Mathematics

Unsupervised learning for texture based prediction on longitudinal medical imaging data

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Biomedical Engineering

eingereicht von

Jeanny Pan
Matrikelnummer 01341597

an der Fakultät für Mathematik und Geoinformation

der Technischen Universität Wien

Betreuung: Prof.Dipl.-Ing. Dr.techn. Georg Langs

Wien, 10. Februar 2023

Jeanny Pan

Georg Langs

Mathematics

Unsupervised learning for texture based prediction on longitudinal medical imaging data

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Biomedical Engineering

by

Jeanny Pan
Registration Number 01341597

to the Faculty of Discrete Mathematics and Geometry

at the TU Wien

Advisor: Prof.Dipl.-Ing. Dr.techn. Georg Langs

Vienna, 10th February, 2023

Jeanny Pan

Georg Langs

Erklärung zur Verfassung der Arbeit

Jeanny Pan

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 10. Februar 2023

Jeanny Pan

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die mich bei meiner Arbeit unterstützt haben. Zuallererst möchte ich mich bei Prof. Georg Langs vom Computational Imaging Research Lab (CIR) an der Medizinischen Universität Wien für seine Hilfe während der gesamten Arbeit bedanken, von der Identifizierung des Forschungsziels bis zum abschließenden Korrekturlesen. Diese Arbeit wäre ohne seine Hilfe beim Austausch von Erkenntnissen, bei der Diskussion der Ergebnisse und bei der ständigen Verbesserung und Weiterentwicklung von Gedanken und Ideen nicht möglich gewesen. Vielen Dank für die ständige Begleitung und Betreuung vom Beginn des Projektes bis zur fertigen Arbeit!!! Außerdem möchte ich mich bei Prof. Helmut Prosch, Dr. Sebastian Roehrich und Dr. Florian Prayer für die Bereitstellung von CT-Daten und die Beantwortung all meiner klinischen Fragen bedanken. Ich danke auch Dr. Johannes Hofmanninger und Dr. Roxane Licandro, die mir bei allen technischen Schwierigkeiten geholfen haben. Meine Eltern verdienen eine besondere Erwähnung für ihre unermüdliche Unterstützung während meines gesamten Lebens, nicht nur während der Arbeit und meiner akademischen Bemühungen. Danke, ohne euch hätte ich es nicht so weit gebracht.

Acknowledgements

At this time, I want to thank everyone who assisted me with my thesis. First of all, I would like to thank Prof. Georg Langs from the Computational Imaging Research Lab (CIR) at the Medical University of Vienna for his help throughout the whole thesis, from identifying the purpose of the research until the final proofreading. This thesis would have been impossible to formulate without his help in exchange of findings, discussion of the results, and the continual improvement and growth of thoughts and ideas. Thank you so much for the constant guidance and supervision from the start of the project to the completed thesis!! Furthermore, I want to thank Prof. Helmut Prosch, Dr Sebastian Roehrich and Dr Florian Prayer for providing CT data and answering all my clinical questions. I also thank Dr Johannes Hofmanninger and Dr Roxane Licandro for helping me with all the technical difficulties. My parents deserve a particular mention for their unwavering support throughout my life, not just throughout the thesis and my academic endeavors. Thank you, I wouldn't have made it that far without you.

Kurzfassung

Die Idiopathischelungenfibrose ist eine schwere und tödliche Krankheit. Die mediane Überlebenszeit liegt bei Patienten ohne Behandlung zwischen 3-5 Jahren. Daher sind die Früherkennung und das Fortschreiten der Krankheit für die Analyse von entscheidender Bedeutung, aber diese Aufgabe kann eine Herausforderung sein, da diese Ergebnisse voneinander abweichen. Die Methode besteht aus sieben Schritten: Bildsegmentierung, Merkmalsextraktion, Merkmalsclusterung mithilfe von *Bag of Visual Words*, Bildregistrierung innerhalb eines Patienten, Identifizierung von Markern für das Fortschreiten der Krankheit, ein Übergangsnetzwerk für das Fortschreiten der Krankheit und Ergebnisvorhersage. Die Bildsegmentierung ist notwendig, um das Lungenvolumen aus den Computertomografiescans zu erhalten. Durch Übersegmentierung wird jedes Voxel auf ein Supervoxel der Größe $0,5cm^3$ statt $0,7mm^3$ reduziert. Jedes Supervoxel wird einem bestimmten Lungenmustertyp zugewiesen. Anhand der Häufigkeit der gegebenen Lungenmustertypen im Klassifikationsmodell haben wir eine Reihe von Krankheitsmuster-Marker-Kandidaten erkannt. Die Marker-Indikatoren wurden in einem Wiederholbarkeits-Setup mit 20 Zufallsläufen gefunden. Für vier Kandidaten mit der höchsten Stabilitätseinstufung ergibt die Überlebensvorhersage unterschiedliche Ergebnisse für Gruppen mit ähnlichen Mustersignaturen, für einen Validierungsdatensatz liefert das Ergebnis eine gleichwertige Aussage.

Abstract

Idiopathic pulmonary fibrosis (IPF) is a severe and lethal disease. The median survival outcome is between 3-5 years for patients without a treatment. Therefore, early detection and quantification of disease progression are essential for the steering of treatment. Treatment decisions are challenging, since we don't understand the relationship between present disease appearance, outcome and treatment response, yet. In this thesis we develop and evaluate a methodology to quantitatively assess changes associated with IPF, and to predict future outcome based on imaging data for patients. The methodology consists of 7 steps, image segmentation, features extraction, features clustering using *Bag of Visual Words*, intra-patient image registration, identification of disease progression markers, a transition network related to disease progression, and outcomes prediction. Image segmentation is necessary to obtain the lung volume from the computer tomography scans. Over-segmentation is applied to reduce each voxel to a supervoxel of size $0.5cm^3$ instead of $0.7mm^3$. Each supervoxel is assigned to a given lung pattern type. We recognized a set of disease pattern marker candidates through the frequency rate of the given lung pattern types in the classification model. The marker indicators were found in a repeatability setup with 20 random runs. For four top stability ranked candidates, the outcome survival prediction yields different outcomes for groups with similar pattern signatures, for a validation dataset, the result delivers an equivalent statement.

Contents

CHAPTER 1

Introduction

Idiopathic pulmonary fibrosis (IPF), is the most frequent form of interstitial lung disease (ILD) [1]. It is a major contributor to roughly 20% of all reported cases of Interstitial Lung Disease (ILD), and it is the most widespread and severe form of Idiopathic Interstitial Pneumonia (IIP) [2]. IIP encompasses a collection of ILDs with undefined causes and differing patterns of inflammation and fibrosis, but they share similar clinical, physiological, and radiologic features [3]. IPF was diagnosed in roughly 40,000 new patients across Europe in 2011, with the United Kingdom accounting for more than 12.5% of all reported cases [4]. The majority of patients diagnosed with IPF suffer from a gradual decline in lung function throughout the course of their illness. However, there is a minority of people who maintain their current level of function and do not exhibit any indications of worsening [5]. IPF patient has a median survival period of between three and five years [6][7][8]. This disease progression cannot be reversed, therefore a possible prediction of the disease progression is crucial. Although this is challenging as disease course in IPF are quite diverse. The CT-lung patterns play a crucial role in the diagnosis and treatment decisions for IPF. The accuracy of CT scans in diagnosing IPF based solely on imaging findings is high, which is why this method is often relied upon in the diagnosis process. These images provide valuable information that can aid in making treatment decisions. However, the difficulty of consistently recognizing these patterns makes it challenging to diagnose IPF accurately. The ability to detect these patterns with more certainty and to understand their relationship to the future course of the disease and the risk of progression is of utmost importance. A better understanding of CT-lung patterns can lead to earlier diagnosis, improved treatment outcomes, and ultimately, a better quality of life for patients with IPF.

1.1 Problem statement

a) An automatic method to process lung CT data.
b) Identify disease patterns that are associated with progression.
c) Learn to predict future disease courses with other markers.

1.2 Aim of the thesis

This work aims to develop an unsupervised machine learning method to identify quantitative radiological imaging markers related to IPF disease progression. The ability of early IPF recognition and prognosis of the disease development course and future outcomes are highly needed in clinical environments. The main contributions are:

- Improvement of image segmentation methods for high density lung patterns such as ground glass opacity, honeycombing and reticular patterns

- Exploration of various methods to extract features from High Resolution Computer Tomography data of the lung

- Identification of novel imaging marker patterns associated with disease progression

- Recognizing pathways of lung tissue transition during the progression of the disease

- Evaluation of these novel markers for outcome regarding their ability to predict

1.3 Thesis Outline

The thesis consists of 8 chapters and is structured as follows:

Chapter 1: Introduction summarizes the purpose of the thesis as well as the rationale behind writing it. In addition to that, it contains an overview of the methodological approach that was employed in this thesis.

Chapter 2: Medical background provides the medical background of the thesis and the latest clinical knowledge about IPF.

Chapter 3: Chest Imaging examines the physical and technological foundations of the relevant chest imaging modalities and their benefits and drawbacks.

Chapter 4: State of the art reviews the state-of-the-art approaches relevant to this work.

Chapter 5: Methodology describes the methods proposed in this thesis. It includes image segmentation, features extraction using bags of visual words, features extraction using StyleGAN, disease progression marker identification, intra-patient image registration, local tissue transition pathway and outcome risk prediction.

Chapter 6: Experiments and Results show the results of the proposed approaches and the evaluation of these methods.

Chapter 8: Conclusion includes a summary of results as well as ideas for further research directions.

1.4 Publications

Parts of this thesis have been published in the journal European Radiology:

[Unsupervised machine learning identifies predictive progression markers of IPF]

CHAPTER 2

Medical background

2.1 Idiopathic pulmonary fibrosis

Idiopathic Pulmonary Fibrosis (IPF) is a progressive and debilitating lung disease characterized by the formation of scar tissue in the lungs. IPF often also referred as cryptogenic fibrosing alveolitis (CFA), is a condition that worsens with time and is typically an untreatable disease [5] [9]. It is a type of interstitial lung disease (ILD) and is considered to be a complex and poorly understood condition with a significant impact on public health[5]. Diagnosis of IPF can be challenging as it requires the exclusion of other causes of ILD, and it can be challenging to differentiate IPF from other forms of fibrotic lung diseases. The criteria for the diagnosis of IPF have been defined by the American Thoracic Society (ATS) and the European Respiratory Society (ERS) [5] [9] as a combination of clinical, radiological, and pathological findings.

2.2 Epidemiology

The epidemiology of IPF is a rapidly evolving field and the exact incidence and prevalence of IPF are not well known. Incidence refers to the number of new cases of a disease that occur in a given population over a specified time period, while prevalence refers to the number of individuals in a population who have the disease at a given point in time. Studies have shown that IPF is a rare disease, with a reported incidence ranging from 2.5 to 16 cases per 100,000 population per year and a prevalence of approximately 11 to 64 cases per 100,000 population[10]. However, it is believed that these numbers may be underestimations of the true incidence and prevalence of IPF, as the disease is often underdiagnosed or misdiagnosed. In Northern Italy, a study found a higher incidence of IPF compared to other European countries, with an estimated incidence of 15.8 cases per 100,000 population per year [11]. In Italy, a study found an incidence of IPF of 7.2 cases per 100,000 population per year and a prevalence of 45 cases per 100,000 population [12].

It is estimated that the disease affects approximately 128,000 individuals in the United States [10].

Within the non-Hispanic white population, the prevalence is around 85.9%. 60.1% of cases are observed in males whereas only 39.9% are observed in females [13]. Males are almost twice as likely as females to be diagnosed with the illness. It is unknown what accounts for the variation in percentages across racial groupings and between the sexes. After the first diagnosis, the average life expectancy with this fatal condition is between three and five years [14] [15] [16].

Diagnosis of IPF is challenging because its symptoms, such as shortness of breath, coughing, and fatigue, are similar to those of other respiratory diseases. Therefore, it is important to rule out other causes of fibrotic lung disease before making a diagnosis of IPF. The diagnosis is typically made through a combination of physical examination, pulmonary function tests, imaging studies, and biopsy of lung tissue. Smoking is a well-established risk factor for the development of IPF and has been shown to increase the incidence of the disease [13]. Other risk factors for IPF include exposure to environmental toxins, such as asbestos, and a family history of the disease.

IPF is a complex and poorly understood disease that affects a relatively small number of individuals, with a higher incidence in Northern Italy compared to other European countries. Further research is needed to understand the epidemiology of IPF and develop effective strategies for the early detection, diagnosis, and treatment of the disease.

2.3 Etiology of IPF

There is still much ambiguity surrounding the etiology of IPF. Despite the fact that the exact risk factors for this disease are yet to be better understood, a variety of exposures have been shown to be correlated with an increased probability of developing IPF [3].

The correlation between cigarette smoking and idiopathic pulmonary fibrosis has been documented in recent study [17]. Baumgartner et al. conducted a case-control study across many sites and discovered that considerably more individuals in the case group (72%) had a smoking history than in the control group (62%). The odds ratio for patients with a smoking history in the past is 1.60. This study also indicated that smokers who consumed between 21 and 40 packs of cigarettes annually had a 2.3% higher risk of developing IPF [13]. This group of smokers had a hazard ratio of 2.3. In a further study where 225 cases of IPF were examined, a similar finding was reported. Along with an average of four controls per case who were matched to the patient in terms of gender, age, and community, the odds ratio for each smoker is higher at 1.57 [18].

Exposure to metal and wood dust was reported to be related to an increased risk of IPF [18]. A history of exposure to metal and wood dust contributed to the development of odds of 0.67 in IPF patients, but in the control cases, the odds were only 0.50 and 0.46.

Genetic factors have been regularly identified as potential causes, despite the fact that there are no known genetic alterations that are directly associated with IPF cases that spread [19]. Patients may have mutations in surfactant protein C, a hydrophobic protein generated only by type II alveolar epithelial cells (AEC II) [20]. In some circumstances, a common polymorphism in the promoter region of the musin 5B gene expression may play a role in the etiology of pulmonary fibrosis, as reported by Seibold et al.[21]. The name for this polymorphism is MUC5B dysregulation.

The exact causes of IPF are not well understood; however, a number of risk factors have been associated with an increased probability of developing IPF, including smoking, exposure to metal and wood dust, and genetic mutations. Studies have shown that cigarette smoking is associated with an increased risk of IPF, with smokers who consume between 21 and 40 packs of cigarettes annually having a higher risk. Exposure to metal and wood dust has also been linked to an increased risk of IPF. Genetic mutations, such as mutations in surfactant protein C or dysregulation in the musin 5B gene, have also been identified as potential causes, but the role of these genetic factors in IPF is not fully understood.

2.4 Diagnosis of IPF

Idiopathic Pulmonary Fibrosis (IPF) is a lung disease characterized by progressive scarring and thickening of lung tissue, leading to shortness of breath, persistent dry cough, fatigue, weight loss, and clubbing of the fingers. Clinical diagnosis of IPF is based on a patient's symptoms, a thorough medical history, and a physical examination, including a lung function test and imaging studies (such as high-resolution computed tomography (CT) scans). According to a review article by Spagnolo et al. [3], the diagnostic criteria for IPF also involve the exclusion of other potential causes of interstitial lung disease. Additionally, in a study by Burrows and Johnson [22], the authors found that patients with IPF typically experience a gradual onset of symptoms and a decline in lung function over time. This study done between 1955 and 1973 revealed that of 220 patients, 92% exhibited dyspnea, 73% had a cough, and 56.8% produced sputum. The great majority of patients' chests had extensive fibrosing. In overall, 145 cases were associated with malformations of the fingernails or toenails.

The diagnosis of Idiopathic Pulmonary Fibrosis (IPF) is based on a combination of clinical symptoms, pulmonary function tests, biopsy results, and medical imaging analysis. Clinical symptoms include progressive shortness of breath, dry cough, and fatigue.

2.4.1 Laboratory analysis

Since specific and adequate quantification methods are yet to be identified, laboratory analysis for IPF patients is relatively limited. Laboratory testing for pulmonary fibrosis is nearly typically restricted to eliminate other recognizable causes[3].

2.4.2 Physiologic of analysis

The results of pulmonary function tests (PFT) conducted on patients with IPF reveal a restrictive pattern during spirometry and a decreased forced vital capacity [1]. This is associated with increased lung stiffness. Patients have a diminished ability for carbon monoxide diffusion, which is one of the few clinically meaningful signals throughout the disease's early and middle stages. A PFT will be conducted during the examination to identify the severity of the condition and predict the outcome [5].

2.4.3 Lung biopsy

Histological diagnosis of IPF was a crucial tool for determining the presence and progression of the disease. The diagnosis involves an examination of lung tissue samples obtained through a biopsy, either through a surgical procedure or a less invasive transbronchial biopsy. These samples are evaluated under a microscope to identify the characteristic patterns of fibrosis and cellular changes indicative of IPF. However, in recent years, a histological examination has lost its reference as the gold standard for IPF diagnosis, as imaging findings are now sufficient for diagnosis in approximately 50% of IPF patients, providing adequate levels of precision and reliability. This has led to a decrease in the necessity for histologic confirmation in most cases [3].

2.4.4 Medical imaging

Medical imaging plays a crucial role in the diagnosis of IPF [5]. In vivo imaging techniques, such as computed tomography (CT) and chest radiographs (x-rays), can provide valuable information regarding the presence and severity of lung fibrosis in IPF patients. Typically, idiopathic pulmonary fibrosis is diagnosed solely based on clinical or radiological imaging findings. This is due to the fact that CT has a high true positive diagnosis accuracy in diagnosing with only imaging findings. The CT scans of IPF patients typically show the pattern of usual interstitial pneumonia (UIP), which is a distinguishing morphologic characteristic of IPF. The UIP pattern is characterized by a honeycomb-like reticular structure and is often associated with traction bronchiectasis. Nevertheless, in more than 90% of cases, high-resolution CT could be used during assured diagnosis of IPF, as demonstrated in Figure 2.1. However, there are some limitations in diagnosing IPF using CT, including inter-observer variability and difficulties in differentiating IPF from other fibrotic lung diseases [5]. Chest radiographs, on the other hand, are more widely available, but have lower diagnostic accuracy for IPF compared to CT scans. In chest x-rays, IPF patients may exhibit small reticular changes, and asymptomatic individuals may also have similar changes. As a result, chest x-rays may not be able to provide a definitive diagnosis of IPF, and further imaging tests or biopsies may be necessary. In conclusion, while medical imaging plays a crucial role in the diagnosis of IPF, both CT and x-ray have their own advantages and limitations. As such, a comprehensive approach that incorporates multiple diagnostic tools is necessary to achieve an accurate and reliable diagnosis of IPF.

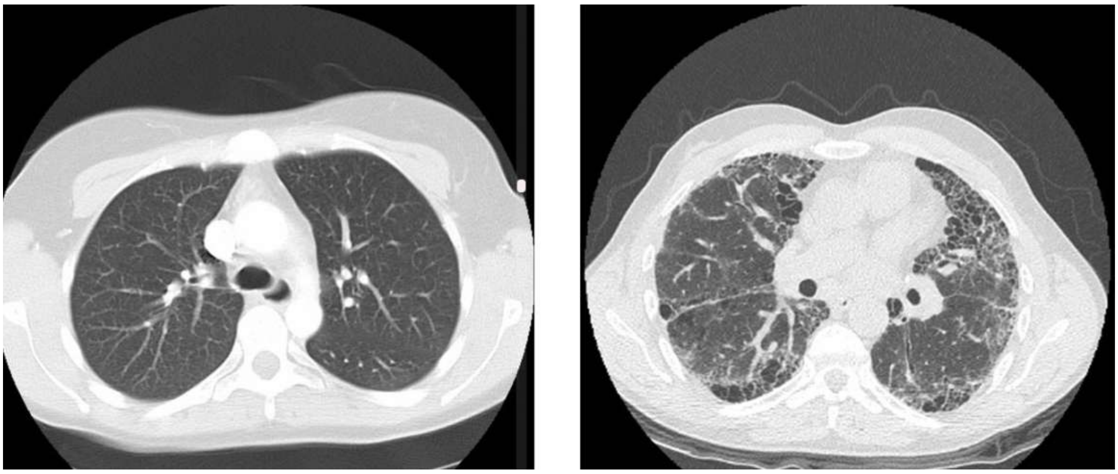


Figure 2.1: The left image shows a high-resolution CT from a healthy person. The right one is a high-resolution CT of an idiopathic pulmonary fibrosis patient. It shows a characteristic combination of predominantly bibasilar reticular abnormalities with associated honeycomb change and traction bronchiectasis.

2.5 Pharmacological Treatment

The most recent evidence-based guidelines from the American Thoracic Society (ATS) GRADE do not include any treatment recommendations for IPF illnesses [5] [23]. The ATS does not recommend the use of all forms of therapies in practice due to insufficient evidence and a lack of sufficient quality reports. There have only been a few medicinal therapies for which there has been received weak approval by FDA and the ATS. [3].

Up until 2014, **pirfenidone** was the only medication permitted for use as a treatment for IPF in nations such as Japan, Europe, Canada, and India. It is a compound containing anti-inflammatory, anti-fibrotic, and antioxidant components [24].

Nintedanib is an inhibitor of intercellular tyrosine kinase, capable of inhibiting both receptor and non-receptor tyrosine kinases [25] [26]. In vitro tests [27] demonstrate that it inhibits receptor tyrosine kinases for vascular endothelial growth factor receptors. G. Keating assessed the efficacy and tolerability of oral nintedanib in IPF patients[28]. In worldwide, randomized, double-blind phase-1 and phase-2 clinical investigations, nintedanib was demonstrated to be superior to placebo in its ability to reduce the pace of forced vital capacity loss. This indicated that the progression of the disease had slowed. On October 15, 2014, the FDA approved the use of Nintedanib as a treatment for idiopathic pulmonary fibrosis.

Computed tomography (CT) imaging is commonly used to monitor the progression of IPF and to guide treatment decisions. CT scans can provide detailed images of lung tissue, allowing physicians to assess the extent of fibrosis and determine the stage of the disease. This information can then be used to make informed decisions about the best course of treatment for the patient [29].

Studies have shown that changes in lung density and fibrosis seen on CT scans are correlated with changes in lung function and clinical outcomes in IPF patients [30]. As a result, CT scans are often used to assess the efficacy of therapeutic interventions and monitor disease progression over time [31]. In this way, CT imaging plays a crucial role in the management of IPF, helping physicians to make informed treatment decisions and track the progression of the disease in individual patients.

CHAPTER 3

Chest Imaging

Medical imaging modalities capture detailed information about the human body and disease. They include Magnetic Resonance Imaging (MRI), Computer Tomography (CT), ultrasound, positron emission tomography (PET), and others. In this thesis, we are primarily concerned with the most widely used 3D modalities for chest imaging: MRI and CT images. The following chapter gives insight into the generation process of the images.

3.1 Chest MRI

Chest MRI is commonly used to detect the following disorders: abnormal lymph nodes, blood vessel problems, thymus tumour, lung masses, oesophagal mass, congenital disabilities of the heart, swollen glands and enlarged lymph nodes in any location of the chest, staging of tumours including invasion of blood vessels, distinguishing between malignant and benign of solitary pulmonary nodules, pulmonary thromboembolic disease, pulmonary hypertension, pneumonia, cystic lung lesions, etc.

In a healthy lung, the tissue density is $0.1g/cm^3$, which is only $\frac{1}{10}$ in comparison with the other soft tissue organs. MRI image quality and signal intensity are indirectly proportional to tissue density. Therefore even under the perfect imaging environment, an MRI image from the lung is still ten times weaker than that from adjacent tissues.

The study from Koyama et al. has shown that non-contrast-enhanced MRI of the lung is as efficient as thin-section CT in distinguishing malignant and benign lung nodules. Even though there is no significant difference between the malignant or benign nodules detection rate, they state the overall detection rate of nodules is lower with MRI images (82.5%) than CT images (97.0%)[32]. Figure 3.1 and 3.2 can visually approve that the image quality is better in a CT image than in MRI images.

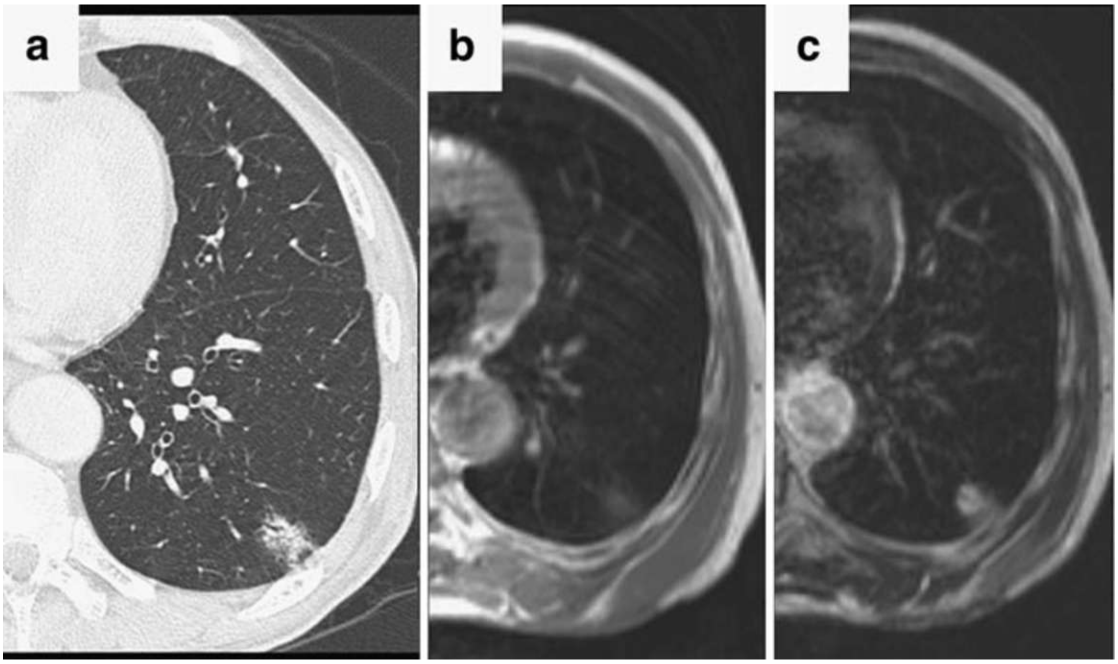


Figure 3.1: A patient with lung cancer in the left lower lobe. a: Thin-section CT shows a partly solid nodule with a diameter of 25.5 mm in the left lower lobe. b: T1-weighted shows low signal intensity for the nodule. c: T2-weighted shows high signal intensity for the nodule. [32]

Typical MR imaging artifacts[33] are radio frequency noise that causes an image's non-uniform, washed-out appearance and can be seen in (Figure 3.3A). Susceptibility artefacts Figure 3.3B) share distortions or local signal change. Zipper artifacts (Figure 3.3C) are a type of artifact where one or more spurious bands of electronic noise extend across the image. Motion artefacts (Figure 3.3D) can occur during scans and result from tissue/fluid movement. Aliasing on MRI occurs when the field of view is smaller than the imaged body part, as seen in Figure 3.3E. Gibbs Ringing typically appears as multiple fine parallel lines immediately adjacent to high-contrast interfaces and can be seen in Figure 3.3F.

MR imaging systems have several primary benefits, including excellent soft tissue imaging capability, a very high resolution of about 1mm cubic voxels, and a good signal-to-noise ratio. Additionally, MR imaging allows for the acquisition of multi-channel images with variable contrast using different pulse sequences, which can be used for segmenting and classifying different structures.

However, MR imaging also has some disadvantages. The acquisition time for MR imaging is significantly longer than that of CT imaging. Additionally, obtaining uniform image quality in MR imaging can be more challenging. MRI for the lung has limitations due to low proton density and fast signal decay from artefacts and air-tissue interfaces. The

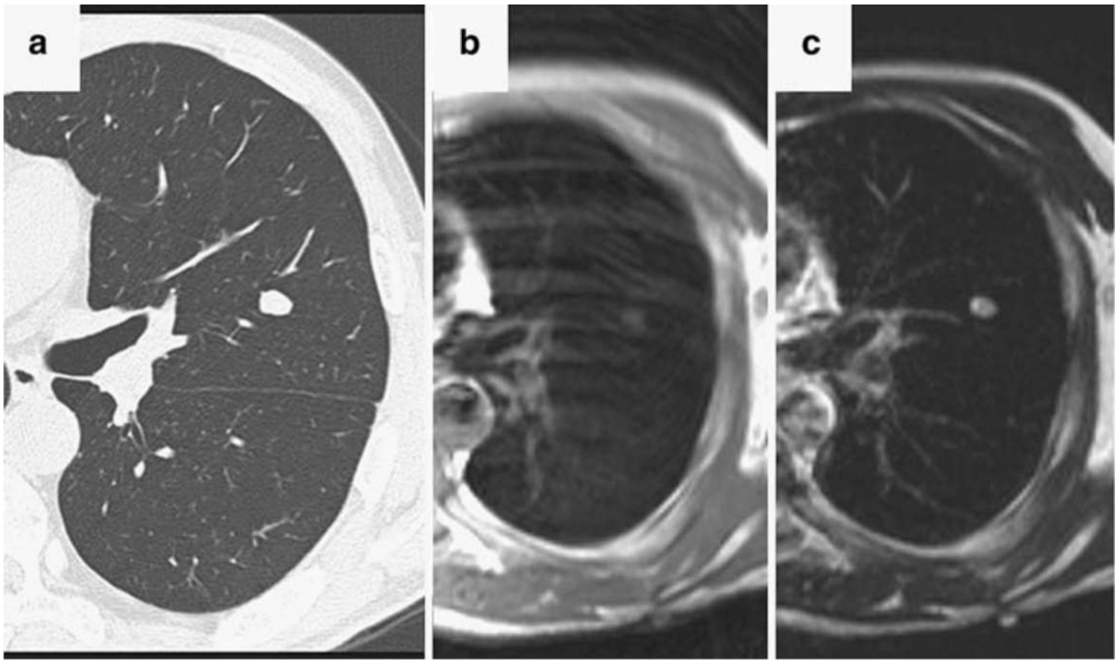


Figure 3.2: A patient with a pathologically unidentified benign nodule in the left upper lobe. a: Thin-section CT shows a solid nodule with a diameter of 9.8 mm in the left upper lobe. b: T1-weighted shows low signal intensity for the nodule. c: T2-weighted shows high signal intensity for the nodule. [32]

lengthy exam time can lead to reduced diagnostic accuracy for patients unable to stay still, additionally the 8mm thickness chosen for MRI in the study [35][36] may not adequately evaluate interstitial changes, as recommended by guidelines and other studies. A thicker slice may fail to detect early signs of fibrosis. Therefore CT scans provide clearer images of lung tissue over MRI, making it easier to identify specific features, such as nodules, which are important in the diagnosis of IPF.

3.2 Chest CT

Computed Tomography (CT) is a diagnostic imaging technique that uses X-rays to produce detailed 3D-reconstructions of internal organs and tissues. CT works by rotating an X-ray source around the patient's body and measuring the amount of X-rays that pass through the tissue. These measurements are then used to generate a series of 2D cross-sectional images of the body, which can be combined to form a 3D-reconstruction.

The CT scan captures multiple X-ray images of the body at different angles. The data collected from these images are processed using specialized algorithms to produce the final 3D image. This image can then be viewed and analyzed by a radiologist to identify any abnormalities or diseases.

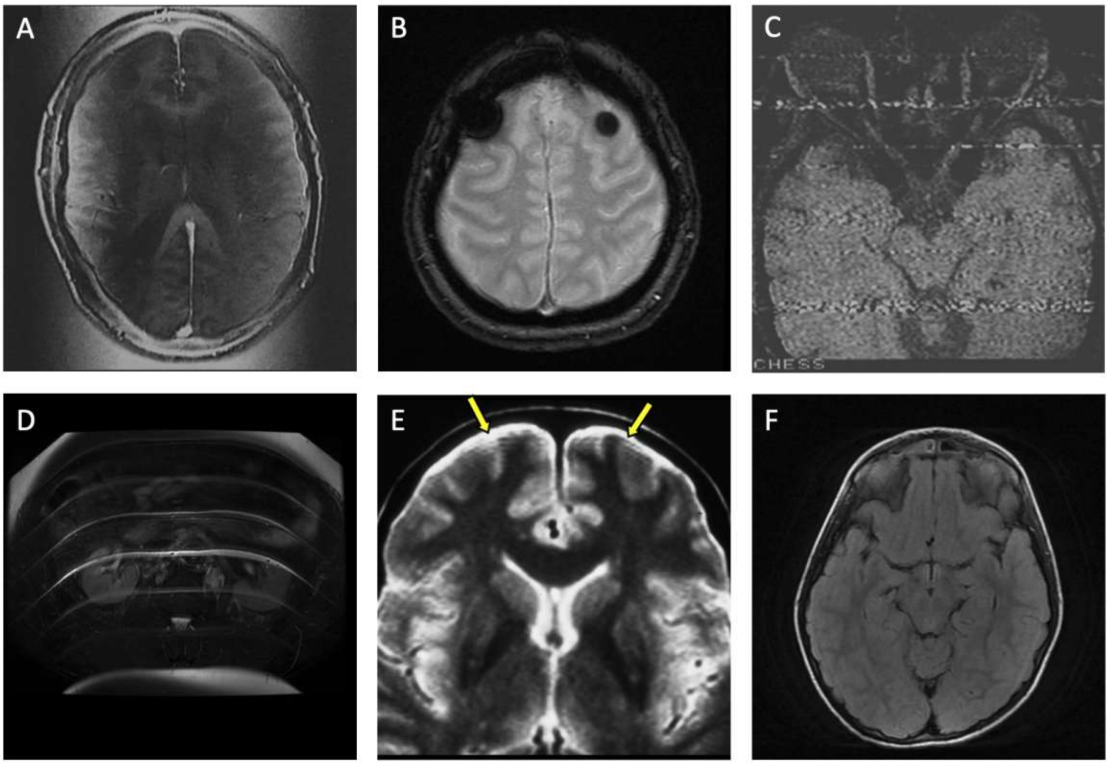


Figure 3.3: Examples of MRI artifacts. A: RF noise, B: Susceptibility artifact, C: Zipper artifacts, D: Motion, E: Aliasing on MRI, and F: Gibbs Ringing.[34]

In CT, exposure to ionizing radiation can increase the risk of developing cancer. However, the benefits of CT scans in diagnosing life-threatening diseases often outweigh the potential risks. CT scans are a non-invasive diagnostic tool that can provide detailed images of internal organs and tissues, making it an essential tool for the diagnosis and treatment of many medical conditions.

High-resolution CT of the lungs has been the tool of choice during the past four decades for evaluating whether or not a patient has a diffuse pulmonary parenchymal abnormality. By combining a large number of two-dimensional chest x-rays with a measurement range of one to two millimeters, high-resolution CT is designed to provide images with great spatial lung information. In 1975, a radiologic-pathologic correlative examination of postmortem lungs was the first application of this technique. In 1978, Itoh et al.[37] published their results about the correlations of minute lung nodules. They concentrated their attention on peribronchiolar. Todo et al.[38] from Kyoto University were the first to disclose high-resolution CT for diagnosing diffuse lung illness in 1982. Their article, which was published in the *Japanese Journal of Clinical Imaging*, examined the use of CT scans on 21 patients with diffuse panbronchiolitis, lymphangitic cancer spreads, sarcoidosis, or TB. All of the individuals were diagnosed with one of these illnesses. Very carefully and thoroughly, this investigation demonstrated the link between the image abnormalities and the inflation-fixed lung specimens. High-resolution CT scans have been utilized to establish a radiologic-pathologic correlation relationship between abnormalities and the architecture of the secondary pulmonary lobule ever since they were established almost fifty years ago.

3.2.1 CT: How it is used for diagnosing ILD/IPF

CT scans are commonly used to diagnose ILD/IPF, as they provide a clear and precise visual representation of the lung tissue, which is essential in identifying characteristic features of IPF. In a study by Raghu et al.[5], it was reported that CT scans are an essential tool in the diagnosis of IPF, with a high level of accuracy, and are considered the gold standard for the assessment of interstitial lung disease.

One of the key advantages of CT scans over other imaging techniques is the ability to produce a 3D reconstruction of the lung tissue, which allows the physician to view the internal structures from multiple angles, providing a more comprehensive understanding of the condition. In the diagnosis of IPF, CT scans are used to evaluate the lung parenchyma for the presence of specific patterns of fibrosis, such as the reticular pattern and honeycombing. The reticular pattern is characterized by the thickening of the interlobular septa, which are the fibrous tissue dividers between lung lobes, leading to a "honeycomb" or "net-like" appearance in the lung parenchyma. This pattern is considered a hallmark of IPF and can help differentiate IPF from other interstitial lung diseases. Honeycombing refers to the characteristic, irregular cystic spaces that are formed in the lung tissue as a result of fibrotic tissue growth. Another feature that is commonly seen on CT scans in IPF is traction bronchiectasis, which is characterized by the thickening of the bronchial walls and widening of the bronchial lumen due to the pull of fibrotic tissue. Those abnormalities usually involve the secondary pulmonary lobule of the lung and can be seen in Figure 3.4. Those patterns are often associated with decreased lung attenuation or air-filled lesions.[39] In addition to identifying these key features, CT scans can also be used to assess the extent and severity of lung involvement in IPF. For example, the CT scan can be used to determine the proportion of lung tissue affected, the thickness of the fibrotic tissue, and the degree of lung volume loss, which are all critical indicators in the diagnosis and management of IPF.

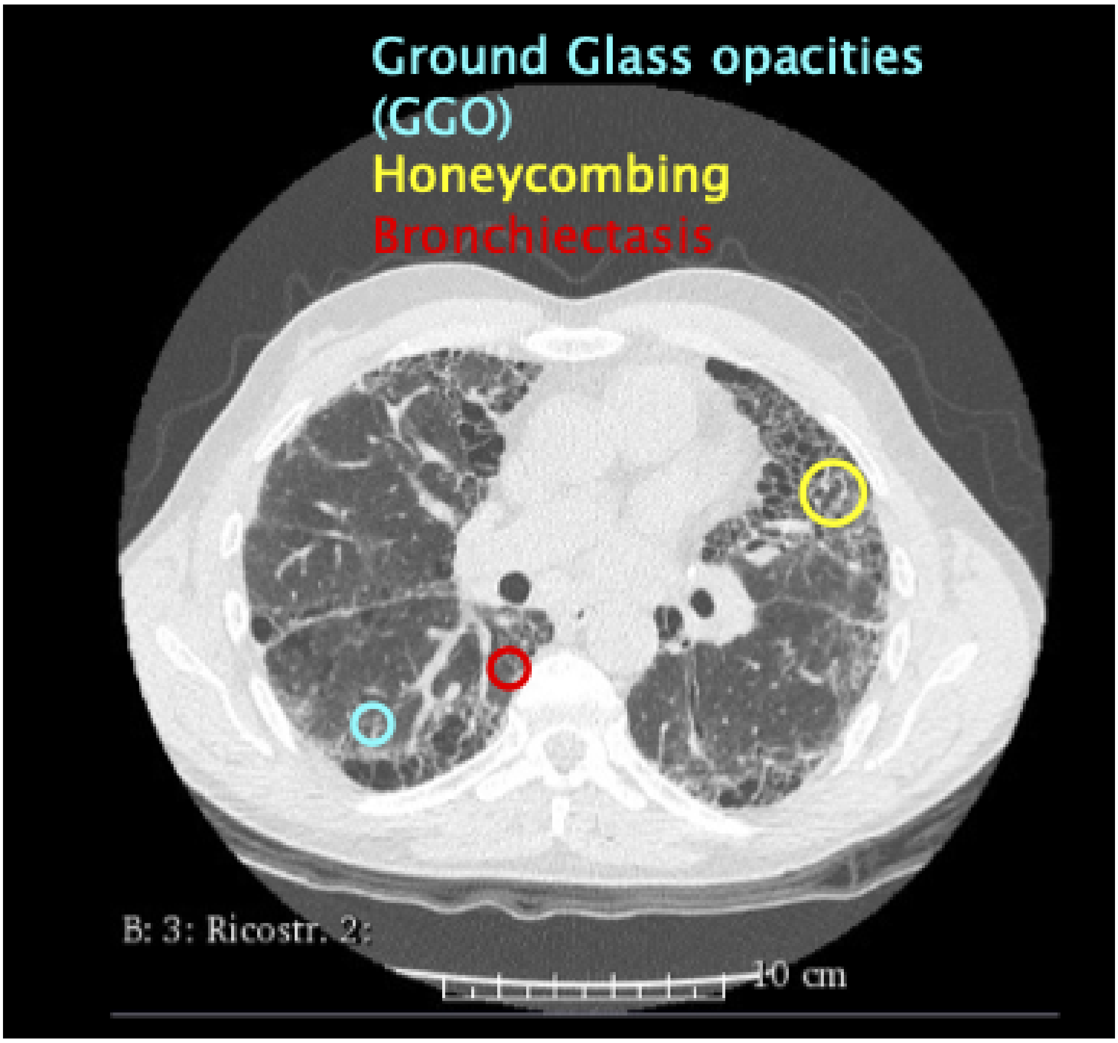


Figure 3.4: Low Attenuation pattern. Most diseases with a low attenuation pattern can be distinguished on the basis of HRCT scans.

CT scans play a crucial role in the diagnosis of ILD/IPF, providing a detailed, clear, and precise image of the lung tissue, which is essential in identifying the hallmark features of the disease. With the ability to produce 3D reconstructions of the lung tissue and assess the extent and severity of lung involvement, CT scans are a valuable tool for the diagnosis and management of IPF.

3.2.2 Artifacts in CT imaging

The majority of high-resolution CTs are performed with the patient supine. When the lung anomaly is widespread in the distribution or severe in profusion, inspiratory pictures are usually adequate [40]. Inspiratory pictures are often performed during complete inspiration, which is easier for the majority of patients to regulate. A notable exception is dynamic and ultrafast CT scans with an electron beam CT scanner, which may be used to monitor the respiratory cycle. End-exhalation is accompanied by a reduction in lung size. During expiration, the posterior membranous wall of the trachea looks concave, in contrast to its convex appearance during inspiration (Figure 3.5 and Figure 3.6). Expiratory pictures may be especially useful for distinguishing the etiology of mosaic attenuation from airway illness, vascular disease, and infiltrative lung disease [41][42].

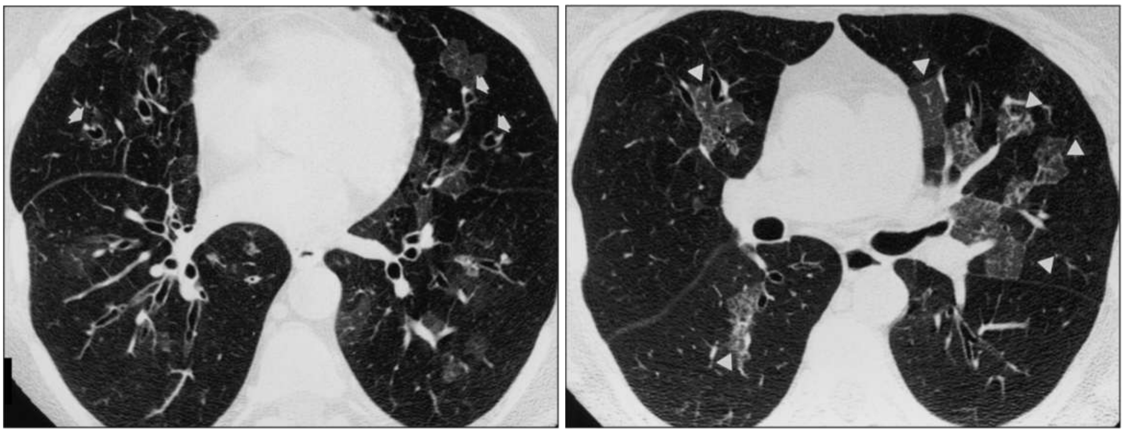


Figure 3.5: A female patient with idiopathic bronchiolitis obliterans. **(Left)** Inspiratory high-resolution CT scan shows diffuse cylindric bronchiectasis, with bronchi larger than adjacent arteries; signet ring sign of bronchiectasis (arrows); and subtle mosaic attenuation. All are findings of small airway disease. **(Right)** Expiratory high-resolution CT scan at the same anatomic level as left image reveals that the expected decrease in lung size is absent, and lungs remain low in attenuation, indicating severe diffuse air trapping, with only normal lung parenchyma found as a few individual secondary pulmonary lobules that increased in attenuation (arrowheads).[40]

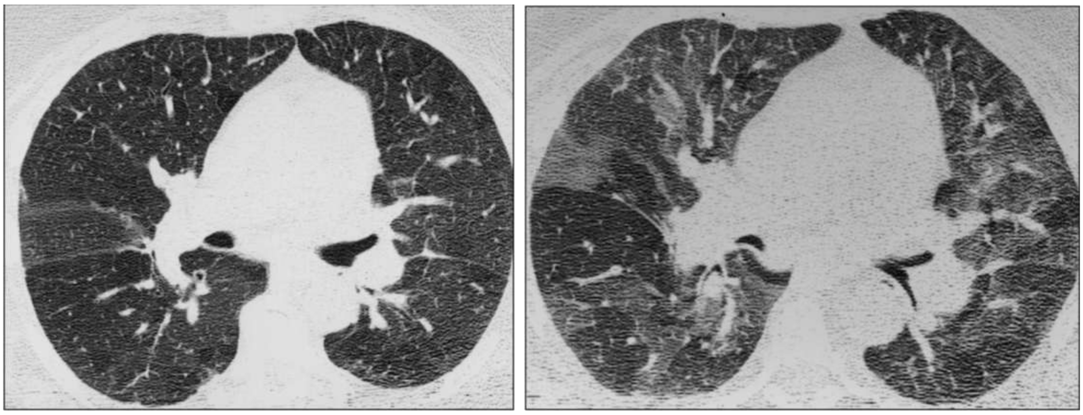


Figure 3.6: A female patient with hypersensitivity pneumonitis. **(Left)** Inspiratory high-resolution CT scan shows a few scattered thickened interlobular septa and a very faint pattern of mosaic attenuation. **(Right)** Expiratory high-resolution CT scan at the same anatomic level as left image reveals multifocal bilateral air trapping represented by low-attenuation lung parenchyma. High-attenuation areas represent normal lung that has developed atelectasis with expiration. Note internal bowing of posterior wall of bronchus intermedius as evidence that scan was taken at expiration.[40]

CT imaging has several artifacts [43] that may appear in the images, such as streak artifacts (Figure 3.7A) that occur when the object of interest is moved during the scanning process. Motion artifacts (Figure 3.7B) are seen when the boundaries of the object are ill-defined and can result in a blurred appearance. Beam hardening artifacts (Figure 3.7C) are due to the nonlinear nature of the x-ray beam and can cause an unnatural appearance in the images. Ring artifacts (Figure 3.7D and E) may be influenced by gain variations, radiation damage to the detector, or irregularities in linearity. Bloom artifacts (Figure 3.7F) result from partial-volume effects or high-density structures, appearing as a bright halo around the object of interest.

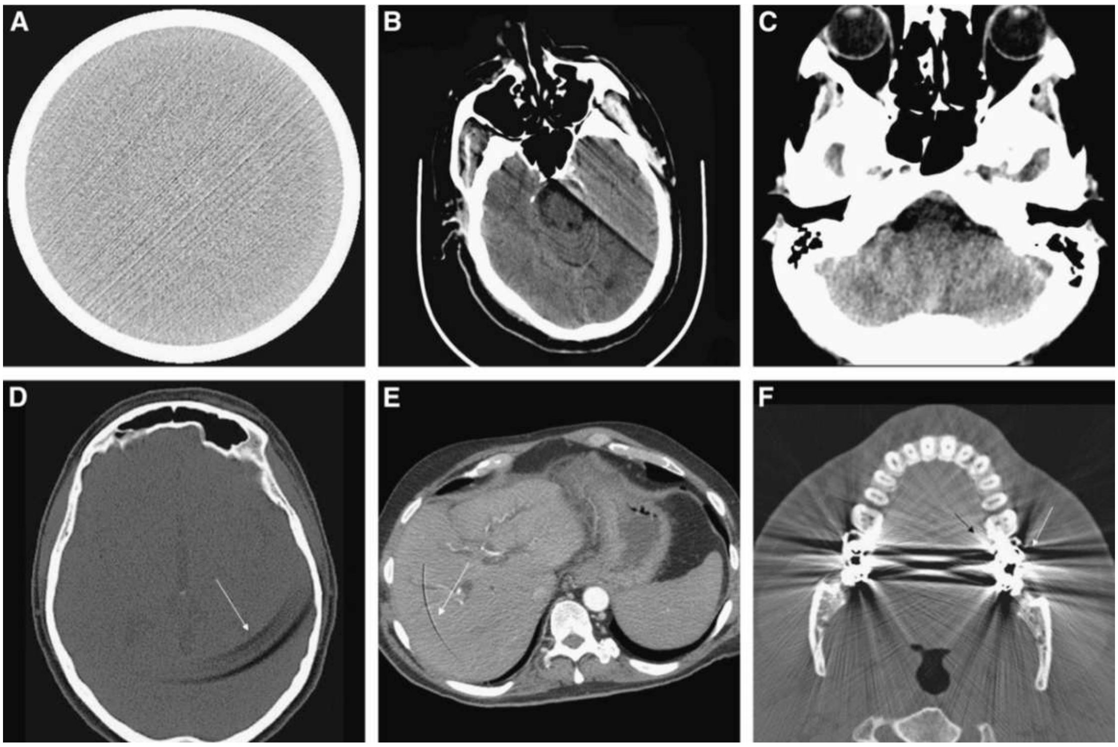


Figure 3.7: Examples of CT artifacts, streak artifacts (A), motion artifact (B), beam-hardening (C), ring artifacts (D and E), and bloom artifacts (F).[44]

CT imaging has several advantages that make it a commonly used diagnostic tool in radiology. These advantages include its cost-effectiveness and accessibility, as well as its high spatial resolution due to the implementation of multi-slice scanners. Furthermore, the short scan duration of CT imaging is a convenient aspect for patients. In terms of imaging sensitivity, CT scans have demonstrated greater detection abilities for subarachnoid hemorrhages compared to MRI, as well as superior abilities for detecting cerebral calcifications.

However, the CT imaging system also has some limitations that must be considered. The most notable disadvantages include the lower soft tissue contrast compared to MRI, which is a result of its X-ray-based imaging method, and the exposure of patients to radiation. Despite these limitations, the advantages of CT imaging have allowed it to remain a widely used diagnostic tool for the examination of the brain, liver, and thorax.

3.2.3 Reconstruction kernel

Computed Tomography (CT) image reconstruction is a statistical process that involves transforming X-ray projection data collected from multiple angles into images. The objective of this process is to produce images that are free of noise, maintain spatial resolution, and have a correct representation of the underlying anatomy while being obtained with the least possible radiation exposure.

There are two main categories of CT image reconstruction techniques, namely Analytical Reconstruction and Iterative Reconstruction. Filtered Back Projection (FBP) is the most widely used analytical reconstruction method. In FBP, a one-dimensional filter is applied to the projection data prior to back projection onto the image space. FBP is favored for its computational efficiency and numerical stability.

Different reconstruction kernels are utilized in CT image reconstruction to enhance the quality of images. The selection of the reconstruction kernel is a crucial step in the process, and must be tailored to the specific clinical application. For instance, sharper kernels are typically employed for examinations of skeletal structures, while smoother kernels are often utilized in brain scans and liver tumor evaluations to reduce noise and improve low-contrast detection.

Slice thickness is another key aspect of CT image reconstruction that has a significant impact on the trade-off between resolution, noise, and radiation dose. The CT user must determine the optimal combination of reconstruction kernel and slice thickness to minimize radiation exposure while preserving image quality. Higher spatial resolution can be obtained by increasing slice thickness [45], but it comes at the cost of increased image noise (Figure 3.8).

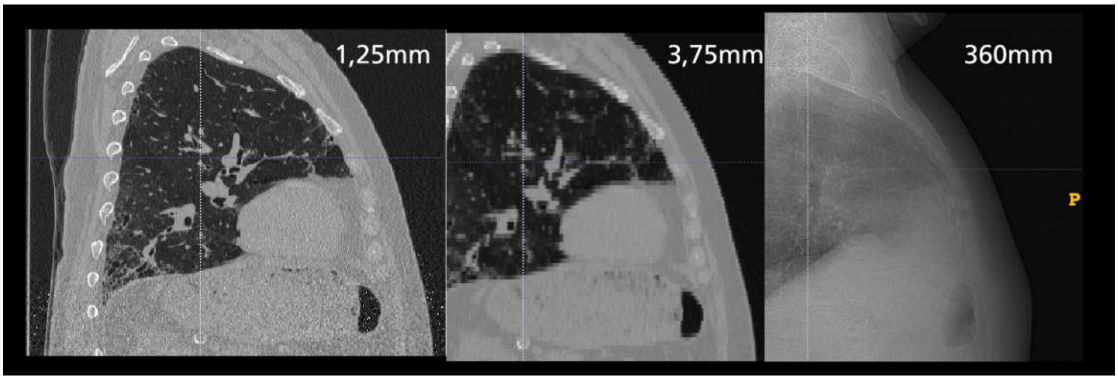


Figure 3.8: CT chest image (sagittal view) with different CT slice thickness.

In our data set, the CT scanner manufacturer provides five types of reconstruction kernels: soft (figure 3.9 A), standard (figure 3.9 B, boneplus (figure 3.9 C), lung (figure 3.9 D, and bone (figure 3.9 E). These kernels can be selected based on specific clinical requirements. Additionally, commercial CT scanners and third-party solutions offer noise reduction algorithms to minimize background noise. Although these algorithms can effectively reduce noise while preserving high-contrast resolution, their diagnostic performance should be thoroughly evaluated before widespread deployment in clinical settings.

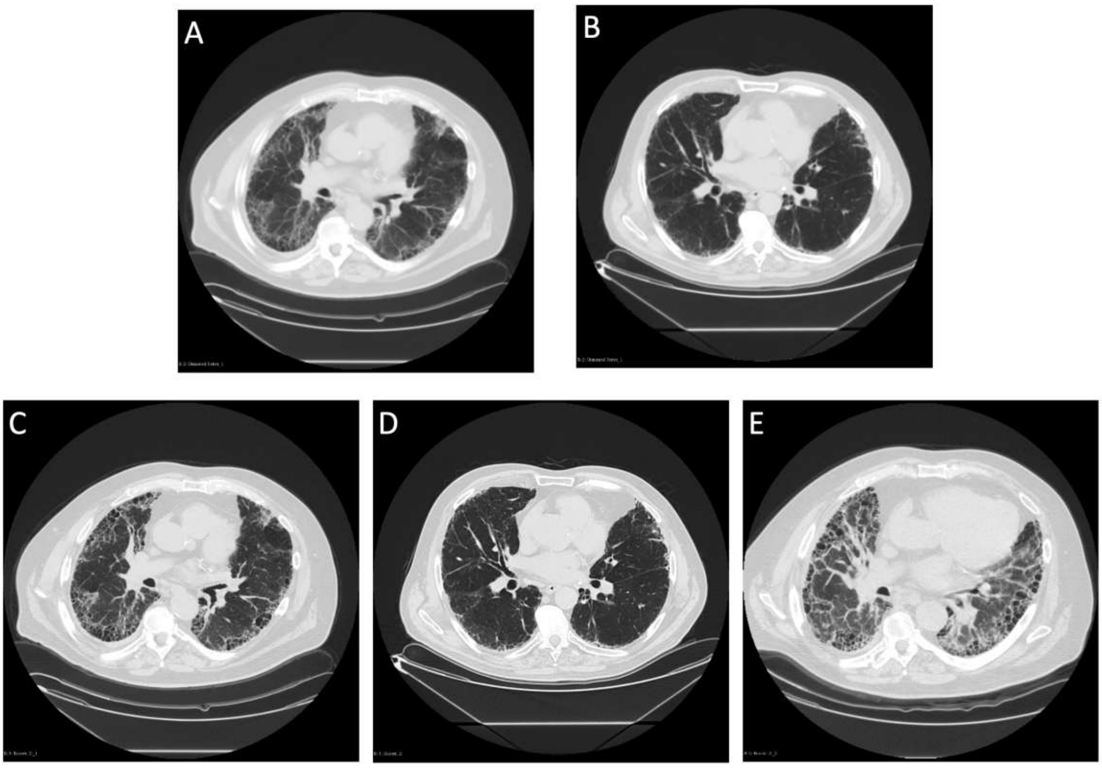


Figure 3.9: Examples of different reconstruction kernels. SOFT kernel (A), STANDARD kernel (B), BONEPLUS kernel (C), LUNG kernel (D) and BONE kernel (E).

The reconstruction kernel plays a crucial role in diagnosing IPF as it determines the quality of images produced by a CT scanner. The selection of a reconstruction kernel has a major impact on the final image quality, as a smooth kernel will produce images with lower noise and smoother tissue boundaries but may also result in the loss of important details, such as small fibrotic changes in the lungs. On the other hand, a sharper kernel may produce images with greater details but may also introduce more noise and artifacts.

Therefore, when diagnosing IPF, it is essential to choose the most appropriate reconstruction kernel to ensure that the CT images produced are of sufficient quality to accurately detect and diagnose the disease. This requires a thorough evaluation of the diagnostic performance of different reconstruction kernels, taking into consideration the trade-off between image quality and radiation exposure.

Ultimately, the selection of the reconstruction kernel will depend on several factors, including the type and stage of the disease, the specific imaging requirements, and the clinical context in which the images will be used.

3.3 Discussion

This chapter provides a comprehensive examination of the underlying principles and technological aspects of chest imaging modalities, as well as their respective advantages and limitations. The two primary chest imaging modalities, which form the basis for radiologists' daily clinical examinations, are thoroughly examined. Despite the widespread use of high-resolution computed tomography (HRCT) as a diagnostic tool for diffuse lung diseases for over two decades, the interpretation of HRCT images remains a challenge for many radiologists. To address this issue, medical education courses continue to be in high demand, as evidenced by the large attendance at events such as the annual meetings of the European Radiological Society and the Radiology Society of North America. It is noteworthy that the distinction between HRCT patterns of various interstitial lung disorders remains a challenging task.

CHAPTER 4

State of the Art

Section 4.1 describes the fundamental concepts and types of machine learning, a subfield of artificial intelligence that enables computers to learn from data and make predictions without human intervention. It covers the three main types of machine learning methods: supervised, unsupervised, and reinforcement learning and explains their respective applications and algorithms. Section 4.2 provides an introduction to regression, a type of supervised machine learning algorithm that is used to determine the statistical relationship between a dependent variable and one or more independent variables. Principal component analysis is introduced in section 4.3, a technique for reducing the dimensions of a data set by identifying the principal components that account for the bulk of the variance in the data. Section 4.4 explains the objective of clustering, a machine learning technique for grouping similar items together, and introduces k-means clustering, one of the most widely used clustering methods that involve an iterative procedure to minimize the sum of squared errors within clusters. Random Forest is a machine learning algorithm that uses a combination of decision trees to make predictions and is widely used for image classification, prediction, and feature selection tasks. It is introduced in section 4.5. A hand-crafted feature extraction technique used in computer vision and image processing, Bag of Visual Words, is presented in section 4.6. In the context of CT scans, local features such as SIFT or SURF descriptors are extracted to create a visual vocabulary, which is then quantized into visual words using a clustering algorithm. Other feature extraction techniques, such as Haralick features, can also be used in BoVW to describe the gray-level co-occurrence matrix of an image. Section 4.7 provides an overview of image segmentation in medical imaging and computer vision, which involves dividing an image into multiple regions based on shared characteristics. The focus is on lung segmentation, which entails identifying the lung regions in medical images such as CT scans and presents the challenges of accurately distinguishing lung tissue from other structures in the image. Over-segmentation, outlined in section 4.8, is a technique used to divide an image into smaller, spatially coherent regions (supervoxels) that correspond to meaningful objects or structures in the image for more efficient processing, improved accuracy, and reduced noise. As explained in section 4.9, Deep Learning refers to the use of neural networks with multiple layers and has roots dating back to the 1940s, but only gained popularity in the early 2000s due to the availability of more powerful hardware, large labelled datasets, and new training algorithms. Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) are two popular types of deep learning neural networks used for tasks such as image classification and generative tasks respectively. There are several quantitative texture analysis tools, explained in section 4.10, that diagnose diffuse lung diseases, including DTA and CALIPER, which are the main state-of-the-art methods using machine learning in the context of IPF and combine supervised and unsupervised machine learning algorithms. The Kaplan-Meier analysis (section 4.11) is a method for estimating the survival function of a population over time by determining the probability that an individual in the population has not experienced an event of interest after a certain time period. A brief summary of the state-of-the-art algorithms is given in Section 4.12.

4.1 Fundamental of Machine Learning

Machine learning (ML) is a subfield of artificial intelligence that deals with the development of algorithms and models that enable computers to learn and improve their performance without being explicitly programmed. It involves the development of algorithms and models that enable computers to learn from their experiences and enhance their performance.

In the context of machine learning, the term "experience" refers to the data and knowledge that is provided to the learning algorithms. The more data and knowledge that is available, the better the algorithms can estimate the outcome of a given task [46]. However, it is important to note that the outcome is not always an exact calculation.

Machine learning algorithms do not begin with a pre-defined system. Instead, they discover patterns and relationships within the provided data sets. The fundamental concepts of ML include representation, generalization, and optimization. Representation refers to the way data and knowledge are presented to the learning algorithm, such as feature extraction and feature engineering. Generalization refers to the ability of the algorithm to make accurate predictions on new, unseen data. Optimization refers to the process of finding the best parameters for the algorithm, such as minimizing the error or maximizing the performance.

There are three main types of machine learning methods: supervised, unsupervised, and reinforcement learning.

Supervised learning is the most common type of machine learning, where the algorithm is trained on a labeled dataset, where the correct output is provided for each input. The algorithm learns to map inputs to outputs, and can be used for tasks such as classification and regression. Examples of supervised learning algorithms include decision trees, k-nearest neighbors and logistic regression.

Unsupervised learning is where the algorithm is trained on an unlabeled dataset, where the correct output is not provided. The algorithm must discover the underlying structure of the data. This type of learning is used for tasks such as clustering, anomaly detection and dimensionality reduction. Examples of unsupervised learning algorithms include k-means and hierarchical clustering, and principal component analysis.

Reinforcement learning (RL) is a type of machine learning where an agent learns to make decisions by interacting with an environment and receiving feedback in the form of rewards or penalties. RL can be used for tasks such as game-playing, robotics, and decision-making.

In summary, machine learning is a field that deals with the development of algorithms and models that enable computers to learn from data, identify patterns and make predictions or decisions without human intervention. The fundamental concepts of ML include representation, generalization and optimization. There are three main types of machine learning methods: supervised, unsupervised and reinforcement learning. Each type of learning is used for specific tasks, and different algorithms can be used for each type.

4.1.1 Supervised Learning

The most important thing is that supervised learning uses labeled dataset for training, where each input is given with a correct output [47]. Supervised learning aims to learn a mapping function that can be used to make predictions or decisions on new, unseen data. This mapping function is often represented as a model, which can be represented as a mathematical equation or a neural network. Typically, datasets for supervised learning contain sets of input-output pairs, where the input is a set of features, and the output is the corresponding label or class. A chosen model is used to learn the mapping function from the input to the output. There are a wide variety of models that can be used in supervised learning, including linear regression, decision trees, k-nearest neighbors, and neural networks. The choice of model will depend on the specific problem and the characteristics of the data.

On the other hand, unsupervised learning, as defined by Alpaydin (2010), is a type of machine learning where the algorithm is trained on an unlabeled dataset, where the correct output is not provided. The goal of unsupervised learning is to discover the underlying structure of the data, such as identifying patterns, grouping similar data points together or detecting anomalies. Unsupervised learning can be used for tasks such as clustering, anomaly detection, and dimensionality reduction. The model is not provided with any specific output, it is up to the model to identify patterns and structure in the data. Examples of unsupervised learning algorithms include k-means, hierarchical clustering, and principal component analysis.

4.1.2 Unsupervised Learning

Unsupervised machine learning is a type of machine learning when the outputs are unknown. It puts its model through training with unpredictable outcomes. These models search for patterns and categorize the data within the supplied sets into groups. When it comes to unsupervised learning, we have no means of knowing if the outcomes are correct or erroneous. These models are capable of identifying pattern correlations. An unsupervised model in the field of bioinformatics, for instance, may establish the relationship and closeness between several gene symbols. Still, it cannot tell which character is the cause of a particular disease. Humans do not yet know which gene mutations cause IPF. Using unsupervised machine learning techniques, it is feasible that physicians may discover that essential gene. Clustering is one of the most often employed techniques in unsupervised learning.

Unsupervised learning, is where the algorithm is trained on an unlabeled dataset, where the correct output is not provided. The algorithm must discover the underlying structure of the data. This type of learning is used for tasks such as clustering, anomaly detection and dimensionality reduction. Examples of unsupervised learning algorithms include k-means and hierarchical clustering, and principal component analysis.

4.1.3 Reinforcement Learning

Reinforcement learning is a type of machine learning where an agent learns to make decisions by interacting with an environment. The agent receives rewards or penalties based on its actions, and the goal is to learn a policy that maximizes the cumulative reward over time. Reinforcement learning can be used for tasks such as game playing, robotics, and recommendation systems. The agent is not provided with any specific output, it learns through trial and error and updates its policy based on the rewards or penalties it receives. Examples of reinforcement learning algorithms include Q-learning, SARSA and DDPG.

4.2 Linear regression

Linear regression, an algorithm of supervised ML, determines the statistical relationship between two or more variables. This technique is used in many ML modelling and analyses for identifying the correlation between a dependent variable y and one or more independent variables x. A regression model relates a dependent variable to a function of independent variables, and unknown parameters β can be written as:

$$y \approx f(x,\ \beta) \tag{4.1}$$

The type of regression employed is determined by the examined function. In linear regression, f represents the linear function. Simple linear regression is a kind of linear regression that only evaluates a single feature. Multiple linear regression, as opposed to fundamental linear regression, contains several distinctive properties. It is the most often employed statistical and machine learning method for predicting the dependent variable y based on the independent variable x.

4.3 Principal Component Analysis

The premise underlying principal component analysis (PCA), commonly abbreviated as PCA, is the reduction of the dimensions of a data set into k main components. A vector v_1 with the most significant variance in the linear function $v_1^T x$ represents the principal component of a data vector x containing p variables. This is the definition of a principal constituent. This method is done j times, where j can vary from 1 to n, for v_2 when it is uncorrelated with v_1. The purpose of this analysis is to discover k components that may account for the bulk of the variance in x. The solution to the eigenvalue problem will result in the identification of the principal components, indicated by v_j

$$(C - \lambda I_n)v_j = 0 \tag{4.2}$$

C is the covariance matrix $C_{ij} = cov(x_i, x_j)$, and I_n is the identity matrix (nxn). Consequently, it is an eigenvalue of C, and v_j is the corresponding eigenvector.

Principal component analysis (PCA) is a powerful technique for dimensional reduction for analyzing medical data, such as chest X-Ray analysis of lung cancer, classification of the pulmonary lesion, and classification of malignancy degree for lung cancer [48][49][50]. The principal component analysis is not restricted limited to lung disease analysis. For neurology, PCA can be used to detect brain signal spikes and action potentials [51][52]. For non-medical utility, it is being applied for facial identification and face analysis [53][54].

4.4 K-mean clustering

The objective of clustering is to group things that are linked to one another but separate from other groups. The aim is to identify a set of n in k groups by collecting n data points in $\mathbb{R}$ d-dimensional space and an integer k. Among the various clustering approaches, k-means clustering is one of the most widely used and studied. K-means is a clustering method that requires an initial set of cluster centers. The number of cluster centers is fixed and pre-defined. The k-means method is an iterative procedure that minimizes the sum of squared errors within clusters [55]. The centers have the smallest mean squared distance between each of the n locations and the center that is closest to them. The algorithm can be mathematically represented as follows:

Initialize K centroids $\mu_1, \mu_2, ..., \mu_K$ randomly from the data points. Assign each data point x_i to the cluster whose centroid it is closest to. This can be represented mathematically as:

$$c_i = \operatorname*{argmin}_{j=1}^{K} |x_i - \mu_j|^2 \tag{4.3}$$

where c_i is the cluster assignment for data point i and $|\cdot|^2$ is the squared Euclidean distance. Update the centroids by taking the mean of all data points assigned to each cluster. This can be represented mathematically as: $\mu_j = \frac{1}{|S_j|} \sum_{i \in S_j} x_i$ where S_j is the set of data points assigned to cluster j and $|S_j|$ is the number of data points in that cluster. Repeat steps 2 and 3 until the cluster assignments no longer change or a stopping criterion is met. The K-means algorithm is sensitive to the initial centroid and the final clusters are dependent on the initial centroids. To counter this problem, K-means is usually run multiple times with different initial centroids and the final clusters are chosen based on the lowest sum of squared distance of points from their respective cluster centroids.

K-means is a widely used algorithm in various fields, such as image compression, image segmentation, speech recognition, and market research.

4.5 Random forest

Random Forest(RF) is a machine learning algorithm that is widely used in image classification, prediction and feature selection tasks. It is a type of ensemble learning method that creates multiple decision trees, which are combined to form a forest of decision trees. Random Forest is a non-parametric method and is flexible to handle complex data structures.

The basic idea behind RF is to use a combination of decision trees to make predictions. Each decision tree in the forest is trained using a random subset of the input data and features. This leads to diversity among decision trees, as each tree is trained on different data and features. The final prediction is made by aggregating the predictions of all decision trees, which can be done by taking a majority vote or averaging the predictions.

Mathematically, let's consider the input data to be a matrix X with m samples and n features. The output target variable Y is a vector with m values. For each tree, we first randomly select a subset of samples with replacement, called bootstrapped samples, and denote this subset as X_{boot}. Then, we randomly select a subset of features for each node in the tree and build the decision tree using X_{boot} and Y. This process is repeated for a specified number of trees, say T.

The prediction for a new sample X_{new} is made by aggregating the predictions of all T decision trees. Let's denote the prediction of the i-th decision tree as $f_i(X_new)$. The final prediction is given by:

$$f(X_{new}) = \frac{1}{T} \sum_{i=1}^{T} f_i(X_{new}) \quad \text{(for regression problems)} \tag{4.4}$$

or

$$f(X_{new}) = majority_{vote}(f_1(X_{new}), f_2(X_{new}), ..., f_T(X_{new})) \quad \text{(for classification problems)} \tag{4.5}$$

RF has been successfully applied to medical image analysis, especially in the context of disease diagnosis and prognosis. In medical imaging, it has been used for image classification, such as classifying CT images into normal or abnormal, and segmentation, such as segmenting lesions in MR images.

One of the benefits of RF in medical imaging is that it can handle large amounts of data, including images, which can be high-dimensional. Additionally, Random Forest is robust to noisy and missing data, which is common in medical imaging, and can handle complex relationships between features.

RF is a powerful machine learning algorithm that has been successfully applied to medical image analysis. Its ability to handle high-dimensional data, robustness to noise and missing data, and flexibility to handle complex relationships between features make it a popular choice for medical imaging tasks.

4.6 Bag of Visual Words

Bag of Visual Words (BoVW) [56] is a feature extraction technique commonly used in computer vision and image processing. It involves creating a visual vocabulary from a set of training images and then representing each image in the dataset as a histogram of visual words. In the context of CT scans, the visual vocabulary is created by extracting local features from the images, such as Scale-Invariant Feature Transform (SIFT) or Speeded Up Robust Features (SURF) descriptors. These descriptors capture the distinctive local characteristics of the image, such as shape, orientation, and texture.

In the context of CT scans, the visual vocabulary is created by extracting local features from the images, such as Scale-Invariant Feature Transform (SIFT) or Speeded Up Robust Features (SURF) descriptors. These descriptors capture the distinctive local characteristics of the image, such as shape, orientation, and texture. SIFT is designed to extract distinctive features from images that are invariant to changes in scale, orientation, and affine distortion. In the context of 3D CT scans, 3D-SIFT is an extension of SIFT that operates on 3D image volumes rather than 2D images.

3D-SIFT features are computed using the following steps. The first step is to detect the **scale-space extrema** in the 3D volume. This is done by constructing a scale-space representation of the volume, where the intensity values of each voxel are filtered using a Gaussian filter with different standard deviations. The scale-space extrema correspond to the local maxima or minima of the filtered volume. The scale-space extrema are then refined to determine the exact **location** and **scale** of the keypoints. This is done by computing the 3D Hessian matrix at each scale-space extrema and using the eigenvalues of the matrix to determine the scale and orientation of the keypoint. Once the keypoints have been localized, a **descriptor** is computed for each keypoint to describe its local appearance. In 3D-SIFT, this is done by dividing the 3D volume around each keypoint into a set of orientation histograms. These histograms capture the gradient information in different directions, and are combined to form a compact and descriptive representation of the local appearance of the keypoint.

The 3D-SIFT descriptor is a vector of orientation histograms, typically represented as a 128-dimensional feature vector. Mathematically, the 3D-SIFT descriptor d_i for keypoint i can be represented as:

$$d_i = [h_{i,1}, h_{i,2}, ..., h_{i,k}],$$

where $h_{i,j}$ is the j^{th} orientation histogram of keypoint i and k is the number of orientation bins. The orientation histograms are computed using gradient information in the vicinity of the keypoint, and are combined to form a compact and descriptive representation of the local appearance of the keypoint.

Once the visual vocabulary has been created, the feature descriptors from each image are quantized into the visual words using a clustering algorithm, such as k-means. The resulting histogram of visual words for each image can be used as a compact and informative representation of the underlying visual content, capturing the unique patterns of lung tissue appearance and texture.

In addition to SIFT and SURF, other feature extraction techniques can also be used in BoVW, such as Haralick features [57]. Haralick features are a set of texture features that describe the gray-level co-occurrence matrix (GLCM) of an image. The GLCM is a matrix that represents the probability of observing a specific gray-level pair in a given direction at a certain spatial distance. Those four points are mainly used as parameters for GLCM [57]. The GLCM can be used to calculate the symbiotic grayscale pixel values of i and j at a specified direction *theta* and distance d, expressed as the number of co-occurrence matrix element.

$$GLCM = \frac{p(i,j|d,\theta)}{\sum_i \sum_j p(i,j|d,\theta)} \tag{4.6}$$

Image **contrast** can be defined as the sharpness of the picture. Contrast increases with the depth of image grooves [58]

$$Constrast = \sum_i \sum_j (i-j)^2 P(i,j) \tag{4.7}$$

Energy can be represented as the measure of gray distribution of an image [58].

$$Engery = \sum_i \sum_j [P(i,j)^2] \tag{4.8}$$

Entropy is defined as the amount of information contained in an image. Low entropy images are blacker; a perfect image would have zero entropy [58].

$$Entropy = \sum_i \sum_j [P(i,j)] log P(i,j) \tag{4.9}$$

Image **correlation** can be described as the degree of similarity of the elements of CT scans [58].

$$Correlatio(d, \theta) = \frac{\sum_{i,j}(i - \mu_x)(j - \mu_y)P(i,j)}{\sigma_x \sigma_y} \tag{4.10}$$

,where

$$\mu_x = \sum_i \sum_j iP(i,j), \mu_y = \sum_i \sum_j jP(i,j), \tag{4.11}$$

$$\sigma_x = \sum_i \sum_j (i - \mu_x)^2(i,j), \sigma_y = \sum_i \sum_j (j - \mu_y)^2(i,j). \tag{4.12}$$

The Haralick features can capture the subtle variations in texture and pattern in an image, and have been successfully used in medical imaging applications to distinguish between different tissue types.

4.7 Image segmentation

Image segmentation is a crucial task in medical imaging and computer vision, as it entails dividing a medical image into multiple regions with shared characteristics, such as color, texture, form, and intensity. This division simplifies the analysis and study of the image, allowing the examination of each region's traits. Despite extensive research and development over the years, image segmentation remains a challenging problem due to the complex anatomy of human bodies and the diverse modalities of medical images [59].

There are several methods employed in image segmentation, including edge detection, thresholding, region growing, and clustering. The choice of method is dependent on the type of image and the task's specific requirements. Lung segmentation, in particular, refers to the identification of lung regions in a medical image, such as a CT scan. This process is essential for studying and analyzing the lungs, aiding in disease detection, diagnosis, and treatment planning. However, lung segmentation is challenging due to the presence of other structures in the image, such as the chest wall, diaphragm, and heart. Therefore, the segmentation algorithms must accurately distinguish between lung tissue and these other structures.

In 2D images, segmentation is relatively simple, and most methods utilize contrast-sensitive information obtained from the gray-level information. Conversely, in 3D images, segmentation can be more challenging, and manual settings and adjustments are often required to achieve accurate results [60] [61] [62]. The segments generated in 3D images are known as supervoxels, which are 3D pixels that represent a small volume of the image. The division of the image into smaller parts facilitates analysis and study, as the traits of each region can be compared [63]. Recent advancements in algorithms, such as MonoSLIC [64], have significantly improved the accuracy and efficiency of image segmentation, particularly in lung segmentation.

Image segmentation and lung segmentation, including supervoxel segmentation, are crucial tasks in medical imaging and computer vision. Accurate and effective segmentation algorithms are vital for the analysis and study of medical images, enabling the detection and diagnosis of diseases. The utilization of image segmentation has the potential to significantly enhance our understanding of human anatomy and aid in medical treatment planning.

4.8 Over-segmentation

Supervoxel is a term used in computer vision and image processing to describe a group of connected 3D voxels (units of 3D pixels) in a volumetric image that represents a contiguous region. Over-segmentation is a technique used to over-segment an image into smaller, spatially coherent regions with the aim of creating supervoxels that correspond to meaningful objects or structures in the image. The current state-of-the-art over-segmentation techniques are primarily designed for 2D real-world images, with only a limited number utilized for medical images. One approach, MonoSLIC, created by Holzer et al. [64], uses k-means clustering in the feature space of spatial coordinates and monogenic local phase extracted from the monogenetic signal [65] [66] and does not require parameter tuning for contrast and brightness, making it more suitable for medical imaging data. Utilizing over-segmentation can address image processing problems in certain applications by grouping pixels into larger, semantically meaningful regions, resulting in more efficient processing, improved accuracy and reduced noise.

4.9 Deep learning

Deep learning, which refers to the use of deep neural networks with multiple layers, has its roots in the 1940s and 1950s with the work of Warren McCulloch and Walter Pitts on artificial neural networks and in the 1960s and 1970s with the work of Frank Rosenblatt and others on perceptrons and backpropagation [67]. However, the development of deep learning was hindered by the lack of computational power and the limited availability of large labeled datasets.

The resurgence of interest in deep learning in the early 2000s can be attributed to several factors. Firstly, the availability of more powerful hardware, such as GPUs, made it possible to train deeper and larger neural networks. Secondly, the release of large labeled datasets, such as ImageNet[68], made it possible to train and evaluate deep neural networks on challenging tasks. Thirdly, the development of new training algorithms, such as Dropout and Rectified Linear Unit (ReLU), made it possible to train deeper and larger neural networks more efficiently.

Convolutional Neural Networks (CNNs) have a long history in computer vision and image processing. The first CNNs were developed in the late 1980s and early 1990s by Yann LeCun, Yoshua Bengio, and their colleagues at AT&T Bell Labs and the University of Montreal [69]. These early CNNs were primarily used for handwritten digit recognition, and they were inspired by the biological visual system of animals and the mathematical concept of convolution.

Generative Adversarial Networks (GANs) were introduced by Ian Goodfellow and his colleagues in 2014 [70]. They proposed a new approach to generative modeling, where two neural networks, a generator and a discriminator, are trained in an adversarial manner. The generator tries to generate new samples that are similar to the real samples, while the discriminator tries to distinguish between the real and generated samples. This approach has been shown to be effective in several generative tasks, such as image synthesis, style transfer, and data augmentation.

CNNs and GANs are both types of deep learning neural networks that have been used for different tasks in computer vision and other fields.

CNNs are designed to perform tasks such as image classification, object detection, and semantic segmentation by learning to extract and classify local features from input images. They consist of multiple layers of convolutional and pooling operations, which are used to extract and down-sample the features, and one or more fully connected layers, which are used to make a final prediction.

GANs, on the other hand, are designed to generate new samples that are similar to a given dataset. They consist of two main components: a generator network and a discriminator network. The generator network learns to generate new samples from a random input, while the discriminator network learns to distinguish between the generated samples and the real samples from the dataset. The two networks are trained in an adversarial manner, where the generator tries to generate samples that the discriminator cannot distinguish from the real samples, and the discriminator tries to improve its ability to distinguish them.

The main use cases of CNNs are image and video recognition tasks, such as object detection, semantic segmentation, and image classification, while GANs are mainly used for generative tasks such, as image and video synthesis, style transfer, and data augmentation.

In recent years, both CNNs and GANs have undergone a lot of developments by introducing new architectures and techniques, such as ResNet[71], Inception Networks [72], and Wasserstein GANs[73], and Transformer-based architectures for CNNs [74], Spectral Normalization GANs [75], and Style-based GANs for GANs [76].

Both CNNs and GANs are key components of deep learning, and they have a rich history of developments and advancements that have led to their current state-of-the-art performance in image and video recognition tasks and generative tasks respectively, while CNNs are used for discriminative tasks, trying to classify or segment an input, GANs are used for generative tasks, creating new samples that can imitate the real ones.

4.9.1 Convolutional Neural Networks

CNNs have been particularly successful in the image and video recognition tasks. They are inspired by the visual system of animals and are designed to automatically and adaptively learn spatial hierarchies of features from input images.

The main building block of a CNN is the convolutional layer, which applies a set of learnable filters (also called kernels or weights) to the input data, in order to extract local features. The filters are small in size (e.g. 2x2 or 3x3 pixels) and slide over the entire input image, computing the dot product between their weights and the overlapping image patches. This operation is called convolution, hence the name of the network.

The output of a convolutional layer is a set of feature maps, which are the same size as the input image, but have a reduced number of channels (e.g. from 3 RGB channels to 64 or 128 feature maps). The feature maps are then passed through a non-linear activation function, such as ReLU, which introduces non-linearity in the network.

Another key component of CNNs is the pooling layer, which performs down-sampling of the feature maps by taking the maximum or average value of small non-overlapping regions (e.g. 2x2 pixels). This reduces the size of the feature maps, reducing the number of parameters and computational cost, while also making the feature maps more robust to small translations of the input image.

After several convolutional and pooling layers, the feature maps are passed through one or more fully connected layers (also called dense layers), which learn a linear combination of the features and output a final prediction. The parameters of the network are learned through backpropagation and stochastic gradient descent.

CNNs have achieved state-of-the-art performance on several image and video recognition benchmarks, and are widely used in computer vision tasks such as object detection, semantic segmentation, and image generation.

4.9.2 Generative adversarial networks

A GAN framework is composed of at least two components: a discriminative model D and a generative model G. The number of inputs may vary. For training purposes, the discriminator is trained on actual examples and random batches of samples generated by the generative model. The goal of the discriminative model D is to reliably identify authentic samples as much as possible (with an output value of "True" or 1), while also recognizing the manufactured samples, also known as the false samples, as much as feasible (output is "False", or 0). These objectives correspond to the first and second terms of the objective function shown below. The goal of the generator is to generate as realistic images as possible so that the discriminative model cannot recognize it as a counterfeit. Therefore, G and D create a min-max game in which both sides optimize themselves during the training process until they achieve equilibrium. Hence, the fake samples are indistinguishable from the actual ones. The following equation illustrates the mathematical theory underlying this game of min-max:

$$\min_{G} \max_{D} V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[log D(x)] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z)))] \tag{4.13}$$

Where G represents the generator, D represents the discriminator, x represents the real images, z represents the random noise vector, $p_{data}(x)$ represents the distribution of real images, and $p_z(z)$ represents the distribution of the random noise vector. We define the mapping to data as $G(z; \theta_g)$, which means that G maps the input noise to the data based on some optimization-sensitive parameters θ_g. In addition, a mapping $D(x; \theta_d)$ is defined from the data space to a scalar output. This is the Discriminator, which assigns the probability that x originated from the data as opposed to p_{data}. For training, the two models compete in a min-max game in which D maximizes the likelihood of correctly labelling data samples from p_g and training data. G is trained to minimize $log(1-D(G(z)))$, which is formalized along with D's objective in the following expression with value function $V(G, D)$.

The result of V, is at its maximal in D, when it is assigned 1 to x and 0 to $D(G(z))$. This can be interpreted as labelling training data as "real" and synthetic data produced by G as "fake". V is minimal in G, which would means that it is able to fool D, a G generated data is labelled as coming from actual data input. With such a clever design, GAN possesses attractive properties. G in GAN, as a generative model, does not require a strict expression for the generated data, as in traditional graph models. This avoids the incomputability that results from excessive growth in complexity when the information is very complex. Also, it does not require some of the substantial computational summation computations of the inference model. The only thing it needs is a noisy input, a bunch of real data without criteria, and two networks that can approximate the function.

4.9.2.1 StyleGan

As chapter 4.9.2 mentioned, GANs are a popular deep learning technique for generative tasks, such as image and video synthesis. A common example of a GAN application is the generation of artificial face images. Over time, GAN images have become more realistic, but one of their main challenges is controlling the output of the generated images, particularly when it comes to specific features such as pose, face shape, and hairstyle. To address this challenge, a new model called StyleGAN was proposed by NVIDIA [76]. This style-based generator architecture for GAN, proposes a new model to address this challenge. It gradually generates artificial images, starting from very low resolution and working up to high resolution (1024×1024). By modifying the input at each level separately, it can control the visual features expressed in that level, from coarse features (pose, facial shape) to fine details (hair color), without affecting the other levels.

The basic components of a GAN are two neural networks: a generator (G) for new samples, and a discriminator (D) that extracts samples from the training data and the generator output and predicts whether they are "true" or "false". The input of the generator is a random vector (noise), so its initial output is also noise. As training progresses, it learns to synthesize more "real" images as it receives feedback from the discriminator. The discriminator also improves as training progresses by comparing the generated samples with the real ones, making it more difficult for the generator to fool it.

The network structure of StyleGAN consists of two parts, the first is the Mapping network, the process of mapping the input to an intermediate latent vector w from the noise variable z. This latent space W is used to control the style of the generated image, the style. The Mapping Network consists of 8 fully connected layers and its output W is the same size as the latent code Z. The second is the Synthesis network, which is used to generate images. The innovation is that each layer of the sub-network is fed with A and B. A is the affine transformation obtained from w conversion, which is used to control the style of the generated image, and B is the converted random noise, which is used to enrich the details of the generated image, i.e., each convolutional layer can adjust the "style" according to the source A.

For the Synthesis network, a key module AdaIN (Adaptive Instance Normalization), is used to control the style of the generated images. This operation adjusts the mean and standard deviation of the feature map of the intermediate latent code to match that of the reference image. This allows the synthesis network to generate images with similar styles to the reference image, while maintaining the structure and content of the intermediate latent code. This module is added in each level of resolution and can define the visual feature changes in each layer of image resolution. The AdaIN algorithm is defined as:

$$AdaIN(x_i, y) = y_{s,i} \frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i} \tag{4.14}$$

Where x_i represented the normalized feature map.

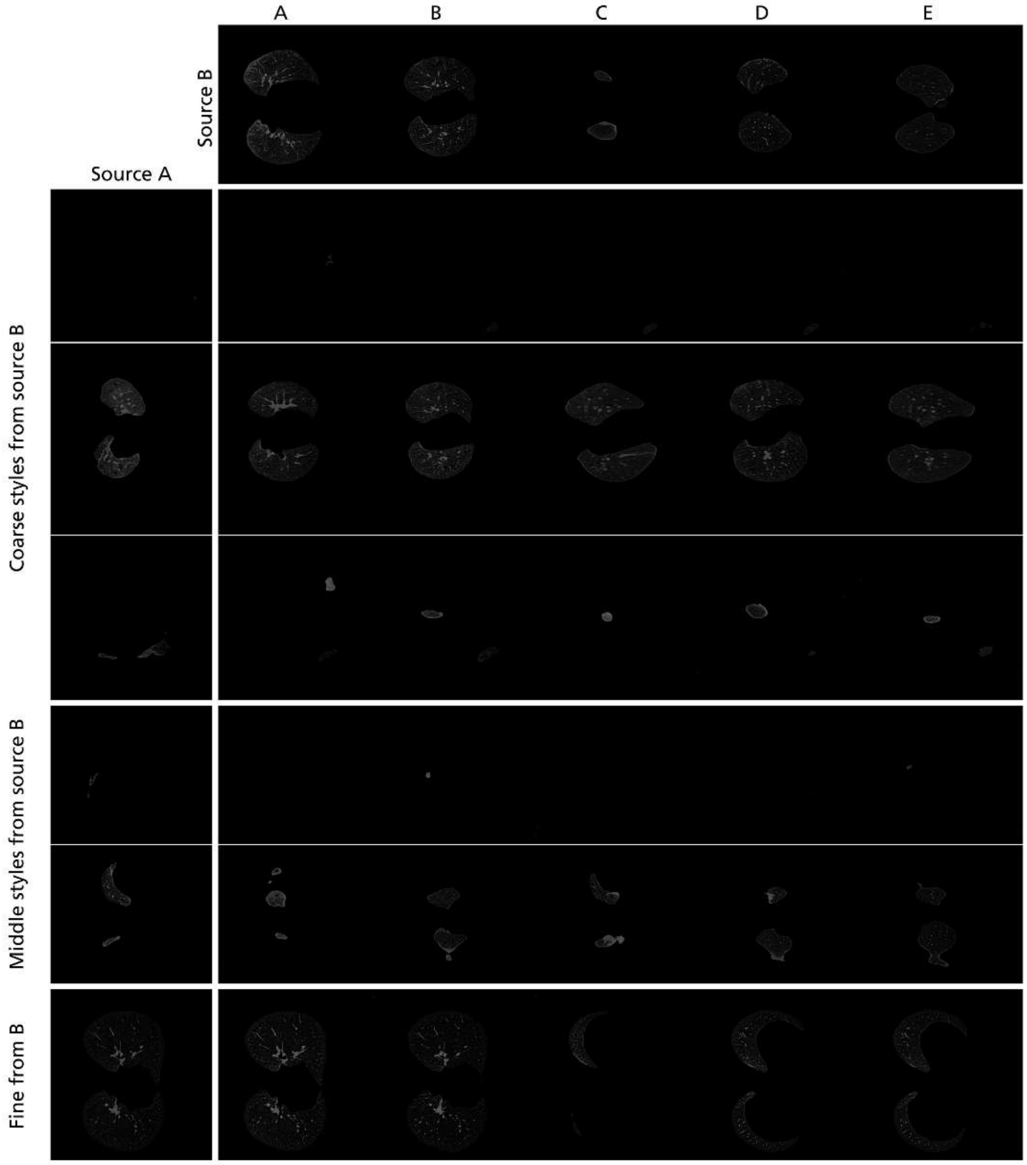


Figure 4.1: Two sets of lung images were generated from their respective latent codes (sources A and B); the rest of the images were generated by copying a specified subset of styles from source B and taking the rest from source A.

x_i is the feature map of the intermediate latent code at a particular layer of the synthesis network.

y_i is the feature map of the reference image at the same layer.

$\mu(x_i)$ and $\sigma(x_i)$ are the mean and standard deviation of the feature map x_i, respectively. They are calculated along the channel dimension of the feature maps.

$y_{s,i}$ and $y_{b,i}$ are the scaling and bias factors, respectively, calculated from the reference image feature map y_i. These factors are used to adjust the mean and standard deviation of the intermediate latent code feature map x_i to match that of the reference image feature map y_i.

$\frac{x_i-\mu(x_i)}{\sigma(x_i)}$ is the normalization step, it normalize the feature maps by subtracting the mean and dividing by the standard deviation.

The final step is to adjust the normalized feature maps by the scaling and bias factors $y_{s,i}$ and $y_{b,i}$, respectively. This is done by element-wise multiplication and addition, represented by the $\odot$ and $+$ operators, respectively. Figure 4.1 presents examples of images synthesized by mixing two latent codes at various scales. We can see that each subset of styles controls meaningful high-level attributes of the image.

4.10 Existing analysis approaches in context of IPF

There are currently several quantitative texture analysis tools[77][78][79][67][80], including DTA[81] and CALIPER[80], which are the main state-of-the-art methods using machine learning in the context of IPF. Both DTA and CALIPER use a combination of supervised and unsupervised machine learning algorithms to diagnose diffuse lung diseases. DTA starts by performing unsupervised clustering analysis on randomly sampled parenchyma from CT images of patients with IPF and non-smoking controls to produce a dictionary of low-level features that distinguish fibrosis from non-fibrotic lungs. The radiologist-labeled regions of interest (ROIs) are then used to train a supervised support vector machine classifier to distinguish fibrosis from normal lungs. The DTA fibrosis score is calculated based on the number of ROIs classified as fibrosis.

CALIPER, on the other hand, begins with supervised learning. Thin-section CT images of patients with pathologically proven ILD are used and divided into volumes of interest (VOIs). The VOIs are then categorized by expert radiologists, and multidimensional scaling is used to discriminate among the different categories. The VOIs are then clustered using an unsupervised machine learning algorithm, and the results are summarized for the entire lung. These findings can be used for statistical analysis or visually represented as a color overlay or summary glyph.

4.11 Kaplan-Meier analysis

The Kaplan-Meier analysis is a statistical method used in medical research to estimate the survival function of a population over time. The survival function represents the probability that an individual in the population has not experienced an event of interest (e.g. death, disease progression) after a certain time period.

Mathematically, the Kaplan-Meier estimate of the survival function $S(t)$ at time t is defined as:

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{4.15}$$

where t_i is the time of the i^{th} event in the population, d_i is the number of events that occurred at time t_i, and n_i is the number of individuals at risk of experiencing the event at time t_i.

In the medical field, the Kaplan-Meier analysis is widely used to study the survival of patients with a specific disease or condition. It provides a way to estimate the average survival time and the probability of survival for a population of patients. The results of the analysis can be used to compare the survival of different patient groups, for example, those receiving different treatments or those with different demographic characteristics.

4.12 Discussion

In this chapter, we have reviewed state of the art in machine learning, covering the fundamental concepts of supervised, unsupervised, and reinforcement learning. We then explored linear regression, principal component analysis, K-mean clustering, random forest, and deep learning. Deep learning, in particular, has seen tremendous growth and success in recent years, with convolutional neural networks and generative adversarial networks playing a key role.

It is worth noting that while deep learning has achieved impressive results in various applications, it is not without limitations. Deep learning models require a large amount of data and computation resources to train, and may not perform well in cases with limited or noisy data. Furthermore, the lack of transparency in deep learning models can pose a challenge in understanding and interpreting their predictions.

Despite these limitations, the continued advancements in deep learning and other machine learning techniques show promise for future advancements and applications. For example, the recent development of StyleGan in generative adversarial networks has shown significant improvements in synthesizing high-quality images.

In conclusion, state of the art in machine learning is rapidly evolving, with deep learning playing a significant role in driving this growth. Further research is needed to address the limitations of deep learning and other machine learning techniques, to make them more widely accessible and applicable for medical images.

CHAPTER 5

Methodology

This chapter provides a detailed description of the methodology used in this thesis. The objective is to use unsupervised learning to predict longitudinal imaging of idiopathic pulmonary fibrosis. Figure 5.1 displays a workflow for the method. Section 5.2 outlines the notation for mathematical objects used in work and provides a brief explanation of their characteristics. Section 5.3 is about the selection of CT scans from a pseudonymized dataset, with statistics showing the most commonly used kernel is BONEPLUS, resulting in the decision to conduct the experiment using only that kernel, with the majority of the CT scans containing background or non-lung organs. Section 5.4, focusing on the segmentation of the lung region, defines and explains this image preprocessing step. Section 5.5 describes the process of extracting hand-crafted image features of texture and shape after over-segmentation to gather complementary visual information that represents the image. The feature extraction from StyleGAN (section 5.6) is achieved by modifying the open-source code from the original StyleGAN paper to fit a training dataset of CT slices, utilizing the generated latent space representation of the images. Section 5.7 describes a study aimed at identifying pattern signature characteristics associated with the development of radiological illness through analyzing consecutive CT scans and utilizing a 500-tree random forest classification model to predict the temporal order of the scans, with the objective of finding markers associated with the onset of radiological disease. Section 5.8 details a process for aligning multiple scans of the same patient to achieve spatial correspondence. The two volumes are initially registered using the ANTs and Ezys software, with an additional affine transformation from ANTs for a good non-rigid transformation initialization. The section 5.9 focuses on the local tissue transition pathway. The goal is to track changes in lung tissue patterns as the illness progresses, and this is done by comparing the image signature components of one scan with the equivalent component in another scan. The section describes the calculation of transition probabilities between different lung tissue clusters by counting the occurrences of tissue transitions between two scans. A network is then created to illustrate these transition probabilities. Section 5.10 introduced a statistical method used to estimate the probability of survival for a population based on subgroup divisions and taking into account censoring events.

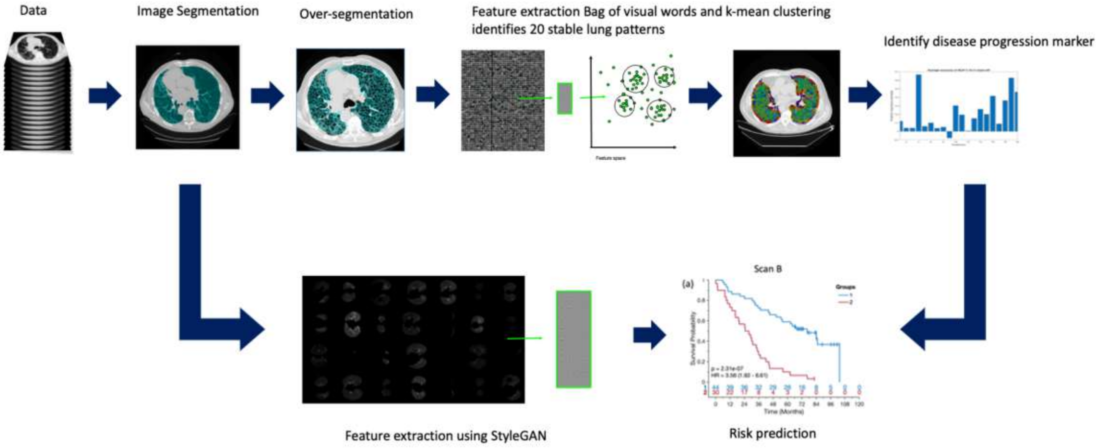

Data
Image Segmentation
Over-segmentation
Feature extraction Bag of visual words and k-mean clustering identifies 20 stable lung patterns
Identify disease progression marker
Feature extraction using StyleGAN
Scan B
(a)
Risk prediction

Figure 5.1: Overview of the algorithm

5.1 Outline of the method

The methodical approach can be separated into the following steps:

- **Image segmentation** is a two-step approach. The first step uses a simple threshold-based method where the threshold is set at -700HU. The next step removes small structures with a morphological area opening. The segmentation is then expanded with a morphological closing operation. If the first approach fails, a second approach is used, which is a multi-template atlas-based segmentation approach. The optimal lung segmentation template is selected from 16 full-body CTs in the VISCERAL Anatomy 3 dataset.

- **Features extraction using bag of visual words** describes a process of hand-crafted image feature extraction for the purpose of determining complementary visual characteristics. The Bag of Visual Words paradigm is used to reduce local features into global volume descriptions, and two vocabularies are trained for microSIFT and macroSIFT features. The hand-crafted features are then mapped to one cluster through the K-means algorithm to minimize the within-cluster sum of squares and obtain the global volume descriptions. The extracted features are reduced using PCA due to the high computation cost.

- **Feature extraction using StyleGAN** illustrates a method for feature extraction from StyleGAN for predicting disease features, specifically for IPF. The process involves modifying the original StyleGAN code to fit a training dataset of CT slices, and mapping the input noise vector to a feature space representation through linear mapping. The resulting feature is then used for prediction and classification tasks.

- The aim of **identifying diseases progression markers** is to predict the temporal order of consecutive CT scans and identify pattern signature characteristics related to the development of radiological illness using a random forest classification model and feature difference between pattern signatures of the scans and the hypothesis was that high RF Gini significance markers indicate the onset of radiological disease.

- **Image registration** process involves aligning multiple examination images of the same patient to establish spatial correspondence. This is achieved by using ANTs and Ezys registration software, which calculates the transformation T_{sij} between two volumes of a patient. This transformation is based on a deformation parameter u_{ij} that transforms the coordinates from one reference frame to the next. The goal is to bring images from previous series into the frame of subsequent series, a process known as intra-subject registration. The aim is to achieve spatial correspondence from one series to the next for a maximum of four examination images of each patient.

- **Local tissue transition pathway** in lung tissue patterns is discussed as the illness progresses. It identifies image signature components at each lung location in two scans and creates a network of transition probabilities based on the counts of the matching components. The network is calculated as the ratio of the matching components to the sum of all components for each transition.

- **Risk progression** Survival outcome prediction based on the pattern marker identified in previous steps is evaluated.

5.2 Notation

In order to offer a clear notation throughout the equations used in this work, the most significant characteristics of all mathematical objects utilized are outlined shortly below. Additional background and detailed explanations are provided wherever equations are introduced.

μ Mean of the population

σ Standard deviation of the population

F_i^t features extracted from image of individual subject i of time point t using Bag of Visual Words

$F_{Style_i}^{t}$ features extracted from image of individual subject i of time point t using StyleGAn

I_i^{t+1} lung mask of individual subject i of the following acquisition series $t+1$

I_i^t image of individual subject i at the acquisition series t

L_i^{t+1} lung mask of individual subject i at the following acquisition series $t+1$

L_i^t lung mask of individual subject i at the acquisition series t

n_{sv} number of supervoxel

P_i^{t+1} global volume description of individual subject i of the following acquisition series $t+1$

P_i^t global volume description of individual subject i of the acquisition series t

S_i^{t+1} supervoxels of individual subject i of the following acquisition series $t+1$

S_i^t supervoxels of individual subject i of the acquisition series t

t Index of acquisition time point

t_n number of acquisition time series

X dataset matrix of all subjects with size n x m

x Data point of the population

z z-scores

5.3 Data

The first step is the selection of the CT scans from the data set. All the scans are pseudonymized and are received in DICOM format. The statistics of the data set in table 6.1 show that for the recommended thin-section CT $\leqslant 1.5mm$, the most commonly used kernel is the BONEPLUS kernel. Therefore we decided to perform the experiment with only the BONEPLUS kernel. The given CT scans from the dataset contain mostly background or non-lung organs (eg. bones, heart, spine). Each patient i has image series $I_i^t \ldots I_i^{tn}$, where the maximum number of tn is 4. A patient must undergo a minimum of $t \geqslant 2$ scans using the boneplus kernel to be eligible for inclusion in the study. To effectively process the radiology information within the lung, it is necessary to perform image segmentation to accurately identify and isolate specific structures or areas of interest.

5.4 Image Segmentation

The initial process of the method described in this thesis is the execution of a lung mask segmentation.

$$f_{lungSeg} : I_i^t(x) \mapsto L_i^t(x) \tag{5.1}$$

This lung mask segmentation $L_i^t(x) \in \mathbb{R}^{m \times n \times h}$ operates automatically via a two-step approach. Human lungs contain mostly air and consequently have a low signal intensity on the HRCT scans. For a given image I of individual subject i at t acquisition time point, $I_i^t(x) \in \mathbb{R}^{m \times n \times h}$ the initial lung mask segmentation is obtained via a simple threshold-based method, where the entry for $B_i^t(x) \in \mathbb{R}^{m \times n \times h}$ is set at -700HU on the CT images[82].

$$B_i^t(x) = \begin{cases} 1, & I_i^t(x) \leq -700HU \\ 0, & I_i^t(x) > -700HU \end{cases} \tag{5.2}$$

The next step is to remove small structures $A_i^t \in \mathbb{R}^3$ such as small bronchi and vessels with a morphological area opening.

$$B_i^t(x) \circ A_i^t = (B_i^t(x) \ominus A_i^t) \oplus A_i^t. \tag{5.3}$$

This operation removes connected components with a pre-defined number of voxels. Additionally, connected components attached to the image border are not considered in the analysis. We divide the airways according to the Lee and Reeves proposed region-growing process with leakage detection and prevention [83]. By selecting the largest two connected components with a volume of at least $200cm^3$, we are able to segment the lungs. The lungs can sometimes touch each other and come together to form a single, connected organ. To find the best cut separating the two lungs in this instance, we employ a graph-cut method developed by Pinho et al. [84]. Lastly, a morphological closing operation with a 3-dimensional spherical structure $D_i^t \in \mathbb{R}^3$ with a radius of 7 mm was performed on each part of the segmented lung (left and right lung).

$$B_i^t(x) \bullet D_i^t = (B_i^t(x) \oplus D_i^t) \ominus D_i^t. \tag{5.4}$$

The segmentation of the lung $\hat{L}_i^t(x) \in \mathbb{R}^{m \times n \times h}$ is expanded by this operation to include small, previously unsegmented structures (such as vessels).

$$\hat{L}_i^t(x) = \begin{cases} B_i^t(x), & \sum B_i^t(x) \geq 200cm^3 \\ \\ 0, & \sum B_i^t(x) < 200cm^3 \end{cases} \tag{5.5}$$

For a healthy subject, this method works flawlessly. Our dataset, however, consists of patients diagnosed with pulmonary fibrosis; in the later stages of the disease, lung fibrosis and high density lung patterns such as ground-glass-opacity appear. Figure 6.3 provides an example. In many cases, the described method of lung mask segmentation fails. Upon failure, a second approach automatically takes over for the lung mask segmentation. We use a multi-template atlas-based segmentation approach to correct the segmentation [85] if the algorithm fails, especially in cases of substantial high density areas and lung scarring. The atlas approach automatically selects an optimal lung segmentation template. The lung transformation template is selected from 16 full-body CTS from the VISCERAL Anatomy 3 dataset [86].

For a given image of individual subject I_i^t and a set of manually annotated template candidates (16 full-body CTs from the VISCERAL Anatomy 3 dataset [86]) $E_1, \cdots, E_E$, which are previously registered to an atlas.

$$T_{ie} : I_i^t(x) \mapsto E_e, \tag{5.6}$$

where T_{ie} indicate a non-linear transformation from I_i^t to a template out of the 16 full-body CTs E_e

$$T_{eA} : E_e \mapsto A, \tag{5.7}$$

and T_{eA} the transformation from E_e to atlas A. I_i^t is then mapped to A by concatenating both non-linear transformations so that

$$A \approx T_{eA}(T_{ie}(I_i^t)) \tag{5.8}$$

Normalized Cross Correlation (NCC) criteria is chosen as the quality criteria to optimizes the non-linear transformation.

$$L_i^t(x) = \begin{cases} \hat{L}_i^t(x), & \hat{L}_i^t(x) > 0 \\ arg\max_{1<e\leq E} NCC(A, T_{eA}(T_{ie}(I_i^t))), & \hat{L}_i^t(x) = 0 \end{cases} \tag{5.9}$$

5.4.1 Over-segmentation

We adapted the concept of MonoSLIC(Section 4.8) to over-segment the lung after lung mask segmentation. The supervoxels $S_i^t \in \left\{1, ..., n_{sv}\right\}^{n\times h}$ is written as:

$$f_{MonoSLIC} : \left\langle I_i^t(x), L_i^t(x) \right\rangle \mapsto S_i^t(x) \tag{5.10}$$

, where n_{sv} represents the number of supervoxels.

This approach was used since it produced a four-dimensional feature space for a three-dimensional picture. The result of the entire image segmentation pipeline can be seen in figure 5.4. The original CT slice scan I_i^t (figure 5.4 A) sized 512x512 pixels is shown before the image segmentation, after applying the lung mask segmentation L_i^t (figure 5.4 B), and lastly the mapped to S_i^t with the over-segmentation method (figure 5.4 B).

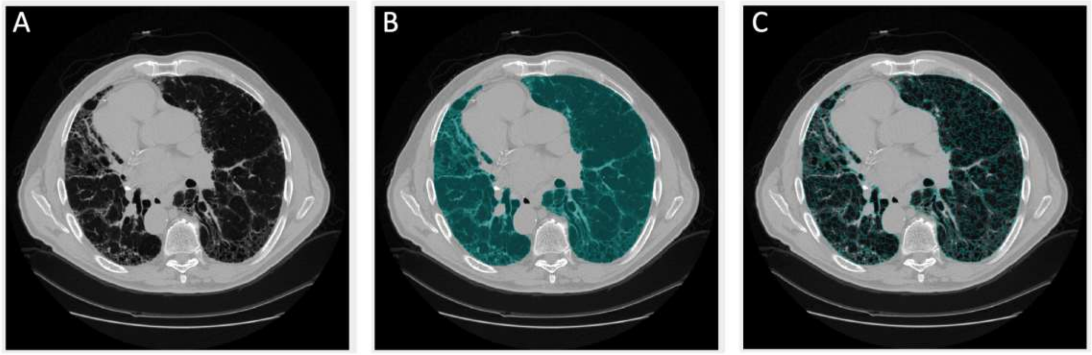


Figure 5.2: The first image shows an example of a CT image I_i^t before segmentation(A). The middle image shows a high-resolution CT image after lung mask segmentation (B) L_i^t. The last image shows the image segmentation with supervoxel S_i^t

5.5 Feature extraction with bag of visual words

Hand-crafted image feature extraction is performed after over-segmentation. For the purpose of the method two types of features were extracted; texture and shape features. This is done in order to determine complementary visual characteristics and to gather visual information that represents the image:

$$f_{feats} : \left\langle I_i^t, S_i^t \right\rangle \mapsto F_i^t \tag{5.11}$$

We adopt a paradigm known as the Bag of Visual Words (Section 4.6) to effectively reduce the specifics of local features into global volume descriptions. Enhancing the characteristics beforehand by assigning them a location in the reference space enables us to be prepared for any circumstance. As a direct result of this, it is possible to cultivate spatial-visual vocabularies. We train two distinct vocabularies, one for microSIFT (3D-SIFT features with a diameter $< 2cm$) and one for macroSIFT (diameter $\geq 2cm$), in order to account for the varied occurrence frequencies of tiny and big 3DSIFT features. MicroSIFT characteristics have a diameter less than 2 cm, whereas macroSIFT characteristics have a diameter greater than 2 cm. The word count feature representations for an over-segmented image S_i^t are designated by the notations $\mathbf{f}_{H}{}_i^t$ (Haralick)(Section 4.6) and $\mathbf{f}_{S}{}_i^t$ (SIFT)(Section 4.6). These notations are interchangeable in practice. The extracted hand-crafted features $\mathbf{F}_i^t = \binom{\mathbf{f}_{H}{}_i^t}{\mathbf{f}_{S}{}_i^t} \in \mathbb{R}^{n_f \times n_{sv}}$ contains n_f statistical value from the texture $\mathbf{f}_{H}{}_i^t$ and shape features $\mathbf{f}_{S}{}_i^t$. PCA (Section 4.3) is used for dimensional reduction, due to the computation cost of high dimensional calculation. The extracted features are mapped to one cluster k, which represents the global volume descriptions:

$$f_{descriptions} : \left\langle S_i^t, F_i^t \right\rangle \mapsto P_i^t \tag{5.12}$$

Let $f = f_1, f_2, \ldots, f_N$ be the set of N feature observations in an M-dimensional feature space, where $f_j \in \mathbb{R}^M$. The K-means algorithm (Section 4.4) assigns each observation f_i to one of the K clusters C_k, where $k = 1, 2, \ldots, K$. The goal is to find the partition of the observations into clusters $C = C_1, C_2, \ldots, C_K$ such that the within-cluster sum of squares (WCSS) is minimized. The WCSS is defined as:

$$WCSS(C) = argmin \sum_{k=1}^{K} \sum_{f_j \in C_k} |f_j - \mu_k|^2 , \tag{5.13}$$

where μ_k is the mean of the observations in cluster C_k.

$$P_i^t = WCSS(F_i^t) \tag{5.14}$$

,

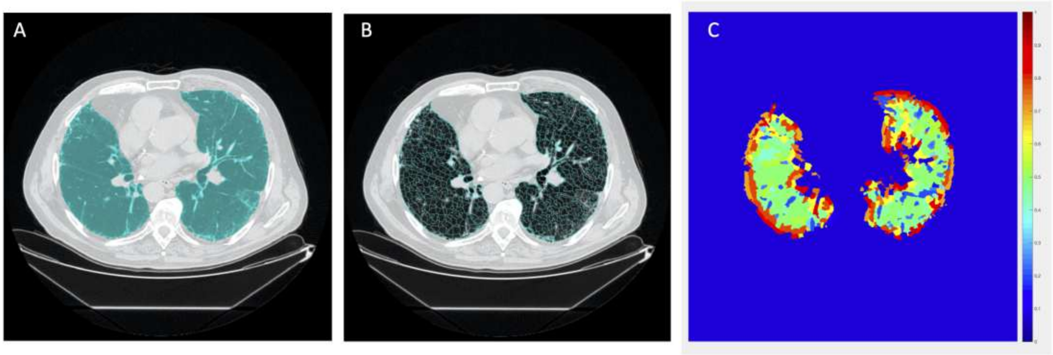


Figure 5.3: The left image shows an example of a CT image after applying the calculated lung mask L_i^t of the usable volume within the lung (A). The middle image shows a high-resolution CT image after over-segmentation S_i^t, where each voxel is define for 5x5mm space resolution (B). The image on the right shows the over-segmentation voxel mapped to one clusters , which represents the global volume descriptions P_i^t.

5.6 Feature extraction using StyleGAN

Feature extraction from StyleGAN can be achieved by leveraging the fact that the model does not require pre-defined latent space input. This allows us to learn disease features from the latent space generated by StyleGAN, thereby enabling the reverse engineering of specific disease features, such as those associated with IPF. To accomplish this, we modified the open-source code from the original StyleGAN paper by Karras et al [73] to fit our training dataset. Our training dataset is comprised of 2D images of CT slices, and after applying the previous lung mask segmentation, each image has a size of 512 X 512 pixels. The number of possible CT slices per subject is represented by s. The input to the StyleGAN model is a noise vector $z \in \mathbb{R}$ which is used to generate a latent space W representation of the image. This latent space representation is then used to generate an output image. The feature extraction process is accomplished by mapping the input noise vector z to the feature space $F_{Style_i^t}$ through linear mapping. The feature extraction process can then be expressed as

$$F_{Style_i^t} = W_z(\left\langle I_i^t, L_i^t \right\rangle) + b \tag{5.15}$$

where $W \in \mathbb{R}$ and $b \in \mathbb{R}$ are the weights and bias of the mapping, respectively. The output feature $F_{Style_i^t}$ is then used as parameter for outcome prediction and classification tasks.

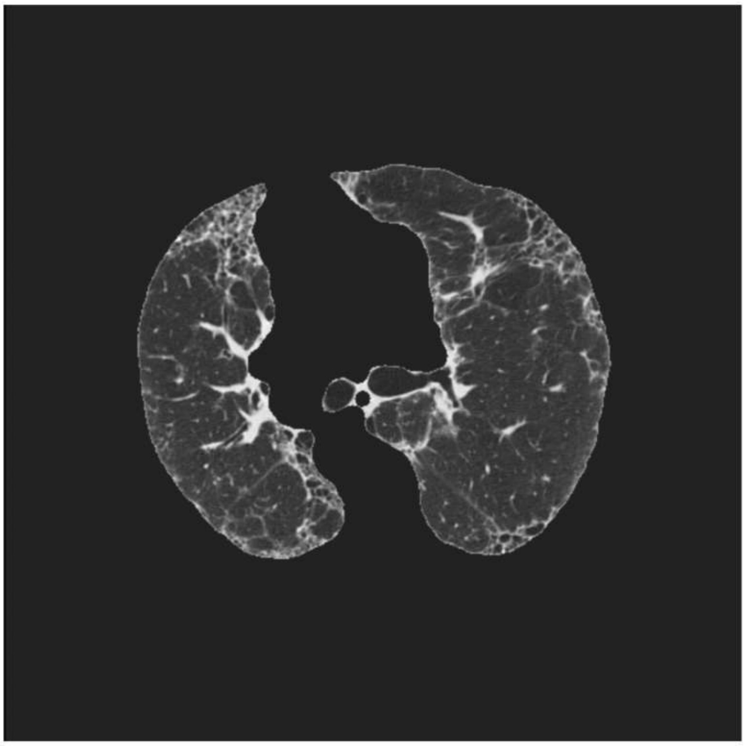


Figure 5.4: Data input from the training of the StyleGAN

5.7 Identify diseases progression marker

In order to identify pattern signature characteristics associated with the development of radiological illness, we analyzed available pairs of consecutive over-segmented CT scans P_i^t and P_i^{t+1}.

$$Q_i^t(j) = count \left\{ P_i^t = j \right\}, \forall j = 1, ...n_{cl} \tag{5.16}$$

, where n_cl is the number of k-mean clusters.

Our objective was to predict the correct temporal order of these scans. To achieve this goal, we utilized a 500-tree random forest (RF) classification model (section 4.5) that was trained on the variance in the pattern signatures of the CT scans. The feature difference ($\Delta Q = Q_i^{t+1} - Q_i^t$) between the pattern signatures served as the basis for this prediction, resulting in the categorization of the scans as either I_i^t acquired after I_i^{t+1} or I_i^{t+1} acquired after I_i^t.

The ground truth for the training data was obtained from the DICOM header of the CT scans, which recorded the acquisition dates. We evaluated the contribution of each feature to the correct categorization of the scans by determining its Gini significance.

Based on this analysis, we assigned a score to each feature to reflect its importance in the prediction.

Our working hypothesis was that markers with high RF Gini significance would be strongly associated with the onset of radiological disease. We supported this hypothesis by observing that IPF is an irreversible illness, and thus the scarring of the lungs shown on the CT scans does not worsen or improve over time. As a result, the features that enable accurate temporal sorting are believed to indicate the development of radiological progression.

An illustration of a training patient's consecutive CT pair is shown in Figure 5.5.

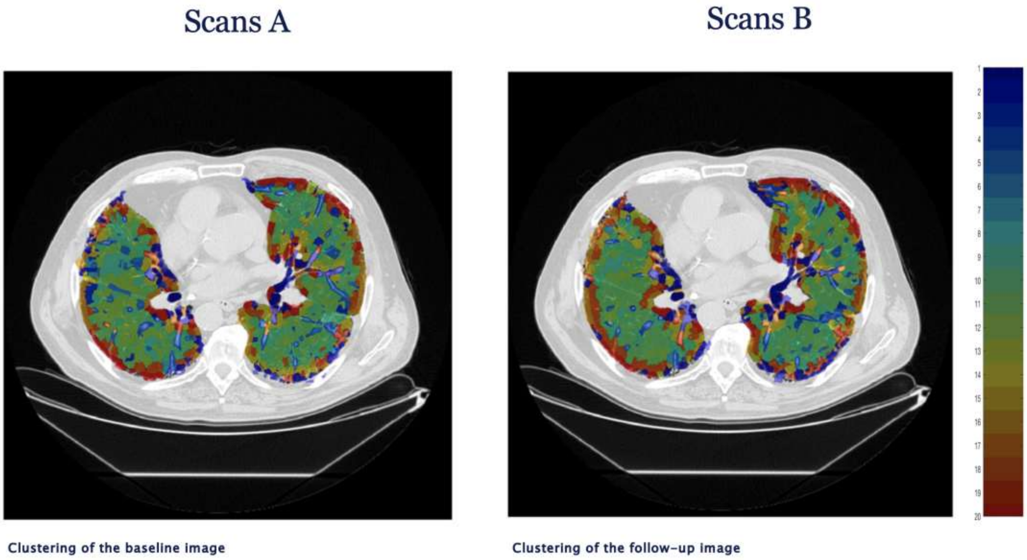

Scans A
Scans B
Clustering of the baseline image
Clustering of the follow-up image
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Figure 5.5: The image on the left shows an indication of the global volume description P_i^t, occurring chronologically prior to the image on the right. The image on the right shows the global volume description of P_i^{t+1} .

5.8 Image Registration

After clustering, image registration is the next processing step. The two volumes of each patient are first registered affine. The transformation T_{sij} is calculated by the ANTs registration software [87], followed by the application of the Ezys[88]. However, the affine transformation calculated using Ezys[88] alone results in too weak an initialization for the non-rigid transformation, which partially invalidates the results. This occurs because Ezys[88] automatically performs an affine transformation before the non-rigid registration. To get a good initialization for the non-rigid transformation portion, ANTs [87] performs an additional affine transformation. After the registration step we have for each follow-up acquisition series j a transformation T_{sij} from a previous acquistion series i, based on a deformation parameter u_{ij} that transforms the coordination from each patient reference to the image:

$$T_{sij}(x) = x + u_{ij}(x), \forall x \epsilon \Omega_s \tag{5.17}$$

For each patient we have e examinations, where $e = 2, ...4$, which can be aligned. The data processed for each patient in this section consists of the initial images I_t of each acquisition timepoints t, and the following acquisition images I_{t+1}. The parameter t identifies the acquisition series by its chronological order starting at one and ending with the number of acquisition series available for the actual patient, in our case a maximum of four. The aim is to reach spatial correspondence from series t to series $t+1$ of the same patient i, which is also called intra-subject registration. The source frame is the chronologically previous scan. If e examinations are available for the subject i, then the goal is to bring $I_i^t(x)$ into the frame of $I_i^{t+1}(T(x)), \forall t < e - 1$.

5.9 Local tissue transition pathway

Lung tissue patterns change from cluster to cluster as the illness progresses. These changes are visible over single or several sequences. To identify the image signature component at each lung location in one scan P_i^t and the equivalent component in the second scan P_i^{t+1}.

$$M(k,l) = counts\left\{P_i^t(x) = k \wedge P_i^{t+1}(T(x)) = l\right\}, \forall i, \forall te - 1 \tag{5.18}$$

As a direct result, a network illustrating transition probabilities $\hat{M}(k,j) \in \mathbb{N}^{n_{cl} \times n_{cl}}$ was created.

$$\hat{M}(k,l) = \frac{M(k,l)}{\sum_{l=1}^{n_{cl}} M(k,l))}, \forall k = 1, .., n_{cl}, \forall l = 1, .., n_{cl} \tag{5.19}$$

5.10 Risk prediction

The Kaplan-Meier analysis(section 4.11) is a statistical method used to estimate the probability of survival (or time to event) for a population over a certain period of time. In this analysis, the population is divided into subgroups based k-means (section 6.2), and the probability of survival for each subgroup is estimated and plotted over time. The Kaplan-Meier curve represents the accumulated probability of survival for the population, taking into account censoring, which refers to cases where the event of interest (such as death or disease progression) has not occurred at the time of analysis. The result of the Kaplan-Meier analysis is often used in medical research to estimate survival probabilities for patients with different diseases, treatments, or risk factors and to compare the outcomes of different interventions.

5.11 Discussion

The methodology utilized in the creation of this thesis has been detailed in this chapter. It consists of several steps, including image segmentation, features extraction, identifying disease progression markers, image registration, local tissue transition pathways, and risk progression.

Image segmentation is the first step in the process, as it involves separating the CT scans into meaningful structures. The first step of this process uses a simple threshold-based method, where the threshold is set at -700HU, followed by a morphological area opening to remove small structures. If the first approach fails, a second approach, which is a multi-template atlas-based segmentation approach, is used. The optimal lung segmentation template is selected from 16 full-body CTs in the VISCERAL Anatomy 3 dataset.

Features extraction using bag of visual words and StyleGAN are also important components of the methodology. The bag of visual words paradigm is used to extract hand-crafted image features and reduce local features into global volume descriptions, which are then mapped to one cluster through the K-means algorithm. On the other hand, the StyleGAN method is used for feature extraction from StyleGAN for the purpose of predicting disease features, specifically for IPF.

The next step involves identifying disease progression markers by predicting the temporal order of consecutive CT scans and identifying pattern signature characteristics related to the development of radiological illness using a random forest classification model and feature differences between pattern signatures of the scans. The hypothesis is that high RF Gini significance markers indicate the onset of radiological disease.

Image registration is the subsequent step in the process, as it involves aligning multiple examination images of the same patient to establish spatial correspondence. This is achieved by using ANTs and Ezys registration software, which calculates the transformation T_{sij} between two volumes of a patient based on a deformation parameter u_{ij}.

The local tissue transition pathway in lung tissue patterns is a crucial aspect of the process, as it identifies image signature components at each lung location in two scans and creates a network of transition probabilities based on the counts of the matching components. This network is calculated as the ratio of the matching components to the sum of all components for each transition.

Finally, the risk progression step involves evaluating the survival outcome prediction based on the pattern marker identified in previous steps. The results of this analysis can provide valuable insights into the progression of radiological diseases and support the development of effective treatments and interventions.

In conclusion, the method presented in this thesis is a complex and multi-step process, which involves image segmentation, features extraction, identifying disease progression markers, image registration, local tissue transition pathways, and risk progression. These steps work together to provide a comprehensive and detailed analysis of disease progression and support the development of effective treatments and interventions.

CHAPTER 6

Experiments and Results

In this chapter, we explain the experiments undertaken for this study. In Section 6.1, the dataset used in this thesis is discussed. The way how the images are acquired and collected can be found in Section 6.2. The application of image segmentation of the proposed method can be found in Section 6.3. Features extracted with different methods is in Section 6.4 and Section 6.5. Identification of the disease related progression marker is demonstrated in Section 6.6. The likelihood of a lung texture pattern changing from the prior pattern to another pattern is interpreted in Section 6.7. A survival prediction of the Kaplan-Meier study is analyzed in Section 6.8. Finally, the whole experiment and results are briefly summarized in Section 6.9.

6.1 Study cohort

The study dataset used for this thesis was retrospectively retrieved from the electronic registers of an Italian referral center (Ospedale Morgagni di Forlì, Italy). It contains a dataset of 106 patients diagnosed with IPF between December 2011 and October 2014. The following inclusion criteria were in place: (1) availability of at least two consecutive HRCT examinations per patient performed at least six months intervals; (2) use of a high-frequency reconstruction kernel (BONEPLUS) with slice thickness =1.25 mm for both exams. Following these inclusion criteria, 76 patients (f/m: 19/57) were included, as follow-up scans with the same reconstruction kernel were only available in these cases. For 74 patients in a sub-cohort, survival data were available. Another retrospective cohort from a different center and country (n=18, Vienna General Hospital, Austria) was used as a validation dataset. This dataset includes patients diagnosed with IPF between April 2007 and April 2017. The inclusion criteria were the same as those of the study referral center. However, the CT reconstruction kernel was different (B60f, B70f, B70s, I70f, I80s) due to the different manufacturers of the scanners. For both cohorts, two expert radiologists determined the CT diagnosis. The multi-disciplinary ILDs boards made the diagnosis of the IPF for both institutions [89].

6.2 Imaging data collection and acquisition

Two CT scanners, a Lightspeed Pro 16 and a BrightSpeed 16, were utilized in order to collect the data for the study's cohort in Italy (both GE Healthcare). The CT exams were carried out with the patients lying in the supine position while maintaining a steady level of deep inspiration. If a patient underwent more than two CT tests, each pair of successive CT scans was considered for inclusion. As a result, a single patient may experience anywhere from one to four sets of scans. Data were collected for the validation cohort using a Siemens Sensation Cardiac 64 scanner while the subjects were in the supine position and taking deep breaths. Each patient underwent two scans: the first one was performed at the time of diagnosis, and the second one was performed at the next hospital visit [89]. The acquired thin-section CT for the study in this work has a slice thickness between 1.25 – 3.75 mm.

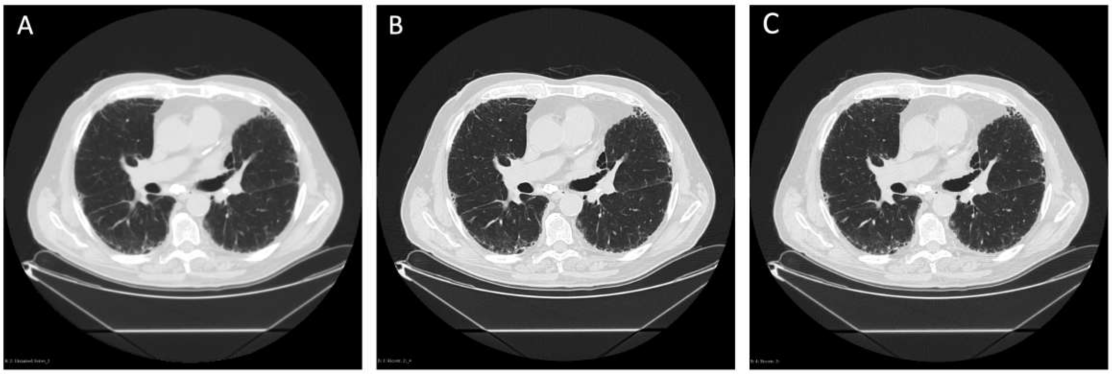


Figure 6.1: Examples of how various slice thicknesses can be combined with different reconstruction kernels. The same patient and same examination were reconstructed with different thicknesses and kernels: (A) 3.75 mm with the standard kernel, (B) 2.5 mm with lung kernel, (C) 1.25 mm with boneplus kernel

According to the official clinical practice guideline for the diagnosis of IPF by Raghu et al.[5], the recommended thin-section CT should be $\leqslant 1.5mm$. Table 6.1 summarizes the essential properties of thickness and kernel from the collected data.

Possible kernels and thickness combination			
Slice thickness	1.25 mm	2.5 mm	3.75 mm
STANDARD	7	8	198
SOFT	2	37	85
LUNG	12	18	5
BONE	98	-	-
BONEPLUS	227	-	-

Table 6.1: Properties of the data set. 106 patients diagnosed with IPF between December 2011 and October 2014. 76 patients with at least two consecutive BONEPLUS kernel scans were selected.

The experiments were done based on the guideline of the recommended thickness of $\leqslant 1.5mm$. Therefore the reconstruction kernel was chosen for BONEPLUS kernel.

6.3 Lung segmentation

All volumes are transformed into isotropic voxels with a resolution of $0.7mmX0.7mmX0.7mm$ to allow homogeneous processing afterwards. A lung mask is created for each volume by applying the simple threshold-based method. The threshold is set at $-700HU$ on the lung CT scans as viewing windows, or in the case of high-density lung patterns, a multi-template atlas-based segmentation approach is used [85]. In theory, the threshold-based methods use morphological area opening can have two most significant connected elements with a volume of $\geq 200cm^3$; if there is no or ≤ 2, the segmentation for lung mask will change to the atlas-based method. After the lung mask, an over-segmentation is created utilizing a 3-D adaption modified by the mono-SLIC superpixel algorithm [85]. The voxel size of the supervoxel is set to be $0.5cm^3$. The total number of supervoxels across the study cohort is $N = 1,578,788$. Figure 6.2 demonstrated the segmentation result of a subject's lung using the simple threshold method

For the patient with late-stage IPF (Figure 6.3) the simple threshold method fail the segmentation, therefore the lung mask segmentation is based on the second method, the multi-template altlas-based approach.

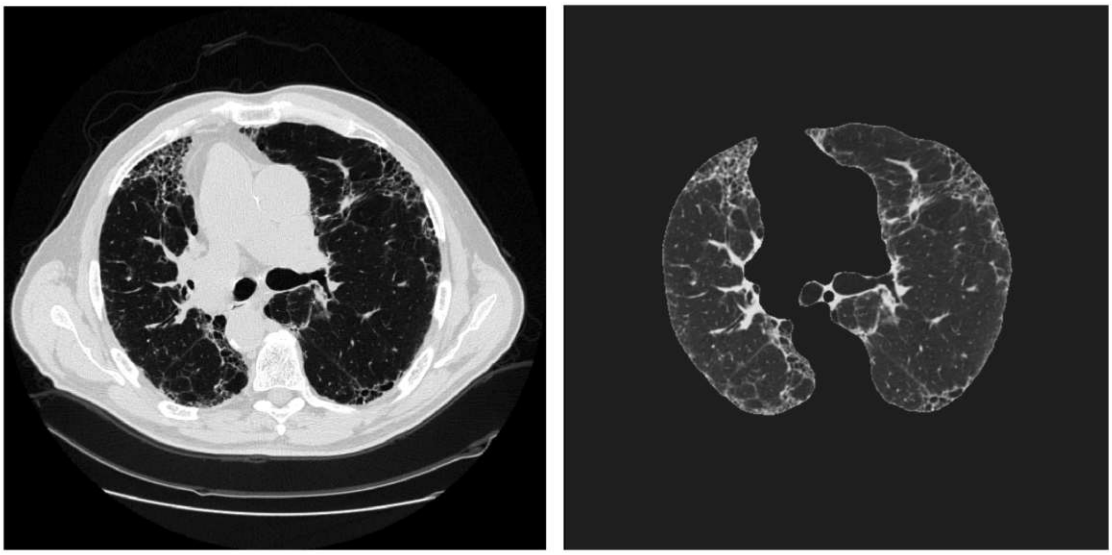


Figure 6.2: The left image shows an axial plane cut of a patient with an early stage of idiopathic pulmonary fibrosis, where the simple threshold method works. The right image showcases the lung foreground images after applying the lung mask.

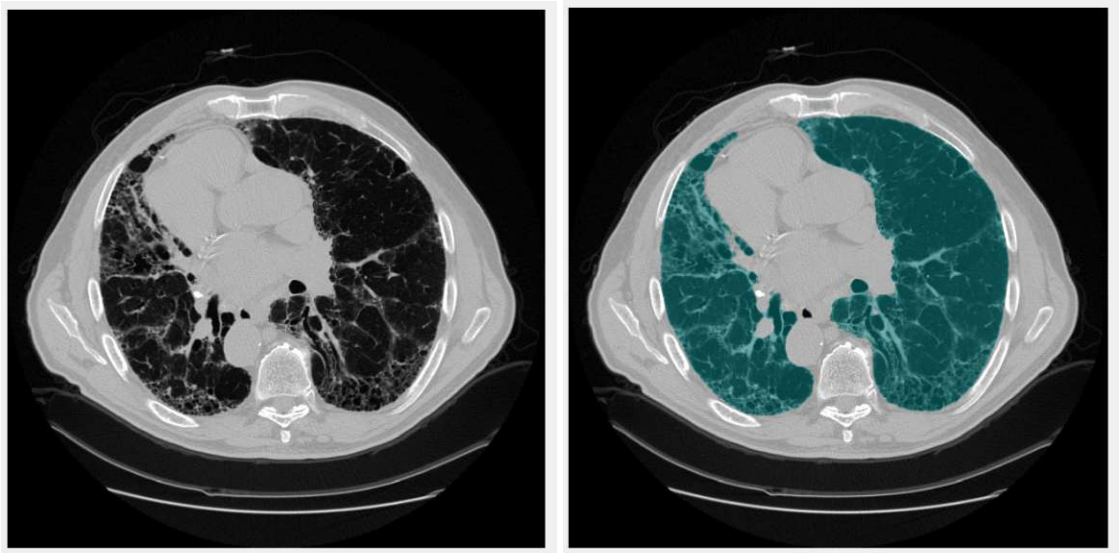


Figure 6.3: A patient in the late-stage of the disease progression. The high-resolution CT images before lung mask segmentation (left). The masked image after multi-template atlas based segmentation.

6.4 Feature extraction with bag of visual words

The aim is to evaluate lung appearance patterns in CT scans to identify disease progression markers. For each supervoxel of size $0.5cmX0.5cmX0.5cm$, a 56-dimensional feature was extracted using Haralick and SIFT features. The total number of supervoxels is 1,578,788, and the feature dimension is 9 after PCA. To determine the optimal number of clusters for the k-means clustering algorithm, a range of values from 2 to 40 clusters were tested. The optimal number of clusters was determined by the Jaccard score [90]. The results showed that the optimal number of clusters was $k = 20$. Each supervoxel was assigned a lung appearance pattern based on its 9-dimensional feature space. Every lung scan was represented by the volume fraction covered by each of the 20 appearance patterns, and this information was used as input to a random forest classifier to identify disease progression markers. The overall texture information of the lung volume was represented as a vector of 20 elements.

6.5 Feature extraction using StyleGan

The training process of the proposed method took a total of 18 days and resulted in a 92% accuracy rate for the training dataset. However, a comparison between the real CT scans(Figure 6.4) and the generated images (Figure 6.5) showed a noticeable difference between the two. This discrepancy suggests that the initial approach of learning and extracting the disease progression features from the high dimensional latent space of the StyleGAN was not effective. The high dimensionality of the latent space and the 18 layers of output may have contributed to this result.

In order to achieve more accurate results, it will be necessary to involve IPF lung expert radiologists in the training process. These experts can provide input and guidance to ensure that the disease progression features are correctly understood from the images. Unfortunately, this collaboration was not possible during the period of this thesis. As a result, future efforts will focus on the extraction of features from the bag of visual words, which may provide a more effective and reliable method for identifying disease progression patterns in CT scans.

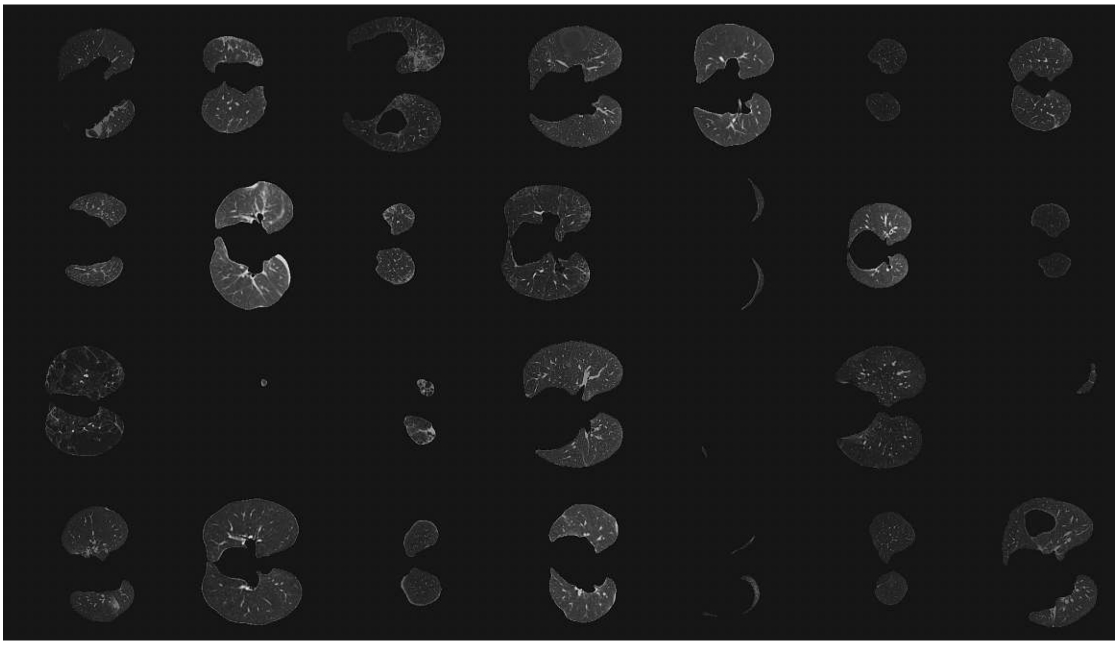


Figure 6.4: The real training image of the styleGAN[76].

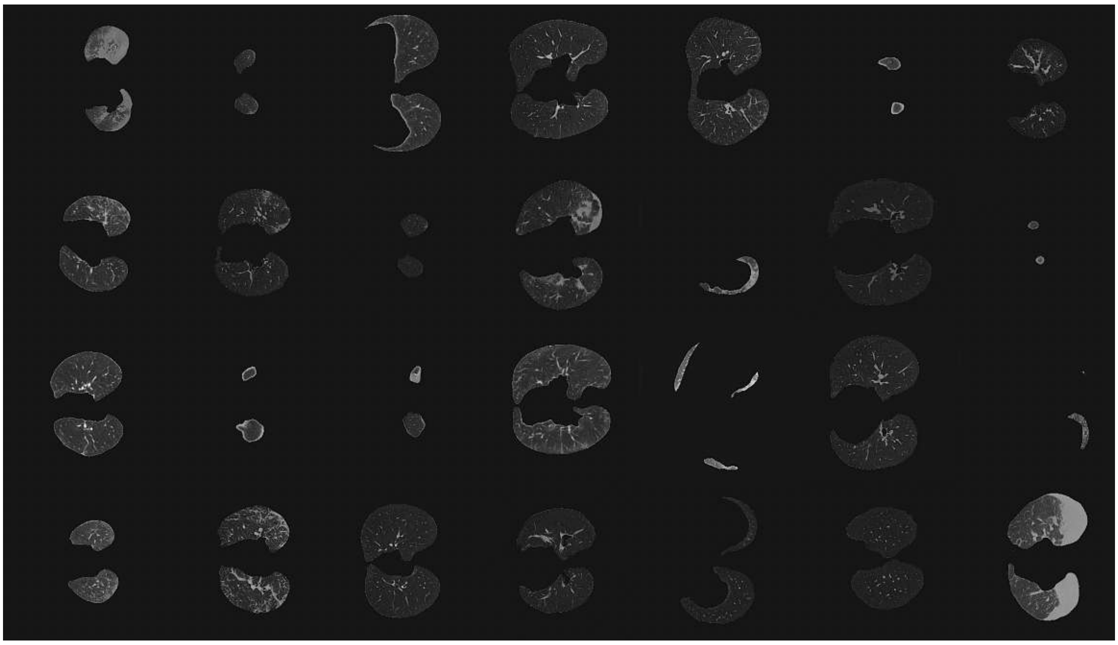


Figure 6.5: The generated image with the latent space space w of figure ??

6.6 Identify diseases progression marker

In general, the quality of prediction outputs of a model requires ground truth labels. However, due to the lack of labels for this particular dataset, an alternative solution to the challenge of constructing an unsupervised model has been proposed. This solution utilizes the characteristic of an abnormal growth path of lung scarring for IPF patients, which tends to increase over time. This characteristic was used to evaluate the model's performance and ensure that its clusters of predictions exhibited commonality. This characteristic was utilized to evaluate our model.

The evaluation was carried out by building a radiological disease progression model that used the overall texture information of the lung volume from pairs of subsequent CTs. To validate this model, we tested if the machine learning model alone could correctly determine the temporal sequence of the scans. The most informative radiological disease progression marker candidates were identified by training a random forest classification model with 500 trees. The stability of the model was evaluated using 20-fold cross-validation with a 95%-5% (training and testing) split on patients.

The measure of the evaluation was the classification accuracy of accurately sorted scans compared to two experts. The average accuracy of the model was 83% for correctly predicting the sequence of CT pairs. The GINI Importance across the 20 runs was ranked by its importance, with the average ranking and the rank standard deviation shown in figure 6.6. The four top candidates (11 - 7 - 10 - 17) consistently ranked as the top four across all runs.

To further illustrate the results, figure 6.7(lower) shows the top four ranked cluster's volume representations from a patient at four different time points, as well as example patches (Figure 6.7(upper)) of those four patterns from the same patient. The top-ranked prototypes were assessed and evaluated as image patches (250x250 pixels) by an expert, with possible biological interpretations provided.

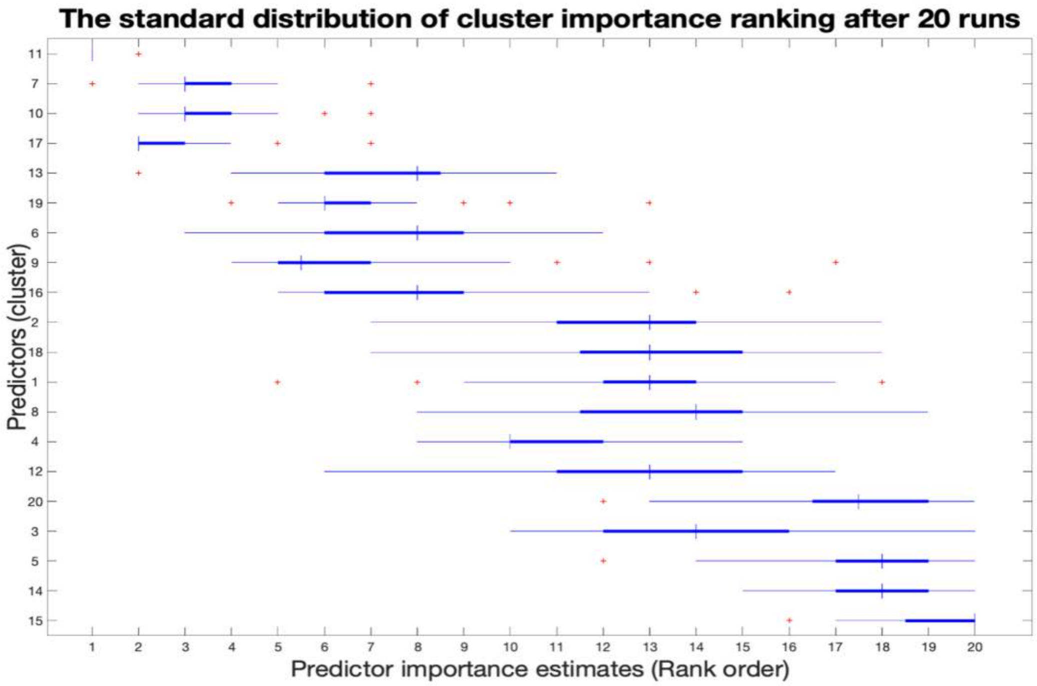


Figure 6.6: Evaluation of the stability of the progression markers. Most informative progression markers identified by the model, and the repeatability of this ranking after 20 runs of random 95%-5% patient splits. The top 4 ranked patterns are stable across all runs. The ranking of less informative patterns fluctuates across runs.

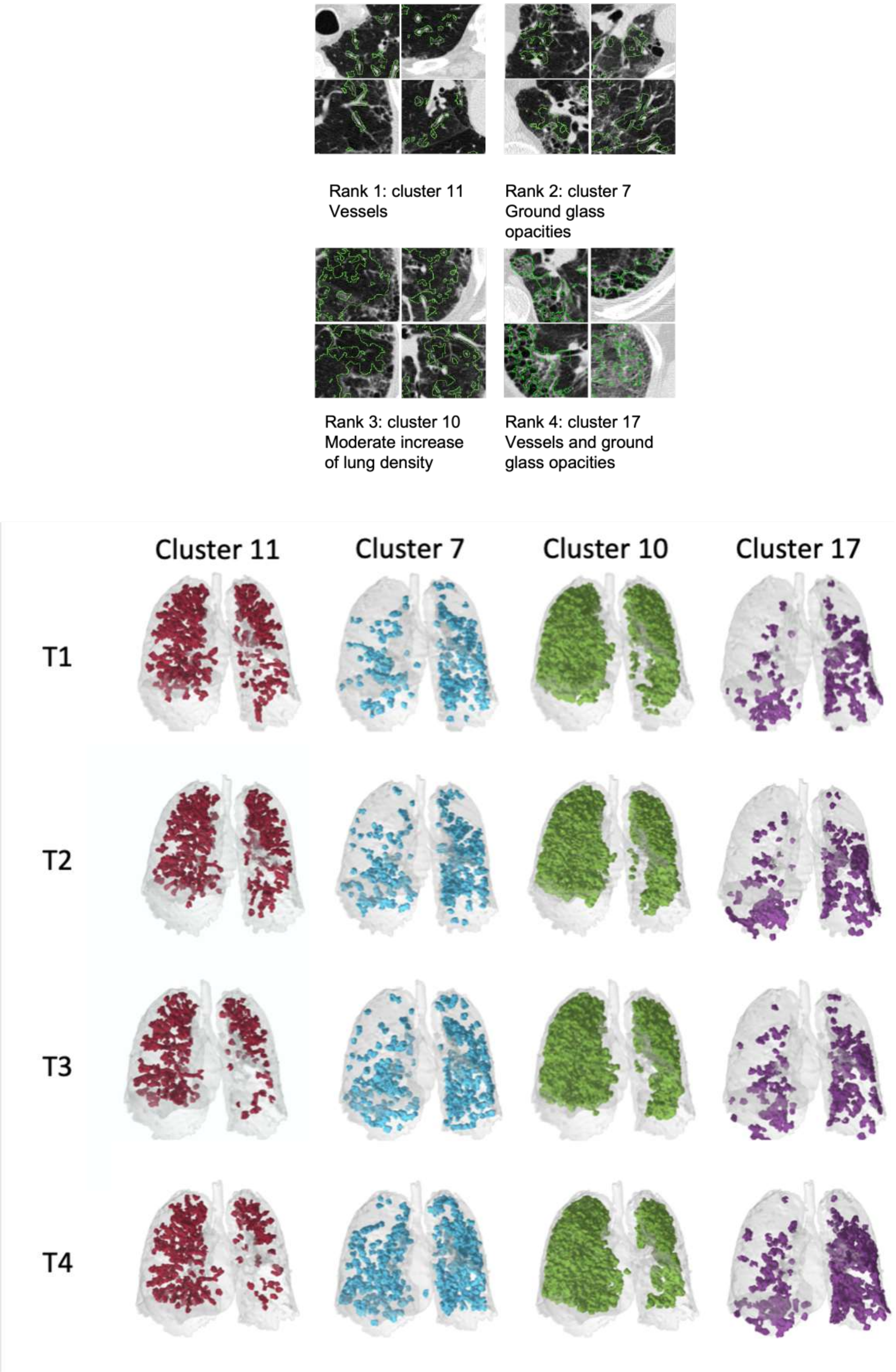


Figure 6.7: (Upper) The pattern example among the top 4 ranked pattern.(Lower) The top 4 ranked cluster volume representation from a patient at 4 different time points.

	Reader 1	Reader 2	Overlap errors
R1 = expert 1, R2 = expert 2	19	38	9(47.3%ofR1,23.6%ofR2)
R1 = expert 1, R2 = ML model	19	19	7 (36.8% of R1)
R1 = expert 2, R2 = ML model	38	19	12 (31.57% of R1)

Table 6.2: Comparison of errors of machine learning models with expert readers

The results show that the model provides a reliable temporal sorting of scans compared to experts. The model was able to correctly sort 95 out of 114 CT scan pairs using the leave-one-patient-out method, and the accuracy was compared to two radiology experts in Table 6.2. The model's ability to correctly sort the scans highlights its potential as a valuable tool in the diagnosis and treatment of IPF.

6.7 Local tissue transition pathway

The evaluation performed in this section focuses on the transition probabilities between different patterns of lung tissue observed in two timepoints (Figure 6.8) of the radiological scans throughout the evolution of a radiological illness. A network of these transition probabilities was found through an exploratory investigation of progression paths. The data used in this analysis includes multiple scans of patients with radiological illness, and the patterns observed in these scans serve as the reference for determining the likelihood of transitions between different patterns. The measurement performed in this section is a qualitative evaluation of the transition probabilities between the different patterns of lung tissue. Three types of patterns were identified by the latent transition network: stable patterns, volatile patterns, and transitory patterns. The stability or likelihood of transitions between these patterns is determined through the use of a simulation of particles undergoing a random walk. The results of the qualitative evaluation are presented in two figures, Figure 6.9 and Figure 6.10. Figure 6.9 presents the transition probabilities between the different patterns of lung tissue, with stable patterns, volatile patterns, and transitory patterns indicated. Figure 6.10 presents two visualizations of potential pathways for the evolution of disease patterns along the transition network. The first visualization launches particles in cluster 9 and monitors their network transitions through 10 patterns, with the most prevalent routes being 9 - 10 - 13 - 17 or 9 - 10 - 13 - 7. The second visualization requests paths that terminate in cluster 17, with the predominant sources being pathways 10 - 13 - 17 or 10 - 13 - 7 - 17, indicating that 7 is a likely intermediary step before 17.

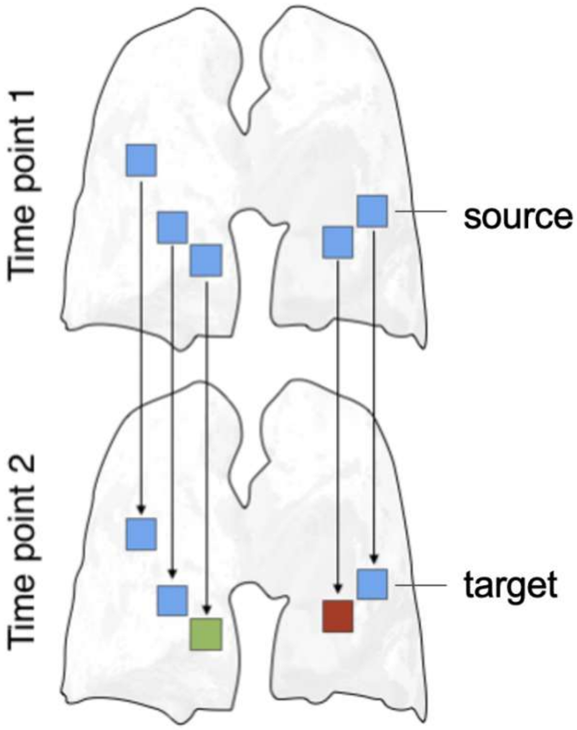


Figure 6.8: From the population of spatially matched follow-up pairs of lungs, we can observe local change of lung tissue from one to another pattern.

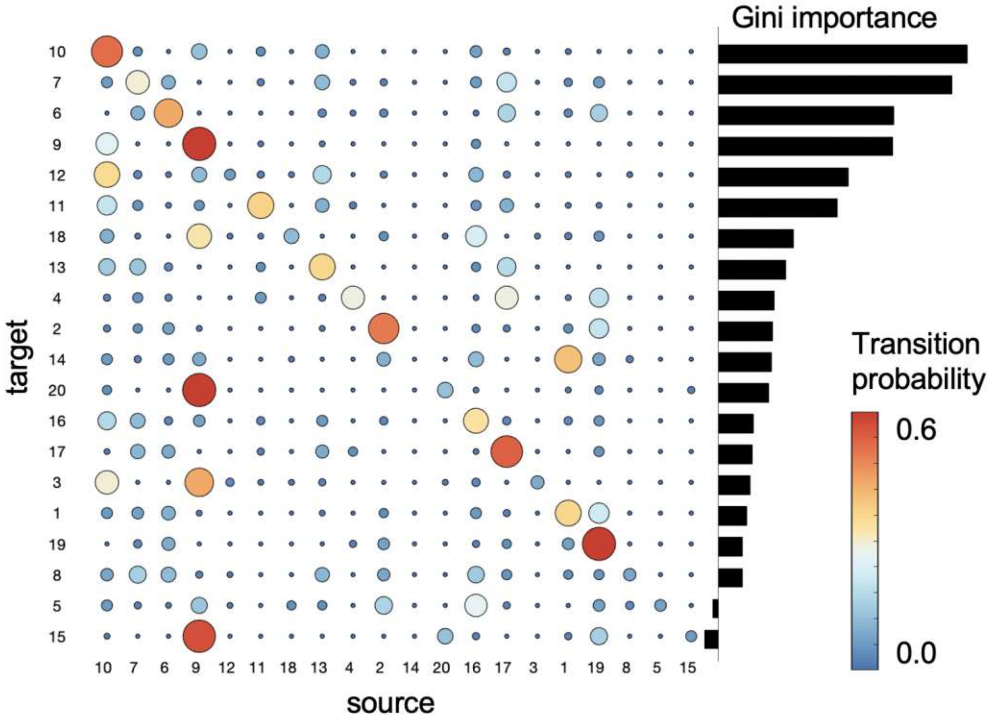
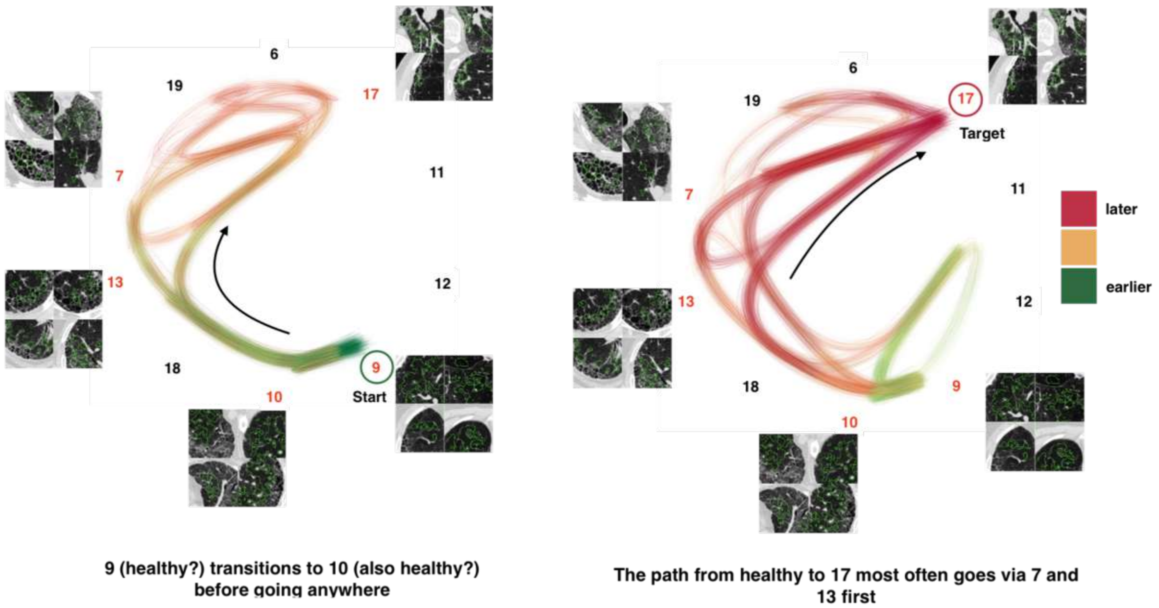


Figure 6.9: This enables obtaining a network of transition probabilities of lung patterns changing to others from one to the next examination time point. The matrix shows how likely a source pattern transitions to a target pattern. Red indicates high probability, blue low probability. These probabilities are generated by an underlying latent transition network that exhibits transition pathways shown in this figure. For the top ranked most informative patterns we plot two pathways to illustrate this model.



9 (healthy?) transitions to 10 (also healthy?) before going anywhere

The path from healthy to 17 most often goes via 7 and 13 first

Figure 6.10: (Right) Pathways originating from a healthy pattern (Cluster 9), and (left) pathways ending in vessels and ground glass pattern (Cluster 17). Arrows point at dominant directions in the graph.

6.8 Risk prediction

The study evaluated the radiological illness progression signature using the top four components (11-7-10-17) and their evolution over two scans. The patients were split into two groups based on disease progression signs using k-means clustering. The overall survival rate for each cluster of patients was determined using Kaplan-Meier analysis[91].

The results showed that clustering individuals on the basis of their radiological illness progression profile yielded two patient groups with distinct outcomes. The Kaplan-Meier analysis was based on the four static disease progression clusters and yielded a hazard ratio of 3.56 (p<.01). By adding the dynamic components (the variation between scans), the hazard ratio increased to 4.14 (p<.01). In the replication cohort, using the same progression signatures and clusters, the static components and the whole progression signature provided hazard ratios of 1.10 and 1.44,(same trend as in study cohort, but not significant), respectively. However, these results were not significant. The training was only done using the study cohort, and no re-training of the cluster patterns was performed on the replication cohort.

The results are presented in Figure 6.11.

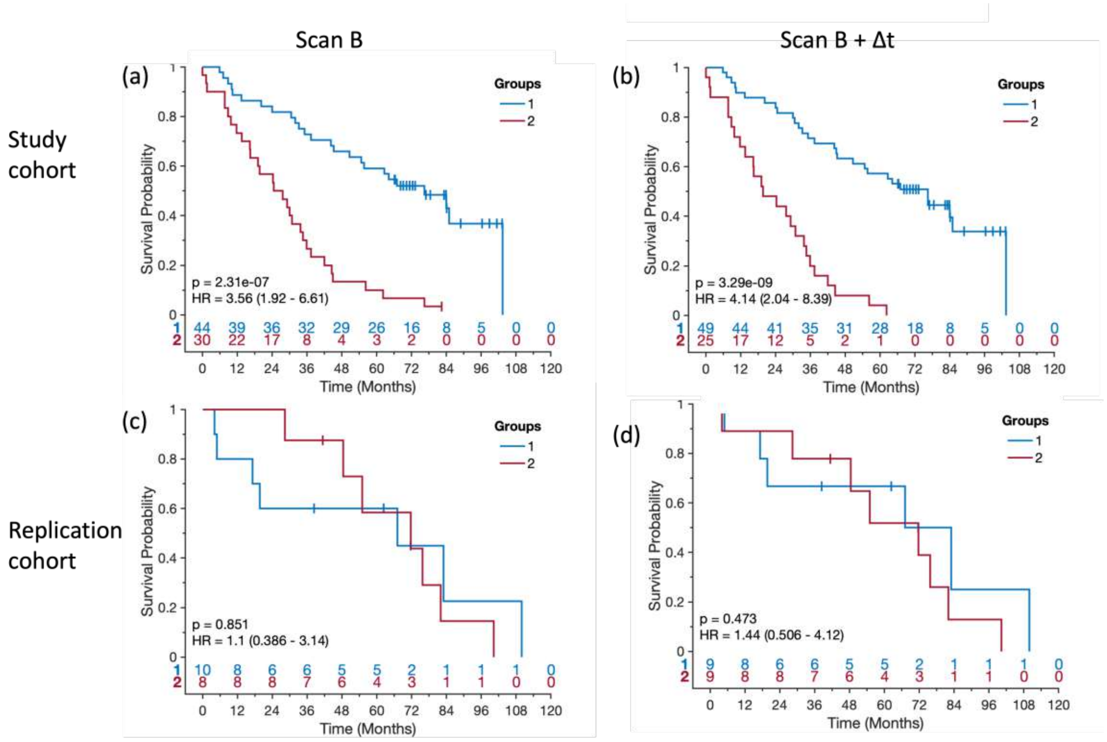


Figure 6.11: The survival study of Kaplan-Meier (KM) estimation of the most informative progression markers. a The KM curve based on markers of the scan B on the study cohort. b The KM curve based on markers of the scan B and the difference of the scan A and B on the study cohort. c The KM curve based on markers of the scan B on the replication cohort. d The KM curve based on markers of the scan B and the difference of the scan A and B on the replication cohort

6.9 Discussion

In this thesis, a study was conducted on patients diagnosed with IPF using HRCT scans. The study dataset was retrospectively retrieved from two referral centers in Italy and Austria, containing 106 and 18 patients respectively, with diagnoses made by multi-disciplinary ILD boards. CT scans were taken using GE Healthcare's Lightspeed Pro 16 and BrightSpeed 16 in Italy and a Siemens Sensation Cardiac 64 in Austria, with a slice thickness of 1.25-3.75 mm. The data was transformed into isotropic voxels with a resolution of $0.7mmx0.7mmx0.7mm$. A lung mask was created using a threshold-based method (-700HU) or a multi-template atlas-based segmentation approach. Over-segmentation was done using the mono-SLIC superpixel algorithm to create supervoxels of size $0.5cm^3$. For each supervoxel, a 56-dimensional feature was extracted using Haralick and SIFT features and reduced to 9 dimensions using PCA. The optimal number of clusters for k-means clustering was determined using the Jaccard score, and the final number of clusters was 20. The bag of visual words was created using the clustered supervoxels, and a machine learning classifier was used to identify markers of disease progression. Random forest classification model were used to identify the most informative radiological progression marker candidates and achieved an average accuracy of 83% in correctly predicting the sequence of CT scan pairs. The study also evaluated the transition probabilities between different patterns of lung tissue and found stable, volatile, and transitory patterns. The study found that clustering patients based on their radiological illness progression profile yielded two groups with distinct outcomes. The hazard ratio was 3.56 for the static components and 4.14 for the dynamic components. The results were not significant in the replication cohort.

CHAPTER 7

Conclusion

This chapter provides a summary of the main points of this thesis. Additionally, suggestions on prospective future work are provided.

7.1 Summary

The main objective of this thesis is to demonstrate the potential of unsupervised machine learning in identifying CT patterns related to the progression of Interstitial Pulmonary Fibrosis disease. The study aims to address three crucial questions regarding the use of machine learning for IPF diagnosis.

The first question is about identifying new disease progression markers in IPF patients beyond the known ones such as ground glass opacities and honeycombing. This is significant because existing markers have limited power in diagnosing IPF, and data-driven models have the potential to expand the marker patterns, thereby improving the accuracy and reliability of IPF diagnosis through imaging. The results from this study indicate that it is possible to discover additional, reliable disease progression markers.

The second question is about visualizing the IPF disease progression transition pathway. The extracted texture pattern signatures provide a visual representation of the disease's transition pathway, which can be used to understand the changes in image features that occur at different stages of the disease. Further investigation of these biologically and pathologically meaningful hypotheses might reveal the histological changes that occur during the disease's progression.

The final question concerns the relationship between different radiological disease progression patterns and future survival outcomes. The study finds that patients with similar pattern signatures have similar survival outcomes, and the reliability of this statement is confirmed through the external validation set. The results of this study highlight the potential of machine learning in identifying and analyzing disease progression patterns and predicting survival outcomes in IPF patients.

7.2 Future Work

The results of this thesis indicate that unsupervised machine learning has the potential to be a valuable tool for the diagnosis and prediction of IPF progression. However, there is much room for future improvement and expansion.

One area for future work is to expand the dataset used for testing. By using a larger dataset with a larger number of patients and scans, the reliability of the results can be further strengthened. Additionally, it would be interesting to test the method on other interstitial lung diseases beyond IPF, to see if the results can be generalized to other diseases. Given the current COVID-19 pandemic, it is also of great interest to explore if these methods can be applied to predict patient outcomes in COVID-19 patients.

Another avenue for improvement is to involve expert radiologists in the evaluation of the image segmentation results. This could involve having the radiologists annotate a subset of the CT scans, which could then be used to evaluate and refine the accuracy of the segmentation. Additionally, the use of K-means clustering is based on Euclidean distance, which is a heuristic metric. Future work could explore the use of alternative metrics for clustering, which may result in improved accuracy and clustering results.

Overall, the results of this thesis provide promising initial results for the use of unsupervised machine learning for the diagnosis and prediction of IPF progression. However, there is much room for further exploration and improvement, and the results of this thesis provide a foundation for future research in this area.

Bibliography

[1] G. I. Sgalla, A. L. Biffi, and L. U. C. A. Richeldi, "INVITED REVIEW SERIES : IDIOPATHIC INTERSTITIAL PNEUMONIA — PART 2 : SPECIFIC DISEASE ENTITIES Idiopathic pulmonary fibrosis : Diagnosis , epidemiology and," *Respirology*, vol. 21, no. November 2015, pp. 427–437, 2016.

[2] W. D. Travis, U. Costabel, D. M. Hansell, T. E. King, D. A. Lynch, A. G. Nicholson, A. U. Wells, J. Behr, D. Bouros, C. J. Ryerson, J. H. Ryu, K. K. Brown, T. V. Colby, H. R. Collard, C. R. Cordeiro, V. Cottin, B. Crestani, M. Drent, R. F. Dudden, J. Egan, K. Flaherty, C. Hogaboam, Y. Inoue, T. Johkoh, D. S. Kim, M. Kitaichi, J. Loyd, F. J. Martinez, J. Myers, S. Protzko, G. Raghu, L. Richeldi, N. Sverzellati, J. Swigris, D. Valeyre, and A. T. S. Ers, "American Thoracic Society Documents An Official American Thoracic Society / European Respiratory Society Statement : Update of the International Multidisciplinary Classification of the Idiopathic Interstitial Pneumonias," vol. 188, pp. 733–748, 2013.

[3] P. Spagnolo, T. M. Maher, and L. Richeldi, "Idiopathic pulmonary fibrosis: Recent advances on pharmacological therapy," *Pharmacology and Therapeutics*, vol. 152, pp. 18–27, 2015.

[4] V. Navaratnam, K. M. Fleming, J. West, C. J. P. Smith, R. G. Jenkins, A. Fogarty, and R. B. Hubbard, "The rising incidence of idiopathic pulmonary fibrosis in the UK," *Thorax*, vol. 66, pp. 462–467, 2011.

[5] G. Raghu, M. Remy-Jardin, L. Richeldi, C. C. Thomson, K. M. Antoniou, B. D. Bissell, D. Bouros, I. Buendia-Roldan, F. Caro, B. Crestani, T. Ewing, M. Ghazipura, D. D. Herman, L. Ho, S. M. Hon, T. Hossain, Y. Inoue, T. Johkoh, S. Jones, F. Kheir, Y. H. Khor, S. L. Knight, M. Kreuter, D. A. Lynch, M. Macrea, T. M. Maher, M. J. Mammen, F. J. Martinez, M. Molina-Molina, J. Morisset, J. L. Myers, A. G. Nicholson, A. L. Olson, A. Podolanczuk, V. Poletti, C. J. Ryerson, M. Selman, M. E. Strek, L. K. Troy, M. Wijsenbeek, and K. C. Wilson, "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*, vol. 205, no. 9, pp. E18–E47, 2022.

[6] E. R. Fernández Pérez, C. E. Daniels, J. St. Sauver, T. E. Hartman, B. J. Bartholmai, E. S. Yi, J. H. Ryu, and D. R. Schroeder, "Incidence, prevalence, and clinical course of idiopathic pulmonary fibrosis: A population-based study," *Chest*, vol. 137, no. 1, pp. 129–137, 2010.

[7] C. J. Ryerson, T. H. Urbania, L. Richeldi, J. J. Mooney, J. S. Lee, K. D. Jones, B. M. Elicker, L. L. Koth, T. E. King, P. J. Wolters, and H. R. Collard, "Prevalence and prognosis of unclassifiable interstitial lung disease," *European Respiratory Journal*, vol. 42, no. 3, pp. 750–757, 2013.

[8] G. Raghu, S.-Y. Chen, W.-S. Yeh, B. Maroni, Q. Li, Y.-C. Lee, and H. R. Collard, "Idiopathic pulmonary fibrosis in us medicare beneficiaries aged 65 years and older: incidence, prevalence, and survival, 2001–11," *The Lancet Respiratory Medicine*, vol. 2, no. 7, pp. 566–572, 2014.

[9] R. S. European, A. T. Society, *et al.*, "American thoracic society/european respiratory society international multidisciplinary consensus classification of the idiopathic interstitial pneumonias. this joint statement of the american thoracic society (ats), and the european respiratory society (ers) was adopted by the ats board of directors, june 2001 and by the ers executive committee, june 2001.," *American Journal of Respiratory and Critical Care Medicine*, vol. 165, no. 2, p. 277, 2002.

[10] A. L. Olson and J. J. Swigris, "Idiopathic pulmonary fibrosis: diagnosis and epidemiology," *Clinics in chest medicine*, vol. 33, no. 1, pp. 41–50, 2012.

[11] S. Harari, F. Madotto, A. Caminati, and S. Conti, "Epidemiology of Idiopathic Pulmonary Fibrosis in Northern Italy," pp. 1–15, 2016.

[12] N. Agabiti, M. A. Porretta, L. Bauleo, A. Coppola, G. Sergiacomi, A. Fusco, F. Cavalli, M. C. Zappa, R. Vignarola, S. Carlone, *et al.*, "Idiopathic pulmonary fibrosis (ipf) incidence and prevalence in italy," *Sarcoidosis vasculitis and diffuse lung disease*, vol. 31, no. 3, pp. 191–197, 2014.

[13] K. B. Baumgartner, J. M. Samet, C. A. Stidley, T. V. Colby, J. A. Waldron, and C. Centers, "Cigarette Smokinv : A Risk Fador for Idiopathic Pulmonary FibrosIs," *Am J Respir Crit Care Med*, vol. 155, no. 242-248, 1997.

[14] K. Flaherty, G. Toews, W. Travis, T. V. Colby, E. Kazerooni, B. Gross, A. Jain, R. Strawderman, R. Paine, A. Flint, *et al.*, "Clinical significance of histological classification of idiopathic interstitial pneumonia," *European Respiratory Journal*, vol. 19, no. 2, pp. 275–283, 2002.

[15] T. E. King, M. I. Schwarz, K. Brown, J. A. Tooze, T. V. Colby, J. A. Waldron, A. Flint, W. Thurlbeck, and R. M. Cherniack, "Relationship between Histopathologic Features and Mortality," *Am J Respir Crit Care Med*, vol. 164, no. 11, pp. 1025–1032, 2001.

[16] A. G. Nicholson, T. V. Colby, R. M. Dubois, D. M. Hansell, and A. U. Wells, “The prognostic significance of the histologic pattern of interstitial pneumonia in patients presenting with the clinical entity of cryptogenic fibrosing alveolitis,” *American Journal of Respiratory and Critical Care Medicine*, vol. 162, no. 6, pp. 2213–2217, 2000.

[17] V. S. Taskar and D. B. Coultas, “Is Idiopathic Pulmonary Fibrosis an Environmental Disease ? FACTORS LIMITING RECOGNITION OF,” *Proc Am Thorac Soc*, vol. 3, no. 293-298, 2006.

[18] R. Hubbard, S. Lewis, K. Richards, I. Johnston, and J. Britton, “Occupational exposure to metal cryptogenic fibrosing alveolitis aetiology of,” *Lancet 1996;*, vol. 347, pp. 284–289, 1996.

[19] J. Grutters and R. Du Bois, “Genetics of fibrosing lung diseases,” *European Respiratory Journal*, vol. 25, no. 5, pp. 915–927, 2005.

[20] A. Q. Thomas, K. Lane, J. P. Iii, M. Prince, C. Markin, M. Speer, D. A. Schwartz, R. Gaddipati, A. Marney, J. Johnson, R. Roberts, J. Haines, M. Stahlman, and J. E. Loyd, “Heterozygosity for a Surfactant Protein C Gene Mutation Associated with Usual Interstitial Pneumonitis and Cellular Nonspecific Interstitial Pneumonitis in One Kindred,” vol. 165, no. 18, pp. 1322–1328, 2002.

[21] M. A. Seibold, A. L. Wise, M. C. Speer, J. E. Loyd, T. E. Fingerlin, D. Ph, W. Zhang, D. Ph, K. B. Adler, D. Ph, B. F. Dickey, R. M. Bois, I. V. Yang, D. Ph, A. Herron, D. Kervitsky, J. L. Talbert, C. Markin, J. Park, A. L. Crews, S. H. Slifer, D. Ph, S. Auerbach, D. Ph, M. G. Roy, J. Lin, C. E. Hennessy, M. I. Schwarz, and D. A. Schwartz, “A Common,” *The New England Journal of Medicine*, vol. 364, pp. 1503–1512, 2011.

[22] B. Burrows and A. Johnson, “Cryptogenic fibrosing alveolitis : clinical features and their influence on survival,” *Thorax*, vol. 35, pp. 171–180, 1980.

[23] H. Schünemann, R. Jaeschke, D. J. Cook, W. F. Bria, A. A. El-solh, A. Ernst, B. F. Fahy, M. K. Gould, K. L. Horan, J. A. Krishnan, C. A. Manthous, J. R. Maurer, W. T. Mcnicholas, A. D. Oxman, G. Rubenfeld, and G. M. Turino, “American Thoracic Society Documents An Official ATS Statement : Grading the Quality of Evidence and Strength of Recommendations in ATS Guidelines and Recommendations,” *Am J Respir Crit Care Med*, vol. 174, pp. 605–614, 2006.

[24] S. N. Iyer, G. Gurujeyalakshmi, and S. N. Giri, “Effects of Pirfenidone on Transforming Growth Factor- Gene Expression at the Transcriptional Level in Bleomycin Hamster Model of Lung Fibrosis 1,” *THE JOURNAL OF PHARMACOLOGY AND EXPERIMENTAL THERAPEUTICS*, vol. 291, no. 1, pp. 367–373, 1999.

[25] G. J. Roth, R. Binder, F. Colbatzky, C. Dallinger, R. Schlenker-herceg, F. Hilberg, S.-l. Wollin, and R. Kaiser, "Nintedanib : From Discovery to the Clinic," *Journal of Medicinal Chemistry*, vol. 58, p. 10531063, 2015.

[26] L. Wollin, I. Maillet, V. Quesniaux, A. Holweg, and B. Ryffel, "Antifibrotic and Anti-inflammatory Activity of the Tyrosine Kinase Inhibitor Nintedanib in Experimental Models of Lung Fibrosis s," *THE JOURNAL OF PHARMACOLOGY AND EXPERIMENTAL*, no. 349, pp. 209–220, 2014.

[27] F. Hilberg, G. J. Roth, M. Krssak, S. Kautschitsch, W. Sommergruber, U. Tontsch-Grunt, P. Garin-Chesa, G. Bader, A. Zoephel, J. Quant, *et al.*, "Bibf 1120: triple angiokinase inhibitor with sustained receptor blockade and good antitumor efficacy," *Cancer research*, vol. 68, no. 12, pp. 4774–4782, 2008.

[28] G. M. Keating, "Nintedanib : A Review of Its Use in Patients with Idiopathic Pulmonary Fibrosis," *Drugs*, 2015.

[29] J. Jacob, B. Bartholmai, S. Rajagopalan, C. Van Moorsel, H. Es, F. van Beek, M. Struik, M. Kokosi, R. Egashira, A. Brun, A. Nair, S. Walsh, G. Cross, J. Barnett, A. de lauretis, E. Judge, S. Desai, R. Karwoski, S. Ourselin, and A. Wells, "Predicting outcomes in idiopathic pulmonary fibrosis using automated ct analysis," *American Journal of Respiratory and Critical Care Medicine*, vol. 198, 04 2018.

[30] J. Jacob, L. Aksman, N. Mogulkoc, A. Proctor, B. Gholipour, G. Cross, J. Barnett, C. Brereton, M. Jones, C. Van Moorsel, W. van Es, F. van Beek, M. Veltkamp, S. Desai, E. Judge, T. Burd, M. Kokosi, R. Savas, S. Bayraktaroglu, and A. Wells, "Serial ct analysis in idiopathic pulmonary fibrosis: Comparison of visual features that determine patient outcome," *Thorax*, vol. 75, pp. thoraxjnl–2019, 04 2020.

[31] H. Robbie, C. Daccord, F. Chua, and A. Devaraj, "Evaluating disease severity in idiopathic pulmonary fibrosis," *European Respiratory Review*, vol. 26, no. 145, 2017.

[32] H. Koyama, Y. Ohno, A. Kono, D. Takenaka, Y. Maniwa, Y. Nishimura, C. Ohbayashi, and K. Sugimura, "Quantitative and qualitative assessment of non-contrast-enhanced pulmonary mr imaging for management of pulmonary nodules in 161 subjects," *European radiology*, vol. 18, pp. 2120–31, 05 2008.

[33] M. Jafar, "Mri artifacts: Radiology reference article," Oct 2021.

[34] Radiopaedia.org, "Case courtesy of usman bashir." [Online; accessed Feb., 2023].

[35] J. Biederer, S. Mirsadraee, M. Beer, F. Molinari, C. Hintze, G. Bauman, M. Both, E. Beek, J. Wild, and M. Puderbach, "Mri of the lung (3/3)—current applications and future perspectives," *Insights into imaging*, vol. 3, pp. 373–86, 01 2012.

[36] G. Lutterbey, C. Grohé, J. Gieseke, M. Falkenhausen, N. Morakkabati, M. Wattjes, R. Manka, D. Trog, and H. Schild, "Initial experience with lung-mri at 3.0t:

Comparison with ct and clinical data in the evaluation of interstitial lung disease activity," *European journal of radiology*, vol. 61, pp. 256–61, 03 2007.

[37] H. Itoh, S. Tokunaga, H. Asamoto, M. Furuta, Y. Funamoto, M. Kitaichi, and K. Torizuka, "Radiologic-pathologic correlations of small lung nodules with special reference to peribronchiolar nodules," *American Journal of Roentgenology*, vol. 130, no. 2, pp. 223–231, 1978.

[38] T. G, I. H, N. Y, D. Y, M. H, M. K, O. T, T. K, I. T, and O. S, "High resolution CT (HR-CT) for the evaluation of pulmonary peripheral disorders," *Rinsho Hoshasen*, vol. 27, no. 12, pp. 1319–1326, 1982.

[39] R. Smithuis, O. Delden, M. Hazewinkel, and J. Bradshaw, "Radiology Assistant," 2017.

[40] E. A. Kazerooni, "High-resolution {CT} of the lungs," *American Journal of Roentgenology*, vol. 177, no. September, pp. 501–519, 2001.

[41] S. Abe and H. Takahashi, "Diffuse lung diseases," *Nihon Naika Gakkai zasshi. The Journal of the Japanese Society of Internal Medicine*, vol. 89, no. 9, pp. 1804–1808, 2000.

[42] S. Aquino, M. Schechter, C. Chiles, D. Ablin, B. Chipps, and R. Webb, "High-resolution inspiratory and expiratory CT in older children and adults with bronchopulmonary dysplasia.," *American Journal of Roentgenology*, vol. 173, no. October, pp. 963–967, 1999.

[43] R. Popilock, K. Sandrasagaren, L. Harris, and K. A. Kaser, "Ct artifact recognition for the nuclear technologist," *Journal of Nuclear Medicine Technology*, vol. 36, pp. 79 – 81, 2008.

[44] L. Breiman, "RANDOM FORESTS," in *Machine Learning*, vol. 45, pp. 5–32, Kluwer Academic Publishers, 5 ed., Oct. 2001.

[45] Z. A. Aziz, S. P. Padley, and D. M. Hansell, "CT techniques for imaging the lung: Recommendations for multislice and single slice computed tomography," *European Journal of Radiology*, vol. 52, no. 2, pp. 119–136, 2004.

[46] R. Shouval, O. Bondi, H. Mishan, a. Shimoni, R. Unger, and a. Nagler, "Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in SCT.," *Bone marrow transplantation*, vol. 49, no. 3, pp. 332–7, 2014.

[47] E. Alpaydin, *Introduction to Machine Learning, fourth edition.* Adaptive Computation and Machine Learning series, MIT Press, 2020.

[48] N. Sakthivel, B. B. Nair, M. Elangovan, V. Sugumaran, and S. Saravanmurugan, "Comparison of dimensionality reduction techniques for the fault diagnosis of mono block centrifugal pump using vibration signals," *Engineering Science and Technology, an International Journal*, vol. 17, no. 1, pp. 30–38, 2014.

[49] W. Xue, "Classification of pulmonary lesions based on cnn and chest x-ray images," *Journal of Physics: Conference Series*, vol. 1952, p. 022025, 06 2021.

[50] E. Kaznowska, J. Depciuch, K. Łach, M. Kołodziej, A. Koziorowska, J. Vongsvivut, I. Zawlik, M. Cholewa, and J. Cebulski, "The classification of lung cancers and their degree of malignancy by ftir, pca-lda analysis, and a physics-based computational model," *Talanta*, vol. 186, pp. 337–345, 2018.

[51] L. K. Hansen, J. Larsen, F. Årup Nielsen, S. C. Strother, E. Rostrup, R. Savoy, N. Lange, J. Sidtis, C. Svarer, and O. B. Paulson, "Generalizable patterns in neuroimaging: How many principal components?," *NeuroImage*, vol. 9, no. 5, pp. 534–544, 1999.

[52] T. Takekawa, Y. Isomura, and T. Fukai, "Spike sorting of heterogeneous neuron types by multimodality-weighted pca and explicit robust variational bayes," *Frontiers in neuroinformatics*, vol. 6, p. 5, 03 2012.

[53] L. Paul and A. Suman, "Face recognition using principal component analysis method," *International Journal of Advanced Research in Computer Engineering Technology (IJARCET)*, vol. 1, pp. 135–139, 11 2012.

[54] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America. A, Optics and image science*, vol. 4, pp. 519–24, 04 1987.

[55] J. MacQueen, "Some methods for classification and analysis of multivariate observations," 1967.

[56] N. Mansoori, M. Nejati, P. Razzaghi, and S. Samavi, "Bag of visual words approach for image retrieval using color information," pp. 1–6, 05 2013.

[57] R. M. Haralick, K. Shanmugam, and I. Dinstein, "TexturalFeatures.pdf," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[58] G. Duan, J. Yang, and Y. Yang, "Content-based image retrieval research," *Physics Procedia*, vol. 22, pp. 471–477, 2011.

[59] Y. Zhuge and J. Udupa, "Intensity standardization simplifies brain mr image segmentation," *Computer vision and image understanding : CVIU*, vol. 113, pp. 1095–1103, 10 2009.

[60] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[61] D. Engel and C. Curio, “Scale-invariant medial features based on gradient vector flow fields,” in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1–4, IEEE, 2008.

[62] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” *International journal of computer vision*, vol. 59, no. 2, pp. 167–181, 2004.

[63] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (2nd Ed)*. Prentice Hall, 2002.

[64] M. Holzer and R. Donner, “Over-Segmentation of 3D Medical Image Volumes based on Monogenic Cues,” tech. rep., 2014.

[65] M. Felsberg and G. Sommer, “The monogenic signal,” *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 3136–3144, 2001.

[66] L. Cohen, “Time-frequency distributions-a review,” *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, 1989.

[67] R. Moreno-Díaz and A. Moreno-Díaz, “On the legacy of w.s. mcculloch,” *Bio Systems*, vol. 88 3, pp. 185–90, 2007.

[68] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[69] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[70] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014.

[71] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.

[72] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” 2015.

[73] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein gan,” 2017. cite arxiv:1701.07875.

[74] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020.

[75] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” 2018.

[76] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” 2018.

[77] H. J. Kim, M. S. Brown, D. Chong, D. W. Gjertson, P. Lu, H. J. Kim, H. Coy, and J. G. Goldin, “Comparison of the quantitative ct imaging biomarkers of idiopathic pulmonary fibrosis at baseline and early change with an interval of 7 months,” *Academic Radiology*, vol. 22, no. 1, pp. 70–80, 2015.

[78] H. Park, S. M. Lee, J. Song, S. Oh, N. Kim, and J. B. Seo, “Texture-based automated quantitative assessment of regional patterns on initial ct in patients with idiopathic pulmonary fibrosis: Relationship to decline in forced vital capacity,” *American Journal of Roentgenology*, vol. 207, pp. 1–8, 08 2016.

[79] G. Raghu, H. R. Collard, J. J. Egan, F. J. Martinez, J. Behr, K. K. Brown, T. V. Colby, J.-F. Cordier, K. R. Flaherty, J. A. Lasky, D. A. Lynch, J. H. Ryu, J. J. Swigris, A. U. Wells, J. Ancochea, D. Bouros, C. Carvalho, U. Costabel, M. Ebina, D. M. Hansell, T. Johkoh, D. S. Kim, T. E. King, Y. Kondoh, J. Myers, N. L. Müller, A. G. Nicholson, L. Richeldi, M. Selman, R. F. Dudden, B. S. Griss, S. L. Protzko, and H. J. Schünemann, “An Official ATS/ERS/JRS/ALAT Statement: Idiopathic Pulmonary Fibrosis: Evidence-based Guidelines for Diagnosis and Management,” *American Journal of Respiratory and Critical Care Medicine*, vol. 183, pp. 788–824, Mar. 2011.

[80] B. Bartholmai, S. Raghunath, R. Karwoski, T. Moua, S. Rajagopalan, F. Maldonado, P. Decker, and R. Robb, “Quantitative computed tomography imaging of interstitial lung diseases,” *Journal of Thoracic Imaging*, vol. 28, pp. 298–307, Sept. 2013.

[81] S. Humphries, K. Yagihashi, J. Huckleberry, B.-H. Rho, J. Schroeder, M. Strand, M. Ischwarz, K. Flaherty, E. Kazerooni, E. Beek, and D. Lynch, “Idiopathic pulmonary fibrosis: Data-driven textural analysis of extent of fibrosis at baseline and 15-month follow-up,” *Radiology*, vol. 285, p. 161177, 05 2017.

[82] W. D. Vogl, H. Prosch, C. Müller-Mang, U. Schmidt-Erfurth, and G. Langs, “Longitudinal alignment of disease progression in fibrosing interstitial lung disease,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8674 LNCS, no. PART 2, pp. 97–104, 2014.

[83] J. Lee and A. P. Reeves, “Segmentation of the airway tree from chest ct using local volume of interest,” 2009.

[84] R. Pinho, V. Delmon, J. Vandemeulebroucke, S. Rit, and D. Sarrut, “Keuhkot : A method for lung segmentation,” 2011.

[85] J. Hofmanninger, M. Krenn, M. Holzer, T. Schlegl, H. Prosch, and G. Langs, “Unsupervised identification of clinically relevant clusters in routine imaging data,”

in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9900, pp. 192–200, 2016.

[86] O. Goksel, A. Foncubierta-Rodríguez, O. A. J. del Toro, and y. Henning Müller and Georg Langs and Marc-André Weber and Bjoern H. Menze and Ivan Eggel and Katharina Gruenberg and Marianne Winterstein and Markus Holzer and Markus Krenn and Georgios Kontokotsios and Sokratis Metallidis and Roger Schaer and Abdel Aziz Taha and András Jakab and Tomas Salas Fernandez and Allan Hanbury, booktitle=VISCERAL Challenge@ISBI, "Overview of the visceral challenge at isbi 2015,"

[87] B. Avants, N. Tustison, and G. Song, "Advanced normalization tools (ants)," *Insight J*, vol. 1–35, 11 2008.

[88] A. Gruslys, J. Acosta-Cabronero, P. Nestor, G. Williams, and R. Ansorge, "A new fast accurate non-linear medical image registration program including surface preserving regularisation.," *IEEE transactions on medical imaging*, vol. 33, 06 2014.

[89] J. Pan, J. Hofmanninger, K.-H. Nenning, F. Prayer, S. Röhrich, N. Sverzellati, V. Poletti, S. Tomassetti, M. Weber, H. Prosch, and et al., "Unsupervised machine learning identifies predictive progression markers of ipf," *European Radiology*, 2022.

[90] P. Jaccard, "THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE," *New Phytologist*, vol. 11, pp. 37–50, Feb. 1912.

[91] M. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-meier estimate," *International journal of Ayurveda research*, vol. 1, pp. 274–8, 10 2010.