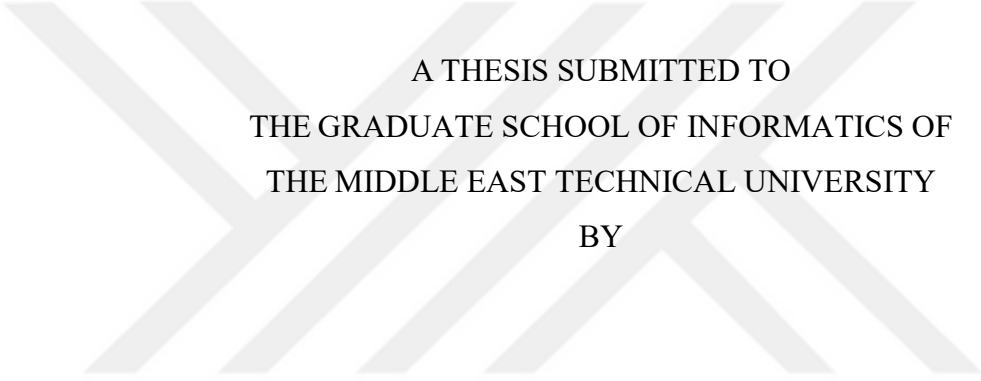


MULTIMODAL DATA FUSION AND MULTICOMPARTMENT IMAGE
ANALYSIS IN ACUTE AND CHRONIC LUNG DISEASES



A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

AHMET GÖRKEM ER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
MEDICAL INFORMATICS

JULY, 2025

**MULTIMODAL DATA FUSION AND MULTICOMPARTMENT IMAGE ANALYSIS IN
ACUTE AND CHRONIC LUNG DISEASES**

Submitted by **Ahmet Grkem Er** in partial fulfillment of the requirements for the degree of **Doctor
of Philosophy in Health Informatics Department, Middle East Technical University** by,



Date:

16.07.2025



I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Ahmet Görkem Er

Signature : _____

ABSTRACT

MULTIMODAL DATA FUSION AND MULTICOMPARTMENT IMAGE ANALYSIS IN ACUTE AND CHRONIC LUNG DISEASES

Er, Ahmet G rkem
Ph.D., Department of Health Informatics
Supervisor: Assoc. Prof. Dr. Yeřim Aydın Son

July 2025, 123 pages

This thesis investigates multimodal data fusion and multicompartment image analysis to enhance diagnostic modeling and clinical prediction in acute and chronic lung diseases. Two distinct patient cohorts were analyzed to address these objectives.

In the first part, we studied a cohort of COVID-19 patients by integrating imaging, clinical, laboratory, and viral genome sequencing data using sparse linear modeling techniques. To identify and quantify relationships across these diverse data modalities, we applied unsupervised sparse canonical correlation analysis, which revealed significant correlations between imaging data and laboratory results and enabled the identification of distinct clinical phenotypes. We also employed supervised cooperative learning to optimize multimodal data fusion for predicting intensive care unit admission. Furthermore, we innovatively leveraged the Word2Vec natural language processing model to encode viral genomes, an approach that not only distinguished major SARS-CoV-2 variants but also preserved phylogenetic relationships between viral strains, enabling the integration of genomic data into downstream clinical modeling.

In the second part, we focused on interstitial lung diseases, applying the concept of multicompartment image analysis by extracting and integrating lung parenchymal and pulmonary artery radiomics features from chest computed tomography scans. We demonstrated that pulmonary artery features are interpretable and hold predictive value

for pulmonary hypertension and transplant-free survival. Moreover, combining pulmonary artery and lung parenchymal features improved predictive performance for these outcomes.

Together, in these parts, we illustrate how multimodal data fusion and multicompartment image analysis can mirror clinical decision-making processes and improve personalized prognostication in complex lung diseases.

Keywords: Multimodal data fusion, multicompartment image analysis, COVID-19, interstitial lung disease, Word2Vec

ÖZ

AKUT VE KRONİK AKCİĞER HASTALIKLARINDA ÇOK MODLU VERİ FÜZYONU VE ÇOK KOMPARTİMANLI GÖRÜNTÜ ANALİZİ

Er, Ahmet Görkem

Doktora, Sağlık Bilişimi Bölümü

Tez Yöneticisi: Doç. Dr. Yeşim Aydın Son

Temmuz 2025, 123 sayfa

Bu tez, akut ve kronik akciğer hastalıklarında tanısal modelleme ve klinik tahmini geliştirmek için çok modlu veri füzyonu ve çok kompartımanlı görüntü analizini araştırmaktadır. Bu hedeflere ulaşmak için iki farklı hasta kohortu analiz edilmiştir.

İlk bölümde, COVID-19 hastalarından oluşan bir kohortu inceleyerek görüntüleme, klinik, laboratuvar ve viral genom dizileme verilerini seyrek doğrusal modelleme teknikleri kullanarak entegre ettik. Farklı veri modaliteleri arasındaki ilişkileri tanımlamak ve nicel olarak belirlemek amacıyla denetimsiz seyrek kanonik korelasyon analizi uyguladık. Bu analiz, görüntüleme verileri ile laboratuvar sonuçları arasında anlamlı korelasyonlar ortaya koymuş ve belirgin klinik fenotiplerin tanımlanmasını sağlamıştır. Yoğun bakım ünitesine kabulü öngörmek üzere çok modlu veri birleştirmeyi optimize etmek için denetimli iş birlikçi öğrenme yöntemini kullandık. Ayrıca, viral genomları kodlayabilmek için Word2Vec doğal dil işleme modelinden yenilikçi bir şekilde faydalandık. Bu yaklaşım yalnızca başlıca SARS-CoV-2 varyantlarını ayırt etmekle kalmamış, aynı zamanda viral suşlar arasındaki filogenetik ilişkilerin korunmasını sağlamış ve genomik verilerin klinik analizlere entegrasyonunu mümkün kılmıştır.

İkinci bölümde, interstisyel akciğer hastalıklarına odaklanarak, göğüs bilgisayarlı tomografi görüntülerinden akciğer parankimi ve pulmoner arterlere ait radyomik öznitelikleri çıkarıp entegre ederek çok kompartımanlı görüntü analizi kavramını uyguladık. Pulmoner arterden elde edilen radyomik özniteliklerin yorumlanabilir

olduğunu ve pulmoner hipertansiyon ile transplantasyonsuz sağkalım açısından öngörü değeri taşıdığını gösterdik. Ayrıca, pulmoner arter ve akciğer parankim öz niteliklerinin bir arada kullanılması bu klinik sonuçlara yönelik tahmin performansını artırdı.

Birlikte ele alındığında, bu bölümlerde, çok modlu veri füzyonu ve çok kompartımanlı görüntü analizinin klinik karar verme süreçlerini yansıtabileceğini ve karmaşık akciğer hastalıklarında kişiselleştirilmiş prognoz tahminini iyileştirebileceğini ortaya koyuyoruz.

Anahtar Sözcükler: Çok modlu veri füzyonu, çok kompartımanlı görüntü analizi, COVID-19, interstisyel akciğer hastalığı, Word2Vec



To My Beloved Family,
Ekin, Evrim, and Berrin

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Assoc. Prof. Yeşim Aydın Son, for her guidance, and confidence in me throughout my doctoral journey. From the earliest stages of this project to its completion, her encouragement and insightful feedback have been invaluable. I am truly grateful for her belief in my work and her dedication to my academic progress.

I am also sincerely thankful to Prof. Serhat Ünal for his vision and support throughout my doctoral studies and medical training. His mentorship has played a pivotal role in shaping both my scientific perspective and professional development. I also extend my appreciation to all the attendings in the Department of Clinical Microbiology and Infectious Diseases at Hacettepe University Faculty of Medicine for their encouragement and support.

I would like to express my sincere thanks to Prof. Mine Durusu Tanrıöver for her invaluable contributions not only to my education in internal medicine but also throughout the course of this project.

I would also like to thank Prof. Hakan S. Orer for his encouragement and mentorship during my PhD journey. His support has been a valuable source of motivation throughout this process.

My sincere appreciation goes to the Turkish Fulbright Commission, whose generous PhD Dissertation Research Grant allowed me to conduct part of my research internationally and significantly broaden its impact.

I am also deeply thankful to my scientific advisors at Stanford University, Prof. Olivier Gevaert and Prof. Rishi Raj, for their invaluable support and collaboration. Their guidance and the resources they provided, including access to patient cohorts and institutional infrastructure, made it possible to expand the scope and depth of this work.

Finally, beyond academia, I owe my deepest gratitude to my family, my children, and my wife, my mother, father, and sister, for their unconditional love, patience, and support. Their belief in me has been the foundation of this journey.

TABLE OF CONTENTS

ABSTRACT	iv
ÖZ.....	vi
DEDICATION	viii
ACKNOWLEDGMENTS.....	ix
TABLE OF CONTENTS	x
LIST OF TABLES	xiii
LIST OF FIGURES.....	xiv
LIST OF ABBREVIATIONS	xvi
CHAPTERS	
1. INTRODUCTION.....	1
2. BACKGROUND.....	7
2.1. COVID-19 Infection.....	7
2.1.1. Word2Vec NLP Model	8
2.1.2. Multimodal Data Fusion.....	9
2.1.3. Sparse Canonical Correlation Analysis.....	10
2.1.4. Cooperative Learning.....	11
2.1.5. The Games-Howell Test.....	11
2.2. Interstitial Lung Diseases	12
2.2.1. Epidemiology	12
2.2.2. Pulmonary Hypertension in ILD: Definition and Hemodynamic Criteria.....	13
2.2.3. Cox Proportional Hazards Model.....	16
2.2.4. Wilcoxon Signed-Rank Test	16
2.3. Concept and Clinical Use of Radiomics.....	16
2.3.1. First-Order Statistics	17
2.3.2. Shape Features.....	18

2.3.3. Texture Features.....	19
3. LITERATURE REVIEW	23
3.1. COVID-19 and Multimodal Data Fusion Related Literature Review	23
3.2. ILD and Radiomics Related Literature Review	25
4. MATERIAL AND METHODS	29
4.1. COVID-19 Cohort.....	29
4.1.1. Sampling and Next-Generation Viral Genome Sequencing Protocol.....	30
4.1.2. Identifying Viral Mutations and Construction of Phylogenetic Tree.....	31
4.1.3. Viral Feature Preprocessing	32
4.1.4. Imaging Data Collection and Analysis	32
4.1.5. Data Fusion Methodologies	33
4.1.6. Statistical Analysis	35
4.2. ILD Cohort	35
4.2.1. Patient Population	35
4.2.2. CT Image Acquisition and Segmentation	36
4.2.3. Pyradiomics Feature Extraction	39
4.2.4. Statistical Analysis	39
5. RESULTS	41
5.1. COVID-19 Cohort.....	41
5.1.1. Descriptive Analysis	41
5.1.2. Visualization of the Global SARS-CoV-2 Strains Using Word2Vec Embedding	44
5.1.3. Unsupervised Pairwise Data Fusion Using Sparse CCA	46
5.1.4. Supervised Multiview Analysis Using Cooperative Learning.....	55
5.2. ILD Cohort	58
5.2.1. Descriptive Analysis	58
5.2.2. Clinical Models in Transplant-Free Survival Analysis.....	60
5.2.3. Radiomics Models in Transplant-Free Survival Analysis	62
5.2.4. Comparison of Clinical and Radiomics Models in Transplant-Free Survival Analysis.....	63

5.2.5.	Interpretation of Radiomics Features in Transplant-Free Survival Analysis	64
5.2.6.	Radiomics Models in Pulmonary Hypertension Analysis.....	67
5.2.7.	Interpretation of Radiomics Features in Pulmonary Hypertension Analysis	68
6.	DISCUSSION	71
	REFERENCES.....	79
	APPENDICES.....	103
	APPENDIX A	103
	APPENDIX B	106
	APPENDIX C	107
	APPENDIX D	108
	APPENDIX E.....	110
	CURRICULUM VITAE	116

LIST OF TABLES

Table 1: Clinical Characteristics of the Patients.	41
Table 2: Sparse CCA for Examining Pairwise Associations Between All Data Modalities.	47
Table 3: Unimodal Prediction Models Evaluated in All Patients.	55
Table 4: Clinical Features of Patients with ILDs	60
Table 5: Performance of Baseline Functional Cox Proportional Hazards Models	61
Table 6: Hazards Ratios of Baseline Functional Cox Proportional Hazards Models for Transplant-Free Survival Analysis.....	62
Table 7: Performance of Lung and Pulmonary Artery Features of Transplant-Free Survival Models.....	63
Table 8: Performance of Lung and Pulmonary Artery Features of Pulmonary Hypertension Models.....	67

LIST OF FIGURES

Figure 1: Types of Multimodal Data Fusion.	10
Figure 2: The Pulmonary Arteries and Veins in the Human.....	14
Figure 3: Visualization of Categories of Radiomics Features in Pyradiomics Package. .	17
Figure 4: Flowchart of the Methods Used in the COVID-19 Cohort.....	31
Figure 5: CT Image Selection Process.	33
Figure 6: Repeated Stratified Nested Cross-Validation.	34
Figure 7: Anterior, Posterior, Superior, and Inferior Views of Binary Masks of Bilateral Lungs and Pulmonary Artery of a Real Patient.....	38
Figure 8: Axial CT Scan of a Real Patient Visualized Using Lung and Vascular Window Settings.	39
Figure 9: The Number of Patients in the Cohort with Different Data Modalities.	43
Figure 10: Phylogenetic Tree, Nucleotide Substitution Matrix, and Word2Vec Encoding Plot of Isolated SARS-CoV-2 Strains.	44
Figure 11: Phylogenetic Tree and Word2Vec Encoding Plot of Global SARS-CoV-2 Strains.....	45
Figure 12: Word2Vec Encoding Plot of Nextclade Clades and Overlapped Pango Lineages in UMAP Plot.....	46
Figure 13: Sparse CCA of Radiomics Features and Laboratory Results.	48
Figure 14: Correlation Plot Derived from Sparse CCA of Radiomics Features and Laboratory Results and Image Density Histogram of Example Patients.	50
Figure 15: Sparse CCA of Laboratory Results and Clinical Data.....	51
Figure 16: Sparse CCA of Viral Genome Sequencing Data with Different Data Modalities.	53
Figure 17: Sparse Multi-CCA of All Data Modalities.	54
Figure 18: The First Canonical Vectors of Multi-CCA.	54
Figure 19: Unimodal Prediction Models for All Data Modalities in All Patients.....	56
Figure 20: Unimodal and Multimodal Prediction Models for the Supervised Task.	57
Figure 21: Standardized Coefficients of the Quadruple Model (CLRW).	58
Figure 22: Flowchart of ILDs Cohort.	59
Figure 23: Comparison of Clinical and Radiomics Models in Transplant-Free Survival Analysis.	64
Figure 24: Forest Plot of Selected Lung Radiomics Features Associated with Transplant-Free Survival	65

Figure 25: Forest Plot of Selected Pulmonary Artery Radiomics Features Associated with Transplant-Free Survival.....	66
Figure 26: Forest Plot of Selected Lung and Pulmonary Artery Radiomics Features from Combined Models Associated with Transplant-Free Survival	66
Figure 27: Comparison of Clinical and Radiomics Models in Pulmonary Hypertension	68
Figure 28: Forest Plot of Selected Features from Pulmonary Hypertension Models Using Lung Radiomics Features.....	69
Figure 29: Forest Plot of Selected Features from Pulmonary Hypertension Models Using Pulmonary Artery Radiomics Features	70
Figure 30: Forest Plot of Selected Features from Pulmonary Hypertension Models Using Lung and Pulmonary Artery Radiomics Features	70

LIST OF ABBREVIATIONS

AUC	Area Under the Curve
C-index	Concordance Index
CCA	Canonical Correlation Analysis
CCI	Charlson Comorbidity Index
CNN	Convolutional Neural Network
COPD	Chronic Obstructive Pulmonary Disease
COVID-19	Coronavirus Disease 2019
CT	Computed Tomography
CTD-ILD	Connective Tissue Disease-Associated Interstitial Lung Disease
CTPA	Computed Tomography Pulmonary Angiography
CRP	C-Reactive Protein
CV	Cross-Validation
DICOM	Digital Imaging and Communications in Medicine
DLCO	Diffusion Capacity for Carbon Monoxide
ESR	Erythrocyte Sedimentation Rate
FVC	Forced Vital Capacity
GISAID	The Global Initiative on Sharing All Influenza Data
GGO	Ground Glass Opacities
GIHSN	Global Influenza Hospital Surveillance Network Project
GLCM	Gray-Level Co-occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Gray-Level Run-Length Matrix
GLSZM	Gray-Level Size Zone Matrix
HP	Hypersensitivity Pneumonitis
HR	Hazard Ratio
HRCT	High Resolution Computed Tomography
HU	Hounsfield Units
ICU	Intensive Care Unit
IDMN	Inverse Difference Moment Normalized
IDN	Inverse Difference Normalized
IIP	Idiopathic Interstitial Pneumonia
ILD	Interstitial Lung Disease
IPF	Idiopathic Pulmonary Fibrosis
LASSO	Least Absolute Shrinkage and Selection Operator
MDS	Multi-dimensional Scaling

ML	Machine Learning
MRI	Magnetic Resonance Imaging
NGTDM	Neighboring Gray-Tone Difference Matrix
NIFTI	Neuroimaging Informatics Technology Initiative
NLP	Natural Language Processing
NSIP	Non-specific Interstitial Pneumonia
OP	Organizing Pneumonia
ORFs	Open Reading Frames
UIP	Usual Interstitial Pneumonia
PCR	Polymerase Chain Reaction
PFT	Pulmonary Function Test
PH	Pulmonary Hypertension
RHC	Right Heart Catheterization
ROI	Regions of Interest
sPAP	Systolic Pulmonary Artery Pressure
TTE	Transthoracic Echocardiography
UMAP	Uniform Manifold Approximation and Projection
WHO	World Health Organization
WL	Window Level
WW	Window Width

CHAPTER 1

INTRODUCTION

Today, in biological sciences and medicine, vast amounts of high-dimensional, high-throughput biomedical data are being generated from a wide range of sources, including, but not limited to, molecular assays, histopathology, radiology, wearable sensors, and clinical records [1]. Advanced machine learning (ML) methodologies have been pivotal in combining and analyzing these data modalities, unveiling a myriad of biomarkers and novel approaches that support personalized medicine and public health applications [2-6]. Complementing these advances, statistical methods, particularly those that promote sparsity and interpretability, remain essential for uncovering meaningful associations across heterogeneous datasets [7].

Multimodal data fusion aims to integrate information from multiple modalities, each offering a unique perspective on the same underlying phenomenon, to address a specific downstream task [8]. It can be classified into three categories based on the state of the inputs: early, intermediate, and late fusion. While in early fusion, the concatenated vector is in the input, in intermediate and late fusion, marginal representations and decisions are combined, respectively. Each strategy has its advantages depending on the nature of the data and the task at hand. Selecting and optimizing the degree of fusion is critical for effectively leveraging multimodal information and improving the model's performance.

Lung diseases are a leading cause of morbidity and mortality worldwide, including a broad spectrum of conditions from acute infections to chronic fibrotic disorders [9]. These conditions can significantly alter the structure and function of the lungs, often leading to impaired gas exchange, respiratory symptoms, and a reduced quality of life. Infections like bacterial and viral respiratory diseases, as well as chronic diseases like asthma, chronic obstructive pulmonary disease (COPD), and interstitial lung diseases (ILDs), are among the most prevalent and clinically relevant forms. For diagnosis, prognosis, and treatment planning, these disorders must be accurately characterized and monitored [10].

The clinical and biological heterogeneity of acute and chronic lung diseases presents significant challenges in developing effective prevention and treatment strategies. For instance, during the COVID-19 pandemic, one of the most challenging factors for clinicians in managing patients with COVID-19 is the variability in clinical course, where the severity determinants include host genetic factors, demographics, comorbidities, viral mutations, and host-pathogen interactions [11-13]. Similarly, ILDs represent a group of chronic lung disorders characterized by progressive scarring of lung tissue with multiple etiologies and variable trajectories. This heterogeneity requires comprehensive approaches, including the use of diverse data modalities in patient management, ranging from histopathology slides and genetic tests to radiological images and cardiac evaluations [14, 15].

Within this spectrum, computed tomography (CT) plays a pivotal role in the imaging assessment of lung disorders. It provides a fast and reliable method for evaluating the extent and severity of acute lung injuries, particularly in cases of viral exacerbations and acute pulmonary embolism [16]. In chronic conditions, high-resolution CT (HRCT) enables the detailed visualization of the airways, vasculature, and lung parenchyma, facilitating the identification of disease subtypes and monitoring the progression. However, despite the critical importance of CT in clinical practice, interpretation remains largely qualitative: The visual assessment is inherently subjective and prone to inter- and intra-observer variability. Even among experienced radiologists, differences in interpretation can lead to inconsistencies in diagnosis and disease staging. Although visual scoring systems offer some standardization, they are still semi-quantitative and may overlook subtle or diffuse changes in the lung parenchyma. This subjectivity limits their utility in research settings and in developing robust, individualized treatment plans [17].

To overcome these limitations, the field of quantitative imaging has gained increasing interest [17]. Radiomics offers various computational methods that are promising for extracting numerous quantitative characteristics from medical images. These characteristics provide details about the intensity, texture, shape, and spatial relationships of a specific region of interest, which transforms imaging data into a mineable, high-dimensional information [18, 19]. Among the computational tools used in radiomics, pyradiomics is a widely adopted open-source library that enables the extraction of handcrafted radiomic features from medical images, capturing shape, texture, and intensity characteristics of a region of interest [20].

However, in medicine, systemic diseases often affect multiple organs and tissues simultaneously [21]. For example, systemic lupus erythematosus is an autoimmune disorder that can affect the skin, joints, kidneys, brain, and lungs [22]. Rheumatoid arthritis, another autoimmune disease, primarily targets the joints but may also impact the lungs, heart, and eyes [23]. Diabetes mellitus, a common metabolic disorder, can lead to complications in the eyes, kidneys, nerves, and the cardiovascular system [24]. Similarly, systemic sclerosis is characterized by widespread vascular abnormalities and fibrosis, commonly affecting the skin, gastrointestinal tract, lungs, and kidneys [25].

These examples underscore the diffuse nature of systemic diseases and the necessity of evaluating multiple anatomical compartments concurrently.

Traditionally, multi-compartment imaging analysis has been adopted in neuroimaging, where imaging signal contributions from different tissue compartments are isolated from distinct brain regions, such as cortical lobes, ventricles, or white matter tracts, to account for spatial heterogeneity in disorders like Alzheimer’s disease or multiple sclerosis [26, 27]. A similar rationale is also applicable for pulmonary diseases, where airways, pulmonary arteries, and lung parenchyma are anatomically distinct, but functionally coupled, located within a shared space, and most importantly, share complementary roles in disease pathogenesis [28]. Despite this, most radiomics pipelines still rely on single-region segmentation and analysis, limiting their ability to model inter-compartmental relationships and spatial heterogeneity in disease pathogenesis.

Word2Vec is a shallow neural network for learning vector representations of words, known as word embeddings, from large, unlabeled corpora [29]. Instead of treating words as discrete and independent symbols, Word2Vec maps each word to a continuous, dense vector in high-dimensional space where semantically similar words are positioned close to each other. Once the model is trained, instead of using the model itself, these embeddings are utilized and enable arithmetic operations on words.

Canonical correlation analysis (CCA) is a classical statistical technique used to identify and quantify linear relationships between two or more sets of variables [30]. This technique is particularly instrumental in scenarios involving paired datasets, such as linking gene expression profiles with clinical measurements. However, traditional CCA assumes that the number of observations exceeds the number of variables, which is often not the case in high-dimensional biomedical data, leading to overfitting and interpretability challenges. To address these limitations, sparse canonical correlation analysis (Sparse CCA) adds LASSO regularization to the canonical vectors, resulting in sparse linear combinations that focus on the most relevant variables. This sparsity enhances model interpretability and is advantageous in high-dimensional settings where the number of variables exceeds the number of observations [31, 32].

Cooperative learning is a recently developed framework for integrating multi-view datasets, where each view offers a distinct representation of the same set of observations with a supervised task [33]. This technique combines the usual squared error loss of predictions with an agreement penalty to encourage the predictions from different data views to align. By tuning the agreement penalty, cooperative learning yields a continuum of solutions between early fusion and late fusion. As an extension to existing solutions, cooperative learning also supports sparsity, enabling variable selection, interpretability, and reducing the risk of overfitting.

The Cox proportional hazards model is a widely used statistical approach for investigating the association between multiple covariates and the time to a specific event, such as death or diagnosis. Unlike parametric survival models, the Cox model does not assume a

specific baseline hazard function, making it a flexible and semi-parametric tool for survival analysis. Its primary advantage lies in estimating hazard ratios (HR), which quantify the relative risk associated with predictors while accounting for censored data [34].

In this thesis, we utilized two distinct patient cohorts to explore multimodal data fusion and multicompartment imaging analysis. The first part focused on an analysis of a unique COVID-19 patient cohort, where we integrated imaging, clinical, laboratory, and viral genome sequencing data of patients. We employed sparse CCA and cooperative learning to investigate the associations between different data modalities and optimize multimodal data fusion for predicting intensive care unit (ICU) admission, respectively. We innovatively leveraged the Word2Vec natural language processing model (NLP) to encode viral genomes to investigate the phylogenetic relationships of viral strains and to integrate viral genomic information into our clinical dataset. In the second part, we conducted radiomics analysis and trained diagnostic and survival models in a cohort of patients with ILDs. By considering the biological and clinical pathways and referring to this approach as multicompartment image analysis, we extracted radiomics features from the pulmonary arteries of patients in addition to the lung parenchyma. We interpreted the findings and elucidated the significance of pulmonary artery radiomics features, thereby extending the application of lung radiomics analysis.

1.1.Motivation

This thesis is motivated by the goal of translating the comprehensive evaluation and management strategies used in clinical practice into the domain of medical informatics, with a focus on multimodal data fusion and multicompartment image analysis. Both COVID-19 and ILDs are suitable candidates for this purpose, especially in terms of multimodal data availability and disease heterogeneity. As is known, the COVID-19 pandemic generated an unprecedented amount of diverse and accessible data worldwide, including viral genomic sequences, medical imaging, electronic health records, vaccination and treatment responses, and population-level public health data [35-37]. Apart from its global importance, COVID-19 affects multiple organ systems and presents a broad spectrum of severity, making it a rich case for studying complex biological interactions [38]. These features make COVID-19 an ideal testbed for developing and validating data fusion methodologies that aim to integrate heterogeneous sources and identify inter-modality patterns, not only in the scale of pandemic response, but also achieving the broader aim of building integrated, data-driven frameworks for future biomedical challenges. To achieve this aim, we utilized sparse CCA and cooperative learning in both unsupervised and supervised tasks. Notably, this study represents the first application of cooperative learning on an external dataset since its initial introduction, demonstrating its potential for generalization beyond the original development setting.

Existing sub-phenotyping schemas for acute and chronic lung diseases primarily focus on either clinical or radiologic subgroups; however, these approaches are rarely integrated in a harmonized manner. This limitation is due to several key factors: Technically, one of

the most critical hurdles is the transformation of rich, high-dimensional data, such as radiographic images, molecular assays, and clinical information, into structured, lower-dimensional representations suitable for modeling. This pressure becomes more apparent day by day, as the declining cost of next-generation sequencing technologies enables researchers to analyze genomics comprehensively. This creates both an opportunity and a necessity to integrate genomics with imaging and clinical data to build robust multimodal predictive models. Such integration highlights the demand for advanced vectorization strategies that move beyond black-box representations to incorporate domain-specific knowledge, context-specific structures, and task relevance. In addition, achieving clinically usable multimodal models requires not only harmonizing data types and standards but also developing interpretable frameworks that align with biology, clinical workflows, and support informed decision-making. In this thesis, the contribution of leveraging the Word2Vec NLP model to the literature is significant, as we aim to explicitly address viral evolution by comparing Word2Vec embeddings with traditional phylogenetic trees and incorporate the core idea of infusing global viral information into the clinical dataset to obtain a better latent representation.

Radiomics provides an opportunity to address inter- and intraobserver variability by extracting reproducible, quantitative data from CT images [39]. The imaging-based evaluation strategies and computational techniques developed in response to COVID-19 have significantly advanced our capabilities in detecting and quantifying acute parenchymal changes. These methodological advances can be effectively extended to chronic lung diseases, where long-term structural alterations require robust and sensitive imaging biomarkers. Studying both acute processes, such as those observed in COVID-19, and chronic fibrotic conditions, such as ILDs, provides a unique opportunity to investigate not only shared imaging characteristics but also features that are distinct to each disease state. Although the same set of radiomic features can be extracted from chest CTs across these conditions, their relevance and existence may differ depending on the disease context. In comparing acute inflammation with chronic fibrosis, distinct radiomic patterns may emerge as meaningful, reflecting the temporal and structural nuances of each disease process. Among chronic conditions, ILDs represent a diverse group characterized by progressive fibrosis, functional decline, and variable clinical trajectories, making them a natural extension for applying and refining radiomics-based analysis.

As highlighted, systemic diseases frequently involve multiple organs and tissues at the same time. Consequently, clinical assessments require a comprehensive evaluation of various anatomical and physiological systems, a principle that has guided medicine for centuries. In oncology, for example, accurate staging and prognostication depend not only on the primary tumor but also on lymph node involvement and distant metastases [40]. The TNM classification system reflects this by assigning structured scores to each component, which are used to guide treatment decisions [41]. Similarly, many pulmonary diseases have both local and systemic complications [42]. Pulmonary hypertension (PH) is one of them, a common comorbidity in advanced lung disease, and is associated with poor prognosis and demands targeted therapeutic strategies. Its evaluation can offer critical prognostic information beyond parenchymal assessment alone [43]. While

radiologists and clinicians routinely interpret chest CT scans by integrating information from the lungs and surrounding structures, most radiomics studies to date have focused on isolated regions of interest in pulmonology. The integration of radiomic features from multiple anatomical compartments into a unified computational framework remains underexplored. This thesis addresses this gap explicitly by extracting and analyzing pulmonary artery radiomics features alongside those from the lung parenchyma.

Hopefully, this work will contribute to the growing field of quantitative imaging and multimodal data fusion, as well as their translation into meaningful clinical applications.



CHAPTER 2

BACKGROUND

2.1.COVID-19 Infection

On March 11, 2020, the World Health Organization (WHO) officially declared COVID-19 a pandemic. Since then, the disease has impacted millions of people worldwide. The pandemic has demonstrated that, even in wealthy regions, overcrowding in hospitals and a lack of resources can put healthcare facilities under pressure [44]. This was particularly evident in key aspects of care, such as emergency triage [45].

COVID-19 presents with a highly variable clinical course [46]. While many individuals experience asymptomatic or mild illness, a substantial proportion develop severe complications, including respiratory failure necessitating ICU admission. The unpredictable nature of disease progression has posed significant challenges to frontline healthcare providers, leading to difficulties in triage and suboptimal resource allocation [47]. Although demographic factors, such as age, sex, and pre-existing comorbidities, have been consistently associated with disease severity, growing evidence has highlighted the influence of viral mutations on transmissibility, immune escape, and patient outcomes [48].

ICU admission in COVID-19 is typically guided by criteria indicating severe respiratory distress or multi-organ dysfunction [49]. These include persistent hypoxemia despite high-flow oxygen therapy, respiratory failure requiring mechanical ventilation, hemodynamic instability, and altered mental status. Mortality rates among ICU patients reached 50% in a systematic review, and this underscores an urgent need for reliable methods to identify patients at high risk of clinical deterioration [50].

In this context, both laboratory markers and chest CT have emerged as valuable tools for assessing disease severity. Elevated levels of laboratory biomarkers, such as C-reactive protein (CRP), D-dimer, ferritin, and interleukin-6, have been associated with severe COVID-19 and poorer outcomes. These markers reflect systemic inflammation, coagulopathy, and immune dysregulation, hallmarks of severe disease. Chest CT most

commonly reveals ground-glass opacities (GGO), with or without accompanying consolidation, consistent with viral pneumonia. In a systematic review evaluating chest CT findings in over 2,700 COVID-19 patients, GGO were reported in 83% of cases, consolidation alongside GGO in 58%, pleural thickening in 52%, interlobular septal thickening in 48%, and air bronchograms in 46% [51]. CT abnormalities are often bilateral, peripherally distributed, and predominantly located in the lower lobes of the lungs. As with radiographs, CT findings may appear normal shortly after symptom onset; however, the likelihood and severity of abnormalities increase as the disease progresses.

The genetic structure of SARS-CoV-2, the causative agent of COVID-19, consists of 13–15 open reading frames (ORFs) and contains approximately 30,000 nucleotides. The genetic organization of these ORFs is highly similar to that of SARS-CoV and MERS-CoV [52, 53]. SARS-CoV-2 has evolved, and most mutations in the SARS-CoV-2 genome have no impact on virulence. However, some variants have attracted widespread attention due to their rapid emergence and evidence related to increased transmission and clinical implications. They have been classified as variants of concern. Each variant has multiple names based on different phylogenetic classification systems. WHO has also assigned labels based on the Greek alphabet for major variants, such as Alpha (B.1.1.7) and Omicron (B.1.1.529) [54].

Global efforts to monitor and understand the evolution of SARS-CoV-2 have led to the rapid development of large-scale viral genome databases. GISAID (The Global Initiative on Sharing All Influenza Data) has become a key platform for the timely sharing of SARS-CoV-2 genomic sequences, enabling researchers worldwide to track mutations, monitor variants of concern, and inform public health decisions [55]. Complementing this, Nextstrain provides a real-time visualization framework that integrates phylogenetic analysis with geographic and temporal data, allowing the dynamic tracking of viral evolution and transmission patterns [56]. Together, these platforms have played a critical role in supporting global genomic surveillance.

2.1.1. Word2Vec NLP Model

Word2Vec is an NLP model, a shallow neural network, that transforms words into continuous vector representations based on their contextual usage within a corpus [29]. The network consists of an input layer, a hidden layer, which defines the embedding size, and an output layer, and it is trained using techniques such as negative sampling or hierarchical SoftMax to efficiently handle large vocabularies. Fine-tuning involves adjusting hyperparameters like window size, vector dimensionality, and the number of negative samples, as well as refining the model on domain-specific corpora to capture more relevant semantic relationships.

The main assumption of Word2Vec is the distributional hypothesis, which states that words that occur in similar contexts tend to have similar meanings. This means that the meaning of a word can be inferred from the context around this word. Word2Vec

leverages this idea by learning to place words that appear in similar contexts closer together in the vector space, thereby capturing semantic relationships based on word co-occurrence patterns. Using distance metrics, such as cosine similarity, enables us to quantify the distance between words.

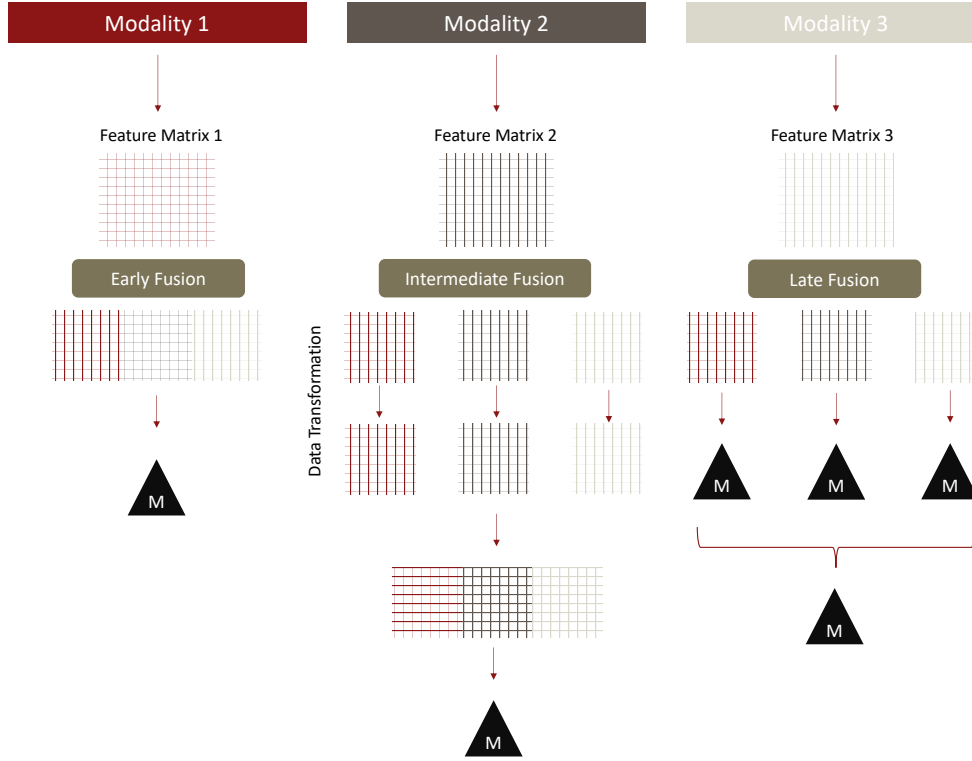
Unlike traditional bag-of-words models, which treat words as discrete and independent units, Word2Vec represents each word as a dense, low-dimensional vector. It operates using one of two architectures: Continuous Bag of Words, which predicts a target word from its context, or Skip-gram, which predicts the surrounding context given a target word. Trained on large corpora, Word2Vec embeddings allow performing algebraic operations on words (e.g., *king* - *man* + *woman* = *queen*).

2.1.2. Multimodal Data Fusion

Multimodal data fusion refers to the process of integrating heterogeneous data types, such as imaging, clinical notes, electronic health records, genomics, and time-series data, into a unified framework for analysis. In the context of ML and healthcare, this approach aims to replicate the clinical diagnostic process, which frequently involves combining information from multiple sources to reach a decision. Unlike unimodal systems that focus on a single type of data input, multimodal data fusion leverages the complementary strengths of each data type to create a more holistic understanding of a patient's health status [6].

The integration of multiple data modalities has become increasingly essential for advancing precision health, where individualized care relies on detailed patient-specific information. In a review, models using multimodal data, rather than unimodal inputs, were found to have an average 6.4% higher area under the curve (AUC), and it was argued that data fusion was generally beneficial and recommended [6]. This enhanced performance is critical for complex conditions such as Alzheimer's disease, cancer, diabetes, and heart failure, where single data types may not sufficiently capture the disease heterogeneity. Furthermore, multimodal data fusion holds promise for addressing biases and disparities in healthcare by leveraging diverse data sources that reflect varying population subgroups. However, practical implementation remains challenged by data heterogeneity, lack of standardization, and limited clinical translation.

Three principal types of data fusion strategies are employed in multimodal ML: early (data-level), intermediate (feature-level or joint), and late (decision-level) fusion (Figure 1) [4, 6]. Early fusion combines raw or pre-processed data from various modalities into a single input before feeding it into a model. Intermediate fusion involves the separate processing of each modality, followed by joint representation learning, which allows models to capture both modality-specific and cross-modal interactions. Late fusion, akin to ensemble learning, builds separate models for each modality and aggregates their predictions through voting or weighted strategies. Each approach offers unique benefits and trade-offs in terms of scalability, interpretability, and performance.



5

Figure 1: Types of Multimodal Data Fusion.

M: Model

2.1.3. Sparse Canonical Correlation Analysis

CCA is a classical multivariate statistical method introduced by Hotelling in 1936 that aims to identify and quantify the associations between two sets of variables [30].

Assume we have two data matrices, X and Z , of dimensions $n \times p$ and $n \times q$ on the same set of n observations given as Eq. (1):

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1q} \\ z_{21} & z_{22} & \cdots & z_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nq} \end{bmatrix} \quad (1)$$

CCA seeks linear combinations (canonical variables) of the variables in X and Z that are maximally correlated. That is, $u_1 = (u_{11}, u_{21}, \dots, u_{p1})^T$ and $v_1 = (v_{11}, v_{21}, \dots, v_{q1})^T$ maximize $\text{corr}(Xu_1, Zv_1)$. Here, we refer to u_1 and v_1 as the canonical vectors and Xu_1 and Zv_1 as the canonical variables. u and v have the dimensions of $p \times K$ and $q \times K$ for K canonical vectors, which are not correlated with each other.

Sparse CCA identifies the linear relationships between two sets of high-dimensional variables while enforcing sparsity. Penalized decomposition analysis extends this framework by incorporating penalization directly into the decomposition process of the cross-covariance matrix between two datasets, using L1 (LASSO) regularization. Thus, sparse CCA aims to find sparse canonical vectors u and v such that $u^T X^T Z v$ is optimized. Sparse CCA inherently finds linear solutions, but also non-linear solutions, including kernel CCA and deep CCA, have been developed as well [31, 57].

2.1.4. Cooperative Learning

Cooperative learning is a recently developed supervised learning framework that aims to optimize the degree of multimodal data fusion [33].

Assume we have two data matrices, X and Z , of dimensions $n \times p_x$ and $n \times p_z$, and a corresponding target vector y .

$$\min E[\frac{1}{2}(y - f_X(X) - f_Z(Z))^2 + \frac{\rho}{2}(f_X(X) - f_Z(Z))^2] \quad (2)$$

As Eq. (2), the first term is the usual prediction error, while the second term is an agreement penalty. This approach systematically combines multiple data views by augmenting the loss function with an agreement penalty that promotes alignment between predictions from each view. By adjusting the weight of this penalty, the method flexibly spans a range of fusion strategies, from early to late fusion. If $\rho = 0$ from Eq. (2), cooperative learning fits f_X and f_Z together. If these functions are additive, it yields an early fusion. Whereas, if $\rho = 1$, the formula yields a solution of an average of marginal fits for X and Z , which is a type of late fusion.

This framework is particularly effective when the different views share latent structure or correlated signals, allowing the model to leverage complementary information across views to improve predictive performance.

2.1.5. The Games-Howell Test

The Games-Howell test is a post-hoc statistical method used to conduct pairwise comparisons between group means following an analysis of variance (ANOVA), particularly when the assumptions of equal variances (homoscedasticity) are violated. Unlike traditional methods such as Tukey's Honest Significant Difference, which assumes homogeneity of variance, the Games-Howell procedure is robust to heteroscedasticity and

unequal sample sizes. The test adjusts both the standard error and the degrees of freedom using a Welch-type correction, thereby controlling for Type I error. As a result, the Games-Howell test is used when ANOVA indicates a significant difference, but Bartlett's test suggests that variances across groups are unequal.

2.2. Interstitial Lung Diseases

2.2.1. Epidemiology

ILDs encompass over 200 distinct pulmonary conditions characterized by varying degrees of inflammation and fibrosis of the lung interstitium [10]. While many ILDs share overlapping clinical, radiological, physiological, and pathological features, they can differ significantly in terms of etiology, progression, and treatment response [58]. Based on cause, ILDs can be broadly classified into those with known etiologies, such as drug-induced ILDs, connective tissue disease-associated ILDs (CTD-ILDs), and ILDs related to occupational or environmental exposures (e.g., asbestosis, silicosis), and those of unknown cause, referred to as idiopathic interstitial pneumonias (IIPs). Among IIPs, idiopathic pulmonary fibrosis (IPF) is the most prevalent and extensively studied subtype, often serving as the model for progressive fibrotic ILD. Other important subtypes include non-specific interstitial pneumonia (NSIP), fibrotic hypersensitivity pneumonitis (HP), and unclassifiable ILD.

The course of ILDs varies; some stay the same, while others worsen quickly. Severe respiratory failure in advance disease is the cause of mortality. Prognostication and treatment choices are influenced by recognition of patients with high risk. Pulmonary function tests (PFTs), especially, forced vital capacity (FVC) and diffusing capacity for carbon monoxide (DLCO) are essential to monitor lung functions [59, 60]. Measures of oxygen desaturation and exercise testing, such the six-minute walk test, can also reveal functional deterioration [61].

The diagnosis and management of ILDs requires a multidisciplinary approach that incorporates clinical assessment, radiological imaging, physiological testing, and, in selected cases, histopathological evaluation [10]. The diagnostic process typically begins with a thorough medical history, focusing on medication use, systemic symptoms, and potential occupational or environmental exposures. HRCT is the cornerstone modality in ILD diagnosis and classification [62]. It offers in-depth imaging of the lung parenchyma, making it possible to recognize distinctive patterns like NSIP, organizing pneumonia (OP), and usual interstitial pneumonia (UIP) [63]. These patterns, described by radiologists using findings like GGOs, honeycombing, traction bronchiectasis, and reticulation, are critical for guiding diagnosis and treatment decisions. However, accurate interpretation of these complex patterns requires substantial radiological expertise, highlighting the growing need for computer-assisted diagnostic tools.

Beyond diagnosis and classification, CT imaging, particularly HRCT, plays a central role in assessing disease severity and progression. Structural changes such as increased fibrotic reticulation, traction bronchiectasis, honeycombing, and volume loss are associated with poor prognosis and often represent irreversible lung damage [64]. Serial HRCT imaging can help quantify the extent of fibrosis, which has been correlated with both lung function decline and reduced survival [65]. Among imaging patterns, the presence of UIP is particularly significant, as it is strongly associated with IPF and yields a worse prognosis.

Despite its advantages, HRCT interpretation in ILD remains challenging. Radiological findings can overlap considerably, particularly between UIP and fibrotic hypersensitivity pneumonitis HP, or between NSIP and early UIP, complicating accurate diagnosis and management [66]. Interobserver variability among radiologists is also a persistent issue, especially in distinguishing between probable and definite UIP patterns [67]. Moreover, identification of the ILD subtype, or even the radiological pattern itself, may not be sufficient for reliable prognostication. For instance, a comparative study involving patients with CTD-ILD, IPF, and HP demonstrated that CTD-ILD patients had better survival than those with IPF, despite both groups sharing the same UIP pattern [68]. Conversely, within a cohort of CTD-ILD patients, those exhibiting a UIP pattern had significantly worse outcomes than those with an NSIP pattern, highlighting the prognostic heterogeneity even within a single disease category [69]. These findings underscore the need for novel prognostic approaches that rely on more robust and quantitative metrics, moving beyond traditional subtype classifications and pattern recognition alone.

2.2.2. Pulmonary Hypertension in ILD: Definition and Hemodynamic Criteria

The pulmonary artery originates from the right ventricle of the heart as the main pulmonary trunk, which subsequently bifurcates into the right and left pulmonary arteries, delivering blood to the respective lungs (Figure 2). It is responsible for carrying deoxygenated blood from the right side of the heart to the lungs for oxygenation [70]. Under normal physiological conditions, the pulmonary circulation is a low-pressure, high-compliance system designed to accommodate the entire cardiac output without significant pressure increases [71]. However, when pulmonary vascular resistance increases due to vasoconstriction, vascular remodeling, or obliteration of pulmonary vessels, pulmonary artery pressure rises.

PH is a pathological condition characterized by elevated blood pressure inside the pulmonary arteries. This increased afterload imposes a significant burden on the right ventricle, which is not structurally adapted to generate high pressures. Over time, this leads to right ventricular hypertrophy, dilation, and ultimately right heart failure if the underlying cause is not addressed. PH is classified into five distinct groups by the WHO, based on etiology, pathophysiology, and treatment approaches [70]. Group 1 includes pulmonary arterial hypertension, which is characterized by pre-capillary PH due to intrinsic abnormalities of the small pulmonary arteries and includes idiopathic, heritable, and connective tissue disease-associated forms. Group 2 PH results from left heart

disease, such as heart failure with preserved or reduced ejection fraction and valvular diseases. Group 3 PH is associated with lung diseases and/or hypoxia, including ILDs, COPD, and sleep-disordered breathing. Group 4 is due to chronic thromboembolic PH, caused by persistent obstruction of pulmonary arteries by thromboembolism. Group 5 includes PH with unclear or multifactorial mechanisms, such as sarcoidosis, hematologic disorders, and metabolic diseases. This classification helps guide diagnostic evaluation and treatment strategies, as the underlying mechanisms and therapeutic options vary widely among groups.

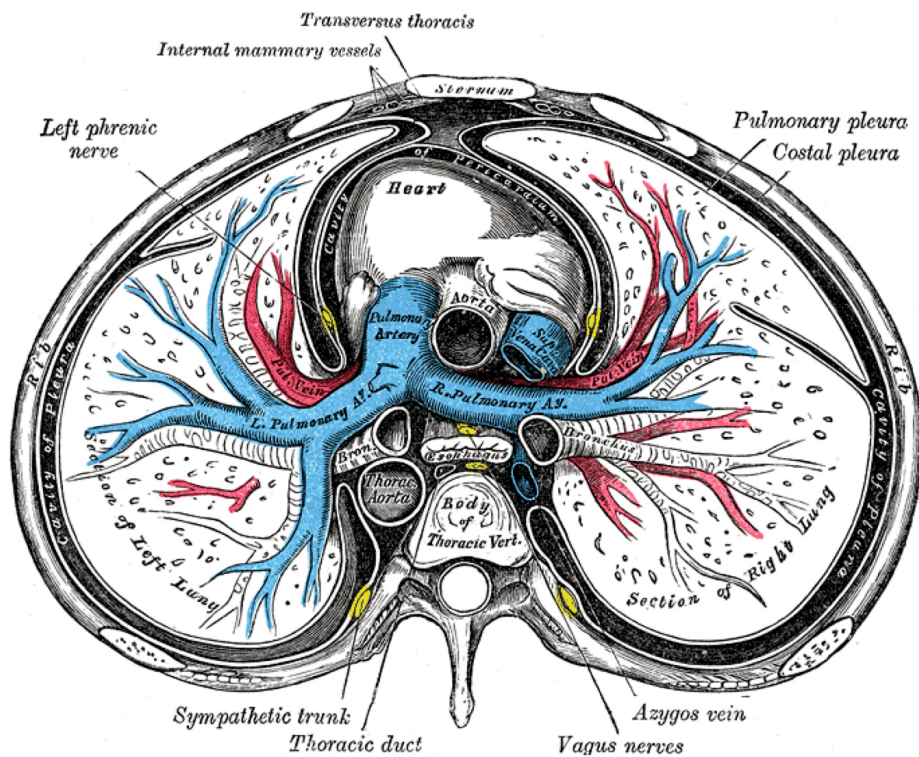


Figure 2: The Pulmonary Arteries and Veins in the Human

(Original diagram from Gray's anatomy, now out of copyright. This diagram has been reproduced from Gray's Anatomy 20th US edition, which has now lapsed into the public domain) [72]

PH associated with ILDs falls under Group 3 PH, where the primary mechanism is chronic lung disease and/or hypoxia-induced vascular remodeling. However, a mixed etiology incorporating group 1 (pulmonary arterial hypertension) might also exist in patients with overlapping risk factors (e.g., systemic sclerosis). In the setting of ILDs, persistent alveolar hypoxia leads to hypoxic pulmonary vasoconstriction, which initially serves to match ventilation with perfusion but, over time, contributes to structural changes in the pulmonary vasculature. These include medial hypertrophy, intimal fibrosis, and

obliteration of capillaries, all of which increase pulmonary vascular resistance. Additionally, extensive fibrosis may directly destroy parts of the pulmonary capillary bed, further worsening vascular load. These changes cumulatively raise pulmonary artery pressure, increasing right ventricular workload and eventually leading to right heart dysfunction or failure if left untreated [73]. The presence of PH in patients with ILDs is associated with markedly worse clinical outcomes. It often manifests with worsening exertional dyspnea, fatigue, and signs of right heart failure, such as peripheral edema and elevated jugular venous pressure. Importantly, ILD-PH may be underdiagnosed, especially in early stages, due to overlapping symptoms with the underlying lung disease.

Early identification of PH is crucial, as it has significant implications for prognosis and management [74]. Timely diagnosis allows for closer monitoring, optimization of supportive therapies such as oxygen supplementation and pulmonary rehabilitation, and consideration of targeted treatments in select cases. Moreover, recognizing PH at an earlier stage may prompt more comprehensive evaluation for transplant eligibility or influence decisions regarding disease-modifying therapies. Given the subtle and nonspecific nature of early PH symptoms, integrating noninvasive imaging, biomarkers, and physiological testing into routine ILD assessment may enhance detection and improve clinical outcomes.

Based on the European Society of Cardiology and European Respiratory Society guidelines, PH is hemodynamically characterized by right heart catheterization (RHC) as a mean pulmonary arterial pressure > 20 mmHg at rest with pulmonary vascular resistance ≥ 2 Wood units [75]. As a noninvasive screening tool, echocardiography plays an important role in the initial assessment of PH, with an estimated systolic pulmonary artery pressure (sPAP) above 35–40 mmHg considered suggestive of PH. [75]. Although they are still not perfect substitutes, CT-derived pulmonary artery parameters, including the major pulmonary artery diameter and the ratio of pulmonary artery to ascending aorta, have become non-invasive indicators of PH [76].

In progressive fibrosing subtypes such as CTD-ILD, IPF, and fibrotic HP, PH is a common and clinically significant consequence of ILDs. In IPF, approximately 50% of patients develop PH during the disease course, particularly in advanced stages [77]. In CTD-ILD, especially systemic sclerosis, PH may occur in up to 20–30% of patients and can present earlier in the disease course [78]. In sarcoidosis or chronic hypersensitivity pneumonitis, PH prevalence is more variable but can be substantial, especially in the fibrotic end-stages. Numerous studies have shown that PH is related to worse exercise tolerance, higher hospitalization rates, and noticeably lower survival rates [79].

Management of ILD-PH is challenging, as most pulmonary vasodilators approved for Group 1 PAH have shown limited or even harmful effects in Group 3 PH. However, recent studies suggest potential benefit from selected therapies (e.g., inhaled treprostinil) in carefully selected patients with ILD-PH [80]. Given its prognostic significance and therapeutic complexity, early recognition and accurate characterization of PH in ILD are essential for optimal care and stratification. Its diagnosis and management are critical, as

the presence of PH often indicates a more severe disease phenotype and may influence treatment decisions and transplant eligibility.

2.2.3. Cox Proportional Hazards Model

The Cox proportional hazards model, introduced by Sir David Cox in 1972, is a widely used statistical technique in survival analysis for examining the relationship between the survival time of subjects and one or more predictor variables [81]. Unlike parametric models, the Cox model is a semi-parametric model, it does not require specification of the baseline hazard function, which allows greater flexibility in modeling complex data. The core assumption of the model is the proportional hazards assumption, meaning that the HRs between different levels of covariates remain constant over time. The model expresses the hazard function for an individual as the product of a baseline hazard function and an exponential function of covariates.

The Cox proportional hazards model's application spans a wide range of fields, including oncology, cardiology, infectious diseases, and pulmonary medicine and enables to evaluate the prognostic value of biomarkers, imaging features and clinical variables while accounting for censored data. Regularized versions of the Cox model (LASSO-Cox) and machine learning based survival models have been developed to handle high-dimensional data as well [82, 83].

2.2.4. Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a non-parametric statistical method used to compare two related samples or matched pairs to assess whether their mean ranks differ. Unlike the paired t-test, which assumes normally distributed differences between paired observations, the Wilcoxon test makes no such assumption. This test evaluates the null hypothesis that the median difference between pairs is zero by ranking the absolute differences and assessing the sign of these ranks. Wilcoxon-signed rank test is generally recommended, instead of using the mean-ranks post-hoc test, for multiple pairwise comparisons [84].

2.3. Concept and Clinical Use of Radiomics

Radiomics is a computational approach that extracts large sets of quantitative features from standard medical images, enabling a detailed characterization of tissue phenotypes [85]. Its clinical potential lies in identifying imaging biomarkers that are difficult or impossible to quantify through conventional visual interpretation. It can also be leveraged to identify patterns related to disease progression and response to treatment [19].

First-order statistics (e.g., mean intensity, standard deviation), shape-based features, and higher-order texture features taken from gray-level matrices like the gray-level run-length

matrix (GLRLM), gray-level size zone matrix (GLSZM) and gray-level co-occurrence matrix (GLCM), are just a few of the many types of features that are extracted from predefined regions of interest (ROIs) [86]. By emphasizing particular spatial frequency patterns or edges in the image, advanced image filters like Wavelet and Laplacian of Gaussian can also be used to improve feature extraction. [87].

Pyradiomics is one of the most widely used open-source tools for radiomic feature extraction [20]. It provides a standardized framework for reproducibly extracting features from segmented ROIs in imaging data by using SimpleITK and NumPy libraries. Numerous radiology and cancer studies have validated and implemented it [88].

In Pyradiomics, radiomic features can be classified in three categories (Figure 3):

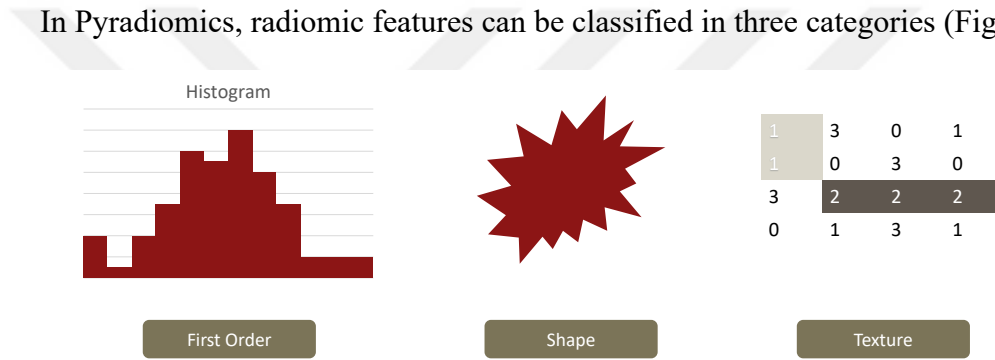


Figure 3: Visualization of Categories of Radiomics Features in Pyradiomics Package.

2.3.1. First-Order Statistics

The most basic statistical descriptors are derived from the global gray-level histograms and include metrics such as mean, maximum, minimum, variance, and percentiles of gray levels. Since these features are calculated from individual pixels or voxels without considering spatial relationships, they are referred to as first-order features. Skewness (Eq. (3)) and Kurtosis (Eq. (4)) describe asymmetry and peakedness of the intensity histogram; entropy (Eq. (5)) reflects image heterogeneity, and energy (Eq. (6)) and uniformity (Eq. (7)) measure the homogeneity of intensity distributions. Example formulas are provided below:

Assuming that $\mathbf{X}$ be a set of N_p voxels included in the ROI, $\mathbf{P}(i)$ be the first order histogram with N_g discrete intensity levels, where N_g is the number of non-zero bins, c is the voxelArrayShift, the hyperparameter for preventing negative values in $\mathbf{X}$, $p(i)$ be the normalized first order histogram and ϵ is an arbitrarily small positive number,

$$skewness = \frac{\mu^3}{\sigma^3} = \frac{\left[\left(\frac{1}{N_p} \right) \sum_{i=1}^{N_p} (X(i) - \bar{X})^3 \right]}{\left[\left(\sqrt{\left(\frac{1}{N_p} \right) \sum_{i=1}^{N_p} (X(i) - \bar{X})^2} \right)^3 \right]} \quad (3)$$

$$kurtosis = \frac{\mu^4}{\sigma^4} = \frac{\left[\left(\frac{1}{N_p} \right) \sum_{i=1}^{N_p} (X(i) - \bar{X})^4 \right]}{\left[\left(\left(\frac{1}{N_p} \right) \sum_{i=1}^{N_p} (X(i) - \bar{X})^2 \right)^2 \right]} \quad (4)$$

$$entropy = - \sum_{i=1}^{N_g} p(i) \log_2(p(i) + \epsilon) \quad (5)$$

$$energy = \sum_{i=1}^{N_p} (X(i) + c)^2 \quad (6)$$

$$uniformity = \sum_{i=1}^{N_g} p(i)^2 \quad (7)$$

2.3.2. Shape Features

Shape features quantify the geometric properties of the segmented region. Examples include volume, surface area, sphericity (Eq. (8)), compactness (Eq. (9)), major axis length (Eq. (10)), elongation (Eq. (11)), flatness (Eq. (12)) etc.

Assuming that V is the volume of the triangle mesh, A is the area of the triangle mesh and λ_{major} , λ_{minor} and λ_{least} are the largest, second largest and smallest principal components in the ROI,

$$sphericity = \frac{\sqrt[3]{36\pi V^2}}{A} \quad (8)$$

$$compactness = \frac{V}{\sqrt{\pi A^3}} \quad (9)$$

$$major\ axis = 4\sqrt{\lambda_{major}} \quad (10)$$

$$elongation = \sqrt{\frac{\lambda_{minor}}{\lambda_{major}}} \quad (11)$$

$$flatness = \sqrt{\frac{\lambda_{least}}{\lambda_{major}}} \quad (12)$$

2.3.3. Texture Features

These features capture patterns and relationships between voxels.

1) GLCM features describe how often pairs of voxel intensities occur in a specified spatial relationship.

- Contrast: Measures local intensity variation.
- Correlation: Assesses linear dependencies between voxel intensities.
- Homogeneity: Higher in smoother regions.
- Entropy: Indicates complexity or randomness.

2) GLRLM: These features quantify the length of consecutive voxels with the same gray level value.

- Short Run Emphasis: Higher in fine textures.
- Long Run Emphasis: Higher in coarse, homogeneous regions.
- Gray Level Nonuniformity: Indicates variability in gray levels.

3) GLSZM: These features measure the size of homogeneous zones of the same gray level. A gray level zone is defined as the number of connected voxels that share the same gray level intensity.

- Small Area Emphasis: High values correspond to fine-grained textures.
- Large Area Emphasis: Linked to broader, more uniform regions.
- Zone Percentage: Reflects the relative number of zones to total voxels. It measures the coarseness of the texture by taking the ratio of number of zones and the number of voxels in the ROI.

4) Gray Level Dependence Matrix (GLDM): It measures the dependence of a voxel on neighboring voxels with similar intensity.

- Small Dependence Emphasis: Reflects fine texture.

- Large Dependence Emphasis: Associated with coarse texture.

5) Neighboring Gray Tone Difference Matrix (NGTDM) features describe differences between a voxel and its neighbors.

- Coarseness: Measures level of fine vs coarse textures.
- Contrast and Busyness: Reflect abrupt changes in texture

Radiomics provides a revolutionary approach to ILDs by making it possible high dimensional, quantitative information extraction from HRCT scans that represent the lung structure [89]. It is particularly a useful technique for risk assessment, treatment response prediction, and longitudinal disease monitoring because it has been demonstrated to correlate with histological findings, lung function, disease progression, and clinical outcomes [90]. This is especially important in lung disorders since radiologic abnormalities such as consolidations, GGOs, and fibrosis are frequently spatially complicated and change over time [91]. Additionally, as the establishment for radiogenomics and multimodal machine learning techniques, radiomics makes it easier to integrate imaging with other data modalities, including genomics, proteomics, and clinical data [92].

Despite its potential, radiomics in ILD faces challenges. These include variability in image acquisition protocols, segmentation accuracy, and feature reproducibility [86]. Standardization, harmonization of preprocessing steps (e.g., resampling, normalization), and robust validation across external datasets are essential to ensure generalizability [93]. Harmonization is a mathematical postprocessing method used to eliminate the batch effect that center-specific variations caused by differences in acquisition parameters on radiomic features following image acquisition, reconstruction, and analysis [94]. Rather than being applied to the images themselves, harmonization targets the numerical values of radiomic features. The most widely used approach today is ComBat harmonization, originally developed for genomic data. It has since been validated for radiomics data to effectively reduce center-related variability while preserving underlying pathophysiological information [95]. Nevertheless, some studies report high model performance and reproducibility without the use of ComBat or similar harmonization techniques despite using the CT scan data from different vendors, particularly when acquisition protocols, slice thicknesses and pixel spaces are consistent and same or similar convolutional kernels are used [96].

Accurate segmentation defines the region of interest from which radiomic features are extracted, directly influencing the reliability and clinical relevance of the analysis. In ILDs, this usually involves the whole lung parenchyma, or fibrotic regions if subregion analysis wants to be performed. The choice of segmentation method significantly influences feature stability and biological interpretability [20]. Manual segmentation by experienced radiologists or trained experts is the traditional gold standard. It allows for precise delineation in complex cases, such as mixed fibrosis and emphysema, traction

bronchiectasis, or honeycombing. However, it is time-consuming, labor-intensive, and prone to inter- and intra-observer variability, which can reduce feature reproducibility [97]. To scale radiomics to larger cohorts, semi-automated or fully automated algorithms are increasingly used. Tools such as lungmask, TotalSegmentator, and U-Net-based deep learning models provide fast and reproducible segmentation of the lungs, lobes, and even detailed regions like airways or vessels [98, 99]. In ILDs, segmentation should ideally exclude major airways and pleural effusions and focus on parenchymal structures. However, fibrosis can distort anatomy, and algorithms trained on healthy datasets may misclassify fibrotic tissue or ground-glass opacities. Segmentation quality control is a critical step in the radiomics pipeline, regardless of whether manual, semi-automated, or fully automated methods are used. Ensuring accuracy involves the visual inspection of a representative subset of cases to identify gross errors or inconsistencies, as well as the use of quantitative overlap metrics such as the Dice similarity coefficient to compare automated segmentations against expert-drawn reference masks [97, 100]. Additionally, incorporating anatomical priors, such as known lung boundaries or fissure locations can help refine segmentation accuracy, especially in the presence of distorted anatomy due to fibrosis or volume loss. Thus, high-quality segmentation directly enhances the reproducibility of radiomics features and contributes to the generalizability of predictive models across datasets [20].

CT scans may vary in voxel size depending on slice thickness and reconstruction parameters. Radiomics features, especially texture-based features are sensitive to voxel size, which can introduce artificial variability [86, 93]. To address this, resampling is commonly performed to isotropic voxels (e.g., $1 \times 1 \times 1$ mm³). This process ensures that spatial relationships among voxels are consistent across scans and that texture features derived from matrices such as GLCM or GLRLM are comparable [101]. The interpolation method used during resampling also matters; linear interpolation is typically used for CT images. Nearest-neighbor interpolation is applied to segmentation masks to preserve discrete label values. It is important to resample both the image and the mask using the same spatial reference grid to avoid misalignment and loss of anatomical precision. CT image intensities are measured in Hounsfield Units, which are theoretically standardized. However, variations in scanner type, acquisition protocol, and patient positioning can still affect the appearance of images.

After completing radiomic image analysis, it is essential to select the most relevant features for building a statistical model to address the clinical question. To include all features extracted would result in a drastic increase in the number of model parameters, as the radiomic features often exhibit high correlations, indicating redundancy in the data. As a result, some features can be eliminated, while others can be grouped and represented by a single surrogate feature using techniques such as principal component analysis, linear discriminant analysis or statistical regularization [102].

Overfitting and underfitting are common pitfalls in machine learning that reflect a model's failure to generalize effectively [19]. Overfitting arises when a model with too many input parameters or excessive flexibility begins to memorize the training data. It captures not

only meaningful, disease-related features but also irrelevant noise and random variations. Despite performing well on the training data, it fails to generalize and performs poorly on unseen data. To prevent overfitting, techniques such as regularization can be used to constrain the model's complexity, or the number of input features can be reduced to limit the number of parameters. Detecting overfitting typically involves monitoring performance on a separate validation dataset: if training error continues to decrease while validation error starts to rise, it indicates that training should be stopped. In contrast, underfitting occurs when the model is too simplistic to accurately classify both training and validation data. In such cases, incorporating more informative data or selecting a more sophisticated model may be necessary. These challenges are particularly relevant in radiomics, where high-dimensional feature sets extracted from medical images increase the risk of overfitting, underscoring the importance of robust feature selection and model validation strategies.

To address overfitting and multicollinearity, penalized regression models introduce regularization terms to the loss function, enforcing constraints on the magnitude of the regression coefficients in radiomics analysis. Common penalties in radiomics are LASSO regression (L1 penalty) (Eq. (13)), ridge regression (L2 penalty) (Eq. (14)) and elastic net (Eq. (15)).

Assuming that $\mathcal{P}(\boldsymbol{\beta})$ is the penalty function, $\boldsymbol{\beta}$ is the model coefficient,

$$\mathcal{P}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j| \quad (13)$$

$$\mathcal{P}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^p \beta_j^2 \quad (14)$$

$$\mathcal{P}(\boldsymbol{\beta}) = \alpha \|\boldsymbol{\beta}\|_1 + (1 - \alpha) \|\boldsymbol{\beta}\|_2^2 \quad (15)$$

Here, LASSO encourages sparsity by shrinking many coefficients to exactly zero. Ridge regression does not yield sparse solutions but stabilizes the coefficient estimates in the presence of multicollinearity by shrinking them toward zero. Elastic Net combines the sparsity of Lasso and the grouping effect of Ridge. The tuning parameters λ (and α in Elastic Net) are typically chosen via cross-validation (CV). AUC, C-index, or mean squared error can be used as the CV metric. Linear regression, logistic regression, or Cox partial log-likelihood can be integrated with penalized models depending on the downstream tasks.

CHAPTER 3

LITERATURE REVIEW

3.1. COVID-19 and Multimodal Data Fusion Related Literature Review

Multiple factors, including age, comorbidities, and male sex, have been linked to severe disease in COVID-19. Radiological findings, laboratory parameters, and specific viral mutations can all influence the disease trajectory [103-108]. The most widely used biomarkers for the severity assessment are lactate dehydrogenase (LDH), aspartate aminotransferase, neutrophil count, neutrophil-to-lymphocyte ratio, troponins, creatine kinase, myoglobin, and D-dimer [109]. High lymphocyte counts and thrombocytosis, along with low ferritin, D-dimer, LDH, and aspartate aminotransferase levels detected at admission, have also been associated with a lower risk of death in COVID-19 patients requiring intubation and mechanical ventilation [110]. In severe COVID-19 cases, ferritin and LDH levels have been identified as prognostic biomarkers for the development of acute respiratory distress syndrome [111]. However, evidence remains limited regarding whether any of these biomarkers can serve as independent predictors of prognosis in COVID-19 [112].

As is well known, pathogen-associated virulence factors can influence the clinical presentation and disease severity by altering tissue tropism or increasing virulence [113]. Key changes, such as the D614G mutation in the spike protein, were identified through genomic surveillance early in the pandemic and linked to increased transmissibility and severe disease [114]. Subsequently, variants of concern, such as Alpha, Delta, and Omicron, emerged, each displaying distinct clinical characteristics, including differences in symptoms, viral load patterns, and risks of hospitalization or ICU admission [115]. These observations underscore the dynamic nature of SARS-CoV-2 and its capacity to alter disease manifestations at both the individual and population levels.

Before applying any machine learning model to genomic or amino acid sequences, it is essential to first convert these sequences into fixed-length numerical representations to enable embedding in a feature space. To achieve this, a range of techniques has been developed, generally categorized into alignment-based and alignment-free methods [116].

Alignment-free methods are further divided into word-based and information theory-based strategies. Word-based approaches analyze the frequency of k-mers (short subsequences) and compute similarity or dissimilarity scores based on these patterns, whereas information theory-based methods quantify the shared information between sequences using entropy or complexity measures. In contrast, alignment-based techniques, despite drawbacks like high computational cost and the assumption that homologous sequences exhibit conserved regions, continue to be widely used, particularly in phylogenetic analyses. During the COVID-19 pandemic, the Pango, Nextclade, and WHO classification systems have been widely used, and all these systems essentially rely on this methodology [56, 117]. They were used in a limited number of machine-learning prediction studies for creating viral feature embeddings [118, 119]. The use of NLP techniques has been recognized in prior research; however, earlier studies were often unimodal and focused on tasks, such as viral classification or evolution tracking [120-123]. Therefore, existing studies are limited in their use of NLP methods to integrate viral genomic data with other modalities and in comparing these approaches with traditional phylogenetic analyses.

Chest CT has been instrumental since the early stages of the COVID-19 pandemic, frequently revealing characteristic patterns such as bilateral GGOs, consolidations, and lesions with peripheral or posterior distribution. However, visual CT scoring systems often fall short in capturing the full extent and heterogeneity of pulmonary involvement, thereby limiting their ability to predict clinical deterioration or the need for intensive care accurately [124]. Radiomics provided an opportunity to address these gaps by extracting quantitative imaging features that capture textural and spatial information beyond the limits of visual perception [125]. Several radiomic studies have been conducted for the detection and screening of COVID-19, including differentiation from other lung infections and prediction of clinical outcomes such as hospital stay [126, 127]. These studies typically utilized CT-derived radiomic features combined with machine learning algorithms to build predictive models. For instance, Qi et al. analyzed CT scans from 52 COVID-19 patients to predict hospital length of stay, using logistic regression and Random Forest models [128]. Their approach achieved AUC values of 0.97 and 0.92 for logistic regression and Random Forest, respectively. In another study, Guiot et al. developed detection models that achieved a sensitivity of 78.9% and a specificity of 91.1% [129]. Similarly, Fang et al. constructed a radiomic signature to distinguish COVID-19 from CT images, yielding an AUC of 0.82 on the test dataset.

In recent years, the power of multimodal data fusion in biomedical research has become increasingly evident, with pioneering applications emerging particularly in oncology. The convergence of next-generation sequencing, whole-slide histopathology imaging, advanced radiological techniques, and the availability of standardized large-scale datasets, such as The Cancer Genome Atlas, has enabled comprehensive integrative analyses that improve diagnostic and prognostic accuracy [5, 130]. Inspired by these successes, the COVID-19 pandemic prompted a wave of multimodal research efforts that integrated imaging data, clinical metrics, plasma multiomics, lung mechanics, and demographic variables to build predictive models for disease severity and outcomes [131]. While these

studies have demonstrated the potential of early, intermediate, and late fusion strategies to outperform unimodal models in terms of risk stratification and mortality prediction, they also highlight critical challenges. These include the scarcity of high-quality, well-annotated datasets, inter-institutional variability in data acquisition, and the need for rigorous domain-specific preprocessing. Moreover, issues related to interpretability, generalizability, and regulatory compliance remain significant barriers to clinical adoption. As such, despite the promising results, the field of multimodal data fusion in COVID-19 is still developing, and further studies are needed to refine fusion strategies, ensure reproducibility, and facilitate integration into routine clinical workflows.

3.2. ILD and Radiomics Related Literature Review

Variables such as gender, age, and FVC have been validated as prognostic indicators across various ILD subtypes, and traditional prognostic models in ILDs, such as ILD-GAP, have long been used to estimate patient outcomes based on clinical and physiological parameters [132]. However, prognostication in ILDs remains a significant clinical challenge due to their heterogeneous nature, variable progression rates, and diverse underlying etiologies. In one study, the ILD-GAP model was enhanced by incorporating the monocyte ratio (ILD-GAPM), which improved 3-year event prediction and provided more accurate risk stratification compared to the original model [133]. In another study focused on microscopic polyangiitis-associated ILD, a model combining FVC, DLCO, and the presence of honeycombing on HRCT, demonstrating its clinical utility by predicting respiratory-related mortality [134]. Additionally, in CTD-ILD, serial measurements of the KL-6 biomarker showed strong correlations with functional decline and radiologic progression, underscoring its potential as a dynamic marker for disease monitoring [135].

Conventional visual assessment of CT scans remains inherently subjective, often limited by qualitative interpretation, interobserver variability, and reduced sensitivity to subtle or early changes in lung parenchyma [17]. Especially, imaging characteristics like reticulation, GGOs, or honeycombing may be difficult for radiologists to detect or measure, and small changes over time may go unrecognized [136]. Visual grading systems are semi-quantitative and might not fully represent the complexity or spatial heterogeneity of fibrotic remodeling, despite their therapeutic utility, as well. Thus, radiomics features may correlate with disease severity, histopathology, functional impairment, and clinical outcomes, and they can be integrated with clinical and functional data to enhance diagnostic precision and prognostication in ILDs [91].

More recent efforts have leveraged radiomics and quantitative imaging methods to enhance prognostication in ILDs, as radiomics plays a complementary role by extracting quantitative features, such as intensity, texture, and wavelet-based patterns, that are not discernible to the human eye, enabling detailed characterization of ILDs and enhancing model reproducibility [137]. By integrating radiomic features with clinical and physiological data (e.g., pulmonary function tests), ML models can identify patterns and

this fusion can uncover associations between fibrotic burden on CT and lung function decline, improving predictive models for disease progression and survival [138]. Studies have shown that certain texture features (e.g., GLCM entropy, GLSZM zone percentage) correlate with functional impairment such as FVC decline and can outperform conventional visual scores in prognostic modeling [90].

Many studies showed that nomogram models including radiomics and clinical features outperformed only radiomics or clinical features in different ILD subtypes [139, 140]. According to a CT-based radiomics nomogram for predicting overall survival in patients with CTD-ILD, the clinical model alone yielded C-indices of 0.747, 0.631, and 0.587 in the training, internal validation, and external validation cohorts, respectively. In contrast, the combined radiomics-clinical nomogram achieved higher C-indices of 0.800, 0.738, and 0.742 across the same cohorts, highlighting the added predictive value of radiomic features in survival modeling [141].

Furthermore, radiomics has the potential to support several key clinical applications in ILDs. It can aid in subtyping disease by differentiating patterns such as UIP, NSIP, or HP based on quantitative analysis of parenchymal texture. Haga et al. utilized pyradiomics features extracted from CT scans and trained models to predict cellular infiltration in fibrotic ILD biopsies. The model achieved 70% accuracy and an F1 score of 0.73 in external validation, demonstrating that texture features like GLCM, GLDM, and GLSZM correlate with histopathology [89]. Ash et al. showed that in their IPF cohort; higher percent high attenuating area, and mean lung density, lower skewness, kurtosis was associated with worse transplant-free survival [142]. In a study using radiomics features for staging scleroderma associated ILD, higher values in GLCM variance and GLSZM entropy features, lower values in the LL-GLCM-cluster shade feature indicated higher probability for classification as ILD, and the mean, skewness and kurtosis features showed a negative correlation with extensive disease [143]. A CT radiomics study (Anti-MDA5-positive dermatomyositis ILD) extracted wavelet and first-order features via pyradiomics, creating a radiomic risk score that was independently associated with 6-month mortality across internal and external cohorts [138].

Two studies highlight the prognostic significance of assessing pulmonary and cardiac pressures in patients with ILDs. A retrospective cohort study by Takahashi et al. on 74 patients with CTD-ILD found that mean pulmonary arterial pressure, measured via right heart catheterization, was an independent predictor of survival, irrespective of the specific type of connective tissue disease. They noted that even mild elevation of the mean pulmonary arterial pressure was common and had a significant impact on mortality [135]. Complementing these findings, Li et al. investigated the value of non-invasive cardiopulmonary ultrasound in a broader ILD population. Their research underscored that beyond pulmonary fibrosis severity, cardiac involvement, particularly right ventricular remodeling and dysfunction, is strongly correlated with a poorer prognosis. They found that right ventricular dimension and right ventricular global longitudinal strain were important prognostic markers and right ventricular global longitudinal strain being capable of detecting early right ventricular impairment even in the absence of overt

pulmonary hypertension. Together, these studies emphasize the critical role of evaluating both pulmonary hemodynamics and right heart function for risk stratification and predicting outcomes in patients with various forms of ILDs [144].

Only a limited number of studies have explored the association between pulmonary hypertension (PH) and radiomics features, and these investigations have predominantly utilized magnetic resonance imaging (MRI). It is not surprising that CT-based studies on PH and radiomics are scarce, as deep learning-based segmentation models for pulmonary arteries and veins have only recently begun to emerge [145]. One cine MRI study demonstrated that radiomic features extracted from the pulmonary artery may predict hemodynamic alterations within the vessel [146]. Another MRI study showed that radiomics features derived from the mid-left ventricular myocardium could non-invasively distinguish patients with and without PH [147]. To the best of our knowledge, pulmonary artery radiomics has not yet been explored using CT imaging in ILDs or other domains, and a comprehensive approach that examines multiple regions of interest within a single scan also appears to be lacking in the current literature.



CHAPTER 4

MATERIAL AND METHODS

4.1.COVID-19 Cohort

Data for this prospective cohort study were collected at Hacettepe University Adult Hospital in Ankara, Turkey, from December 22, 2020, to May 5, 2021.

The cohort consisted of two groups of patients: one group recruited as a part of the project entitled “Viral Genome Analysis in COVID-19 Patients and Genotyping of Genetic Variants Shown in the Literature Related to the Severe Course of the Disease in Humans” and the other group recruited within the scope of the Global Influenza Hospital Surveillance Network Project (GIHSN)-2020-21 in Turkey. Ethical approvals were obtained from the Institutional Ethics Committee with the code numbers GO 2021/02-22, GO 20-102, and GO 22-1211. Good clinical and laboratory practices were followed throughout the study in accordance with the Helsinki Declaration. Adult patients (≥ 18 years of age) hospitalized in the medical wards or ICUs who were positive for SARS-CoV-2 polymerase chain reaction (PCR) test within the last 120 hours and gave written informed consent were included in the study. COVID-19 patients with at least one of the below criteria were admitted to the ICU:

- Dyspnea and respiratory distress,
- Respiratory rate $>30/\text{min}$,
- Partial Oxygen Pressure / Inhaled Fractioned Oxygen $< 300 \text{ mmHg}$,
- Increased oxygen demand during follow-up,

- Oxygen saturation < 90% or Partial Oxygen Pressure < 70 mmHg under 5 l/min O₂ support,
- Hypotension (Systolic blood pressure <90 mmHg or more than 40 mmHg drop from usual systolic blood pressure level or mean arterial blood pressure <65 mmHg),
- Development of acute organ dysfunction such as acute kidney injury, acute elevation in liver function tests, confusion, acute bleeding diathesis,
- Elevated serum troponin with arrhythmia,
- Lactate >2 mmol/l.

Relevant clinical information was gathered through face-to-face interviews with patients and attending physicians and by reviewing clinical records. Age, sex, comorbidities such as diabetes, heart failure, coronary artery disease, hypertension, malignancy, PCR test results, vaccination history, medications and therapies, outcomes, laboratory results, GISAID EPI_SET identifiers, and imaging data were recorded. Age-adjusted Charlson comorbidity index (CCI) was calculated for each patient. Min-max normalization was performed for age and CCI. Since laboratory results can change within days or even hours in COVID-19 patients, only laboratory results at the time of imaging were included. Among recorded 22 laboratory features out of 127 patients, erythrocyte sedimentation rate (ESR) was missing in 5 (4%), and troponin-I and myoglobin were absent in 11 (8.7%) patients. The mean substitution was performed to handle these missing values.

4.1.1. Sampling and Next-Generation Viral Genome Sequencing Protocol

A nasopharyngeal swab or nasal specimen combined with an oropharyngeal swab was obtained from conscious patients, and a tracheal aspirate from intubated patients, in case they comply with inclusion criteria. Medical Wire M40-A Compliant Sigma-Virocult™ Viral Collection and Transport System combining open-bud Sigma-Swab™ with Virocult™ medium was used. EZ1 Virus mini kit V2.0 (Catalog number: 955134, Qiagen, Germany) was used for total nucleic acid extraction. Samples were directly introduced to the sequencing platform after nucleic acid extraction. Library preparation was performed using Respiratory Virus Oligos Panel V2 (Illumina Inc., #20044311) and Illumina RNA Prep with Enrichment, (L) Tagmentation (Illumina Inc., #20040537) kits according to official protocol. Before sequencing, the libraries were quantified and checked on Qubit 4 Fluorometer (ThermoFisher Inc.) and 2100 Bioanalyzer systems (Agilent Inc.). The sequencing was performed on Illumina NextSeq 550 platforms with 1,000,000 reads (2 × 150 bp) per sample on average. Quality control of the raw data was analyzed using the FASTQC tool. The quality passed samples were uploaded to BaseSpace (Illumina Inc.)

for bioinformatics analysis. The data were analyzed using the DRAGEN COVID Lineage app (v.3.5.4) on the BaseSpace platform. The low-quality and low-variant fraction variants were filtered out (coverage >10). The filtered variants were submitted to the GISAID database [55].

4.1.2. Identifying Viral Mutations and Construction of Phylogenetic Tree

Viral genome sequences were retrieved from the GISAID database. Nextclade CLI (v.2.12.0) was used for sequence alignment and identifying the nucleotide and amino acid mutations and variant clades [55]. The reference genome was determined as SARS-CoV-2 isolate Wuhan-Hu-1, GenBank: MN908947.3. Isolated study strains' phylogenetic tree construction was performed using an optimized substitution model (GTR + F + I) according to the lowest Bayesian Information Criterion (BIC) score obtained by the ModelFinder approach and followed by ultrafast bootstrap analysis (175 iterations) on the IQ-TREE software (v.1.6.12) [148-150]. Consensus tree annotation and visualization were then completed using the ggtree (v.3.8.0) R package [151].

Our approach contained the following steps (Figure 4):

- Using binary encoding and leveraging the Word2Vec NLP model for viral encoding
- Extraction of radiomics features from CT images.
- Performing sparse CCA to investigate relationships between data modalities and identify clinical phenotypes [31]
- Using cooperative learning to build prediction models for ICU admission [33]

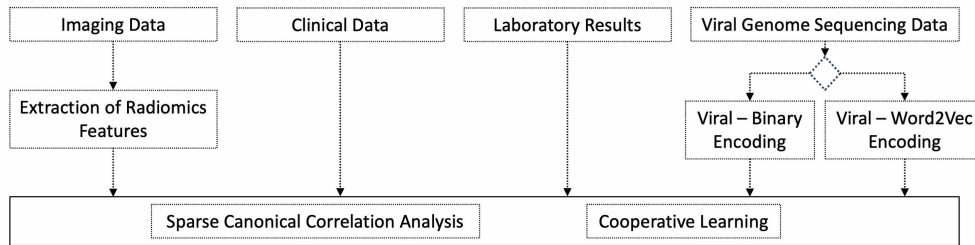


Figure 4: Flowchart of the Methods Used in the COVID-19 Cohort.

Unsupervised sparse canonical correlation analysis and supervised cooperative learning are used for multi-modal data fusion. Two different viral encoding techniques are performed and compared in these analyses.

Reprinted from Er et al. (2024), “Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study”, NPJ Digital Medicine, under CC BY license [152].

4.1.3. Viral Feature Preprocessing

We tried and compared two different techniques for viral encoding:

- **Viral-Binary encoding:** as a typical example of alignment-based methods, according to whether mutations were present or were not, each mutation was encoded as “0” or “1”. In total, 439 unique amino acid mutations were identified in 105 isolated study strains. This created a binary column for each mutation and returned a sparse matrix.
- **Viral-Word2Vec encoding:** since well-established viral genome databases exist on a global scale, we aimed to combine alignment-free and alignment-based models leveraging the Word2Vec NLP model to reduce the size of the embedding space and extract the semantic relationship between each of the mutations and the strains themselves [153]. We treated amino acid mutations as words and strains as sentences and used the Skip-Gram model architecture that predicts surrounding mutations in a context window given the current mutation. To construct the corpus, viral strains whose outcomes were known and collected between December 30, 2019, and March 2, 2023, on the GISAID database were used. After collecting the data, we defined the outlier strains as falling below $Q1 - 1.5 \text{ IQR}$ or above $Q3 + 1.5 \text{ IQR}$ in terms of the number of mutations. As a result, 653,134 viral genomes were used for constructing the corpus. The maximum number of mutations per strain was found to be 115, and accordingly, this number was chosen as the context window to train the Word2Vec model with vector dimension 300. The vocabulary size (the number of unique amino acid mutations) was found to be 52,311. Strain embeddings were calculated by getting the strains’ mean vector of amino acid mutations. Multi-dimensional scaling (MDS) and Uniform Manifold Approximation and Projection (UMAP) were used to reduce dimensionality and visualize 2D plots [154]. Cosine similarities between the strains were computed to construct a dissimilarity matrix, and this matrix was used as the precomputed dissimilarity metric in MDS and UMAP.

4.1.4. Imaging Data Collection and Analysis

For each patient with longitudinal CT images, the images were selected based on the following criterion: Since the study outcome for the supervised task is discharge from the medical ward (non-ICU group) or ICU admission (ICU group), we selected images closest to the outcome (Figure 5). Image data were obtained from SIEMENS scanners. CT section thickness was as follows: less than 1 mm (7 patients, 5.5%), greater than or equal to 1, less than 2 mm (119 patients, 93.7%), and greater than or equal to 2 mm, less than 3 mm (1 patient, 0.01%). All images were visualized in 3D Slicer (v.5.3.0), and the right and left lungs were segmented with U-net-based pre-trained lungmask R231 [98, 155]. To ensure all voxel intensities were positive before feature extraction, a voxel array shift of 1000 was applied to the CT images. Using pyradiomics, 18 first-order, 14 shape, 22 GLCM, 16

GLSZM, and 14 GLDM features were taken from each lung's original and wavelet-filtered images (v.3.0.1) [20]. All images were normalized with a scale of 500. “sitkBSpline” was used as the interpolator for resampling pixel space to (1.0,1.0,1.0) mm. A fixed bin number of 64 was used for all analyses.

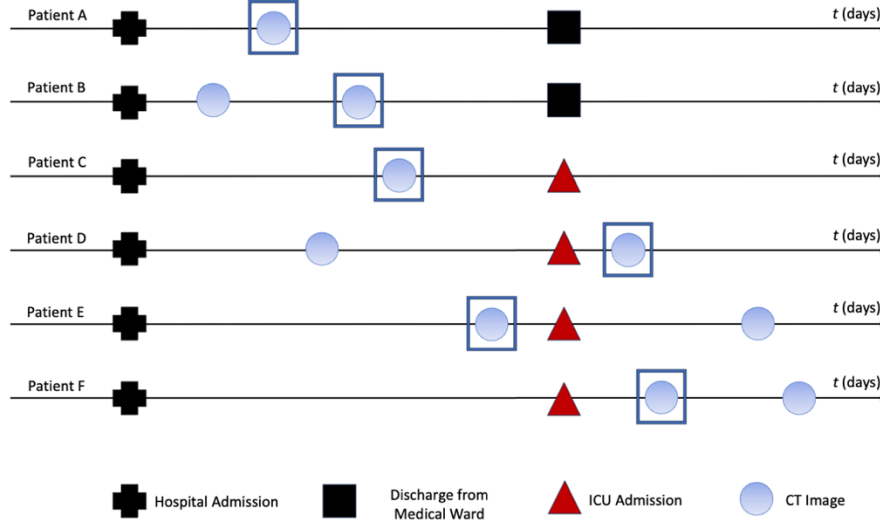


Figure 5: CT Image Selection Process.

Patients A and B are examples of the non-ICU group, whereas Patients C, D, E, and F are examples of the ICU group. The selected CT images for the study are shown in the frame. If a patient has more than one CT image, images closest to the discharge date from the medical ward or ICU admission were selected. For the sake of simplicity, hospital admission and outcome dates are aligned. *Reprinted from Er et al. (2024), “Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study”, NPJ Digital Medicine, under CC BY license [152].*

4.1.5. Data Fusion Methodologies

Sparse CCA was conducted using the PMA (v.1.2.1) package, where we used the CCA.permute function for selecting the sparsity parameters [31]. Multi-CCA was used as an extension where we assessed the relationship between more than two data modalities. The p values were calculated through permutation with the MultiCCA.permute function.

Cooperative learning was used for the supervised learning task to build prediction models for ICU admission with multimodal data, which uses an agreement penalty to encourage

alignment between predictions from different data modalities. We used logistic regression, as the response is binary.

Assume we have two data matrices, X and Z , of dimensions $n \times p_x$ and $n \times p_z$, and a corresponding target vector y . β_x is the vector of coefficients in X , and β_z is the vector of coefficients in Z . Logistic regression has the objective, which is given in Eq. (16), to be minimized

$$\ell(X\beta_x + Z\beta_z, y) + \rho/2 \|X\beta_x - Z\beta_z\|^2 + \lambda \left[(1 - \alpha) (\|\beta_x\|_1 + \|\beta_z\|_1) + \alpha (\|\beta_x\|_2^2 / 2 + \|\beta_z\|_2^2 / 2) \right] \quad (16)$$

This function consists of three parts. The first term is the loss function, which is the cross-entropy loss in our settings. The second term is the agreement penalty. It calculates the difference between the predictions made by each view independently. The hyperparameter ρ controls the strength of the penalty. The third term is the standard Elastic-Net regularization penalty. $\|\beta_x\|_1$ is the L1 norm, and $\|\beta_x\|_2^2$ is the L2 norm. λ is the overall strength of regularization, and α is the mixing parameter that balances the L1 and L2 penalties. Using CV, we chose the degree of agreement. Analyses were performed with the multiview (v.0.8) package [33].

We used a repeated stratified nested CV framework for hyperparameter tuning, model selection, and assessment (Figure 6). Ten-fold CV was performed, with the loss function as “deviance” for tuning the elastic-net mixing parameter and the weight of the agreement penalty in the inner loop. The outer loop assessed the performance of models trained in the inner loop. The final performance scores were averaged after a 5-fold CV. We conducted each experiment 30 times without matching patient splits in inner and outer loops.

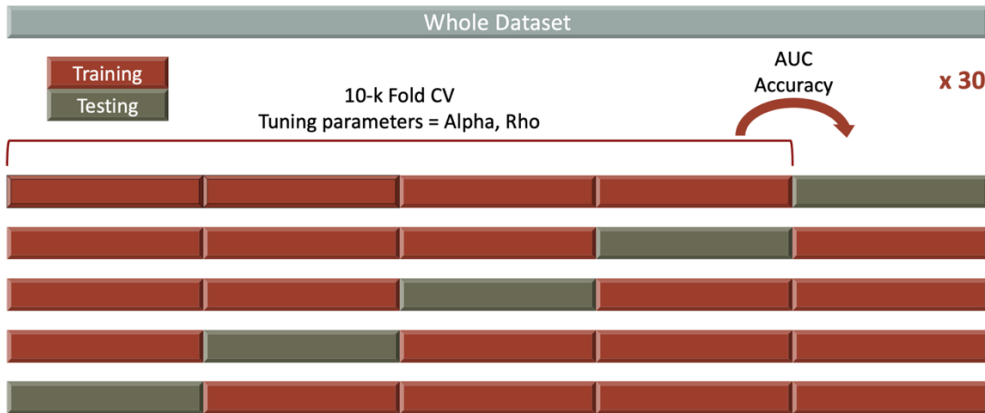


Figure 6: Repeated Stratified Nested Cross-Validation.

Reprinted from Er et al. (2024), “Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study”, NPJ Digital Medicine, under CC BY license [152]. 10-fold cross-validation (CV) was performed, with the loss function as “deviance” for tuning the elastic-net mixing parameter and the weight of the agreement penalty in the inner loop. The outer loop assessed the performance of models trained in the inner loop. AUC and accuracy metrics were calculated. The final performance scores were averaged after a 5-fold CV. We conducted each experiment 30 times.

4.1.6. Statistical Analysis

All the information related to patients was anonymized. Analyses were performed in a high-performance computing cluster provided by Stanford Research Computing Center. Descriptive statistics were used to calculate frequency and percent distributions. After testing assumptions of normality, the mean and standard deviation were used for continuous variables with normal distribution, and the median and interquartile range for continuous variables without normal distribution. The statistical significance of the differences between groups was tested using Chi-square and Fischer’s exact Chi-square tests for categorical variables, an independent two-sided t-test for continuous variables with normal distribution, and a Mann–Whitney U test for continuous variables without normal distribution. The equality of variances of the results of cooperative learning was assessed with Bartlett’s test. If the hypothesis of equal variances was rejected, Welch’s ANOVA was used to test the significance between three or more groups. The Games–Howell post hoc test was used as a nonparametric approach to compare pairwise results of cooperative learning. Type I error was set at 0.05 for all analyses. The source code and documentation for this part are publicly available at: https://github.com/ahmetgorkemer/multimodal_covid19_study.

4.2.ILD Cohort

4.2.1. Patient Population

This study utilized data from Stanford University’s ILD registry that includes clinical, functional test, and imaging information. Ethical approval was obtained from the Stanford University Institutional Review Board (IRB-57306). Patients were eligible if they had a confirmed diagnosis of ILD, as determined through multidisciplinary clinical discussions and evaluations. Demographic features (age, sex), ILD subtypes, FVC values, estimated sPAP from TTE, and follow-up time were recorded. The first CT scan taken during follow-up was considered the baseline. Each FVC was evaluated, and sequential FVC measurements underwent outlier analysis within themselves. FVC values below $Q1 - 1.5 \text{ IQR}$ or above $Q3 + 1.5 \text{ IQR}$ were defined as outliers. After removing outliers, only the FVCs performed within one year of the CT date and their corresponding CT scans were selected and analyzed. These FVCs were accepted as the baseline. Inclusion required the availability of follow-up data on mortality or lung transplantation. Patients were excluded

if imaging quality was insufficient, follow-up data were missing, or a lung transplant had occurred before the baseline date. All patient information was anonymized. Analyses were performed on a high-performance computing platform equipped with 87 CPUs, 503 GB of RAM, and a GPU with 40 GB of dedicated memory at Stanford University.

4.2.2. *CT Image Acquisition and Segmentation*

CT images were obtained in Digital Imaging and Communications in Medicine (DICOM) format, which contains both imaging data and associated metadata. We developed an algorithm to select the CT scans with the most diagnostically relevant images while filtering out low-quality, redundant, or non-lung-related slices. All HRCT scans were reviewed and preprocessed before radiomic feature extraction. Using the pydicom package (v.3.0.1), we extracted DICOM metadata including study and series descriptions, slice thickness, image type, reconstruction kernel, manufacturer, patient positioning, and voxel spacing to identify suitable CT series for analysis. Candidate series were selected based on the following criteria:

- The series type contained the keyword “original”
- Axial image resolution was 512×512 pixels
- The series included at least 30 DICOM slices
- The recorded patient position was either "FFS" (Feet-First Supine) or "HFS" (Head-First Supine)

As part of our image preprocessing pipeline, we aimed to ensure that each chest CT scan included the full extent of the lungs, rather than partial coverage limited to specific regions such as the lung bases. To achieve this, we utilized TotalSegmentator v2 (v.2.7.0) in its fastest mode to identify key anatomical structures. Specifically, we required the presence of the following labels in candidate series: `clavicula_left` (index 73), `clavicula_right` (index 74), `trachea` (index 16), `liver` (index 5), `lung_upper_lobe_left` (index 10), `lung_lower_lobe_left` (index 11), `lung_upper_lobe_right` (index 12), `lung_middle_lobe_right` (index 13), and `lung_lower_lobe_right` (index 14). Only series that contained all these anatomical regions were retained for subsequent analysis.

After anatomical filtering, we categorized the CT series based on their reconstruction kernels (Appendix A). Scans with sharp reconstruction kernels were grouped separately from those using non-sharp kernels to account for differences in image texture and sharpness that could affect downstream analysis. If the patient had multiple images in the same reconstruction kernel group, the one with the least slice thickness was selected. If the selected CT series had multiple reconstruction kernel groups, sharp kernels were selected.

We then converted the filtered DICOM series to Neuroimaging Informatics Technology Initiative (NIfTI) -1 format using the dicom2nifti package (v.2.6.2), ensuring consistent orientation in the (Left, Anterior, Superior) coordinate system. Series with irregular slice spacing or missing slices were excluded to maintain data integrity. After generating NIfTI-1 files, we tested four lung segmentation tools:

- Lungmask: A vanilla U-Net-based architecture [98].
- TotalSegmentator V2: An nn-Unet-based architecture trained on the TotalSegmentator dataset. We evaluated the default high-resolution version (resampling to $1.5 \times 1.5 \times 1.5$ mm) [156].
- MONAI Auto3Dseg: A SegResNet-based architecture, also trained on the TotalSegmentator dataset.
- MONAI Model Zoo Whole Body Segmentator: Another SegResNet-based architecture trained on the TotalSegmentator dataset.

Based on visual assessments of segmented lungs and binary masks using 3DSlicer, the TotalSegmentatorV2 (v.2.7.0) provided the best results. We also segmented the pulmonary arteries using TotalSegmentatorV2 (heartchambers_highres) after obtaining a non-commercial license from the authors of TotalSegmentator V2. (Figure 7)

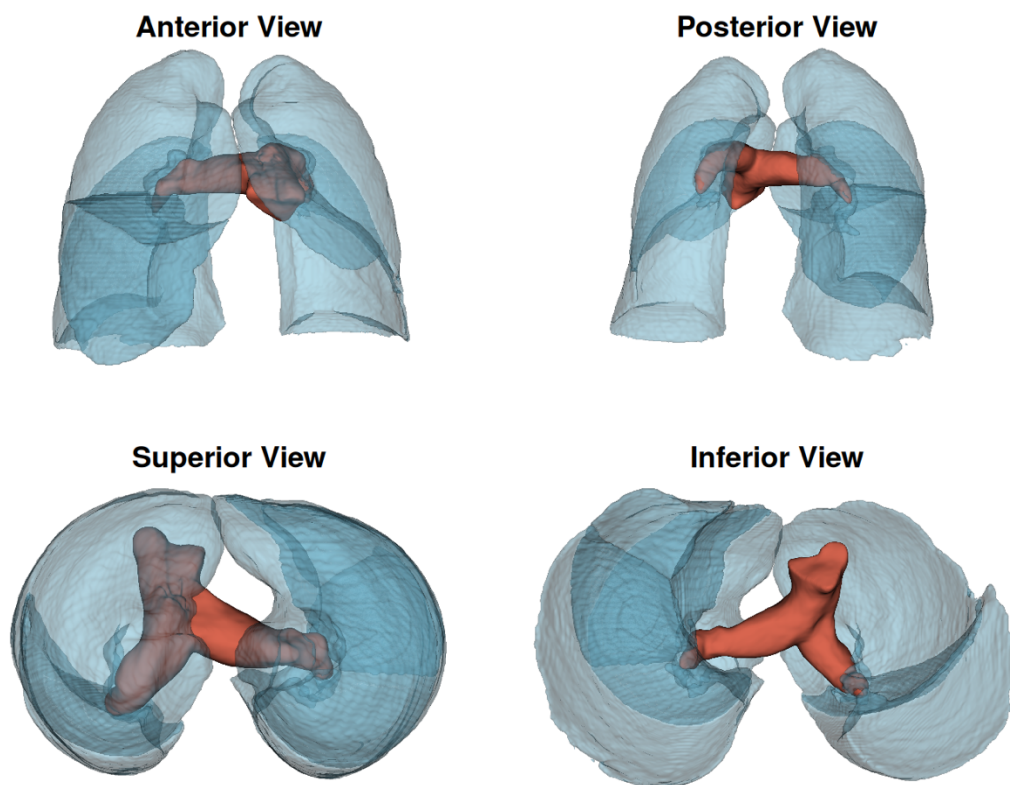


Figure 7: Anterior, Posterior, Superior, and Inferior Views of Binary Masks of Bilateral Lungs and Pulmonary Artery of a Real Patient.

To improve visualization of lung parenchyma and ensure consistent intensity representation across scans, lung windowing was applied using a window level (WL) of -600 and a window width (WW) of 1500 (Figure 8). Voxel intensities were clipped to this range while maintaining the original affine matrix. For the pulmonary artery feature extraction, a vascular windowing was applied with a WL of 200 and WW of 600 [157]. For labeling the contrast status of the images, we used the classifier XGBoost classifier developed from the TotalSegmentator team, which was designed to predict post-injection time. We labeled CT scans accordingly into native, early arterial, late arterial, and portal venous phases

Lung window

Vascular window

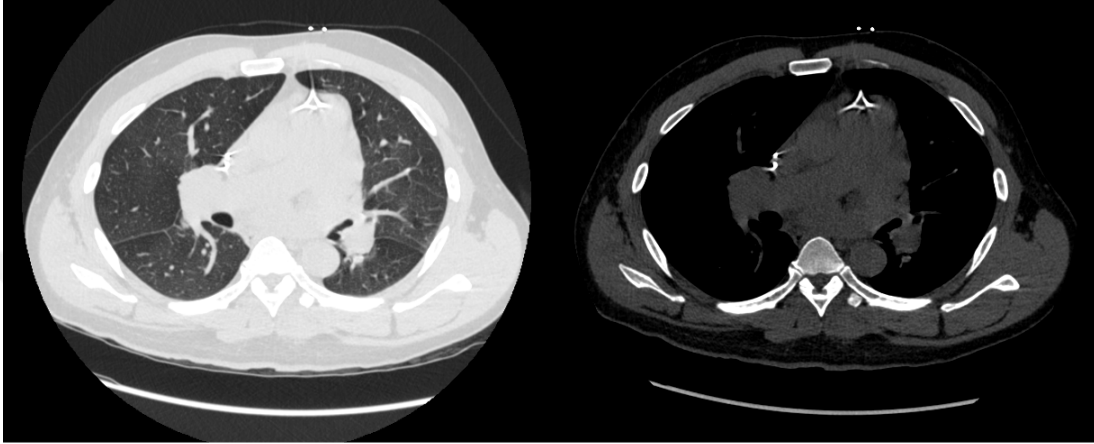


Figure 8: Axial CT Scan of a Real Patient Visualized Using Lung and Vascular Window Settings.

4.2.3. *Pyradiomics Feature Extraction*

We extracted handcrafted radiomics features using the pyradiomics library (v.3.0.1). These features were extracted from the original CT images using both original and wavelet filters. All images were normalized to a scale of 500. 'sitkLinear' was used as the interpolator for resampling pixel space to (1.0, 1.0, 1.0) mm. A fixed bin number of 25 was used for all analyses. Lung segmentation masks were interpolated using nearest-neighbor interpolation. To ensure all voxel intensities were positive before feature extraction, a voxel array shift of 1000 was applied to the CT images. A total of 14 shape-related features were extracted to quantify the geometric properties of the segmented regions from the lungs, as well as from the pulmonary artery. Furthermore, 18 first-order, 24 GLCM, 16 GLRLM, 16 GLSZM, 14 GLDM, and 5 NGTDM features were extracted from the original images and wavelet decompositions. To capture multi-scale texture patterns within the chest CT images, Wavelet decomposition was applied as part of the feature extraction pipeline using pyradiomics. Specifically, we used a discrete wavelet transform to decompose each image into eight frequency sub-bands, corresponding to combinations of high-pass and low-pass filtering along the x, y, and z axes (e.g., HHH, HHL, HLH, etc.). For each decomposed sub-band, radiomics features were computed separately. The only difference between lung parenchyma and pulmonary artery feature extraction was the initial windowing; the rest of the pipeline was identical.

4.2.4. *Statistical Analysis*

Descriptive statistics were used to summarize the data as frequencies and percentages for categorical variables. For continuous variables, normality was assessed. Variables with normal distributions were reported as mean and standard deviation, while those without normal distributions were summarized using median and interquartile range. Group differences were evaluated using the Chi-square or Fisher's exact test for categorical variables, and either an independent two-sided *t*-test or the Mann–Whitney *U* test for continuous variables, depending on distribution.

Cox proportional hazards regression models were used to investigate two outcomes: the development of pulmonary hypertension and transplant-free survival. Individuals who did not experience an event during the observation period were censored at the time of their most recent functional or clinical evaluation. To perform fixed-time survival analysis, all outcome times were administratively censored at five years. Observations with follow-up longer than five years were truncated, and their event indicators were censored beyond this time point. Multivariable penalized Cox models were used to evaluate the association between covariates and each outcome, using the glmnet package (v.4.1.9). To compare model performances, we employed a nested, stratified CV framework with five outer folds and ten inner folds. Stratification was based on the contrast phase. Within each training set, we tuned the elastic net penalty. For clinical variables, we employed the survival package (v.3.5.8). When combining clinical variables with radiomics features in the Cox model, we assigned a penalty factor of zero to the clinical variables to ensure they were not penalized during model fitting. Radiomics features were standardized using Z-score normalization. For each fold in the CV framework, mean and standard deviation were computed exclusively from the training set. These parameters were then applied to transform both the training and corresponding test sets to ensure that feature scaling was based solely on the training data to prevent information leakage into the test set. The concordance index (C-index) was used as the CV metric inside the inner loop. After selecting the models by choosing the one with the highest C-index in the inner loop, the final model performance was estimated by averaging the C-index across the outer test folds. Time-dependent AUC values were calculated at multiple time points in the outer test folds of the nested CV framework to assess the discriminative performance of the models over time. To statistically compare models trained on identical folds, we used the Wilcoxon signed-rank test on paired results. P-values were adjusted for multiple comparisons using the Benjamini–Hochberg procedure. Type I error was set at 0.05 for all analyses.

CHAPTER 5

RESULTS

5.1.COVID-19 Cohort

5.1.1. Descriptive Analysis

In total, 149 patients were enrolled in the study, and 63 patients (42.3%) were admitted to the ICU (Table 1). The mean age was 57.6 ± 16.2 years, and 61 patients (40.9%) were female. Hypertension (48.3%), diabetes mellitus (29.5%), and coronary artery disease (22.1%) were the most common comorbidities. The mean age was higher in the ICU group compared to the non-ICU group (63.9 ± 15.0 vs. 53.1 ± 15.6 , $p < 0.001$). The median CCI was 4 (3–6) in patients hospitalized in ICU compared to 2 (1–4) in those not hospitalized in ICU ($p < 0.001$). Age, sex, comorbidities, and CCI were used as clinical variables in downstream tasks.

Table 1: Clinical Characteristics of the Patients.

Modified from the Author's Clinical Microbiology and Infectious Diseases Thesis, Hacettepe University, 2024. Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study", NPJ Digital Medicine, under CC BY license [152].

Variable	ICU (+) (n: 63)	ICU (-) (n: 86)	p-value
Age, years, mean (SD)	63.9 (15.0)	53.1 (15.6)	< 0.001
Female sex, n (%)	26 (41.3)	35 (40.7)	1

Table 1 (cont.)

Comorbidities, <i>n</i> (%)			
Hypertension	38 (60.3)	34 (39.6)	0.02
Diabetes mellitus	22 (34.9)	22 (25.6)	0.29
Coronary artery disease	22 (34.9)	11 (12.8)	0.003
Solid organ malignancy	13 (20.6)	19 (22.1)	0.99
Kidney disease	13 (20.6)	10 (11.6)	0.20
Cardiac failure	14 (22.2)	6 (7.0)	0.01
Chronic lung disease	12 (19.0)	7 (8.1)	0.08
Hematologic malignancy	4 (6.3)	7 (8.1)	0.76
Rheumatologic disease	4 (6.3)	3 (3.5)	0.46
Cerebrovascular disease	2 (3.2)	3 (3.5)	1
Peripheral arterial disease	1 (1.6)	1 (1.2)	1
Dementia	2 (3.2)	0 (0.0)	0.18
Liver disease	0 (0.0)	2 (2.3)	0.50
Immunocompromised, <i>n</i> (%)	12 (19.0)	16 (18.6)	1
COVID-19 vaccination history, <i>n</i> (%)	14 (22.2)	8 (9.3)	0.05
Age-adjusted Charlson comorbidity index, median (IQR)	2 (1-3)	1 (0-1)	< 0.001
Steroid treatment (Minimum 6 mg/day dexamethasone), <i>n</i> (%)	51 (81.0)	23 (26.7)	< 0.001
Hospital LOS, days, median (IQR)	15 (10.5-31.5)	10 (6-14.75)	< 0.001
In-hospital mortality, <i>n</i> (%)	19 (30.2)	1 (1.2)	< 0.001
Abbreviations: ICU, Intensive care unit; LOS, Length of stay; IQR, Interquartile range			

The number of patients who underwent chest CT at least once was 127 (Figure 9). In 105 isolated viral genome samples, the number of unique nucleotide mutations was 710, and the number of unique amino acid mutations was 439. The median of number of nucleotide mutations per strain was 29.0 (21.0–33.0), and the median of number of amino acid

mutations was 22.5 (12.0–27.8) (Appendix B). Fifty-two strains (49.5%) were assigned to Variant 20I (Alpha, V1) according to the Nextclade clades (Figure 10).

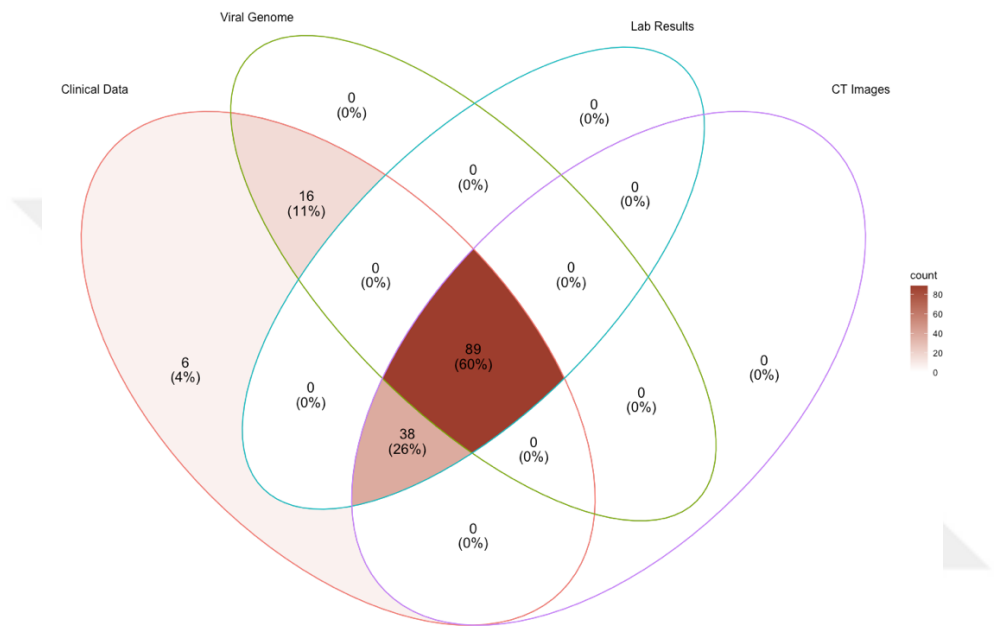


Figure 9: The Number of Patients in the Cohort with Different Data Modalities.

Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study", NPJ Digital Medicine, under CC BY license [152].

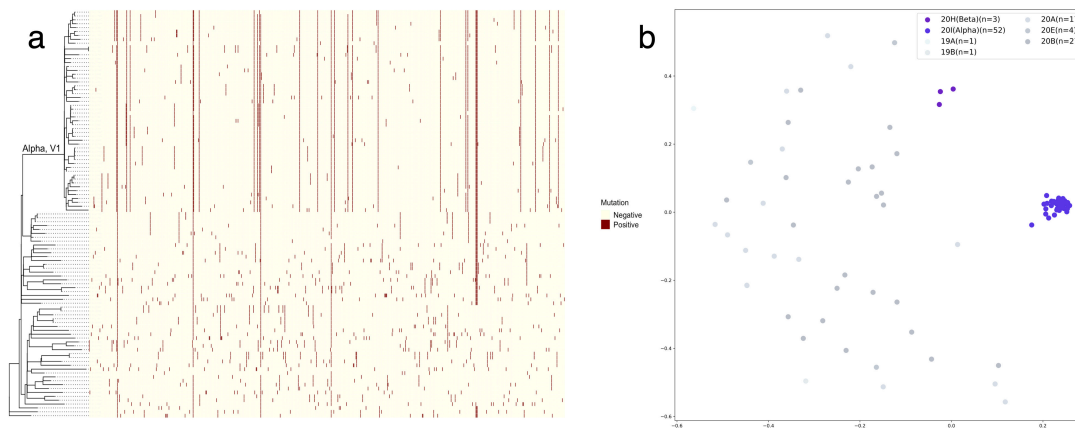


Figure 10: Phylogenetic Tree, Nucleotide Substitution Matrix, and Word2Vec Encoding Plot of Isolated SARS-CoV-2 Strains.

a The phylogenetic tree of isolated SARS-CoV-2 strains and nucleotide substitutions in matrix form, in which the presence of substitutions is shown in dark red. b The Word2Vec encoding plot of the same strains. The nucleotide substitution matrix and Word2Vec encoding plot represent that Alpha strains appear more similar compared to non-Alpha strains. *Modified from the Author's Clinical Microbiology and Infectious Diseases Thesis, Hacettepe University, 2024. Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study", NPJ Digital Medicine, under CC BY license [152].*

5.1.2. Visualization of the Global SARS-CoV-2 Strains Using Word2Vec Embedding

We visualized 300 randomly selected viral strains from each Nextclade clade in the corpus, generated with global viral genome sequences from the GISAID database. This shows that major variants, such as Variants 20I (Alpha, V1), 20H (Beta, V2), 21I, and 21J (Deltas), and Omicron clades, were successfully separated (Figure 11). Not only separation but also some of the characteristic features seen in phylogenetic relationships were observed in the embedding space. While more ancestral clades, for instance, Variants 19A, 19B, and 20A, representing the early days of the pandemic, had a wider distribution, clades dating to later periods tend to be observed within clusters. Furthermore, clades that have closer evolutionary relationships, for example, Variants 20H (Beta, V2), 21I (Delta), and 21H (Mu) were located closer in the embedding space [158]. On the other hand, Omicron variants were separated from these groups, as highlighted in the phylogenetics literature [159]. It was also seen that some recombinant strains were located close to the

Omicron variants, while others tended to spread toward other variants, representing that recombinant variants were not homogenous.

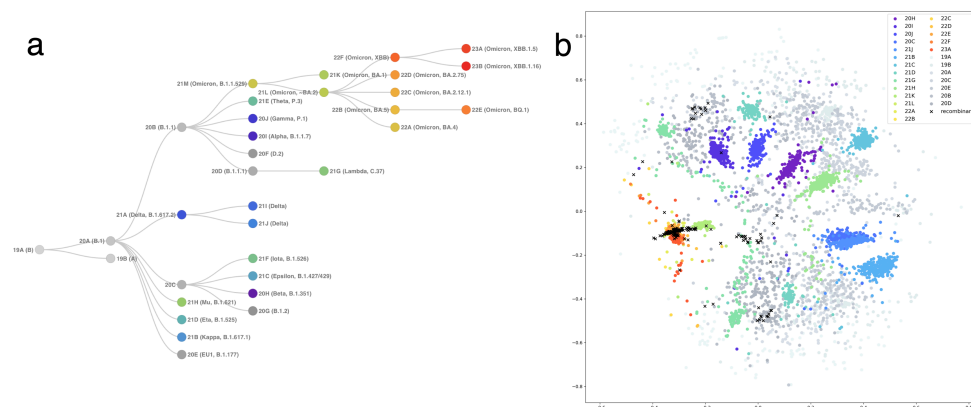


Figure 11: Phylogenetic Tree and Word2Vec Encoding Plot of Global SARS-CoV-2 Strains.

a The phylogenetic relationships of the global SARS-CoV-2 clades as defined by Nextstrain. The screenshot was taken from CoVariants.org 26. b the Word2Vec encoding plot of 300 randomly selected viral strains from each Nextclade clade. Major variants, such as Variants 20I (Alpha, V1), 20H (Beta, V2), 21I, and 21J (Delta's), and Omicron clades, are successfully separated. *Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study", NPJ Digital Medicine, under CC BY license [152].*

We also computed the Word2Vec embeddings of Nextclade clades in the UMAP plot shown in the Figure 12. As seen in the plot, the variants were well-separated: recombinant strains tend to cluster near the center, while Omicron (BA.1) and other clades were localized in distinct regions.

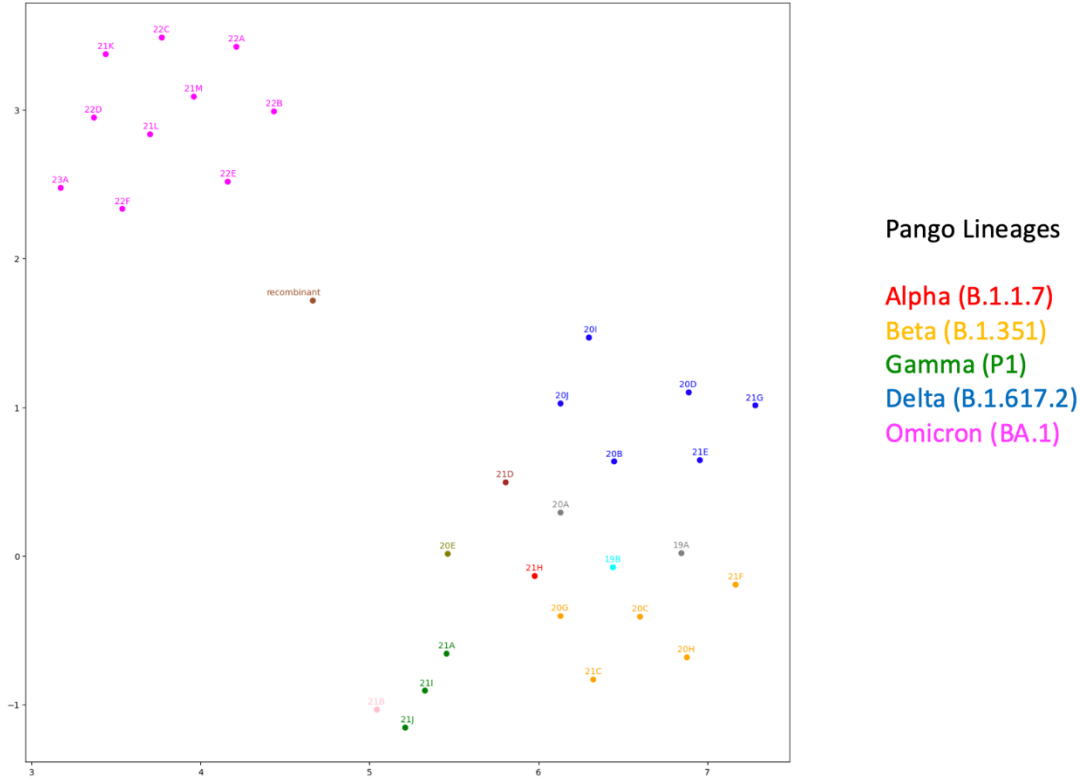


Figure 12: Word2Vec Encoding Plot of Nextclade Clades and Overlapped Pango Lineages in UMAP Plot.

5.1.3. Unsupervised Pairwise Data Fusion Using Sparse CCA

Next, we performed sparse CCA to examine the pairwise associations between all data modalities (Table 2). Relevant sparsity parameters corresponding to the highest Z-stat score were determined with the number of non-zero weights of X and Z. We first report the results for combining laboratory results and radiomics features for 127 patients ($\text{cor}(X_{u1}, Z_{v1}) = 0.596$). In the laboratory results group, lactate dehydrogenase (LDH), which relates to disease progression and worse outcome, had the highest coefficient value (0.47), followed by erythrocyte sedimentation rate, D-dimer, polymorphonuclear leukocytes, white blood cell count, and acute phase reactants such as CRP and fibrinogen [160]. At the same time, albumin had the lowest coefficient value (-0.46) as a negative acute phase reactant, along with hemoglobin, lymphocyte, and sodium levels. In the radiomics features group, original and wavelet features in the LLL frequency channel had the highest absolute values of coefficients (Figure 13). Among them, histogram-based first-order features reporting the skewness, kurtosis, and uniformity had the lowest negative coefficients, whereas entropy-related features had the highest positive coefficients. The negative coefficients of skewness and kurtosis features indicated that, as laboratory results worsened, since image density is measured approximately 1000

Hounsfield units (HU) for air, and 0 to +70 HU for various tissue types such as blood, pleural effusion, abscess, and mucus, depending on the extent and content of the lung abnormalities, the image intensity histogram became flatter and more right-skewed [161, 162]. This led to a loss of homogeneity and increased entropy in the radiomics features (Figure 14).

Table 2: Sparse CCA for Examining Pairwise Associations Between All Data Modalities.

Sparsity parameters and correlations were calculated with the CCA.permute function. We chose the L1 bound for X and Z at the highest value of the Z-stat for each pairwise data. *Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study", NPJ Digital Medicine, under CC BY license [152].*

Data Modalities		n	Best L1 bound for X and Z	Z-stat	p-value	Number of non-zero weights		Correlation
Radiomics	Lab Results	127	0.70 / 0.70	2.882	<0.01	1199	17	0.596
Radiomics	Clinical Data	127	0.10 / 0.10	3.423	<0.01	26	1	0.646
Radiomics	Viral-Binary E.	89	0.50 / 0.50	0.558	0.16	462	235	0.761
Radiomics	Viral-Word2Vec E.	89	0.70 / 0.70	0.584	0.20	775	241	0.524
Lab Results	Clinical Data	127	0.70 / 0.70	1.392	0.04	16	21	0.628
Lab Results	Viral-Binary E.	89	0.23 / 0.23	1.012	0.24	2	201	0.915
Lab Results	Viral-Word2Vec E.	89	0.10 / 0.10	0.489	0.20	1	6	0.576
Clinical Data	Viral-Binary E.	105	0.3 / 0.3	3.281	<0.01	17	328	0.982
Clinical Data	Viral-Word2Vec E.	105	0.7 / 0.7	0.214	0.36	20	206	0.487
Abbreviations: n, number; E., encoding								

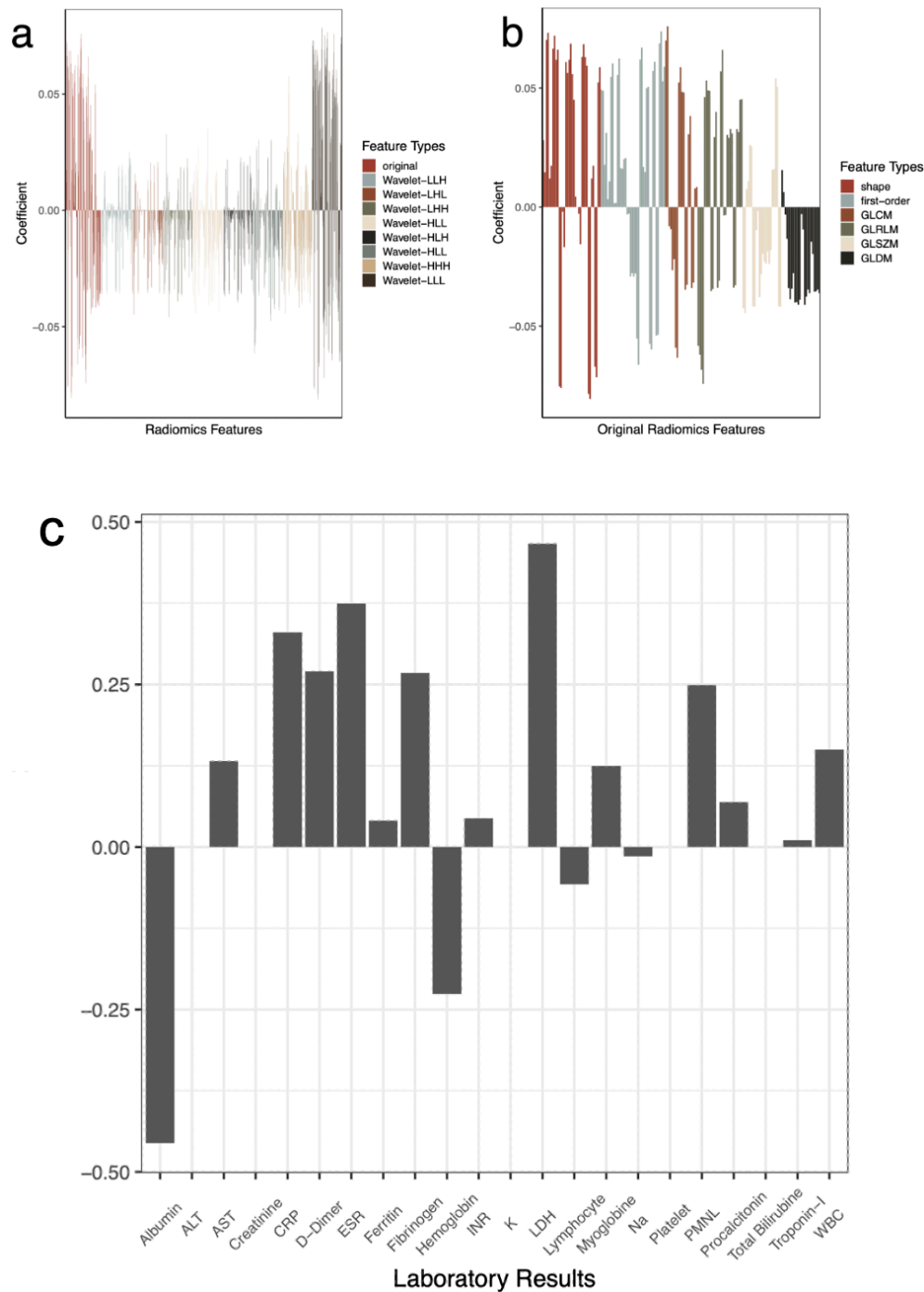


Figure 13: Sparse CCA of Radiomics Features and Laboratory Results.

a Correlated radiomics features. Original and wavelet features in the LLL frequency channel have the highest absolute values of coefficients. b Coefficients of the original radiomics features. c Correlated

laboratory results. Coefficients of laboratory results align with serum biomarkers related to severe disease and acute phase response.



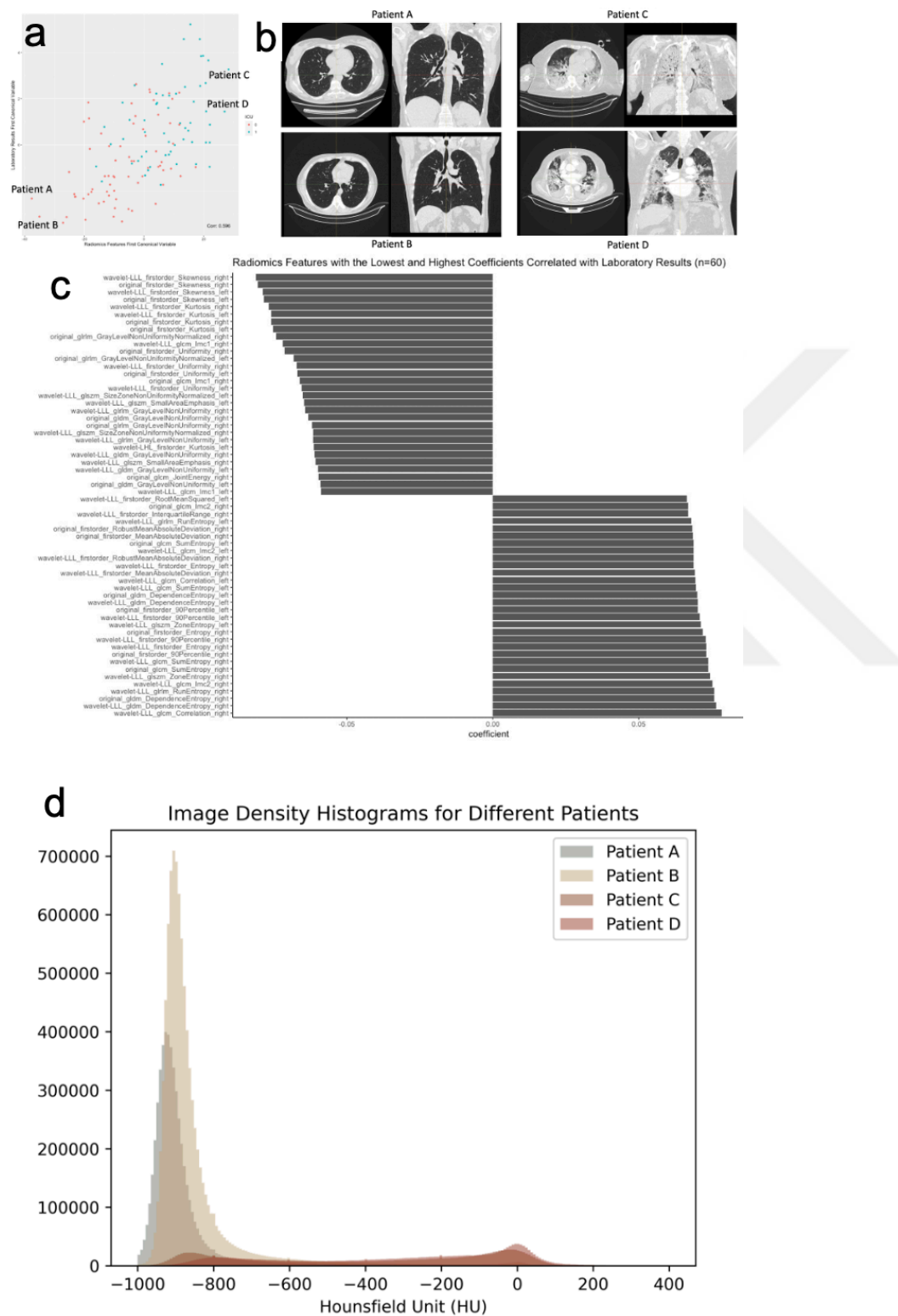


Figure 14: Correlation Plot Derived from Sparse CCA of Radiomics Features and Laboratory Results and Image Density Histogram of Example Patients.

a The correlation between the first set of canonical variables shows that the first pair can capture the ICU outcome. We select two patients (Patient A and Patient B) with the lowest and two patients (Patient C and Patient D) with the highest canonical variables for radiomics features b Patient A and Patient B's CT images in axial and coronal planes have no pulmonary infiltration, whereas there are apparent findings on Patient C and Patient D's CT images for COVID-19 pneumonia. c We select and visualize thirty variables with the highest and thirty variables with the lowest coefficients among the radiomics features. d The image intensity histograms of the patients show that Patient A and Patient B have left-skewed histograms peaking around -1000 to -800 HU, consistent with air and lung parenchyma densities; however, histograms of Patient C and Patient D are flatter and more right-skewed, consistent with negative coefficients for skewness and kurtosis features and revealing a wider distribution of HU values. *Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study", NPJ Digital Medicine, under CC BY license [152].*

Sparse CCA of laboratory results and clinical data revealed different clinical phenotypes. ($\text{cor}(X_{u1}, Z_{v1}) = 0.63$, best L1 bound for X and Z: 0.7, $p = 0.04$) (Figure 15). While laboratory results' first canonical coefficients revealed a phenotype related to high creatinine levels with low albumin levels and anemia, on the clinical side, the patient appeared to be elderly and multi-morbid with moderate to severe renal disease. The second canonical variables represented a different patient phenotype who was young and immunocompromised with high ferritin levels. The third canonical variables also characterized a phenotype that likely presents a patient with a history of liver disease with elevated INR, total bilirubin levels, and hypoalbuminemia.

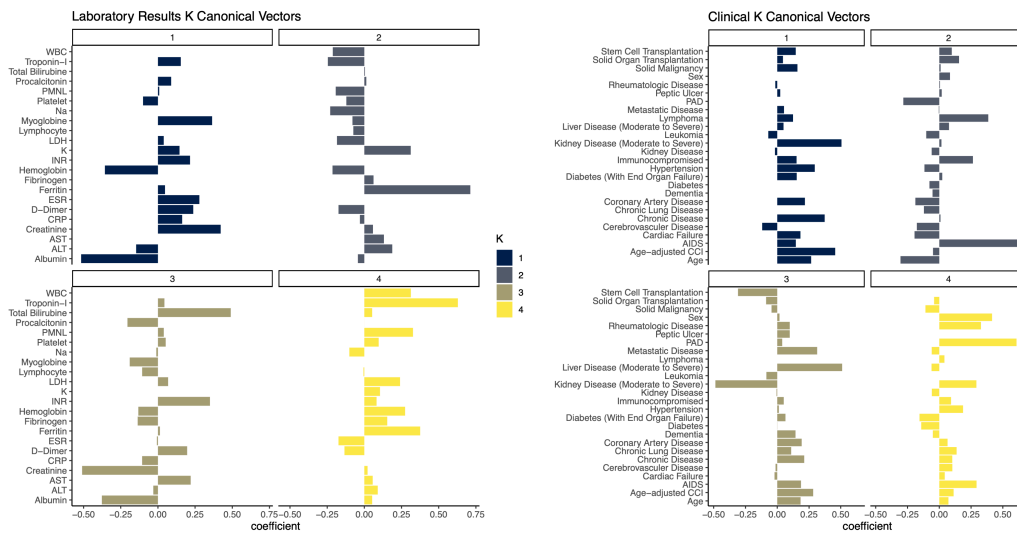


Figure 15: Sparse CCA of Laboratory Results and Clinical Data.

The first four canonical variables are provided. Different canonical variables provide different clinical phenotypes: The first canonical variables represent a patient phenotype who is elderly, multi-morbid, and has moderate to severe renal disease with high creatinine and myoglobin levels, whereas the third canonical variables represent a different patient phenotype with moderate to severe liver disease with high bilirubin and INR levels. *Adapted from the Author's Clinical Microbiology and Infectious Diseases Thesis, Hacettepe University, 2024. Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical*

correlation analysis and cooperative learning: a COVID-19 cohort study”, NPJ Digital Medicine, under CC BY license [152].

Next, we focused on the sparse CCA of imaging and clinical data (Appendix C). In this experiment, the permutation-based approach for choosing parameters provided a sparser solution than previous ones; the correlation coefficient was 0.65 ($p < 0.01$). The first and third canonical vectors were found to be associated with sex, and the second and fourth ones were with lymphoma and CCI, respectively. Mainly, the first sex-related canonical vector coefficients belonged to the left, and the second sex-related canonical vector coefficients belonged to the right lung and consisted of first-order and shape radiomics classes. In combinations of pairwise sparse CCA of viral genome sequencing data with other data modalities, different encoding techniques for the viral genome altered the correlation plots, with better separation obtained with Viral-Word2Vec encoding (Figure 16).

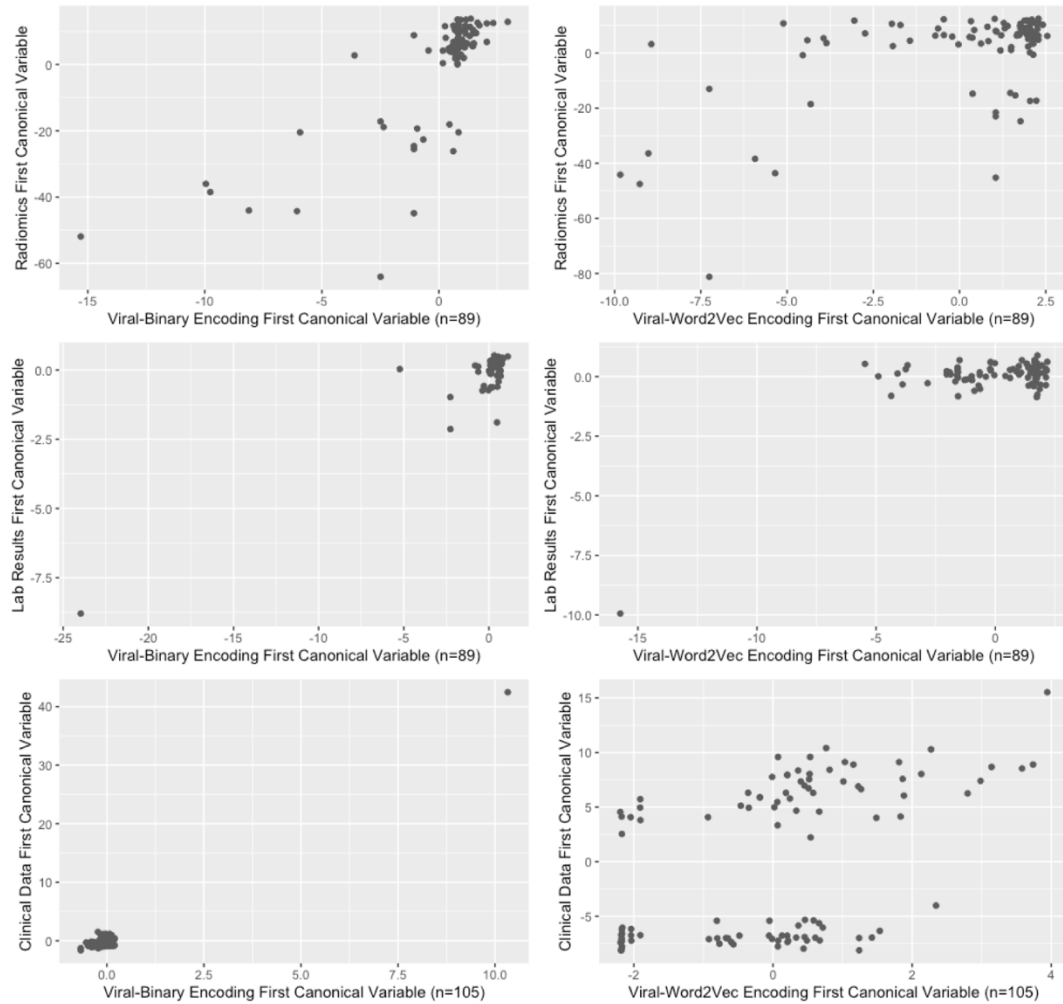


Figure 16: Sparse CCA of Viral Genome Sequencing Data with Different Data Modalities.

The left side shows the first canonical variables of Viral-Binary encoding, and the right side shows Viral-Word2Vec encoding. As it is seen, different encoding techniques for the viral genome change the correlation plots, yielding a better separation with Viral-Word2Vec encoding. Only the correlation plot at the bottom left was found to be significant ($p < 0.01$). Reprinted from Er et al. (2024), “Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study”, *NPJ Digital Medicine*, under CC BY license [152].

Finally, we performed sparse multi-CCA on patients with all data modalities ($n = 89$). When Viral-Binary encoding was used for viral encoding, the highest Z-score was 2.15 (penalties = 12.22, 6.64, 1.56, 1.7, $p < 0.01$). The highest Z-score was 3.14 (penalties = 17.923, 8.468, 2.293, 2.493, $p < 0.01$) when Viral-Word2Vec was used for viral encoding. We showed that with the Viral-Word2Vec embedding, the viral genome projection of the patients was more homogeneously distributed, and pairwise correlation values of viral

features and other data modalities were reduced (Figure 17). While the dominance of histogram-based first-order radiomics features and the importance of albumin and LDH among the laboratory results persisted in the analysis of these 89 patients, clinical data's first canonical vectors slightly revealed a different clinical phenotype, which was an elderly multimorbid patient with dementia (Figure 18).

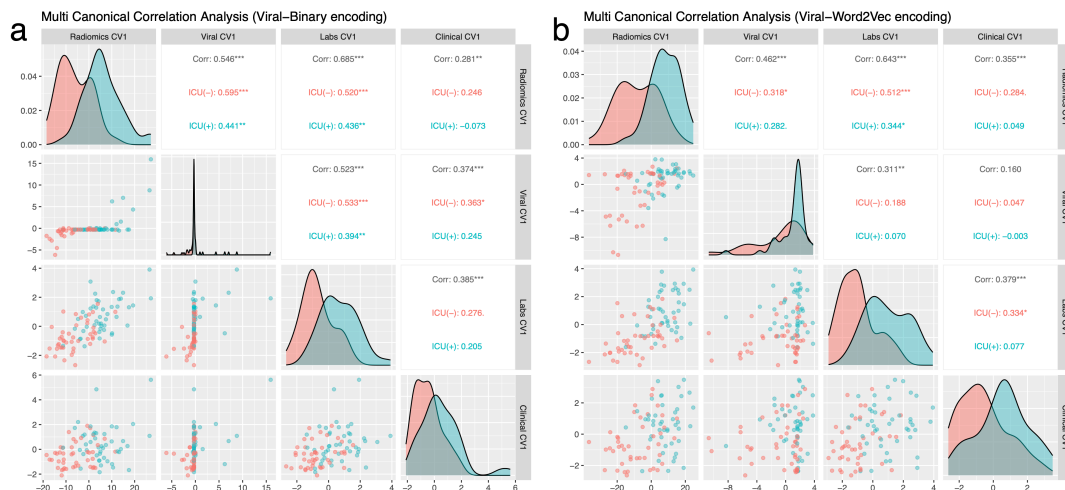


Figure 17: Sparse Multi-CCA of All Data Modalities.

a The correlation pairs plot of the first canonical vectors of four data modalities, including Viral-Binary encoding. b Using Viral-Word2Vec encoding instead of Viral-Binary encoding provides a more homogenous distribution and better separation among canonical variables. *Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study", NPJ Digital Medicine, under CC BY license [152].*

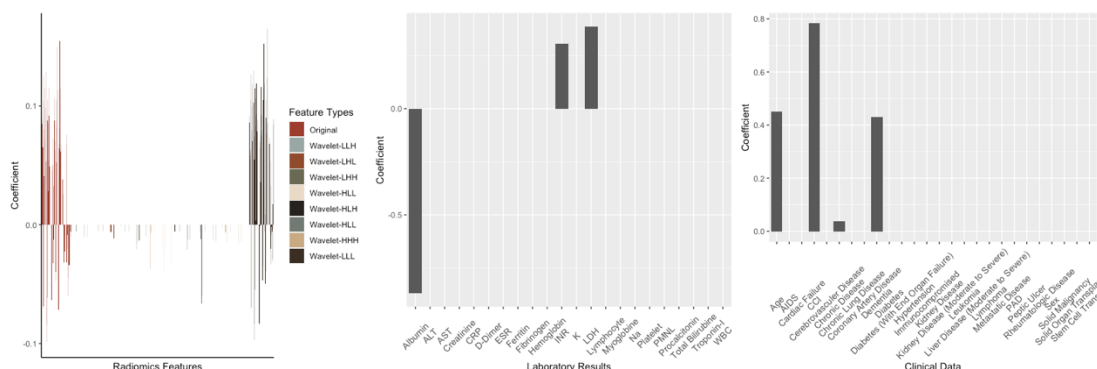


Figure 18: The First Canonical Vectors of Multi-CCA.

While the dominance of histogram-based first-order radiomics features and the importance of albumin and LDH among the laboratory results persist in the analysis of these 89 patients, as seen in the example of sparse CCA of radiomics features and laboratory results, clinical data's first canonical vectors slightly reveal

a different clinical phenotype of an elderly multimorbid patient with dementia. *Reprinted from Er et al. (2024), “Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study”, NPJ Digital Medicine, under CC BY license [152].*

5.1.4. Supervised Multiview Analysis Using Cooperative Learning

We performed cooperative learning to build prediction models for ICU admission. First, we used all five data modalities to train unimodal prediction models in all patients. The highest score was achieved with the model using radiomics features ($AUC = 0.83 \pm 0.01$), followed by laboratory results and clinical data ($AUC = 0.77 \pm 0.02$ vs. 0.67 ± 0.02), respectively (Table 3 and Figure 19).

Table 3: Unimodal Prediction Models Evaluated in All Patients.

Reprinted from Er et al. (2024), “Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study”, NPJ Digital Medicine, under CC BY license [152].

Model	<i>n</i>		Accuracy	AUC
	ICU (+)	ICU (-)		
Clinical Data	63 (42%)	86 (58%)	0.62 ± 0.02	0.67 ± 0.02
Radiomics Features	57 (45%)	70 (55%)	0.72 ± 0.02	0.83 ± 0.01
Laboratory Results	57 (45%)	70 (55%)	0.68 ± 0.02	0.77 ± 0.02
Viral-Binary Encoding	49 (47%)	56 (53%)	0.54 ± 0.03	0.58 ± 0.05
Viral-Word2Vec Encoding	49 (47%)	56 (53%)	0.50 ± 0.03	0.52 ± 0.04
Abbreviations: ICU, Intensive care unit; AUC, Area under the curve				

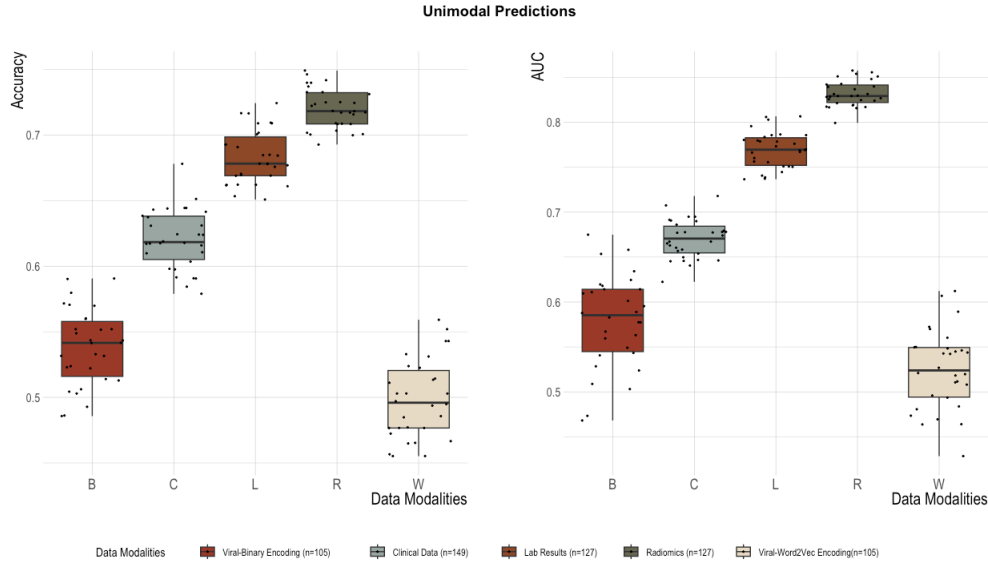


Figure 19: Unimodal Prediction Models for All Data Modalities in All Patients.

Radiomics features have the highest score ($AUC=0.83 \pm 0.01$), followed by laboratory results and clinical data ($AUC=0.77 \pm 0.02$ vs. 0.67 ± 0.02), respectively. *Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study", NPJ Digital Medicine, under CC BY license [152].*

Then, the models were trained with all available singular, dual, and quad data combinations and evaluated for the same task in 89 patients. The best accuracy and AUC values were achieved with the quadruple model, which combines clinical, laboratory, radiomics, and viral genome sequencing data using Viral-Word2Vec encoding (CLRW) (Appendix D and Figure 20). While unimodal prediction models of radiomics features and laboratory results had a mean AUC score of 0.83, the quadruple prediction model using Viral-Word2Vec encoding achieved a mean AUC score of 0.87. The mean AUC score for the quadruple model using Viral-Binary encoding was 0.83 ± 0.2 . There was a statistical difference between these two quadruple models ($p < 0.001$). The model that combines radiomics and Viral-Word2Vec encoding also outperformed the model that combines radiomics and Viral-Binary encoding in terms of AUC scores (0.86 vs. 0.83 , $p < 0.001$). There was no statistical difference when we used different viral embedding techniques in other models with dual data combination.

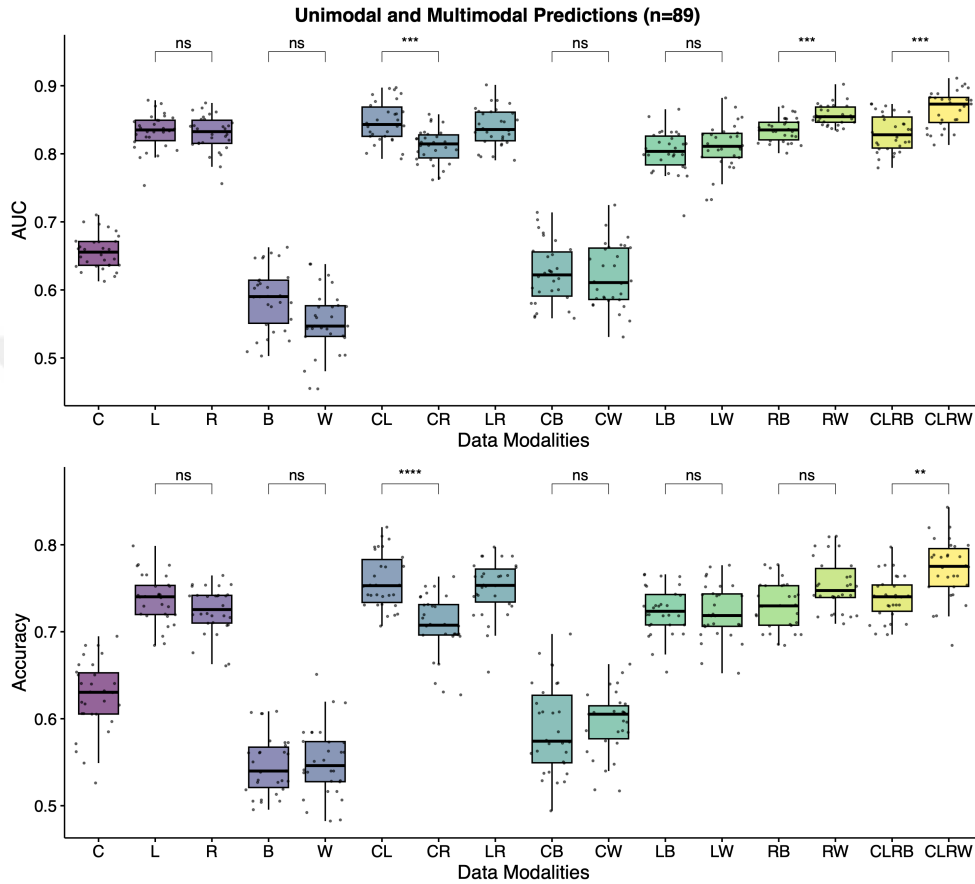


Figure 20: Unimodal and Multimodal Prediction Models for the Supervised Task.

The best accuracy and AUC values are achieved with the quadruple model using Word2Vec Encoding (CLRW). Abbreviations: C, Clinical data; L, Laboratory results; R, Radiomics; B, Viral-Binary encoding; W, Viral-Word2Vec encoding; AUC, area under the curve; ns, non-significant. *Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study", NPJ Digital Medicine, under CC BY license [152].*

To understand which features were the most important for the performance of the CLRW model, we used all the available data as the training set. We performed a 5-fold CV to get the optimal hyperparameters and found Alpha and rho as 0.2 and 0.1, respectively. Model standardized coefficients were extracted at $\lambda = 0.1$ (Figure 21). Again, like the results of the unsupervised sparse CCC analysis of radiomics and laboratory results, the original and wavelet features in the LLL frequency channel had the highest absolute values for the standardized coefficients. LDH, ESR, CRP, albumin, and total bilirubin were the selected serum biomarkers, and age, chronic disease, and CCI were the clinical variables of the CLRW model. Word2Vec encoding also contributed to the supervised task with its 4 out of 300 dimensions.

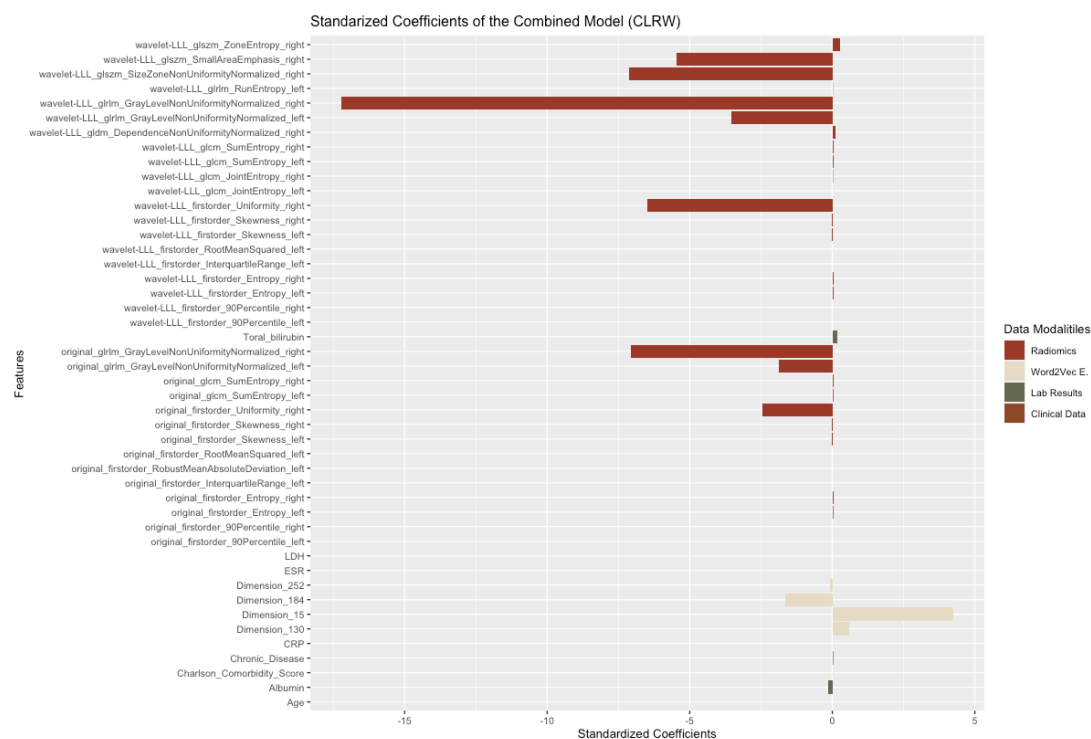


Figure 21: Standardized Coefficients of the Quadruple Model (CLRW).

Model standardized coefficients were extracted at $\lambda = 0.1$. The original and wavelet features in the LLL frequency channel have the highest absolute values for the standardized coefficients. Word2Vec embedding also contributes to the supervised task with its four dimensions. *Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study", NPJ Digital Medicine, under CC BY license [152].*

5.2.ILD Cohort

5.2.1. Descriptive Analysis

A total of 2,638 patients diagnosed with ILDs between January 1998 and November 2023 were retrospectively screened for inclusion. Of these, 1,622 patients with available FVC measurements within one year of their baseline chest CT scan were included for transplant-free survival analysis. Additionally, 556 patients who underwent transthoracic echocardiography (TTE) during follow-up were included in pulmonary artery radiomics analysis using Cox regression to predict the development of pulmonary hypertension (Figure 22). Baseline characteristics of deceased and surviving patients are summarized in Table 4. Significant differences ($p < 0.05$) were observed between the two groups in

age, ILD subtypes, FVC values, ILD-related treatments, and lung transplantation rates in univariate analysis.

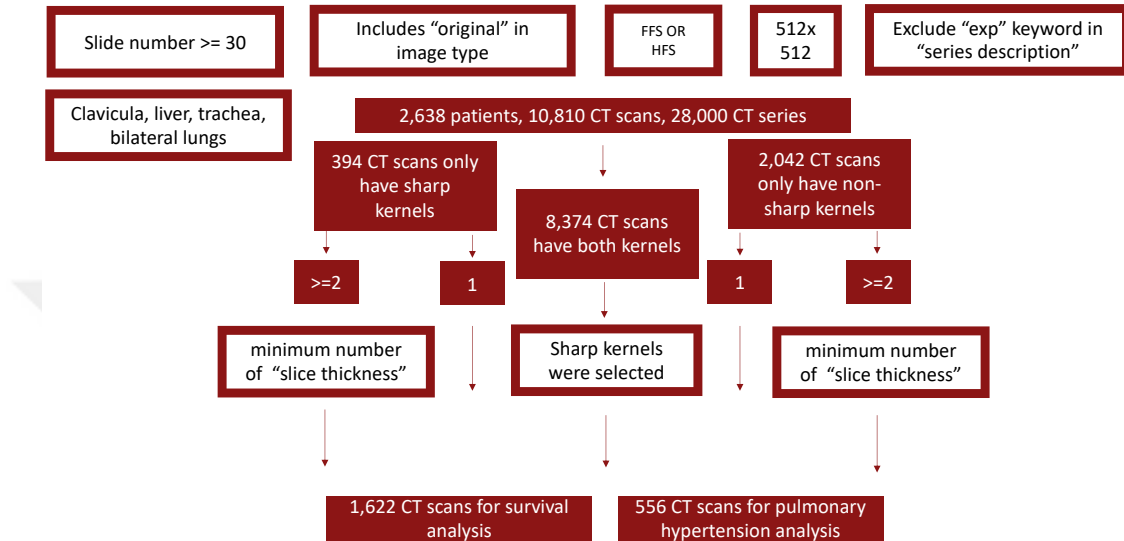


Figure 22: Flowchart of ILDs Cohort.

In total, 1,622 CT scans were selected for transplantation-free survival analysis and 556 CT scans were selected for pulmonary hypertension analysis.

The median slice thickness was 1.25 mm (interquartile range [IQR], 1–2.5 mm), and the median number of slices per scan was 228 (IQR, 101–296). A sharp reconstruction kernel was used in 1,327 cases (81.8%). Regarding contrast timing, 758 scans (46.7%) were performed without contrast (native), 429 (26.4%) during the early arterial phase, 309 (19.1%) during the late arterial phase, and 126 (7.8%) during the portal venous phase.

Table 4: Clinical Features of Patients with ILDs

	All patients N: 1,622	Deceased N: 211 (13%)	Survived N:1,411(87%)	P value
Age, mean±SD	64.2±13.9	66.5±12.6	63.9±14.1	0.006
Female Sex (%)	889 (54.8)	100 (47.4)	789 (55.9)	0.06
ILD subtypes, n (%)				<0.001
CTD_ILD	349 (21.5)	29 (13.7)	320 (22.7)	
IPF	293 (18.1)	52 (24.6)	241 (17.1)	
HP	196 (12.1)	27 (12.8)	169 (12.0)	
Sarcoidosis	86 (5.3)	3 (1.4)	83 (5.9)	
Unclassified	697 (43.0)	100 (47.4)	597 (42.3)	
FVC (L), mean±SD	2.54±0.95	2.26±0.85	2.58±0.96	<0.001
Follow up time, days, median	1,258 (563-2,093)	1,183 (543-2,024)	1,276 (570-2,124)	0.76
Transplantation, n(%)	97 (5.3)	38 (18.1)	59 (4.2)	<0.001
Mortality, n (%)	211(13)			
SD: standard deviation, CTD-ILD: Connective tissue related interstitial lung disease, IPF: idiopathic pulmonary fibrosis, HP: hypersensitivity pneumonitis, FVC: forced vital capacity				

5.2.2. Clinical Models in Transplant-Free Survival Analysis

We evaluated two Cox proportional hazards regression models to figure out the baseline performance across five outer folds: a baseline model that included age and sex, and an extended baseline model that included age, sex, and FVC (Table 5). For the age and sex model, the C-index values across the five folds ranged from 0.536 to 0.653. Time-dependent AUCs at year 1 ranged from 0.529 to 0.673, and AUCs across years 2 to 5 varied from 0.510 to 0.727. Log-rank p-values were below 0.01 in four folds. For the age, sex, and FVC model, the C-index ranged from 0.666 to 0.772 across the five folds. Time-dependent AUCs at year 1 ranged from 0.767 to 0.793, with AUCs at years 2 to 5 ranging from 0.682 to 0.845. All folds showed statistically significant survival differences between risk groups based on log-rank tests ($p < 0.01$ in all folds).

Table 5: Performance of Baseline Functional Cox Proportional Hazards Models

Fold	C-index	AUC (t=1)	AUC (t=2)	AUC (t=3)	AUC (t=4)	AUC (t=5)	Log- Rank p-value	Test Size	Events
Age and Sex Model									
1	0.6531	0.6568	0.6806	0.7152	0.6773	0.7012	< 0.001	327	51
2	0.6463	0.6376	0.5806	0.6345	0.7337	0.7273	< 0.001	323	47
3	0.5361	0.5329	0.5257	0.5500	0.5105	0.5574	0.23	324	37
4	0.5993	0.5291	0.5509	0.6457	0.6411	0.7167	< 0.001	324	39
5	0.6098	0.6725	0.6816	0.6407	0.5917	0.5919	0.01	324	34
Age, Sex and FVC Model									
1	0.7716	0.7834	0.8452	0.8129	0.8087	0.7913	< 0.001	327	51
2	0.7339	0.7810	0.7327	0.7556	0.7818	0.7313	< 0.001	323	47
3	0.7022	0.7927	0.7432	0.7447	0.6822	0.6922	0.002	324	37
4	0.6661	0.6342	0.6287	0.6755	0.6802	0.7417	< 0.001	324	39
5	0.7462	0.7673	0.7844	0.7654	0.7412	0.7429	< 0.001	324	34
C-index: Concordance index, AUC: Area under the curve									

In Table 6, for the age, sex, and FVC model, across the five outer folds, the median HR for age was 0.996 (IQR: 0.993–0.997). For male sex, the HR was elevated, with a median of 4.40 (IQR: 4.15–4.52). The HR for FVC was below 1 across all folds, with a median of 0.47 (IQR: 0.44–0.48), indicating a protective association between higher FVC and a lower risk of the outcome.

Table 6: Hazards Ratios of Baseline Functional Cox Proportional Hazards Models for Transplant-Free Survival Analysis

Variable	HR (F1)	HR (F2)	HR (F3)	HR (F4)	HR (F5)
Age	0.995 (0.984 - 1.007)	0.993 (0.981 - 1.005)	0.997 (0.986 - 1.008)	1.003 (0.991 - 1.015)	0.996 (0.985 - 1.007)
Gender	4.152 (2.920 - 5.902)	4.401 (3.092 - 6.263)	4.683 (3.351 - 6.545)	4.381 (3.117 - 6.158)	4.520 (3.223 - 6.332)
FVC	0.481 (0.399 - 0.580)	0.439 (0.363 - 0.531)	0.470 (0.393 - 0.561)	0.386 (0.315 - 0.474)	0.475 (0.396 - 0.570)
HR: Hazards Ratio					

5.2.3. Radiomics Models in Transplant-Free Survival Analysis

For lung and pulmonary artery radiomics-based transplant-free survival analysis, we trained models with and without clinical variables for age and sex. Lung and pulmonary artery radiomics outer loop results are provided in Table 7. The lung radiomics model achieved a median C-index of 0.78 (range: 0.73–0.81) and demonstrated high time-dependent AUCs across years 1–5, peaking at 0.89 in Fold 3. The pulmonary artery radiomics model demonstrated predictive performance with a median C-index of 0.71 (range: 0.69–0.73) and corresponding AUCs ranging from 0.61 to 0.78. The combined model, incorporating both lung and pulmonary artery radiomics, achieved the highest overall median C-index of 0.79.

Table 7: Performance of Lung and Pulmonary Artery Features of Transplant-Free Survival Models

Fold	Alpha	C-index (min)	C-index (1se)	AUC (t=1)	AUC (t=2)	AUC (t=3)	AUC (t=4)	AUC (t=5)	Log-Rank p-value	Test Size	Events
Lung Radiomics											
1	0.4	0.7878	0.7714	0.8100	0.8258	0.7731	0.8341	0.7801	<0.001	327	51
2	0.2	0.7307	0.6829	0.7762	0.7258	0.7082	0.6863	0.6549	0.005	323	47
3	0.2	0.7975	0.7893	0.8918	0.8881	0.8737	0.7439	0.7397	< 0.001	324	37
4	0.2	0.7280	0.7051	0.7118	0.7334	0.7040	0.7192	0.6856	< 0.001	324	39
5	0.6	0.8053	0.8169	0.7724	0.8322	0.8315	0.8585	0.8494	< 0.001	324	34
Pulmonary Artery Radiomics											
1	0.2	0.7120	0.6972	0.6866	0.7429	0.7301	0.6938	0.7102	<0.001	327	51
2	0.2	0.6934	0.6644	0.7243	0.6560	0.6690	0.6925	0.6322	0.005	323	47
3	0.2	0.7267	0.7054	0.7820	0.7995	0.8148	0.6605	0.6175	< 0.001	324	37
4	0.2	0.7035	0.6882	0.6059	0.7073	0.7200	0.7150	0.6719	0.002	324	39
5	0.2	0.7136	0.6986	0.7324	0.7856	0.7330	0.6765	0.6796	< 0.001	324	34
Lung – Pulmonary Artery Radiomics											
1	0.2	0.7915	0.7803	0.8129	0.8467	0.7878	0.8394	0.7967	<0.001	327	51
2	0.2	0.7329	0.7079	0.7823	0.7405	0.7245	0.7160	0.6784	<0.001	323	47
3	0.2	0.8124	0.7783	0.8733	0.8796	0.8718	0.7406	0.7103	<0.001	324	37
4	0.2	0.7527	0.7358	0.7345	0.7600	0.7376	0.7584	0.7289	<0.001	324	39
5	0.2	0.7986	0.8186	0.7684	0.8543	0.8332	0.8586	0.8570	<0.001	324	34
AUC: Area under the cure, C-index: Concordance index											

5.2.4. Comparison of Clinical and Radiomics Models in Transplant-Free Survival Analysis

To ensure a fair and valid comparison between clinical and radiomics models, each experiment was repeated 10 times using identical outer and inner CV splits across all models within each repetition. As illustrated in Figure 23, the mean C-index values were 60.9 (IQR: 60.4 – 62.0) for clinical variables (age and sex), 72.7 (IQR: 72.4 – 72.9) for age-sex combined with FVC, 70.8 (IQR: 70.0 – 71.0) for pulmonary artery radiomic

features, and 76.8 (IQR: 76.4 – 77.0) for lung radiomics alone. Combining age, sex, and lung radiomics yielded a higher C-index of 77.7 (IQR: 76.0 – 78.5), while the combination of age, sex, and pulmonary artery features resulted in 70.8 (IQR: 70.0 – 71.0). The highest predictive performance was achieved when all four feature groups of age, sex, lung, and pulmonary artery radiomics features were integrated, reaching a C-index of 78.5 (IQR: 77.9 – 78.7).

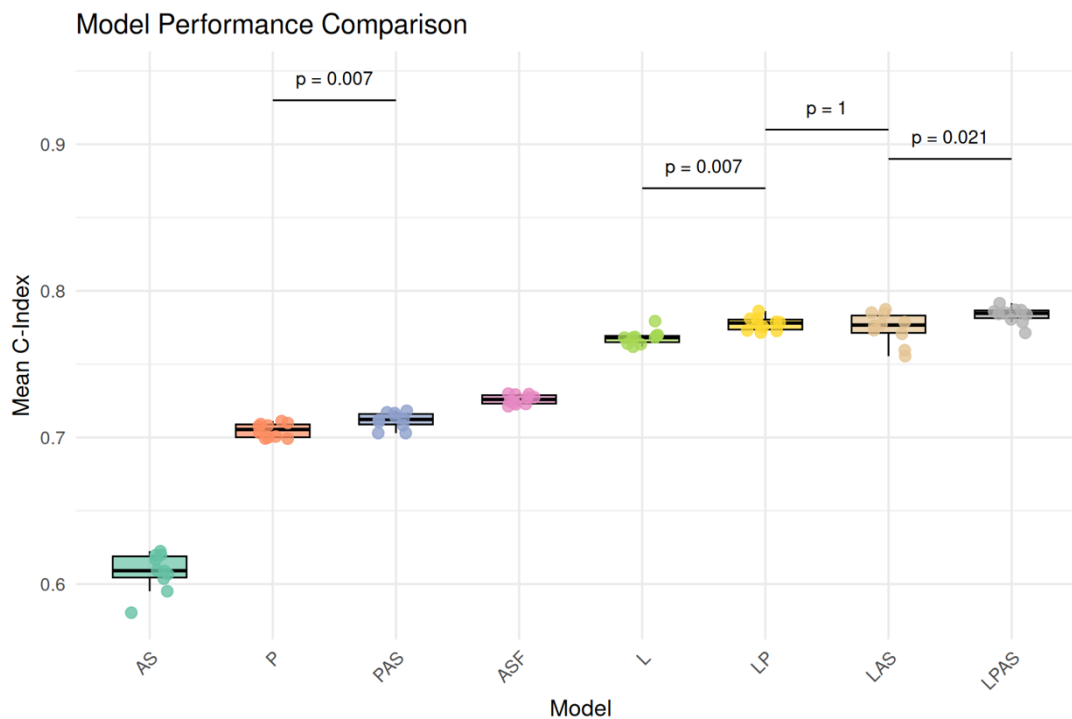


Figure 23: Comparison of Clinical and Radiomics Models in Transplant-Free Survival Analysis.

A: age; S: sex; F: FVC, L: lung radiomics; P: pulmonary artery radiomics

5.2.5. Interpretation of Radiomics Features in Transplant-Free Survival Analysis

Age and sex were included as covariates in the multivariate radiomics Cox analysis to account for potential demographic confounding when interpreting lung and pulmonary artery radiomics features. Consistently selected features across the outer loops of the survival models are summarized in Figures 24, 25, and 26. Features selected in at least 3 out of 5 outer loops are presented in tabular form in Appendix 5. Notably, in 4 out of 5 outer loops, all pulmonary artery features were selected (with $\alpha = 0$, corresponding to ridge regression), and therefore, features selected in all outer loops are presented. Texture

features reflecting greater heterogeneity, such as GLDM Small Dependence Low Gray Level Emphasis, GLCM Difference Average, GLDM Small Dependence Emphasis, and GLSZM Zone Percentage, were associated with higher hazard ratios. Conversely, features linked to more homogeneous textures, including GLCM Inverse Difference Moment Normalized (IDMN) and GLCM Inverse Difference Normalized (IDN), demonstrated lower hazard ratios. In the analysis of pulmonary artery features, in addition to texture-based features, shape-related features such as Major Axis Length, Elongation, and Flatness were consistently selected by the prognostic models. Among the texture features, GLCM ICM2 was notably associated with lower hazard rates. In the combined model using both lung and pulmonary artery features, lung-related features were predominant; however, texture-based pulmonary artery features also contributed to the overall model performance.

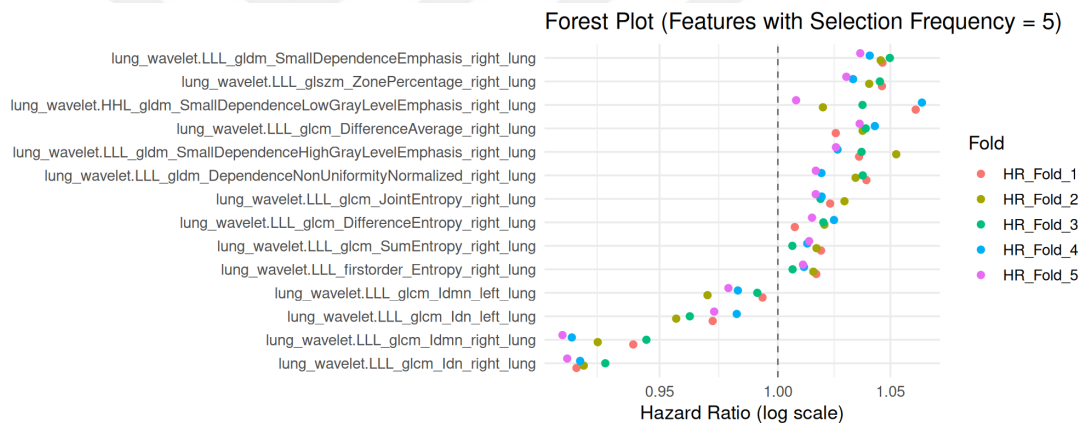


Figure 24: Forest Plot of Selected Lung Radiomics Features Associated with Transplant-Free Survival

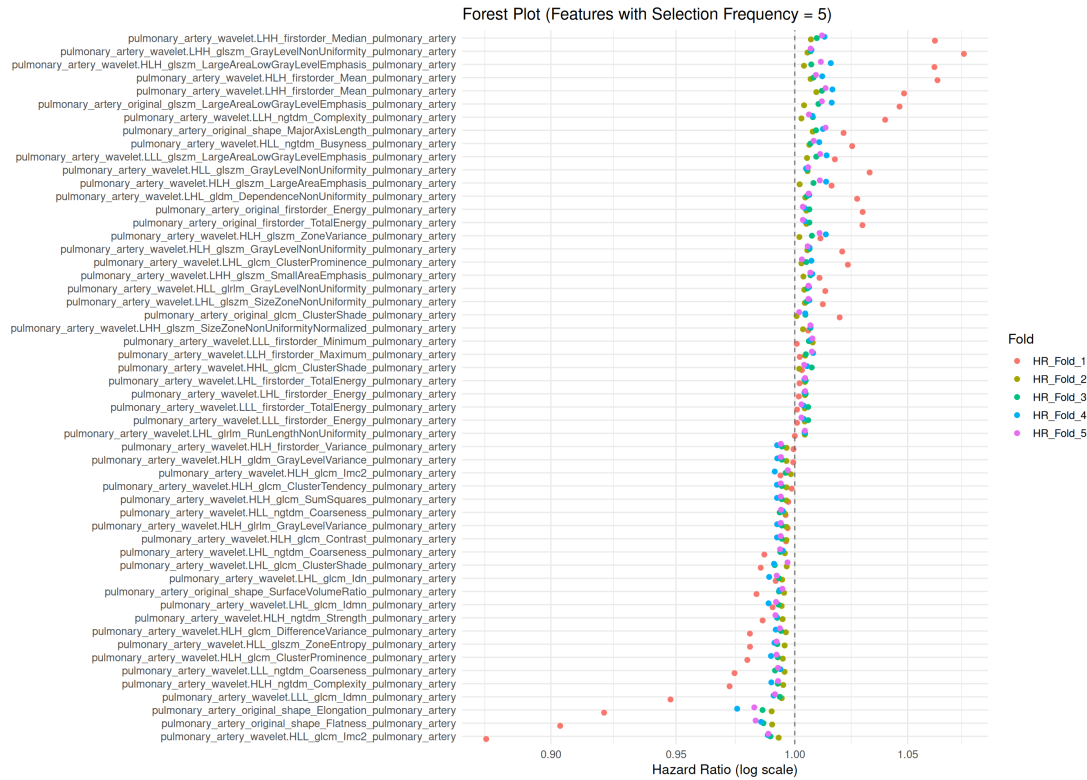


Figure 25: Forest Plot of Selected Pulmonary Artery Radiomics Features Associated with Transplant-Free Survival

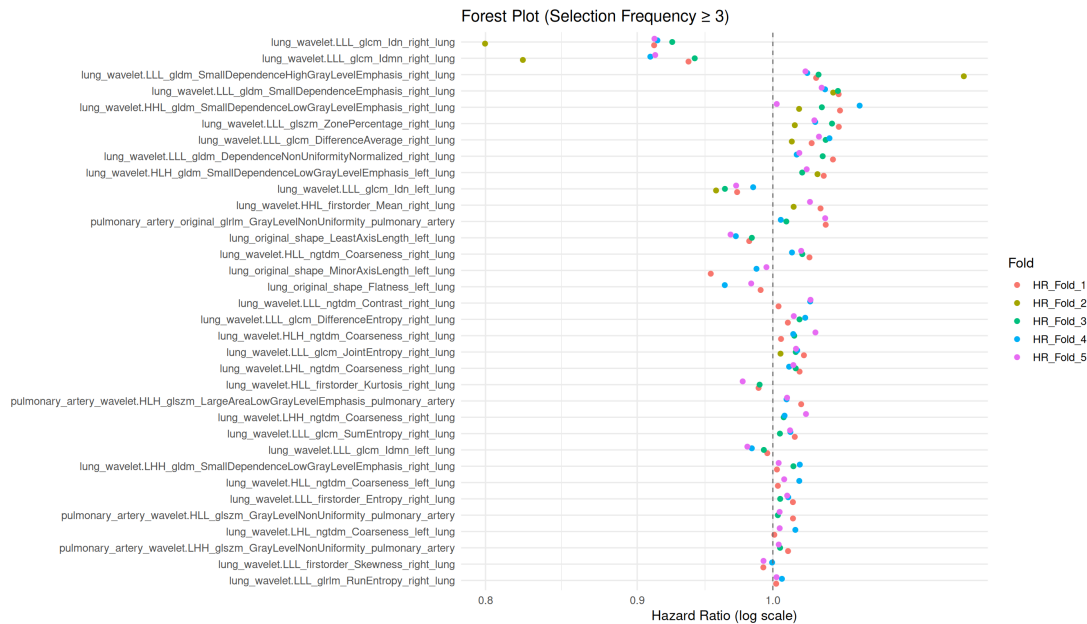


Figure 26: Forest Plot of Selected Lung and Pulmonary Artery Radiomics Features from Combined Models Associated with Transplant-Free Survival

5.2.6. Radiomics Models in Pulmonary Hypertension Analysis

Among 556 patients with available sPAP measurements, 211 (38%) had at least one sPAP value exceeding 40 mmHg. The date on which an sPAP > 40 mmHg was first observed was defined as the event date. The mortality rate among these patients was 23.7% (50/211) compared to 5.5% (19/345) in those without ($p < 0.001$).

Across the five outer folds, the lung radiomics models achieved C-index values ranging from 0.69 to 0.76, with significant log-rank p-values observed in four of the five folds (Table 8).

Table 8: Performance of Lung and Pulmonary Artery Features of Pulmonary Hypertension Models

Fold	Alpha	C-index (min)	C-index (1se)	Log-Rank p-value	Test Size	Event
Lung Radiomics						
1	0	0.7388	0.6785	0.031	112	22
2	0.4	0.7402	0.7231	0.012	113	21
3	0.2	0.7348	0.7531	0.008	108	18
4	0.8	0.6884	0.6658	0.169	110	21
5	0	0.7599	0.7379	0.002	113	17
Pulmonary Artery Radiomics						
1	0.2	0.6923	0.6763	0.021	112	22
2	0.8	0.6752	0.6980	0.093	113	21
3	0.2	0.7619	0.7575	<0.001	108	18
4	0.2	0.6551	0.6526	0.05	110	21
5	1	0.7108	0.7372	0.013	113	17
Lung and Pulmonary Artery Radiomics						
1	0.2	0.7714	0.6798	0.018	112	22
2	0	0.7345	0.7150	0.080	113	21
3	0	0.7766	0.7641	0.001	108	18
4	1	0.7293	0.7167	<0.001	110	21
5	0.4	0.7613	0.7540	0.023	113	17
C-index: Concordance index						

Pulmonary artery radiomics models exhibited greater variability, with C-index values ranging from 0.66 to 0.76. Statistically significant log-rank p-values were obtained in three folds. The combined model, incorporating both lung and pulmonary artery radiomics features, outperformed the individual models, achieving the highest C-index values (up to 0.78) and significant log-rank p-values in four of the five folds.

As was done in the transplant-free survival analysis, each experiment was repeated 10 times, and within each repetition, identical inner and outer split assignments were used across models. As illustrated in Figure 27, the median C-index values were 69.5 (IQR: 67.6 – 69.9) for pulmonary artery radiomic features and 71.5 (IQR: 71.1 – 73.0) for lung radiomics alone. The median index for the combination of lung and pulmonary artery features was 75.1 (IQR: 74.8 – 75.5).

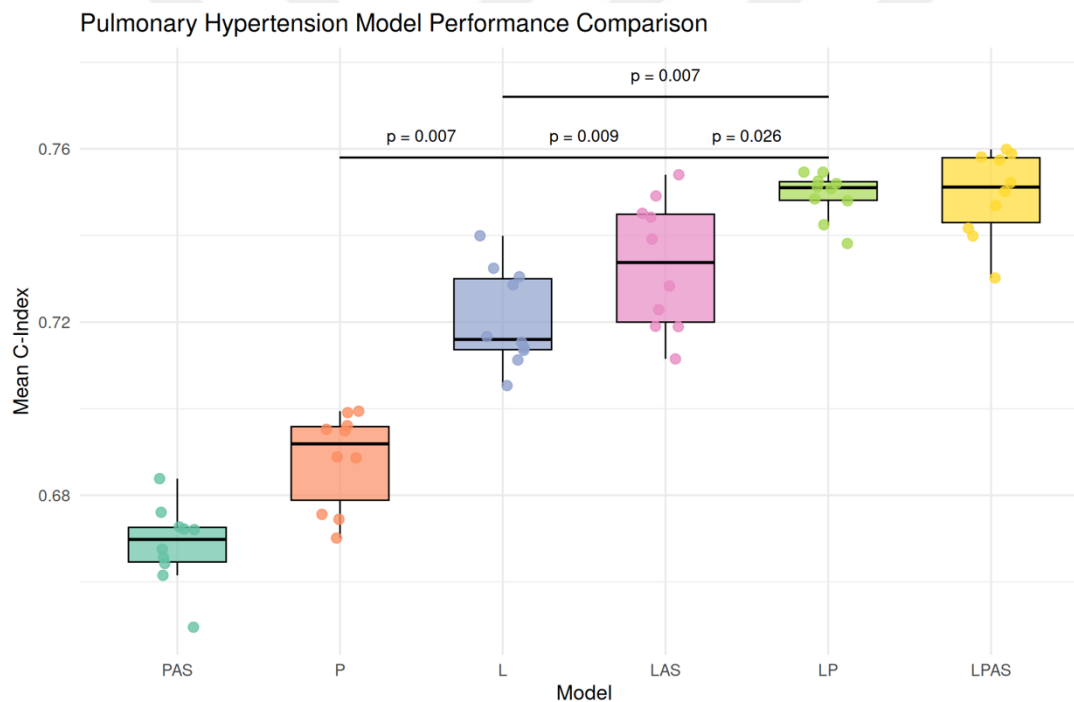


Figure 27: Comparison of Clinical and Radiomics Models in Pulmonary Hypertension

A: age; S: sex; L: lung radiomics; P: pulmonary artery radiomics

5.2.7. Interpretation of Radiomics Features in Pulmonary Hypertension Analysis

As was done in the transplant-free survival analysis, age and sex were included as covariates in the multivariate radiomics Cox regression analysis to adjust for potential demographic confounding when evaluating the diagnostic value of lung and pulmonary artery radiomics features. Consistently selected features from outer loops in pulmonary hypertension models were summarized in Figures 28, 29, 30. Features selected in at least 3 out of 5 outer loops are presented in tabular form in Appendix 5. While first-order

features, such as Skewness and Mean, tended to be dominant across lung features, pulmonary artery features continued to be prominently selected, such as texture and shape-related features, reinforcing the patterns observed in the survival analysis. Interestingly, as different from transplant-free survival analysis, volume-related features, such as mesh volume and voxel volume, were selected as pulmonary artery features. The combined models, both pulmonary artery and lung-related features, contribute to the prediction models.

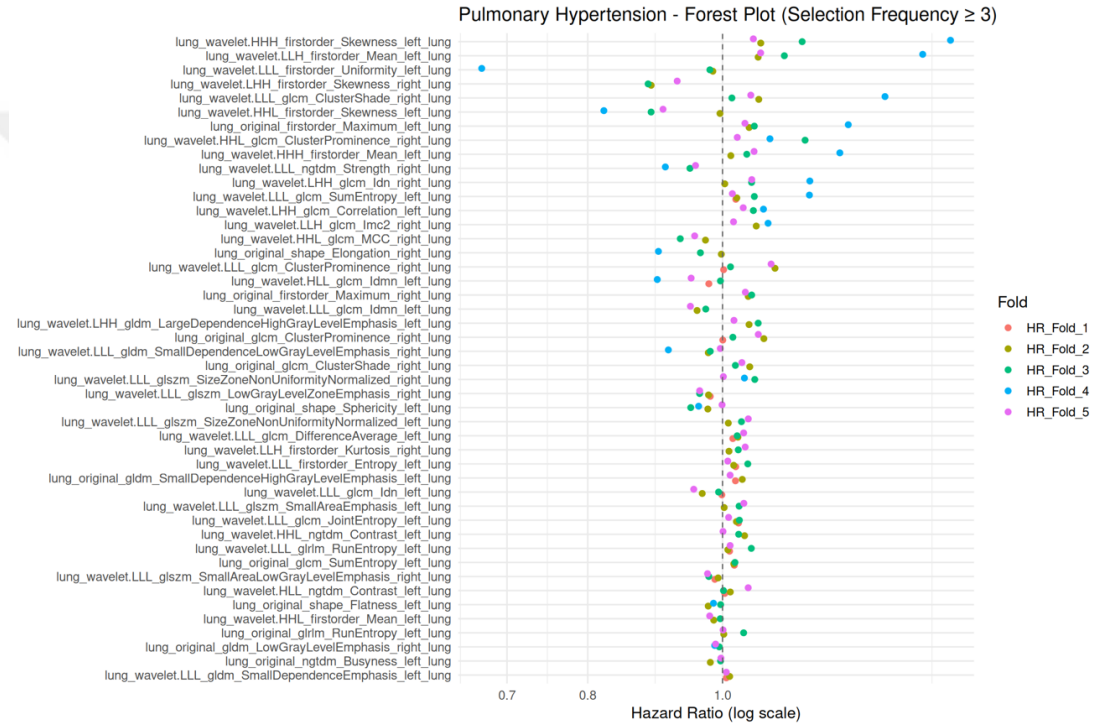


Figure 28: Forest Plot of Selected Features from Pulmonary Hypertension Models Using Lung Radiomics Features

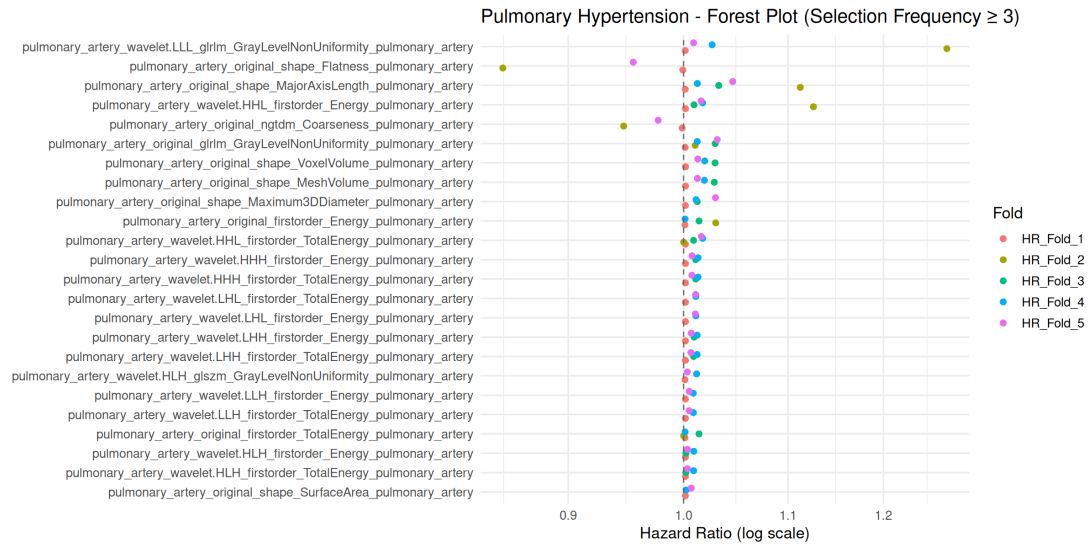


Figure 29: Forest Plot of Selected Features from Pulmonary Hypertension Models Using Pulmonary Artery Radiomics Features

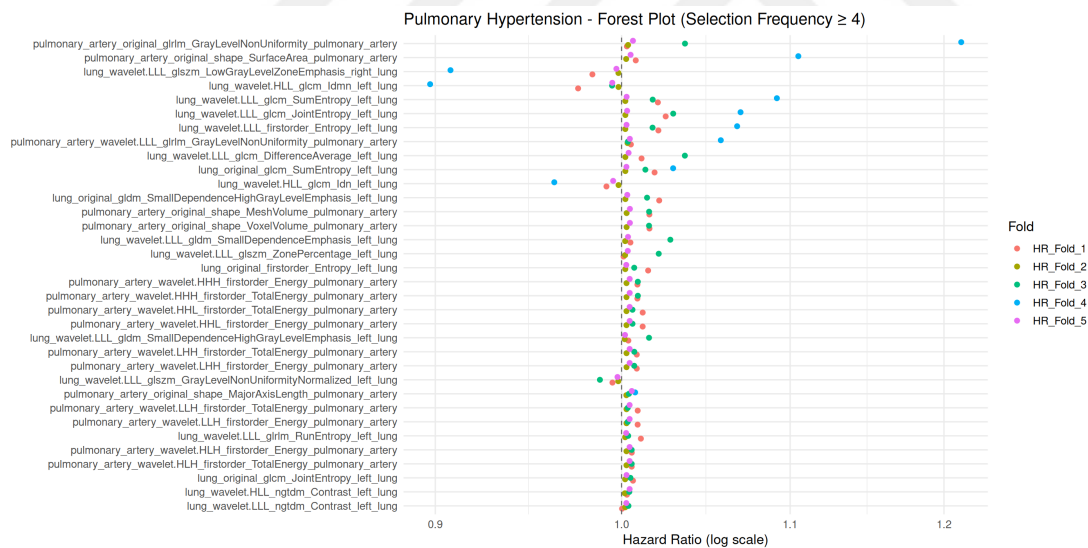


Figure 30: Forest Plot of Selected Features from Pulmonary Hypertension Models Using Lung and Pulmonary Artery Radiomics Features

CHAPTER 6

DISCUSSION

In this thesis, we investigated multimodal data fusion and multicompartiment image analysis using two distinct patient cohorts. In the first part, we integrated imaging, clinical, laboratory, and viral genome sequencing data from a COVID-19 patient cohort. Using unsupervised sparse CCA, we identified and quantified relationships across these diverse data modalities. For instance, between imaging data and laboratory results, histogram- and entropy-based first-order radiomic features correlated with serum biomarkers indicative of the acute phase response. Additionally, sparse CCA enabled the identification of distinct clinical phenotypes. We further applied cooperative learning to optimize multimodal data fusion, demonstrating its robustness in predicting ICU admission. For incorporating viral genomic information, we leveraged the Word2Vec NLP model, which effectively encoded the SARS-CoV-2 genome, distinguished major viral variants, preserved phylogenetic structure, and enabled the integration of viral sequence data into downstream clinical analyses. In the second part, we conducted radiomics analysis in an ILD cohort. In addition to traditional lung-based radiomic analysis, we extracted radiomics features from the pulmonary arteries and demonstrated that these features were interpretable and carried significant predictive value. Furthermore, integrating pulmonary artery and lung radiomics features improved model performance in predicting PH and transplant-free survival.

The complexities of modern biomedical challenges demand more holistic approaches to data analysis. This study spotlights the potential of multimodal data fusion, harnessing diverse data modalities to derive deeper insights into health phenomena. By integrating disparate data types, we can uncover nuanced patterns and relationships that unimodal analyses might miss. In particular, the potential of integrating viral genome sequencing, imaging, clinical data, and laboratory results is immense. Sparse CCA and cooperative learning, as highlighted in our study, are instrumental in combining these data strands, offering a multi-faceted view of a complex disease such as COVID-19. Our exploration into using the Word2Vec NLP model for viral embedding further underscores the value of innovative techniques in transforming raw data into meaningful representations.

As is well recognized, viral mutations typically evolving in clusters, giving rise to distinct viral lineages. In this study, rather than analyzing individual mutations in isolation, we aimed to capture their collective effect using Word2Vec encoding. To achieve this, we leveraged the global knowledge related to viral evolution and infused it into our local clinical dataset, enabling a more informative and biologically meaningful latent representation of viral genomes. We illustrated that Word2Vec encoding not only separates major SARS-CoV-2 variants but also allows the preservation of phylogenetic relationships among them. We also found that Word2Vec encoding was particularly beneficial in visualizing the relationships between recombinant strains and major viral lineages, an insight that is often difficult to capture using traditional phylogenetic trees.

Next, imaging data, particularly CT scans, holds potentially more information than can be observed by radiologists. Our analysis shows that serum biomarkers that represent positive and negative acute phase responses are found to be correlated with COVID-19 radiomics features related to the distribution of voxel intensities. As known, in lung nodule and cancer studies, mainly shape-related features are robust and provide information about disease phenotypes and prognosis; however, our results emphasize the importance of histogram- and entropy-related features because a higher proportion of involved lung parenchyma, i.e., diffuse pulmonary infiltrates, is associated with severe disease [163-165]. Furthermore, we reveal that biomarkers not playing a role in acute phase response, such as alanine transaminase, potassium, and creatinine, are not correlated with radiomics features.

Sparse CCA has been championed in various biomedical domains, from understanding eating disorders via imaging data to categorizing clinical subtypes in dementia [166, 167]. Our application to COVID-19 aligns with this trajectory, delineating clinical phenotypes like kidney disease, liver dysfunction, and age-related vulnerabilities, supported by existing literature and our clinical observations [168-170]. However, challenges persist. Multimodal data fusion demands rigorous computational techniques, nuanced strategies for feature extraction, and the transformation of raw data into a structured format. Our experiments show that when these elements align, as seen with the optimal results using Word2Vec encoding, the outcomes are promising. This performance edge over viral-binary encoding likely arises because Word2Vec benefits from an expansive external database, capturing nuances not specific to our dataset.

Several studies have focused on predicting supervised clinical outcomes in COVID-19 patients by integrating imaging and clinical data [171-176]. Our study extends beyond these efforts by incorporating viral genome sequencing data into the analysis, enabling a more comprehensive multimodal approach. It is foreseeable that viral genome sequencing data will be directly integrated into our clinical patient management pipelines in the coming years [177]. Yet, decisions surrounding data fusion techniques and staging remain paramount. We employed cooperative learning, enhancing alignment across modalities, to determine the degree of optimal fusion, guided by the agreement penalty tuning to distinguish which patients should be admitted to the medical ward or ICU when the relevant patient data was collected. At the end of the analysis, a review of the features

selected in both our unsupervised and supervised models revealed a consistency, highlighting a harmony between an unsupervised use of sparse CCA and supervised predictive cooperative learning.

A few caveats need to be noted. First, our relatively small sample size of our COVID-19 cohort might be suffering from overfitting. To overcome this issue, we used a stratified nested CV framework for estimating the generalization performances of the trained models. The sparsity penalty in both sparse CCA and cooperative learning also helped mitigate the risk of overfitting. Next, from a clinical point of view, this study is dated in the early days of the pandemic, and a small number of patients were vaccinated for COVID-19. Because it is not possible to quantify the vaccination effect properly, we had to ignore this effect.

Our analysis of ILD patients revealed statistically significant differences between survivors and deceased individuals across key demographic and clinical variables, including age, sex, ILD subtype, FVC, treatment status, and lung transplantation rates. Interestingly, in the baseline functional Cox proportional hazards models assessing transplant-free survival, age was not found to be a statistically significant predictor. This seemingly paradoxical result may be explained by the underlying age distribution and natural history of fibrotic ILDs. For example, IPF, a rapidly progressive disease that primarily affects males in their 50s, was more prevalent among deceased patients [178]. In contrast, CTD-ILD, which was more common among survivors, often presents earlier in life and typically follows a more indolent trajectory, particularly in subtypes such as systemic sclerosis with limited fibrotic involvement [68]. Furthermore, it is important to note that transplant-free survival as an endpoint complicates direct age comparisons, as younger patients are more likely to be eligible for lung transplantation, thereby altering survival patterns in age-stratified analyses.

The significantly higher rate of lung transplantation among deceased patients is consistent with the clinical practice of prioritizing transplantation for patients with advanced disease and poor prognosis. While transplantation is a potentially life-saving intervention, its presence in the mortality group likely reflects the severity of disease leading up to the procedure, rather than its failure. Taken together, these findings emphasize the heterogeneity of ILD outcomes and the prognostic importance of demographic and disease-specific factors.

A major strength of our study lies in the high-quality preprocessing and segmentation of CT images, which we believe provided a robust foundation for subsequent radiomic analyses. We implemented a comprehensive, standardized pipeline to ensure the selection of diagnostically relevant HRCT scans while systematically excluding low-quality or incomplete series. Anatomical completeness was verified with anatomical labeling of clavicae, trachea, liver, and the lungs using TotalSegmentatorv2, allowing us to include only scans covering the full extent of the lungs and key thoracic structures. Careful categorization by reconstruction kernel types and selection of series with minimal slice thickness further enhanced image consistency. Segmentation performance was rigorously

compared across multiple state-of-the-art tools, with TotalSegmentatorv2 ultimately chosen based on visual inspection of binary masks. Additionally, we applied tailored windowing protocols for lung parenchyma and vascular structures and ensured spatial normalization and intensity scaling prior to feature extraction. This meticulous preprocessing protocol, including contrast phase labeling and voxel-wise standardization, minimized technical variability and increased the reliability of the radiomic features extracted from both lung parenchyma and pulmonary arteries.

Our imaging dataset exhibits high overall quality, with a median slice thickness of 1.25 mm and a substantial number of slices per scan, providing sufficient spatial resolution for reliable radiomic feature extraction. Thin-slice CT images are particularly important in ILD radiomics, as they enhance visualization of fine parenchymal structures and minimize partial volume effects. Notably, more than 80% of the scans were reconstructed using a sharp kernel, which improves edge definition and texture contrast, attributes that can enhance the sensitivity of radiomic analysis. The dataset also reflects real-world clinical heterogeneity in contrast enhancement, with approximately half of the scans acquired without contrast (native phase) and the remainder distributed across early arterial, late arterial, and portal venous phases. This variability in contrast timing can influence both vascular and parenchymal attenuation, particularly affecting pulmonary artery radiomics features. To address this, we incorporated contrast phase stratification into our CV framework and preprocessing pipeline to mitigate potential confounding effects and improve the robustness of feature interpretation and model performance.

Our baseline Cox regression hazard models confirmed the clinical relevance of traditional predictors such as age, sex, and FVC in estimating transplant-free survival in ILD patients. The model incorporating only age and sex demonstrated modest discriminative ability, with C-index values ranging from 0.536 to 0.653 and time-dependent AUCs reflecting limited performance, particularly beyond the first year of follow-up. These findings suggest that demographic variables alone offer insufficient granularity for robust risk stratification in this heterogeneous population. The addition of FVC substantially improved model performance across all evaluation metrics. This improvement reflects the central role of lung function tests as a well-established predictor of prognosis in ILDs.

Our findings demonstrate that radiomic features extracted from both lung parenchyma and pulmonary arteries hold strong prognostic value for predicting transplant-free survival in ILD patients, with and without traditional demographic variables. The lung radiomics model alone achieved a median C-index of 0.78, outperforming the baseline clinical models and showing consistently high time-dependent AUCs throughout the five-year follow-up period. This result highlights the potential of lung radiomics features to capture disease severity and progression risk beyond what is measurable by clinical metrics such as FVC. In addition, the pulmonary artery radiomics model also demonstrated meaningful predictive ability, with a median C-index of 0.71. This is a noteworthy finding, as it emphasizes the prognostic relevance of vascular morphology and attenuation characteristics in ILDs, particularly given the well-established role of pulmonary hypertension in worsening outcomes.

The model integrating both lung and pulmonary artery radiomic features achieved the highest median C-index of 0.79. In repeated experiments, the combined model demonstrated statistically significant improvement over the lung-only radiomics model, underscoring the value of a multicompartiment image analysis approach. Overall, the results provide strong evidence for the clinical utility of radiomics-based models in risk stratification and highlight the added value of including vascular imaging features in predictive modeling for ILDs.

In our multivariate transplant-free survival analysis, age and sex were included as covariates to control for potential demographic confounding, allowing clearer interpretation of lung and pulmonary artery radiomic features. Features consistently selected across outer loops in lung radiomics analysis are indicative of increased heterogeneity, such as GLDM Small Dependence Low Gray Level Emphasis, GLCM Difference Average, GLDM Small Dependence Emphasis, and GLSZM Zone Percentage, were associated with elevated hazard ratios. This suggests that greater tissue heterogeneity corresponds to worse prognosis. In contrast, features reflecting more homogeneous textures, including GLCM IDMN and IDN, correlated with lower hazard ratios, indicating a more favorable outcome.

Regarding pulmonary artery radiomics, prognostic models consistently selected both texture and shape features. GLSZM Gray Level Non-Uniformity was found to be informative as a texture feature, which quantifies the variability of gray-level intensity values. A high value for this feature could theoretically indicate vascular remodeling, heterogenous contrast enhancement, and intraluminal flow artifacts. Shape descriptors like Major Axis Length, Elongation, and Flatness were notably prognostic, highlighting the importance of vascular morphology. Major Axis Length was found to be associated with positive hazards, whereas Elongation and Flatness were associated with negative hazards. According to the pyradiomics documentation, Elongation quantifies the relationship between the two largest principal components of the ROI shape and is defined as the inverse of true elongation for computational purposes [20]. Thus, higher values (approaching 1) indicate a more circular shape, while lower values (approaching 0) reflect greater elongation. Similarly, Flatness follows the same computational convention, where values near 1 represent a more spherical object, and values closer to 0 indicate a flatter, more compressed structure. Taken together, these findings suggest that pulmonary arteries with long, elongated, and flat geometries, in other words, those that deviate from a spherical shape, are associated with worse prognosis. This morphological alteration may reflect underlying vascular remodeling processes. Notably, previous studies have shown that aging is positively correlated with increased length of the proximal pulmonary artery, and pulmonary artery dilatation is commonly observed in patients with pulmonary hypertension [179, 180]. These observations support the idea that pulmonary artery radiomic features capture clinically meaningful structural changes related to vascular pathology in patients with ILDs.

It is important to note that the lung radiomics-based model outperformed the pulmonary artery-based model in the analysis of PH. This finding is not unexpected, given the well-

established association between pulmonary HT and mortality, as seen in our dataset as well. Lung radiomic features likely capture a broad range of pathological patterns related to disease severity and mortality, and in doing so, may also indirectly reflect signals associated with pulmonary HT. It should also be noted that anomalies, such as mosaic attenuation, can be observed in lung parenchyma in patients with PH; thus, models trained with lung radiomics may leverage these findings as well [181]. In the analysis of pulmonary hypertension, distinct from the transplant-free survival analysis, the selected pulmonary artery features included not only shape-related metrics such as Major Axis Length and Flatness, but also volume-related features like Mesh Volume, Voxel Volume, and Maximum 3D Diameter. These features are considered highly interpretable, as radiologists commonly measure pulmonary artery diameter on CT scans to assess the likelihood of PH. However, as demonstrated in the transplant-free survival analysis, the combined model, integrating both lung and pulmonary artery radiomics features, outperformed models based on either compartment alone. This highlights the complementary nature of multicompartment radiomics analysis and underscores the added value of incorporating arterial features alongside parenchymal imaging biomarkers for more accurate prognostication in ILDs. This becomes evident when examining the selected features of the combined model, where predictive contributions arise from both lung and pulmonary artery compartments.

Our ILD study has several limitations that should be acknowledged. First, its retrospective design may limit the ability to establish causal relationships. Although we applied a standardized and rigorous image preprocessing pipeline to enhance consistency, inherent variability in CT acquisition parameters, such as differences in scanner types, slice thickness, and reconstruction kernels, may still have influenced the extracted radiomic features and introduced noise into the analysis. Additionally, the diagnosis of PH in this study was based on TTE findings, which, while widely used in clinical practice, lack the diagnostic accuracy of RHC, the gold standard for PH assessment. Furthermore, we used only FVC as the functional parameter in our models to be able to include the maximum number of patients possible in our study. However, the absence of other physiological measures, such as FVC predicted and DLCO, may have limited our ability to capture the full extent of functional status. Lastly, the patient cohort was assembled over a prolonged period during which treatment strategies, diagnostic criteria, and supportive care practices may have evolved, potentially impacting clinical outcomes and radiologic patterns. Future prospective multicenter studies, and particularly performing experiments in external cohorts, incorporating standardized imaging protocols and confirmatory hemodynamic assessments, are needed to validate our findings to better elucidate the prognostic role of pulmonary artery radiomics in ILDs, especially related to PH diagnosis and treatment response.

In conclusion, by utilizing two distinct patient cohorts, this thesis has demonstrated how multimodal data fusion and multicompartment image analysis can emulate clinical decision-making and enhance personalized prognostication in complex lung diseases. Our findings highlight the importance of integrating diverse data sources, as exemplified by our work in COVID-19, and anatomically distinct yet functionally connected

compartments, such as the lung parenchyma and pulmonary arteries, as illustrated in the case of ILDs.

Looking forward, the methodologies applied here, such as Word2Vec latent representation and multimodal data fusion, hold significant promise beyond COVID-19. These approaches can be extended to other critical areas, including HIV, influenza, and antimicrobial resistance, where they could be used to encode microbial genomic data for integrating genomics with other data modalities. They can be leveraged in microbiome analysis, as well. The comparison of Word2Vec's encoding performance with other language models trained on nucleic acid and protein sequences, both semantically and across various supervised tasks, can also yield valuable insights [182].

Moreover, beyond the use of hand-crafted radiomics features, both the concept of multicompartment image analysis and multimodal data fusion are inherently compatible with transformer-based architectures, whose attention mechanisms enable the modeling of contextual relationships across anatomical regions and data modalities. In the context of ILDs, such integration holds great potential to uncover novel clinical phenotypes that may not be apparent through unimodal analysis.

Future research should also further explore the role of pulmonary artery radiomics, particularly in diseases where parenchymal findings are subtle or absent, such as pulmonary embolism. Moreover, there is a potential for differentiating different PH subgroups, which have different treatment opportunities. In these contexts, pulmonary vascular imaging biomarkers may offer critical diagnostic and prognostic value. The ability to predict PH using a fast, widely available modality like CT is particularly valuable. Accurate identification of PH in ILD patients has become increasingly important, especially as therapeutic options are being explored for Group 3 PH. While RHC remains the gold standard for diagnosis, it is invasive and not always feasible. Similarly, cardiovascular MRI, though informative, is costly, time-consuming, and often unsuitable for patients with advanced lung disease. Thus, using of CT-based radiomics offers a promising and practical alternative for non-invasive patient management.



REFERENCES

1. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2022;28(9):1773-84. <https://doi.org/10.1038/s41591-022-01981-2>
2. Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*. 2019;35(14):i446-i54. <https://doi.org/10.1093/bioinformatics/btz342>
3. Hartmann K, Sadee CY, Satwah I, Carrillo-Perez F, Gevaert O. Imaging genomics: data fusion in uncovering disease heritability. *Trends Mol Med*. 2023;29(2):141-51. <https://doi.org/10.1016/j.molmed.2022.11.002>
4. Steyaert S, Pizurica M, Nagaraj D, Khandelwal P, Hernandez-Boussard T, Gentles AJ, Gevaert O. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat Mach Intell*. 2023;5(4):351-62. <https://doi.org/10.1038/s42256-023-00633-5>
5. Steyaert S, Qiu YL, Zheng Y, Mukherjee P, Vogel H, Gevaert O. Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. *Commun Med (Lond)*. 2023;3(1):44. <https://doi.org/10.1038/s43856-023-00276-y>
6. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, Wang F, Cheng F, Luo Y. Multimodal machine learning in precision health: A scoping review. *NPJ Digit Med*. 2022;5(1):171. <https://doi.org/10.1038/s41746-022-00712-8>
7. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci U S A*. 2019;116(44):22071-80. <https://doi.org/10.1073/pnas.1900654116>
8. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinform*. 2022;23(2). <https://doi.org/10.1093/bib/bbab569>

9. Venkatesan P. GOLD COPD report: 2025 update. *Lancet Respir Med.* 2025;13(1):e7-e8. [https://doi.org/10.1016/S2213-2600\(24\)00413-2](https://doi.org/10.1016/S2213-2600(24)00413-2)

10. Travis WD, Costabel U, Hansell DM, King TE, Jr., Lynch DA, Nicholson AG, Ryerson CJ, Ryu JH, Selman M, Wells AU, Behr J, Bouros D, Brown KK, Colby TV, Collard HR, Cordeiro CR, Cottin V, Crestani B, Drent M, Dudden RF, Egan J, Flaherty K, Hogaboam C, Inoue Y, Johkoh T, Kim DS, Kitaichi M, Loyd J, Martinez FJ, Myers J, Protzko S, Raghu G, Richeldi L, Sverzellati N, Swigris J, Valeyre D, Pneumonias AEColl. An official American Thoracic Society/European Respiratory Society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. *Am J Respir Crit Care Med.* 2013;188(6):733-48. <https://doi.org/10.1164/rccm.201308-1483ST>

11. Anastassopoulou C, Gkizarioti Z, Patrinos GP, Tsakris A. Human genetic factors associated with susceptibility to SARS-CoV-2 infection and COVID-19 disease severity. *Hum Genomics.* 2020;14(1):40. <https://doi.org/10.1186/s40246-020-00290-4>

12. ER A, ÜNAL S. 2019 Coronavirus Pandemic in Turkey and Across the World. *FLORA INFEKSIYON HASTALIKLARI VE KLINIK MIKROBIOLOJİ DERGİSİ.* 2020;25(1).

13. Severe Covid GG, Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, Fernandez J, Prati D, Baselli G, Asselta R, Grimsrud MM, Milani C, Aziz F, Kassens J, May S, Wendorff M, Wienbrandt L, Uellendahl-Werth F, Zheng T, Yi X, de Pablo R, Chercoles AG, Palom A, Garcia-Fernandez AE, Rodriguez-Frias F, Zanella A, Bandera A, Protti A, Aghemo A, Lleo A, Biondi A, Caballero-Garralda A, Gori A, Tanck A, Carreras Nolla A, Latiano A, Fracanzani AL, Peschuck A, Julia A, Pesenti A, Voza A, Jimenez D, Mateos B, Nafria Jimenez B, Quereda C, Paccapelo C, Gassner C, Angelini C, Cea C, Solier A, Pestana D, Muniz-Diaz E, Sandoval E, Paraboschi EM, Navas E, Garcia Sanchez F, Ceriotti F, Martinelli-Boneschi F, Peyvandi F, Blasi F, Tellez L, Blanco-Grau A, Hemmrich-Stanisak G, Grasselli G, Costantino G, Cardamone G, Foti G, Aneli S, Kurihara H, ElAbd H, My I, Galvan-Femenia I, Martin J, Erdmann J, Ferrusquia-Acosta J, Garcia-Etxebarria K, Izquierdo-Sanchez L, Bettini LR, Sumoy L, Terranova L, Moreira L, Santoro L, Scudeller L, Mesonero F, Roade L, Ruhlemann MC, Schaefer M, Carrabba M, Riveiro-Barciela M, Figuera Basso ME, Valsecchi MG, Hernandez-Tejero M, Acosta-Herrera M, D'Angio M, Baldini M, Cazzaniga M, Schulzky M, Cecconi M, Wittig M, Ciccarelli M, Rodriguez-Gandia M, Bocciolone M, Miozzo M, Montano N, Braun N, Sacchi N, Martinez N, Ozer O, Palmieri O, Faverio P, Preatoni P, Bonfanti P, Omodei P, Tentorio P, Castro P, Rodrigues PM, Blandino Ortiz A, de Cid R, Ferrer R, Gualtierotti R, Nieto R, Goerg S, Badalamenti S, Marsal S, Matullo G, Pelusi S, Juzenas S, Aliberti S, Monzani V, Moreno V, Wesse T, Lenz TL, Pumarola T, Rimoldi V, Bosari S, Albrecht W, Peter W, Romero-Gomez M, D'Amato M, Duga S, Banales JM,

Hov JR, Folseraas T, Valenti L, Franke A, Karlsen TH. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N Engl J Med*. 2020;383(16):1522-34. <https://doi.org/10.1056/NEJMoa2020283>

14. Kropski JA. Familial Interstitial Lung Disease. *Semin Respir Crit Care Med*. 2020;41(2):229-37. <https://doi.org/10.1055/s-0040-1708054>

15. Molgat-Seon Y, Schaeffer MR, Ryerson CJ, Guenette JA. Cardiopulmonary Exercise Testing in Patients With Interstitial Lung Disease. *Front Physiol*. 2020;11:832. <https://doi.org/10.3389/fphys.2020.00832>

16. Revel MP, Parkar AP, Prosch H, Silva M, Sverzellati N, Gleeson F, Brady A, European Society of R, the European Society of Thoracic I. COVID-19 patients and the radiology department - advice from the European Society of Radiology (ESR) and the European Society of Thoracic Imaging (ESTI). *Eur Radiol*. 2020;30(9):4903-9. <https://doi.org/10.1007/s00330-020-06865-y>

17. Walsh SLF, Wells AU, Desai SR, Poletti V, Piciucchi S, Dubini A, Nunes H, Valeyre D, Brillet PY, Kambouchner M, Morais A, Pereira JM, Moura CS, Grutters JC, van den Heuvel DA, van Es HW, van Oosterhout MF, Seldenrijk CA, Bendstrup E, Rasmussen F, Madsen LB, Gooptu B, Pomplun S, Taniguchi H, Fukuoka J, Johkoh T, Nicholson AG, Sayer C, Edmunds L, Jacob J, Kokosi MA, Myers JL, Flaherty KR, Hansell DM. Multicentre evaluation of multidisciplinary team meeting agreement on diagnosis in diffuse parenchymal lung disease: a case-cohort study. *Lancet Respir Med*. 2016;4(7):557-65. [https://doi.org/10.1016/S2213-2600\(16\)30033-9](https://doi.org/10.1016/S2213-2600(16)30033-9)

18. Frix AN, Cousin F, Refaee T, Bottari F, Vaidyanathan A, Desir C, Vos W, Walsh S, Occhipinti M, Lovinfosse P, Leijenaar RTH, Hustinx R, Meunier P, Louis R, Lambin P, Guiot J. Radiomics in Lung Diseases Imaging: State-of-the-Art for Clinicians. *J Pers Med*. 2021;11(7). <https://doi.org/10.3390/jpm11070602>

19. Mayerhoefer ME, Materka A, Langs G, Haggstrom I, Szczypinski P, Gibbs P, Cook G. Introduction to Radiomics. *J Nucl Med*. 2020;61(4):488-95. <https://doi.org/10.2967/jnumed.118.222893>

20. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin JC, Pieper S, Aerts H. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. 2017;77(21):e104-e7. <https://doi.org/10.1158/0008-5472.CAN-17-0339>

21. Yao L, Li Q, Liu Y, Li Q, Wang T, Zhou Z, Yin J. How to assess multimorbidity: a systematic review. *Front Public Health*. 2025;13:1525593. <https://doi.org/10.3389/fpubh.2025.1525593>
22. Siegel CH, Sammaritano LR. Systemic Lupus Erythematosus: A Review. *Jama*. 2024;331(17):1480-91. <https://doi.org/10.1001/jama.2024.2315>
23. Aletaha D, Smolen JS. Diagnosis and Management of Rheumatoid Arthritis: A Review. *Jama*. 2018;320(13):1360-72. <https://doi.org/10.1001/jama.2018.13103>
24. Banday MZ, Sameer AS, Nissar S. Pathophysiology of diabetes: An overview. *Avicenna J Med*. 2020;10(4):174-88. https://doi.org/10.4103/ajm.ajm_53_20
25. Volkmann ER, Andréasson K, Smith V. Systemic sclerosis. *Lancet*. 2023;401(10373):304-18. [https://doi.org/10.1016/s0140-6736\(22\)01692-0](https://doi.org/10.1016/s0140-6736(22)01692-0)
26. Lakhani DA, Schilling KG, Xu J, Bagnato F. Advanced Multicompartment Diffusion MRI Models and Their Application in Multiple Sclerosis. *American Journal of Neuroradiology*. 2020;41(5):751-7. <https://doi.org/10.3174/ajnr.A6484>
27. Lo J, Du K, Lee D, Zeng C, Athertya JS, Silva ML, Flechner R, Bydder GM, Ma Y. Multicompartment imaging of the brain using a comprehensive MR imaging protocol. *Neuroimage*. 2024;298:120800. <https://doi.org/10.1016/j.neuroimage.2024.120800>
28. Zeder K, Marsh LM, Avian A, Brcic L, Birnhuber A, Douschan P, Foris V, Sassmann T, Hoetzenecker K, Boehm PM, Kwapiszewska G, Olschewski A, Olschewski H, Kovacs G. Compartment-specific remodeling patterns in end-stage chronic obstructive pulmonary disease with and without severe pulmonary hypertension. *J Heart Lung Transplant*. 2024;43(7):1090-101. <https://doi.org/10.1016/j.healun.2024.02.1044>
29. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:13013781*. 2013.
30. HOTELLING H. RELATIONS BETWEEN TWO SETS OF VARIATES*. *Biometrika*. 1936;28(3-4):321-77. <https://doi.org/10.1093/biomet/28.3-4.321>

31. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*. 2009;10(3):515-34. <https://doi.org/10.1093/biostatistics/kxp008>
32. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*. 2009;8(1):Article28. <https://doi.org/10.2202/1544-6115.1470>
33. Ding DY, Li S, Narasimhan B, Tibshirani R. Cooperative learning for multiview analysis. *Proc Natl Acad Sci U S A*. 2022;119(38):e2202113119. <https://doi.org/10.1073/pnas.2202113119>
34. George B, Seals S, Aban I. Survival analysis and regression models. *J Nucl Cardiol*. 2014;21(4):686-94. <https://doi.org/10.1007/s12350-014-9908-2>
35. Dron L, Kalatharan V, Gupta A, Haggstrom J, Zariffa N, Morris AD, Arora P, Park J. Data capture and sharing in the COVID-19 pandemic: a cause for concern. *Lancet Digit Health*. 2022;4(10):e748-e56. [https://doi.org/10.1016/S2589-7500\(22\)00147-9](https://doi.org/10.1016/S2589-7500(22)00147-9)
36. Shuja J, Alanazi E, Alasmay W, Alashaikh A. COVID-19 open source data sets: a comprehensive survey. *Appl Intell (Dordr)*. 2021;51(3):1296-325. <https://doi.org/10.1007/s10489-020-01862-6>
37. Alderman JE, Charalambides M, Sachdeva G, Laws E, Palmer J, Lee E, Menon V, Malik Q, Vadera S, Calvert M, Ghassemi M, McCradden MD, Ordish J, Mateen B, Summers C, Gath J, Matin RN, Denniston AK, Liu X. Revealing transparency gaps in publicly available COVID-19 datasets used for medical artificial intelligence development-a systematic review. *Lancet Digit Health*. 2024;6(11):e827-e47. [https://doi.org/10.1016/S2589-7500\(24\)00146-8](https://doi.org/10.1016/S2589-7500(24)00146-8)
38. Brodin P. Immune determinants of COVID-19 disease presentation and severity. *Nat Med*. 2021;27(1):28-33. <https://doi.org/10.1038/s41591-020-01202-8>
39. Silva M, Milanese G, Seletti V, Ariani A, Sverzellati N. Pulmonary quantitative CT imaging in focal and diffuse disease: current research and clinical applications. *Br J Radiol*. 2018;91(1083):20170644. <https://doi.org/10.1259/bjr.20170644>

40. Ji H, Hu C, Yang X, Liu Y, Ji G, Ge S, Wang X, Wang M. Lymph node metastasis in cancer progression: molecular mechanisms, clinical significance and therapeutic interventions. *Signal Transduct Target Ther.* 2023;8(1):367. <https://doi.org/10.1038/s41392-023-01576-4>
41. Greene FL, Sobin LH. The staging of cancer: a retrospective and prospective appraisal. *CA Cancer J Clin.* 2008;58(3):180-90. <https://doi.org/10.3322/CA.2008.0001>
42. Dinwiddie R, Sonnappa S. Systemic diseases and the lung. *Paediatr Respir Rev.* 2005;6(3):181-9. <https://doi.org/10.1016/j.prrv.2005.06.007>
43. Nikkho SM, Richter MJ, Shen E, Abman SH, Antoniou K, Chung J, Fernandes P, Hassoun P, Lazarus HM, Olschewski H, Piccari L, Psotka M, Saggat R, Shlobin OA, Stockbridge N, Vitulo P, Vizza CD, Wort SJ, Nathan SD. Clinical significance of pulmonary hypertension in interstitial lung disease: A consensus statement from the Pulmonary Vascular Research Institute's innovative drug development initiative-Group 3 pulmonary hypertension. *Pulm Circ.* 2022;12(3):e12127. <https://doi.org/10.1002/pul2.12127>
44. El Bcheraoui C, Weishaar H, Pozo-Martin F, Hanefeld J. Assessing COVID-19 through the lens of health systems' preparedness: time for a change. *Global Health.* 2020;16(1):112. <https://doi.org/10.1186/s12992-020-00645-5>
45. Wu L, Kong X. COVID-19 pandemic: ethical issues and recommendations for emergency triage. *Front Public Health.* 2023;11:1160769. <https://doi.org/10.3389/fpubh.2023.1160769>
46. Choi BY, Grace AR, Tsai J. Heterogeneity of COVID-19 symptoms and associated factors: Longitudinal analysis of laboratory-confirmed COVID-19 cases in San Antonio. *PLoS One.* 2023;18(12):e0295418. <https://doi.org/10.1371/journal.pone.0295418>
47. Kirkpatrick JN, Hull SC, Fedson S, Mullen B, Goodlin SJ. Scarce-Resource Allocation and Patient Triage During the COVID-19 Pandemic: JACC Review Topic of the Week. *J Am Coll Cardiol.* 2020;76(1):85-92. <https://doi.org/10.1016/j.jacc.2020.05.006>

48. Kharroubi SA, Diab-El-Harake M. Sex-differences in COVID-19 diagnosis, risk factors and disease comorbidities: A large US-based cohort study. *Front Public Health*. 2022;10:1029190. <https://doi.org/10.3389/fpubh.2022.1029190>
49. Ceruti S, Glotta A, Biggiogero M, Maida PA, Marzano M, Urso P, Bona G, Garzoni C, Molnar Z. Admission criteria in critically ill COVID-19 patients: A physiology-based approach. *PLoS One*. 2021;16(11):e0260318. <https://doi.org/10.1371/journal.pone.0260318>
50. Lavrentieva A, Kaimakamis E, Voutsas V, Bitzani M. An observational study on factors associated with ICU mortality in Covid-19 patients and critical review of the literature. *Sci Rep*. 2023;13(1):7804. <https://doi.org/10.1038/s41598-023-34613-x>
51. Miro O, Llorens P, Jimenez S, Pinera P, Burillo-Putze G, Martin A, Martin-Sanchez FJ, Garcia-Lamberetchs EJ, Jacob J, Alquezar-Arbe A, Modol JM, Lopez-Diez MP, Guardiola JM, Cardozo C, Lucas Imbernon FJ, Aguirre Tejedo A, Garcia Garcia A, Ruiz Grinspan M, Llopis Roca F, Gonzalez Del Castillo J, Spanish Investigators on Emergency Situations Team N. Frequency, Risk Factors, Clinical Characteristics, and Outcomes of Spontaneous Pneumothorax in Patients With Coronavirus Disease 2019: A Case-Control, Emergency Medicine-Based Multicenter Study. *Chest*. 2021;159(3):1241-55. <https://doi.org/10.1016/j.chest.2020.11.013>
52. Lu R, Wang Y, Wang W, Nie K, Zhao Y, Su J, Deng Y, Zhou W, Li Y, Wang H, Wang W, Ke C, Ma X, Wu G, Tan W. Complete Genome Sequence of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) from the First Imported MERS-CoV Case in China. *Genome Announc*. 2015;3(4). <https://doi.org/10.1128/genomeA.00818-15>
53. Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Penaranda S, Bankamp B, Maher K, Chen MH, Tong S, Tamin A, Lowe L, Frace M, DeRisi JL, Chen Q, Wang D, Erdman DD, Peret TC, Burns C, Ksiazek TG, Rollin PE, Sanchez A, Liffick S, Holloway B, Limor J, McCaustland K, Olsen-Rasmussen M, Fouchier R, Gunther S, Osterhaus AD, Drosten C, Pallansch MA, Anderson LJ, Bellini WJ. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*. 2003;300(5624):1394-9. <https://doi.org/10.1126/science.1085952>
54. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, Pearson CAB, Russell TW, Tully DC, Washburne AD, Wenseleers T, Gimma A, Waite W, Wong KLM, van Zandvoort K, Silverman JD, Group CC-W, Consortium C-GU, Diaz-Ordaz K, Keogh R, Eggo RM, Funk S, Jit M, Atkins KE, Edmunds WJ. Estimated transmissibility

and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*. 2021;372(6538). <https://doi.org/10.1126/science.abg3055>

55. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill*. 2017;22(13). <https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494>

56. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;34(23):4121-3. <https://doi.org/10.1093/bioinformatics/bty407>

57. Andrew G, Arora R, Bilmes J, Livescu K. Deep Canonical Correlation Analysis. *PMLR*. 2013;28(3):1247-55.

58. Cottin V, Hirani NA, Hotchkiss DL, Nambiar AM, Ogura T, Otaola M, Skowasch D, Park JS, Poonyagariyagorn HK, Wuyts W, Wells AU. Presentation, diagnosis and clinical course of the spectrum of progressive-fibrosing interstitial lung diseases. *Eur Respir Rev*. 2018;27(150). <https://doi.org/10.1183/16000617.0076-2018>

59. Raghu G, Remy-Jardin M, Myers JL, Richeldi L, Ryerson CJ, Lederer DJ, Behr J, Cottin V, Danoff SK, Morell F, Flaherty KR, Wells A, Martinez FJ, Azuma A, Bice TJ, Bouros D, Brown KK, Collard HR, Duggal A, Galvin L, Inoue Y, Jenkins RG, Johkoh T, Kazerooni EA, Kitaichi M, Knight SL, Mansour G, Nicholson AG, Pipavath SNJ, Buendia-Roldan I, Selman M, Travis WD, Walsh S, Wilson KC, American Thoracic Society ERSJRS, Latin American Thoracic S. Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am J Respir Crit Care Med*. 2018;198(5):e44-e68. <https://doi.org/10.1164/rccm.201807-1255ST>

60. Amin R, Vaishali K, Maiya GA, Mohapatra AK, Acharya V, Dale MT, Alison JA. Relationships between disease severity and measures of health status in people with interstitial lung disease in India: an observational study. *Sci Rep*. 2025;15(1):16985. <https://doi.org/10.1038/s41598-025-01877-4>

61. Flaherty KR, Andrei AC, Murray S, Fraley C, Colby TV, Travis WD, Lama V, Kazerooni EA, Gross BH, Toews GB, Martinez FJ. Idiopathic pulmonary fibrosis: prognostic value of changes in physiology and six-minute-walk test. *Am J Respir Crit Care Med*. 2006;174(7):803-9. <https://doi.org/10.1164/rccm.200604-488OC>

62. Walsh SLF, Mackintosh JA, Calandriello L, Silva M, Sverzellati N, Larici AR, Humphries SM, Lynch DA, Jo HE, Glaspole I, Grainge C, Goh N, Hopkins PMA, Moodley Y, Reynolds PN, Zappala C, Keir G, Cooper WA, Mahar AM, Ellis S, Wells AU, Corte TJ. Deep Learning-based Outcome Prediction in Progressive Fibrotic Lung Disease Using High-Resolution Computed Tomography. *Am J Respir Crit Care Med*. 2022;206(7):883-91. <https://doi.org/10.1164/rccm.202112-2684OC>
63. Mueller-Mang C, Grosse C, Schmid K, Stiebellehner L, Bankier AA. What every radiologist should know about idiopathic interstitial pneumonias. *Radiographics*. 2007;27(3):595-615. <https://doi.org/10.1148/rg.273065130>
64. Romei C, Tavanti LM, Taliani A, De Liperi A, Karwoski R, Celi A, Palla A, Bartholmai BJ, Falaschi F. Automated Computed Tomography analysis in the assessment of Idiopathic Pulmonary Fibrosis severity and progression. *Eur J Radiol*. 2020;124:108852. <https://doi.org/10.1016/j.ejrad.2020.108852>
65. Jacob J, Bartholmai BJ, Rajagopalan S, Kokosi M, Nair A, Karwoski R, Raghunath SM, Walsh SL, Wells AU, Hansell DM. Automated Quantitative Computed Tomography Versus Visual Computed Tomography Scoring in Idiopathic Pulmonary Fibrosis: Validation Against Pulmonary Function. *J Thorac Imaging*. 2016;31(5):304-11. <https://doi.org/10.1097/RTI.0000000000000220>
66. Lynch DA, Sverzellati N, Travis WD, Brown KK, Colby TV, Galvin JR, Goldin JG, Hansell DM, Inoue Y, Johkoh T, Nicholson AG, Knight SL, Raoof S, Richeldi L, Ryerson CJ, Ryu JH, Wells AU. Diagnostic criteria for idiopathic pulmonary fibrosis: a Fleischner Society White Paper. *Lancet Respir Med*. 2018;6(2):138-53. [https://doi.org/10.1016/S2213-2600\(17\)30433-2](https://doi.org/10.1016/S2213-2600(17)30433-2)
67. Walsh SL, Calandriello L, Sverzellati N, Wells AU, Hansell DM, Consort UIPO. Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT. *Thorax*. 2016;71(1):45-51. <https://doi.org/10.1136/thoraxjnl-2015-207252>
68. Kim JS, Pugashetti J, Ma SF, Huang Y, Podolanczuk AJ, Lynch DA, Oh A, Mychaleckyj J, Manichaikul A, Rich S, Chua F, Adams TM, Flaherty KR, Lee JS, Lasky JA, Raghu G, Murray S, Spino C, Martinez FJ, Noth I, Humphries SM, Adegunsoye A, Molyneaux PL, Oldham JM, Newton CA. Associations of interstitial lung disease subtype and CT pattern with lung function and survival. *Thorax*. 2025. <https://doi.org/10.1136/thorax-2024-222149>

69. Zamora-Legoff JA, Krause ML, Crowson CS, Ryu JH, Matteson EL. Progressive Decline of Lung Function in Rheumatoid Arthritis-Associated Interstitial Lung Disease. *Arthritis Rheumatol.* 2017;69(3):542-9. <https://doi.org/10.1002/art.39971>

70. Maron BA. Revised Definition of Pulmonary Hypertension and Approach to Management: A Clinical Primer. *J Am Heart Assoc.* 2023;12(8):e029024. <https://doi.org/10.1161/JAHA.122.029024>

71. Sharara RS, Hattab Y, Patel K, DiSilvio B, Singh AC, Malik K. Introduction to the Anatomy and Physiology of Pulmonary Circulation. *Crit Care Nurs Q.* 2017;40(3):181-90. <https://doi.org/10.1097/CNQ.0000000000000157>

72. C H. Visceral nerves of the thorax (Gray's illustration): Radiopaedia.org; 2021 [Available from: <https://doi.org/10.5334/rID-92532>].

73. Nathan SD, Stinchon MR, Jr., Atcheson S, Simone L, Nelson M. Shining a spotlight on pulmonary hypertension associated with interstitial lung disease care: The latest advances in diagnosis and treatment. *J Manag Care Spec Pharm.* 2025;31(1-a Suppl):S2-S17. <https://doi.org/10.18553/jmcp.2025.31.1-a.s2>

74. Ley L, Grimminger F, Richter M, Tello K, Ghofrani A, Bandorski D. The Early Detection of Pulmonary Hypertension. *Dtsch Arztebl Int.* 2023;120(48):823-30. <https://doi.org/10.3238/arztebl.m2023.0222>

75. Humbert M, Kovacs G, Hoeper MM, Badagliacca R, Berger RMF, Brida M, Carlsen J, Coats AJS, Escribano-Subias P, Ferrari P, Ferreira DS, Ghofrani HA, Giannakoulas G, Kiely DG, Mayer E, Meszaros G, Nagavci B, Olsson KM, Pepke-Zaba J, Quint JK, Radegran G, Simonneau G, Sitbon O, Tonia T, Toshner M, Vachiery JL, Vonk Noordegraaf A, Delcroix M, Rosenkranz S, Group EESD. 2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension. *Eur Heart J.* 2022;43(38):3618-731. <https://doi.org/10.1093/eurheartj/ehac237>

76. Chen R, Liao H, Deng Z, He Z, Zheng Z, Lu J, Jiang M, Wu X, Guo W, Huang Z, Chen H, Hong C, Zhong N. Efficacy of computed tomography in diagnosing pulmonary hypertension: A systematic review and meta-analysis. *Front Cardiovasc Med.* 2022;9:966257. <https://doi.org/10.3389/fcvm.2022.966257>

77. Lettieri CJ, Nathan SD, Barnett SD, Ahmad S, Shorr AF. Prevalence and outcomes of pulmonary arterial hypertension in advanced idiopathic pulmonary fibrosis. *Chest*. 2006;129(3):746-52. <https://doi.org/10.1378/chest.129.3.746>
78. Proudman SM, Stevens WM, Sahhar J, Celermajer D. Pulmonary arterial hypertension in systemic sclerosis: the need for early detection and treatment. *Intern Med J*. 2007;37(7):485-94. <https://doi.org/10.1111/j.1445-5994.2007.01370.x>
79. Behr J, Ryu JH. Pulmonary hypertension in interstitial lung disease. *Eur Respir J*. 2008;31(6):1357-67. <https://doi.org/10.1183/09031936.00171307>
80. Waxman A, Restrepo-Jaramillo R, Thenappan T, Ravichandran A, Engel P, Bajwa A, Allen R, Feldman J, Argula R, Smith P, Rollins K, Deng C, Peterson L, Bell H, Tapson V, Nathan SD. Inhaled Treprostinil in Pulmonary Hypertension Due to Interstitial Lung Disease. *N Engl J Med*. 2021;384(4):325-34. <https://doi.org/10.1056/NEJMoa2008470>
81. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1972;34(2):187-202. <https://doi.org/https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>
82. Simon N, Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*. 2011;39(5):1 - 13. <https://doi.org/10.18637/jss.v039.i05>
83. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. *Journal of machine learning research*. 2019;20(129):1-30.
84. Benavoli A, Corani G, Mangili F. Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research*. 2016;17(1):152-61.
85. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. 2016;278(2):563-77. <https://doi.org/10.1148/radiol.2015151169>
86. Zwanenburg A, Vallieres M, Abdalah MA, Aerts H, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, Bogowicz M, Boldrini L, Buvat I, Cook GJR, Davatzikos C, Depeursinge A, Desseroit MC, Dinapoli N, Dinh CV, Echegaray S, El Naqa I, Fedorov AY, Gatta R, Gillies RJ, Goh V, Gotz M, Guckenberger M, Ha SM,

Hatt M, Isensee F, Lambin P, Leger S, Leijenaar RTH, Lenkiewicz J, Lippert F, Losnegard A, Maier-Hein KH, Morin O, Muller H, Napel S, Nioche C, Orlhac F, Pati S, Pfaehler EAG, Rahmim A, Rao AUK, Scherer J, Siddique MM, Sijtsema NM, Socarras Fernandez J, Spezi E, Steenbakkers R, Tanadini-Lang S, Thorwarth D, Troost EGC, Upadhaya T, Valentini V, van Dijk LV, van Griethuysen J, van Velden FHP, Whybra P, Richter C, Lock S. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology*. 2020;295(2):328-38. <https://doi.org/10.1148/radiol.2020191145>

87. Yip SS, Aerts HJ. Applications and limitations of radiomics. *Phys Med Biol*. 2016;61(13):R150-66. <https://doi.org/10.1088/0031-9155/61/13/R150>

88. Russo L, Charles-Davies D, Bottazzi S, Sala E, Boldrini L. Radiomics for clinical decision support in radiation oncology. *Clin Oncol (R Coll Radiol)*. 2024;36(8):e269-e81. <https://doi.org/10.1016/j.clon.2024.03.003>

89. Haga A, Iwasawa T, Misumi T, Okudela K, Oda T, Kitamura H, Saka T, Matsushita S, Baba T, Natsume-Kitatani Y, Utsunomiya D, Ogura T. Correlation of CT-based radiomics analysis with pathological cellular infiltration in fibrosing interstitial lung diseases. *Jpn J Radiol*. 2024;42(10):1157-67. <https://doi.org/10.1007/s11604-024-01607-2>

90. Best AC, Meng J, Lynch AM, Bozic CM, Miller D, Grunwald GK, Lynch DA. Idiopathic pulmonary fibrosis: physiologic tests, quantitative CT indexes, and CT visual scores as predictors of mortality. *Radiology*. 2008;246(3):935-40. <https://doi.org/10.1148/radiol.2463062200>

91. Rea G, Sverzellati N, Bocchino M, Lieto R, Milanese G, D'Alto M, Bocchini G, Maniscalco M, Valente T, Sica G. Beyond Visual Interpretation: Quantitative Analysis and Artificial Intelligence in Interstitial Lung Disease Diagnosis "Expanding Horizons in Radiology". *Diagnostics* (Basel). 2023;13(14). <https://doi.org/10.3390/diagnostics13142333>

92. He W, Huang W, Zhang L, Wu X, Zhang S, Zhang B. Radiogenomics: bridging the gap between imaging and genomics for precision oncology. *MedComm* (2020). 2024;5(9):e722. <https://doi.org/10.1002/mco2.722>

93. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJ. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. *Front Oncol.* 2015;5:272. <https://doi.org/10.3389/fonc.2015.00272>
94. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, Soussan M, Frouin F, Frouin V, Buvat I. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. *J Nucl Med.* 2018;59(8):1321-8. <https://doi.org/10.2967/jnumed.117.199935>
95. Orlhac F, Eertink JJ, Cottreau AS, Zijlstra JM, Thieblemont C, Meignan M, Boellaard R, Buvat I. A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies. *J Nucl Med.* 2022;63(2):172-9. <https://doi.org/10.2967/jnumed.121.262464>
96. Ibrahim A, Barufaldi B, Refaee T, Silva Filho TM, Acciavatti RJ, Salahuddin Z, Hustinx R, Mottaghy FM, Maidment ADA, Lambin P. MaasPenn Radiomics Reproducibility Score: A Novel Quantitative Measure for Evaluating the Reproducibility of CT-Based Handcrafted Radiomic Features. *Cancers.* 2022;14(7):1599.
97. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, Abdalah MA, Schabath MB, Goldgof DG, Mackin D, Court LE, Gillies RJ, Moros EG. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys.* 2017;44(3):1050-62. <https://doi.org/10.1002/mp.12123>
98. Hofmanninger J, Prayer F, Pan J, Rohrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *Eur Radiol Exp.* 2020;4(1):50. <https://doi.org/10.1186/s41747-020-00173-2>
99. Wasserthal J, Breit H-C, Meyer MT, Pradella M, Hinck D, Sauter AW, Heye T, Boll DT, Cyriac J, Yang S, Bach M, Segeroth M. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence.* 2023;5(5):e230024. <https://doi.org/10.1148/ryai.230024>
100. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal.* 2017;36:61-78. <https://doi.org/10.1016/j.media.2016.10.004>

101. Zhou M, Scott J, Chaudhury B, Hall L, Goldgof D, Yeom KW, Iv M, Ou Y, Kalpathy-Cramer J, Napel S, Gillies R, Gevaert O, Gatenby R. Radiomics in Brain Tumor: Image Assessment, Quantitative Feature Descriptors, and Machine-Learning Approaches. *AJNR Am J Neuroradiol*. 2018;39(2):208-16. <https://doi.org/10.3174/ajnr.A5391>
102. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Forster K, Aerts HJ, Dekker A, Fenstermacher D, Goldgof DB, Hall LO, Lambin P, Balagurunathan Y, Gatenby RA, Gillies RJ. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30(9):1234-48. <https://doi.org/10.1016/j.mri.2012.06.010>
103. Wu C, Chen X, Cai Y, Xia J, Zhou X, Xu S, Huang H, Zhang L, Zhou X, Du C, Zhang Y, Song J, Wang S, Chao Y, Yang Z, Xu J, Zhou X, Chen D, Xiong W, Xu L, Zhou F, Jiang J, Bai C, Zheng J, Song Y. Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA Intern Med*. 2020;180(7):934-43. <https://doi.org/10.1001/jamainternmed.2020.0994>
104. Kwee TC, Kwee RM. Chest CT in COVID-19: What the Radiologist Needs to Know. *Radiographics*. 2020;40(7):1848-65. <https://doi.org/10.1148/rg.2020200159>
105. Liao D, Zhou F, Luo L, Xu M, Wang H, Xia J, Gao Y, Cai L, Wang Z, Yin P, Wang Y, Tang L, Deng J, Mei H, Hu Y. Haematological characteristics and risk factors in the classification and prognosis evaluation of COVID-19: a retrospective cohort study. *Lancet Haematol*. 2020;7(9):e671-e8. [https://doi.org/10.1016/S2352-3026\(20\)30217-9](https://doi.org/10.1016/S2352-3026(20)30217-9)
106. Bao C, Liu X, Zhang H, Li Y, Liu J. Coronavirus Disease 2019 (COVID-19) CT Findings: A Systematic Review and Meta-analysis. *J Am Coll Radiol*. 2020;17(6):701-9. <https://doi.org/10.1016/j.jacr.2020.03.006>
107. Flores-Vega VR, Monroy-Molina JV, Jimenez-Hernandez LE, Torres AG, Santos-Preciado JJ, Rosales-Reyes R. SARS-CoV-2: Evolution and Emergence of New Viral Variants. *Viruses*. 2022;14(4). <https://doi.org/10.3390/v14040653>
108. Young BE, Fong SW, Chan YH, Mak TM, Ang LW, Anderson DE, Lee CY, Amrun SN, Lee B, Goh YS, Su YCF, Wei WE, Kalimuddin S, Chai LYA, Pada S, Tan SY, Sun L, Parthasarathy P, Chen YYC, Barkham T, Lin RTP, Maurer-Stroh S, Leo YS, Wang LF, Renia L, Lee VJ, Smith GJD, Lye DC, Ng LFP. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an

observational cohort study. Lancet. 2020;396(10251):603-11.
[https://doi.org/10.1016/S0140-6736\(20\)31757-8](https://doi.org/10.1016/S0140-6736(20)31757-8)

109. Andrianto, Al-Farabi MJ, Nugraha RA, Marsudi BA, Azmi Y. Biomarkers of endothelial dysfunction and outcomes in coronavirus disease 2019 (COVID-19) patients: A systematic review and meta-analysis. Microvasc Res. 2021;138:104224.
<https://doi.org/10.1016/j.mvr.2021.104224>

110. Topp G, Bouyea M, Cochran-Caggiano N, Ata A, Torres P, Jacob J, Wales D. Biomarkers Predictive of Extubation and Survival of COVID-19 Patients. Cureus. 2021;13(6):e15462. <https://doi.org/10.7759/cureus.15462>

111. Martinez Mesa A, Cabrera Cesar E, Martin-Montanez E, Sanchez Alvarez E, Lopez PM, Romero-Zerbo Y, Garcia-Fernandez M, Velasco Garrido JL. Acute Lung Injury Biomarkers in the Prediction of COVID-19 Severity: Total Thiol, Ferritin and Lactate Dehydrogenase. Antioxidants (Basel). 2021;10(8).
<https://doi.org/10.3390/antiox10081221>

112. Sorrentino S, Cacia M, Leo I, Polimeni A, Sabatino J, Spaccarotella CAM, Mongiardo A, De Rosa S, Indolfi C. B-Type Natriuretic Peptide as Biomarker of COVID-19 Disease Severity-A Meta-Analysis. J Clin Med. 2020;9(9).
<https://doi.org/10.3390/jcm9092957>

113. Stolp B, Stern M, Ambiel I, Hofmann K, Morath K, Gallucci L, Cortese M, Bartenschlager R, Ruggieri A, Graw F, Rudelius M, Keppler OT, Fackler OT. SARS-CoV-2 variants of concern display enhanced intrinsic pathogenic properties and expanded organ tropism in mouse models. Cell Rep. 2022;38(7):110387.
<https://doi.org/10.1016/j.celrep.2022.110387>

114. Weissman D, Alameh MG, de Silva T, Collini P, Hornsby H, Brown R, LaBranche CC, Edwards RJ, Sutherland L, Santra S, Mansouri K, Gobeil S, McDanal C, Pardi N, Hengartner N, Lin PJC, Tam Y, Shaw PA, Lewis MG, Boesler C, Sahin U, Acharya P, Haynes BF, Korber B, Montefiori DC. D614G Spike Mutation Increases SARS CoV-2 Susceptibility to Neutralization. Cell Host Microbe. 2021;29(1):23-31 e4.
<https://doi.org/10.1016/j.chom.2020.11.012>

115. Esper FP, Adhikari TM, Tu ZJ, Cheng YW, El-Haddad K, Farkas DH, Bosler D, Rhoads D, Procop GW, Ko JS, Jehi L, Li J, Rubin BP. Alpha to Omicron: Disease Severity

and Clinical Outcomes of Major SARS-CoV-2 Variants. *J Infect Dis.* 2023;227(3):344-52. <https://doi.org/10.1093/infdis/jiac411>

116. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.* 2017;18(1):186. <https://doi.org/10.1186/s13059-017-1319-7>

117. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol.* 2020;5(11):1403-7. <https://doi.org/10.1038/s41564-020-0770-5>

118. Kuzmin K, Adeniyi AE, DaSouza AK, Jr., Lim D, Nguyen H, Molina NR, Xiong L, Weber IT, Harrison RW. Machine learning methods accurately predict host specificity of coronaviruses based on spike sequences alone. *Biochem Biophys Res Commun.* 2020;533(3):553-8. <https://doi.org/10.1016/j.bbrc.2020.09.010>

119. Sokhansanj BA, Rosen GL. Predicting COVID-19 disease severity from SARS-CoV-2 spike protein sequence by mixed effects machine learning. *Comput Biol Med.* 2022;149:105969. <https://doi.org/10.1016/j.compbiomed.2022.105969>

120. Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. *Science.* 2021;371(6526):284-8. <https://doi.org/doi:10.1126/science.abd7331>

121. Singh R, Nagpal S, Pinna NK, Mande SS. Tracking mutational semantics of SARS-CoV-2 genomes. *Sci Rep.* 2022;12(1):15704. <https://doi.org/10.1038/s41598-022-20000-5>

122. Adjuik TA, Ananey-Obiri D. Word2vec neural model-based technique to generate protein vectors for combating COVID-19: a machine learning approach. *Int J Inf Technol.* 2022;14(7):3291-9. <https://doi.org/10.1007/s41870-022-00949-2>

123. Nagpal S, Pal R, Ashima, Tyagi A, Tripathi S, Nagori A, Ahmad S, Mishra HP, Malhotra R, Kutum R, Sethi T. Genomic Surveillance of COVID-19 Variants With Language Models and Machine Learning. *Front Genet.* 2022;13:858252. <https://doi.org/10.3389/fgene.2022.858252>

124. Gaia C, Maria Chiara C, Silvia L, Chiara A, Maria Luisa C, Giulia B, Silvia P, Lucia C, Alessandra T, Annarita S, Cristina V, Maria A, Maria Rosaria D, Giacinta A, Riccardo G, Zaher K, Andrea L, Maddalena B, Catalano C, Paolo R. Chest CT for early detection and management of coronavirus disease (COVID-19): a report of 314 patients admitted to Emergency Department with suspected pneumonia. *Radiol Med*. 2020;125(10):931-42. <https://doi.org/10.1007/s11547-020-01256-1>
125. McCague C, Ramlee S, Reinius M, Selby I, Hulse D, Piyatissa P, Bura V, Crispin-Ortuzar M, Sala E, Woitek R. Introduction to radiomics for a clinical audience. *Clin Radiol*. 2023;78(2):83-98. <https://doi.org/10.1016/j.crad.2022.08.149>
126. Shiri I, Salimi Y, Pakbin M, Hajianfar G, Avval AH, Sanaat A, Mostafaei S, Akhavanallaf A, Saberi A, Mansouri Z, Askari D, Ghasemian M, Sharifipour E, Sandoughdaran S, Sohrabi A, Sadati E, Livani S, Iranpour P, Kolahi S, Khateri M, Bijari S, Atashzar MR, Shayesteh SP, Khosravi B, Babaei MR, Jenabi E, Hasanian M, Shahhamzeh A, Foroghi Ghomi SY, Mozafari A, Teimouri A, Movaseghi F, Ahmari A, Goharpey N, Bozorgmehr R, Shirzad-Aski H, Mortazavi R, Karimi J, Mortazavi N, Besharat S, Afsharpad M, Abdollahi H, Geramifar P, Radmard AR, Arabi H, Rezaei-Kalantari K, Oveisi M, Rahmim A, Zaidi H. COVID-19 prognostic modeling using CT radiomic features and machine learning algorithms: Analysis of a multi-institutional dataset of 14,339 patients. *Comput Biol Med*. 2022;145:105467. <https://doi.org/10.1016/j.combiomed.2022.105467>
127. Sun Y, Salerno S, He X, Pan Z, Yang E, Sujimongkol C, Song J, Wang X, Han P, Kang J, Sjoding MW, Jolly S, Christiani DC, Li Y. Use of machine learning to assess the prognostic utility of radiomic features for in-hospital COVID-19 mortality. *Sci Rep*. 2023;13(1):7318. <https://doi.org/10.1038/s41598-023-34559-0>
128. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Tao Q, Sun Z, Xia L. Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. *Radiology*. 2020;296(2):E32-E40. <https://doi.org/10.1148/radiol.2020200642>
129. Guiot J, Vaidyanathan A, Deprez L, Zerka F, Danthine D, Frix AN, Thys M, Henket M, Canivet G, Mathieu S, Eftaxia E, Lambin P, Tsoutzidis N, Miraglio B, Walsh S, Moutschen M, Louis R, Meunier P, Vos W, Leijenaar RTH, Lovinfosse P. Development and Validation of an Automated Radiomic CT Signature for Detecting COVID-19. *Diagnostics* (Basel). 2020;11(1). <https://doi.org/10.3390/diagnostics11010041>

130. Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell*. 2018;173(2):283-5. <https://doi.org/10.1016/j.cell.2018.03.042>
131. Chen W, Sá RC, Bai Y, Napel S, Gevaert O, Lauderdale DS, Giger ML. Machine learning with multimodal data for COVID-19. *Heliyon*. 2023;9(7):e17934. <https://doi.org/https://doi.org/10.1016/j.heliyon.2023.e17934>
132. Ryerson CJ, Vittinghoff E, Ley B, Lee JS, Mooney JJ, Jones KD, Elicker BM, Wolters PJ, Koth LL, King TE, Collard HR. Predicting Survival Across Chronic Interstitial Lung Disease: The ILD-GAP Model. *Chest*. 2014;145(4):723-8. <https://doi.org/https://doi.org/10.1378/chest.13-1474>
133. Hirata M, Hara Y, Fujii H, Murohashi K, Saigusa Y, Zhao S, Kobayashi M, Nagasawa R, Tagami Y, Izawa A, Otsu Y, Watanabe K, Horita N, Kobayashi N, Kaneko T. ILD-GAP combined with the monocyte ratio could be a better prognostic prediction model than ILD-GAP in patients with interstitial lung diseases. *BMC Pulmonary Medicine*. 2024;24(1):16. <https://doi.org/10.1186/s12890-023-02833-6>
134. Matsuda S, Kotani T, Okazaki A, Nishioka D, Watanabe R, Gon T, Manabe A, Shoji M, Kadoba K, Hiwa R, Yamamoto W, Hashimoto M, Takeuchi T. Prediction model for respiratory-related mortality in microscopic polyangiitis with interstitial lung disease: multicentre REVEAL cohort study. *Rheumatology (Oxford)*. 2024;63(6):1607-15. <https://doi.org/10.1093/rheumatology/kead444>
135. Álvarez Troncoso J, Porto Fuentes Ó, Fernández Velilla M, Gómez Carrera L, Soto Abánades C, Martínez Robles E, Sorriguieta Torre R, Ríos Blanco JJ. Evaluating the diagnostic and prognostic utility of serial KL-6 measurements in connective tissue disease patients at risk for interstitial lung disease: correlations with pulmonary function tests and high-resolution computed tomography. *BMC Pulmonary Medicine*. 2024;24(1):603. <https://doi.org/10.1186/s12890-024-03264-7>
136. Lange M, Boddu P, Singh A, Gross BD, Mei X, Liu Z, Bernheim A, Chung M, Huang M, Masseaux J, Dua S, Platt S, Sivakumar G, DeMarco C, Lee J, Fayad ZA, Yang Y, Padilla M, Jacobi A. Influence of thoracic radiology training on classification of interstitial lung diseases. *Clin Imaging*. 2023;97:14-21. <https://doi.org/10.1016/j.clinimag.2022.12.010>
137. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more

information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-6. <https://doi.org/10.1016/j.ejca.2011.11.036>

138. Xu W, Wu W, Zheng Y, Chen Z, Tao X, Zhang D, Zhao J, Wang K, Guo B, Luo Q, Han Q, Zhou Y, Ye S. A Computed Tomography Radiomics-Based Prediction Model on Interstitial Lung Disease in Anti-MDA5-Positive Dermatomyositis. *Front Med (Lausanne)*. 2021;8:768052. <https://doi.org/10.3389/fmed.2021.768052>

139. Han N, Guo Z, Zhu D, Zhang Y, Qin Y, Li G, Gu X, Jin L. A nomogram model combining computed tomography-based radiomics and Krebs von den Lungen-6 for identifying low-risk rheumatoid arthritis-associated interstitial lung disease. *Front Immunol*. 2024;15:1417156. <https://doi.org/10.3389/fimmu.2024.1417156>

140. He W, Cui B, Chu Z, Chen X, Liu J, Pang X, Huang X, Yin H, Lin H, Peng L. Radiomics based on HRCT can predict RP-ILD and mortality in anti-MDA5 + dermatomyositis patients: a multi-center retrospective study. *Respir Res*. 2024;25(1):252. <https://doi.org/10.1186/s12931-024-02843-w>

141. Qin S, Kang B, Liu H, Ji C, Li H, Zhang J, Wang X. A computed tomography-based radiomics nomogram for predicting overall survival in patients with connective tissue disease-associated interstitial lung disease. *Eur J Radiol*. 2023;165:110963. <https://doi.org/10.1016/j.ejrad.2023.110963>

142. Ash SY, Harmouche R, Vallejo DL, Villalba JA, Ostridge K, Gunville R, Come CE, Onieva Onieva J, Ross JC, Hunninghake GM, El-Chemaly SY, Doyle TJ, Nardelli P, Sanchez-Ferrero GV, Goldberg HJ, Rosas IO, San Jose Estepar R, Washko GR. Densitometric and local histogram based analysis of computed tomography images in patients with idiopathic pulmonary fibrosis. *Respir Res*. 2017;18(1):45. <https://doi.org/10.1186/s12931-017-0527-8>

143. Joye AA, Bogowicz M, Gote-Schniering J, Frauenfelder T, Guckenberger M, Maurer B, Tanadini-Lang S, Gabrys HS. Radiomics on slice-reduced versus full-chest computed tomography for diagnosis and staging of interstitial lung disease in systemic sclerosis: A comparative analysis. *Eur J Radiol Open*. 2024;13:100596. <https://doi.org/10.1016/j.ejro.2024.100596>

144. Zhu WW, Li H, Li YD, Sun L, Kong L, Ye X, Cai Q, Lv XZ. Risk assessment in interstitial lung disease: the incremental prognostic value of cardiopulmonary ultrasound. *BMC Pulm Med*. 2021;21(1):237. <https://doi.org/10.1186/s12890-021-01606-3>

145. Chu Y, Luo G, Zhou L, Cao S, Ma G, Meng X, Zhou J, Yang C, Xie D, Mu D, Henao R, Setti G, Xiao X, Wu L, Qiu Z, Gao X. Deep learning-driven pulmonary artery and vein segmentation reveals demography-associated vasculature anatomical differences. *Nature Communications*. 2025;16(1):2262. <https://doi.org/10.1038/s41467-025-56505-6>
146. Lin K, Sarnari R, Gordon D, Markl M, Carr J. Cine Mri-derived Radiomics Features Indicate Hemodynamic Changes in the Pulmonary Artery. *Journal of Cardiovascular Magnetic Resonance*. 2024;26. <https://doi.org/10.1016/j.jocmr.2024.100727>
147. Priya S, Aggarwal T, Ward C, Bathla G, Jacob M, Gerke A, Hoffman EA, Nagpal P. Radiomics Detection of Pulmonary Hypertension via Texture-Based Assessments of Cardiac MRI: A Machine-Learning Model Comparison-Cardiac MRI Radiomics in Pulmonary Hypertension. *J Clin Med*. 2021;10(9). <https://doi.org/10.3390/jcm10091921>
148. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermini LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14(6):587-9. <https://doi.org/10.1038/nmeth.4285>
149. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol*. 2018;35(2):518-22. <https://doi.org/10.1093/molbev/msx281>
150. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32(1):268-74. <https://doi.org/10.1093/molbev/msu300>
151. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinformatics*. 2020;69(1):e96. <https://doi.org/10.1002/cpbi.96>
152. Er AG, Ding DY, Er B, Uzun M, Cakmak M, Sadee C, Durhan G, Ozmen MN, Tanriover MD, Topeli A, Aydin Son Y, Tibshirani R, Unal S, Gevaert O. Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study. *NPJ Digit Med*. 2024;7(1):117. <https://doi.org/10.1038/s41746-024-01128-2>

153. Mallory JD, Mallory XF, Kolomeisky AB, Igoshin OA. Theoretical Analysis Reveals the Cost and Benefit of Proofreading in Coronavirus Genome Replication. *J Phys Chem Lett.* 2021;12(10):2691-8. <https://doi.org/10.1021/acs.jpcllett.1c00190>

154. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011;12(null):2825–30.

155. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, Buatti J, Aylward S, Miller JV, Pieper S, Kikinis R. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging.* 2012;30(9):1323-41. <https://doi.org/10.1016/j.mri.2012.05.001>

156. Wasserthal J, Breit HC, Meyer MT, Pradella M, Hinck D, Sauter AW, Heye T, Boll DT, Cyriac J, Yang S, Bach M, Segeroth M. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiol Artif Intell.* 2023;5(5):e230024. <https://doi.org/10.1148/ryai.230024>

157. Murphy A, Feger J, Ismail M. Windowing (CT): Radiopedia.org; 2025 [Available from: <https://radiopaedia.org/articles/52108>].

158. Hodcroft EB. CoVariants: SARS-CoV-2 Mutations and Variants of Interest 2021 [Available from: <https://covariants.org/>].

159. Kandeel M, Mohamed MEM, Abd El-Lateef HM, Venugopala KN, El-Beltagi HS. Omicron variant genome evolution and phylogenetics. *J Med Virol.* 2022;94(4):1627-32. <https://doi.org/10.1002/jmv.27515>

160. Gruys E, Toussaint MJ, Niewold TA, Koopmans SJ. Acute phase reaction and acute phase proteins. *J Zhejiang Univ Sci B.* 2005;6(11):1045-56. <https://doi.org/10.1631/jzus.2005.B1045>

161. Simon BA, Christensen GE, Low DA, Reinhardt JM. Computed tomography studies of lung mechanics. *Proc Am Thorac Soc.* 2005;2(6):517-21, 06-7. <https://doi.org/10.1513/pats.200507-076DS>

162. Cullu N, Kalemci S, Karakas O, Eser I, Yalcin F, Boyaci FN, Karakas E. Efficacy of CT in diagnosis of transudates and exudates in patients with pleural effusion. *Diagn Interv Radiol*. 2014;20(2):116-20. <https://doi.org/10.5152/dir.2013.13066>
163. Zhou M, Leung A, Echegaray S, Gentles A, Shrager JB, Jensen KC, Berry GJ, Plevritis SK, Rubin DL, Napel S, Gevaert O. Non-Small Cell Lung Cancer Radiogenomics Map Identifies Relationships between Molecular and Imaging Phenotypes with Prognostic Implications. *Radiology*. 2018;286(1):307-15. <https://doi.org/10.1148/radiol.2017161845>
164. Bartholomeus GA, van Amsterdam WAC, Harder AMD, Willemink MJ, van Hamersvelt RW, de Jong PA, Leiner T. Robustness of pulmonary nodule radiomic features on computed tomography as a function of varying radiation dose levels-a multi-dose in vivo patient study. *Eur Radiol*. 2023;33(10):7044-55. <https://doi.org/10.1007/s00330-023-09643-8>
165. Laino ME, Ammirabile A, Lofino L, Lundon DJ, Chiti A, Francone M, Savevski V. Prognostic findings for ICU admission in patients with COVID-19 pneumonia: baseline and follow-up chest CT and the added value of artificial intelligence. *Emerg Radiol*. 2022;29(2):243-62. <https://doi.org/10.1007/s10140-021-02008-y>
166. Zhao K, Xie H, Fonzo GA, Carlisle N, Osorio RS, Zhang Y. Defining Dementia Subtypes Through Neuropsychiatric Symptom-Linked Brain Connectivity Patterns. *bioRxiv*. 2023. <https://doi.org/10.1101/2023.07.02.547427>
167. Lee H, Park BY, Byeon K, Won JH, Kim M, Kim SH, Park H. Multivariate association between brain function and eating disorders using sparse canonical correlation analysis. *PLoS One*. 2020;15(8):e0237511. <https://doi.org/10.1371/journal.pone.0237511>
168. Mehta P, McAuley DF, Brown M, Sanchez E, Tattersall RS, Manson JJ, Hlth Across Speciality Collaboration UK. COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet*. 2020;395(10229):1033-4. [https://doi.org/10.1016/S0140-6736\(20\)30628-0](https://doi.org/10.1016/S0140-6736(20)30628-0)
169. Levey AS, Eckardt KU, Tsukamoto Y, Levin A, Coresh J, Rossert J, De Zeeuw D, Hostetter TH, Lameire N, Eknoyan G. Definition and classification of chronic kidney disease: a position statement from Kidney Disease: Improving Global Outcomes (KDIGO). *Kidney Int*. 2005;67(6):2089-100. <https://doi.org/10.1111/j.1523-1755.2005.00365.x>

170. Newsome PN, Cramb R, Davison SM, Dillon JF, Foulerton M, Godfrey EM, Hall R, Harrower U, Hudson M, Langford A, Mackie A, Mitchell-Thain R, Sennett K, Sheron NC, Verne J, Walmsley M, Yeoman A. Guidelines on the management of abnormal liver blood tests. *Gut*. 2018;67(1):6-19. <https://doi.org/10.1136/gutjnl-2017-314924>
171. Xu Q, Zhan X, Zhou Z, Li Y, Xie P, Zhang S, Li X, Yu Y, Zhou C, Zhang L, Gevaert O, Lu G. AI-based analysis of CT images for rapid triage of COVID-19 patients. *NPJ Digit Med*. 2021;4(1):75. <https://doi.org/10.1038/s41746-021-00446-z>
172. Gong K, Wu D, Arru CD, Homayounieh F, Neumark N, Guan J, Buch V, Kim K, Bizzo BC, Ren H, Tak WY, Park SY, Lee YR, Kang MK, Park JG, Carriero A, Saba L, Masjedi M, Talari H, Babaei R, Mobin HK, Ebrahimian S, Guo N, Digumarthy SR, Dayan I, Kalra MK, Li Q. A multi-center study of COVID-19 patient prognosis using deep learning-based CT image analysis and electronic health records. *Eur J Radiol*. 2021;139:109583. <https://doi.org/10.1016/j.ejrad.2021.109583>
173. Oi Y, Ogawa F, Yamashiro T, Matsushita S, Oguri A, Utada S, Misawa N, Honzawa H, Abe T, Takeuchi I. Prediction of prognosis in patients with severe COVID-19 pneumonia using CT score by emergency physicians: a single-center retrospective study. *Sci Rep*. 2023;13(1):4045. <https://doi.org/10.1038/s41598-023-31312-5>
174. Butler L, Karabayir I, Samie Tootooni M, Afshar M, Goldberg A, Akbilgic O. Image and structured data analysis for prognostication of health outcomes in patients presenting to the ED during the COVID-19 pandemic. *Int J Med Inform*. 2021;158:104662. <https://doi.org/10.1016/j.ijmedinf.2021.104662>
175. Chao H, Fang X, Zhang J, Homayounieh F, Arru CD, Digumarthy SR, Babaei R, Mobin HK, Mohseni I, Saba L, Carriero A, Falaschi Z, Pasche A, Wang G, Kalra MK, Yan P. Integrative analysis for COVID-19 patient outcome prediction. *Med Image Anal*. 2021;67:101844. <https://doi.org/10.1016/j.media.2020.101844>
176. Jiao Z, Choi JW, Halsey K, Tran TML, Hsieh B, Wang D, Eweje F, Wang R, Chang K, Wu J, Collins SA, Yi TY, Delworth AT, Liu T, Healey TT, Lu S, Wang J, Feng X, Atalay MK, Yang L, Feldman M, Zhang PJL, Liao WH, Fan Y, Bai HX. Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study. *Lancet Digit Health*. 2021;3(5):e286-e94. [https://doi.org/10.1016/S2589-7500\(21\)00039-X](https://doi.org/10.1016/S2589-7500(21)00039-X)

177. Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. *Nat Rev Microbiol.* 2017;15(3):183-92. <https://doi.org/10.1038/nrmicro.2016.182>
178. Sese L, Nunes H, Cottin V, Israel-Biet D, Crestani B, Guillot-Dudoret S, Cadranel J, Wallaert B, Tazi A, Maitre B, Prevot G, Marchand-Adam S, Hirschi S, Dury S, Giraud V, Gondouin A, Bonniaud P, Traclet J, Juvin K, Borie R, Carton Z, Freynet O, Gille T, Planes C, Valeyre D, Uzunhan Y. Gender Differences in Idiopathic Pulmonary Fibrosis: Are Men and Women Equal? *Front Med (Lausanne).* 2021;8:713698. <https://doi.org/10.3389/fmed.2021.713698>
179. Manning EP, Mishall P, Ramachandra AB, Hassab AHM, Lamy J, Peters DC, Murphy TE, Heerdt P, Singh I, Downie S, Choudhary G, Tellides G, Humphrey JD. Stiffening of the human proximal pulmonary artery with increasing age. *Physiological Reports.* 2024;12(12):e16090. <https://doi.org/https://doi.org/10.14814/phy2.16090>
180. Akagi S, Nakamura K, Sarashina T, Ejiri K, Kasahara S, Ito H. Progression of pulmonary artery dilatation in patients with pulmonary hypertension coexisting with a pulmonary artery aneurysm. *Journal of Cardiology.* 2018;71(5):517-22. <https://doi.org/https://doi.org/10.1016/j.jjcc.2017.11.005>
181. Kligerman SJ, Henry T, Lin CT, Franks TJ, Galvin JR. Mosaic Attenuation: Etiology, Methods of Differentiation, and Pitfalls. *Radiographics.* 2015;35(5):1360-80. <https://doi.org/10.1148/rg.2015140308>
182. Unsal S, Atas H, Albayrak M, Turhan K, Acar AC, Doğan T. Learning functional properties of proteins with language models. *Nature Machine Intelligence.* 2022;4(3):227-45. <https://doi.org/10.1038/s42256-022-00457-9>

APPENDICES

APPENDIX A

Table A: Reconstruction Kernels

Kernel type	Resolution	Manufacturer
B80s	Sharp	Siemens
B80f	Sharp	Siemens
B70f	Sharp	Siemens
B70s	Sharp	Siemens
B60f	Sharp	Siemens
B60s	Sharp	Siemens
B54d	Sharp	Siemens
B50f	Sharp	Siemens
B45s	Sharp	Siemens
Br58f	Sharp	Siemens
Br54s	Sharp	Siemens

YB	Sharp	Philips
YC	Sharp	Philips
YD	Sharp	Philips
I70f	Sharp	Siemens

Table A (cont.)

FC86	Sharp	Toshiba-Canon
FC56	Sharp	Toshiba-Canon
FC55	Sharp	Toshiba-Canon
FC52	Sharp	Toshiba-Canon
FC51	Sharp	Toshiba-Canon
FC30	Sharp	Toshiba-Canon
FC12	Sharp	Toshiba-Canon
FC03	Sharp	Toshiba-Canon
LUNG	Sharp	GE
B41f	Non-sharp	Siemens
B32d	Non-sharp	Siemens
B31f	Non-sharp	Siemens
B31s	Non-sharp	Siemens
B30f	Non-sharp	Siemens
B30s	Non-sharp	Siemens
B25f	Non-sharp	Siemens
B20f	Non-sharp	Siemens
Br40f	Non-sharp	Siemens
Br40d	Non-sharp	Siemens
Br36f	Non-sharp	Siemens

Br32f	Non-sharp	Siemens
Bv40f	Non-sharp	Siemens
Bv40d	Non-sharp	Siemens
Bv36d	Non-sharp	Siemens

Table: Reconstruction Kernels (cont.)

Table A (cont.)

B	Non-sharp	Philips
C	Non-sharp	Philips
YA	Non-sharp	Philips
FL04	Non-sharp	Philips
I50f	Non-sharp	Siemens
I31f	Non-sharp	Siemens
I26f	Non-sharp	Siemens
FC18	Non-sharp	Toshiba-Canon
FC13	Non-sharp	Toshiba-Canon
FC08	Non-sharp	Toshiba-Canon
STANDARD	Non-sharp	GE
CHST	Non-sharp	GE

APPENDIX B

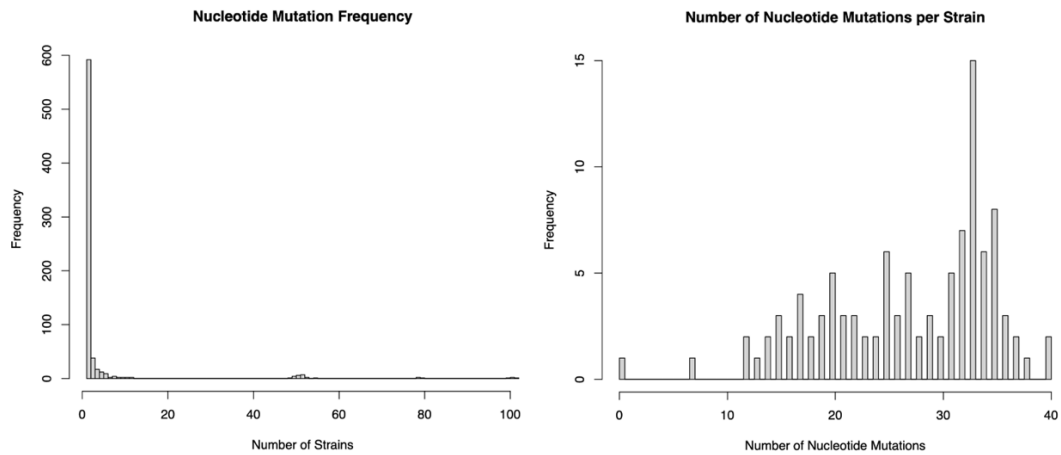


Figure B: Nucleotide Mutation Frequency and Number of Nucleotide mutations per strain histograms.

The histogram on the left side shows the nucleotide mutation frequency. 592 nucleotide mutations were observed at once. The histogram on the right side shows the number of nucleotide mutations per strain. *Reprinted from Er et al. (2024), "Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study", NPJ Digital Medicine, under CC BY license [152].*

APPENDIX C

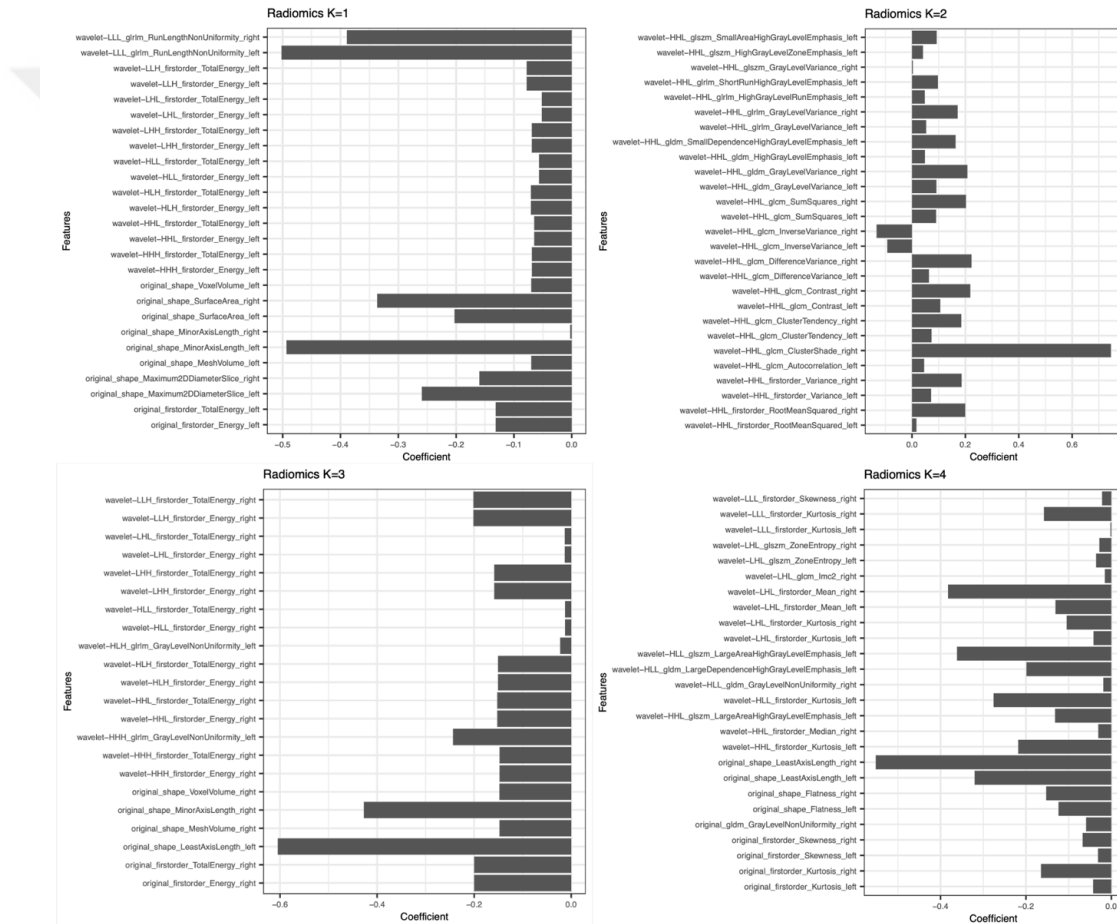


Figure C: Sparse CCA of Imaging and Clinical Data.

The first four canonical variates are provided. The first and third canonical vectors are found to be associated with sex, and the second and fourth ones were with lymphoma and CCI, respectively. *Reprinted from Er et al. (2024), “Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study”, NPJ Digital Medicine, under CC BY license [152].*

APPENDIX D

Table D: Unimodal and Multimodal Prediction Models Evaluated in 89 Patients with Post-hoc Analysis Results (n=89).

Reprinted from Er et al. (2024), “Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study”, NPJ Digital Medicine, under CC BY license [152].

M	Acc.	Adjusted p-values															
		C	L	R	D	W	CL	CR	LR	CD	C W	LD	LW	RD	R W	CLR D	CLR W
C	0.63 ± 0.04	-	.00 0	.00 0	.00 0	.00 0	.00 0	.00 0	.00 0	.13 8	.16 7	.00 0	.00 0	.00 0	.00 0	.000	.000
L	0.74 ± 0.03	.00 0	-	.87 4	.00 0	.00 0	.36 6	.03	.95 8	.00 0	.00 0	.85 4	.82 4	.99 9	.77 4	1	.005
R	0.72 ± 0.03	.00 0	.87 4	-	.00 0	.00 0	.00 3	.73 1	.08 2	.00 0	.00 0	1	1	1	.01 9	.693	.000
B	0.55 ± 0.03	.00 0	.00 0	.00 0	-	1	.00 0	.00 0	.00 0	.03 6	.00 0	.00 0	.00 0	.00 0	.00 0	.000	.000
W	0.55 ± 0.04	.00 0	.00 0	.00 0	1	-	.00 0	.00 0	.00 0	.14 8	.00 3	.00 0	.00 0	.00 0		.000	.000
CL	0.76 ± 0.03	.00 0	.36 6	.00 3	.00 0	.00 0	-	.00 0	1	.00 0	.00 0	.00 3	.00 4	.03	1	.431	.936
CR	0.71 ± 0.04	.00 0	.03 0	.73 1	.00 0	.00 0	.00 0	-	.00 0	.00 0	.00 0	.77 5	.90 6	.31 1	.00 0	.013	.000
LR	0.75 ± 0.03	.00 0	.95 8	.08 2	.00 0	.00 0	1	.00 0	-	.00 0	.00 0	.07 4	.08 1	.36 1	1	.981	.383
CB	0.59 ± 0.05	.13 8	.00 0	.00 6	.03 8	.14 0	.00 0	.00 0	.00 0	-	1	.00 0	.00 0	.00 0	.00 0	.000	.000
CW	0.60 ± 0.04	.16 7	.00 0	.00 0	.00 0	.00 3	.00 0	.00 0	.00 0	1	-	.00 0	.00 0	.00 0	.00 0	.000	.000
LB	0.72 ± 0.03	.00 0	.85 0	1	.00 0	.00 0	.00 3	.77 5	.07 4	.00 0	.00 0	-	1	1	.01 7	.658	.000
LW	0.72 ± 0.03	.00 0	.82 4	1	.00 0	.00 0	.00 4	.90 6	.08 6	.00 0	.00 0	1	-	1	.02 1	.642	.000
RB	0.73 ± 0.03	.00 0	.99 9	1	.00 0	.00 0	.03 1	.31 1	.36 1	.00 0	.00 0	1	1	-	.13 2	.987	.000
RW	0.75 ± 0.03	.00 0	.77 4	.01 9	.00 0	.00 0	1	.00 0	1	.00 0	.00 0	.01 7	.02 1	.13 2	-	.844	.567
CLRB	0.74 ± 0.03	.00 0	1	.69 3	.00 0	.00 0	.43 1	.01 3	.98 1	.00 0	.00 0	.65 8	.64 2	.98 7	.84 4	-	.006
CLR W	0.77 ± 0.03	.00 0	.00 5	.00 0	.00 0	.00 0	.93 6	.00 0	.38 3	.00 0	.00 0	.00 0	.00 0	.00 0	.56 7	.006	-

M	AUC	Adjusted p-values															
		C	L	R	D	W	CL	CR	LR	CD	C W	LD	LW	RD	R W	CLR D	CLR W
C	0.66 ± 0.03	-	.00 0	.00 0	.00 0	.00 0	.00 0	.00 0	.00 0	.08 3	.03	.00 0	.00 0	.00 0	.00 0	.000	.000

Table D (cont.)

L	0.83 ± 0.03	.00 0	-	1	.00 0	.00 0	.84 9	.13 4	1	.00 0	.00 0	.01 9	.32 9	1	.00 1	1	.000
R	0.83 ± 0.03	.00 0	1	-	.00 0	.00 0	.59 3	.35 4	.99 3	.00 0	.00 0	.06 4	.59 2	1	.00 2	1	.000
B	0.59 ± 0.05	.00 0	.00 0	.00 0	-	.24 0	.00 0	.00 0	.00 0	.17 0	.37 5	.00 0	.00 0	.00 0	.00 0	.000	.000
W	0.55 ± 0.05	.00 0	.00 0	.00 0	.24 0	-	.00 0	.00 0	.00 0	.00 0	.00 0	.00 0	.00 0	.00 0	.00 0	.000	.000
CL	0.85 ± 0.03	.00 0	.84 9	.59 3	.00 0	.00 0	-	.00 0	.01 5	.00 0	.00 0	.00 0	.00 6	.86 2	.97 2	.665	.462
CR	0.81 ± 0.02	.00 0	.13 4	.35 4	.00 0	.00 0	.00 0	-	.99 9	.00 0	.00 0	.99 9	1	.00 8	.00 0	.304	.000
LR	0.84 ± 0.03	.00 0	1	.99 3	.00 0	.00 0	.01 5	.99 9	-	.00 0	.00 0	.00 2	.06 9	1	.21 9	.997	.03
CB	0.62 ± 0.04	.08 3	.00 0	.00 0	.17 0	.00 0	.00 0	.00 0	.00 0	-	1	.00 0	.00 0	.00 0	.00 0	.000	.000
CW	0.62 ± 0.05	.03 0	.00 0	.00 0	.37 5	.00 0	.00 0	.00 0	.00 0	1	-	.00 0	.00 0	.00 0	.00 0	.000	.000
LB	0.80 ± 0.03	.00 0	.01 9	.06 4	.00 0	.00 0	.00 0	.99 9	.00 2	.00 0	.00 0	-	1	.00 1	.00 0	.052	.000
LW	0.81 ± 0.03	.00 0	.32 9	.59 2	.00 0	.00 0	.00 6	1	.06 9	.00 0	.00 0	1	-	.08 5	.00 0	.539	.000
RB	0.83 ± 0.02	.00 0	1	1	.00 0	.00 0	.86 2	.00 8	1	.00 0	.00 0	.00 1	.08 5	-	.00 0	1	.000
RW	0.86 ± 0.02	.00 0	.00 1	.00 2	.00 0	.00 0	.97 2	.00 0	.21 9	.00 0	.00 0	.00 0	.00 0	.00 0	-	.003	.956
CLRB	0.83 ± 0.02	.00 0	1	1	.00 0	.00 0	.66 5	.30 4	.99 7	.00 0	.00 0	.05 2	.53 9	1	.00 3	-	.000
CLR W	0.87 ± 0.03	.00 0	.00 0	.00 0	.00 0	.00 0	.46 2	.00 0	.03 0	.00 0	.00 0	.00 0	.00 0	.00 0	.95 6	.000	-
Abbreviations: M, Model; Acc, Accuracy; ICU, Intensive care unit; AUC, Area under the curve; C, Clinical data; L, Laboratory results; R, Radiomics; B, Viral-Binary encoding; W, Viral-Word2Vec encoding																	

APPENDIX E

Table E1: Lung features Associated with Transplant-Free Survival in ≥ 3 of 5 Outer Folds

Feature	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Selection Frequency
age_train	1.002	1.003	1.006	1.011	1.003	5
gender_train	2.474	2.314	2.785	2.200	2.838	5
lung_wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis_right_lung	1.062	1.02	1.037	1.065	1.008	5
lung_wavelet.LLL_firstorder_Entropy_right_lung	1.017	1.016	1.006	1.011	1.011	5
lung_wavelet.LLL_gldm_DifferenceAverage_right_lung	1.025	1.037	1.039	1.043	1.036	5
lung_wavelet.LLL_gldm_DifferenceEntropy_right_lung	1.007	1.02	1.020	1.025	1.015	5
lung_wavelet.LLL_gldm_Idmn_left_lung	0.993	0.97	0.991	0.983	0.979	5
lung_wavelet.LLL_gldm_Idmn_right_lung	0.939	0.925	0.945	0.915	0.911	5
lung_wavelet.LLL_gldm_Idn_left_lung	0.972	0.957	0.963	0.982	0.973	5
lung_wavelet.LLL_gldm_Idn_right_lung	0.916	0.919	0.928	0.918	0.913	5
lung_wavelet.LLL_gldm_JointEntropy_right_lung	1.023	1.029	1.019	1.019	1.017	5
lung_wavelet.LLL_gldm_SumEntropy_right_lung	1.019	1.017	1.006	1.013	1.014	5
lung_wavelet.LLL_gldm_DependenceNonUniformityNormalized_right_lung	1.039	1.034	1.038	1.019	1.017	5
lung_wavelet.LLL_gldm_SmallDependenceEmphasis_right_lung	1.046	1.046	1.050	1.041	1.036	5
lung_wavelet.LLL_gldm_SmallDependenceHighGrayLevelEmphasis_right_lung	1.036	1.053	1.037	1.026	1.026	5
lung_wavelet.LLL_glszm_ZonePercentage_right_lung	1.046	1.04	1.045	1.033	1.03	5
lung_original_shape_LeastAxisLength_left_lung	0.987	NA	0.990	0.975	0.971	4
lung_wavelet.HLH_gldm_SmallDependenceLowGrayLevelEmphasis_left_lung	1.036	1.003	1.016	NA	1.027	4
lung_wavelet.HLH_ngtdm_Coarseness_right_lung	1.002	NA	1.013	1.015	1.033	4
lung_wavelet.HLL_ngtdm_Coarseness_right_lung	1.025	NA	1.016	1.012	1.021	4
lung_wavelet.LHH_gldm_SmallDependenceLowGrayLevelEmphasis_right_lung	1.002	NA	1.011	1.014	1.004	4
lung_wavelet.LHL_ngtdm_Coarseness_right_lung	1.018	NA	1.010	1.010	1.016	4
lung_wavelet.LLL_grlm_RunEntropy_right_lung	1.007	1.005	NA	1.006	1.005	4
lung_wavelet.LLL_ngtdm_Contrast_right_lung	1.006	1.027	NA	1.028	1.03	4
lung_original_shape_Flatness_left_lung	0.989	NA	NA	0.973	0.978	3
lung_original_shape_MinorAxisLength_left_lung	0.962	NA	NA	0.990	0.998	3
lung_wavelet.HLL_firstorder_Kurtosis_right_lung	0.993	NA	0.992	NA	0.979	3
lung_wavelet.HLL_ngtdm_Coarseness_left_lung	1.004	NA	NA	1.019	1.008	3
lung_wavelet.LHH_ngtdm_Coarseness_right_lung	NA	NA	1.005	1.008	1.025	3
lung_wavelet.LHL_ngtdm_Coarseness_left_lung	1.001	NA	NA	1.016	1.005	3

Table E2: Pulmonary Artery Features Associated with Transplant-Free Survival in ≥ 5 of 5 Outer Folds

Feature	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Selection Frequency
age_train	1.001	1.002	1.004	1.011	1.003	5
gender_train	1.898	1.940	2.297	1.855	2.096	5
pulmonary_artery_original_firstorder_Energy_pulmonary_artery	1.030	1.005	1.006	1.004	1.003	5
pulmonary_artery_original_firstorder_TotalEnergy_pulmonary_artery	1.030	1.005	1.006	1.004	1.003	5
pulmonary_artery_original_glcml_ClusterShade_pulmonary_artery	1.020	1.001	1.005	1.005	1.002	5
pulmonary_artery_original_glszm_LargeAreaLowGrayLevelEmphasis_pulmonary_artery	1.046	1.004	1.010	1.016	1.012	5
pulmonary_artery_original_shape_Elongation_pulmonary_artery	0.921	0.990	0.986	0.975	0.983	5
pulmonary_artery_original_shape_Flatness_pulmonary_artery	0.904	0.990	0.987	0.986	0.983	5
pulmonary_artery_original_shape_MajorAxisLength_pulmonary_artery	1.021	1.008	1.009	1.012	1.013	5
pulmonary_artery_original_shape_SurfaceVolumeRatio_pulmonary_artery	0.984	0.995	0.993	0.993	0.995	5
pulmonary_artery_wavelet.HLL_glcml_ClusterShade_pulmonary_artery	1.003	1.002	1.007	1.005	1.004	5
pulmonary_artery_wavelet.HLH_firstorder_Mean_pulmonary_artery	1.064	1.007	1.008	1.012	1.009	5
pulmonary_artery_wavelet.HLH_firstorder_Variance_pulmonary_artery	0.999	0.996	0.994	0.992	0.994	5
pulmonary_artery_wavelet.HLH_glcml_ClusterProminence_pulmonary_artery	0.980	0.995	0.993	0.990	0.992	5
pulmonary_artery_wavelet.HLH_glcml_ClusterTendency_pulmonary_artery	0.999	0.996	0.994	0.992	0.994	5
pulmonary_artery_wavelet.HLH_glcml_Contrast_pulmonary_artery	0.996	0.996	0.994	0.992	0.994	5
pulmonary_artery_wavelet.HLH_glcml_DifferenceVariance_pulmonary_artery	0.981	0.996	0.994	0.992	0.994	5
pulmonary_artery_wavelet.HLH_glcml_Imc2_pulmonary_artery	0.994	0.998	0.996	0.991	0.997	5
pulmonary_artery_wavelet.HLH_glcml_SumSquares_pulmonary_artery	0.997	0.996	0.994	0.992	0.994	5
pulmonary_artery_wavelet.HLH_gldm_GrayLevelVariance_pulmonary_artery	0.999	0.996	0.994	0.992	0.994	5
pulmonary_artery_wavelet.HLH_gldm_GrayLevelVariance_pulmonary_artery	0.997	0.996	0.994	0.992	0.994	5
pulmonary_artery_wavelet.HLH_glszm_GrayLevelNonUniformity_pulmonary_artery	1.021	1.005	1.006	1.006	1.006	5
pulmonary_artery_wavelet.HLH_glszm_LargeAreaEmphasis_pulmonary_artery	1.016	1.002	1.008	1.014	1.011	5
pulmonary_artery_wavelet.HLH_glszm_LargeAreaLowGrayLevelEmphasis_pulmonary_artery	1.062	1.004	1.007	1.016	1.011	5
pulmonary_artery_wavelet.HLH_glszm_ZoneVariance_pulmonary_artery	1.011	1.002	1.007	1.014	1.011	5
pulmonary_artery_wavelet.HLH_ngtdm_Complexity_pulmonary_artery	0.972	0.995	0.993	0.990	0.993	5
pulmonary_artery_wavelet.HLH_ngtdm_Strength_pulmonary_artery	0.986	0.995	0.992	0.992	0.992	5
pulmonary_artery_wavelet.HLH_glcml_Imc2_pulmonary_artery	0.875	0.993	0.990	0.988	0.989	5
pulmonary_artery_wavelet.HLL_gldm_GrayLevelNonUniformity_pulmonary_artery	1.013	1.004	1.005	1.006	1.006	5
pulmonary_artery_wavelet.HLL_glszm_GrayLevelNonUniformity_pulmonary_artery	1.033	1.006	1.005	1.005	1.006	5
pulmonary_artery_wavelet.HLL_glszm_ZoneEntropy_pulmonary_artery	0.981	0.996	0.992	0.991	0.992	5
pulmonary_artery_wavelet.HLL_ngtdm_Busyness_pulmonary_artery	1.025	1.006	1.007	1.011	1.008	5
pulmonary_artery_wavelet.HLL_ngtdm_Coarseness_pulmonary_artery	0.996	0.996	0.993	0.995	0.994	5
pulmonary_artery_wavelet.LHH_firstorder_Mean_pulmonary_artery	1.048	1.009	1.012	1.016	1.013	5
pulmonary_artery_wavelet.LHH_firstorder_Median_pulmonary_artery	1.062	1.007	1.010	1.013	1.012	5
pulmonary_artery_wavelet.LHH_glszm_GrayLevelNonUniformity_pulmonary_artery	1.076	1.005	1.007	1.007	1.007	5
pulmonary_artery_wavelet.LHH_glszm_SizeZoneNonUniformityNormalized_pulmonary_artery	1.006	1.004	1.007	1.007	1.007	5
pulmonary_artery_wavelet.LHH_glszm_SmallAreaEmphasis_pulmonary_artery	1.011	1.004	1.007	1.008	1.007	5
pulmonary_artery_wavelet.LHL_firstorder_Energy_pulmonary_artery	1.002	1.004	1.005	1.004	1.004	5
pulmonary_artery_wavelet.LHL_firstorder_TotalEnergy_pulmonary_artery	1.002	1.004	1.005	1.004	1.004	5
pulmonary_artery_wavelet.LHL_glcml_ClusterProminence_pulmonary_artery	1.023	1.003	1.005	1.007	1.003	5
pulmonary_artery_wavelet.LHL_glcml_ClusterShade_pulmonary_artery	0.985	0.997	0.991	0.991	0.997	5
pulmonary_artery_wavelet.LHL_glcml_Idmn_pulmonary_artery	0.991	0.994	0.993	0.989	0.992	5
pulmonary_artery_wavelet.LHL_glcml_Idn_pulmonary_artery	0.992	0.995	0.993	0.989	0.992	5
pulmonary_artery_wavelet.LHL_gldm_DependenceNonUniformity_pulmonary_artery	1.027	1.005	1.005	1.006	1.006	5
pulmonary_artery_wavelet.LHL_gldm_RunLengthNonUniformity_pulmonary_artery	1.000	1.004	1.004	1.004	1.004	5
pulmonary_artery_wavelet.LHL_glszm_SizeZoneNonUniformity_pulmonary_artery	1.012	1.004	1.005	1.006	1.006	5
pulmonary_artery_wavelet.LHL_ngtdm_Coarseness_pulmonary_artery	0.987	0.996	0.994	0.995	0.994	5
pulmonary_artery_wavelet.LLH_firstorder_Maximum_pulmonary_artery	1.002	1.004	1.005	1.008	1.007	5
pulmonary_artery_wavelet.LLH_ngtdm_Complexity_pulmonary_artery	1.040	1.003	1.008	1.008	1.006	5
pulmonary_artery_wavelet.LLL_firstorder_Energy_pulmonary_artery	1.001	1.004	1.006	1.004	1.003	5
pulmonary_artery_wavelet.LLL_firstorder_Minimum_pulmonary_artery	1.001	1.008	1.006	1.007	1.008	5
pulmonary_artery_wavelet.LLL_firstorder_TotalEnergy_pulmonary_artery	1.001	1.004	1.006	1.004	1.003	5
pulmonary_artery_wavelet.LLL_glcml_Idmn_pulmonary_artery	0.948	0.994	0.994	0.991	0.991	5
pulmonary_artery_wavelet.LLL_glszm_LargeAreaLowGrayLevelEmphasis_pulmonary_artery	1.018	1.005	1.009	1.014	1.011	5
pulmonary_artery_wavelet.LLL_ngtdm_Coarseness_pulmonary_artery	0.974	0.996	0.992	0.994	0.993	5

Table E3: Lung and Pulmonary Artery Features Associated with Transplant-Free Survival in ≥ 3 of 5 Outer Folds

Feature	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Selection Frequency
age_train	1.001	1.004	1.005	1.010	1.002	5
gender_train	2.404	2.390	2.739	2.230	2.757	5
lung_wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis_right_lung	1.054	1.021	1.039	1.070	1.003	5
lung_wavelet.LLL_gldm_DifferenceAverage_right_lung	1.031	1.015	1.042	1.045	1.036	5
lung_wavelet.LLL_gldm_Idmn_right_lung	0.937	0.824	0.941	0.909	0.913	5
lung_wavelet.LLL_gldm_Idn_left_lung	0.973	0.957	0.963	0.985	0.972	5
lung_wavelet.LLL_gldm_Idn_right_lung	0.912	0.800	0.925	0.914	0.912	5
lung_wavelet.LLL_gldm_JointEntropy_right_lung	1.024	1.006	1.018	1.019	1.018	5
lung_wavelet.LLL_gldm_SmallDependenceEmphasis_right_lung	1.052	1.048	1.052	1.041	1.039	5
lung_wavelet.LLL_gldm_SmallDependenceHighGrayLevelEmphasis_right_lung	1.034	1.160	1.036	1.027	1.026	5
lung_wavelet.LLL_glszm_ZonePercentage_right_lung	1.053	1.017	1.047	1.033	1.033	5
lung_original_shape_LeastAxisLength_left_lung	0.982	NA	0.984	0.972	0.968	4
lung_wavelet.HLH_gldm_SmallDependenceLowGrayLevelEmphasis_left_lung	1.040	1.035	1.023	NA	1.027	4
lung_wavelet.HLH_ngtdm_Coarseness_right_lung	1.006	NA	1.017	1.016	1.034	4
lung_wavelet.LLL_ngtdm_Coarseness_right_lung	1.029	NA	1.023	1.015	1.022	4
lung_wavelet.LHH_gldm_SmallDependenceLowGrayLevelEmphasis_right_lung	1.003	NA	1.016	1.021	1.005	4
lung_wavelet.LHH_ngtdm_Coarseness_right_lung	1.021	NA	1.018	1.013	1.016	4
lung_wavelet.LLL_firstorder_Entropy_right_lung	1.016	NA	1.006	1.012	1.011	4
lung_wavelet.LLL_gldm_DifferenceEntropy_right_lung	1.012	NA	1.021	1.025	1.016	4
lung_wavelet.LLL_gldm_Idmn_left_lung	0.996	NA	0.993	0.984	0.980	4
lung_wavelet.LLL_gldm_SumEntropy_right_lung	1.017	NA	1.006	1.014	1.013	4
lung_wavelet.LLL_gldm_DependenceNonUniformityNormalized_right_lung	1.048	NA	1.040	1.019	1.021	4
pulmonary_artery_original_gldm_GrayLevelNonUniformity_pulmonary_artery	1.042	NA	1.011	1.006	1.042	4
lung_original_shape_Flatness_left_lung	0.991	NA	NA	0.963	0.983	3
lung_original_shape_MinorAxisLength_left_lung	0.953	NA	NA	0.987	0.995	3
lung_wavelet.HHL_firstorder_Mean_right_lung	1.038	1.016	NA	NA	1.029	3
lung_wavelet.HHL_firstorder_Kurtosis_right_lung	0.989	NA	0.990	NA	0.977	3
lung_wavelet.HLL_ngtdm_Coarseness_left_lung	1.004	NA	NA	1.021	1.009	3
lung_wavelet.LHH_ngtdm_Coarseness_right_lung	NA	NA	1.008	1.009	1.026	3
lung_wavelet.LHH_ngtdm_Coarseness_left_lung	1.001	NA	NA	1.018	1.005	3
lung_wavelet.LLL_firstorder_Skewness_right_lung	0.993	NA	NA	0.999	0.993	3
lung_wavelet.LLL_gldm_RunEntropy_right_lung	1.003	NA	NA	1.007	1.003	3
lung_wavelet.LLL_ngtdm_Contrast_right_lung	1.004	NA	NA	1.029	1.030	3
pulmonary_artery_wavelet.HLH_glszm_LargeAreaLowGrayLevelEmphasis_pulmonary_artery	1.022	NA	NA	1.011	1.011	3
pulmonary_artery_wavelet.HLL_glszm_GrayLevelNonUniformity_pulmonary_artery	1.016	NA	1.004	NA	1.005	3
pulmonary_artery_wavelet.LHH_glszm_GrayLevelNonUniformity_pulmonary_artery	1.012	NA	1.006	NA	1.005	3

Table E4: Lung Features Associated with Pulmonary Hypertension in ≥ 3 of 5 Outer Folds

Feature	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Selection Frequency
age	1.029	1.023	1.019	1.026	1.026	5
gender	1.472	2.249	2.447	2.290	1.899	5
lung_wavelet.LLL_glcM_SumEntropy_left_lung	1.022	1.024	1.054	1.155	1.016	5
lung_original_firstorder_Maximum_left_lung	NA	1.045	1.054	1.231	1.038	4
lung_original_glcM_ClusterProminence_right_lung	1.000	1.071	1.017	NA	1.061	4
lung_original_shape_Sphericity_left_lung	NA	0.976	0.949	0.961	0.999	4
lung_wavelet.HHH_firstorder_Mean_left_lung	NA	1.014	1.041	1.214	1.054	4
lung_wavelet.HHH_firstorder_Skewness_left_lung	NA	1.065	1.141	1.458	1.052	4
lung_wavelet.HHL_firstorder_Skewness_left_lung	NA	0.996	0.889	0.822	0.906	4
lung_wavelet.HLL_glcM_idm_left_lung	0.978	NA	0.997	0.898	0.949	4
lung_wavelet.HLL_ngtdm_Contrast_left_lung	1.004	1.013	1.002	NA	1.043	4
lung_wavelet.LHH_glcM_idn_right_lung	NA	1.004	1.049	1.155	1.050	4
lung_wavelet.LLL_firstorder_Mean_left_lung	NA	1.061	1.108	1.393	1.065	4
lung_wavelet.LLL_firstorder_Entropy_left_lung	1.022	1.019	1.043	NA	1.009	4
lung_wavelet.LLL_glcM_ClusterProminence_right_lung	1.002	1.091	1.013	NA	1.084	4
lung_wavelet.LLL_glcM_ClusterShade_right_lung	NA	1.062	1.016	1.308	1.048	4
lung_wavelet.LLL_glcM_DifferenceAverage_left_lung	1.017	1.026	1.025	NA	1.036	4
lung_wavelet.LLL_glcM_idn_left_lung	0.999	0.967	0.994	NA	0.953	4
lung_wavelet.LLL_glcM_JointEntropy_left_lung	1.027	1.023	1.028	NA	1.010	4
lung_wavelet.LLL_gldm_SmallDependenceLowGrayLevelEmphasis_right_lung	NA	0.977	0.980	0.914	0.996	4
lung_wavelet.LLL_glrM_RunEntropy_left_lung	1.012	1.009	1.049	NA	1.013	4
lung_wavelet.LLL_glszm_LowGrayLevelZoneEmphasis_right_lung	0.980	0.977	0.963	NA	0.963	4
lung_wavelet.LLL_glszm_SmallAreaLowGrayLevelEmphasis_right_lung	0.987	0.993	0.978	NA	0.976	4
lung_original_firstorder_Maximum_right_lung	NA	1.043	1.049	NA	1.038	3
lung_original_glcM_ClusterShade_right_lung	NA	1.046	1.021	NA	1.033	3
lung_original_glcM_SumEntropy_left_lung	1.020	1.019	1.021	NA	NA	3
lung_original_gldm_LowGrayLevelEmphasis_right_lung	NA	NA	0.995	0.987	0.989	3
lung_original_gldm_SmallDependenceHighGrayLevelEmphasis_left_lung	1.022	1.033	NA	NA	1.013	3
lung_original_glrM_RunEntropy_left_lung	NA	1.002	1.036	NA	1.001	3
lung_original_ngtdm_Busyness_left_lung	NA	0.980	0.997	NA	0.997	3
lung_original_shape_Elongation_right_lung	NA	0.998	0.964	0.899	NA	3
lung_original_shape_Flatness_left_lung	NA	0.977	0.997	0.985	NA	3
lung_wavelet.HHL_firstorder_Mean_left_lung	NA	0.986	0.996	NA	0.979	3
lung_wavelet.HHL_glcM_ClusterProminence_right_lung	NA	NA	1.146	1.082	1.025	3
lung_wavelet.HHL_glcM_MCC_right_lung	NA	0.972	0.932	NA	0.955	3
lung_wavelet.HHL_ngtdm_Contrast_left_lung	NA	1.037	1.027	NA	1.001	3
lung_wavelet.LHH_firstorder_Skewness_right_lung	NA	0.889	0.884	NA	0.928	3
lung_wavelet.LHH_glcM_Correlation_left_lung	NA	NA	1.052	1.070	1.035	3
lung_wavelet.LHH_gldm_LargeDependenceHighGrayLevelEmphasis_left_lung	NA	1.045	1.061	NA	1.019	3
lung_wavelet.LLL_firstorder_Kurtosis_right_lung	NA	1.011	1.026	NA	1.038	3
lung_wavelet.LLL_glcM_lmc2_right_lung	NA	1.057	NA	1.078	1.018	3
lung_wavelet.LLL_firstorder_Uniformity_left_lung	NA	0.984	NA	0.671	NA	3
lung_wavelet.LLL_glcM_idm_left_lung	NA	0.959	NA	NA	0.948	3
lung_wavelet.LLL_gldm_SmallDependenceEmphasis_left_lung	1.006	1.012	NA	NA	1.006	3
lung_wavelet.LLL_glszm_SizeZoneNonUniformityNormalized_left_lung	NA	1.010	1.032	NA	1.044	3
lung_wavelet.LLL_glszm_SizeZoneNonUniformityNormalized_right_lung	NA	NA	1.055	1.037	1.001	3
lung_wavelet.LLL_glszm_SmallAreaEmphasis_left_lung	NA	1.003	1.028	NA	1.036	3
lung_wavelet.LLL_ngtdm_Strength_right_lung	NA	NA	0.948	0.910	0.956	3

Table E5: Pulmonary Artery Features Associated with Pulmonary Hypertension in ≥ 3 of 5 Outer Folds

Feature	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Selection Frequency
age	1.029	1.016	1.022	1.023	1.024	5
gender	1.284	1.237	1.625	1.267	1.291	5
pulmonary_artery_original_glrIm_GrayLevelNonUniformity_pulmonary_artery	1.001	1.011	1.029	1.013	1.031	5
pulmonary_artery_original_shape_MajorAxisLength_pulmonary_artery	1.002	1.112	1.033	1.013	1.046	5
pulmonary_artery_wavelet.HHL_firstorder_Energy_pulmonary_artery	1.002	1.126	1.010	1.018	1.016	5
pulmonary_artery_wavelet.HHL_firstorder_TotalEnergy_pulmonary_artery	1.002	1.000	1.009	1.018	1.016	5
pulmonary_artery_original_firstorder_Energy_pulmonary_artery	1.001	1.030	1.014	1.001	NA	4
pulmonary_artery_original_firstorder_TotalEnergy_pulmonary_artery	1.001	1.000	1.014	1.001	NA	4
pulmonary_artery_original_shape_Maximum3DDiameter_pulmonary_artery	1.002	NA	1.013	1.011	1.030	4
pulmonary_artery_original_shape_MeshVolume_pulmonary_artery	1.002	NA	1.028	1.019	1.013	4
pulmonary_artery_original_shape_VoxelVolume_pulmonary_artery	1.002	NA	1.029	1.020	1.013	4
pulmonary_artery_wavelet.HHH_firstorder_Energy_pulmonary_artery	1.002	NA	1.011	1.013	1.008	4
pulmonary_artery_wavelet.HHH_firstorder_TotalEnergy_pulmonary_artery	1.002	NA	1.011	1.013	1.008	4
pulmonary_artery_wavelet.HLH_firstorder_Energy_pulmonary_artery	1.002	NA	1.002	1.009	1.004	4
pulmonary_artery_wavelet.HLH_firstorder_TotalEnergy_pulmonary_artery	1.002	NA	1.002	1.009	1.004	4
pulmonary_artery_wavelet.LHH_firstorder_Energy_pulmonary_artery	1.002	NA	1.010	1.013	1.007	4
pulmonary_artery_wavelet.LHH_firstorder_TotalEnergy_pulmonary_artery	1.002	NA	1.009	1.012	1.007	4
pulmonary_artery_wavelet.LLL_glrIm_GrayLevelNonUniformity_pulmonary_artery	1.002	1.272	NA	1.026	1.009	4
pulmonary_artery_original_ngtDm_Coarseness_pulmonary_artery	0.999	0.947	NA	NA	0.977	3
pulmonary_artery_original_shape_Flatness_pulmonary_artery	0.999	0.848	NA	NA	0.955	3
pulmonary_artery_original_shape_SurfaceArea_pulmonary_artery	1.002	NA	NA	1.002	1.007	3
pulmonary_artery_wavelet.HLH_glszm_GrayLevelNonUniformity_pulmonary_artery	1.001	NA	NA	1.012	1.003	3
pulmonary_artery_wavelet.LHL_firstorder_Energy_pulmonary_artery	1.002	NA	NA	1.011	1.011	3
pulmonary_artery_wavelet.LHL_firstorder_TotalEnergy_pulmonary_artery	1.002	NA	NA	1.011	1.011	3
pulmonary_artery_wavelet.LLH_firstorder_Energy_pulmonary_artery	1.002	NA	NA	1.009	1.005	3
pulmonary_artery_wavelet.LLH_firstorder_TotalEnergy_pulmonary_artery	1.002	NA	NA	1.009	1.005	3

Table E6: Lung and Pulmonary Artery Features Associated with Pulmonary Hypertension in ≥ 3 of 5 Outer Folds

Feature	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Selection Frequency
age	1.026	1.021	1.020	1.019	1.022	5
gender	1.280	1.999	1.840	1.554	1.537	5
lung_original_glcM_SumEntropy_left_lung	1.019	1.002	1.014	1.030	1.003	5
lung_wavelet.HLL_glcM_Idm_left_lung	0.976	0.998	0.995	0.897	0.995	5
lung_wavelet.LLL_firstorder_Entropy_left_lung	1.021	1.002	1.018	1.068	1.003	5
lung_wavelet.LLL_glcM_JointEntropy_left_lung	1.025	1.002	1.030	1.070	1.003	5
lung_wavelet.LLL_glcM_SumEntropy_left_lung	1.021	1.002	1.018	1.092	1.003	5
pulmonary_artery_original_glrIm_GrayLevelNonUniformity_pulmonary_artery	1.003	1.004	1.037	1.212	1.006	5
pulmonary_artery_wavelet.LLL_glrIm_GrayLevelNonUniformity_pulmonary_artery	1.005	1.003	1.004	1.058	1.005	5
lung_original_firstorder_Entropy_left_lung	1.015	1.002	1.007	NA	1.003	4
lung_original_glcM_JointEntropy_left_lung	1.006	1.002	1.005	NA	1.003	4
lung_original_gldm_SmallDependenceHighGrayLevelEmphasis_left_lung	1.021	1.002	1.015	NA	1.003	4
lung_wavelet.HLL_glcM_Idm_left_lung	0.991	0.998	NA	0.963	0.995	4
lung_wavelet.HLL_ngtdm_Contrast_left_lung	1.003	1.002	1.004	NA	1.005	4
lung_wavelet.LLL_glcM_DifferenceAverage_left_lung	1.011	1.002	1.036	NA	1.004	4
lung_wavelet.LLL_gldm_SmallDependenceEmphasis_left_lung	1.005	1.002	1.028	NA	1.004	4
lung_wavelet.LLL_gldm_SmallDependenceHighGrayLevelEmphasis_left_lung	1.004	1.002	1.016	NA	1.002	4
lung_wavelet.LLL_glrIm_RunEntropy_left_lung	1.011	1.002	1.004	NA	1.003	4
lung_wavelet.LLL_glszm_GrayLevelNonUniformityNormalized_left_lung	0.995	0.998	0.988	NA	0.998	4
lung_wavelet.LLL_glszm_LowGrayLevelZoneEmphasis_right_lung	0.984	0.998	NA	0.908	0.997	4
lung_wavelet.LLL_glszm_ZonePercentage_left_lung	1.001	1.002	1.021	NA	1.004	4
lung_wavelet.LLL_ngtdm_Contrast_left_lung	1.000	1.002	1.004	NA	1.003	4
pulmonary_artery_original_shape_MajorAxisLength_pulmonary_artery	NA	1.003	1.004	1.008	1.006	4
pulmonary_artery_original_shape_MeshVolume_pulmonary_artery	1.016	1.003	1.016	NA	1.005	4
pulmonary_artery_original_shape_SurfaceArea_pulmonary_artery	1.008	1.003	NA	1.105	1.005	4
pulmonary_artery_original_shape_VoxelVolume_pulmonary_artery	1.016	1.003	1.016	NA	1.005	4
pulmonary_artery_wavelet.HHH_firstorder_Energy_pulmonary_artery	1.009	1.003	1.009	NA	1.005	4
pulmonary_artery_wavelet.HHH_firstorder_TotalEnergy_pulmonary_artery	1.009	1.003	1.009	NA	1.005	4
pulmonary_artery_wavelet.HHL_firstorder_Energy_pulmonary_artery	1.012	1.003	1.006	NA	1.005	4
pulmonary_artery_wavelet.HHL_firstorder_TotalEnergy_pulmonary_artery	1.012	1.003	1.006	NA	1.005	4
pulmonary_artery_wavelet.HLH_firstorder_Energy_pulmonary_artery	1.006	1.003	1.006	NA	1.005	4
pulmonary_artery_wavelet.HLH_firstorder_TotalEnergy_pulmonary_artery	1.006	1.003	1.006	NA	1.005	4
pulmonary_artery_wavelet.LHH_firstorder_Energy_pulmonary_artery	1.009	1.003	1.007	NA	1.005	4
pulmonary_artery_wavelet.LHH_firstorder_TotalEnergy_pulmonary_artery	1.009	1.003	1.007	NA	1.005	4
pulmonary_artery_wavelet.LLH_firstorder_Energy_pulmonary_artery	1.009	1.003	1.003	NA	1.005	4
pulmonary_artery_wavelet.LLH_firstorder_TotalEnergy_pulmonary_artery	1.009	1.003	1.003	NA	1.005	4
lung_original_firstorder_Maximum_left_lung	NA	1.002	NA	1.067	1.004	3
lung_original_firstorder_Maximum_right_lung	NA	1.002	NA	1.058	1.004	3
lung_original_firstorder_Variance_right_lung	1.007	1.002	NA	NA	1.002	3
lung_original_glcM_ClusterShade_right_lung	NA	1.003	NA	1.048	1.004	3
lung_original_glcM_SumSquares_right_lung	1.006	1.002	NA	NA	1.002	3
lung_original_gldm_GrayLevelVariance_right_lung	1.007	1.002	NA	NA	1.002	3
lung_original_gldm_SmallDependenceHighGrayLevelEmphasis_right_lung	1.001	1.001	NA	NA	1.003	3
lung_original_glrIm_GrayLevelVariance_right_lung	1.008	1.002	NA	NA	1.002	3
lung_original_glrIm_LongRunLowGrayLevelEmphasis_left_lung	NA	0.999	NA	0.980	0.997	3
lung_original_glrIm_LongRunLowGrayLevelEmphasis_right_lung	NA	0.999	NA	0.997	0.997	3
lung_original_glszm_GrayLevelNonUniformityNormalized_left_lung	0.998	0.998	NA	NA	0.998	3
lung_original_glszm_SmallAreaHighGrayLevelEmphasis_right_lung	1.003	1.002	NA	NA	1.003	3
lung_original_shape_Elongation_right_lung	NA	0.999	NA	0.989	0.998	3
lung_wavelet.HHH_firstorder_Mean_left_lung	NA	1.002	NA	1.010	1.004	3
lung_wavelet.HHH_firstorder_Skewness_left_lung	NA	1.002	NA	1.160	1.004	3
lung_wavelet.HHL_firstorder_Skewness_left_lung	NA	0.998	NA	0.884	0.995	3
lung_wavelet.HLL_firstorder_Median_left_lung	NA	1.002	NA	1.021	1.003	3
lung_wavelet.LHH_glcM_Idm_right_lung	NA	1.001	NA	1.002	1.003	3
lung_wavelet.LHH_glcM_Idm_right_lung	NA	1.001	NA	1.041	1.003	3
lung_wavelet.LLH_firstorder_Mean_left_lung	NA	1.002	NA	1.124	1.005	3
lung_wavelet.LLL_firstorder_Uniformity_left_lung	NA	0.998	NA	0.955	0.997	3
lung_wavelet.LLL_glcM_ClusterProminence_right_lung	NA	1.003	NA	1.028	1.005	3
lung_wavelet.LLL_glcM_ClusterShade_right_lung	NA	1.003	NA	1.074	1.004	3
lung_wavelet.LLL_glcM_DifferenceEntropy_left_lung	NA	1.002	1.018	NA	1.003	3
lung_wavelet.LLL_glcM_Idm_left_lung	NA	0.998	0.991	NA	0.996	3
lung_wavelet.LLL_glcM_Idm_left_lung	NA	0.998	0.984	NA	0.996	3
lung_wavelet.LLL_glcM_MaximumProbability_right_lung	NA	0.999	NA	0.987	0.998	3
lung_wavelet.LLL_gldm_DependenceNonUniformityNormalized_left_lung	NA	1.002	1.016	NA	1.004	3
lung_wavelet.LLL_glrIm_GrayLevelNonUniformityNormalized_left_lung	NA	0.998	NA	0.973	0.997	3
lung_wavelet.LLL_glszm_SmallAreaHighGrayLevelEmphasis_right_lung	1.009	1.002	NA	NA	1.002	3
lung_wavelet.LLL_glszm_SmallAreaLowGrayLevelEmphasis_right_lung	0.990	0.998	NA	NA	0.997	3
lung_wavelet.LLL_ngtdm_Busyness_left_lung	NA	0.998	NA	0.974	0.998	3
pulmonary_artery_original_firstorder_Energy_pulmonary_artery	NA	1.002	NA	1.047	1.004	3
pulmonary_artery_original_firstorder_TotalEnergy_pulmonary_artery	NA	1.002	NA	1.047	1.004	3
pulmonary_artery_original_shape_Maximum3DDiameter_pulmonary_artery	1.007	1.003	NA	NA	1.005	3
pulmonary_artery_wavelet.HLL_firstorder_Energy_pulmonary_artery	1.001	1.003	NA	NA	1.004	3
pulmonary_artery_wavelet.HLL_firstorder_TotalEnergy_pulmonary_artery	1.001	1.003	NA	NA	1.004	3
pulmonary_artery_wavelet.LHL_firstorder_Energy_pulmonary_artery	1.008	1.003	NA	NA	1.004	3
pulmonary_artery_wavelet.LHL_firstorder_TotalEnergy_pulmonary_artery	1.008	1.003	NA	NA	1.004	3
pulmonary_artery_wavelet.LHL_glrIm_GrayLevelNonUniformity_pulmonary_artery	NA	1.003	NA	1.016	1.006	3

CURRICULUM VITAE

Name Ahmet Görkem Er

ORCID ID 0000-0001-6638-7955

EDUCATION

2018 - 2024 Infectious Diseases Specialist
Department of Clinical Microbiology and Infectious Diseases
Hacettepe University School of Medicine
Ankara, Turkey

2022 - 2023 Visiting Student Researcher
Stanford Center for Biomedical Informatics Research (BMIR)
Stanford University
Stanford, CA

2011 - 2016 Internal Medicine Specialist
Department of Internal Medicine
Hacettepe University School of Medicine
Ankara, Turkey

2004 - 2010 Medical Doctor (M.D.)
Istanbul University School of Medicine
Istanbul, Turkey

THESIS IN CLINICAL MICROBIOLOGY AND INFECTIOUS DISEASES

2024 Viral Genome Analysis and Genotyping of Human Genetic Variants Associated with Severe Disease in COVID-19 Patients.
Thesis Advisors: Prof. Dr. Yesim Aydin Son, Prof. Dr. Serhat Unal
Hacettepe University School of Medicine
Ankara, Turkey

THESIS IN INTERNAL MEDICINE

2016 Determining *Clostridium difficile* Infection Rate and Risk Factors in Hospitalized Patients in Hacettepe University Adult and Oncology Hospitals
Thesis Advisors: Prof. Dr. Gülay Sain Güven, Prof. Dr. Serhat Unal
Hacettepe University School of Medicine
Ankara, Turkey

CERTIFICATIONS

2022 Turkish Society of Clinical Microbiology and Infectious Diseases Infectious Diseases Board
2014 Turkish Internal Medicine Society Internal Medicine Board

HONORS AND AWARDS

2025 Hacettepe University Prof. Dr. M. Şeref Zileli Thesis of the Year Award
2025 Turkish Society of Clinical Microbiology and Infectious Diseases Young Investigator Award
2023 TÜBİTAK 2219-International Postdoctoral Research Fellowship Program
2022 2022-2023 Academic Year Fulbright Ph.D. Dissertation Grant

2016 Hacettepe University Prof. Dr. M. Şeref Zileli
Resident of the Year Award

PROFESSIONAL EXPERIENCE

01.2016 present	-	National Tuberculosis Scientific Committee Member Ministry of Health General Directorate of Public Health Ankara, Turkey
04.2018 11.2020	-	Turkish Thoracic Society Tuberculosis Study Group Executive Committee Member
03.2018 11.2020	-	Independent Expert Evaluator PROEMPOWER – Pre-commercial Procurement of Innovative ICT for Patient Empowerment and Self-management for Patients with Type 2 Diabetes Mellitus Horizon 2020 Program
10.2016 06.2018	-	Internist Karaman State Hospital Department of Internal Medicine Karaman, Turkey
01.2016 06.2016	-	Medical Doctor National Tuberculosis Department Ministry of Health General Directorate of Public Health Ankara, Turkey
05.2015 09.2015	-	Project Manager Project Code: TBI-2015-5751 Project Title: “The Role of <i>Triangular Clinic Model</i> in the Follow Up HIV/AIDS and Tuberculosis Patients and Evaluation of Adaptability This Model to Primary Health Care System in Turkey” funded by Hacettepe University

		Global Institute for Health and Human Rights, SUNY at Albany New York, USA
11.2010	-	Medical Doctor
08.2011		112 Emergency Department Tekirdag, Turkey

MEMBERSHIPS IN PROFESSIONAL ASSOCIATIONS

2023	-	Turkish Society of Clinical Microbiology and Infectious Diseases
present		
2017	-	Turkish Internal Medicine Society
present		
2017	-	European Society of Clinical Microbiology and Infectious Diseases
present		
2016	-	Turkish Thoracic Society
present		
2010	-	Turkish Medical Association
present		

COURSES AND WORKSHOPS

2023	Sanger Advanced Bioinformatics Workshop for the GPS and JUNO project Ankara, Turkey
2021	Innovations in treatment and management of drug- resistant tuberculosis Latvian WHO Collaborating Center for Research and Training in Management of Multidrug-Resistant Tuberculosis Riga, Latvia - ONLINE
2018	HUKAM Clinical Studies Education Meeting Ankara, Turkey
2018	MIRCIM 2018: 5 th McMaster International Review Course in Internal Medicine Krakow, Poland

- 2017 Tuberculosis Field Team Education Program
Ministry of Health General Directorate of Public Health
Antalya, Turkey
- 2016 Field Epidemiology Education Program Introduction Course
Ministry of Health General Directorate of Public Health
Ankara, Turkey
- 2016 WHO Interregional Workshop for Tuberculosis Control and Care Among Refugee and Migrant Populations
Catania, Italy
- 2015 HIV Update: Contemporary Issues in Management
Harvard Medical School
Boston, MA
- 2014 Operational Research Skills in One Day
45th Union World Congress on Lung Health
Barcelona, Spain

PUBLICATIONS

- 2024 **Er AG**, Çubuk C, Durusu-Tanrioer M, et al. Influenza Surveillance Under the Shadow of COVID-19 Pandemic: Phylogenetic Analysis of Influenza A (H1N1)pdm09 Viruses in Türkiye in 2019-20 Season. *Flora*. 2024;29(4):488-501. doi: 10.5578/flora.2024041131
- 2024 Kaygusuz Y, Üstün C, **Er AG**, et al. An Adult-Onset Chronic Granulomatous Disease Case with Hemophagocytic Lymphohistiocytosis Caused by Burkholderia and Aspergillus Infections. *Infect Dis Clin Microbiol*. 2024;6(4):334-338. doi: 10.36519/idcm.2024.381
- 2024 Bakir Ekinci P, Kara E, **Er AG**, Pinar A, Inkaya AC, Demirkan K, Metan G, Uzun O. Effect of inflammation on voriconazole levels in patients with invasive pulmonary aspergillosis. *J Infect Dev Ctries*. 2024;18(10):1617-1624. doi: 10.3855/jidc.19290

- 2024 **Er AG**, Aslan AT, Mikulska M, Akova M. Prevention and treatment of bacterial infections in patients with haematological cancers and hematopoietic stem cell transplantation: headways and shortcomings. *Clin Microbiol Infect.* 2024; S1198-743X(24)00447-6. doi: 10.1016/j.cmi.2024.09.015.
- 2024 **Er AG**, Ding DY, Er B, et al. Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study. *NPJ Digit Med.* 2024;7(1):117. doi:10.1038/s41746-024-01128-2
- 2023 Er B, **Er AG**, Gulmez D, et al. Diagnostic performance and longitudinal analysis of fungal biomarkers in COVID-19 associated pulmonary aspergillosis. *Heliyon.* 2023;9(11):e21721. doi:10.1016/j.heliyon.2023.e21721
- 2022 **Er AG**, Alonso AAR, Marin-Leon I, et al. Community-acquired pneumonia - An EFIM guideline critical appraisal adaptation for internists. *Eur J Intern Med.* 2022;106:1-8. doi:10.1016/j.ejim.2022.10.009
- 2022 Sayinalp-Arslan B, **Er AG**, Yildirim M, Kilicaslan B, Akinci SB, Uzun O. A case of Legionnaires' disease with severe rhabdomyolysis misdiagnosed as COVID-19. *Heliyon.* 2022;8(12):e12341. doi:10.1016/j.heliyon.2022.e12341
- 2022 Er B, **Er AG**, Gulmez D, Sahin TK, Halacli B, Durhan G, et al. A screening study for COVID-19-associated pulmonary aspergillosis in critically ill patients during the third wave of the pandemic. *Mycoses.* 2022;65(7):724-32. DOI: 10.1111/myc.13466
- 2021 Bahap M, Kara E, Er AG, Demirkan K, Unal S. Investigation and Comparison of Diagnosis and Treatment Guidelines for Community Acquired Pneumonia in terms of Pharmaceutical Care Requirements. *Flora the Journal of Infectious Diseases and Clinical Microbiology* 2021;26(1):50-66. DOI: 10.5578/flora.20219905.
- 2021 Er B, **Er AG**, Metan G, et al. Is COVID-19 a risk factor for invasive pulmonary aspergillosis in critically ill patients? *Tuberk Toraks* 2021;69(1):118-120. DOI: 10.5578/tt.20219918.

- 2021 Demir Cuha M, **Er AG**, Durukan BM, Karagoz A, Hazirolan G, Unal S. A rare case of necrotizing fasciitis of the abdominal wall due to *Hungatella effluvii* and *Streptococcus constellatus*. *Anaerobe* 2021;70:102409. DOI: 10.1016/j.anaerobe.2021.102409.
- 2021 Bakir Ekinci P, Kara E, **Er AG**, Inkaya AC, Demirkan K, Uzun O. Challenge in treating COVID-19 associate pulmonary aspergillosis: Supratherapeutic voriconazole levels. *Br J Clin Pharmacol* 2021. DOI: 10.1111/bcp.14953.
- 2021 Jakob CEM, Mahajan UM, Oswald M, ...**Er AG**..., et al. Prediction of COVID-19 deterioration in high-risk patients at diagnosis: an early warning score for advanced COVID-19 developed by machine learning. *Infection* 2021. DOI: 10.1007/s15010-021-01656-z.
- 2021 Telli Dizman G, Metan G, Ayaz Ceylan CM, ...**Er AG**..., et al. A COVID-19 First Evaluation Clinic at a University Hospital in Turkey. *Turk J Med Sci* 2021. DOI: 10.3906/sag-2104-152.
- 2021 Consortium C-CC, Group LS. Clinical presentation, disease course, and outcome of COVID-19 in hospitalized patients with and without pre-existing cardiac disease: a cohort study across 18 countries. *Eur Heart J* 2021. DOI: 10.1093/eurheartj/ehab656.
- 2020 Ozsurekci Y, Aykac K, **Er AG**, et al. Predictive value of cytokine/chemokine responses for the disease severity and management in children and adult cases with COVID-19. *J Med Virol* 2021;93(5):2828-2837. DOI: 10.1002/jmv.26683.
- 2020 Acar AC, **Er AG**, Burduglu HC, et al. Projecting the course of COVID-19 in Turkey: A probabilistic modeling approach. *Turk J Med Sci* 2021;51(1):16-27. DOI: 10.3906/sag-2005-378.
- 2020 **Er AG**, Unal S. 2019 Coronavirus Pandemic in Turkey and Across the World. *Flora the Journal of Infectious Diseases and Clinical Microbiology* 2020;25(1):1-8. DOI: 10.5578/flora.202001.

BOOK CHAPTERS

- 2021 COVID-19: Epidemiology, Global Projection.
Hacettepe University COVID-19 Pandemic Report.
ISBN: 978-975-97501-5-2. e-ISBN: 978-975-97501-
6-9
- 2019 Ministry of Health General Directorate of Public
Health Tuberculosis Diagnosis and Treatment
Guideline, ISBN: 978-975-590-717-8



TEZ İZİN FORMU / THESIS

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences ☐

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ☐

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics ☐

Enformatik Enstitüsü / Graduate School of Informatics ☒

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences ☐

YAZARIN / AUTHOR

Soyadı / Surname : ER

Adı / Name : AHMET GÖRKEM

Bölümü / Department : TIP BİLİŞİMİ

TEZİN ADI / TITLE OF THE THESIS (İngilizce / English) : MULTIMODAL DATA FUSION AND MULTICOMPARTMENT IMAGE ANALYSIS IN ACUTE AND CHRONIC LUNG DISEASES

TEZİN TÜRÜ / DEGREE: Yüksek Lisans / Master ☐ **Doktora / PhD** ☒

1. Tezin tamamı dünya çapında erişime açılacaktır. / Release the entire work immediately for access worldwide. ☒
2. Tez iki yıl süreyle erişime kapalı olacaktır. / Secure the entire work for patent and/or proprietary purposes for a period of two year. ☐
3. Tez altı ay süreyle erişime kapalı olacaktır. / Secure the entire work for period of six months. ☐
4. Tezim ile ilgili bir ambargo kararı bulunmamakla birlikte, tezimin OpenMETU'de erişime açılmasının mezuniyet tarihimden itibaren 1 (bir) yıl süreyle ertelenmesini talep ederim (Bu seçenek tezin sadece OpenMETU'de erişimini sınırlandırır). / Although there is no embargo decision regarding my thesis, I request that the public access to my thesis on OpenMETU be postponed for a period of one (1) year following my graduation date (This option only restricts access to the thesis on OpenMETU) ☐

Yazarın imzası / Signature

Tarih