Margaret L Salisbury, Meng Xia, Yueren Zhou, Susan Murray, Nabihah Tayob, Kevin K Brown, Athol U Wells, Shelley L Schmidt, Fernando J Martinez, and Kevin R Flaherty. Idiopathic Pulmonary Fibrosis: Gender-Age-Physiology Index Stage for Predicting Future Lung Function Decline. *Chest*, 149(2):491–498, 2016. 73

[168] Jing Gao, D. Kalafatis, Lisa Carlson, Ida Pesonen, Chuan xing Li, Å. Wheelock, J. Magnusson, and C. Sköld. Baseline Characteristics and Survival of Patients of Idiopathic Pulmonary Fibrosis: a Longitudinal Analysis of the Swedish IPF Registry. *Respiratory Research*, 22, 2021. 74

[169] Jürgen Behr, Antje Prasse, Hubert Wirtz, Dirk Koschel, David Pittrow, Matthias Held, Jens Klotsche, Stefan Andreas, Martin Claussen, Christian Grohé, et al. Survival and Course of Lung Function in the Presence or Absence of Antifibrotic Treatment in Patients with Idiopathic Pulmonary Fibrosis: Long-Term Results of the INSIGHTS-IPF Registry. *European Respiratory Journal*, 56(2), 2020. 74

[170] Elizabeth R Volkmann, Donald P Tashkin, Myung Sim, Ning Li, Ellen Goldmuntz, Lynette Keyes-Elstein, Ashley Pinckney, Daniel E Furst, Philip J Clements, Dinesh Khanna, et al. Short-Term Progression of Interstitial Lung Disease in Systemic Sclerosis Predicts Long-Term Survival in Two Independent Clinical Trial Cohorts. *Annals of the Rheumatic Diseases*, 78(1):122–130, 2019. 74

[171] Jean Pastre, Scott Barnett, Inga Ksovreli, Jeannie Taylor, A Whitney Brown, Oksana A Shlobin, Kareem Ahmad, Vikramjit Khangoora, Shambhu Aryal,

Christopher S King, et al. Idiopathic Pulmonary Fibrosis Patients with Severe Physiologic Impairment: Characteristics and Outcomes. *Respiratory Research*, 22:1–10, 2021. 74

[172] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum Likelihood From Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 75, 76

[173] Charles R Harris, K Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. 80

[174] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and Statistical Modeling with Python. In *9th Python in Science Conference*, 2010. 80

[175] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised Feature Learning Via Non-Parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 84

[176] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic Lung Segmentation in Routine Imaging Is Primarily a Data Diversity Problem, Not a Methodolog Problem. *European Radiology Experimental*, 4(1):1–13, 2020. 88, 108

[177] 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*, pages 770–778, 2016. 89

[178] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: the Missing Ingredient for Fast Stylization. *arXiv Preprint arXiv:1607.08022*, 2016. 89

[179] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *International Conference on Machine Learning*, volume 30. PMLR, 2013. 89, 90

[180] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training By Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015. 90

[181] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2018. 90, 104

[182] Sang Hoon Lee, Jong Sun Park, Song Yee Kim, Dong Soon Kim, Young Whan Kim, Man Pyo Chung, Soo Taek Uh, Choon Sik Park, Sung Woo Park, Sung Hwan Jeong, et al. Comparison of CPI and GAP Models in Patients with Idiopathic Pulmonary Fibrosis: a Nationwide Cohort Study. *Scientific Reports*, 8(1):4784, 2018. 95

[183] Charles Sharp, Huzaifa I Adamali, and Ann B Millar. A Comparison of Published Multidimensional Indices to Predict Outcome in Idiopathic Pulmonary Fibrosis. *European Respiratory Journal Open Research*, 3(1), 2017. 95

[184] Miia Kärkkäinen, Hannu-Pekka Kettunen, Hanna Nurmi, Tuomas Selander, Minna Purokivi, and Riitta Kaarteenaho. Comparison of Disease Progression Subgroups in Idiopathic Pulmonary Fibrosis. *BMC Pulmonary Medicine*, 19:1–9, 2019. 95

[185] Harold R Collard, Talmadge E King, Becki Bucher Bartelson, Jason S Vourlekis, Marvin I Schwarz, and Kevin K Brown. Changes in Clinical

and Physiologic Variables Predict Survival in Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 168(5):538–542, 2003. 95

[186] Brian J. Bartholmai, Sushravya Raghunath, Ronald A. Karwoski, Teng Moua, Srinivasan Rajagopalan, Fabien Maldonado, Paul A. Decker, and Richard A. Robb. Quantitative Computed Tomography Imaging of Interstitial Lung Diseases. *Journal of Thoracic Imaging*, 28(5):298–307, 2013. 95

[187] Chiara Romei, Laura M Tavanti, Alessandro Taliani, Annalisa De Liperi, Ronald Karwoski, Alessandro Celi, Antonio Palla, Brian J Bartholmai, and Fabio Falaschi. Automated Computed Tomography Analysis in the Assessment of Idiopathic Pulmonary Fibrosis Severity and Progression. *European Journal of Radiology*, 124, 2020. 95

[188] David A. Nix and Andreas S. Weigend. Estimating the Mean and Variance of the Target Probability Distribution. *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, 1:55–60 vol.1, 1994. 104

[189] Abhinav Agrawal, Isha Verma, Varun Shah, Abhishek Agarwal, and Rutuja R Sikachi. Cardiac Manifestations of Idiopathic Pulmonary Fibrosis. *Intractable & Rare Diseases Research*, 5(2):70–75, 2016. 108

[190] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computing*, 9:1735–1780, 1997. 114

[191] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 114, 119

[192] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. 116

[193] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. 116

[194] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org, 2017. 116

[195] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017. 116

[196] Simon Walsh, Jan De Backer, Helmut Prosch, Georg Langs, Lucio Calandriello, Vincent Cottin, Kevin K. Brown, Yoshikazu Inoue, Vasilios Tzilas, and Elizabeth Estes. Towards the adoption of quantitative computed tomography in the management of interstitial lung disease. *European Respiratory Review*, 33, 2024. 117

[197] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv Preprint arXiv:2303.08774*, 2023. 118, 119

[198] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro,

Faisal Azhar, et al. Llama: Open and Efficient Foundation Language Models. *arXiv Preprint arXiv:2302.13971*, 2023. 118, 119

[199] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv Preprint arXiv:2310.06825*, 2023. 118, 119

[200] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 2023. 118

[201] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 118

[202] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022. 118

[203] Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical Image Understanding with Pretrained Vision Language Models: A Comprehensive Study. *ArXiv*, abs/2209.15517, 2022. 118

[204] Shenmin Zhang, Yanbo Xu, Naoto Usuyama, J. Bagga, Robert Tinn, Sam Preston, Rajesh N. Rao, Mu-Hsin Wei, Naveen Valluri, Cliff Wong, M. Lungren, Tristan Naumann, and Hoifung Poon. Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing. *ArXiv*, abs/2303.00915, 2023. 118

[205] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34:50–70, 2018. 120

[206] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020. 123

[207] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 123

[208] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 123