UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior De Ingenieros De Telecomunicación

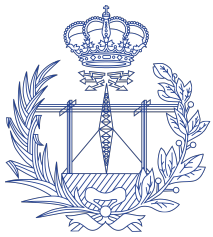
Spatio-temporal image analysis methods for lung cancer screening and therapy response prediction

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Benito Farina
MSc in Biomedical Engineering

Madrid, 2024



UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior De Ingenieros De Telecomunicación

Doctoral Degree in Electronic Systems Engineering

Spatio-temporal image analysis methods for lung cancer screening and therapy response prediction

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Benito Farina
MSc in Biomedical Engineering

Under the supervision of:
Dr. María Jesús Ledesma Carbayo

Madrid, 2024

Title: Spatio-temporal image analysis methods for lung cancer screening and therapy response prediction

Author: Benito Farina

Doctoral Programme: Electronic Systems Engineering

Thesis Supervision:

Dr. María Jesús Ledesma Carbayo, Doctora Ingeniera de Telecomunicación, Universidad Politécnica de Madrid

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

This thesis has been supported by an FPI grant from Spain's Ministry of Education.

A mio padre

Agradecimientos

Es difícil resumir en pocas líneas cinco años tan intensos y significativos. Un recorrido lleno de experiencias extraordinarias y desafíos difíciles.

En primer lugar, quiero agradecer a mi tutora Chus, quien me dió la inestimable oportunidad de vivir en Madrid. Por su constante apoyo, sus sabios consejos, las estimulantes discusiones y su gran optimismo, que fueron clave para completar mi doctorado.

Un sincero agradecimiento a todas las personas excepcionales que formaron parte de nuestro grupo de investigación BIT a lo largo de los años. Me gustaría mencionar en particular a Andrés, Eva, Juan, Juanjo, David, Ana, Daniel, Pedro, David. Por los preciosos momentos compartidos juntos, tanto dentro como fuera del laboratorio.

También quiero agradecer a Raúl por hacer posible mi experiencia en Boston. A Gonzalo, por el apoyo y la ayuda que me ofreció en Boston y en el último período del doctorado.

No puedo expresar más que un agradecimiento infinito a Feli, constante en mi vida desde que llegué a Madrid. Su cariño, su premura, su empatía y su paciencia han hecho estos años difíciles más sencillos, enriquecidos con recuerdos inolvidables.

A mis seres queridos, mi familia, que estuvo lejos y a quienes extrañé cada día. Siempre a mi lado, apoyándome en cada decisión. A mi querida madre Emilia y a mis dos extraordinarias hermanas Filomena y Federica, las mujeres más fuertes que conozco. Y a mi padre, que vive en mi corazón, cuya presencia siempre está viva a mi lado.

Ringraziamenti

È difficile riassumere in poche righe cinque anni così intensi e significativi. Un percorso ricco di esperienze straordinarie e sfide impegnative.

Innanzitutto, desidero ringraziare la mia tutor Chus, che mi ha offerto l'inestimabile opportunità di vivere a Madrid. Per il suo sostegno costante, i consigli saggi, le discussioni stimolanti e il suo grande ottimismo, che sono stati chiave per concludere il mio percorso di dottorato.

Un sentito ringraziamento va a tutte le persone eccezionali che hanno fatto parte del nostro gruppo di ricerca BIT nel corso degli anni. Vorrei menzionare in particolare Andrés, Eva, Juan, Juanjo, David, Ana, Daniel, Pedro, David. Per i momenti preziosi trascorsi insieme, sia dentro che fuori dal laboratorio.

Desidero anche ringraziare Raul che ha reso possibile la mia esperienza a Boston. A Gonzalo, per il supporto e l'aiuto che mi ha dato a Boston e nell'ultimo periodo del dottorato.

Non posso che esprimere un infinito ringraziamento a Feli, costante della mia vita da quando sono arrivato a Madrid. Il suo affetto, la premura, l'empatia e la pazienza hanno reso questi difficili anni più semplici, arricchiti da ricordi indimenticabili.

Ai miei cari, la mia famiglia, lontana, che mi è mancata ogni giorno. Sempre al mio fianco, sostenendomi in ogni scelta. Alla mia adorata mamma Emilia e alle mie due straordinarie sorelle Filomena e Federica, le donne più forti che conosca. E al mio papà, che vive nel mio cuore, la cui presenza è sempre viva accanto a me.

Abstract

Lung cancer remains a significant global health challenge, being the leading cause of cancer-related deaths worldwide. Despite a decline in incidence due to decreased smoking rates, the burden of lung cancer persists, particularly in regions with ongoing tobacco consumption. Early detection remains critical, as prognosis varies greatly depending on disease stage, with localized disease showing significantly higher survival rates compared to metastatic cases. Immunotherapy has revolutionized cancer treatment, becoming the new standard for locally advanced and metastatic non-small cell lung cancer (NSCLC) patients. However, its efficacy is limited to a subset of patients (20-30%), with potential immune-related adverse effects.

Computed Tomography (CT) imaging plays a crucial role in lung cancer management, offering valuable information for disease prognosis, treatment planning, and intervention. Quantitative assessment of CT scans is gaining attention for its ability to gauge lung pathologies and correlate with tumor tissue phenotypes. It provides advantages such as low invasiveness, cost-effectiveness, and reduction of radiologist subjectivity and variability. Additionally, analyzing images at different time points may reveal temporal patterns, offering potential assistance in achieving more accurate assessment.

In this Ph.D. thesis, we propose a method for predicting treatment response in advanced lung cancer patients undergoing immunotherapy based on CT imaging and clinical information. Utilizing longitudinal data acquired during early treatment significantly improves prediction performance over baseline data models, underscoring the importance of incorporating patient information over time. Integration of imaging and clinical data provides a better understanding of treatment response, reflecting the close relationship between response and patient clinical condition. Moreover, we study the effects of image and feature harmonization on the radiomics features in order to enhance radiomics robustness against confounding factors associated with CT acquisition parameters. Finally, we propose a novel spatio-temporal deep learning network and methods for handling missing data, applied for predicting indeterminate lung nodule malignancy.

The results demonstrate the potential of the proposed methodologies for clinical use, effectively leveraging spatio-temporal information to improve treatment response prediction and to identify potential lung tumors.

Resumen

El cáncer de pulmón sigue siendo un importante desafío de salud global, siendo la principal causa de muerte relacionada con el cáncer en todo el mundo. A pesar de una disminución en la incidencia debido a la reducción de las tasas de tabaquismo, la carga del cáncer de pulmón persiste, especialmente en regiones en las que persisete el consumo de tabaco. La detección temprana sigue siendo fundamental, ya que el pronóstico varía considerablemente según el estadio de la enfermedad, con tasas de supervivencia significativamente más altas en casos de enfermedad localizada en comparación con casos metastásicos. La inmunoterapia ha revolucionado el tratamiento del cáncer, convirtiéndose en el nuevo estándar para pacientes con cáncer de pulmón no microcítico localmente avanzado y metastásico. Sin embargo, su eficacia se limita a un subconjunto de pacientes (20-30%), con posibles efectos adversos relacionados con el sistema inmunológico.

La Tomografía Computarizada (TC) juega un papel crucial en el manejo del cáncer de pulmón, ofreciendo información valiosa para el pronóstico de la enfermedad, la planificación del tratamiento y la intervención. La evaluación cuantitativa de las imágenes de TC está ganando atención por su capacidad para evaluar patologías pulmonares y correlacionarlas con fenotipos de tejidos tumorales. Ofrece ventajas como baja invasividad, rentabilidad y reducción de la subjetividad y variabilidad del radiólogo. Además, el análisis de las imágenes en diferentes instantes temporales puede permitir la identificación de patrones temporales, lo que podría ayudar en una valoración más precisa.

En esta Tesis Doctoral se propone un método para predecir la respuesta al tratamiento en pacientes con cáncer de pulmón avanzado tratados con inmunoterapia utilizando imágenes de TC y datos clínicos. La utilización de datos longitudinales adquiridos durante el inicio de la terapia mejora significativamente la predicción en comparación con los modelos basados en datos basales, subrayando la importancia de incorporar información del paciente a lo largo del tiempo. La integración de datos de imagen y clínicos proporciona una mejor comprensión de la respuesta al tratamiento, reflejando la estrecha relación entre la respuesta y la condición clínica del paciente. Además, estudiamos los efectos de la armonización de la imagen y de las características de radiómica para mejorar la robustez de las técnicas de radiómica frente a factores de confusión asociados con los parámetros de adquisición de TC. Finalmente, proponemos una nueva red de aprendizaje profundo espacio-temporal y métodos para procesar los datos que faltan, aplicados para predecir la malignidad de nódulos pulmonares indeterminados.

Los resultados demuestran el potencial de las metodologías propuestas para uso clínico, aprovechando eficazmente la información espacio-temporal para mejorar la predicción de la respuesta al tratamiento e identificar posibles tumores pulmonares.

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

CAD Computer-aided Detection and Diagnosis

CNN Convolutional Neural Network

COPD Chronic Obstructive Pulmonary Disease

CRNN Convolutional Recurrent Neural Network

CT Computed Tomography

CUN Clínica Universidad de Navarra

FJD Hospital Universitario Fundación Jiménez Díaz

H12O Hospital Universitario 12 de Octubre

HU Hounsfield Units

IBSI Image Biomarker Standardization Initiative

LDCT Low-dose Computed Tomography

LIDC-IDRI Lung Image Database Consortium and Image Database Resource Initiation Data Set

MRI Magnetic Resonance Imaging

NLST National Lung Screening Trial

NSCLC Non-small Cell Lung Cancer

OS Overall Survival

PD-1 Programmed Death-1

PD-L1 Programmed Death Ligand-1

PET Positron Emission Tomography

PFS Progression-free Survival

RIDER Reference Image Database to Evaluate Therapy Response

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

CHAPTER 1

Motivation and Objectives

1.1 Motivation

Lung cancer represents a significant global health challenge, being the leading cause of cancer-related deaths worldwide. Despite a declining trend in incidence due to decreased smoking rates, lung cancer's burden persists, particularly in regions with ongoing tobacco consumption. Various factors, including active and passive smoking, occupational exposures to carcinogens, genetic predisposition, and environmental pollutants, contribute to the complex etiology of lung cancer, underscoring the need for comprehensive prevention and management strategies (Collins et al., 2007; Spyratos et al., 2013; Stojšić, 2018).

Lung cancer, comprising non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), originates primarily in the respiratory tract. Despite improvements in survival rates, early detection remains critical, as prognosis varies greatly depending on disease stage, with localized disease showing significantly higher survival rates compared to metastatic cases. Immunotherapy has radically changed the therapeutic paradigm in cancer, becoming the new standard for treating locally advanced and metastatic NSCLC patients (Gridelli et al., 2022). While immunotherapy demonstrates promising outcomes in terms of long-term survival, its efficacy is limited to a subset of patients with NSCLC (20-30%), with potential immune-related adverse effects (Berghmans et al., 2020; Blons et al., 2019; Kanwal et al., 2018; Punekar et al., 2022).

In this scenario, there's a pressing need to enhance early lung cancer detection through screening initiatives and optimize management strategies for late-stage disease to improve patient outcomes.

Diagnosis and prognosis assessment of lung cancer typically rely on a combination of imaging modalities, laboratory analysis, and histopathological examination. While chest X-rays are commonly used for initial screening, low-dose computed tomography (LDCT) scans have emerged as a superior option. LDCT scans offer greater accuracy, reduced mortality rates, and improved survival outcomes in disease diagnosis (de Koning et al., 2020; Team, 2011b). On the other hand, high-dose CT scans are extensively employed in modern healthcare, providing clinicians with valuable information

for disease prognosis, treatment planning, and intervention. High-dose CT scans are particularly beneficial in situations where detailed anatomical information and precise characterization of lung pathologies are required. The quantitative assessment of both low- and high-dose CT scans is increasingly garnering attention for its ability to evaluate the extent of lung pathologies and correlate with tumor tissue phenotypes (Aerts et al., 2014; Sun et al., 2018).

The objective quantification of lung diseases via CT imaging holds immense promise, offering advantages such as low invasiveness, cost-effectiveness, mitigation of radiologist subjectivity and inter- and intra-observer variability. Additionally, CT interpretation is often time-consuming, insensitive to subtle changes, and demands a high level of radiological expertise. Furthermore, quantitative analysis has the potential to discern temporal patterns when multiple images of the same patient are available over time, potentially improving the accuracy and consistency of medical imaging evaluation (Cousin et al., 2023; Hanaoka et al., 2024; Lee et al., 2017).

In recent research, artificial intelligence has gained popularity for analyzing lung lesions, with both radiomics and deep learning approaches emerging as effective tools for facilitating objective diagnosis, predicting disease progression, and anticipating treatment response to conventional and novel therapies (Dercle et al., 2022; Ghaffari Laleh et al., 2023).

1.2 Objectives and Original Contributions

This PhD Thesis focuses on the study and implementation of new techniques for lung lesion characterization, leveraging radiomics and deep learning methodologies for the quantitative analysis of CT scans. **The main objective of this thesis is to develop, implement, and validate spatio-temporal methods to facilitate automated diagnosis of lung cancer and to improve the prediction of tumor response to immunotherapy.** Additionally, the study aims to contribute to the understanding of lung cancer and its prognosis.

The proposed approach relies on two main datasets with longitudinal data: the National Lung Screening Trial (NLST) dataset for lung cancer screening over three years of screening program and a retrospective dataset comprising CT images from patients with locally advanced NSCLC acquired at baseline and during early immunotherapy treatment collected from three different institutions. The research has primarily focus on implementing spatio-temporal algorithms to predict malignancy in nodules from the screening dataset and forecast treatment response in the in-house dataset using both radiomics and deep learning techniques. Additionally, the study has explore the reproducibility of radiomics features, which are known to be highly sensitive to technical variations such as image acquisition parameters. Ensuring their high reproducibility against confounding factors is crucial to establish their reliability and gain trust within the medical community.

Achieving the objectives of this work could represent a significant advancement toward a non-invasive method for characterizing lung lesions. Such an approach has the potential to reduce unnecessary interventions, avoid ineffective or harmful treatments, and ultimately improve patient outcomes in lung cancer screening and treatment.

The objectives outlined above can be divided into the following sub-objectives:

1. Design, implementation and validation of spatio-temporal methods for the prediction of treatment response in lung cancer patients in advanced stage treated with immunotherapy. Immunotherapy is a novel treatment that has shown remarkable advancement in the treatment of advanced metastatic patients enhancing survival but with the drawback of only a 20-30% of response and possible immune-related adverse events, which can be life threatening. This sub-objective aims to develop methods to predict treatment response exploiting the spatial information of CT scans and the temporal information from follow-up CT during early treatment. The combination with clinical information has been also studied.
2. Design, implementation and validation of harmonization methods to improve the quantification of lung lesions through radiomics features extracted from CT scans. This sub-objective seeks to address the integration of radiomics frameworks into clinical routine limited by the sensitivity to technical variations, which significantly impact their reproducibility. This sub-objective aims to mitigate biases introduced by technical parameters, thereby enhancing reproducibility, temporal consistency and predictive performance.
3. Design, implementation and validation of spatio-temporal deep-learning methods for predicting the malignancy of nodules detected during lung cancer screening programs. This sub-objective aims to explore the potential of analyzing repeated annual exams of indeterminate lung nodules to enhance screening accuracy. The goal is to develop automated methods for identifying and classifying malignant nodules using longitudinal data from screening trials.

CHAPTER 2

Clinical Context and State of the Art

2.1 Clinical Context

2.1.1 Epidemiology

Lung cancer is the second leading cause of death after cardiovascular diseases and the leading cause of cancer-related death worldwide in both men and women (Siegel et al., 2020). With an estimated 1,435,943 cases, lung cancer is the most common cancer in men (14.3%), and the third most common cancer in women, after breast and colorectum cancer, with 770,828 cases (8.4%) every year. In the last years, the incidence of lung cancer has exhibited a consistent and steady decline, with an annual decrease of 2.6% among men and 1.1% among women (Siegel et al., 2023). The decline in lung cancer incidence and mortality is largely due to a rapid and constant decline of smoking prevalence since 1990: globally, the smoking prevalence has declined by 27.2% for men and by 37.9% for women, respectively (Dai et al., 2022). However, despite the progress made in reducing lung cancer rates in certain regions, it is important to recognize that the tobacco smoking epidemic continues to unfold in parts of Asia and Africa, with multiple countries still experiencing an increase in cases.

Tobacco smoking, both current and historical, stands as the primary determinant of lung cancer burden: around 80% lung cancer deaths are caused by smoking. In addition to direct smoking, exposure to second-hand smoke has also been identified as a significant risk factor for developing lung cancer in non-smokers (Steliga & Dresler, 2011). Inhalation of the smoke emitted by others who are actively smoking increases the risk of lung cancer, further emphasizing the importance of comprehensive tobacco control measures (Dyba et al., 2021). Certain occupational exposures to carcinogens have posed significant risks to various worker groups, and some of these occupational agents (such as asbestos and radon) have shown synergistic effects with smoking, further increasing the likelihood of developing lung cancer (Spyratos et al., 2013). Moreover, observations of familial aggregation of lung cancer have suggested the involvement of genetic factors in determining risks among smokers, although the precise genes involved are still subject to active investigation.

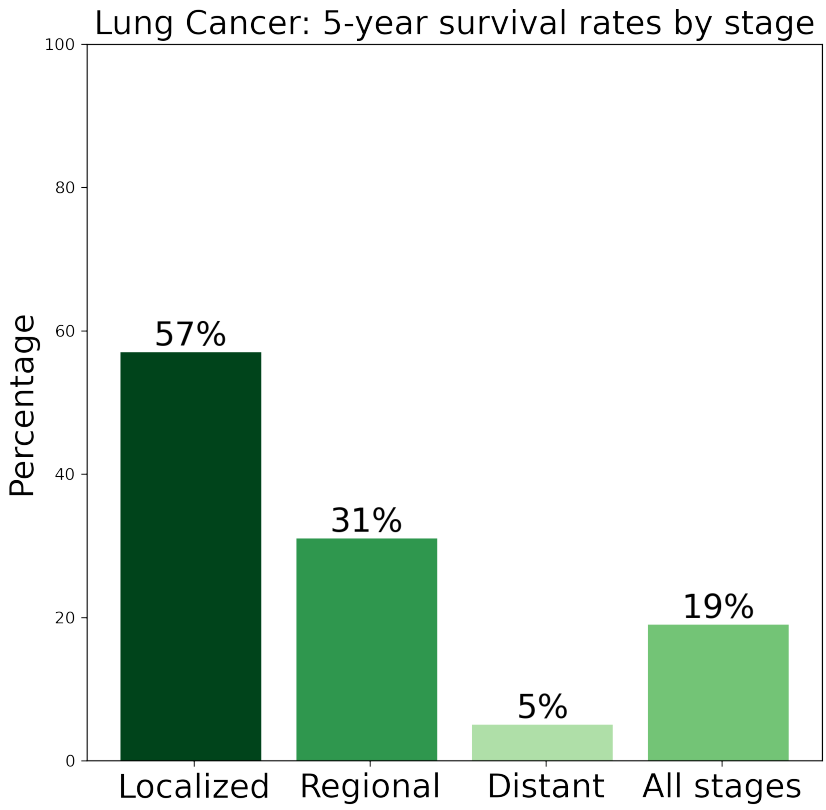


Figure 2.1: Five-year survival rates by stage in the US from 2009 to 2015. Source data from (Siegel et al., 2020).

Additionally, there is evidence that long-term and cumulative exposure to environmental pollutants contribute to an overall increase in lung cancer risks, especially in those patients with high genetic risk (Y. Huang et al., 2021). Previous non-cancer-related pulmonary diseases, for instance, chronic obstructive pulmonary disease and tuberculosis, have also been linked to lung cancer.

Lung cancer encompasses a wide range of histologically and clinically diverse malignancies that originates within the respiratory tract, primarily but not exclusively in the cells lining in the lung airways. The World Health Organization (WHO) classifies lung tumors into two main categories in relation to the nature of the cells where it originates: non-small cell lung cancer (NSCLC), the most common form, which accounts for approximately 80-85% of all lung cancer cases, and small cell lung cancer (SCLC) representing the remaining 15-20% of cases (Collins et al., 2007). NSCLC typically grows and spreads more slowly than SCLC, which usually has cells that are smaller in size than most other cancer cells and tend to grow very fast and spread rapidly to other parts of the body. NSCLC can be further subdivided into distinct histological sub-types, each one with its variants, specific immunophenotype and biological behavior, including adenocarcinoma (LUAD), squamous cell carcinoma (LUSC), and large cell carcinoma (LCC).

The assessment of NSCLC severity is carried out by doctors through a comprehensive staging process, which aims to determine the extent of cancer spread within the lungs and its potential involvement

in other areas of the body. Staging involves various diagnostic tests, such as imaging scans (e.g., CT, PET), biopsies, and sometimes surgical exploration, to accurately evaluate tumor size, location, and potential metastasis. The staging system commonly used for NSCLC is the TNM system, which takes into account the characteristics of the primary tumor (T), the involvement of nearby lymph nodes (N), and the presence of distant metastasis (M). The resulting stage classification provides valuable information for treatment planning, prognosis estimation, and determining appropriate therapeutic strategies for patients with NSCLC. At the time of the diagnosis, about 75% of lung cancers are inoperable, meaning that surgical treatment is not a viable option. In such cases, the identification of the histological type and stage of NSCLC becomes crucial as it serves as a critical factor in determining the most suitable personalized therapy approach (Stojšić, 2018).

Lung cancer survival rates vary depending on several factors, including the stage of cancer at diagnosis, the specific type of lung cancer, and individual patient characteristics. Unfortunately, lung cancer is often diagnosed at an advanced stage when treatment options are limited. Although the overall 5-year survival rate for lung cancer has improved since the mid-1970s, it remains relatively low, ranging from approximately 10% to 20% in most countries (Allemani et al., 2018). In Europe, the 5-year survival rate is of 13% (11.2% in males, 13.9% in females) (De Angelis et al., 2014). However, it's important to note that this bad prognosis is linked to its late diagnosis and it can vary significantly between different stages of the disease. For localized lung cancer that is confined to the lungs, the 5-year survival rate is higher than for cases where the cancer has spread to distant organs or lymph nodes, where it drops from 57% to about 5%. In the US, the 5-year survival for those with localized disease at diagnosis (stage I-II) is 57%, decreasing to 31% among those with regional disease (stage III) and 5% among those with metastatic (stage IV) disease (Siegel et al., 2020) (Figure 2.1).

Considering the significantly low 5-year survival rates observed in patients with advanced-stage lung cancer, and the unfortunate reality that a substantial number of lung cancer cases are diagnosed at late stages, there is an urgent need for effective strategies to promote early detection of lung cancer, including better risk assessment and screening campaigns, and to cure or manage late-stage lung cancer.

2.1.2 Lung Cancer Diagnosis

Various imaging modalities and approaches are employed to achieve accurate diagnosis and to inform treatment decisions improving patient outcome. Imaging techniques, including computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI), play pivotal roles in lung cancer diagnosis. CT scans offer detailed anatomical information, facilitating the detection and characterization of lung nodules as well as assessment of pathologies at the organ level.

PET-CT scans provide functional data by assessing metabolic activity, aiding in lesion characterization and staging. MRI complements CT scans by evaluating tumor invasion into surrounding tissues, enhancing diagnostic accuracy.

Biomarker testing is integral to lung cancer diagnosis, guiding treatment selection and prognostication. Tumor genetic testing identifies actionable mutations in genes such as EGFR, ALK,

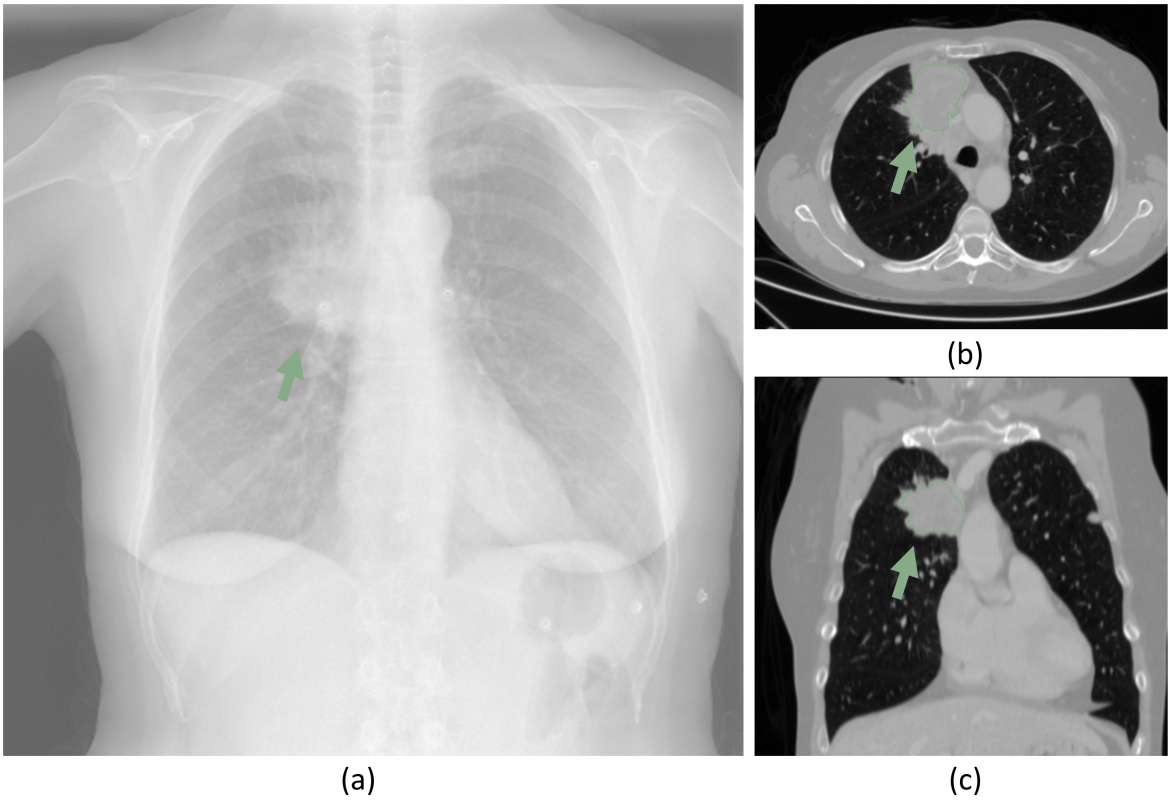


Figure 2.2: Example of an X-ray image (a) alongside axial (b) and sagittal (c) views of a CT scan from the same patient. The tumor is indicated with a green arrow and delineated and highlighted in green in the CT image. This illustration was generated in-house using data from our cohort.

ROS1, and BRAF, informing targeted therapy decisions. Protein biomarkers, such as programmed death-1 (PD-1) or programmed death ligand 1 (PD-L1) expression levels, predict response to immunotherapy and inform treatment strategies.

Histopathological examination remains indispensable for confirming lung cancer diagnosis and guiding treatment decisions. Tissue samples obtained via biopsy procedures, including bronchoscopy, needle biopsy, and surgical resection, undergo histopathological evaluation to determine tumor type, grade, and stage.

A multidisciplinary approach is essential in lung cancer diagnosis and personalized treatment, involving collaboration among pulmonologists, radiologists, oncologists, pathologists, and thoracic surgeons. This collaborative effort ensures comprehensive evaluation, accurate diagnosis, and personalized therapy management for each patient.

2.1.3 Lung Cancer Screening

In recent decades, chest radiography and sputum cytology have been the primary approaches used to reduce lung cancer-related deaths through screening. These methods have successfully detected

early-stage cancers that were resectable in the majority of cases, but their use in early detection has not demonstrated a reduction in mortality rates (Nanavaty et al., 2014). Clinical trials focusing on lung cancer screening, such as low-dose computed tomography (LDCT) scans for high-risk individuals, have shown promise in identifying lung cancer at earlier, more treatable stages. In Figure 2.2, examples of the two main modalities currently used, X-ray and LDCT, are illustrated.

The National Lung Screening Trial (NLST), launched in 2002 by the National Cancer Institute, demonstrated a remarkable 20% reduction in mortality rates among high-risk individuals (former/current smokers with a $\geq$ 30 pack-year history) being screened with LDCT scans compared to chest radiography (Team, 2011a). Moreover, the lung cancer detection rate was 13% higher for patients screened with LDCT compared to those screened with chest radiography (Team, 2011b). A more recent study, the Multicentric Italian Lung Detection (MILD) trial, which included current or former smokers with a $\geq$ 20 pack-years, demonstrated a 39% reduction in lung cancer mortality compared to no intervention (Pastorino et al., 2019).

Despite these results, several concerns persist regarding the establishment of lung cancer screening using LDCT as a standard of care. One concern is the high false-positive rates associated with LDCT screening, which can lead to unnecessary follow-up tests and potential patient anxiety (Hammer et al., 2022). Additionally, there is a risk of adverse events resulting from over-diagnosis, where indolent or slow-growing cancers are detected and treated, leading to potential harm without improving overall outcomes (Patz et al., 2014). Another concern is the significant percentage of lung tumors that are still detected at advanced stages despite screening efforts, indicating the need for more effective early detection strategies involving new risk biomarkers (Ruano-Ravina et al., 2015). Finally, there is a potential risk of radiation-induced neoplasms associated with repeated LDCT scans, emphasizing the importance of carefully balancing the benefits and risks of screening in terms of radiation exposure (Brenner, 2004).

However, the prognosis of NSCLC poses significant challenges as early diagnosis through screening has been proven to be challenging in terms on identifying eligible patients and training physicians on decision-making (Kale et al., 2024; Martini et al., 2021).

2.1.4 Traditional Treatments and Immunotherapy

Lung cancer treatment depends on several factors, including lung cancer type and stage, size and position of primary tumor and overall health of the patient. Traditional treatments include surgical resection, chemotherapy and radiation therapy. Surgical resection represent the best treatment for patients with early stage disease and some patients with locally advanced disease. The majority of patients have metastatic disease at the time of the diagnosis and they are prone to relapse even after surgery. Chemotherapy and radiotherapy are recommended for locally advanced and metastatic cancers. Platinum-based chemotherapy is the standard-of-care first-line treatment for advanced NSCLC, but it has shown poor response rate of 15-30% and its impact on overall survival has plateaued in recent years (Schiller et al., 2002). The inability of chemotherapy and radiotherapy to eliminate tumor cells is partly due to the necessity of administrating low doses, due to their irreparable damages to the normal tissues and their toxic side effects (Albano et al., 2021; H. Chen et al., 2022).

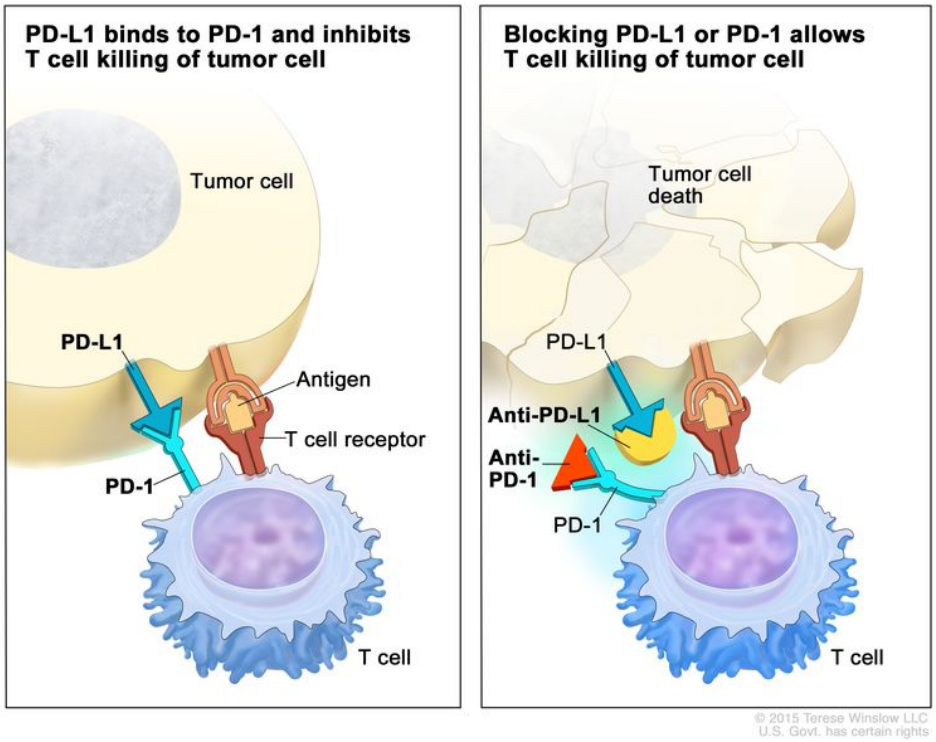


Figure 2.3: Illustration of immune checkpoint inhibition. Left panel: the interaction between PD-L1 on tumor cells and PD-1 on T cells inhibits the T cells from attacking tumor cells in the body. Right panel: blocking this interaction with an immune checkpoint inhibitor (such as anti-PD-L1 or anti-PD-1) enables T cells to effectively kill tumor cells. Source: National Cancer Institute, adapted from https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq

However, there is optimism in the field of lung cancer management with the emergence of novel treatments, like immunotherapy and targeted therapies. The advent of molecularly targeted therapies has changed the management of metastatic NSCLC among patients possessing specific genetic alterations, such as EGFR mutations and EML4-ALK translocations. Targeted therapies act directly over specific protein mutations within oncogenic proteins and have demonstrated remarkable efficacy in suppressing tumor growth and improving patient outcomes (Korpanty et al., 2014). However, despite these significant advances, a notable proportion of patients still face challenges in achieving durable and stable disease control, resulting in persistently low survival rates. Furthermore, another critical issue associated with these therapies is the development of acquired resistance, which poses a substantial clinical obstacle (Sequist et al., 2011).

More recently, immune checkpoint inhibition therapy targeting PD-1 or PD-L1 pathways has emerged as a groundbreaking approach for lung cancer management. PD-1, a negative regulatory receptor found on the surface of T cells, inhibits the cytotoxic T-cell response when it interacts with its ligands PD-L1 or PD-L2 on tumor cells. The main aim of these innovative drugs is to disrupt the interaction between PD-1 and the tumor cell ligands, thereby boosting T-cell anti-tumor activity. As shown in Figure 2.3, immune checkpoint inhibitors target proteins like PD-L1 on tumor cells and PD-1 on T cells, which normally help regulate immune responses. By leveraging the body's immune

system to combat cancer cells, immunotherapy has demonstrated promising outcomes, including enhanced survival rates and durable responses, surpassing the efficacy of conventional treatments, like chemotherapy or radiation therapy, either alone or in combination with these treatments (Kanwal et al., 2018). Initially approved as second-line treatments in 2015, single-agent PD-1 pathway inhibitors have shown improved overall survival compared to chemotherapy for patients with advanced squamous cell carcinoma lacking EGFR and ALK mutations (Borghaei et al., 2015). Furthermore, recent studies have demonstrated enhanced progression-free survival and overall survival when metastatic NSCLC patients received a combination of standard chemotherapy and immunotherapy (Gandhi et al., 2018; Paz-Ares et al., 2018; X. Wang et al., 2021). Immunotherapy has also made significant strides in neoadjuvant or adjuvant settings, being introduced as a treatment option for resectable tumors at earlier stages. Despite the promise of immunotherapy, not all patients exhibit a positive response, and there can be potential serious side effects (Mantia & Buchbinder, 2019). However, approximately 20% of patients experience substantial and often durable responses, presenting a remarkable opportunity for treatment (Kanwal et al., 2018). Exploring the mechanisms underlying the antitumor immune response and resistance will enable identifying patients who can benefit most from immunotherapy, striving for optimal efficacy with minimal toxicity.

2.1.5 Quantitative Imaging

Medical imaging employs various techniques, including CT, MRI, ultrasound, and digital pathology, to visualize internal and external tissues, aiding diagnosis, treatment planning, and intervention. While these technologies offer valuable insights, the increasing volume of medical images presents challenges to the healthcare system. With billions of imaging studies conducted annually worldwide, the interpretation of such vast datasets by human experts, notably radiologists, is subject to limitations like human subjectivity and fatigue (Schöckel et al., 2020). Moreover, the reliance on traditional protocols, like the response evaluation criteria in solid tumors (RECIST) criteria, which focus solely on tumor diameter, may not capture all lesion characteristics (Eisenhauer et al., 2009).

Computer-based methods for quantitative analysis of medical images address these challenges by providing rapid and objective measurements of abnormal lesions, facilitating early detection and accurate tracking of changes over time. Image-based biomarkers offer diagnostic and prognostic information beyond lesion shape, predicting treatment response and guiding personalized therapy. This approach reduces reliance on subjective assessments and enhances efficiency in diagnosis and treatment planning.

Understanding the appearance and characteristics of lung nodules and tumors is essential for accurate diagnosis and treatment planning. Lung nodules can vary in size, shape, density, and location within the lungs, and their appearance on imaging modalities like CT scans can provide crucial information about their likelihood of malignancy (Vlahos et al., 2018). Typically, lung nodules fall into one of four categories: solid nodules, the most common type, characterized by homogeneous soft-tissue attenuation; ground-glass nodules, which lack uniform appearance and do not obscure underlying bronchial structures or pulmonary vessels; part-solid nodules, featuring both solid and ground glass characteristics; and heterogeneous nodules with a complex distribution of solid-like portions (Figure 2.4). The presence of fat or eccentric calcification within the lesion

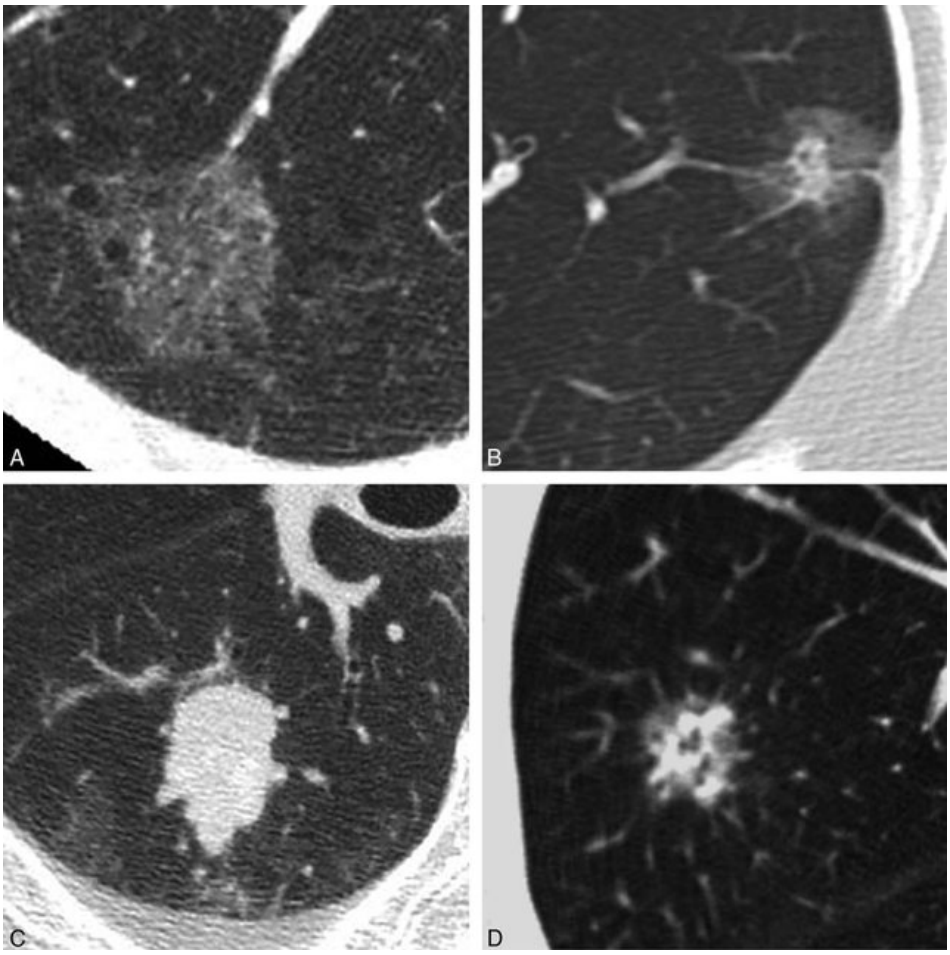


Figure 2.4: A visual representation of different types of lung nodules on CT scans: A) Ground-glass nodule, (B) part-solid nodule, (C) solid nodule, (D) heterogeneous nodule. Source data from Yanagawa et al. (2017)

typically indicates a benign behavior. Conversely, malignant nodules tend to exhibit rapid growth, doubling in size much faster than benign nodules. Additionally, ground glass nodules may develop a solid component over time, indicating malignant transformation.

Similarly, tumors may present diverse characteristics on imaging, such as enhancement patterns on contrast-enhanced CT or MRI scans, which can help determine their histological subtype and guide treatment decisions. Figure 2.5 illustrates typical radiographic manifestations observed in cases of lung cancer.

Quantitative imaging enables a comprehensive evaluation of tumor characteristics, providing a holistic view beyond tumor size. While many current biomarkers rely on tumor diameter, they may not adequately capture the complexity of lung lesions. For instance, standards for nodule assessment and treatment response primarily focus on tumor size evolution. However, image-based characteristics show promise in providing information about tumor phenotype and its microenvironment (Gillies et al., 2016). These features offer easily obtainable, non-invasive,

Figure 2.5: Typical manifestations of lung cancer in CT scans. A) A centrally located mass with mediastinal invasion; B) a peripherally situated mass abutting the pleura; C) mass characterized by smooth, lobulated margins; D) a mass exhibiting spiculated, irregular margins. Source data from Purandare and Rangarajan (2015)

cost-effective biomarkers that, when combined with clinical or genomic data, can be correlated with clinical outcomes, providing a powerful tool for evidence-based clinical decision support.

2.2 State-of-the-art in Lung Cancer Outcome Prediction

While medical imaging is crucial for disease detection and diagnosis due to its low invasiveness, cost-effectiveness, and information richness, qualitative interpretation often leads to non-reproducible and highly subjective results. To address this, quantitative imaging analysis has gained traction, enabling clinicians to adopt simple metrics for assessing medical imaging. Nevertheless, metrics like the RECIST criteria (Eisenhauer et al., 2009) and its immunotherapy update iRECIST (Seymour et al., 2017), utilized for determining optimal treatment strategies, along with lungRADs (Chelala

et al., 2021) and tumor doubling time (Usuda et al., 1994) for lung cancer screening, may not fully utilize the rich information present in images and may overlook intricate pixel and voxel relationships.

Machine learning algorithms offer a promising solution by effectively analyzing medical images, extracting relevant features, and supporting the detection, characterization, and quantification of abnormalities. Through extensive image datasets, machine learning models identify intricate patterns, learn from examples, and provide accurate predictions. This has led to the development of computer-aided detection and diagnosis (CAD) systems and clinical decision support systems (CDSS), assisting clinicians in interpreting medical imaging and making accurate diagnostic decisions.

Machine learning methods leverage spatial and temporal image features that are often imperceptible to the human eye, thereby offering deeper insight into patient-specific characteristics and enhancing treatment response prediction. Additionally, they provide valuable insights into therapy mechanisms and potential resistance factors.

2.2.1 Radiomics in Lung Cancer

The vast amount of available medical imaging data have opened a new era where numerous algorithms have been developed to extract information from these images. Radiomics involves the extraction of a vast amount of data, providing an enhanced characterization of properties within the heterogeneous tumor region, which has been shown to correlate with tumor biology and clinical outcomes (Gillies et al., 2016; Lambin et al., 2017).

Radiomics comprises two primary approaches for feature extraction from images: hand-crafted radiomics (traditional radiomics) and deep-radiomics. Hand-crafted radiomics involves extracting features using formulas based on intensity histograms, shape characteristics, and texture, which can describe the relationship between image pixel in a meaningful way. Conversely, deep radiomics features are extracted using deep learning techniques. These are typically extracted through supervised methods, which require outcome information, making these features specific to particular outcomes, image types, and population characteristics (Ghaffari Laleh et al., 2023). Alternatively, unsupervised or self-supervised methods can produce more general features that are not dependent on outcomes (Z. Li et al., 2023).

Recent studies have highlighted the utility of radiomics in lung cancer screening for early lung cancer detection (Choi et al., 2018; Gao et al., 2020), malignancy prediction, and staging. For instance, Hawkins et al. (2016) proposed a Random Forest model with 23 features capable of predicting nodules that would progress to cancer within 1 or 2 years . Similarly, Liu et al. (2017) demonstrated the effectiveness of radiomics in predicting lung nodule malignancy. B. T. Chen et al. (2020) used radiomics to differentiate between NSCLC and peripherally located SCLC.

Radiomics signatures have proven valuable also in predicting patient response to therapy. Sun et al. (2018) developed a radiomics-based biomarker to predict tumor-infiltrating CD8 cell expression, yielding promising results in inferring clinical outcomes across multiple cohorts. Their CT-based radiomics signature was specifically designed to predict the tumor immune environment and

distinguish between tumor-inflamed and tumor-desert phenotypes. It was found to be correlated with CD8 infiltration, with higher values associated with a higher likelihood of achieving response at 3 and 6 months after treatment. Tang et al. (2018) combined radiomics features with other clinical biomarkers, such as tumor PD-L1 expression and tumor-infiltrating lymphocyte density, to develop a signature that significantly associated with 5-year overall survival. Trebeschi et al. (2019) developed a biomarker to distinguish between immunotherapy responding and non-responding patients in NSCLC and advanced melanoma. Nardone et al. (2020) predicted survival in NSCLC patients undergoing PD-1/PD-L1 blockade treatment using a combination of radiomics features extracted from pre- and post-contrast CT scans. Tunali et al. (2019) aimed to build a predictor of rapid disease progression phenotypes in NSCLC patients treated with immune-checkpoint inhibitors, incorporating demographics, clinical, driver mutations, hematology, and radiomics data.

Deep radiomics methodologies often require a large dataset for model training and may lack interpretability compared to traditional radiomics techniques. Furthermore, traditional radiomics methods rely on features selected by domain experts, ensuring that the algorithm learns in the correct direction. In contrast, deep radiomics approaches have the advantage that with proper training, it may achieve higher performance, and the model can continue learning as more data becomes available.

Despite promising results reported in the literature, the integration of radiomics frameworks into clinical practice remains limited. This limitation arises from the sensitivity of radiomics-based features to technical variations, which significantly impact their reproducibility (B. Zhao et al., 2016). The establishment of reliable clinical biomarker depends on its reproducibility (Califf, 2018). The robustness of radiomics features is influenced by inherent variability across the radiomics workflow, including image acquisition, reconstruction and post-processing, volume segmentation, and feature calculation. Distinguishing between quantitative changes in radiomics attributed to biological variations and those arising from the heterogeneity of image parameters, often termed as batch effects, poses a significant challenge to the radiomics community.

Multiple deep learning approaches have been proposed to characterize lung nodules and predict treatment response. To assist the screening process,Paul et al. (2018) proposed an ensemble of pre-trained convolutional neural network (CNN) and radiomics features to predict nodule malignancy. Causey et al. (2018) employed a similar strategy, introducing a CNN architecture named NoduleX, which achieved high accuracy for nodule malignancy classification integrating traditional and deep radiomics features. Zhu et al. (2018) introduced DeepLung, a fully automated lung cancer diagnosis system for CT images, which incorporated a specific component for automatic nodule detection alongside a nodule classification component. Ciompi et al. (2017) proposed a multi-scale with a multi-stream CNN for lung nodule type prediction simultaneously processing multiple slices of the same nodule at multiple scales. Y.-S. Huang et al. (2023) implemented a 3D-CNN model with an attention scheme based on the squeeze-and-excitation module block to focus on important features, integrating nodules features at different scale. Mikhael et al. (2023) designed a 3D-CNN to predict future lung cancer risk up to 6 years after the CT scan, utilizing both local nodule details and global CT information.

Additionally, Tian et al. (2021) addressed the challenge of predicting treatment response by proposing a deep convolutional neural network. This network aimed to derive a PD-L1 expression

signature, which was found capable of predicting high PD-L1 expression levels. High PD-L1 expression has been associated with the efficacy of immunotherapy. Yang et al. (2021) proposed a deep learning network integrating multi-modal data sources (imaging, laboratory, and baseline clinical information) to predict response to anti-PD-1/PD-L1 treatment, which also demonstrated the ability to predict survival benefit from therapy in certain patients with stable disease. Similarly, X. Wang et al. (2023) adopted a multi-modal approach, combining clinical and CT image data in gastric cancer patients undergoing immunotherapy. They demonstrated that integrating their model with PD-L1 expression further enhanced the area under the curve (AUC) by 4-7% in absolute terms. Saad et al. (2023) introduced a CT-based deep learning signature, comprised of an ensemble of various deep learning architectures, a supervised learning network, two hybrid networks that merge supervised and unsupervised learning differently, and an unsupervised learning network, capable of predicting clinical benefit from immunotherapy in patients with NSCLC.

2.2.2 Spatio-Temporal Analysis

In clinical practice, radiographic imaging is pivotal for assessing nodule malignancy and treatment response. Typically, this involves measuring changes in nodule or tumor sizes over time. However, such methods may oversimplify the complex biological changes induced by treatments and are often evaluated subjectively by radiologists. Despite efforts to enhance standard practices, reliably evaluating temporal changes in lung lesions remains challenging.

For lung cancer screening, the Lung-RADS method is commonly used to classify lung lesions, providing radiologists with a standardized nodule follow-up protocol based on nodule size, growth, and morphology using 2D measurements. However, 3D measurements and texture changes over time in lung lesions may offer a more accurate estimate of nodule malignancy (Chelala et al., 2021). Similarly, in predicting lung treatment response, RECIST guidelines are commonly employed, focusing solely on changes in tumor size. However, considering tumor heterogeneity may provide deeper insights into tumor biology and evolution during therapy. Additionally, relying solely on RECIST endpoints, such as overall response rate, may not correlate with overall survival and should not solely determine treatment duration or cessation (Mushti et al., 2018; Nie et al., 2019). Furthermore, the unusual response patterns, like pseudoprogression, in immunotherapy make this problem even more challenging.

Despite the medical importance of monitoring the evolution of pulmonary lesions for determining malignancy likelihood or treatment response, few works have really taken into account the temporal dimension to provide better estimates. Several studies have investigated the use of radiomics to quantify changes in lung lesions over time and assess heterogeneity.

One such approach is delta-radiomics, which involves extracting radiomics features from the same region of interest at different time intervals to study variations over time. Compared to standard radiomics, which reflect a static situation, delta-radiomics biomarkers study phenotypical modifications of a tissue or lesion that may occur after treatment introduction. This approach may allow the detection of early signs of tumor response before any size modification. Delta-radiomics features have been used by Khorrami et al. (2020) to recognize patterns predicting immunotherapy response and overall survival in NSCLC patients, achieving promising results with an AUC of 0.81

and 0.85 in different validation sets.

Similarly, Gong et al. (2022) proposed a delta-radiomics signature that significantly improved response prediction in two validation cohorts compared to models based on RECIST measurements. Delta-radiomics has also been found effective in predicting nodule malignancy, as demonstrated by Alahmari et al. (2018), who showed that integrating delta radiomics features with conventional radiomics improves nodule malignancy classification.

Recently, deep learning approaches have been explored to analyze longitudinal data, including imaging and clinical data. Ardila et al. (2019) proposed a deep learning algorithm that predicted the risk of lung cancer based on previous and current CT images of the patients. The model was compared with the performance of 6 radiologists, showing on-par performance when prior images were available, and outperforming them with a reduction of 11% and 5% in false positives and false negatives rates respectively when previous scans were not available. Veasey et al. (2020) implemented a siamese-style convolutional attention network capable of effectively integrating relevant 2D features from multiple time points. By dynamically weighting the 2D features extracted from 10 slices within the same nodule, they demonstrated that the model was able to prioritize the features from the slices that contained information that was most relevant for accurate predictions. Furthermore, L. Zhang et al. (2019) implemented a spatio-temporal convolutional LSTM (long-short term memory) segmentation model for the prediction of tumor growth from 4D CT images. They demonstrated that the model was able to jointly learn the inter-slice and intra-slice nodule's features along with its temporal dynamic. Xu et al. (2019a) proposed a CRNN (convolutional recurrent neural network) to predict chemotherapy treatment response in lung cancer patients through analyzing time series CT images, demonstrating that the model's performance improved with the inclusion of more time points. Additionally, Yang et al. (2021) proposed a Simple Temporal Attention (SimTA) module for processing asynchronous time-series imaging and laboratory data, showing promising results in evaluating treatment response in patients with lung cancer treated with anti-PD-1/PD-L1 immunotherapy at 3 months. The SimTA module prioritized measurements closer to the assessment time over those further away.

CHAPTER 3

Radiomics in Immunotherapy Treatment Response

Immunotherapy has revolutionized cancer treatment, emerging as the standard approach for managing locally advanced and metastatic non-small cell lung cancer (NSCLC). While numerous studies have demonstrated its efficacy, with notable improvements in long-term survival either as a standalone therapy or in combination with other modalities, only a fraction of patients (20–30%) exhibit a positive response (Berghmans et al., 2020; Punekar et al., 2022). Given the unreliability of current clinical biomarkers, our focus shifts towards exploring a novel non-invasive biomarker for predicting durable clinical benefit from immunotherapy. This involves integrating radiomics and clinical data gathered during early anti-PD-1/PD-L1 monoclonal antibody treatment in patients with advanced NSCLC. The content of this chapter has been published in the Journal of Translational Medicine.[1]

3.1 Introduction

Immunotherapy has radically changed the therapeutic paradigm in cancer, becoming the new standard for treating locally advanced and metastatic NSCLC patients (Gridelli et al., 2022). Many studies have shown positive results in terms of improved long-term survival when used alone or in combination with other treatments (Broderick, 2020; Doroshow et al., 2019; Patel & Weiss, 2020; Paz-Ares et al., 2021), but only a small proportion of patients (20-30%) respond to therapy (Berghmans et al., 2020; Blons et al., 2019; Kanwal et al., 2018; Punekar et al., 2022). Due to immunotherapy's unconventional response pattern, including delayed response or pseudoprogression, traditional approaches to defining response are no longer adequate. Furthermore, patients may

[1]Farina, Benito, et al. "Integration of longitudinal deep-radiomics and clinical data improves the prediction of durable benefits to anti-PD-1/PD-L1 immunotherapy in advanced NSCLC patients." *Journal of Translational Medicine* 21.1 (2023): 174.

experience immune-related adverse events, which can be life threatening (Suresh et al., 2018).

It has become crucial to identify biomarkers that could predict long-term clinical benefit patients to monitor their condition over time effectively. Different biomarkers have been investigated, such as PD-L1 expression and tumor mutational burden, and their association with treatment response has been reported in previous studies with mixed results (Bai et al., 2020; Dong et al., 2021). Furthermore, tumor heterogeneity could influence the reliability of these biomarkers, as they depend on biopsied tissue, which cannot cover the entire tumor microenvironment.

The use of non-invasive image-based biomarkers has gained increased attention during the past few years because of their availability and non-invasiveness. Typically, the effectiveness of treatment has been evaluated using the response evaluation criteria in solid tumors (RECIST) (Eisenhauer et al., 2009) or its adaptation to immunotherapy (iRECIST) (Seymour et al., 2017). However, these criteria are often subjective and do not consider changes in tumor heterogeneity.

Radiomics involves the high-throughput extraction of a large number of quantitative characteristics from medical imaging, which can provide complete information on tumor radiophenotype and microenvironment heterogeneity (Gillies et al., 2016). Several studies have demonstrated the ability of radiomics features to predict the immunotherapy response for advanced NSCLC patients, uncovering characteristics that otherwise could not be identified by human observers (Gong et al., 2022; Khorrami et al., 2020; Trebeschi et al., 2019, 2021; Tunali et al., 2019). In addition, recent advances in deep learning have shown that radiomics features can be automatically extracted using neural networks without human feature interaction, resulting in better prediction performance (deep-radiomics) (Mu et al., 2021; Tian et al., 2021). Most of these studies have focused on the development of biomarkers considering only baseline and first follow-up information. However, given that tumors are heterogeneous in terms of both spatial heterogeneity and temporal evolution, it could be beneficial to consider more temporal information during early treatment to understand better tumor response patterns.

Furthermore, integrating multimodal data, such as clinical and imaging data, could provide complementary patient- and tumor-specific information for better patient monitoring (Vanguri et al., 2022).

The present chapter aimed to investigate the potential improvements in prediction performance by integrating imaging and clinical data monitored through early treatment. The ability of deep learning to extract more complex and response-related features was also explored and compared with traditional radiomics. An ensemble model based on the integration of longitudinal radiomics and clinical data has been developed and validated in an independent test set to predict the clinical durable benefit of immunotherapy in patients with NSCLC at 6 and 9 months after the start of treatment.

3.2 Materials and Methods

3.2.1 Datasets and Patient Selection

Immunotherapy Dataset

A total of 291 patients with pathologically confirmed stage IV NSCLC treated with anti-PD-1/PD-L1 monoclonal antibodies from January 2013 to December 2021 were retrospectively collected at the Hospital Universitario Fundación Jiménez Díaz (FJD, 154 patients) and Clínica Universidad de Navarra (CUN, 137 patients). Their institutional review boards approved the study, and informed consent was collected accordingly. Inclusion criteria were: (a) confirmed advanced NSCLC; (b) patients were treated with immunotherapy as monotherapy, a combination of immuno-based agents, or in combination with traditional treatment such as chemotherapy or radiation therapy; (c) availability of clinical and epidemiological information; (d) patient data were not right-censored. Finally, 264 patients were enrolled in this study.

The institutional medical records systems were searched to identify those patients with imaging data. CT images were available for 186 patients and were collected following these inclusion criteria: (a) availability of chest CT scans; (b) availability of baseline CT within 2 months before the start of immunotherapy. Exclusion criteria were as follows: (a) lung resection during treatment; (b) an experienced radiologist could not detect and segment the primary tumor in the baseline CT; (c) poor quality image; (d) patient data were not right-censored. Finally, 171 patients were enrolled for imaging data analysis.

According to the clinical protocol, during the first 4 months of immunotherapy, CT scans were acquired after every two or three treatment cycles. Conventional clinical evaluations (including hemograms) were performed after each treatment cycle within the first 2 months of treatment. As a result, our data included demographic, epidemiological, hemogram and other conventional clinical data from this period, at least a baseline CT scan and up to two follow-up CT scans.

The cohort was divided into a training set and an independent test set balancing the availability of baseline and follow-up data (Figure 3.1). To compare the performance of the different models, 43 patients ($43/171 = 25\%$) with baseline imaging and clinical data were randomly selected as an independent test set to maximize the number of patients available to test all the implemented models. All remaining patients were used as the discovery set ($n = 221$). Among the independent test cohort patients, 40 had longitudinal imaging data, 33 longitudinal clinical data and 32 had both longitudinal imaging and clinical data.

Characterization Dataset

In addition to the immunotherapy dataset, a characterization dataset was considered in order to train a deep learning model on nodule characteristics for feature extraction. Benign and confirmed malignant nodules were collected from 719 patients of The Lung Image Database Consortium and Image Database Resource Initiation Data Set (LIDC-IDRI), which consists of annotated chest CT

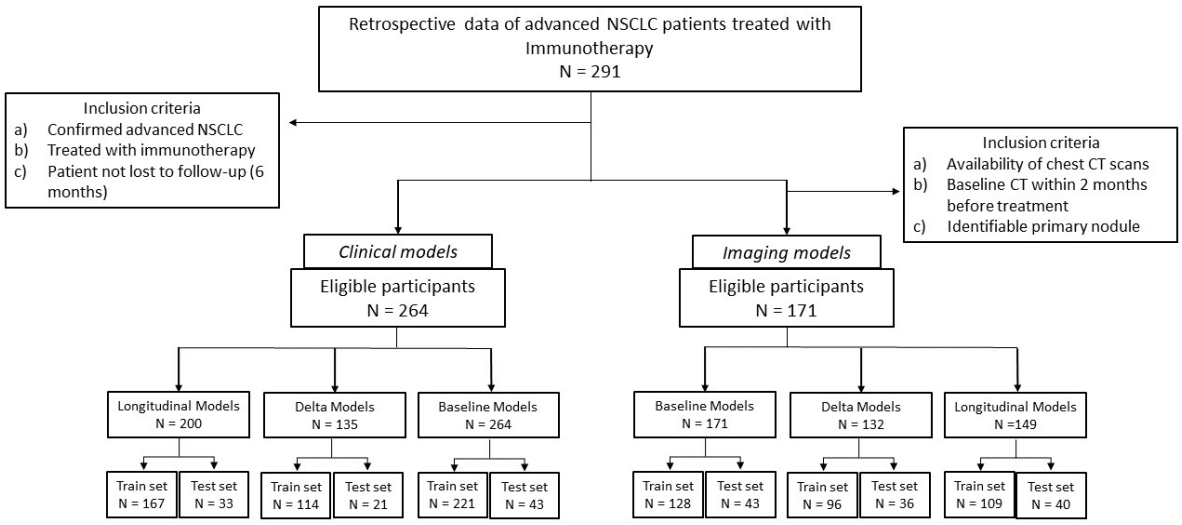


Figure 3.1: Flowchart showing the inclusion and exclusion criteria considering the endpoint PFS6. Details of the number of patients in the training and independent test set are provided.

scans for lung cancer screening. Fourteen patients who did not meet the inclusion criteria of the immunotherapy dataset were also added to this collection for a total of 733 patients. Their nodules were all malignant. Given that each patient may have more than one nodule, the characterization dataset contained 1,528 nodules, 1024 of which were benign and 504 malignant.

3.2.2 Clinical Endpoints

The primary endpoint of this study was the durable clinical benefit defined by progression-free survival (PFS). It measures the time from the first cycle of immunotherapy to death/disease progression or last follow-up. Disease progression was defined based on the patient's general clinical status and iRECIST criteria derived from the imaging evaluation. Patients with durable clinical benefit that had a PFS longer than 6 (PFS6) or 9 (PFS9) months were denominated as responders, while the others as non-responders (Dercle et al., 2022). Patients with censored data 6 or 9 months after treatment were excluded from the analysis. The maximum follow-up period was 48 months.

The secondary endpoint was overall survival (OS), defined as the time in months between the initiation of immunotherapy and death or censored to the last follow-up visit for survivors.

Metadata	Value range	FJD	CUN
Manufacturer	Siemens Toshiba GE Medical Systems Philips	Siemens Toshiba Philips	Siemens Toshiba GE Medical Systems
kVp (kV)	[80, 140]	[80, 140]	[80, 140]
X-ray current tube (mA)	[48, 1481]	[60, 1481]	[48, 998]
Pixel Spacing (mm)	[0.359, 1.523]	[0.359, 0.975]	[0.525, 1.523]
Slice thickness (mm)	[0.625, 5.0]	[0.900, 3.0]	[0.625, 5.0]

Table 3.1: CT image acquisition and reconstruction parameters of the two institutions involved in the study: FJD and CUN.

3.2.3 Image Acquisition and Preprocessing

All patients underwent a CT scan within 2 months before the immunotherapy treatment start date. When available, follow-up CT scans were acquired within 4 months after treatment (up to three temporal time points per patient). All CT images were acquired after contrast injection during a patient inspiratory breath hold, following the contrast-enhanced CT chest protocol. CT scans were reconstructed using a standard kernel. A description of CT parameters is available in Table 3.1. Further details regarding the CT manufacturers' models from which the images were extracted can be found in 6.3 Table A1.

For each case, the primary tumor was selected as the target lesion. 3D tumors were identified and segmented by an experienced radiologist on the baseline and follow-up CT images using either the syngo.via Siemens Healthineers software or 3D Slicer (Fedorov et al., 2012). The largest lesion was considered if a patient had an ambiguous primary tumor. Follow-up CT scans were discarded if the tumor found in the baseline CT scan was no longer visible.

For preprocessing, Hounsfield units of all CT images were clipped between -1000 and 3050, and z-score normalization was then applied.

3.2.4 Feature Extraction

Radiomics Analysis

Radiomics features were extracted by using Pyradiomics (version 3.0.1) (Van Griethuysen et al., 2017). The voxel intensity values were discretized when computing some texture features using a bin width of 25 Hounsfield units (Larue et al., 2017). To reduce the effect of low resolution along the z-axis in part of the data, the radiomics features were computed only by applying 2D filters.

The spectrum of radiomics features encompassed shape features, first-order features (intensity statistics), and texture features. Texture features included those extracted from the gray-level co-occurrence matrix (GLCM), gray-level size zone matrix (GLSZM), gray-level run length matrix (GLRLM), gray-level dependence matrix (GLDM), and neighboring gray-tone difference matrix (NGTDM). Moreover, features were extracted from both the original and transformed images,

involving wavelet, LoG, SquareRoot, square, logarithm, and exponential transformations.

Many studies have shown evidence that radiomics features are influenced by image acquisition parameters and image segmentation. For this reason, we performed a reproducibility analysis to select only those features that were stable with respect image acquisition parameters and changes in segmentations. Feature repeatability against segmentation was verified using two datasets: QIN Lung CT Segmentation and a random subset of the immunotherapy dataset. In the first case, two automatic segmentations with two different algorithms for each nodule were considered. In the second case, an experienced radiologist refined two segmentations of the same nodule obtained with two different modules of syngo.via software. A total of 56 nodules were analyzed. Feature reproducibility was assessed through a test-retest analysis using the Reference Image Database to Evaluate Therapy Response (RIDER) dataset. The dataset included 31 patients who underwent two chest CT scans, acquired 15 minutes apart with the same image protocol. To assess feature repeatability and reproducibility, the Lin's concordance correlation coefficient (CCC) was used. Features presenting a high CCC (≥ 0.85) for both tests were considered stable and used for further analysis. Reproducible and repeatable features are potentially more robust to variations in CT scanners, acquisition parameters, and segmentation.

After feature extraction and reproducibility selection, delta-radiomics features were calculated as the relative net change between features at baseline and first follow-up CTs. Patients without first follow-up CT were discarded from this analysis.

A standard scaler was applied to normalize each radiomics feature. The transformation was learned in training and then applied to the test set.

Deep Feature Extraction

To extract high-level and domain-related representations (e.g., texture, morphology) of the tumors' deep learning-based features, the convolutional neural network (CNN) architecture NoduleX (Causey et al., 2018) was used as a reference implementation to predict the response to immunotherapy. NoduleX input consists of a small 3D volume of 47×47 pixels $\times$ 5 slices centered in the centroid of the tumor that was sampled and resized from a square of 10×10 cm^2. Image intensities were clipped to the range [–1000, 3050] HU and then normalized.

A transfer learning approach was used to pretrain NoduleX CNN architecture weights. Namely, the network was pre-trained to predict the malignancy of tumors collected from 719 patients of LIDC-IDRI dataset (Armato III et al., 2011) and 14 patients who did not meet the inclusion criteria of the immunotherapy dataset (1528 nodules). Then, the last two convolutional layers and the classification layers were fine-tuned to predict the response (defined by the endpoint PFS6). For network fine-tuning, all primary tumors of all available CT images from the immunotherapy training data set (357 tumors - 128 patients) were used. Fine-tuning allowed the efficient transfer of malignant-related spatial features to more complicated high-level semantic features related to immunotherapy response.

After training, deep features were extracted for each tumor from the first fully connected layers of the network (500 deep features), referred to as DF-imm. Similarly to delta-radiomics, delta

Clinical variables	
Categorical variables	**Continuous variables**
Sex	Platelets (cells/microL)
Age	Lymphocytes (cells/microL)
Weight	Monocytes (cells/microL)
Height	Eosinophiles (cells/microL)
Body mass index (BMI)	Hemoglobin (Hb, g/dL)
Surgery	Neutrophil absolute count (NT, cells/microL)
Smoking	Neutrophil-to-lymphocyte ratio (NLR)
Tumor histology	Monocyte-to-lymphocyte ratio (MLR)
Steroids	Platelet-to-lymphocyte ratio (PLR)
Antibiotics	Systemic immune-inflammation index (SII)
COPD	
Central nervous system (CNS) metastases	
Adrenal metastases	
Liver metastases	
Bone metastases	

Table 3.2: Clinical variables used for the implementation of the clinical and ensemble models.

DF-imm features were also calculated.

3.2.5 Clinical Data

Baseline demographic, epidemiological, clinical and laboratory data were collected from electronic patient records, as well as hemogram-related data after the second and third treatment cycles. From the electronic medical record, it has been retrieved information about demographics, tumor histology, smoking habits, stage of disease, presence of metastases per site prior the treatment, chronic obstructive pulmonary disease (COPD), etc. Hematology data were obtained from the blood test performed after the second and third cycles of immunotherapy and included: platelets, lymphocytes, monocytes, eosinophiles, hemoglobin, Neutrophil absolute count (NT), Neutrophil-to-lymphocyte ratio (NLR), Monocyte-to-lymphocyte ratio (MLR), Platelet-to-lymphocyte ratio (PLR), Systemic immune-inflammation index (SII). More details about clinical variables are shown in Table 3.2.

One hot encoding was applied to categorical or constant variables. Z-score normalization was applied to continuous variables, and missing data were imputed using the k-means algorithm. Delta features were also calculated.

3.2.6 Model Design and Analysis

Random Forest (RF) models were built for each primary endpoint in the training set using stratified three-fold cross-validation. Feature selection and RF hyperparameter optimization were performed using a Bayesian optimization approach. The optimized hyperparameters were the number of estimators, the maximum depth, and the number of features.

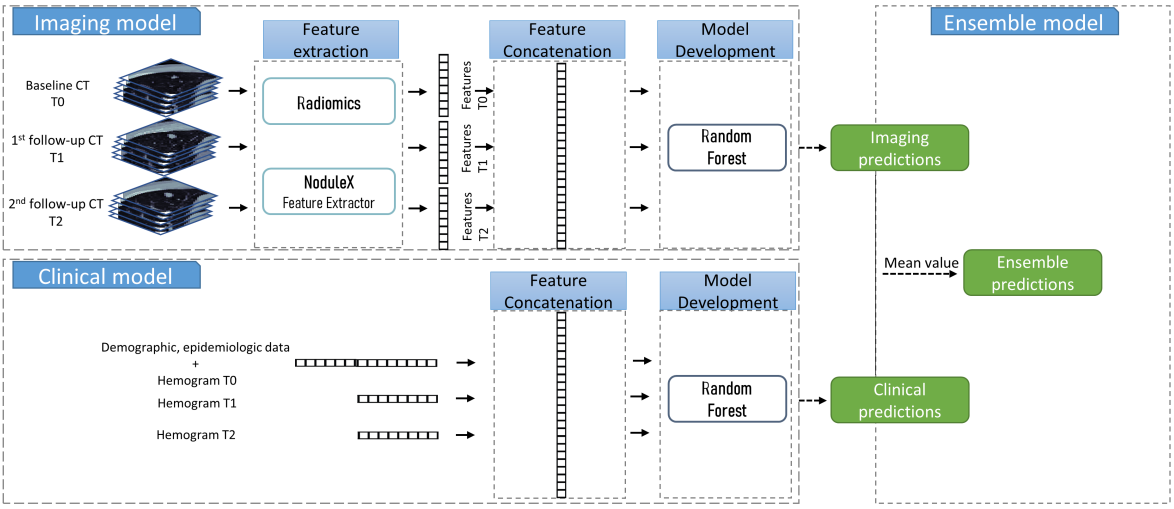


Figure 3.2: Implementation workflow of the longitudinal and ensemble models.

Radiomics, deep features, and clinical data were used to implement baseline, delta, and longitudinal RF models trained for predicting the immunotherapy response. Baseline models' (RF-baseline) inputs were only the data before the start of treatment, whereas longitudinal models used baseline and early treatment data. Patients who did not have follow-up data were excluded from the longitudinal analysis. Two types of longitudinal models were constructed: RF-delta and RF-longitudinal. RF-delta model had delta features as input and considered only patients with baseline and first follow-up data. On the other hand, RF-longitudinal input was the concatenation of all available features over time for each patient (number of features multiplied by the number of time points). Missing time points were imputed as the closest in time available data.

For comparison, the NoduleX architecture pre-trained for malignancy prediction was fine-tuned with the baseline training data of the immunotherapy dataset to predict treatment durable response (CNN-baseline).

For predicting PFS9, because the training was imbalanced, a synthetic minority oversampling technique (SMOTE) was used during the training phase to resample the minority class ("responders"). As SMOTE was configured to generate synthetic samples in training considering five nearest neighbors, the numbers of responders and nonresponders were equal.

Once the models were trained, ensemble RF models were implemented as the mean value of the predictions of the imaging and clinical models alone (ensemble RF). They allowed integrating both clinical and image information. The workflow is shown in Figure 3.2.

3.2.7 Model Interpretation

The SHAP (or SHapley Additive exPlanations) algorithm was employed to visualize each feature's contribution to producing the final prediction of the model (Lundberg & Lee, 2017). SHAP assigns an importance value to each feature for each individual predicted value based on concepts from

Cooperative Game Theory and local explanations. We applied the SHAP algorithm to the clinical model of the ensemble RF model. SHAP values were calculated to understand how much each feature impacted the model output or how much it increased or decreased the probability of a single outcome. SHAP values allowed us to determine whether the relationship between a feature and the output was correlative or anticorrelative. SHAP analysis was performed in Python using the KernelExplainer in the SHAP module (version 0.40.0).

3.2.8 Statistical and Survival Analysis

Stratified three-fold cross-validation was performed in the training set to train all the implemented models and optimize the RF hyperparameters. Model performance was evaluated by the area under the receiver operating characteristic (ROC) curve (AUC) and the corresponding 95% confidence interval (CI) was estimated with a bootstrap resampling approach (1000 iterations). The differences between ROC curves were assessed using the DeLong test. Kaplan–Meier survival analysis was performed for patients' stratification based on the model's predictions (threshold $= 0.5$). The significance of differences between survival curves was assessed with the log-rank test. Hazard ratios (HRs) and concordance index were calculated using the Cox proportional-hazards model. p-values less than 0.05 (two-sided tests) were considered significant. R (version 4.1.1) and Python (version 3.7.10) were used for statistical analysis and model implementation.

3.3 Experiments and Results

3.3.1 Patient Characteristics

The clinical characteristics of patients in the training and independent test cohorts in the baseline and longitudinal analysis for PFS6 are summarized in Table 3.3. The characteristics of a subset of patients with imaging data are summarized Tables 3.4 and 3.5. The same distributions were also verified for PFS9.

Among the selected 264 patients, 80 were female (mean age, 62.6 $\pm$9.8 [standard deviation]) and 184 were male (mean age, 65.7 $\pm$9.7 [standard deviation]). Regarding our cohort, we found the following: 43.9% of the patients responded to immunotherapy after 6 months of treatment, while only 33.2% responded after 9 months; adenocarcinoma was the most prevalent histological variant of advanced NSCLC (76.9%); and 89.7% of the patients were current or former smokers. Immunotherapy treatment included monotherapy (58.3%), immunotherapy combined with radiation therapy (6.4%), immunotherapy combined with chemotherapy (18.9%) and a combination of different immunological agents (14.8%). No demographic or clinical characteristics had significant differences (p-value < 0.05) between the training and test set after the two samples of T-tests for continuous variables and Chi-square tests for categorical variables.

For the subcohort of patients with imaging data (171 over 264 patients), the training and the independent test sets had identical distributions of demographics and clinical characteristics (no statistical difference $p > 0.05$).

Characteristic	Baseline Analysis				Longitudinal Analysis			
	All patients (N= 264)	Train set (N = 221)	Test set (N = 43)	P-value	All patients (N= 200)	Train set (N = 167)	Test set (N = 33)	p-value
PFS, mean (SD)	9.0 (11.1)	9.3 (11.6)	7.6 (8.1)	0.242	11.1 (11.8)	11.6 (12.3)	9.0 (8.6)	0.147
OS, mean (SD)	13.3 (12.2)	13.3 (12.5)	13.5 (10.5)	0.903	16.0 (12.4)	16.0 (12.8)	15.7 (10.6)	0.889
Status								
Alive	107 (40.5%)	91 (41.2%)	16 (37.2%)	0.753	91 (45.5)	78 (46.7)	13 (39.4)	0.562
Dead	157 (59.5%)	130 (58.8%)	27 (62.8%)		109 (54.5)	89 (53.3)	20 (60.6)	
Response								
Non-responders	148 (56.1%)	124 (56.1%)	24 (55.8%)	1.000	90 (45.0%)	75 (44.9%)	15 (45.5%)	1.000
Responders	116 (43.9%)	97 (43.9%)	19 (44.2%)		110 (55.0%)	92 (55.1%)	18 (54.5%)	
Progression								
No progression	45 (17.0%)	40 (18.1%)	5 (11.6%)	0.417	42 (21.0%)	38 (22.8%)	4 (12.1%)	0.256
Progression	219 (83.0%)	181 (81.9%)	38 (88.4%)		158 (79.0%)	129 (77.2%)	29 (87.9%)	
Age, median [Q1,Q3]	65.0 [59.0,71.0]	65.0 [58.0,71.0]	67.0 [60.5,72.5]	0.204	65.0 [58.0,70.2]	64.0 [57.0,70.0]	67.0 [60.0,72.0]	0.266
Gender								
Female	80 (30.3%)	66 (29.9%)	14 (32.6%)	0.865	58 (29.0%)	47 (28.1%)	11 (33.3%)	0.696
Male	184 (69.7%)	155 (70.1%)	29 (67.4%)		142 (71.0%)	120 (71.9%)	22 (66.7%)	
IPA, mean (SD)	45.2 (33.4)	45.1 (33.8)	45.4 (31.5)	0.958	44.0 (34.1)	44.9 (34.6)	39.0 (31.2)	0.357
Smoking								
Current smoker	55 (21.0%)	50 (22.7%)	5 (11.9%)	0.258	39 (19.7%)	35 (21.1%)	4 (12.5%)	0.530
Former smoker	180 (68.7%)	147 (66.8%)	33 (78.6%)		135 (68.2%)	111 (66.9%)	24 (75.0%)	
Non-smoker	27 (10.3%)	23 (10.5%)	4 (9.5%)		24 (12.1%)	20 (12.0%)	4 (12.5%)	
Tumor Histology								
Adenocarcinoma	203 (76.9%)	170 (76.9%)	33 (76.7%)	0.897	151 (75.5%)	126 (75.4%)	25 (75.8%)	0.896
Squamous cell carcinoma	52 (19.7%)	43 (19.5%)	9 (20.9%)		40 (20.0%)	33 (19.8%)	7 (21.2%)	
Other	9 (3.4%)	8 (3.6%)	1 (2.3%)		9 (4.5%)	8 (4.8%)	1 (3.0%)	
PDL1, mean (SD)	0.4 (0.4)	0.4 (0.4)	0.4 (0.4)	0.876	0.4 (0.4)	0.4 (0.4)	0.3 (0.3)	0.194
Surgery								
No	227 (86.0%)	190 (86.0%)	37 (86.0%)	1.000	171 (85.5%)	142 (85.0%)	29 (87.9%)	0.792
Yes	37 (14.0%)	31 (14.0%)	6 (14.0%)		29 (14.5%)	25 (15.0%)	4 (12.1%)	
Treatment								
Combined Immunological Agents	39 (14.8%)	29 (13.1%)	10 (23.3%)	0.393	31 (15.5%)	24 (14.4%)	7 (21.2%)	0.276
Immunotherapy+Chemotherapy	50 (18.9%)	41 (18.6%)	9 (20.9%)		39 (19.5%)	30 (18.0%)	9 (27.3%)	
Immunotherapy+Radiotherapy	17 (6.4%)	15 (6.8%)	2 (4.7%)		11 (5.5%)	11 (6.6%)	0 (0%)	
Monotherapy	154 (58.3%)	132 (59.7%)	22 (51.2%)		116 (58.0%)	99 (59.3%)	17 (51.5%)	
Other	4 (1.5%)	4 (1.8%)	0 (0%)		3 (1.5%)	3 (1.8%)	0 (0%)	

Table 3.3: Demographic and clinical characteristics of the patients in the baseline and longitudinal analyses. Adjusted p-values (using Bonferroni correction) were calculated for comparisons between the two cohorts using two-sample T-tests for continuous variables and Chi-square tests for categorical variables. SD represents the standard deviation, and Q1 and Q3 represent the first and third quartiles, respectively.

PFS6	Train set			Test set		
Data	RF-baseline	RF-delta	RF-longitudinal	RF-baseline	RF-delta	RF-longitudinal
Only clinical data	221	114	167	43	21	33
Only imaging data	128	96	109	43	36	40
Clinical and imaging data	128	-	92	43	-	32

Table 3.4: Number of patients in the training and independent test set for each model considering the endpoint PFS6.

PFS9	Train set			Test set		
Data	RF-baseline	RF-delta	RF-longitudinal	RF-baseline	RF-delta	RF-longitudinal
Only clinical data	216	112	163	43	21	33
Only imaging data	125	93	106	43	36	40
Clinical and imaging data	125	-	90	43	-	32

Table 3.5: Number of patients in the training and independent test set for each model considering the endpoint PFS9.

3.3.2 Model Development and Response Prediction Performance

From the initial set of 1365 radiomics features, only 173 (13%) verified both reproducibility and repeatability against segmentation tests. Furthermore, a total of 500 DF-imm were extracted for each tumor using the NoduleX architecture. After feature selection, each model had a different number of features as input.

Figures 3.3 and 3.4 compare the ROC curves of CNN-baseline and the baseline, delta and longitudinal RF models using clinical, radiomics and DF-imm data in the independent test cohort for PFS6 and PFS9, respectively.

Longitudinal models performed better than baseline or delta models in the independent test cohort, achieving an AUC of 0.740 (95% CI: 0.563-0.833) with DF-imm and an AUC of 0.700 (95% CI: 0.508-0.877) with clinical data for PFS6 and an AUC of 0.702 (95% CI: 0.515-0.867) with DF-imm and an AUC of 0.585 (95% CI: 0.367-0.783) with clinical data for PFS9. In both cases, the automatically extracted features performed better than the hand-crafted radiomics features and clinical data.

Tables 3.6 and 3.7 compare the evaluation metrics of all implemented models, showing great improvement when using the longitudinal models.

3.3.3 Integration of Imaging and Clinical Data

Table 3.8 shows the performance in the independent test set of the ensemble RF models that used both clinical and imaging information. The comparison with baseline and longitudinal RF models tested on the same patients is shown in Appendix A: Tables A2 and A3 for endpoint PFS6 and PFS9, respectively. The ensemble RF-longitudinal achieved an AUC of 0.824 (95% CI: 0.658-0.953) for PFS6 with a 41% improvement for RF models with only clinical data (DeLong test: p-value = 0.001) and 13% for the RF model with deep features data (DeLong test: p-value = 0.013). When

Model	Features	N test	AUC [95% CI]	ACC [95% CI]	SENS [95% CI]	SPES [95% CI]	PREC [95% CI]	bACC [95% CI]
CNN-baseline	Image data	43	0.518 [0.329,0.696]	0.535 [0.372,0.674]	0.750 [0.565,0.909]	0.263 [0.067,0.478]	0.562 [0.387,0.737]	0.507 [0.377,0.643]
RF-baseline	Clinical data	43	0.667 [0.485,0.833]	0.651 [0.512,0.791]	**0.833** [0.667,0.962]	0.421 [0.200,0.650]	0.645 [0.480,0.812]	0.627 [0.488,0.774]
RF-baseline	Radiomics	43	0.448 [0.291,0.607]	0.442 [0.302,0.605]	0.333 [0.150,0.526]	0.579 [0.350,0.800]	0.500 [0.250,0.750]	0.456 [0.306,0.601]
RF-baseline	DF-imm	43	0.588 [0.409,0.767]	0.558 [0.419,0.698]	**0.833** [0.679,0.960]	0.211 [0.050,0.417]	0.571 [0.406,0.735]	0.522 [0.403,0.638]
RF-delta	Clinical data	21	0.435 [0.173,0.714]	0.333 [0.143,0.571]	0.167 [0.000,0.417]	0.556 [0.200,0.875]	0.333 [0.000,0.750]	0.361 [0.163,0.559]
RF-delta	Radiomics	36	0.489 [0.276,0.706]	0.528 [0.361,0.694]	0.524 [0.304,0.737]	0.533 [0.273,0.786]	0.611 [0.389,0.833]	0.529 [0.357,0.688]
RF-delta	DF-imm	36	0.660 [0.451,0.846]	0.611 [0.444,0.778]	0.714 [0.500,0.900]	0.467 [0.200,0.733]	0.652 [0.455,0.842]	0.590 [0.433,0.750]
RF-longitudinal	Clinical data	33	0.700 [0.508,0.877]	0.576 [0.394,0.727]	0.467 [0.200,0.733]	**0.667** [0.438,0.875]	0.530 [0.250,0.800]	0.567 [0.405,0.733]
RF-longitudinal	Radiomics	40	0.581 [0.407,0.749]	0.628 [0.488,0.767]	0.667 [0.464,0.850]	0.579 [0.348,0.800]	0.667 [0.474,0.852]	0.623 [0.466,0.763]
RF-longitudinal	DF-imm	40	**0.740** [0.563,0.883]	**0.700** [0.550,0.825]	0.818 [0.647,0.958]	0.556 [0.312,0.783]	**0.692** [0.500,0.864]	**0.687** [0.550,0.827]

Table 3.6: Response prediction performance comparison between baseline, delta and longitudinal models in the independent test set for endpoint PFS6 by evaluating AUC, ACC, SENS, SPEC, PREC and bACC, respectively. For each metric, the 95% confidence interval is shown and the highest value is highlighted in bold.

Model	Features	N test	AUC [95% CI]	ACC [95% CI]	SENS [95% CI]	SPES [95% CI]	PREC [95% CI]	bACC [95% CI]
CNN-baseline	Image data	43	0.429 [0.249,0.616]	0.674 [0.535,0.814]	1.000 [1.000,1.000]	0.000 [0.000,0.000]	0.674 [0.535,0.814]	0.500 [0.500, 0.500]
RF-baseline	Clinical data	43	0.563 [0.392,0.735]	0.581 [0.442,0.721]	0.793 [0.636,0.929]	0.143 [0.000,0.357]	0.657 [0.500,0.811]	0.468 [0.352,0.591]
RF-baseline	Radiomics	43	0.286 [0.112,0.494]	0.512 [0.372,0.651]	0.655 [0.480,0.815]	0.214 [0.000,0.455]	0.633 [0.464,0.800]	0.435 [0.303,0.576]
RF-baseline	DF-imm	43	0.541 [0.359,0.724]	0.628 [0.488,0.767]	0.759 [0.600,0.903]	0.357 [0.118,0.600]	0.710 [0.533,0.867]	0.558 [0.405,0.711]
RF-delta	Clinical data	21	0.550 [0.301,0.795]	0.524 [0.333,0.762]	0.636 [0.333,0.900]	0.400 [0.100,0.714]	0.538 [0.250,0.818]	0.518 [0.306,0.750]
RF-delta	Radiomics	36	0.598 [0.353,0.848]	0.639 [0.472,0.778]	0.680 [0.500,0.857]	**0.545** [0.231,0.857]	**0.773** [0.588,0.947]	0.613 [0.429,0.788]
RF-delta	DF-imm	36	0.525 [0.315,0.743]	0.556 [0.389,0.722]	0.760 [0.571,0.920]	0.091 [0.000,0.300]	0.655 [0.481,0.824]	0.425 [0.315,0.554]
RF-longitudinal	Clinical data	33	0.585 [0.367,0.783]	0.545 [0.364,0.697]	0.600 [0.381,0.812]	0.462 [0.182,0.727]	0.632 [0.412,0.850]	0.531 [0.360,0.698]
RF-longitudinal	Radiomics	40	0.528 [0.341,0.701]	0.558 [0.395,0.698]	0.724 [0.562,0.88]	0.214 [0.000,0.455]	0.656 [0.484,0.818]	0.469 [0.338,0.612]
RF-longitudinal	DF-imm	40	**0.702** [0.515,0.867]	**0.750** [0.625,0.875]	**0.885** [0.750,1.000]	0.500 [0.214,0.769]	0.767 [0.606,0.914]	**0.692** [0.540,0.840]

Table 3.7: Response prediction performance comparison between baseline, delta and longitudinal models in the independent test set for endpoint PFS9 by evaluating AUC, ACC, SENS, SPEC, PREC and bACC, respectively. For each metric, the 95% confidence interval is shown and the highest value is highlighted in bold.

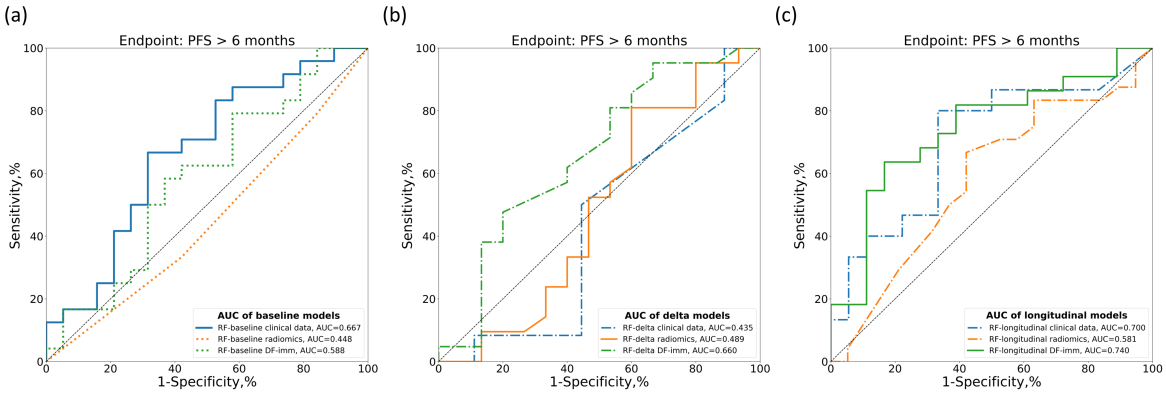


Figure 3.3: Comparisons of the ROC curves for endpoint PFS6 for the baseline (a), delta (b), and longitudinal RF models (c) based on clinical, radiomics, or deep-radiomics data.

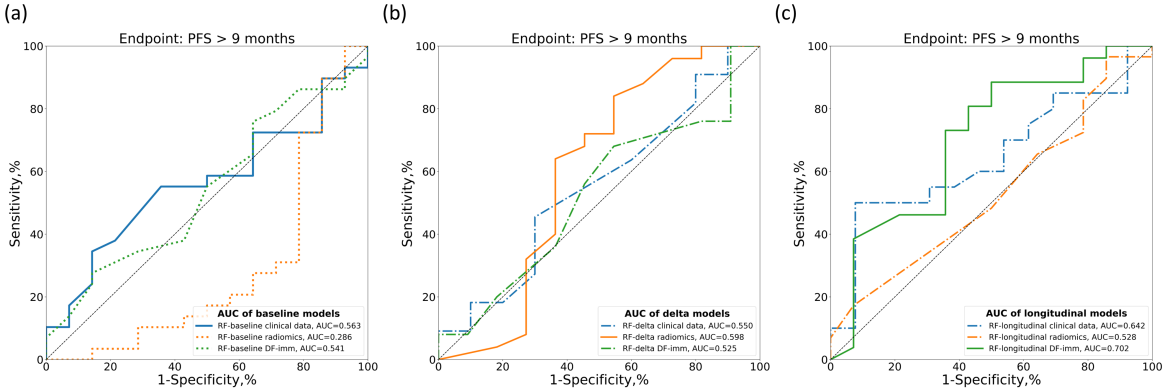


Figure 3.4: Comparisons of the ROC curves for endpoint PFS9 for the baseline (a), delta (b), and longitudinal RF models (c) based on clinical, radiomics, or deep-radiomics data.

considering PFS9, the ensemble model achieved an AUC of 0.753 (95% CI: 0.549-0.931) with a 31% improvement compared to RF models with only clinical data (DeLong test: p-value = 0.053) and 5% for the RF model based on deep features data (DeLong test: p-value = 0.058) (Figure 3.5).

Furthermore, the ensemble models scores were significantly associated with progression-free survival and overall survival in the independent test set (6 months: C-index 4.68, 95% CI: [1.52,7.84], p-value < 0.004; 9 months: C-index 2.38, 95% CI: [0.23,4.54], p-value < 0.030). The HRs with their corresponding 95% CIs and the C-indexes of longitudinal and ensemble RF models for PFS and OS are shown in Tables 3.9 (endpoint PFS6) and 3.10 (endpoint PFS9). The integration of clinical and DF-imm data appeared to be a more robust approach compared to the radiomics or clinical models.

Figure 3.6 shows the Kaplan-Meier survival curves for PFS and OS on the independent test set for the ensemble RF models. The ensemble RF could significantly stratify PFS and OS for both endpoints compared to the other models (p-value < 0.05). The comparisons between Kaplan-Meier curves for longitudinal RF and ensemble RF models are shown in Appendix A: Figures A1 (endpoint

Endpoint	Model	Features	N test	AUC [95% CI]	ACC [95% CI]	SENS [95% CI]	SPES [95% CI]	PREC [95% CI]	bACC [95% CI]
PFS6	Ensemble RF-baseline	DF-imm Clinical data	43	0.678 [0.513,0.836]	0.605 [0.442,0.744]	**0.875** [0.731,1.000]	0.263 [0.071,0.467]	0.600 [0.436,0.758]	0.569 [0.448,0.684]
	Ensemble RF-longitudinal	DF-imm Clinical data	32	**0.824** [0.658,0.953]	**0.750** [0.594,0.906]	0.733 [0.500,0.938]	**0.765** [0.533,0.947]	**0.733** [0.471,0.933]	**0.749** [0.594,0.897]
PFS9	Ensemble RF-baseline	DF-imm Clinical data	43	0.560 [0.377,0.731]	0.581 [0.442,0.721]	0.793 [0.643,0.933]	0.143 [0.000,0.364]	0.657 [0.487,0.811]	0.468 [0.360,0.590]
	Ensemble RF-longitudinal	DF-imm Clinical data	32	**0.753** [0.549,0.931]	**0.813** [0.656,0.938]	**0.947** [0.826,1.000]	**0.615** [0.357,0.889]	**0.783** [0.609,0.950]	**0.781** [0.631,0.923]

Table 3.8: Response prediction performance comparison between longitudinal and ensemble models in the independent test set for endpoint PFS6 and PFS9 by evaluating AUC, ACC, SENS, SPEC, PREC and bACC, respectively. For each metric, the 95% confidence interval is shown and the highest value for each endpoint is highlighted in bold.

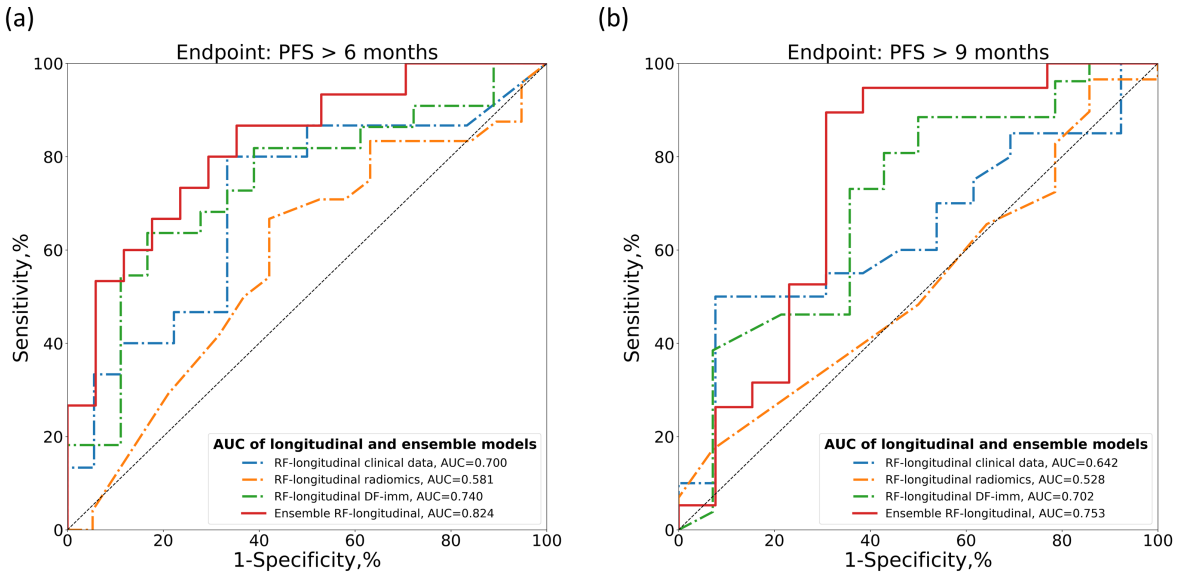


Figure 3.5: Comparisons of ROC curves of longitudinal and ensemble RF models with clinical and radiomics data. (a) ROC curves for PFS6: PFS $>$ 6 months. (b) ROC curve for PFS9: PFS $>$ 9 months.

		PFS			OS		
Model	Features	HR [95% CI]	p-value	C-index	HR [95% CI]	p-value	C-index
RF-longitudinal	Clinical data	1.63 [-1.00,4.25]	0.224	0.615	3.49 [0.28,6.69]	0.033	0.656
RF-longitudinal	DF-imm	3.30 [1.02,5.59]	0.005	0.687	4.31 [1.43,7.12]	0.003	0.709
Ensemble RF-longitudinal	DF-imm Clinical data	**4.68** [1.52,7.84]	**0.004**	**0.723**	**6.00** [2.27,9.73]	**0.002**	**0.768**

Table 3.9: Hazard ratios and C-indexes of longitudinal and ensemble models trained for endpoint PFS6 to predict PFS and OS in the independent test set. The best value for each metric is highlighted in bold.

		PFS			OS		
Model	Features	HR [95% CI]	p-value	C-index	HR [95% CI]	p-value	C-index
RF-longitudinal	Clinical data	0.52 [-1.16,2.20]	0.542	0.575	1.73 [-0.67,4.13]	0.157	0.613
RF-longitudinal	DF-imm	1.35 [-0.23,2.92]	0.093	0.642	1.72 [-0.18,3.62]	0.076	0.641
Ensemble RF-longitudinal	DF-imm Clinical data	**2.38** [0.23,4.54]	**0.030**	**0.685**	**2.94** [0.40,5.48]	**0.023**	**0.736**

Table 3.10: Hazard ratios and C-indexes of longitudinal and ensemble models trained for endpoint PFS9 to predict PFS and OS in the independent test set. The best value for each metric is highlighted in bold.

PFS6) and A2 (endpoint PFS9).

3.3.4 Model Interpretation

The SHAP algorithm was employed to visualize each feature's contribution to producing the final prediction of the model. The SHAP algorithm was applied to the clinical model of the ensemble RF. A positive SHAP value indicated an increased risk of progression for each prediction. As observed in Figures 3.7a and 3.7b, the most important clinical variables were the NLR and the SII: for both endpoints, the higher the values in the second time step (around 1–2 months after treatment), the higher the probability of progression. Moreover, the presence of liver metastases appeared to be related to a worse outcome.

3.4 Discussion and Conclusion

In immuno-oncology, the traditional approach of manually measuring the size changes of the target lesions during treatment is no longer adequate because the tumor unconventionally responds to treatment (Borcoman et al., 2019). Therefore, identifying unusual tumor response patterns could avoid premature treatment interruptions or ineffective prolongation. Automatic extraction of imaging biomarkers that capture changes in tumor radiophenotypes during treatment in association with clinical information can potentially aid in patient evaluation and ultimately monitor and adapt therapy dynamically.

In this two-institutional study, longitudinal information from clinical data and radiomics was used to predict clinical durable benefit at 6 and 9 months after the start of anti-PD-1/PD-L1 monoclonal antibodies treatment in advanced NSCLC patients using an ensemble approach.

A deep-learning method was used to automatically extract spatial information from CT scans without manual or semiautomatic segmentation and with the advantage of extracting features closely associated with response. Previous studies have demonstrated the ability of deep learning

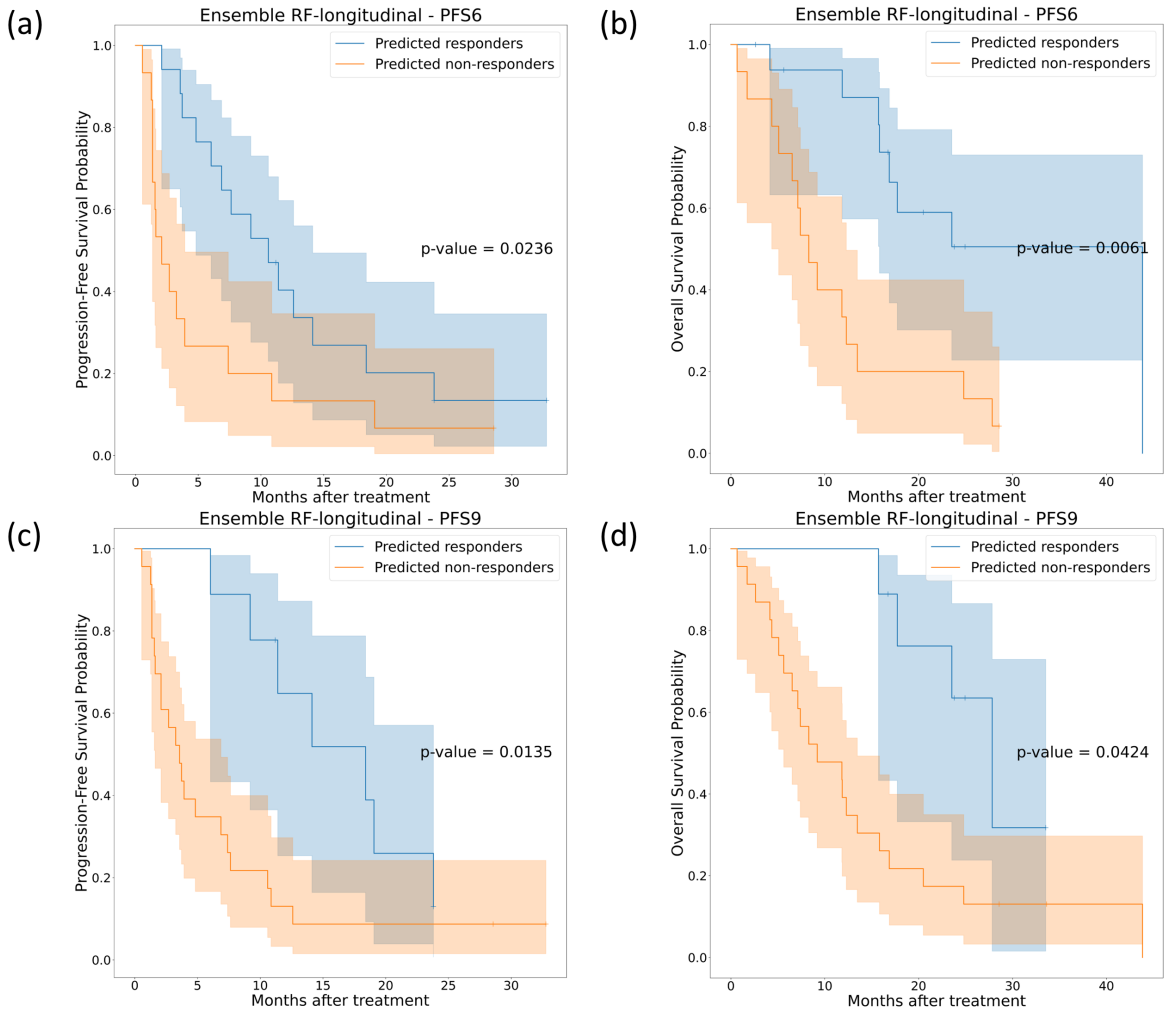
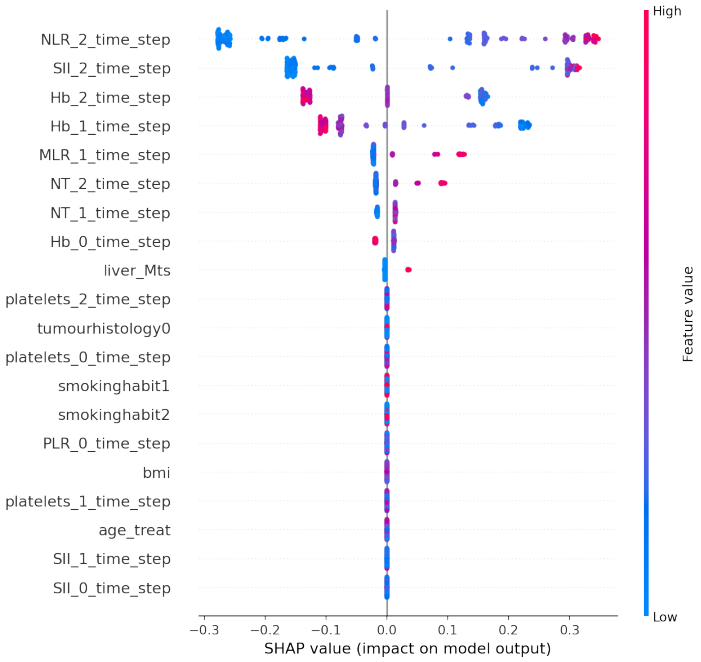


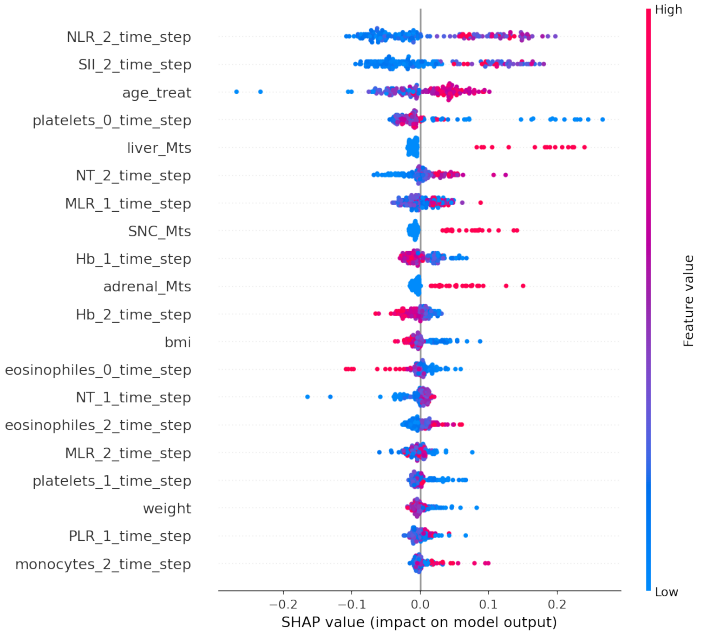
Figure 3.6: Kaplan-Meier survival curves on the independent test cohort for ensemble RF models trained for endpoint PFS6 (first row) and PFS9 (second row). (a) and (c) represent the PFS Kaplan-Meier curves, while (b) and (d) represent the OS Kaplan-Meier curves.

to capture higher-level features related to the immunotherapy response (B. He et al., 2020; Mu et al., 2021; Tian et al., 2021; Trebeschi et al., 2021). The results of this study demonstrated that the deep features were more robust than traditional radiomics in predicting immunotherapy clinical durable benefit in advanced NSCLC, as well as in survival prediction and patient stratification. This confirms the hypothesis that deep-learning techniques allow the extraction of higher-level spatial features that are deeply related to response to treatment. They might represent properties of the tumors that are indicative of treatment response, such as changes in shape, size or intensity.

Moreover, a multiple time-point analysis was performed. Typically, only data before the start of treatment is used for prediction, without including any information during treatment. In previous studies, longitudinal data have been used to predict immunotherapy response from baseline and first follow-up CT scans (Gong et al., 2022; Khorrami et al., 2020; Liu et al., 2021; Trebeschi et al.,



(a) Endpoint PFS6



(b) Endpoint PFS9

Figure 3.7: Clinical model interpretation using SHAP. The summary plots show each clinical data impact on longitudinal RF model for endpoint PFS6 (a) and PFS9 (b). A positive SHAP value indicates an increased risk of progression. Each point in the summary plot represents a patient.

2021). However, using data before treatment and up to four months after treatment (up to three time points per patient), we were able to improve the predictions of durable clinical benefit of immunotherapy.

To the best of our knowledge, no previous studies have demonstrated that the integration of complementary longitudinal clinical and imaging data can significantly improve immunotherapy clinical benefit prediction. The ensembles of longitudinal models with deep-radiomics (DF-imm) and clinical data significantly improved prediction performance, achieving an AUC of 0.824 for PFS6 and an AUC of 0.753 for PFS9. These models significantly stratified patients in high- and low-risk groups for both PFS and OS (p-value < 0.05), and their predictions significantly correlated with PFS (PFS6 model: C-index 0.723, p-value $= 0.004$; PFS9 model: C-index 0.685, p-value $= 0.030$) and OS (PFS6 models: C-index 0.768, p-value $= 0.002$; PFS9 model: C-index 0.736, p-value $= 0.023$). After attempting to identify any unique characteristics among the patients with better survival, we found no significant differences in their clinical data. As a result, we have determined that the accurate predictions result from the model effectively integrating information from both the deep-features and clinical variables. As a comparison, Vanguri et al. (2022) showed that integrating baseline medical imaging, histopathological and genomic features (multimodal model) outperformed unimodal models, achieving an AUC of 0.80 for the immunotherapy response prediction.

The final ensemble models considered changes in imaging tumor radiophenotypes and clinical covariates during early treatment. The SHAP analysis shows that for both PFS6 and PFS9 endpoints, the most important clinical variables were the NLR and the SII. High values of NLR and SII after the second cycle of therapy were highly associated with poor prognosis probably because of a reduced antitumor effect of the immune system. This is consistent with the literature in which baseline NLR is considered a prognostic factor associated with a lower likelihood of treatment response (Valero et al., 2021), and inflammation markers, such as SII, are related to tumor growth, progression, and poor OS (Fu et al., 2021). In our study, both NLR and SII early follow-up values are shown to be important for the clinical durable benefit of the therapy. Furthermore, the models considered that the presence of metastases in the liver before treatment was related to a worse outcome. On the other hand, higher levels of hemoglobin before and during treatment were associated with a better response to treatment.

Our study had some limitations. First, the retrospective and multi-center nature of the work implies a heterogeneity of the cohort in terms of treatment and imaging protocols. Second, the sample size of the two cohorts (FJD and CUN) was relatively large, but a relevant number of cases did not have longitudinal imaging data. Third, there was an important unbalance between responders and nonresponders for PFS9. The SMOTE technique was used to partially reduce this imbalance during the model training, but it did not result in performance comparable to the PFS6 models. To further improve the prediction of treatment response, it may be necessary to collect more data from patients with prolonged responses to treatment and/or include more time points in the analysis. Forth, the interpretation of the deep-features is often not straightforward since they are optimized to minimize the prediction error and are not designed to match human intuition or knowledge. Despite the limitations, they can still offer insights into the relationships between the tumors' image information and response prediction and contribute to making accurate predictions.

Finally, no comparison with other prognostic biomarkers was made, such as PDL1 or tumor mutational burden, due to their inaccessibility. Similarly, for the definition of radiological progression, the iRECIST criteria were not quantitatively evaluated by the radiologists, so that no comparison could have been performed. In addition, the integration of these biomarkers, as well as other new molecular parameters from liquid biopsies such as circulating tumor DNA, circulating tumor cells, circulating endothelial cells or the changes in variant allele frequencies with the deep features and clinical data used in the study, may enhance the performance of the models even further (Kato et al., 2022; Sinoquet et al., 2023).

In conclusion, an ensemble of longitudinal deep-radiomics and clinical data has been used to predict the durable clinical benefit of immunotherapy at 6 and 9 months after treatment. Our results demonstrate that integrating multidimensional and longitudinal data improves prediction performance. The model may be used as a prognostic biomarker and decision-support tool that can assist oncologists in identifying patients for whom the therapy is effective, avoiding premature interruptions or, on the other hand, the lengthening of an ineffective treatment.

CHAPTER 4

Harmonization Impact on Radiomics

Identifying predictive non-invasive biomarkers of immunotherapy response is crucial to avoid premature treatment interruptions or ineffective prolongation. In this chapter, our focus centers on the harmonization of radiomics features across centers as well as in time, probing the interplay between image and feature harmonization and their influence on radiomics features. Introducing a statistical image harmonization method coupled with subsequent feature harmonization, we evaluate their efficacy in predicting immunotherapy treatment response. The content of this chapter has been drafted for a publication that is being submitted to an international journal.[1]

4.1 Introduction

The field of radiomics has been gaining great attention in recent years due to its ability to extract a multitude number of features from medical images, including magnetic resonance (MR) or computed tomography (CT). These features have demonstrated correlations with histological and molecular tumor phenotypes (Grossmann et al., 2017; Sun et al., 2018). Radiomics features may unveil patterns not visually discernible to the human eye, providing a enhanced characterization of properties within the heterogeneous tumor region. One of the primary advantages of radiomics lies in its non-invasive and cost-effective nature, enabling a comprehensive longitudinal analysis of the heterogeneity present in the entire tumor and its surrounding tissues. The fundamental hypothesis behind radiomics is that information about abnormal mechanisms and the overall condition of a patient is encoded in the patterns of image intensity.

Radiomics features play a crucial role in constructing predictive models for diagnostics and prognostics, significantly contributing to improving clinical decision-making. Numerous studies underscore the substantial value of radiomics in lung cancer screening, diagnosis, and predicting

[1] Farina, Benito, et al. "Image and feature harmonization in radiomics for mitigation of heterogeneity in CT image parameters: impact on a longitudinal study of advanced NSCLC patients treated with anti-PD-1/PD-L1 immunotherapy." to be submitted to Computer Methods and Programs in Biomedicine.

treatment responses (Thawani et al., 2018). Globally, lung cancer stands as the second leading cause of death and the foremost cause of cancer-related death in both men and women (Siegel et al., 2020). Lung cancer survival rates vary depending on several factors, including the stage of cancer at diagnosis, the specific type of lung cancer, and individual patient characteristics. Unfortunately, diagnosis often occurs at an advanced stage, limiting treatment options. Despite improved 5-year survival rates since the mid-1970s, it remains relatively low, ranging from approximately 10% to 20% in most countries (Allemani et al., 2018). In recent developments, immune checkpoint inhibition therapy, targeting PD-1 or PD-L1 pathways, has emerged as a groundbreaking approach for lung cancer management. By leveraging the body's immune system to combat cancer cells, immunotherapy has demonstrated promising outcomes, including enhanced survival rates and durable responses, surpassing conventional treatments like chemotherapy or radiation therapy, either alone or in combination (Kanwal et al., 2018). However, a critical challenge lies in the lack of a reliable biomarker for predicting tumor response to immunotherapy. To avoid potential drawbacks such as toxicity or hyperprogression, identifying additional predictive biomarkers becomes essential for better stratifying patient groups likely to respond to treatment.

The limited integration of radiomics frameworks into clinical practice is attributed to the sensitivity of radiomics features to technical variations, affecting their reproducibility (Califf, 2018; B. Zhao et al., 2016). Variability across the radiomics workflow, including image acquisition, reconstruction, segmentation, and feature calculation, influences the robustness of radiomics features. Distinguishing quantitative changes in radiomics due to biological variations from those caused by image parameter heterogeneity, termed batch effects, is a key challenge. Overcoming this challenge is essential for clinician confidence and the adoption of radiomics tools in clinical decision-making.

Moreover, to develop generalizable algorithms, radiomics studies need to be constructed and validated using large image datasets, which require the collection of data from multiple sites. While a large and varied sample can potentially yield more robust inferences and enhance the statistical relevance of results, it also introduces potential variation in population demographic, imaging protocols, and acquisition methodologies (Kumar et al., 2012). Due to the sensitivity of radiomics features to these variations, a feature that proves valuable in one dataset may lose its significance in another, potentially compromising the quality of results.

To alleviate these discrepancies, various harmonization methods are employed, including the standardization of image acquisition and the implementation of post-acquisition harmonization. Although guidelines promoting standardized image acquisition have been proposed, their adoption faces constraints due to difficulty in implementation across diverse institutions (Clarke et al., 2014). Despite adherence to such guidelines may potentially mitigate the variability in data, its effectiveness is limited in radiomics studies where images are extracted using varied protocols and scanners. Moreover, even when the acquisition protocol and scanners are consistent, variations in doses may still occur. Furthermore, in many retrospective studies this guidelines might not have been followed. Many initiative have also recognized the necessity for radiomics standardization to enhance repeatability. For instance, the Image Biomarker Standardization Initiative (IBSI) (Zwanenburg et al., 2020) has proposed guidelines for standardizing radiomics workflow, including feature extraction (using IBSI-compliant software like Pyradiomics) and image preprocessing, to guarantee reproducibility. However, few studies follow this guidelines (Brancato et al., 2022).

Therefore, post-acquisition harmonization is usually needed to ensure the consistency and reliability of radiomics analysis. This becomes particularly crucial when handling data sourced from diverse origins, including different institutions, where variations in imaging protocols, scanners, and doses can introduce significant heterogeneity. Harmonization in the image domain involves adjusting image intensities to ensure uniformity in image properties, aiming to eliminate variations caused by technical differences in image acquisition and noise. Various approaches have been proposed for harmonization in the image domain, including image resampling to homogenize image resolution across multiple datasets (Van Timmeren et al., 2020), image normalization to standardize signal intensities across multiple scans (Larue et al., 2017), and image denoising and calibration to remove noise sources in the image signal and calibrate image intensities based on known references such as the intensity in the trachea. The primary goal of the latter approach is to mitigate the impact of spatially variant noise and biases caused by different acquisition conditions (Vegas-Sánchez-Ferrero et al., 2017, 2018, 2019, 2024).

In the feature domain, harmonization involves identifying sources of radiomics variability not related to biological aspects. The correction of these batch effects is designed to eliminate such sources of variation, with the primary objective of enhancing the biological signal. The main goal is to reduce the signal variance caused by variability between batches. This correction process is crucial for improving the biological signal and involves removing biases associated with various factors in the image acquisition process (Mali et al., 2021). Different batch effect correction techniques have been explored and developed for feature harmonization. These techniques aim to mitigate variations in features unrelated to biological differences, thereby enhancing feature reproducibility. This can be achieved through preprocessing steps, such as resampling images to a fixed resolution, and/or post-processing methods, such as features harmonization techniques like ComBat (Johnson et al., 2007).

Many studies have investigated the effects of different preprocessing methods, such as image resampling, kernel normalization, and image discretization. Additionally, post-processing methods such as singular value decomposition (SVD)-based batch effect removal or ComBat have been investigated, either individually or in combination with preprocessing methods (Alter et al., 2000; L. He et al., 2016; Ibrahim et al., 2021; Ligero et al., 2021; Refaee et al., 2022; Singh, Horng, Chitalia, et al., 2022).

Specifically, Ligero et al. (2021) investigated sources of radiomics variability related to image-acquisition parameters and aimed to mitigate this variability by employing a combination of image resampling and post-acquisition correction methods like SVD and ComBat. Their study revealed that voxel resampling effectively reduced variability caused by acquisition voxel size, while ComBat successfully addressed differences in convolutional kernel and slice thickness across datasets. Similarly, Singh, Horng, Roshkovan, et al. (2022) proposed image resampling to mitigate variability in voxel spacing, along with a nestedCombat technique to address acquisition parameters such as contrast enhancement and CT reconstruction kernel differences. Their research demonstrated that harmonization efforts led to improved prognostic performance of radiomics features, resulting in enhanced survival prediction for patients with lung cancer undergoing first-line immunotherapy.

The present chapter aimed to investigate the impact of image and feature harmonization on radiomics features extracted from CT images of patients with advanced non-small cell lung cancer

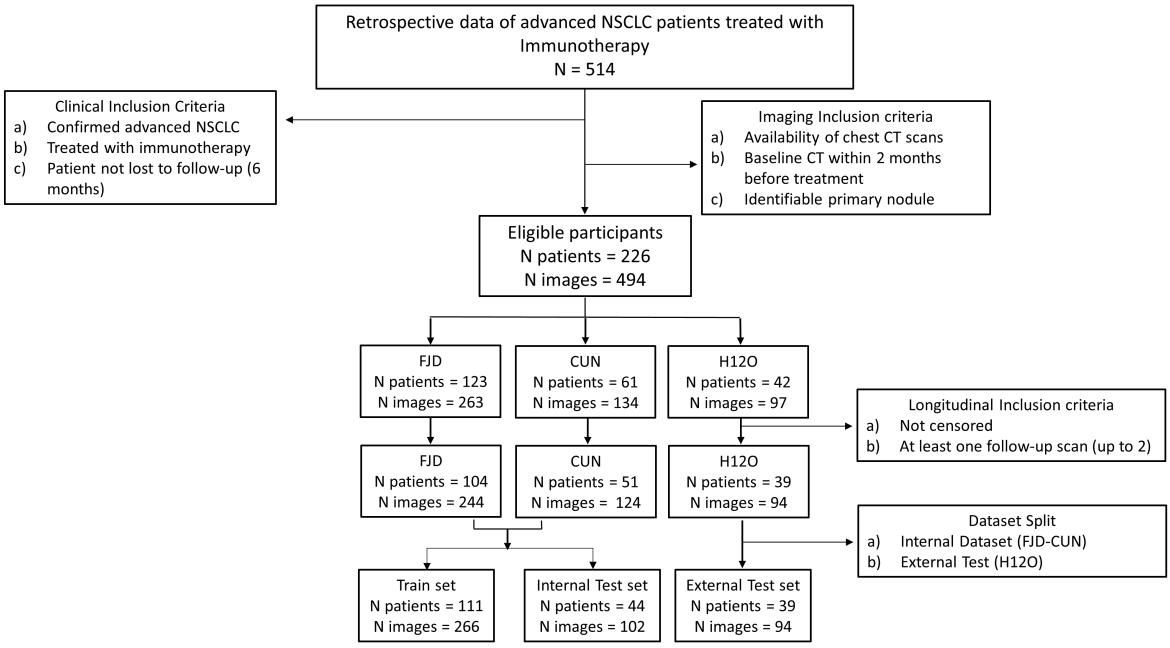


Figure 4.1: Flowchart of the patients included in the analysis considering the inclusion and exclusion criteria. Details of the number of patients in the training and internal and external test sets are provided.

(NSCLC) undergoing immunotherapy across three different institutions. We aim to assess the individual effectiveness of each harmonization technique in mitigating batch effects associated with acquisition parameters such as manufacturer, kVp, noise variance, and slice thickness. We also aim to explore the combined use of these techniques. Additionally, we compare their performance in predicting treatment response with scenarios where no harmonization is applied. Longitudinal models will be developed using radiomics data extracted from CT images acquired at baseline and during follow-up of immunotherapy for up to three time points per patient. Ensuring the comparability of radiomics features across different time points is particularly crucial in this context. Previous research by Plautz et al. (2019) demonstrated that changes observed in features extracted from longitudinal CT scans, acquired under consistent manufacturer and protocol conditions, were stable and reproducible. However, in our scenario, both manufacturer and protocol can vary over time, with changes occurring at least once during the follow-up period for approximately 30% of the patients. To the best of our knowledge, no prior research has systematically examined the concurrent impact of both image and feature harmonization on CT images and longitudinal data, particularly in the context of predicting immunotherapy outcomes using longitudinal data.

4.2 Materials and Methods

This study received approval from the institutional ethical review boards of all participating hospitals. Written informed consent was obtained from all included patients.

	FJD	CUN	H12O	Total
Responders	56	26	26	108
Non-responders	48	25	13	86
Total	104	51	39	194

Table 4.1: Summary of patient distribution across FJD, CUN, and H12O.

4.2.1 Dataset

We retrospectively collected a multicohort of 514 patients with pathologically confirmed NSCLC treated with anti-PD-1/PD-L1 monoclonal antibodies from January 2013 to December 2023. The data were gathered from three distinct institutions: Hospital Universitario Fundación Jiménez Díaz (FJD, 295 patients), Clínica Universidad de Navarra (CUN, 174 patients), and Hospital Universitario 12 de Octubre (H12O, 121 patients).

The inclusion criteria for this study comprised: (a) confirmed stage III-IV NSCLC, excluding the neuroendocrine subtype due to its unique treatment response (Garcia-Alvarez et al., 2022); (b) patients treated with immunotherapy, either as monotherapy, in combination with immune-based agents, or in combination with traditional (e.g. chemotherapy, radiation therapy) or mutation-driven treatments (e.g. tyrosine kinase inhibitors); (c) availability of at least baseline clinical and epidemiological information; (d) absence of right-censored patient data. Finally, 424 patients met these criteria and were included in the study.

To identify patients with relevant imaging data, the institutional medical records systems were searched. Inclusion criteria for radiomics analysis were: (a) availability of chest CT scans; (b) availability of baseline CT within 2 months before the start of immunotherapy. Exclusion criteria were defined as: (a) lung resection during treatment; (b) inability of an experienced radiologist to detect and segment the primary tumor in the baseline CT; (c) CT performed without contrast injection; (d) poor image quality. Consequently, 226 patients were enrolled for imaging data analysis as shown in Figure 4.1.

As additional inclusion criteria, patients who did not have at least one follow-up CT scan after starting treatment were excluded from our longitudinal analysis. For each patient, we considered up to two follow-up CT scans acquired up to 4 months after treatment. Additionally, censored data were excluded from subsequent analyses.

The selection of endpoints in this study aligns with the previous chapter. Responders to immunotherapy were identified as patients exhibiting a durable clinical benefit for at least 6 months following the start of treatment. Durable benefit was assessed through progression-free survival (PFS), measuring the time from the initiation of immunotherapy to either death, disease progression, or the last follow-up. Disease progression was determined based on the patient's overall clinical status and iRECIST criteria derived from imaging assessments. Patients with censored data at 6 months after treatment initiation were excluded. The study's maximum follow-up period extended to 48 months. Additionally, overall survival (OS) was considered, defined as the duration in months from the initiation of immunotherapy to death or censoring at the last follow-up for survivors.

	Train		Independent Internal Test		Independent External Test
	FJD	CUN	FJD	CUN	H12O
Responders	59		22		26
	40	16	16	7	
Non-responders	52		22		13
	33	19	15	6	

Table 4.2: Summary of responders and non-responders across train and internal (FJD, CUN) and external (H12O) independent test sets (H12O).

		Missing	Overall	Internal Cohort	External Cohort	P-Value (adjusted)
n images			462	94	368	
Manufacturer, n (%)	GE Medical Systems	0	7 (1.5)	1 (1.1)	6 (1.6)	<0.001
	Philips		140 (30.3)	48 (51.1)	92 (25.0)	
	Siemens		287 (62.1)	45 (47.9)	242 (65.8)	
	Toshiba		28 (6.1)		28 (7.6)	
Exposure (mAs), mean (SD)		0	177.7 (73.6)	120.6 (77.2)	192.2 (65.2)	<0.001
kVp (kV), mean (SD)		0	118.5 (14.5)	137.9 (6.7)	113.6 (11.6)	<0.001
Pixel Spacing (mm), mean (SD)		0	0.7 (0.1)	0.8 (0.1)	0.7 (0.1)	<0.001
Slice Thickness (mm), mean (SD)		0	2.3 (1.3)	3.0 (0.2)	2.2 (1.4)	<0.001
Exposure Time (ms), mean (SD)		28	539.1 (130.1)	779.0 (162.1)	496.0 (55.8)	<0.001
stdNoise, mean (SD)		0	26.2 (4.8)	24.3 (4.2)	26.7 (4.8)	<0.001

Table 4.3: CT image acquisition and reconstruction parameters across for images from the internal and external datasets. Adjusted p-values (using Bonferroni correction) were calculated using two-sample T-tests for continuous variables and Chi-square tests for categorical variables. SD represents the standard deviation.

Table 4.1 provides an overview of the patients' distribution across the three institutions.

Following the methodology outlined in the previous chapter, patients from CUN and FJD were divided into a training and an internal test set (25%), considering the balance in the availability of baseline and follow-up data, institution, and endpoint. This resulted in an independent internal test set comprising 44 patients. Additionally, 39 patients from H12O were used as an independent external test set. Additional information about patient split can be found in Table 4.2. The training set was utilized for model development, while both the internal and external test sets were employed to evaluate the performance of the implemented models.

4.2.2 Data Acquisition

A baseline CT scan was acquired for each patient within 2 months before starting immunotherapy. Follow-up CT scans were acquired within 4 months post-treatment, allowing for up to three temporal assessments per patient. CT images were acquired after contrast injection during a patient inspiratory breath hold, following the contrast-enhanced CT chest protocol. CT scans were reconstructed using a standard kernel. The scans were conducted in the axial plane, with tube voltage ranging from 80 to 140 kVp, slice thickness ranging from 0.6 to 5 mm, and pixel

spacing ranging from 0.4 to 1.5 mm. Further details regarding the CT manufacturers' models from which the images were extracted can be found in 6.3 Table B4. Additional details about the CT parameters can be found in Table 4.3. Notably, significant differences were observed in all metadata between the internal and external cohorts. However, no differences were found between responders and non-responders in internal and external cohorts as shown in Appendix B Tables B5 and B6, respectively. Conversely, significant differences in metadata were noted between the internal and external test sets, as illustrated in Appendix B Table B7.

Primary tumors were delineated by experienced radiologists at each participating hospital, utilizing either 3DSlicer or institution-specific software. Each tumor was segmented at baseline and follow-up, prioritizing the largest lesion in cases of ambiguous primary tumors. Follow-up CT scans were excluded if the tumor identified in the baseline CT scan was no longer visible.

To standardize voxel spacing, images were resampled to isotropic voxels with the minimum in-plane pixel spacing prior to feature extraction, utilizing bilinear interpolation. Segmentation was resampled accordingly using a nearest neighbor interpolation method. No resampling was performed if the z-spacing was smaller than the in-plane spacing. This was essential to fulfill the fundamental assumption of the image harmonization method, which requires isotropic or quasi-isotropic voxels. Failure to satisfy this assumption resulted in the occurrence of the partial volume effect, where a single voxel may contain a mix of tissue types, potentially distorting image features. Resampling ensured uniform voxel dimensions in all directions, thereby preserving spatial consistency and mitigating the risk of partial volume effects.

4.2.3 Image Harmonization

To mitigate the confounding effects arising from variations in scanner manufacturer and models and variations in tube voltage, we implement a two-step harmonization methodology, as introduced in Vegas-Sánchez-Ferrero et al. (2017). This methodology comprises a noise stabilization stage which aims at estimating the noiseless signal and the spatially-variant noise standard deviation, thereby identifying the specific noise pattern affecting the signal. The primary objective is to transform the non-stationary noise into a stationary Gaussian process, facilitating the comparison of images acquired with different doses and reconstruction kernels. These outcomes can then be combined to produce a new CT image with the desired characteristics. To ensure the preservation of structural details, one can integrate the noise with a stabilized noise standard deviation, ensuring homogeneous characteristics throughout the image.

The second stage of the harmonization methodology involves autocalibration (Vegas-Sánchez-Ferrero et al., 2019). Here, a functional relationship between noise variance and the noise-induced bias affecting the signal is utilized to estimate the bias based on the results obtained during the noise stabilization stage. The autocalibration process effectively eliminates the bias introduced by the spatially-variant noise and aligns density values of well-defined structures, such as the trachea, with their reference attenuation values (e.g., -1000 HU for the trachea). Importantly, this methodology do not require any phantom.

In the original methodology, the calibration was done with two reference density levels (blood in the aorta and air in the trachea) assumed to have relatively homogeneous attenuation levels.

However, due to the observed density variability in blood introduced by contrast administration, we opted to calibrate with respect to the tumor rather than the aorta. The tumor reference value was calculated as the median value of the mode distribution of tumor intensities across the entire population. Additionally, recognizing that resolution variability along the z-axis introduces errors in noise estimation that subsequently impact bias estimation due to noise, we decided to discard the estimation of spatially-variant bias, assuming that the bias in the tumor would not be significantly affected due to its high density. This is a reasonable assumption, as bias due to noise declines as tissue density increases Vegas-Sánchez-Ferrero et al. (2018). In conclusion, for calibration, we employed a simplified scheme consisting of linear interpolation that transforms densities linearly with two reference density levels (trachea and tumor) common to the entire cohort.

The overall outcome of this harmonization methodology is a low-noise, calibrated CT scan that remains consistent across various scanner parameters that may otherwise influence CT density values.

4.2.4 Feature Extraction and Preprocessing

Radiomics features were derived from segmented tumors in both the original and harmonized images using the Pyradiomics package, following the IBSI guidelines (Van Griethuysen et al., 2017). Before extraction, image intensity values were clipped within the range of [-1000, 400] HU and discretized with a fixed bin width of 25 (Larue et al., 2017). Given the varying slice thickness across datasets, with a median slice thickness of 3 mm, it was decided to use only 2D filters for radiomics feature extraction in order to minimize the impact of low resolution along the z-axis on a portion of the data.

Radiomics features with variance lower than 0.001, indicating quasi-constant values, were identified in the training set and subsequently removed to enhance the robustness of the analysis. Additionally, features were standardized to have zero-mean and unit-variance. The transformation was learned during training and subsequently applied to the test sets.

Feature Stability

To implement a robust model, it is crucial to assess the robustness of radiomics features, aiming for models that generalize well across diverse datasets from various institutions. In this study, we investigated the reproducibility and repeatability of radiomics features. Reproducible features show limited variation under consistent image protocols and acquisition parameters, whereas repeatable features maintain consistency despite slight variations in image segmentation. The choice of segmentation software and the expertise of the reader can significantly influence the segmentation process. Tumor delineation, susceptible to inter-observer variability, can notably impact specific radiomics features even with slight segmentation changes. This challenge is more pronounced in complex cases, such as advanced lung tumors, where distinguishing the tumor from surrounding tissues is not always clear (Mercieca et al., 2021).

Stable features are both reproducible and repeatable. Identifying stable features is essential for

constructing high-fidelity models, preventing unreliable findings and ensuring model repeatability. Stable features are more likely to contain clinically relevant information and maintain their discriminating power across different institutions. While this method may lead to a potential loss of clinically relevant information, it's essential to note that not all information may be relevant and transferable from one dataset to another.

We assessed feature reproducibility using a test-retest analysis on the Reference Image Database to Evaluate Therapy Response (RIDER) dataset (Armato III et al., 2008). This dataset, designed to establish a consensus on data collection and analysis for quantitative lung cancer imaging, comprised 32 patients who underwent two chest CT scans. These scans were acquired 15 minutes apart, employing the same image protocol, scanner, and acquisition and processing parameters. Any variability between the test and retest scans was solely attributed to non-biological and non-batch-effects factors, such as patient orientation, respiration, and overall patient movements.

We also evaluated feature repeatability using two datasets: the QIN Lung CT Segmentation dataset (Kalpathy-Cramer et al., 2016) and a random subset of our internal cohort. The first dataset aimed to compare bias and repeatability among automatic, semi-automatic, and manual segmentations for 30 lung CT studies. We considered two automatic segmentations with two different algorithms for each nodule. In the second dataset, an experienced radiologist from FJD independently performed semi-automatic segmentation on 26 tumors. The segmentation process was performed twice using two different modules of syngo.via software, resulting in an initial segmentation that was subsequently refined by the radiologist.

To evaluate the repeatability and reproducibility of features, we employed Lin's concordance correlation coefficient (CCC) proposed by Lawrence and Lin (1989). Features with a CCC of 0.85 or higher for both tests were considered stable and included in further analysis, while those with a CCC below 0.85 were excluded. Reproducible and repeatable features are expected to demonstrate higher robustness against variations in image acquisition conditions and segmentation, thereby enhancing their reliability for subsequent analyses.

All datasets were resampled to isotropic voxels with the minimum in-plane pixel spacing to ensure consistency with the imaging conditions of the immunotherapy study dataset.

4.2.5 Feature Harmonization

We explored significant sources of radiomics variability attributed to technical parameters, commonly referred to as batch effects. These batch effects typically arise from variations in acquisition parameters, protocols, acquisition site, scanner configuration, and imaging reconstruction techniques, among other factors. Radiomics feature harmonization involves correcting for batch effects to eliminate bias associated with these sources of variation.

In this study, we identified batch effects related to specific technical parameters, including slice thickness, kVp, and manufacturer. Additionally, we identified a novel batch effect concerning the variance of noise within each image, quantified as the standard deviation of noise in the trachea, denoted as stdNoise. Trachea segmentations were generated using the TotalSegmentator method proposed by Wasserthal et al. (2023).

To mitigate this batch effects, while retaining informative covariates and eliminating residual heterogeneity, we applied the ComBat correction only to the features identified as stable and repeatable following the description of previous section.

ComBat Definition

ComBat is a statistical harmonization method originally introduced in genomics by Johnson et al. (2007) and then applied to imaging data by Fortin et al. (2017). It has been designed to correct variation in imaging features due to imaging parameters. Many studies in radiomics have demonstrated that this method can harmonize features extracted from different CT protocols and reduce the number of features with significantly different distributions due to batch effects. ComBat uses an empirical Bayes linear model framework to estimate location and scale parameters to shift data.

Let y_{ijv} be the feature vectors of observed data, with $i = 1, 2, ..., M$, $j = 1, 2, ..., n_i$, $v = 1, 2, ..., V$, where i denotes the indexes of the batch effect, j the indexes of the subjects within the batch effect i, n_i the number of subjects within the batch effect i, and V the number of features. ComBat assumes that features are affected by additive and multiplicative effects that depend on a batch effect, following this equation:

$$y_{ijv} = \alpha_v + \boldsymbol{X}_{ij}^T \beta_v + \gamma_{iv} + \delta_{iv}\epsilon_{ijv} \tag{4.1}$$

where α_v is the overall, grand mean, $\boldsymbol{X}_{ij}^T$ is the design matrix of biological covariates, β_v is the vector of regression coefficient, γ_{iv} is the mean batch effect, and δ_{iv} is the variance batch effect. The errors ϵ_{ijv} are assumed to follow $e_{ijv} \sim N(0, \sigma_v^2)$. Assuming that the batch effect follow the same distribution across all features, the mean batch effect γ_{iv} follows an independent normal distribution and the variance batch effect δ_{iv} follows an independent inverse gamma distribution, they can be estimated using the empirical Bayes method via method of moments using data across all features as the means of the posterior distributions. The ComBat harmonized features are then obtained as:

$$y_{ijv}^{ComBat} = \frac{y_{ijv} - \hat{\alpha}_v - \boldsymbol{X}_{ij}^T \hat{\beta}_v - \gamma_{iv}^*}{\delta_{iv}^*} + \hat{\alpha}_v + \boldsymbol{X}_{ij}^T \hat{\beta}_v \tag{4.2}$$

ComBat has several limitations that should be considered, which include:

- ComBat assumes errors from standardized input features to be normally distributed, an assumption that is violated in case of multimodal distributions.
- ComBat requires all batch effects and clinical covariates to be known.
- Since ComBat is a data-driven method, the data shift is specific to input data. If a new batch effect is introduced in data, the shift has to be recalculated.

Batch effects	Category	N samples (%)	
		Internal Cohort	External Cohort
Manufacturer	Siemens	66%	48%
	Philips	24%	51%
	Toshiba	7%	0%
	GE Medical Systems	2%	1%
kVp (kV)	<= 100	31%	1%
	>100	69%	99%
Slice Thickness (mm)	<=1	30%	1%
	1<x<=2	38%	0%
	2<x<=3	14%	99%
	>3	18%	0%
stdNoise	<=26	52%	66%
	>26	48%	34%

Table 4.4: List of batch effects and their respective categories, along with the percentage of images from the external and internal cohorts corresponding to each category.

ComBat Implementation

To address potential sources of variability, ComBat correction was applied individually to predefined batches.

ComBat cannot handle continuous batch effects, so they were encoded according to the categories outlined in Table 4.4. Batch effects were categorized in order to comply with ComBat's recommendation of having at least 20-30 samples per batch (Orlhac et al., 2022), even tough it has shown robustness with smaller sample sizes (Da-Ano, Visvikis, & Hatt, 2020). Although one of the manufacturer categories had low representation, with only 2% of cases (6 samples) scanned with a GE scanner, these cases were retained instead of being discarded. For the stdNoise, the threshold was determined based on the median value to ensure adequate representation for both high and low noise levels.

A OPNested ComBat approach was employed to consider multiple batch effects simultaneously, facilitating sequential harmonization (Horng, Singh, Yousefi, Cohen, Haghighi, Katz, Noël, Kontos, & Shinohara, 2022; Horng, Singh, Yousefi, Cohen, Haghighi, Katz, Noël, Shinohara, & Kontos, 2022). Two main combinations were examined: NestedComBat1 (manufacturer, kVp, stdNoise) and NestedComBat2 (manufacturer, stdNoise, slice thickness). While other combinations were explored, these were prioritized for their significance. During ComBat, the outcome variable (response to treatment) and clinical variables (gender and treatment type) were protected to retain biological variability.

An Anderson-Darling (AD) test evaluated distributional differences for each batch effect. A non-significant p-value indicates well-harmonized features (Anderson & Darling, 1954). It is important to notice that only features demonstrating significant differences in distribution for a specific batch effect underwent harmonization procedures.

In OPNested ComBat, the harmonization order was determined by iterating through all possible

permutations of batch effects. The best order was chosen as the one minimizing the number of features with significant differences in distribution across at least one batch effect, assessed through the AD test. Features persistently affected by batch effects after harmonization were considered non-robust and excluded from subsequent analysis.

Additionally, traditional ComBat centers data to the overall mean of all samples, potentially resulting in a loss of the original physical feature meaning due to arbitrary shifts. To address this, a modification of ComBat, M-ComBat, was applied, ensuring data was centered to the location and scale of a predetermined reference batch (Da-Ano, Masson, et al., 2020). The reference batch for each batch effect was selected based on the most frequent batch value, such as the most common scanner in the case of the manufacturer.

To define the features harmonization we used all the images from the patients in the training set to calculate ComBat estimates assuming that these data is representative of the variability that could be found in the test data. These estimates were saved post-training for application to the internal and external test sets. In previous literature, harmonization techniques have been applied to all available data, including both training and test sets. Although this approach may offer better results, we aimed to test a more practical strategy in which there is no need to redefine the harmonization transformation to study new cases as it would be in the real-world clinical scenario. We need to ensure that all the variables that should to be considered are taken into account in the Combat estimation. Notably, if a new variable is introduced in the test set data, applying the ComBat method is not feasible. For example, we chose not to incorporate the site as batch effect, a common practice in the literature.

4.2.6 Immunotherapy Response Prediction

To evaluate the effectiveness of the proposed data harmonization techniques, we investigated their ability to predict immunotherapy response defined as the durable clinical benefit at 6 months after the start of the treatment. Various types of inputs were considered for constructing the predictive models, including radiomics features extracted from original, stabilized and harmonized data, and radiomics features post-processed with ComBat harmonization.

Feature selection was implemented to enhance the model's performance by reducing the number of features and improving interpretability. Initially, the Pearson correlation coefficient was calculated to identify features with high-collinearity, where correlations exceeding 95% were considered indicative of strong redundancy. Subsequently, only one feature from each highly correlated pair was retained for further analysis, thereby reducing multi-collinearity. Next, a Random Forest (RF) feature selector was utilized to further reduce the feature space to the top 20 most informative features. More in detail, to optimize RF hyperparameters, a Bayesian optimization approach was applied using 3-fold stratified cross-validation. This optimization aimed to find the best combination of hyperparameters including:

- Number of estimators: the number of decision trees in the forest.
- Maximum depth: the maximum depth of each decision tree.
- Minimum samples split: the minimum number of samples required to split an internal node.

- Minimum samples leaf: the minimum number of samples required to be at a leaf node.
- Maximum number of features: the maximum number of features to consider when looking for the best split.

The resulting 20 most significant features, along with the best hyperparameters, were selected to construct the final RF model.

A RF-longitudinal model incorporating both baseline and follow-up data was implemented. Patients without follow-up information were excluded from this analysis. The input of RF-longitudinal model comprised the concatenation of all available features across different time points for each patient, resulting in an expanded feature set (the number of features multiplied by the number of time points). Missing data points were imputed using the most recent available data.

4.2.7 Statistical Analysis

The statistical analysis, feature extraction, harmonization, and model implementation were conducted using Python (version 3.9). A significance level of 0.05 (two-sided tests) was used.

The Anderson-Darling test was employed to detect differences in distribution associated with batch effects. This test was chosen due to its ability to handle the multi-modality of the same feature distribution and its capability to test for differences among more than two batches associated with a specific batch effect. A feature was considered to be successfully harmonized if the associated test yielded a p-value greater than 0.05, indicating no significant difference between the distributions associated with each batch variable. Additionally, principal component analysis (PCA) was performed to further evaluate the impact of image and feature harmonization.

The performance of the models was evaluated using stratified 3-fold cross-validation in the training set. The best model was selected based on the highest accuracy among the three folds. Model performance was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC), along with other metrics including accuracy (ACC), specificity (SPEC), sensitivity (SENS), balanced accuracy (bACC), and precision (PREC). These metrics provide a comprehensive evaluation of the model's performance. 95% confidence interval (CI) for all metrics were estimated using a bootstrap resampling approach with 1000 iterations. Differences between ROC curves were evaluated using the DeLong test. Models were compared based on whether the features were extracted from harmonized or non-harmonized images and whether the features were subsequently harmonized using ComBat harmonization.

Kaplan–Meier survival analysis was conducted to stratify patients in low- and high- risk of progression based on the model's predictions (threshold $= 0.5$). The significance of differences between survival curves was evaluated using the log-rank test. A p-value less than 0.05 indicated a statistically significant separation between patients with low- and high- risk of both progression-free survival and overall survival.

4.3 Results

4.3.1 Patients Characteristics

In our study, 158 patients from the internal dataset (FJD and CUN) and 39 from the external dataset (H12O) were identified for longitudinal analysis. Notably, within the internal dataset, three patients were censored at the six-month follow-up and subsequently excluded from the analysis, resulting in 155 patients available for the longitudinal analysis. Conversely, no patients were censored in the external validation cohort. A detailed comparison between demographics and clinical characteristics of internal and external datasets is provided in Table 4.5. Comparisons between responders and non-responders for both the internal and external datasets are provided in Appendix B Tables B1 and Table B2, respectively.

As shown in Table 4.5 there are significant differences between the internal and external cohorts. These disparities encompassed various clinical variables, including treatment types (notably, H12O exclusively received monotherapy), PFS, status, tumor histology, pre-treatment ECOG scores, and steroid usage. More details about the differences between train, internal and external test sets are presented in Appendix B Table B3.

In the internal dataset comprising 155 patients, 74 were responders, while 81 were non-responders. The majority of patients received immunotherapy as monotherapy (51.6%), followed by a combination of immunotherapy with chemotherapy (24.5%). Regarding disease stage, 92.6% of patients were in stage IV, and 67.7% were male. Adenocarcinoma was the most prevalent histological variant of NSCLC (72.5%), and 92.7% of patients were current or former smokers. No significant differences (p-value < 0.05) in demographic or clinical characteristics were observed between the training and internal test set.

In the external cohort comprising 39 patients, 13 were responders, and 26 were non-responders. All patients in this cohort received immunotherapy as monotherapy. Among them, 94.9% were in stage IV, and 84.6% were male. Squamous cell carcinoma was the most prevalent histological variant of NSCLC (46.2%), and 95.5% were current or former smokers.

4.3.2 Radiomics Stability Assessment

Out of the initial set of 1379 features, only 324 (23.5%) demonstrated both reproducibility and repeatability (Table 4.6). Given the low repeatability of features, we conducted further evaluations focusing solely on the QIN dataset and a random subset of our internal cohort. As detailed in Table 4.7, only 13.3% of features exhibited reproducibility in the immunotherapy dataset, contrasting with the 56.9% observed in the QIN dataset. This disparity can be attributed to the complexity of segmenting tumors in patients with advanced NSCLC. In such cases, defining tumor boundaries is often challenging due to the presence of adjacent post-obstructive atelectasis or pneumonia. Additionally, some lung tumors manifest both solid and ground glass components, further complicating the delineation process and potentially introducing variability in boundary definition (Q. Huang et al., 2018).

		Missing	Overall	External Cohort	Internal Cohort	P-Value (adjusted)
n patients			194	39	155	
Treatment Response, n (%)	Non-responders	0	100 (51.5)	26 (66.7)	74 (47.7)	1.000
	Responders		94 (48.5)	13 (33.3)	81 (52.3)	
Treatment, n (%)	Monotherapy	0	119 (61.3)	39 (100.0)	80 (51.6)	<0.001
	Combined Immunological Agents		20 (10.3)		20 (12.9)	
	Immuno+Chemotherapy		38 (19.6)		38 (24.5)	
	Immunotherapy+Radiotherapy		16 (8.2)		16 (10.3)	
	Other		1 (0.5)		1 (0.6)	
Stage, n (%)	III	6	13 (6.9)	2 (5.1)	11 (7.4)	1.000
	IV		175 (93.1)	37 (94.9)	138 (92.6)	
Gender, n (%)	Female	0	56 (28.9)	6 (15.4)	50 (32.3)	1.000
	Male		138 (71.1)	33 (84.6)	105 (67.7)	
Age, mean (SD)		0	65.3 (9.7)	65.8 (10.4)	65.2 (9.5)	1.000
PFS (months), mean (SD)		0	10.6 (14.3)	5.8 (6.1)	11.7 (15.5)	0.005
OS (months), mean (SD)		0	16.7 (16.1)	12.4 (9.1)	17.8 (17.2)	0.160
Status, n (%)	Alive	0	73 (37.6)	5 (12.8)	68 (43.9)	0.014
	Dead		121 (62.4)	34 (87.2)	87 (56.1)	
Progression, n (%)	No	0	31 (16.0)	1 (2.6)	30 (19.4)	0.414
	Yes		163 (84.0)	38 (97.4)	125 (80.6)	
IPA, mean (SD)		12	45.5 (27.9)	46.4 (0.0)	45.3 (31.5)	1.000
Smoking habit, n (%)	Current smoker	1	44 (22.8)	9 (23.1)	35 (22.7)	1.000
	Former smoker		137 (71.0)	29 (74.4)	108 (70.1)	
	Non-smoker		12 (6.2)	1 (2.6)	11 (7.1)	
Tumor histology, n (%)	Adenocarcinoma	2	127 (66.1)	16 (41.0)	111 (72.5)	0.013
	Squamous cell carcinoma		54 (28.1)	18 (46.2)	36 (23.5)	
	SCLC		11 (5.7)	5 (12.8)	6 (3.9)	
PDL1, mean (SD)		73	0.3 (0.4)	0.1 (0.2)	0.4 (0.4)	<0.001
Surgery, n (%)	No	29	140 (84.8)	39 (100.0)	101 (80.2)	0.114
	Yes		25 (15.2)		25 (19.8)	
Pre-treatment ECOG, mean (SD)		30	0.7 (0.5)	1.0 (0.4)	0.6 (0.5)	<0.001
Weight, mean (SD)		8	70.9 (13.1)	73.6 (12.5)	70.2 (13.2)	1.000
Height, mean (SD)		12	167.7 (7.7)	166.8 (6.7)	168.0 (7.9)	1.000
Steroids, n (%)	No	17	113 (63.8)	36 (92.3)	77 (55.8)	0.001
	Yes		64 (36.2)	3 (7.7)	61 (44.2)	
Antibiotics, n (%)	No	26	134 (79.8)	32 (82.1)	102 (79.1)	1.000
	Yes		34 (20.2)	7 (17.9)	27 (20.9)	
COPD, n (%)	No	69	69 (55.2)	29 (74.4)	40 (46.5)	0.136
	Yes		56 (44.8)	10 (25.6)	46 (53.5)	

Table 4.5: Demographic and clinical characteristics of the patients in the internal and external cohorts. Adjusted p-values (using Bonferroni correction) were calculated for comparisons between the two cohorts using two-sample T-tests for continuous variables and Chi-square tests for categorical variables. SD represents the standard deviation, and Q1 and Q3 represent the first and third quarterlies, respectively.

Feature type	N features	N reproducible features (%)	N repeatable features (%)	N stable features (%)
Original	105	49 (46.7%)	15 (14.3%)	11 (10.5%)
LoG	455	322 (70.8%)	242 (53.2%)	198(43.5%)
Wavelet	364	165 (45.3%)	91 (25%)	54(14.8%)
SquareRoot	91	23 (25.2%)	9 (9.9%)	6(6.6%)
Square	91	49 (53.8%)	30 (33.0 %)	19(20.9%)
Logarithm	91	21 (23.1%)	8 (8.8%)	5(5.5%)
Exponential	91	56 (61.5%)	27 (29.7%)	14(15.4%)
All features	1379	719 (52.1%)	443 (32.1%)	324 (23.5%)

Table 4.6: Summary of stable features based on repeatability and reproducibility analyses. Stable features exhibit CCC greater than 0.85 for both analyses.

Feature Type	N features	N repeatable features (%)	
		QIN	IMMUNO
Original	105	50 (47.6)	5 (4.76)
LoG	455	325 (71.5)	103 (22.6)
Wavelet	364	220 (60.4)	43 (11.8)
SquareRoot	91	17 (18.7)	2 (2.20)
Square	91	47 (51.6)	10 (11)
Logarithm	91	17 (18.7)	2 (2.20)
Exponential	91	69 (75.5)	4 (4.4)
All features	1379	784 (56.9)	183 (13.3)

Table 4.7: Comparison of feature repeatability between the QIN dataset and a subset of our internal cohort. The analysis was performed separately for each dataset.

4.3.3 Impact of Image and Feature Harmonization

Figure 4.2 shows the results obtained through the image harmonization method that has been described above in section 4.2.3. Two images are shown for one acquisition manufacturer (Siemens) and high original inter-slice resolution (1 mm before resampling) and a different acquisition manufacturer (Philips) and low original inter-slice resolution (3 mm before resampling). In the residual images (obtained by taking the difference between the original and noise-stabilized images), a noticeable reduction in local noise variance is observed for both scenarios. However, it's important to note that residual images still contain some signal, which is undesired when estimating noise. These examples highlight the impact of image stabilization process on tumor visualization and consistency.

In Table 4.8, we present the results of feature harmonization between original, stabilized and harmonized images. The table illustrates the percentage of features exhibiting significantly different distributions attributed to each batch effect alone or in combination with others (as in the nestedComBat techniques) before and after applying feature harmonization. Notably, in the case of nestedComBat, a feature was considered batch-dependent if it varied for at least one batch effect.

The table reveals that ComBat successfully harmonizes the majority of features for original, stabilized and harmonized images, leading to an increase in the percentage of robust features. For instance,

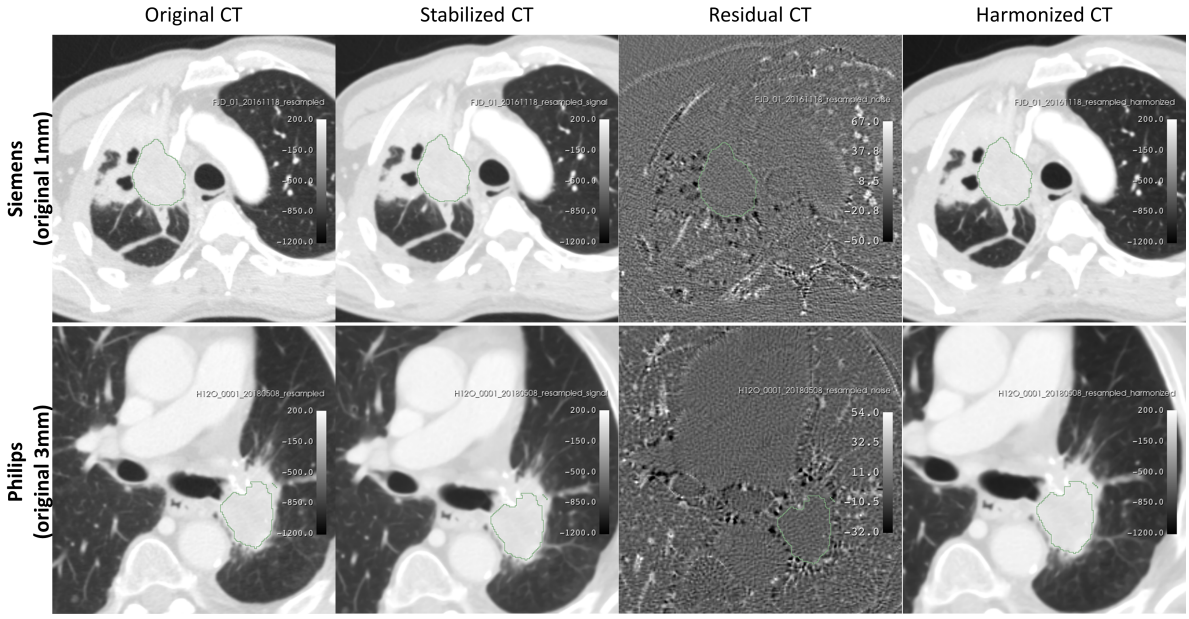


Figure 4.2: Example demonstrating the appearance of a tumor in the original image (first column), after image stabilization (second column), their residual (third column) and after the corresponding autocalibration (harmonized images - forth column). The top row displays an image acquired with a Siemens scan with initial high resolution in all dimensions, while the bottom row shows an image acquired with a Philips scan with initial low inter-slice resolution (note that images were resampled to the minimum in-plane resolution).

in the case of manufacturer batch effect alone, the percentage of robust features rose from 88% for original images to 99.97 % for stabilized images and 99.93% for harmonized images. Similarly, when considering combinations of batch effects together, the same trend is observed. For instance, in the case of original images, considering the nestedComBat1, which included manufacturer, kVp and StdNoise the proportion of robust features increased from 27.3% to 88.3%. However, as features extracted from stabilized and harmonized images were already more robust against batch effects alone or in combination, therefore the enhancement of robust features after feature harmonization was less pronounced. For stabilized images, the proportion of robust features increased from 82.1% to 99.3% and for harmonized images, it was from 81.1% to 99.0%.

Furthermore, it is evident that features extracted from stabilized and harmonized images exhibit greater robustness with respect to manufacturer variation compared to original images, with percentages increasing from 31% to 86.1% and 85.8%, respectively. However, it's important to note that the image stabilization process is affected by slice thickness, as robustness decreases from 57% for original images to 53.7% and 53.4% for stabilized and harmonized images, respectively.

It is important to note that the feature stability analysis conducted in the preceding paragraph retained only the most robust features. This may explain why the number of batch-dependent features is relatively low. Interestingly, features do not seem to depend significantly on noise levels, with the percentage of robust features being greater than or equal to 97% in both for original and stabilized images. The harmonized images appear to exhibit slightly higher dependence on noise

Data type	Harmonization type	Batch effect	N features	N batch-dependent features (%)	N features failed to harmonize (%)
Original Images	ComBat	kVp	300	21 (0.07)	2 (0.007)
Stabilized	ComBat	kVp	296	19 (0.064)	1 (0.003)
Harmonized	ComBat	kVp	296	19 (0.064)	1 (0.003)
Original Images	ComBat	Manufacturer	300	207 (0.69)	36 (0.12)
Stabilized	ComBat	Manufacturer	296	41 (0.139)	1 (0.003)
Harmonized	ComBat	Manufacturer	296	42 (0.142)	2 (0.007)
Original Images	ComBat	Slice Thickness	300	129 (0.43)	19 (0.063)
Stabilized	ComBat	Slice Thickness	296	137 (0.463)	24 (0.081)
Harmonized	ComBat	Slice Thickness	296	138 (0.466)	24 (0.081)
Original Images	ComBat	StdNoise	300	4 (0.013)	0 (0.0)
Stabilized	ComBat	StdNoise	296	8 (0.027)	0 (0.0)
Harmonized	ComBat	StdNoise	296	19 (0.064)	1 (0.003)
Original Images	NestedComBat1	Man-kVp-stdNoise	300	218 (0.727)	35 (0.117)
Stabilized	NestedComBat1	Man-kVp-stdNoise	296	53 (0.179)	2 (0.007)
Harmonized	NestedComBat1	Man-kVp-stdNoise	296	56 (0.189)	3 (0.01)
Original Images	NestedCombat2	Man-SliceThick-stdNoise	300	236 (0.787)	64 (0.213)
Stabilized	NestedCombat2	stdNoise-Man-SliceThick	296	142 (0.48)	35 (0.118)
Harmonized	NestedCombat2	Man-SliceThick-stdNoise	296	145 (0.49)	38 (0.128)

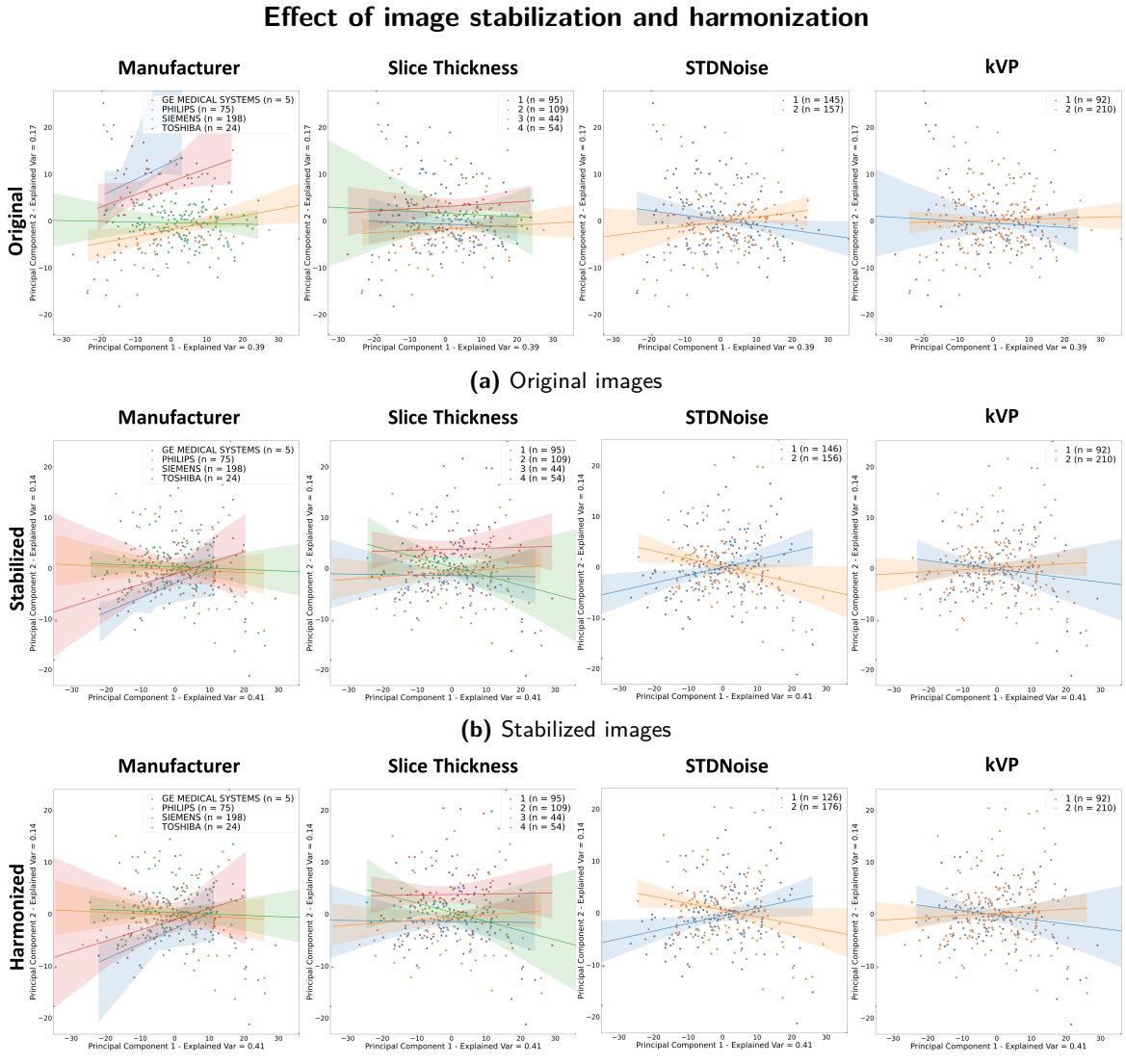
Table 4.8: Percentage of features exhibiting significantly different distributions attributed to batch effects, before and after applying ComBat, across various categories including manufacturer, slice thickness, kVp, stdNoise, and nestedComBat1 and nestedComBat2. In the case of NestedComBat harmonization, the order denotes the sequence of batch effects used in sequential harmonization for multiple batch effects.

levels, with 93.6% of the features demonstrating robustness.

A visual representation of the impact of image and feature harmonization on features extracted from original, stabilized and harmonized images in the training set is presented in Figure 4.3 and 4.5. Similar comparisons for the internal and external test sets are presented in Figures 4.4 and 4.6. The figures display the first two principal components of PCA, stratified by each batch effect, with regression lines approximating the distribution of points for each batch effect.

In Figures 4.3 and 4.4, it can be noticed that features extracted from the original image exhibit a dependency on manufacturer across both training and internal/external test sets. Specifically, features from GE Medical System and Toshiba showing slightly higher regression lines compared to the other two manufacturers (Figure 4.3a and Figure 4.4a). However, this difference is not observed in features extracted from stabilized and harmonized images, enabling a more consistent comparison between manufacturers, leaving the distributions with respect to the other batch effects largely unaltered (Figures 4.3b and 4.3c for the training set, and Figures 4.4b and 4.4c for the test sets). Interestingly, features from stabilized and harmonized images show a slight dependence on slice thickness, particularly for images with lower resolution (slice thickness greater than 3 mm), a problem that was not evident in the original features. This dependency might have been introduced during the harmonization process, specifically through the noise stabilization step. The image stabilization and harmonization was primary proposed for the context of voxel isotropy, in this case we have address this requirement resampling the data to the minimum in-plane pixel spacing, however, the local variance stabilization seems to be slightly influenced by the anisotropy.

The impact of ComBat harmonization on feature distribution is evident when comparing Figures

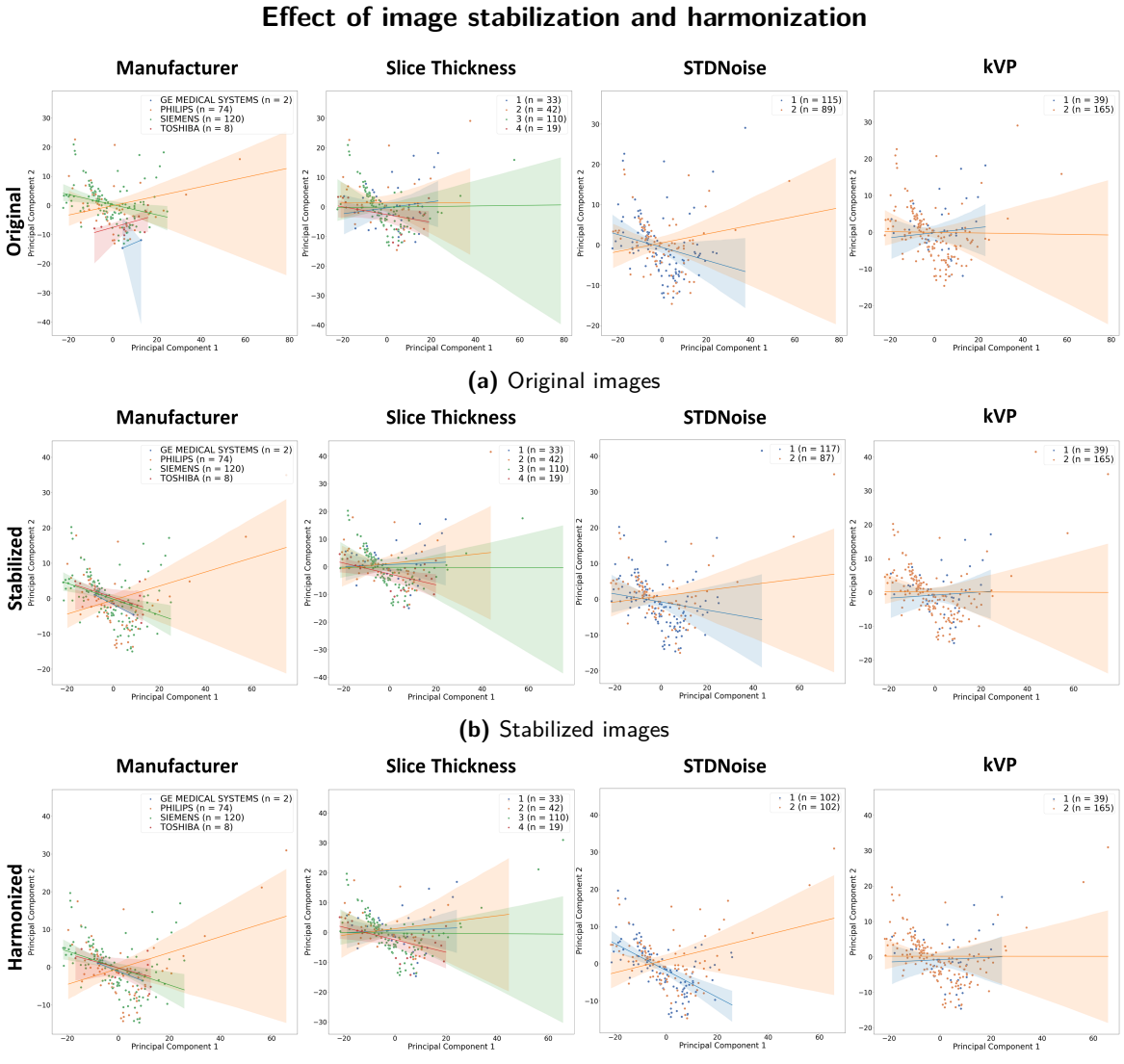


(a) Original images

(b) Stabilized images

(c) Harmonized images

Figure 4.3: Principal Component analysis in the training set for the original (a), stabilized images (b) and harmonized images (c) features for each one of the selected batch effects.

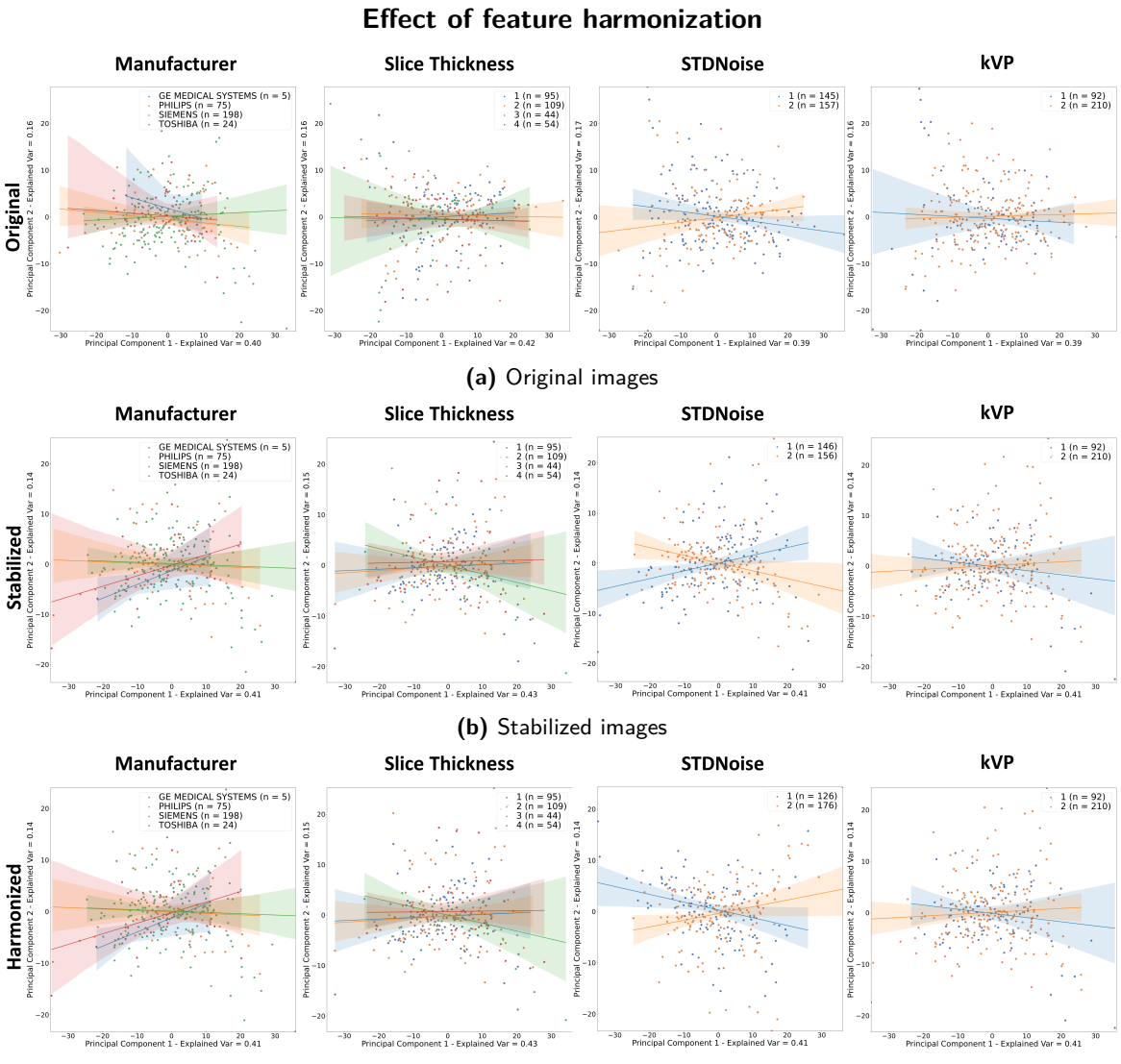


(a) Original images

(b) Stabilized images

(c) Harmonized images

Figure 4.4: Principal Component analysis in the internal and external test sets for the original (a), stabilized (b) and harmonized images (c) features for each one of the selected batch effects.

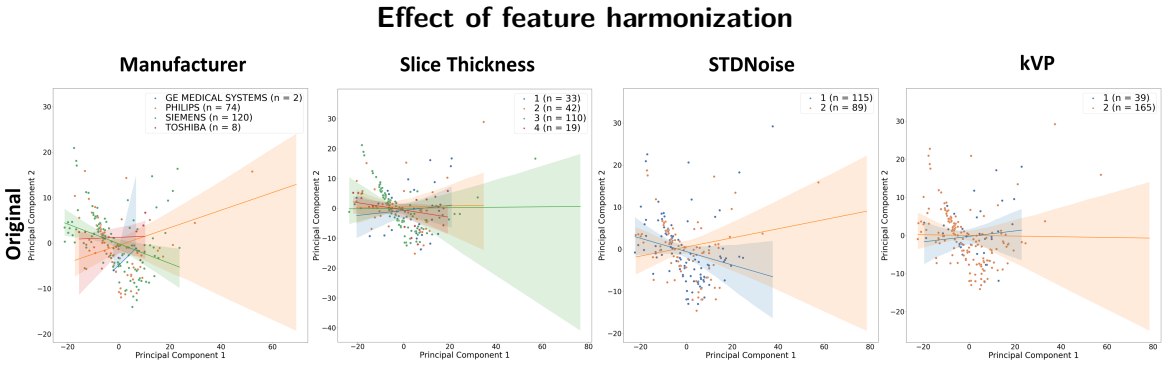


(a) Original images

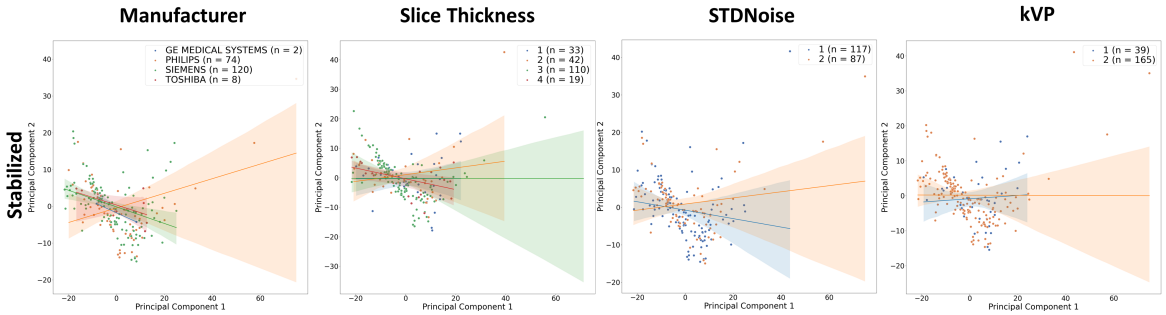
(b) Stabilized images

(c) Harmonized images

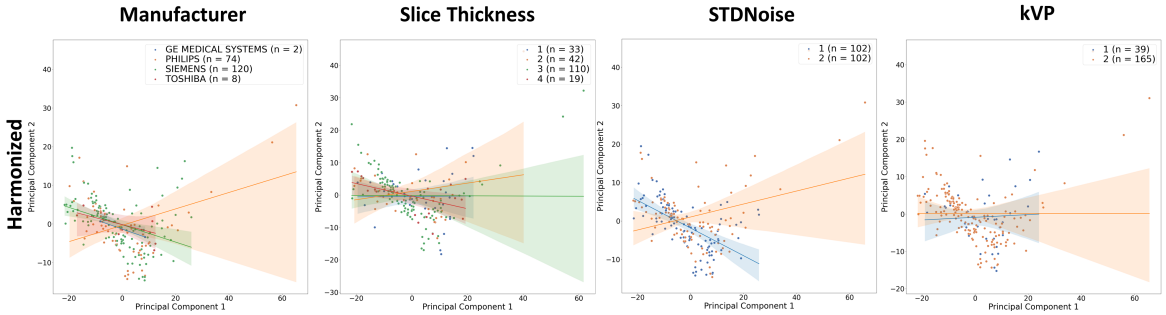
Figure 4.5: Principal Component analysis in the training set for the original (a), stabilized (b) and harmonized images (c) features after feature harmonization with Combat with respect to each one of the selected batch effects.



(a) Original images



(b) Stabilized images



(c) Harmonized images

Figure 4.6: Principal Component analysis in the internal and external test sets for the original (a), stabilized (b) and harmonized images (c) features after feature harmonization with Combat with respect to each one of the selected batch effects.

4.3 and 4.5, which illustrate a reduction of differences across the considered batch effects in the training set. Similar observations can be made for the test sets by comparing 4.4 with 4.6. After feature harmonization, the features exhibit independence from any specific batch effect, indicating successful alignment and harmonization across the dataset.

4.3.4 Immunotherapy Response Analysis

Prediction Performance

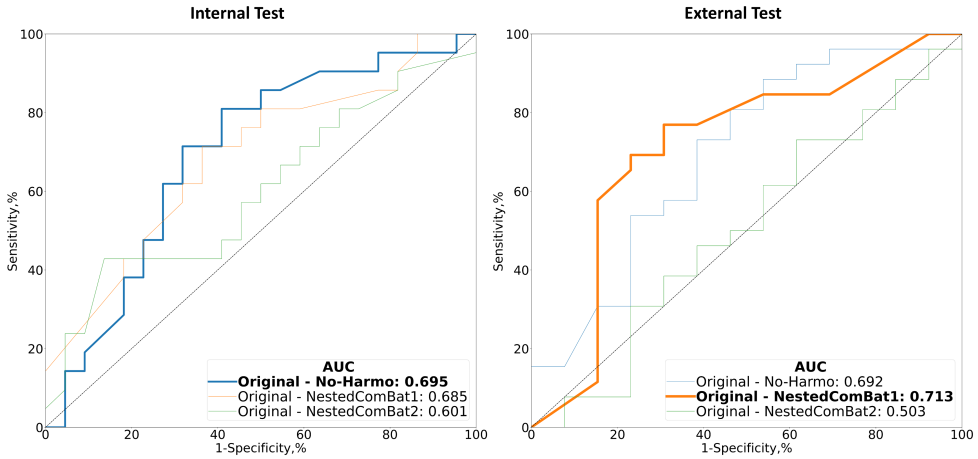
Data type	Feature Harmonization Type	Batch effect	AUC [95% CI]	ACC [95% CI]	SENS [95% CI]	SPEC [95% CI]	PREC [95% CI]	bACC [95% CI]
Original	No harmo	-	0.695 [0.529, 0.848]	0.674 [0.535, 0.814]	0.667 [0.450, 0.864]	0.682 [0.480, 0.875]	0.667 [0.450, 0.857]	0.674 [0.533, 0.813]
Stabilized	No harmo	-	0.641 [0.463, 0.801]	0.535 [0.372, 0.674]	0.571 [0.350, 0.789]	0.500 [0.286, 0.714]	0.522 [0.300, 0.720]	0.536 [0.377, 0.690]
Harmonized	No harmo	-	0.671 [0.495, 0.844]	0.721 [0.581, 0.860]	**0.762** [0.571, 0.933]	0.682 [0.476, 0.867]	0.696 [0.500, 0.875]	0.722 [0.581, 0.857]
Original	NestedCombat1	Man-kVp-stdNoise	0.685 [0.519, 0.841]	0.651 [0.512, 0.791]	0.619 [0.407, 0.833]	0.682 [0.480, 0.875]	0.650 [0.417, 0.857]	0.650 [0.508, 0.796]
Stabilized	NestedCombat1	Man-kVp-stdNoise	0.734 [0.571, 0.885]	**0.744** [0.628, 0.884]	**0.762** [0.571, 0.933]	0.727 [0.536, 0.909]	0.727 [0.524, 0.913]	0.745 [0.611, 0.877]
Harmonized	NestedCombat1	Man-kVp-stdNoise	0.665 [0.489, 0.829]	0.605 [0.465, 0.744]	0.429 [0.217, 0.647]	**0.773** [0.579, 0.941]	0.643 [0.364, 0.889]	0.643 [0.364, 0.889]
Original	NestedCombat2	Man-SliceThick-stdNoise	0.601 [0.424, 0.774]	0.512 [0.372, 0.651]	0.476 [0.269, 0.682]	0.545 [0.348, 0.760]	0.500 [0.278, 0.727]	0.511 [0.361, 0.666]
Stabilized	NestedCombat2	stdNoise-Man-SliceThick	0.684 [0.516, 0.846]	0.674 [0.535, 0.814]	0.619 [0.409, 0.826]	0.727 [0.524, 0.909]	0.684 [0.450, 0.889]	0.673 [0.528, 0.813]
Harmonized	NestedCombat2	Man-SliceThick-stdNoise	**0.768** [0.618, 0.904]	**0.744** [0.605, 0.860]	**0.762** [0.571, 0.933]	0.727 [0.542, 0.895]	**0.727** [0.526, 0.905]	**0.745** [0.612, 0.863]

Table 4.9: Comparison of response prediction performance of different longitudinal models in the internal test set based on the resulting features after various image and feature harmonization techniques. Peformance is assessed by evaluating AUC, ACC, SENS, SPEC, PREC and bACC, respectively. For each metric, the 95% confidence interval is shown and the highest value is highlighted in bold.

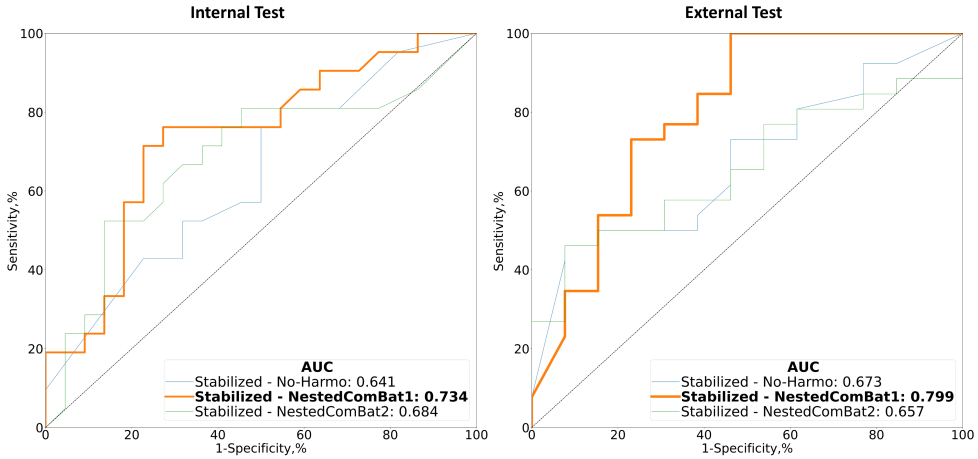
Tables 4.9 and 4.10 present the performance metrics observed in both the independent internal and external test sets, respectively. The tables detail the results obtained from longitudinal RF models, with variations in input features, both pre- and post- image and feature harmonization. Harmonization techniques such as ComBat and nestedComBat were employed across different batch effects. To enhance clarity in the presentation of results, only the nestedComBat results are displayed here. Results for ComBat are shown in Appendix B Tables B8 and B9.

Additionally, Figure 4.7 offers a summary of these results, illustrating a comparative analysis of the ROC curves for each of the presented models applied on the original, stabilized and harmonized images.

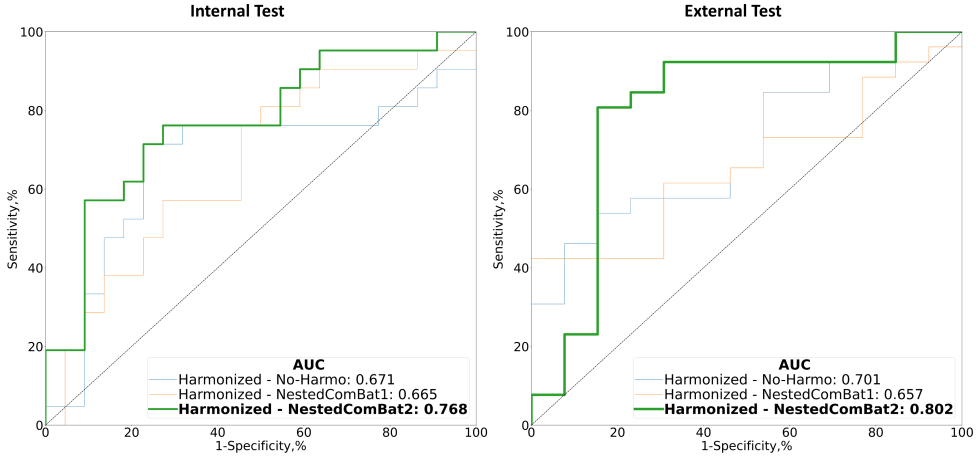
The longitudinal RF model, utilizing features extracted from stabilized images post nestedComBat harmonization—including Manufacturer, kVp, and StdNoise—achieved promising performance metrics in both the internal and external test sets. In the internal test set, it achieved an AUC of 0.743, ACC of 0.744, SENS of 0.762, and SPEC of 0.727. Similarly, in the external test set, it yielded an AUC of 0.799, ACC of 0.744, SENS of 0.731, and SPEC of 0.769. Similarly, the longitudinal RF, utilizing feautres extracted from harmonized images post nestedComBat harmonization -including manufacturer, slice thickness and StdNoise- achived an AUC of 0.768,



(a) Original images



(b) Stabilized images



(c) Harmonized images

Figure 4.7: Comparisons of ROC curves of different longitudinal models evaluated in the internal (left) and external (right) test sets. The curves represent the performance of models using original (a), stabilized (b), and harmonized (images) with and without feature harmonization. The highest AUC value is highlighted in bold.

Data type	Feature Harmonization Type	Batch effect	AUC [95% CI]	ACC [95% CI]	SENS [95% CI]	SPEC [95% CI]	PREC [95% CI]	bACC [95% CI]
Original	No harmo	-	0.692 [0.476, 0.875]	0.590[0.436, 0.744]	0.538 [0.346, 0.733]	0.692 [0.417, 0.923]	0.778 [0.562, 0.947]	0.615 [0.442, 0.769]
Stabilized	No harmo	-	0.673 [0.489, 0.840]	0.564 [0.410, 0.718]	0.500 [0.308, 0.692]	0.692 [0.412, 0.923]	0.765 [0.533, 0.947]	0.596 [0.424, 0.760]
Harmonized	No harmo	-	0.701 [0.518, 0.859]	0.641 [0.487, 0.795]	0.577 [0.400, 0.762]	0.769 [0.500, 1.000]	0.833 [0.643, 1.000]	0.673 [0.512, 0.822]
Original	NestedCombat1	Man-kVp-StdN	0.713 [0.507, 0.895]	0.718 [0.564, 0.846]	0.692 [0.500, 0.870]	0.769 [0.500, 1.000]	0.857 [0.692, 1.000]	0.731 [0.579, 0.878]
Stabilized	NestedCombat1	Man-kVp-StdN	0.799 [0.600, 0.951]	0.744 [0.590, 0.872]	**0.731** [0.548, 0.900]	0.769 [0.500, 1.000]	0.864 [0.700, 1.000]	0.750 [0.589, 0.893]
Harmonized	NestedCombat1	Man-kVp-StdN	0.657 [0.483, 0.819]	0.538 [0.385, 0.692]	0.462 [0.276, 0.667]	0.692 [0.417, 0.923]	0.750 [0.500, 0.938]	0.577 [0.404, 0.731]
Original	NestedCombat2	Man-SliceThick-StdN	0.503 [0.290, 0.705]	0.487 [0.308, 0.641]	0.462 [0.259, 0.655]	0.462 [0.259, 0.655]	0.667 [0.429, 0.875]	0.500 [0.322, 0.673]
Stabilized	NestedCombat2	StdN-Man-SliceThick	0.657 [0.480, 0.826]	0.615 [0.462, 0.769]	0.577 [0.391, 0.769]	0.692 [0.429, 0.929]	0.789 [0.588, 0.955]	0.635 [0.469, 0.800]
Harmonized	NestedCombat2	Man-SliceThick-StdN	**0.802** [0.617, 0.959]	**0.769** [0.641, 0.897]	**0.731** [0.552, 0.897]	**0.846** [0.625, 1.000]	**0.905** [0.765, 1.000]	**0.788** [0.655, 0.914]

Table 4.10: Comparison of response prediction performance of different longitudinal models in the external test set based on the resulting features after various image and feature harmonization techniques. Performance is assessed by evaluating AUC, ACC, SENS, SPEC, PREC and bACC, respectively. For each metric, the 95% confidence interval is shown and the highest value is highlighted in bold.

ACC of 0.744, SENS of 0.762, SPEC of 0.727 in the internal test and an AUC of 0.802, ACC of 0.769, SENS of 0.731, SPEC of 0.848 in the external test set.

Of particular significance is the impact of selecting StdNoise as a batch effect, which played a pivotal role in achieving robust prediction performance. Notably, the model trained on original images, with features harmonized by this batch effect, also demonstrated good results in both internal and external test sets, with AUC values of 0.708 and 0.790, respectively.

It is worth to mention that none of the models exhibited significantly superior results upon evaluation using the deLong test ($p > 0.05$).

Figure 4.8 and 4.9 illustrate the Kaplan-Meier survival curves for progression-free survival and overall survival on both the independent internal and external test sets for stabilized images after nestedComBat1 and harmonized images after nestedComBat2. Notably, these models demonstrated significant stratification for both progression-free survival and overall survival, with p-values below 0.05. These findings suggest promising potential for the model as a reliable biomarker for predicting patient response to immunotherapy treatment.

4.4 Discussion and Conclusion

Computed Tomography (CT) imaging plays a crucial role in oncology by offering precise tumor staging and monitoring treatment response (Baumann et al., 2016). Radiomics holds promise for extracting valuable features from CT images, potentially revealing tumor heterogeneity that may not be discernible to the human eye. They may contain information on tumor heterogeneity that reflect the underlying tumor structure and molecular tumor phenotypes. These features could serve as low-cost, quantitative, and reproducible biomarkers for prognostic and predictive purposes

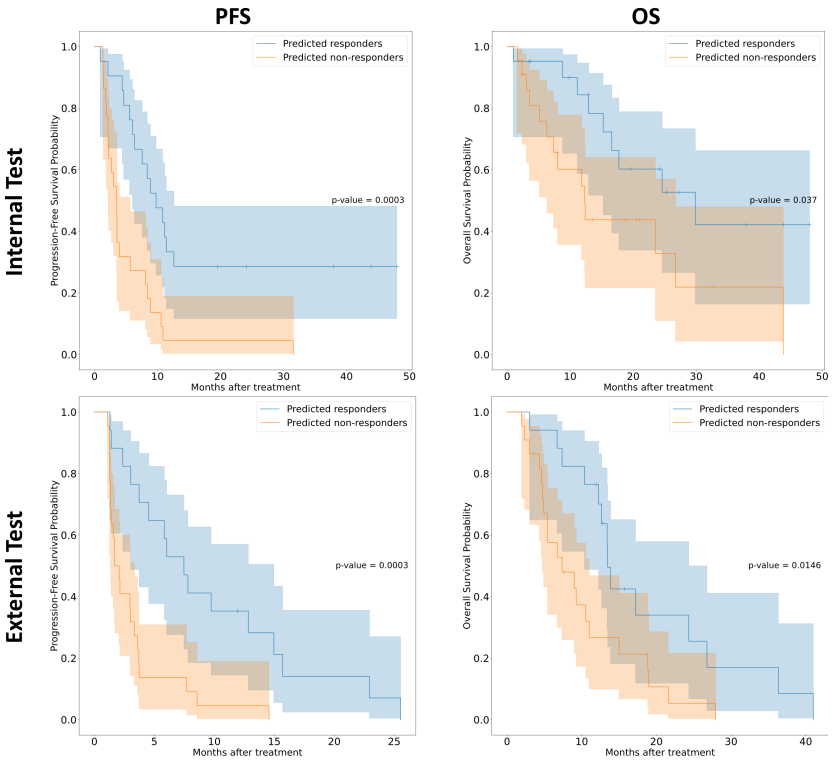


Figure 4.8: Kaplan-Meier survival curves for the internal (top row) and external (bottom row) test sets, illustrating PFS in the first column and OS in the second column. The curves are based on the model utilizing features extracted from the stabilized CTs after NestedComBat1 harmonization.

(Gillies et al., 2016). However, the translation of radiomics into clinical practice has been limited by challenges related to their reproducibility.

Numerous studies have highlighted the variability introduced in radiomics features due to factors such as acquisition techniques, reconstruction methods, and scanner types. Moreover, inconsistencies in tumor segmentation further worsen this issue (Califf, 2018; B. Zhao et al., 2016). Implementing generalizable models based on radiomics necessitates multi-center studies, making it even more difficult to ensure uniformity in acquisition protocols and parameters. This challenge becomes even more critical when analyzing data over time, as features extracted from CT scans acquired with different acquisition settings can exhibit significant variations.

In our study, we aimed to investigate potential sources of variation in radiomics features, specifically in NSCLC patients undergoing immunotherapy across a three-center cohort monitored through CT evaluations during the first months of treatment. We analyzed images from each patient's baseline CT scan, taken before treatment initiation, as well as up to two follow-up scans during early treatment. To mitigate these sources of variation, we employed two approaches for data harmonization: (a) image harmonization, aimed at reducing noise in the images before feature extraction, and (b) feature harmonization, targeting the variability of radiomics features due to confounders, often referred to as batch effects. For image harmonization, we first resampled the data to the minimum in-plane resolution, followed by implementing the harmonization technique

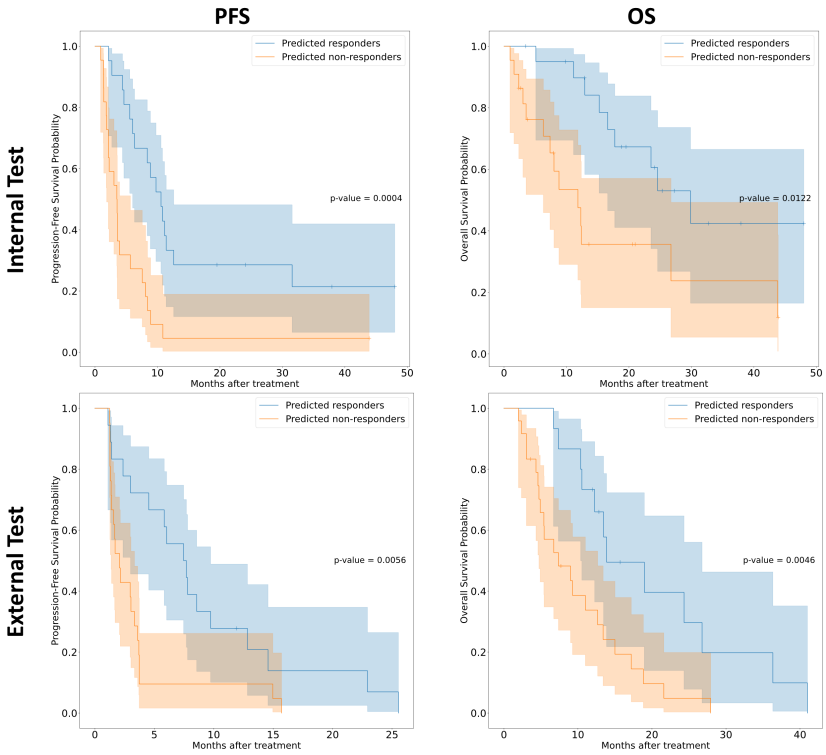


Figure 4.9: Kaplan-Meier survival curves for the internal (top row) and external (bottom row) test sets, illustrating PFS in the first column and OS in the second column. The curves are based on the model utilizing features extracted from the harmonized CTs after NestedComBat2 harmonization.

proposed by Vegas-Sánchez-Ferrero et al. (2019). This method involves statistically characterizing the signal-to-noise relationship within CT image intensities. By estimating the noise-filtered signal and integrating it with an autocalibration technique, this approach aims to mitigate biases caused by noise and reconstruction techniques. To further harmonize the features, we employed ComBat, a method known for its ability to adjust mean and variance across batches using Bayesian estimation (Fortin et al., 2017; Johnson et al., 2007). ComBat is particularly useful for small datasets but it corrects only one batch effect at a time. To address multiple batch effects simultaneously, we employed Nested ComBat, which sequentially harmonize these effects. Furthermore, we introduced a strategy not extensively explored in the literature. This involves learning the ComBat estimates in the training set and applying them subsequently to the test set.

We assessed the synergistic value of combining image and feature harmonization to derive more robust radiomics features. To our knowledge, no previous studies have investigated the combination of both image and feature harmonization. Our analysis identified four primary sources of batch effects, encompassing acquisition parameters such as manufacturer, kVp, slice thickness, and image noise, quantified by the standard deviation of image intensities in the trachea.

Initially, features extracted from original images exhibited significant variability, with approximately 69% influenced by manufacturer-related batch effect. Conversely, features derived from stabilized

and harmonized images demonstrated enhanced robustness, with less than 15% dependent on manufacturer (Table 4.8). Interestingly, original images appeared slightly more robust to slice thickness variations compared to features from harmonized images. This discrepancy may be due to our proposed stabilization process, potentially affected by the low inter-slice resolution in part of the dataset. It suggests that resampling may not entirely compensate for confounding effects associated with differing inter-slicer resolutions across the dataset. As the image stabilization was primary proposed for the context of voxel isotropy, the local variance estimation may have been slightly influenced by this anisotropy.

Moreover, harmonized features exhibited increased robustness when multiple batch effects were considered simultaneously, with less than 19% of features dependent on either manufacturer, kVp, or stdNoise. Harmonization effectively reduced feature variability against CT acquisition parameters; however, complete compensation for these batch effects required feature harmonization. Notably, both ComBat and nestedComBat harmonizations demonstrated remarkable efficacy in addressing all potential batch effects and their combinations, elevating the proportion of robust features to over 90% for almost all batch effects. These results are depicted in both Table 4.8 and Figure 4.5. It's worth noting that employing both image and feature harmonization led to the highest proportion of robust features, especially when compared to feature harmonization applied to images without harmonization.

Following image and feature harmonization, we investigated whether the enhanced robustness of features against batch effects corresponded to improved prediction performance in longitudinal models trained to predict immunotherapy treatment outcomes. Our hypothesis was that reducing the reliance of features on batch effects and enhancing feature comparability over time — by aligning tumor intensities during calibration and feature distributions for each batch effect —would significantly enhance model performance.

Our findings revealed improvement in prediction performance for models utilizing radiomics features extracted from either stabilized or harmonized images, particularly those harmonized with nestedComBat, compared to features from original images. Specifically, the model using features extracted from stabilized images and harmonized with nestedComBat1 achieved an AUC of 0.743, ACC of 0.744, SENS of 0.762, and SPEC of 0.727 in the internal test set, and an AUC of 0.799, ACC of 0.744, SENS of 0.731, and SPEC of 0.769 in the external test set. Similarly, the model using features extracted from harmonized images and corrected with nestedComBat2 yielded favorable results in both test sets, with an AUC of 0.768, ACC of 0.744, SENS of 0.762, and SPEC of 0.727 in the internal test set, and an AUC of 0.802, ACC of 0.769, SENS of 0.731, and SPEC of 0.848 in the external test set. However, these improvements were not statistically significant compared to the results obtained using original images without feature harmonization, as confirmed by testing with the deLong test.

These results aligns with previous research. Ibrahim et al. (2022) observed a decrease in the performance of radiomics features in predicting survival in patients with renal cell carcinoma following ComBat harmonization. Similarly, Singh, Horng, Chitalia, et al. (2022) noted that while harmonization techniques enhance the reproducibility of radiomics models, this improvement does not necessarily translate into enhanced prediction performance. Instead, harmonization facilitates the comparability of prognostic scores between subject groups categorized by batch variables,

thereby enhancing reproducibility. The satisfactory results in the external test set enlightens that one of the main gains of image and feature harmonization in radiomics studies could be improved generalizability and robustness. Further investigation with larger cohorts are guaranteed in this line of research.

Moreover, we proved that a scheme that does not require re-estimation or re-training for neither image nor feature harmonization to be applied new data is satisfactory and this is a crucial aspect for the clinical deployment of radiomics-based models.

This study has several limitations that need to be addressed. Firstly, our study sample size was relatively small, and it exhibited significant heterogeneity in both image parameters and clinical characteristics among the patient population. Additionally, the inclusion of patients with varying treatment regimens may have introduced confounding variables. Despite utilizing two independent test sets, they were still relatively small, potentially affecting the representativeness of our results. Secondly, we observed that the image stabilization was influenced by slice thickness. While this dependency was significantly alleviated by resampling images to the minimum in-plane resolution of each scan, and the effect is minimal, we will certainly explore to improve the image stabilization to ensure the methodology can be applied in real world data with different inter-slice resolutions. Thirdly, not all categories within each batch effect were equally represented in our dataset, particularly notable in the manufacturer category. This imbalance could have impacted the estimation of Combat parameters, especially in cases where certain manufacturers were underrepresented, such as GE. Moreover, despite implementing feature and image harmonization techniques to reduce feature dependency on batch effects, our prediction results did not demonstrate a significant improvement over those obtained using original features alone. Our evaluation in the internal and external cohort shows an slight improvement trend. Further experiments are needed to confirm these trends in larger data. Additionally, we need to investigate the implications of employing less stringent criteria to define stable features. Leveraging a broader range of features, which may be corrected by image and feature harmonization, could potentially enhance prediction performance without the risk of utilizing batch-dependent features.

Lastly, we did not account for contrast doses or administration patterns as an additional batch effect due to the unavailability of this information in our dataset. This omission may have overlooked a potential source of variability that could impact the reproducibility of radiomics features in the tumor.

In conclusion, we have identified slice thickness and manufacturer as the primary sources of variability in CT acquisition parameters affecting radiomics. While image preprocessing through image harmonization successfully reduced manufacturer variability, it was insufficient to compensate for all the variability related the batch effects. Therefore, post feature harmonization using ComBat was essential to minimize all sources of variability, either individually or in combination. Additionally, we investigated whether the enhancement in radiomics robustness correlated with improved predictive model performance. Although we observed a slight increase in prediction accuracy, it did not reach statistical significance. Overall, this study contributes to a deeper understanding of data harmonization techniques in medical imaging, offering potential for more reliable and reproducible radiomics analyses in clinical practice.

CHAPTER 5

Spatio-temporal Deep Learning in Indeterminate Lung Nodules

As the foremost contributor to cancer-related mortality globally, lung cancer necessitates innovative approaches for diagnosis and management. Deep learning-based computer-aided diagnosis (CAD) systems, integrated into screening programs, offer promise in enhancing malignancy prediction, aiding radiologists in decision-making, and mitigating inter-reader variability. However, limited research has explored the potential of analyzing repeated annual exams of indeterminate lung nodules to enhance accuracy. In this chapter, we introduce a novel spatio-temporal deep learning framework to predict indeterminate lung nodule malignancy using serial screening computed tomography (CT) images from the National Lung Screening Trial (NLST) dataset. The content of this chapter has been drafted for a publication that is being submitted to an international journal.[1]

5.1 Introduction

Lung cancer is one of the leading causes of death and the most common cause of cancer-related death worldwide, with an overall 5-year survival rate ranging from approximately 10% to 20% in most countries since its diagnosis (Allemani et al., 2018). Most patients are already in advanced stages when diagnosed, which contributes to a poor prognosis. However, this poor prognosis is related to its late diagnosis and can vary significantly between different stages of the disease (Siegel et al., 2020). In this scenario, there's a pressing need for early lung cancer detection via screening campaigns and improved late-stage management to enhance outcomes.

The diagnosis of lung cancer is based on a combination of imaging techniques, laboratory tests, and histopathological examination. Chest X-rays are commonly used as a screening tool, providing an

[1] Farina, Benito, et al. "A spatio-temporal deep learning framework with temporal global attention for nodule malignancy prediction from longitudinal screening CT images." to be submitted to Computerized Medical Imaging and Graphics.

initial assessment of lung abnormalities. However, low-dose computed tomography (LDCT) scans have been shown to be more accurate, reduce mortality rate, and increase survival (de Koning et al., 2020; Team, 2011b). The NLST demonstrated a 20% reduction in mortality rates among high-risk individuals (former/current smokers with a $\geq$ 30 pack-year history) being screened with LDCT scans compared to chest radiography (Team, 2011a). Nevertheless, several concerns remain regarding the establishment of lung cancer screening using LDCT as standard of care. One concern is the high false-positive rates associated with LDCT screening, which can lead to unnecessary follow-up tests and potential patient anxiety (Hammer et al., 2022). Additionally, there is a risk of adverse events resulting from overdiagnosis, in which slow-growing or indolent cancers are detected and treated, leading to potential harm without improving overall outcomes (Patz et al., 2014). Another concern is the significant percentage of lung tumors that are still detected in advanced stages, indicating the need for more effective early detection strategies (Ruano-Ravina et al., 2015). Finally, there is a potential risk of radiation-induced neoplasms associated with repeated LDCT scans and additional testing such as PET CT (Brenner, 2004). Notably, a considerable number of identified nodules remain indeterminate after conventional evaluation, posing a challenge in terms of assessment and management, especially considering the uncertain potential for malignancy (Massion & Walker, 2014). These uncertain cases often necessitate repeated annual examinations and further interventions to rule out malignancy, imposing a significant burden on healthcare systems.

Imaging modalities such as CT play a crucial role in modern healthcare, providing clinicians with valuable information for diagnosis, treatment planning and assessment, and intervention. Nevertheless, the growing volume of medical images generated in recent years poses significant challenges (Schöckel et al., 2020). Interpreting such vast quantities of images requires extensive expertise. However, this approach has limitations such as human subjectivity, inter- and intra-observer variability, and fatigue. The increasing volume of medical images places significant time constraints on radiologists, resulting in high financial and clinical costs due to missed or late diagnosis, as well as unnecessary biopsy procedures.

To address these challenges, CAD system has been shown to be effective in improving the inter-reader consistency and assisting radiologists in decision making (Munir et al., 2019). In recent years, these systems have demonstrated their capability to automatically detect lung nodules and extract essential information to track the evolution of the nodule over time (Lampeter, 1985; Lu, 2021; X. Wang et al., 2019; Y. Zhao et al., 2012). The integration of CAD systems not only saves radiologists time on each case but also ensures more consistent and reproducible results during nodule follow-up. Y. Zhao et al. (2012) demonstrated CAD systems effectiveness in reducing reading time and diagnostic errors in lung screening compared to the time-consuming and expensive double-reading process.

In recent research, artificial intelligence has gained popularity for analyzing lung nodules, with both radiomics and deep learning approaches emerging as effective methods for predicting nodule malignancy. They have demonstrated significant advancements in estimating risk and categorizing patients with lung cancer considering various comprehensive nodule characteristics rather than just its size (Alahmari et al., 2018; Beig et al., 2019; Causey et al., 2018; Hawkins et al., 2016; Mikhael et al., 2023). Causey et al. (2018) proposed a deep learning framework for the prediction of nodule malignancy by integrating traditional radiomics features with features extracted from a convolutional

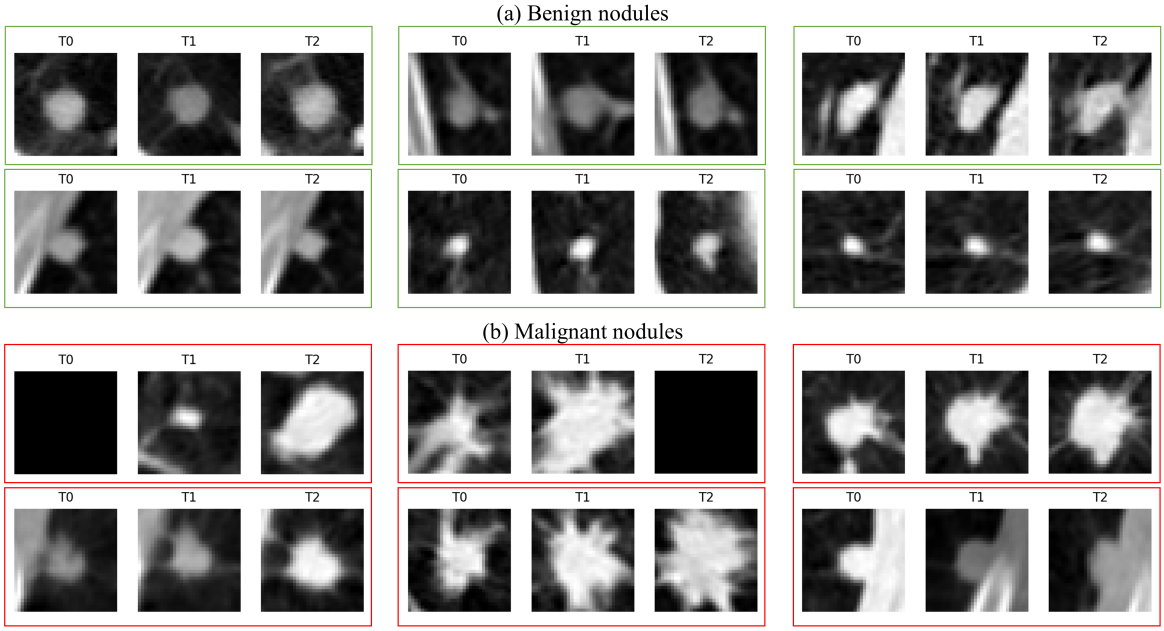


Figure 5.1: Examples of CT lung nodules: a) benign nodules, b) malignant nodules. Three images of the same nodule, extracted from CT scans at different time points (T0, T1, and T2) during patient follow-up, are displayed. Missing time points are represented by black images, indicating the absence of CT scan at those instances. The figure illustrate distinct characteristics between the two groups.

neural network (CNN) model trained for malignancy classification. Ciompi et al. (2017) proposed a multi-scale with a multi-stream CNN for lung nodule type prediction simultaneously processing multiple slices of the same nodule at multiple scales. Y.-S. Huang et al. (2023) implemented a 3D-CNN model with an attention scheme based on the squeeze-and-excitation module block to focus on important features, integrating nodules features at different scale. Mikhael et al. (2023) designed a 3D-CNN to predict future lung cancer risk up to 6 years after the CT scan, utilizing both local nodule details and global CT information.

Many CAD systems proposed for cancer prediction and nodule classification have primarily focused on analyzing single instances. However, accurate outcome prediction, especially in screening, necessitates considering not only the spatial domain, but also the temporal changes in nodule size, shape, and density, especially for indeterminate nodules. For instance, in terms of spatial characteristics, malignant nodules often exhibit more irregular margins and greater internal heterogeneity compared to benign nodules (F. Li et al., 2004). Malignant nodules typically demonstrate relatively rapid growth, with their size doubling approximately every four months or even faster (see Figure 5.1). Ground glass nodules may develop a solid component over time, indicating malignant transformation. That notwithstanding, it is important to emphasize that volume doubling time is not sufficient to predict nodule malignancy, since some malignant nodules can grow slowly or even remain stable over time, as clinically observed by R. Zhang et al. (2020).

Therefore, considering the temporal evolution becomes crucial in assessing nodule malignancy,

given the diverse growth patterns observed in clinical practice. This involves monitoring changes in various nodule characteristics across repeated CT scans (Kastner et al., 2021). While capturing both spatial and temporal information is essential, only a limited number of previous studies have introduced deep learning architectures designed to effectively utilize such spatio-temporal data. Most of the proposed spatio-temporal deep learning architectures focus on video processing or dynamic imaging, primarily for lesion detection or segmentation (Amador et al., 2022; Su et al., 2022). Few works process longitudinal data, primarily for predicting future lesion appearance. For instance, in lung cancer, (L. Zhang et al., 2019) implemented a spatio-temporal convolutional LSTM segmentation model for the prediction of tumor growth from 4D CT images. Very few studies have addressed longitudinal data for predicting outcomes or classifying lesions (Veasey et al., 2020; Xu et al., 2019b).

Veasey et al. (2020) implemented a siamese-style convolutional attention network capable of effectively concatenating relevant 2D features from multiple time points. Ardila et al. (2019) proposed a spatial deep learning framework that extracted local and global features from the whole scan at two time instants to determine the probability of cancer. While these methods incorporate spatio-temporal data, their prediction models often do not utilize all available time points, potentially overlooking important temporal patterns. Moreover, they commonly operate on highly imbalanced datasets, skewed towards benign cases, which can introduce biases into the results. Additionally, the disparity in missing data distribution between cancer and non-cancer patients is often overlooked. Training large networks on such imbalanced datasets poses a significant challenge, requiring careful consideration of dataset selection and network design to ensure effective training and convergence.

The present chapter aimed to propose a novel spatio-temporal deep learning framework featuring a temporal global attention module for the prediction of indeterminate nodule malignancy using serial CT images. Our framework integrates both spatial and temporal dimensions by simultaneously processing multiple time points for each nodule. We incorporate a global attention module to capture temporal changes in nodule characteristics and employ strategies to handle missing data in the temporal dimension, thereby mitigating potential biases arising from missing time steps. Unlike previous methods that consider all available nodules indiscriminately, we specifically design, train, and validate our methodology for indeterminate nodules. Focusing solely on indeterminate nodules has reduced the size of our training and test sets. However, this targeted approach improves the accuracy and interpretability of predictions for indeterminate nodules, leading to more precise and interpretable assessments of indeterminate nodule malignancy. Additionally, it tackles challenges linked to longitudinal data, thereby enhancing classification accuracy and prediction reliability. Additionally, we contribute to the field by publicly sharing longitudinal information, including nodule position and diagnosis, for the indeterminate nodules studied in this work.

5.2 Materials and Methods

5.2.1 NLST Dataset

The data was obtained from NLST trail database after approval and signature of data transfer agreement with the National Cancer Institute (Bethesda, MD, USA), comprising a large repository of lung cancer screening data. The trial aimed to assess the efficacy of LDCT scans in detecting lung cancer in an early stage compared to chest X-rays. Participants were randomly assigned to receive either a standard chest X-ray or a LDCT annually for three years as part of the screening process. The dataset consists of imaging data from more than 50,000 participants identified as high risk for lung cancer, enrolled between 2002 and 2004. These individuals were current or former smokers who quit smoking within the last 15 years, aged 55 to 74 years, with a smoking history of at least 30 pack-years and no signs, symptoms, or history of lung cancer, among others (Team, 2011a). The original dataset contained the annotations of the screening exams performed by radiologists. An exam was considered as positive (suspicious of lung cancer) if the radiologist identified a non-calcified nodule or mass with a diameter of 4 mm or more, or detected other suspicious abnormalities (Team, 2011a).

Within the available data, which included approximately 15,000 subjects, most scans were nodule free. We restricted our analysis to the indeterminate lung nodules, subsequently confirmed either benign or malignant (through biopsy) during the 3-year screening phase. Importantly, these nodules were consistently monitored across the three rounds of LDCT screening. Subjects with cancer detected in the post-screening phase (beyond the initial 3-year follow-up period) were excluded.

Participants in a lung cancer screening program may present with multiple nodules during baseline or follow-up scans. Our study focuses on indeterminate nodules, which include suspicious lesions initially considered positive at baseline but later confirmed as benign, as well as incidental findings detected during follow-up. Consequently, malignant nodules identified at baseline were excluded from the analysis. Incidental nodules are particularly important due to their relatively high probability of malignancy despite their smaller size (Walter et al., 2016). The risk-stratification of these lesions poses a significant unmet clinical challenge in lung cancer screening programs. Managing and predicting their behavior, let alone conducting invasive testing, is highly complicated due to their small size when compared to prevalent nodules detected during initial LDCT screenings. Our methodology specifically targets indeterminate nodules, as they pose the most significant challenge. However, this focused approach, often overlooked by previous methods, significantly impacts the size of the development and testing dataset.

In addition, our patient selection process ensured a minimum of two CT scans per participant over the screening duration. Furthermore, we restricted our analysis to malignant nodules explicitly identified by NLST, considering their anatomical locations as described in the NLST participant data sheet.

Regarding image quality, each subject may have multiple images for each screening examination due to the application of various filters in the CT reconstruction algorithm. To ensure consistency, we specifically utilized soft kernels for the reconstruction filters, which are commonly employed for detecting nodules and assessing nodule morphology (Vonder et al., 2021). However, it should be

	Total	Train	Test
Patients	443	333	110
Non-cancer	235	176	59
Cancer	208	157	51
Nodules	703	528	175
Benign	486	365	121
Malignant	217	163	54

Table 5.1: Number of patients and nodules in the training and test sets.

emphasized that only CT scans with sufficient visual quality were included in our study cohort.

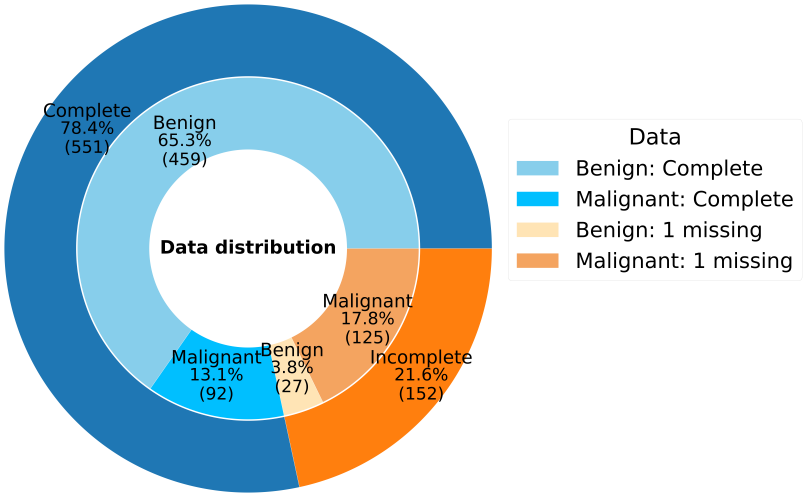


Figure 5.2: Pie chart displaying the distribution of complete and incomplete data (missing one time point) for benign and malignant nodules. The outer ring represents the distribution for the entire dataset, while the inner ring provides a detailed breakdown by nodule type.

In total, 443 NLST subjects were selected for the purpose of our analysis, including 235 non-cancer cases and 208 cancer cases. Each subject had up to three LDCTs (three time points), a baseline scan (T0), and two follow-up scans (T1 and T2) performed at one year intervals. For non-cancer cases, all nodules identified by LDCT were confirmed as benign during follow-up. In contrast, for malignant nodules, we relied on the nodule location annotations provided in the NLST dataset, which typically included only one nodule per patient. As a result, our study analyzed a total of 703 nodules. The total number of patients and nodules selected are shown in Table 5.1. It is worth mentioning that, since some follow-up CT scans are missing, not all nodules had three time points, as it can be seen in Figure 5.1. Nevertheless, we guaranteed that at least two scans were available for each nodule. Figure 5.2 shows the distribution of complete and incomplete data for the benign and malignant nodules studied. It reveals that 94% of benign nodules has complete data, whereas malignant nodules exhibit a higher proportion of incomplete data (one time point missing as one follow-up scan is missing) at 58%. The presence of these substantial differences in missing data in malignant nodules raises concerns about potential temporal bias, raising the need to address this

2D CNN			globAttCRNN		
Layers	Details	Parameters	Layers	Details	Parameters
Input	32 x 32	0	Input	3 x 32 x 32	0
Conv	48, 3x3, pad 0, stride 1, l1l2 regularization, ReLU	480	2D CNN		106452
Conv	48, 3x3, pad 0, stride 1, l1l2 regularization, ReLU	20784	Masking		0
MaxPool	2 x 2, pad 0, stride 2	0	GRU	4, dropout 0.05	504
BatchNorm		192	BatchNorm		16
Conv	48, 3x3, pad 0, stride 1, l1l2 regularization, ReLU	20784	Global Attention		48
Conv	48, 3x3, pad 0, stride 1, l1l2 regularization, ReLU	20784	FC	2, softmax	10
MaxPool	2 x 2, pad 0, stride 2	0			
BatchNorm		192			
Flatten		0			
FC	36, ReLU	43236			
Dropout	p = 0.05	0			
FC	2, softmax	74			
Total parameters		106,526			106,830

Table 5.2: Architectural details and parameters of the 2D-CNN (left) and globAttCRNN networks (right).

issue in order to ensure reliable and unbiased analyses.

The dataset was randomly divided, ensuring patient-wise stratification, into a training set (333 patients) and an independent test set (110 patients). This split maintained a balanced representation of cancer patients in both sets. Each nodule was assigned to its corresponding patient group to maintain stratification at the patient level. The primary endpoint of this study was the prediction of nodule malignancy. Overall survival (OS) was a secondary endpoint, defined as the time in months between patient enrollment in the trial and either death or censored to 2009 (end of data collection).

5.2.2 Global Attention CRNN

In this study, we introduce a novel architecture called globAttCRNN, which combines a lightweight convolutional recurrent neural network (CRNN) with a temporal global attention module. This network leverages the strengths of a 2D-CNN to capture spatial nodule features and a recurrent neural network (RNN) with a temporal global attention module to effectively integrate information across multiple time points.

The proposed approach has several advantages. First, the RNN relies not only on the spatial features extracted from the latest available image, but also incorporates temporal information from the entire sequence. This enables the model to dynamically capture the temporal evolution of the nodule characteristics, providing a more comprehensive representation of the nodule's behavior over time.

The integration of a temporal global attention module further enhances the model's performance. The attention module enables the network to focus on the most informative spatial and temporal features, while selectively ignoring irrelevant or redundant information. This module promotes more effective feature learning and contributes to improve accuracy in nodule classification. Further insights into the functioning and implementation of the temporal global attention module will be provided in the following section.

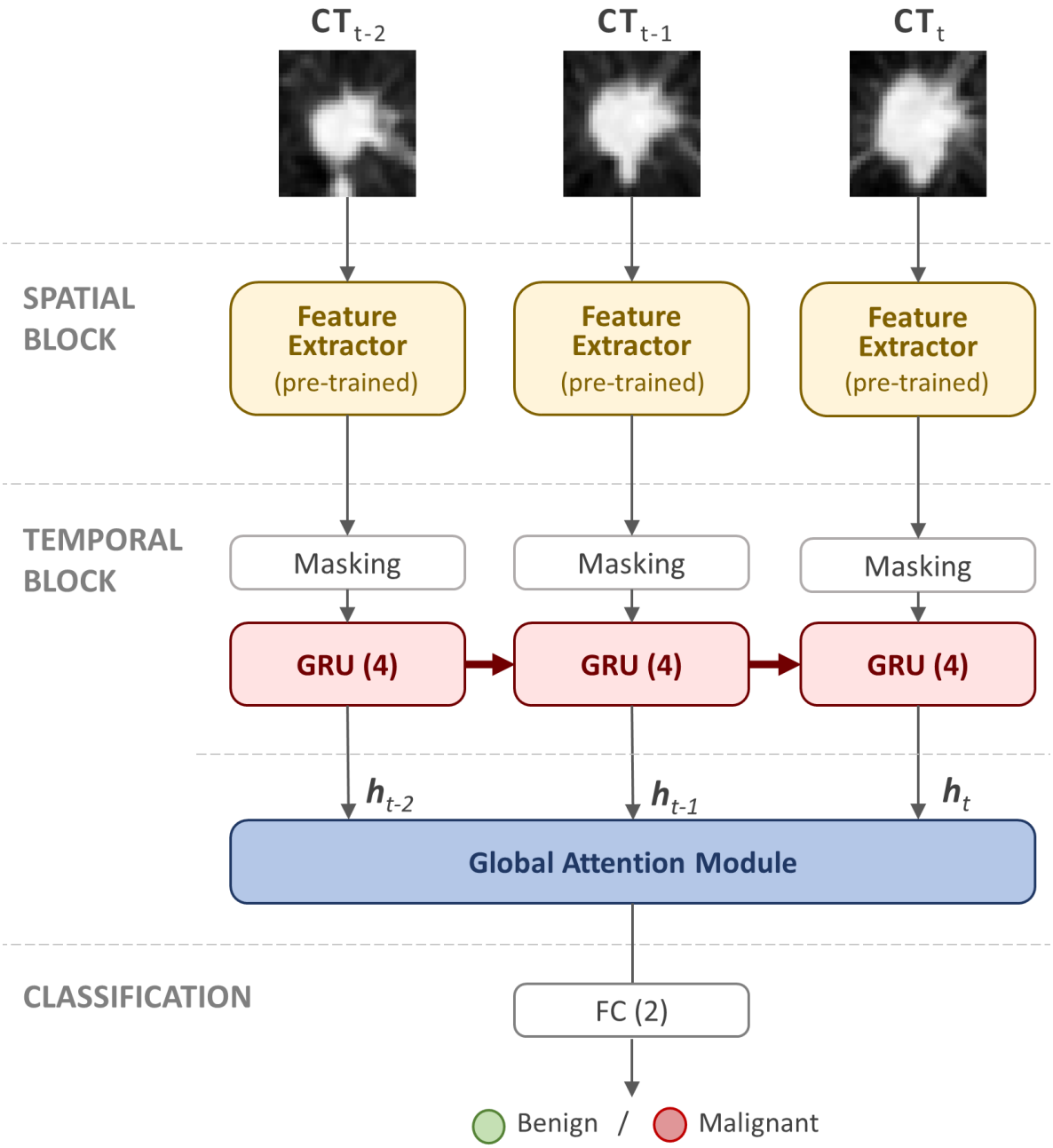


Figure 5.3: General overview of the globAttCRNN architecture for the prediction of nodule malignancy. It comprises two primary components: a spatial block for capturing spatial nodule features and a temporal block for integrating temporal information from multiple time points.

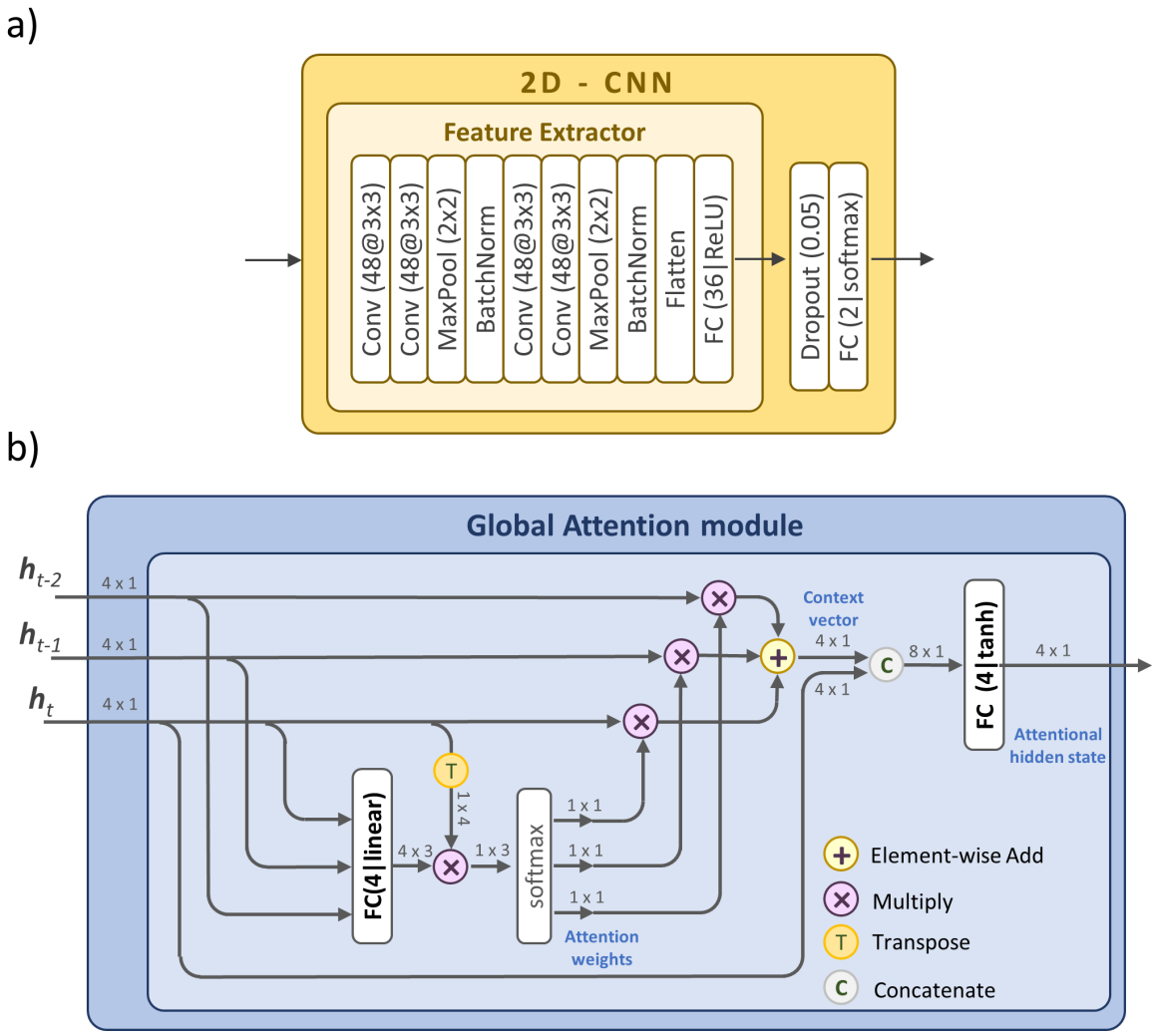


Figure 5.4: General overview of the modules of globAttCRNN architecture a) Focus on the 2D-CNN architecture for nodule malignancy prediction, consisting of a feature extractor responsible for spatial nodule features extraction and a classification layer for making predictions. b) In-depth focus on the global attention module, which dynamically weights the significance of each time point in the malignancy prediction process.

The architecture of our proposed model is illustrated in Figure 5.3. This architecture is designed to be adaptable to any number of time instants. However, for our study, we focused on the NLST dataset, which includes only three time instants: T0, T1, and T2, in figure corresponding to t-2, t-1 and t, respectively. Table 5.2 provides an overview of the parameters and weights associated with the spatial and temporal blocks within our approach.

The model comprised a 2D-CNN to extract spatial features and capture spatial information from the images of each nodule at different time points (Figure 5.4a). The 2D-CNN was designed to be very small, with few convolutional blocks in order to keep the number of network weights low and prevent overfitting. Remarkably, this component of the model comprises only 106,526 parameters, with a small additional increase of 578 parameters introduced by the temporal block. Compared to conventional CNN models such as ResNet and VGG, our model has a significantly smaller number of weights, which played a crucial role in mitigating overfitting. This feature not only improves the computational efficiency of the whole network, but also highlights its practical deployability in various computing environments. It is worth noting that we experimented with different conventional CNN architectures, but encountered overfitting problems when the number of network weights increased substantially. Hence, our emphasis on weight efficiency and compactness proved critical to the success of the model.

For the temporal block, we introduced a RNN, specifically a (Gated Recurrent Unit) GRU layer with 4 hidden units, to capture the temporal dependencies among the extracted features from the image sequence. This layer takes the features of each image at different time steps as input, and its hidden states encode the sequential information present within the data. Following the RNN, we incorporate a temporal global attention module, which plays a crucial role in our model. At each time step, this module calculates attention scores to determine the relative importance of the image features at each point in the sequence. To further elaborate, the hidden states of the RNN at each time point are passed through the global attention module. This module allows to integrate the information from the last hidden state and a weighted sum of all hidden states, thereby enabling the model to make informed decisions based on both the recent information and the collective understanding of the entire sequence.

To address the issue related to missing scans, we introduced a masking layer positioned on top of the RNN to skip the missing time points during network processing. This ensures that the model appropriately handles and accounts for the absence of certain scans in the input sequence.

The output of the temporal global attention module is a feature vector of dimension 4, encompassing the integrated temporal and spatial information pertaining to the nodule. By combining these aspects, our model achieved a comprehensive representation that captured the relevant nodule features across both the temporal and spatial domains.

The lightweight CRNN configuration ensured feasible training, especially when dealing with the inherent challenge of limited data typically encountered in medical imaging studies.

Temporal Global Attention Module

The global attention module was implemented to dynamically weight each input so that the model could focus on and give greater importance to the time steps with the most relevant information for the prediction of malignancy. The global attention module was defined according to the concepts presented in Luong et al. (2015). Figure 5.4b provides a visual representation of the global attention module within our proposed globAttCRNN architecture.

This module takes as input all the hidden states of the RNN and generates a concatenation of the last hidden state and a *context vector* calculated as the weighted sum of each hidden state by the corresponding attention values. These features are finally passed through a dense layer with a Tanh activation function to produce a new feature vector that integrates both spatial and temporal information.

Let $\boldsymbol{h}_t$ be the last hidden state and $\boldsymbol{H}$ a matrix comprising all hidden states. The context vector $\boldsymbol{c}_t$ and the attention weights $\boldsymbol{a}_t$ are calculated as follows:

$$\text{score}\left(\boldsymbol{h}_t, \boldsymbol{H}\right) = \boldsymbol{h}_t^T \boldsymbol{W_a} \boldsymbol{H} \tag{5.1}$$

$$\boldsymbol{a}_t = \text{softmax}\left(\text{score}\left(\boldsymbol{h}_t, \boldsymbol{H}\right)\right) \tag{5.2}$$

$$\boldsymbol{c}_t = \boldsymbol{a}_t \boldsymbol{H} \tag{5.3}$$

$$\widetilde{\boldsymbol{h}_t} = f\left(\boldsymbol{c}_t, \boldsymbol{h}_t\right) = \tanh\left(\boldsymbol{W_c}\left[\boldsymbol{c}_t, \boldsymbol{h}_t\right]\right) \tag{5.4}$$

where $\widetilde{\boldsymbol{h}_t}$ represents the attentional hidden state, which is subsequently used for making predictions. In (5.1), the attention scores are calculated by comparing the last hidden state of the RNN with all previous hidden states. These scores quantify the relevance or importance of each image features for the current time step in the classification task. Higher scores indicate that the corresponding image is more relevant or informative for the current time step. To determine the attention weights, the attention scores are normalized using a softmax function to ensure that they sum up to 1 (5.2). It was ensured that the attention weights for missing time steps were 0. The context vector is then computed by taking the weighted sum of each hidden state using the attention weights (5.3). Finally, an attentional hidden state was computed by combining the context vector and the last hidden state as in (5.4).

The attentional hidden state is subsequently fed into a fully-connected classifier to make predictions.

This module facilitates gaining insights into model predictions by allowing the assessment of the contribution of each time point to the model output through the analysis of its attention weights, thereby enhancing model interpretation.

5.3 Experiments

5.3.1 Image Preprocessing

This study primarily aimed to classify indeterminate lung nodules as either benign or malignant.

Due to the lack of explicit information regarding the position of each nodule in the NLST dataset, we relied on a visual determination of the centroid of the nodule. We took special care to ensure consistency in the centroid positions across different time points for the same nodule, thereby establishing coherence in their localization. It is worth noting that registration techniques could have potentially addressed this issue; however, such an approach was beyond the scope of our study. Furthermore, existing commercial tools, such as the Tumor Tracking of Philips IntelliSpace Portal, have partially solved this issue by automatically detecting and segmenting the nodule across all available images for a given patient.

For each lung nodule at each temporal instant, we extracted a single 2D image measuring 20×20 mm^2, centered around the respective centroid. This image size of the nodule provided us with sufficient contextual information for our experiments. In particular, we observed that using smaller (10 mm) or larger (40 mm, 80 mm) nodule images did not yield performance improvements and, in some cases, even led to a degradation in performance.

We conducted additional experiments to explore the potential performance improvement by incorporating multiple slices per nodule (3D fashion) at each time step. However, these experiments did not reveal any enhancements compared to using a single slice. We attribute this to the characteristic nature of screening nodules, which are typically small and exhibit limited heterogeneity along the z-axis.

Subsequently, we resized each image to fit the input size of the 2D-CNN, resulting in an image dimension of 32×32 pixels.

To ensure standardized intensity values, we applied a two-step process. First, we clipped the image intensities between -1000 and 3050 HU. Then, we performed normalization, scaling the values to the range of [0,1] based on the minimum and maximum values observed in the training nodules.

5.3.2 Experiment Settings

Model Training

The purpose of the model was to analyze a nodule image across multiple time points and produce a malignancy score, ultimately predicting whether the patient would receive a cancer diagnosis within one year following the last available temporal instance.

All the models were trained using stratified 5-fold cross-validation, ensuring an equal balance of cancer and non-cancer patients in each fold. Folds were created based on subjects, meaning that nodules from the same subject were assigned to the same fold. Furthermore, the models were trained using the same set of hyperparameters until their performance on the validation fold

reached a plateau. This standardized approach allowed for fair comparison and selection of the best-performing model.

We employed a two-phase training strategy for globAttCRNN. In the first phase, to create a robust feature extractor, the 2D-CNN was pre-trained on all available nodules within the NLST training set, including data from T0, T1, and T2 scans. This pretraining involved 333 patients with a total of 528 nodules. Since each nodule could appear in up to three different time points, the pretraining dataset consisted of 1497 nodule instances. During training, a batch size of 32 samples was used, along with a Adadelta optimizer and a learning rate of 0.05. The objective was to minimize the cross-entropy error over a maximum of 500 epochs. To optimize both the performance and generalization capabilities of the model, we employed L1L2 regularization (with L1 = 0.0025 and L2 = 0.001) and implemented data augmentation techniques during the training phase. This involved random flipping, translation, and rotation of the images. The feature extraction involved the output of the last fully connected layer before the classification stage (Figure 5.4b). This feature extractor underwent training using a stratified 5-fold cross-validation approach. We considered the model with the minimum validation loss during cross-validation as the best-performing model. Although we attempted an ensemble approach by combining all models from the 5-fold cross-validation, no significant improvement was observed. Therefore, to keep the network's parameter count low, we decided to exclude the ensemble approach from our final implementation.

In the second phase, our focus shifted to training the temporal part of the globAttCRNN architecture, without updating the feature extractor's weights. The training process in the second phase closely resembled the feature extractor pretraining, with the exception of data augmentation. To augment data, we applied consistent transformations across all instances of each nodule. This ensures that the data augmentation process maintains consistency across the temporal instances of each nodule, preventing the network from learning spatial transformations that are not representative of reality.

The models were implemented using the TensorFlow framework on a personal computer equipped with a GPU configuration, including an NVIDIA GeForce GTX 1080 with 8 GB of memory and an NVIDIA Quadro P6000 with 24 GB of memory. This hardware setup provided the computational power necessary for efficient model training and inference.

Temporal Augmentation and Dropout

Given the disparity in numbers between malignant and benign nodules (Figure 5.5), we implemented a unique augmentation strategy tailored for malignant nodules with complete data, referred to as (*Temporal malignant nodule augmentation*). This technique involved recursively removing each time instance from the original set of images, generating multiple augmented samples by removing one time instance at a time. For example, if a malignant nodule X had images at time points T0, T1 and T2 ($X_{\text{T0-T1-T2}}$), then we added three other malignant nodules to the training set: $X_{\text{T0-T1}}$,$X_{\text{T0-T2}}$,$X_{\text{T1-T2}}$. This approach increased the representation of malignant nodules in the training set and allowed the network to better capture the temporal variations in their spatial characteristics.

To mitigate potential biases associated with missing data, especially prevalent in malignant nodules, we implemented a specific strategy within the training set. This approach aimed to achieve a

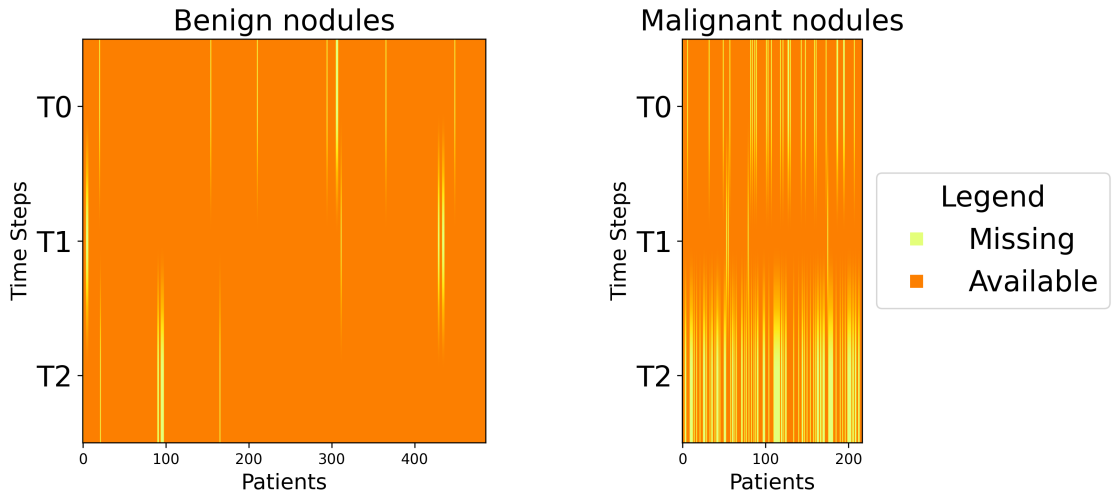


Figure 5.5: Missing data heatmap: comparison of benign (left) and malignant (right) nodules across time intervals in the selected NLST dataset.

more balanced representation of missing data during model training. As illustrated in Figure 5.2, benign nodules typically had fewer instances of missing data compared to malignant nodules. This discrepancy can be attributed, in part, to patients with detected cancer discontinuing screening, resulting in absent data in subsequent scans. This practice aligns with established clinical protocols, such as Lung-RADS (McKee et al., 2016), where radiologists adopt a conservative approach. Nodules considered non-suspicious lead to a postponement of the final decision, with follow-up visits scheduled. In contrast, the detection of suspicious findings leads to patient referral for invasive procedures. On the other hand, missing data for benign nodules may arise due to factors like poor scan quality or fluctuations in nodule appearance during subsequent follow-up scans, particularly if the etiology is found to be infectious or inflammatory. However, missing data in benign nodules is less frequent, making them more likely to have a complete set of three rounds of CT images available.

To address this discrepancy in missing data, a temporal dropout strategy was implemented during training (*Temporal dropout*). For benign nodules that had all available temporal instances, we randomly removed one temporal instance at a time until the number of complete benign data matched that of malignant nodules. This strategy was intended to create a more balanced training scenario, particularly in terms of longitudinal data representation. By doing so, we aimed to ensure that the network learns the spatial changes in the nodules over time, rather than being influenced by the distribution of number of time points or missing data. These techniques were crucial in improving the model's ability to capture meaningful temporal and spatial characteristics of the nodules during training.

These augmentation techniques were exclusively applied to the training set, ensuring that the test set remained unaltered and unaffected by these transformations.

5.3.3 Ablation Study

We conducted an ablation study to comprehensively assess the globAttCRNN model, focusing on understanding the role of its individual components and to evaluate whether the temporal global attention module effectively directed the network's attention to the most informative time steps.

Multiple architectures were examined by progressively removing specific layers of the network and comparing their performance. First, the global attention module was eliminated to create a CRNN network (CRNN). This allowed us to assess the performance of the original model in comparison to a simplified version without the global attention module.

The RNN was then removed while retaining the global attention module (globAttCNN), which directly encoded the spatial features of each nodule at different time steps.

Lastly, a straightforward approach was adopted in which the features extracted from each temporal instant were concatenated and used for classification (concatCNN). This enabled an examination of the impact of discarding the temporal relationship between follow-up images and relying solely on the spatial features.

To ensure that any performance differences between models in the ablation study were solely due to architectural variations and not influenced by differences in training settings, all models were trained using the exact same methodology as the one proposed for the globAttCRNN model.

5.3.4 Clinical Relevance Analysis

We conducted an additional analysis to evaluate the clinical utility of our proposed model, emphasizing the importance of using longitudinal information to predict nodule malignancy. Initially, we compared our longitudinal model with a single time point model, where the already trained feature extractor (2D-CNN) was used as the nodule malignancy predictor. Instead of considering all available time points collectively, we assessed the model's performance when making predictions solely based on the most recent time instant for each nodule (Tlast). This comparison accounted for the possibility that, in certain cases, the T1 scan could serve as the latest available time point for the nodule, instead of the T2 scan. This comparison is crucial given the clinical practice of radiologists who often rely on the most recent imaging study as it may reveal novel suspicious nodule characteristics or a temporal change that might inform additional imaging or invasive testing. We also explored combining the predictions from each temporal instant through average calculation or majority voting, but the approach using the last available time instant yielded superior results.

5.3.5 Statistical Analysis

To assess model performance, we used an independent test set and employed the area under the receiver operating characteristic (ROC) curve (AUC) as the primary evaluation metric. The DeLong test was employed to compare the area under the ROC curves and to assess any significant differences.

In addition to the AUC, several other performance metrics were calculated, including accuracy (ACC), specificity (SPEC), sensitivity (SENS), balanced accuracy (bACC), and precision (PREC). These metrics offer a more detailed assessment of the model's performance. The 95% confidence interval (CI) for all metrics were estimated using a bootstrap resampling approach with 1000 iterations.

To evaluate the ability of the model to stratify patients according to the risk of developing lung cancer, we performed a Kaplan-Meier survival analysis. For each patient, the probability of cancer was predicted by averaging the predicted malignancy probabilities of all their nodules. Patients were subsequently classified into low- and high- mortality risk groups using a cutoff value of 0.5. The log-rank test was used to determine the significance of differences between the survival curves of these two groups.

Statistical significance was considered for p-values less than 0.05 in two-sided tests, indicating the presence of significant differences or associations. This rigorous statistical analysis was designed to provide robust and reliable conclusions regarding the model's performance and patient risk stratification.

5.4 Results

5.4.1 Comparative Analysis of Model Performance

The 2D-CNN feature extractor was pre-trained on all available nodules from the NLST training set, including data from T0, T1, and T2 scans. In the 5-fold cross-validation it achieved an ACC of 0.820 $\pm$ 0.002 (n = 1497). On the other hand, the proposed globAttCRNN model achieved an ACC of 0.854 $\pm$ 0.003 (n = 528). We evaluated the robustness of globAttCRNN model by testing its performance on an independent test set. Nodules in the test set did not undergo the preprocessing step, resulting in a slightly different distribution of missing time steps compared to the training set, particularly with minimal missing data in benign nodules. Detailed visualization of the distributions of missing data in the test set can be found in Figure 5.6.

Model	Time instants	N nodules test	AUC	ACC	SENS	SPES	PREC	BACC	DeLong p-value
CNN	Tlast	175	0.916 [0.861,0.964]	0.846 [0.789,0.897]	0.796 [0.685,0.900]	0.868 [0.805,0.924]	0.729 [0.614,0.837]	0.832 [0.768,0.893]	0.0064
concatCNN	T0 - T1 - T2	175	0.903 [0.854,0.944]	0.817 [0.760,0.874]	0.796 [0.682,0.900]	0.826 [0.758,0.890]	0.672 [0.564,0.779]	0.811 [0.746,0.872]	0.0001
CRNN	T0 - T1 - T2	175	0.926 [0.878,0.969]	0.857 [0.806,0.909]	0.852 [0.754,0.944]	0.860 [0.795,0.916]	0.730 [0.619,0.833]	0.856 [0.799,0.912]	0.0090
globAttCNN	T0 - T1 - T2	175	0.914 [0.864,0.958]	0.851 [0.794,0.903]	0.833 [0.729,0.929]	0.860 [0.794,0.916]	0.726 [0.617,0.828]	0.846 [0.785,0.905]	0.0017
globAttCRNN	T0 - T1 - T2	175	**0.954** [0.922,0.983]	**0.897** [0.851,0.937]	**0.870** [0.774,0.959]	**0.909** [0.856,0.957]	**0.810** [0.707,0.905]	**0.890** [0.838,0.941]	-

Table 5.3: Quantitative performance of the implemented architectures assessed on the independent test set. For each metric, the 95% confidence interval is shown in brackets, and the highest value is highlighted in bold. The DeLong test was utilized to compare the AUCs of all models against the proposed architecture.

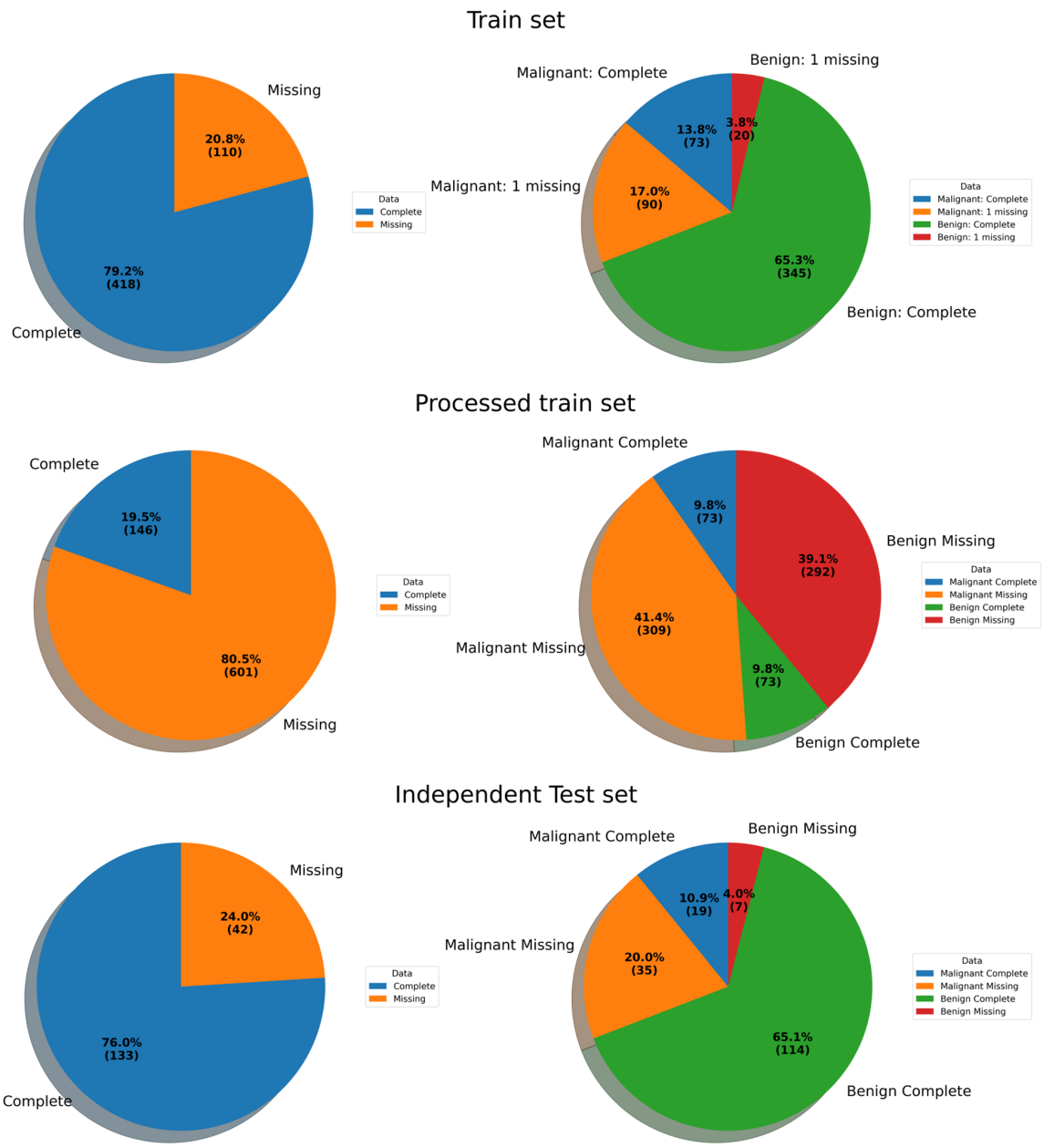


Figure 5.6: Pie charts displaying the distributions of complete and incomplete data (missing one time point) for benign and malignant nodules in the training set (first row), the training set after data preprocessing (second row) and independent test set (third row). Left image represents the distribution for the entire set, while the right image provide a detailed breakdown by nodule type.

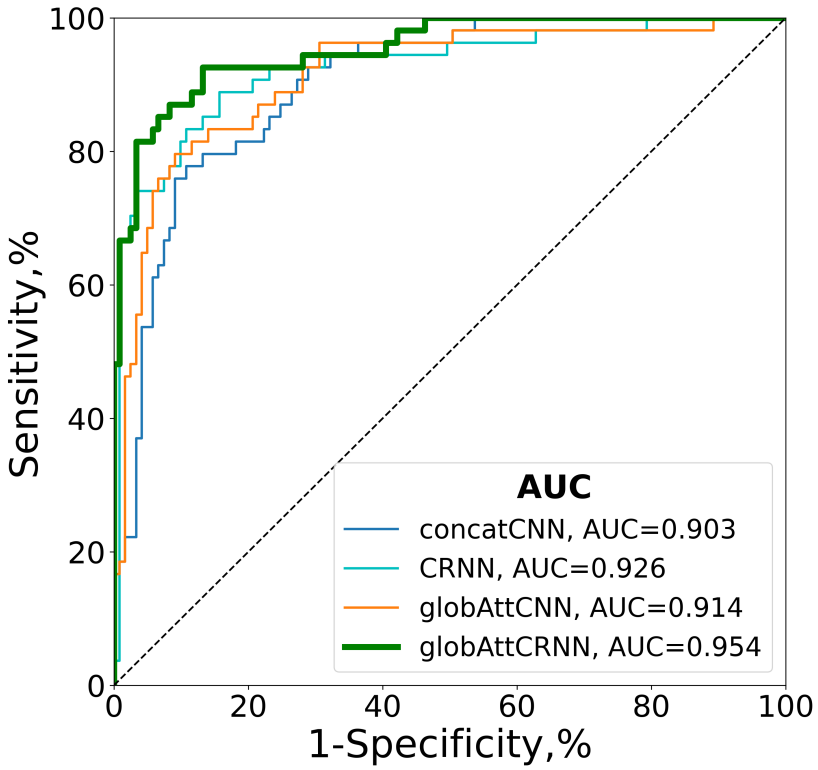


Figure 5.7: Comparison of ROC curves for all the models implemented as part of the ablation study.

Table 5.3 and Figure 5.7 present a comprehensive comparison of the different models implemented. Sequentially removing each component of our proposed globAttCRNN, we evaluated the performance of each model. The proposed model achieved superior results, with an AUC of 0.954, significantly outperforming CNNlast (DeLong test: p-value = 0.0064), concatCNN (DeLong test: p-value = 0.0014), CRNN (DeLong test: p-value = 0.0090), and globAttCNN (DeLong test: p-value = 0.0017) models.

The high performance on the independent test set underscores the model's robust ability to generalize across various distributions of missing data. It prioritizes capturing spatial changes in nodule characteristics over time rather than focusing on hidden data distributions.

Furthermore, as shown in (Table 5.4) globAttCRNN exhibited significantly superior performance in predicting cancer risk (patient-level analysis), achieving an AUC of 0.934 compared to the CNN Tlast model's AUC of 0.834 (DeLong test: p-value = 0.0390).

Figure 5.8 shows the Kaplan-Meier survival curves for the globAttCRNN and CNN Tlast models. Both models effectively stratified patients based on overall survival, yet the proposed model showed a superior ability to differentiate between low- and high-risk patients (log-rank p-value = 0.0002).

Model	Time instants	N patients test	AUC [95% CI]	ACC [95% CI]	SENS [95% CI]	SPES [95% CI]	PREC [95% CI]	BACC [95% CI]	De Long p-value
CNN	Tlast	110	0.834 [0.762,0.900]	0.836 [0.764,0.900]	0.804 [0.690,0.902]	0.864 [0.772,0.949]	0.837 [0.727,0.935]	0.834 [0.762,0.900]	0.0390
concatCNN	T0 - T1 - T2	110	0.873 [0.801,0.934]	0.809 [0.736,0.882]	0.784 [0.667,0.889]	0.831 [0.731,0.926]	0.800 [0.681,0.909]	0.807 [0.73,0.881]	0.001
CRNN	T0 - T1 - T2	110	0.903 [0.833,0.96]	0.836 [0.764,0.9]	0.843 [0.737,0.938]	0.831 [0.729,0.925]	0.811 [0.698,0.915]	0.837 [0.764,0.903]	0.033
globAttCNN	T0 - T1 - T2	110	0.892 [0.822,0.949]	0.836 [0.764,0.9]	0.824 [0.711,0.92]	0.847 [0.754,0.934]	0.824 [0.711,0.923]	0.835 [0.764,0.902]	0.008
globAttCRNN	T0 - T1 - T2	110	**0.934** [0.881,0.974]	**0.864** [0.800,0.927]	**0.863** [0.761,0.948]	**0.864** [0.772,0.947]	**0.846** [0.739,0.941]	**0.864** [0.793,0.926]	-

Table 5.4: Quantitative performance of the implemented architectures assessed on the independent test set at patient level. For each metric, the 95% confidence interval has been shown, and the highest value has been highlighted in bold. The DeLong test was utilized to compare the AUCs of all models against the proposed architecture.

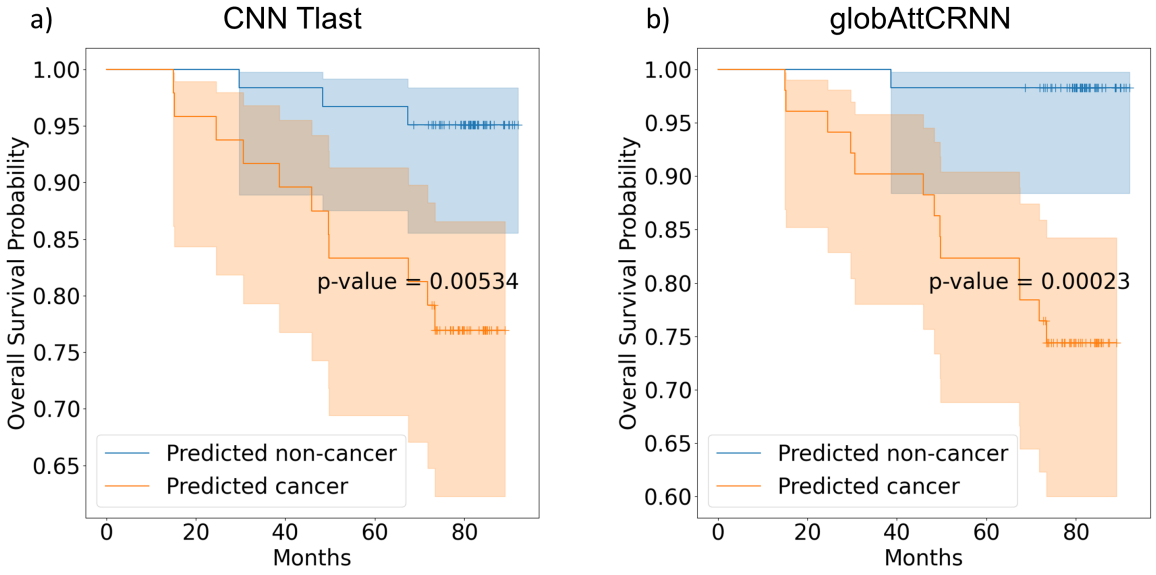


Figure 5.8: Kaplan-Meier survival curves on the independent test cohort for the best single- and multiple-time-points models: a) CNN Tlast and b) globAttCRNN

5.4.2 Model Interpretation

In this study, the use of the global attention module enabled the assessment of the contribution of each time point to the model output through the analysis of its attention weights (Figure 5.9). By assigning different weights to each image at each time instant, the model focused on the most informative time points, enhancing its ability to capture relevant features.

Moreover, Figure 5.10 show the distribution of activation weights for the last and second-to-last available time instants, respectively. Notably, the last available follow-up is consistently weighted higher than the previous scans. This observation is consistent with clinical practice, as radiologists often rely on the most recent observations to make informed decisions about nodule classification.

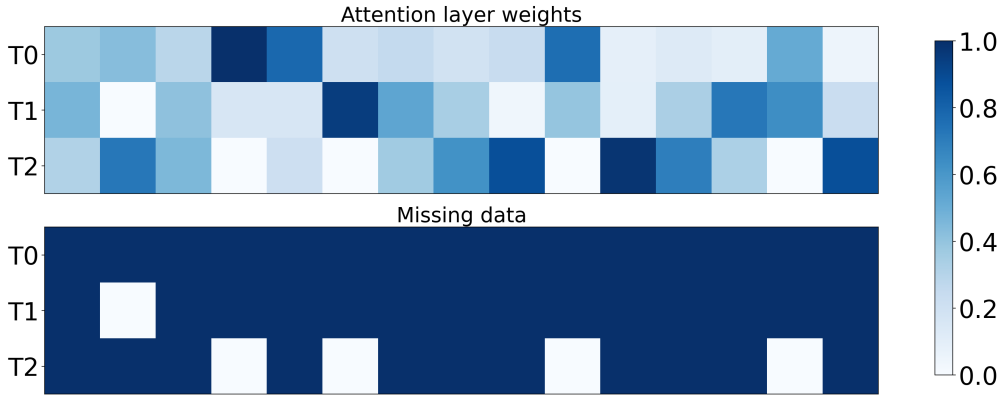


Figure 5.9: Example of the activation weights in a subset of 15 test nodules. First row represents the activation weights for T0, T1, and T2. Second row shows missing data for each nodule.

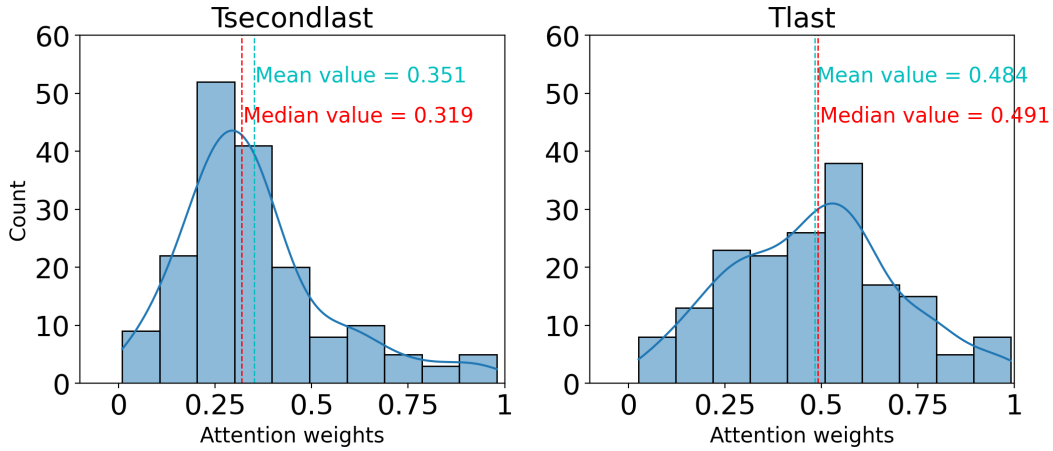


Figure 5.10: Distribution of the activation weights for the last (Tlast) and second-to-last (Tsecondtolast) available time instants.

It is worth to notice that also the 2D-CNN performed better using the last available time instant, aligning with the fact that it is the most representative time instant.

Figure 5.11 presents input data for the globAttCRNN model, alongside predictions and corresponding global attention weights. Classification confidence values are provided, ranging from 0 (high insecurity) to 1 (high confidence). The examples show that nodules with noticeable growth are more likely to be predicted malignant, while smaller nodules with minimal growth are more likely to be predicted benign. Importantly, the model performs well even in cases where malignant nodules exhibit subtle changes over time, addressing a challenging aspect of clinical diagnosis. The examples demonstrate that nodules that exhibit significant growth or noticeable changes over time tend to have a higher likelihood of being predicted malignant. On the contrary, smaller nodules that show minimal growth over time are more likely to be predicted benign. Importantly, the model exhibits robust performance in cases where malignant nodules undergo subtle changes over time, a particularly challenging aspect of clinical diagnosis.

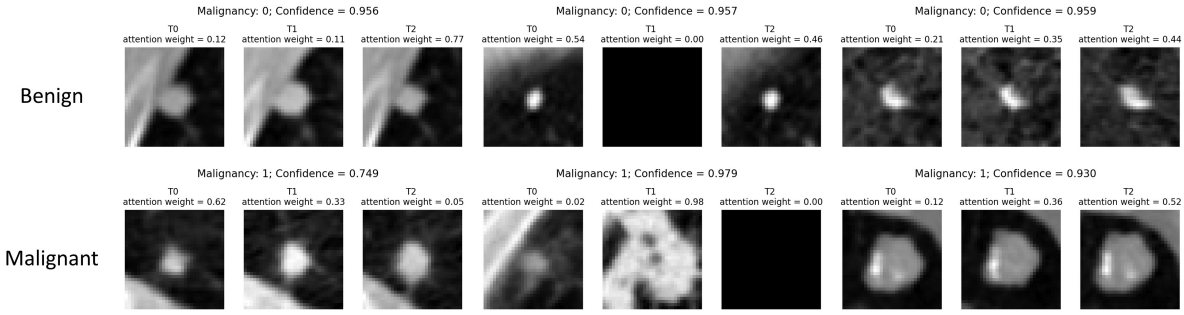


Figure 5.11: Visualization of some examples of the predictions of the globAttCRNN model with their classification confidence. The top row displays predictions for benign nodules, while the bottom row shows predictions for malignant nodules.

5.5 Discussion and Conclusion

In this study, we introduce a spatio-temporal deep learning framework that incorporates a temporal attention module. This framework is designed to predict indeterminate lung nodule malignancy using serial CT images acquired during the follow-up period of a lung cancer screening program.

We employed LDCT scans from the NLST cohort, focusing on patients with indeterminate nodules. Our globAttCRNN model utilizes all available CT scans per patient across the screening period to predict nodule malignancy probability within a year after the last scan. Our model integrates a CNN-based spatial feature extractor for encoding nodule images and an RNN with a temporal global attention module to capture temporal patterns. The entire network was designed to minimize the number of weights, ensuring efficient convergence and effective training. We explored alternative solutions using more conventional CNN models for feature extraction and a temporal block with additional layers and units, which would have resulted in a higher number of weights. Furthermore, we explored the use of 3D networks instead of 2D networks; however, these experiments did not show any improvement. It could be attributed to the inherent characteristic of screening nodules, which are typically small and exhibit limited heterogeneity along the z-axis. It is worth mentioning that our framework can easily accommodate more than three temporal instants with minimal adjustments to the spatial and temporal network blocks. These approaches were not presented as they exhibited significant performance degradation due to overfitting.

Our longitudinal model achieved remarkable results in an independent test set, with an AUC of 0.954, ACC of 0.897, SENS of 0.870, and SPEC of 0.909. These metrics outperformed single-time models, highlighting the clinical significance of leveraging longitudinal data for enhanced nodule malignancy prediction.

To mitigate temporal biases in model training, we carefully curated and preprocessed the data, ensuring its reliability and consistency. We proposed two approaches to handle missing data and achieve a balanced temporal distribution for both benign and malignant nodules, as they exhibit distinct missing data patterns. Data augmentation increased the diversity of training data, reducing the risk of overfitting and improving generalization. Specifically, we introduced a temporal data augmentation for malignant nodules to increase their representation in the training set. In addition,

a temporal dropout technique was applied to benign nodules to harmonize their distribution of complete and incomplete data with that of malignant nodules. The successful use of these techniques highlights the importance of addressing the challenge of missing data in longitudinal analyses.

The clinical relevance of our proposed model is greatly enhanced by its interpretability. The inclusion of a global attention module facilitated the assessment of each time point's contribution to the model's output through attention weights. By assigning time-specific weights, the model prioritized the most informative time points, enhancing its ability to capture relevant features and improve prediction accuracy. Furthermore, the distribution of activation weights revealed a consistent prioritization of the last available follow-up, which is consistent with standard clinical practice where the latest observations heavily influence nodule classification.

Visual examples further demonstrated the model's ability to distinguish between malignant and benign nodules based on their temporal changes. Nodules with substantial growth or noticeable alterations were more likely to be predicted malignant. In contrast, smaller nodules with minimal growth tended to be benign. Moreover, the model effectively identified malignant nodules undergoing subtle changes, offering promise in addressing the uncertainties frequently encountered by radiologists in such cases. Leveraging the capabilities of this model could provide clinicians with valuable support, potentially mitigating ambiguity in challenging diagnostic scenarios.

Comparisons can be made with previous studies on lung nodule classification. However, it's crucial to highlight that our focus was specifically on utilizing indeterminate nodules, known for their increased difficulty in prediction. Fair comparisons should be conducted using the same test data. In Veasey et al. (2020), they reported an AUC of 0.882 for their best longitudinal models using three time instants for the prediction of nodule malignancy. Their study focused on malignant nodules with 2 or 3 time instants, whereas benign nodules were limited to 3 time instants without implementing any precautions to address potential biases. In our experience this approach may be more susceptible to learning temporal biases in the data, resulting in a higher likelihood of predicting images with fewer time instants as malignant. Furthermore, our proposed model results in a significantly lower number of parameters (approximately 1k) compared to the 18.4M parameters reported in their paper. Additionally, Ardila et al. (2019) implemented a longitudinal model with two temporal instants, achieving an AUC of 0.944, SENS of 0.837, and SPEC of 0.950 in cancer prediction at patient level. Upon evaluating their predictions on a subset of 16 cases matching our test set, considering only the last available CT scan for each patient, our model achieved an ACC of 0.938, SENS of 1.0 and SPEC of 0.857 compared to their ACC of 0.938, SENS of 0.889 and SPEC of 1.0. Notably, our model achieved better sensitivity despite not being specifically trained for patient-level cancer prediction. Importantly, their model allowed the input to be either a single CT scan or both the current and previous scans. Although our model could also accommodate this flexibility, we intentionally chose to focus on cases with at least two time points. Our specific emphasis was on indeterminate nodules, encompassing positive baseline scans with follow-ups or incidental nodules, as the latter pose the most significant challenge to radiologists during screening. The same 16 patients were present in the test set of Mikhael et al. (2023). Despite their use of only one image per patient, they achieved an ACC of 0.875, SENS of 0.6, and SPEC of 1.0. Also in this case, our model outperformed their cancer prediction. It is noteworthy that in the previous two studies, they utilized all patients available in the NLST dataset, resulting in a significant

disproportion between cancer and non-cancer cases (3.7% and 4.8%, respectively). This could potentially lead the model to be more specific than sensitive.

Overall, our proposed model demonstrates superior performance compared to these two previous studies using multiple time points for the classification of lung nodules. Incorporating precautions to address potential biases, along with the utilization of a feature extractor with a small parameter configuration, contributes to the improved accuracy and effectiveness of our model in capturing relevant features and predicting malignancy.

We acknowledge several limitations in our study. Firstly, despite implementing techniques like data augmentation and temporal dropout to address missing data, handling missing data in longitudinal models remains a complex challenge. Inherent biases and limitations in addressing missing data could potentially impact the model's performance and its ability to generalize. Secondly, while our model's performance was evaluated using an independent test set, external validation on diverse datasets from various screening programs or institutions was not possible due to the lack of available longitudinal open datasets for comparison. Future research with prospective multicenter cohorts is needed to address these limitations comprehensively.

In conclusion, the main contribution of our work is the development a spatio-temporal deep neural network for the prediction of indeterminate nodule malignancy using serial CT images. The proposed architecture, featuring a temporal global attention module, effectively captures both the spatial and temporal dynamics of lung nodules, resulting in improved performance in the prediction of nodule malignancy. Notably, our findings demonstrate the model's proficiency in revealing meaningful temporal malignancy patterns over time, overcoming the limitations associated with models reliant solely on single CT scans. Moreover, the integration of data augmentation and temporal dropout techniques mitigated temporal biases, improving the model's robustness and generalization capabilities. Our evaluation emphasizes its potential as a valuable tool for the diagnosis and stratification of patients at risk of lung cancer.

CHAPTER 6

Concluding Remarks

6.1 Discussion and Conclusion

The concept of precision medicine in lung cancer has gained significant attention in recent years due to its potential to provide personalized prevention, diagnosis, and treatment strategies tailored to individual patient characteristics. Given the low 5-year survival rates for patients diagnosed at late stages, there is a critical need for effective screening methods to detect cancer at early stages. Emerging treatments such as immunotherapy offer new hope for managing lung cancer, particularly among patients in advanced stages. However, determining which patients will respond to these therapies remains challenging.

Medical imaging plays a crucial role in precision medicine by enabling screening, early detection, treatment response evaluation, and recurrence assessment. Its non-invasive nature, ability to generate 3D visual representations of the body and disease, longitudinal tracking of patient progression, and cost-effectiveness make it uniquely suited for this purpose. Radiomics and artificial intelligence show great promise in extracting valuable information from medical images, revealing intricate imaging features that may not be discernible to the human eye. These features have the potential to serve as biomarkers with significant implications for clinical practice, enhancing the accuracy of diagnosis, prognosis, and overall disease assessment.

Chapter 3 described a method for the prediction of treatment response in patients with advanced lung cancer undergoing immunotherapy. We proposed different novelties in the methodology for the prediction of response. First, we hypothesized that the use of longitudinal data could give more information about patient evolution through the immunotherapy, mainly considering the fact that this new treatment may introduce new response patterns that are related to the immune related response and that had not been seen previously with other treatments. In fact, patients are likely to experience the so called pseudoprogression, where the tumor can grow in the first months of treatment even if they are responding to the treatment (M. Y. Chen & Zeng, 2022). We demonstrated that the use of longitudinal data significantly improves the prediction performance

over baseline model, highlighting the importance of using patient information during early treatment to have an early confirmation of response. Furthermore, as the response to treatment is closely related to patient's clinical condition, we demonstrated that integrating both imaging and clinical data allows a better understanding of treatment response. Additionally, we introduced a deep learning architecture to extract imaging features closely associated to treatment response instead of more general features that can be extracted from the tumor through radiomics. The drawback of this methodology is that the features that have been extracted could only be applied to this problem and can not be generalized to all the tumors.

Chapter 4 examined the robustness of imaging features against various CT acquisition parameters and noise. In this study, we chose to focus solely on traditional radiomics features, given their fixed nature and independence from the specific problem at hand. Our investigation revealed distinct sources of variation for radiomics features, notably highlighting the strong dependency on CT scanner (manufacturer) and CT scan inter-slice resolution (slice thickness). Conversely, features appeared to exhibit greater robustness to X-ray dosage (kVp) and image noise (stdNoise). We hypothesized that image harmonization techniques could alleviate the impact of these batch effects on radiomics features, with our chosen methods notably reducing dependency on the CT scanner. However, complete correction of the dependence of batch-effects was not achieved, leading us to apply a nestedComBat feature harmonization approach to mitigate residual dependencies. Furthermore, we explored whether the attainment of more robust features corresponded to enhanced response precision of immunotherapy. Performance improvement trends were observed but there were not statistically significant.

The longitudinal methods employed in these chapters primarily relied on traditional machine learning models, leveraging temporal information by concatenating all data from each time instant into a single array. This approach was motivated by the small dataset, which was unsuitable for implementing a deep learning model. To fully harness both spatial and temporal information, we utilized a corresponding cohort from lung screening trials, where spatio-temporal data on indeterminate nodules extracted from CT scans during screening were available. In Chapter 5, we introduced a novel spatio-temporal architecture capable of effectively processing and exploiting this spatio-temporal information, with a particular focus on the most representative time instances facilitated by an attention mechanism. Additionally, we underscored the importance of handling missing data carefully, as the architecture's performance could be influenced by patterns in the missing data. Through the application of specific techniques, we successfully addressed this challenge.

6.2 Contributions

The major contributions of this work are:

- Collection of a three-institution dataset comprising a total of 226 patients undergoing immunotherapy. This dataset includes CT images along with tumor segmentation, complemented by comprehensive clinical information for each patient. Specifically, CT images were obtained both before and during treatment. Among the 226 patients, longitudinal data were

available for 194 patients, making this dataset unique in its composition.

- Collection of a dataset consisting of 443 patients and 703 nodules, derived from follow-up scans conducted during screening trials. Notably, this dataset exclusively features longitudinal images of indeterminate nodules, which pose significant challenges for clinical characterization. The positions and diagnoses of these indeterminate nodules will be publicly accessible, contributing to the field of spatio-temporal analysis. This dataset is unique and has not been previously made available.
- Design, implementation and validation of an ensemble model that could integrate longitudinal radiomics and clinical data that was able to predict durable clinical benefit of immunotherapy at 6 and 9 months after treatment. Furthermore, we demonstrated the added value of radiomics features extracted with a deep learning approach instead of the traditional one which used hand-crafted features.
- Introduction of a novel pipeline to ensure radiomics robustness with respect to technical acquisition parameters. This pipeline combines image and feature harmonization strategies and has been validated to be effective in predicting immunotherapy response.
- Design, implementation and validation of a spatio-temporal deep neural network for predicting indeterminate nodule malignancy using serial CT images obtained during screening trials. We introduced preprocessing techniques such as class-based temporal data augmentation and temporal dropout to handle missing data effectively.

6.3 Future Work

The primary objective of this thesis is to leverage both spatial and temporal information derived from longitudinal CT images for addressing challenges in immunotherapy treatment prediction and lung nodule screening.

In the context of immunotherapy treatment prediction, in Chapter 3 we have demonstrated the significance of integrating multimodal data to enhance our understanding of treatment response patterns. However, our analysis was limited to two sources of information: imaging data (CT images) and clinical data (patient demographics and blood test results). Incorporating additional data sources such as features extracted from pathology images or genomics data could potentially enhance prediction accuracy. Further research is guaranteed, particularly in the context of a new retrospective study, which will provide additional sources of information for integration.

Moreover, due to our small sample size, we were constrained in training more complex deep learning models. Such models could offer dual benefits: extracting more representative features and building more effective predictive models. Despite addressing the sample size limitation by utilizing a pretrained approach, it may not be optimal to pretrain a network for extracting features from advanced lung tumors with data from early cancer nodules which may present different appearance. Hence, exploring the training of new architectures in a self- or unsupervised manner to generate deep features capable of generalizing across lesions of the same imaging type could be advantageous over creating deep features tailored to specific outcomes.

Due to the high heterogeneity of our dataset and the extensive literature demonstrating the low reproducibility of radiomics features, in Chapter 4 we conducted an analysis to assess how confounding factors related to image acquisition (including manufacturer, slice thickness, and kVp) and noise impacted the stability of radiomics features and their ability to assess response. While we considered the effect of X-ray doses through X-ray tube voltage on radiomics stability, we did not evaluate the potential impact of contrast agents on this stability. This is particularly intriguing because CT intensities in regions with high contrast, such as the aorta, exhibit significant variability between subjects. We hypothesize that similar variability may be observed in tumor regions, suggesting that higher contrast injection may result in elevated values in the tumor not solely attributable to tumor characteristics but also influenced by the contrast agent. Thus, compensating for the effect of contrast administration when comparing tumor distributions over time is essential. However, estimating this effect is challenging due to various factors influencing the amount of contrast that is perfusing the tumor, including the type and composition of the contrast agent, specific tumor characteristics (e.g., vascularity, size, location), microenvironment, and lymphatic system involvement. A deeper understanding of these mechanisms could enhance the calibration of CT intensities. In our study, we simplified this process by aligning tumor probability density functions (PDFs) to the same mode (54 HU). Moreover, with access to larger datasets, it would be advantageous to explore newer image harmonization approaches based on deep learning self-supervised techniques, such as the one introduced by Cackowski et al. (2023). This model, originally developed for MRI, has shown superior performance compared to other methods. It excels in generating high-quality images from traveling subjects, effectively eliminating biases from different sites or scanners while enhancing patient classification. Additionally, it harmonizes data from new sites or scanners without requiring additional fine-tuning. Furthermore, considering the longitudinal nature of our dataset, as recommended by Beer et al. (2020), could yield further enhancements in feature harmonization. This approach may facilitate a more precise estimation of parameters, enabling the alignment of longitudinal features based on individual patients rather than the entire dataset. Such refinements hold the potential to significantly improve the robustness and effectiveness of our harmonization techniques.

Moreover, all the models developed for predicting immunotherapy treatment response relied on a straightforward approach of aggregating features from the same tumor over time. This simplicity was required by the limited size of our dataset. In contrast, for predicting indeterminate nodule malignancy, in Chapter 5 we employed a more complex architecture due to the availability of a larger dataset, which was better suited for deep learning analysis. However, despite our efforts to focus on a specific type of nodule, which reduced the dataset size, it would be advantageous to validate our model on larger datasets and extend its applicability to nodules with only one available scan, including prevalent nodules. Although our architecture was designed to accommodate varying numbers of time points, we were constrained to training it with only three time instants, with the requirement that each nodule had at least two time points available. The dataset size also constrained us from developing CNN-based models for feature extraction from tumors. Despite adopting transfer learning techniques, training more robust features in larger datasets is imperative. We are in the process of acquiring access to a larger dataset to explore more sophisticated and data-greedy architectures capable of handling spatio-temporal data, such as transformers, as proposed by Ahmadi et al. (2023).

Publications

Journal Articles

Farina, B., Ramos Guerra, A. D., Bermejo-Peláez, D., Palacios Miras, C., Alcazar Peral, A., Gallardo Madueño, G., Corral Jaime J., Vilalta-Lacarra, A., Rubio Pérez J., Muñoz-Barrutia, A., Peces-Barba, G. R., Seijo Maceiras, L., Gil-Bazo, I., Dómine Gómez, M. and Ledesma-Carbayo, M.J. (2023). "Integration of longitudinal deep-radiomics and clinical data improves the prediction of durable benefits to anti-PD-1/PD-L1 immunotherapy in advanced NSCLC patients.".In: *Journal of Translational Medicine*, 21.1, pp.1-15. ISSN: 1479-5876. DOI: 10.1186/s12967-023-04004-x.

Farina, B., Carbajo Benito, R., Montalvo-García, D., Bermejo-Peláez, D., Seijo Maceiras, L., Ledesma-Carbayo, M.J. "A spatio-temporal deep learning framework with temporal global attention for nodule malignancy prediction from longitudinal screening CT images." submitted to Computerized Medical Imaging and Graphics.

Farina, B., Vegas-Sánchez-Ferrero, G., Ramos Guerra, A. D., Bermejo-Peláez, D., Palacios Miras, C., Alcazar Peral, A., Gallardo Madueño, G., Corral Jaime J., Vilalta-Lacarra, A., Rubio Pérez J., Zugazagoitia, J., Peces-Barba, G. R., Seijo Maceiras, L., Paz-Ares, L., Gil-Bazo, I., Dómine Gómez, M., San José Estépar, R. and Ledesma-Carbayo, M.J. "Image and feature harmonization in radiomics for mitigation of heterogeneity in CT image parameters: impact on a longitudinal study of advanced NSCLC patients treated with anti-PD-1/PD-L1 immunotherapy." to be submitted to Computer Methods and Programs in Biomedicine.

Ramos Guerra, A. D., **Farina, B.**, Peces-Barba, G. R., Seijo Maceiras, L., Corral Jaime J., Gil-Bazo, I., Zugazagoitia, J., Paz-Ares, L., Rubio Pérez J.,Dómine Gómez, M., and Ledesma-Carbayo, M.J. "Multivariate longitudinal Bayesian Joint Modelling of real-world data to predict survival and response to immunotherapy in advanced Non-small Cell Lung Cancer." to be submitted to Clinical Cancer Research.

Conference Articles

Farina, B., Ramos Guerra, A. D., Palacios Miras, C., Gallardo Madueño, G., Corral Jaime J., Muñoz-Barrutia, A., Peces-Barba, G. R., Seijo Maceiras, L., Gil-Bazo, I., Dómine Gómez, M. and Ledesma-Carbayo, M.J. (2021). "Delta-Radiomics Signature For Prediction Of Survival In Advanced Nsclc Patients Treated With Immunotherapy.".In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI 2021)*, pp.886-890. ISSN: 1945-8452. DOI: 10.1109/ISBI48211.2021.9434076.

Corral Bolaños, M., **Farina, B.**, Ramos Guerra, A. D., Palacios Miras, C., Gallardo Madueño, G., Corral Jaime J., Muñoz-Barrutia, A., Peces-Barba, G. R., Seijo Maceiras, L., Gil-Bazo, I., Dómine Gómez, M. and Ledesma-Carbayo, M.J. (2020). "Delta-Radiomics Signature For Prediction Of Survival In Advanced NSCLC Patients Treated With Immunotherapy.".In: *XXXVIII Congreso Anual de la Sociedad Española de Ingeniería Biomédica. CASEIB 2020: Libro de actas*, pp.181-184.

References

Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, *5*(1), 4006. https://doi.org/10.1038/ncomms5006

Ahmadi, N., Tsang, M., Gu, A., Tsang, T., & Abolmaesumi, P. (2023). Transformer-based spatio-temporal analysis for classification of aortic stenosis severity from echocardiography cine series. *IEEE Transactions on Medical Imaging*. https://doi.org/10.1109/TMI.2023.3305384

Alahmari, S. S., Cherezov, D., Goldgof, D. B., Hall, L. O., Gillies, R. J., & Schabath, M. B. (2018). Delta radiomics improves pulmonary nodule malignancy prediction in lung cancer screening. *IEEE Access*, *6*, 77796–77806. https://doi.org/10.1109/ACCESS.2018.2884126

Albano, D., Benenati, M., Bruno, A., Bruno, F., Calandri, M., Caruso, D., Cozzi, D., De Robertis, R., Gentili, F., Grazzini, I., et al. (2021). Imaging side effects and complications of chemotherapy and radiation therapy: A pictorial review from head to toe. *Insights into Imaging*, *12*(1), 76. https://doi.org/10.1186/s13244-021-01017-2

Allemani, C., Matsuda, T., Di Carlo, V., Harewood, R., Matz, M., Nikšić, M., Bonaventure, A., Valkov, M., Johnson, C. J., Estève, J., et al. (2018). Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): Analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet*, *391*(10125), 1023–1075. https://doi.org/10.1016/S0140-6736(17)33326-3

Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, *97*(18), 10101–10106. https://doi.org/10.1073/pnas.97.18.10101

Amador, K., Wilms, M., Winder, A., Fiehler, J., & Forkert, N. D. (2022). Predicting treatment-specific lesion outcomes in acute ischemic stroke from 4D CT perfusion imaging using spatio-temporal convolutional neural networks. *Medical Image Analysis*, *82*, 102610. https://doi.org/10.1016/j.media.2022.102610

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, *49*(268), 765–769.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, *25*(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x

Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., et al. (2011). The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, *38*(2), 915–931. https://doi.org/10.1118/1.3528204

Armato III, S. G., Meyer, C. R., McNitt-Gray, M. F., McLennan, G., Reeves, A., Croft, B. Y., Clarke, L. P., & Group, R. R. (2008). The reference image database to evaluate response to therapy in lung cancer (rider) project: A resource for the development of change-analysis software. *Clinical Pharmacology & Therapeutics*, *84*(4), 448–456. https://doi.org/10.1038/clpt.2008.161

Bai, R., Lv, Z., Xu, D., & Cui, J. (2020). Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors. *Biomarker Research*, *8*, 1–17. https://doi.org/10.1186/s40364-020-00209-0

Baumann, M., Krause, M., Overgaard, J., Debus, J., Bentzen, S. M., Daartz, J., Richter, C., Zips, D., & Bortfeld, T. (2016). Radiation oncology in the era of precision medicine. *Nature Reviews Cancer*, *16*(4), 234–249. https://doi.org/10.1038/nrc.2016.18

Beer, J. C., Tustison, N. J., Cook, P. A., Davatzikos, C., Sheline, Y. I., Shinohara, R. T., Linn, K. A., Initiative, A. D. N., et al. (2020). Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. *Neuroimage*, *220*, 117129. https://doi.org/10.1016/j.neuroimage.2020.117129

Beig, N., Khorrami, M., Alilou, M., Prasanna, P., Braman, N., Orooji, M., Rakshit, S., Bera, K., Rajiah, P., Ginsberg, J., et al. (2019). Perinodular and intranodular radiomic features on lung CT images distinguish adenocarcinomas from granulomas. *Radiology*, *290*(3), 783–792. https://doi.org/10.1148/radiol.2018180910

Berghmans, T., Durieux, V., Hendriks, L. E., & Dingemans, A.-M. (2020). Immunotherapy: From advanced nsclc to early stages, an evolving concept. *Frontiers in Medicine*, *7*, 90. https://doi.org/10.3389/fmed.2020.00090

Blons, H., Garinet, S., Laurent-Puig, P., & Oudart, J.-B. (2019). Molecular markers and prediction of response to immunotherapy in non–small cell lung cancer, an update. *Journal of thoracic disease*, *11*(Suppl 1), S25. https://doi.org/10.21037/jtd.2018.12.48

Borcoman, E., Kanjanapan, Y., Champiat, S., Kato, S., Servois, V., Kurzrock, R., Goel, S., Bedard, P., & Le Tourneau, C. (2019). Novel patterns of response under immunotherapy. *Annals of Oncology*, *30*(3), 385–396. https://doi.org/10.1093/annonc/mdz003

Borghaei, H., Paz-Ares, L., Horn, L., Spigel, D. R., Steins, M., Ready, N. E., Chow, L. Q., Vokes, E. E., Felip, E., Holgado, E., et al. (2015). Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer. *New England Journal of Medicine*, *373*(17), 1627–1639. https://doi.org/10.1056/NEJMoa1507643

Brancato, V., Cerrone, M., Lavitrano, M., Salvatore, M., & Cavaliere, C. (2022). A systematic review of the current status and quality of radiomics for glioma differential diagnosis. *Cancers*, *14*(11), 2731. https://doi.org/10.3390/cancers14112731

Brenner, D. J. (2004). Radiation risks potentially associated with low-dose CT screening of adult smokers for lung cancer. *Radiology*, *231*(2), 440–445. https://doi.org/10.1148/radiol.2312030880

Broderick, S. R. (2020). Adjuvant and neoadjuvant immunotherapy in non–small cell lung cancer. *Thoracic surgery clinics*, *30*(2), 215–220. https://doi.org/10.1016/j.thorsurg.2020.01.001

Cackowski, S., Barbier, E. L., Dojat, M., & Christen, T. (2023). Imunity: A generalizable vae-gan solution for multicenter mr image harmonization. *Medical Image Analysis*, *88*, 102799. https://doi.org/10.1016/j.media.2023.102799

Califf, R. M. (2018). Biomarker definitions and their applications. *Experimental Biology and Medicine*, *243*(3), 213–221. https://doi.org/10.1177/1535370217750088

Causey, J. L., Zhang, J., Ma, S., Jiang, B., Qualls, J. A., Politte, D. G., Prior, F., Zhang, S., & Huang, X. (2018). Highly accurate model for prediction of lung nodule malignancy with ct scans. *Scientific Reports*, *8*(1), 9286. https://doi.org/10.1038/s41598-018-27569-w

Chelala, L., Hossain, R., Kazerooni, E. A., Christensen, J. D., Dyer, D. S., & White, C. S. (2021). Lung-rads version 1.1: Challenges and a look ahead, from the ajr special series on radiology reporting and data systems. *American Journal of Roentgenology*, *216*(6), 1411–1422. https://doi.org/10.2214/AJR.20.24807

Chen, B. T., Chen, Z., Ye, N., Mambetsariev, I., Fricke, J., Daniel, E., Wang, G., Wong, C. W., Rockne, R. C., Colen, R. R., et al. (2020). Differentiating peripherally-located small cell lung cancer from non–small cell lung cancer using a ct radiomic approach. *Frontiers in Oncology*, *10*, 593. https://doi.org/10.3389/fonc.2020.00593

Chen, H., Han, Z., Luo, Q., Wang, Y., Li, Q., Zhou, L., & Zuo, H. (2022). Radiotherapy modulates tumor cell fate decisions: A review. *Radiation Oncology*, *17*(1), 196. https://doi.org/10.1186/s13014-022-02171-7

Chen, M. Y., & Zeng, Y.-C. (2022). Pseudoprogression in lung cancer patients treated with immunotherapy. *Critical reviews in oncology/hematology*, *169*, 103531. https://doi.org/10.1016/j.critrevonc.2021.103531

Choi, W., Oh, J. H., Riyahi, S., Liu, C.-J., Jiang, F., Chen, W., White, C., Rimner, A., Mechalakos, J. G., Deasy, J. O., et al. (2018). Radiomics analysis of pulmonary nodules in low-dose ct for early detection of lung cancer. *Medical Physics*, *45*(4), 1537–1549. https://doi.org/10.1002/mp.12820

Ciompi, F., Chung, K., Van Riel, S. J., Setio, A. A. A., Gerke, P. K., Jacobs, C., Scholten, E. T., Schaefer-Prokop, C., Wille, M. M., Marchiano, A., et al. (2017). Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Scientific Reports*, *7*(1), 46479. https://doi.org/10.1038/srep46479

Clarke, L. P., Nordstrom, R. J., Zhang, H., Tandon, P., Zhang, Y., Redmond, G., Farahani, K., Kelloff, G., Henderson, L., Shankar, L., et al. (2014). The quantitative imaging network: NCI's historical perspective and planned goals. *Translational Oncology*, *7*(1), 1–4. https://doi.org/10.1593/tlo.13832

Collins, L. G., Haines, C., Perkel, R., & Enck, R. E. (2007). Lung cancer: Diagnosis and management. *American family physician*, *75*(1), 56–63.

Cousin, F., Louis, T., Dheur, S., Aboubakar, F., Ghaye, B., Occhipinti, M., Vos, W., Bottari, F., Paulus, A., Sibille, A., et al. (2023). Radiomics and delta-radiomics signatures to predict response and survival in patients with non-small-cell lung cancer treated with immune checkpoint inhibitors. *Cancers*, *15*(7), 1968. https://doi.org/10.3390/cancers15071968

Da-Ano, R., Masson, I., Lucia, F., Doré, M., Robin, P., Alfieri, J., Rousseau, C., Mervoyer, A., Reinhold, C., Castelli, J., et al. (2020). Performance comparison of modified combat for harmonization of radiomic features for multicenter studies. *Scientific Reports*, *10*(1), 10248. https://doi.org/10.1038/s41598-020-66110-w

Da-Ano, R., Visvikis, D., & Hatt, M. (2020). Harmonization strategies for multicenter radiomics investigations. *Physics in Medicine & Biology*, *65*(24), 24TR02. https://doi.org/10.1088/1361-6560/aba798

Dai, X., Gakidou, E., & Lopez, A. D. (2022). Evolution of the global smoking epidemic over the past half century: Strengthening the evidence base for policy action. *Tobacco control*, *31*(2), 129–137. https://doi.org/10.1136/tobaccocontrol-2021-056535

De Angelis, R., Sant, M., Coleman, M. P., Francisci, S., Baili, P., Pierannunzio, D., Trama, A., Visser, O., Brenner, H., Ardanaz, E., et al. (2014). Cancer survival in europe 1999-2007 by country and age: Results of EUROCARE-5—a population-based study. *Lancet Oncology*, *15*(1), 23–34. https://doi.org/10.1016/S1470-2045(13)70546-1

de Koning, H. J., van Der Aalst, C. M., de Jong, P. A., Scholten, E. T., Nackaerts, K., Heuvelmans, M. A., Lammers, J.-W. J., Weenink, C., Yousaf-Khan, U., Horeweg, N., et al. (2020). Reduced lung-cancer mortality with volume CT screening in a randomized trial. *The New England Journal of Medicine*, *382*(6), 503–513. https://doi.org/10.1056/NEJMoa1911793

Dercle, L., McGale, J., Sun, S., Marabelle, A., Yeh, R., Deutsch, E., Mokrane, F.-Z., Farwell, M., Ammari, S., Schoder, H., et al. (2022). Artificial intelligence and radiomics: Fundamentals, applications, and challenges in immunotherapy. *Journal for Immunotherapy of Cancer*, *10*(9). https://doi.org/10.1136/jitc-2022-005292

Dong, A., Zhao, Y., Li, Z., & Hu, H. (2021). PD-L1 versus tumor mutation burden: Which is the better immunotherapy biomarker in advanced non–small cell lung cancer? *The Journal of Gene Medicine*, *23*(2), e3294. https://doi.org/10.1002/jgm.3294

Doroshow, D. B., Sanmamed, M. F., Hastings, K., Politi, K., Rimm, D. L., Chen, L., Melero, I., Schalper, K. A., & Herbst, R. S. (2019). Immunotherapy in non–small cell lung cancer: Facts and hopes. *Clinical Cancer Research*, *25*(15), 4592–4602. https://doi.org/10.1158/1078-0432.CCR-18-1538

Dyba, T., Randi, G., Bray, F., Martos, C., Giusti, F., Nicholson, N., Gavin, A., Flego, M., Neamtiu, L., Dimitrova, N., et al. (2021). The european cancer burden in 2020: Incidence and mortality estimates for 40 countries and 25 major cancers. *European Journal of Cancer*, *157*, 308–347. https://doi.org/10.1016/j.ejca.2021.07.039

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al. (2009). New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). *European Journal of Cancer*, *45*(2), 228–247. https://doi.org/10.1016/j.ejca.2008.10.026

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al. (2012). 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, *30*(9), 1323–1341. https://doi.org/10.1016/j.mri.2012.05.001

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, *161*, 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Fu, F., Deng, C., Wen, Z., Gao, Z., Zhao, Y., Han, H., Zheng, S., Wang, S., Li, Y., Hu, H., et al. (2021). Systemic immune-inflammation index is a stage-dependent prognostic factor in patients with operable non–small cell lung cancer. *Translational Lung Cancer Research*, *10*(7), 3144. https://doi.org/10.21037/tlcr-21-267

Gandhi, L., Rodríguez-Abreu, D., Gadgeel, S., Esteban, E., Felip, E., De Angelis, F., Domine, M., Clingan, P., Hochmair, M. J., Powell, S. F., et al. (2018). Pembrolizumab plus chemotherapy in metastatic non–small-cell lung cancer. *New England journal of medicine*, *378*(22), 2078–2092. https://doi.org/10.1056/NEJMoa1801005

Gao, N., Tian, S., Li, X., Huang, J., Wang, J., Chen, S., Ma, Y., Liu, X., & Guo, X. (2020). Three-dimensional texture feature analysis of pulmonary nodules in ct images: Lung cancer predictive models based on support vector machine classifier. *Journal of Digital Imaging*, *33*, 414–422. https://doi.org/10.1007/s10278-019-00238-8

Garcia-Alvarez, A., Cubero, J. H., & Capdevila, J. (2022). What is the status of immunotherapy in neuroendocrine neoplasms? *Current Oncology Reports*, *24*(4), 451–461. https://doi.org/10.1007/s11912-022-01235-x

Ghaffari Laleh, N., Ligero, M., Perez-Lopez, R., & Kather, J. N. (2023). Facts and hopes on the use of artificial intelligence for predictive immunotherapy biomarkers in cancer. *Clinical Cancer Research*, *29*(2), 316–323. https://doi.org/10.1158/1078-0432.CCR-22-0390

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology*, *278*(2), 563–577. https://doi.org/10.1148/radiol.2015151169

Gong, J., Bao, X., Wang, T., Liu, J., Peng, W., Shi, J., Wu, F., & Gu, Y. (2022). A short-term follow-up ct based radiomics approach to predict response to immunotherapy in advanced non–small-cell lung cancer. *Oncoimmunology*, *11*(1), 2028962. https://doi.org/10.1080/2162402X.2022.2028962

Gridelli, C., Peters, S., Mok, T., Forde, P., Reck, M., Attili, I., & de Marinis, F. (2022). First-line immunotherapy in advanced non–small-cell lung cancer patients with ecog performance status 2: Results of an international expert panel meeting by the italian association of thoracic oncology. *ESMO open*, *7*(1), 100355. https://doi.org/10.1016/j.esmoop.2021.100355

Grossmann, P., Stringfield, O., El-Hachem, N., Bui, M. M., Rios Velazquez, E., Parmar, C., Leijenaar, R. T., Haibe-Kains, B., Lambin, P., Gillies, R. J., et al. (2017). Defining the biological basis of radiomic phenotypes in lung cancer. *eLife*, *6*, e23421. https://doi.org/10.7554/eLife.23421

Hammer, M. M., Byrne, S. C., & Kong, C. Y. (2022). Factors influencing the false positive rate in CT lung cancer screening. *Academic Radiology*, *29*, S18–S22. https://doi.org/10.1016/j.acra.2020.07.040

Hanaoka, T., Matoba, H., Nakayama, J., Ono, S., Ikegawa, K., & Okada, M. (2024). A spatio-temporal image analysis for growth of indeterminate pulmonary nodules detected by ct scan. *Radiological Physics and Technology*, *17*(1), 71–82. https://doi.org/10.1007/s12194-023-00750-1

Hawkins, S., Wang, H., Liu, Y., Garcia, A., Stringfield, O., Krewer, H., Li, Q., Cherezov, D., Gatenby, R. A., Balagurunathan, Y., et al. (2016). Predicting malignant nodules from screening CT scans. *Journal of Thoracic Oncology*, *11*(12), 2120–2128. https://doi.org/10.1016/j.jtho.2016.07.002

He, B., Dong, D., She, Y., Zhou, C., Fang, M., Zhu, Y., Zhang, H., Huang, Z., Jiang, T., Tian, J., et al. (2020). Predicting response to immunotherapy in advanced non–small-cell lung cancer using tumor mutational burden radiomic biomarker. *Journal for Immunotherapy of Cancer*, *8*(2). https://doi.org/10.1136/jitc-2020-000550

He, L., Huang, Y., Ma, Z., Liang, C., Liang, C., & Liu, Z. (2016). Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of

radiomics signature in solitary pulmonary nodule. *Scientific Reports*, *6*(1), 34921. https://doi.org/10.1038/srep34921

Horng, H., Singh, A., Yousefi, B., Cohen, E. A., Haghighi, B., Katz, S., Noël, P. B., Kontos, D., & Shinohara, R. T. (2022). Improved generalized combat methods for harmonization of radiomic features. *Scientific Reports*, *12*(1), 19009. https://doi.org/10.1038/s41598-022-23328-0

Horng, H., Singh, A., Yousefi, B., Cohen, E. A., Haghighi, B., Katz, S., Noël, P. B., Shinohara, R. T., & Kontos, D. (2022). Iterative combat methods for harmonization of radiomic features. *Medical Imaging 2022: Computer-Aided Diagnosis*, *12033*, 386–390. https://doi.org/10.1117/12.2610831

Huang, Q., Lu, L., Dercle, L., Lichtenstein, P., Li, Y., Yin, Q., Zong, M., Schwartz, L., & Zhao, B. (2018). Interobserver variability in tumor contouring affects the use of radiomics to predict mutational status. *Journal of Medical Imaging*, *5*(1), 011005–011005. https://doi.org/10.1117/1.JMI.5.1.011005

Huang, Y., Zhu, M., Ji, M., Fan, J., Xie, J., Wei, X., Jiang, X., Xu, J., Chen, L., Yin, R., et al. (2021). Air pollution, genetic factors, and the risk of lung cancer: A prospective study in the uk biobank. *American journal of respiratory and critical care medicine*, *204*(7), 817–825. https://doi.org/10.1164/rccm.202011-4063OC

Huang, Y.-S., Wang, T.-C., Huang, S.-Z., Zhang, J., Chen, H.-M., Chang, Y.-C., & Chang, R.-F. (2023). An improved 3-D attention CNN with hybrid loss and feature fusion for pulmonary nodule classification. *Computer Methods and Programs in Biomedicine*, *229*, 107278. https://doi.org/10.1016/j.cmpb.2022.107278

Ibrahim, A., Lu, L., Yang, H., Akin, O., Schwartz, L. H., & Zhao, B. (2022). The impact of image acquisition parameters and combat harmonization on the predictive performance of radiomics: A renal cell carcinoma model. *Applied Sciences*, *12*(19), 9824. https://doi.org/10.3390/app12199824

Ibrahim, A., Refaee, T., Primakov, S., Barufaldi, B., Acciavatti, R. J., Granzier, R. W., Hustinx, R., Mottaghy, F. M., Woodruff, H. C., Wildberger, J. E., et al. (2021). The effects of in-plane spatial resolution on ct-based radiomic features' stability with and without combat harmonization. *Cancers*, *13*(8), 1848. https://doi.org/10.3390/cancers13081848

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Kale, M. S., Morgan, O., Wisnivesky, J., Schnur, J., & Diefenbach, M. A. (2024). Challenges addressing lung cancer screening for patients with multimorbidity in primary care: A qualitative study. *The Annals of Family Medicine*, *22*(2), 103–112. https://doi.org/10.1370/afm.3080

Kalpathy-Cramer, J., Zhao, B., Goldgof, D., Gu, Y., Wang, X., Yang, H., Tan, Y., Gillies, R., & Napel, S. (2016). A comparison of lung nodule segmentation algorithms: Methods and results from a multi-institutional study. *Journal of Digital Imaging*, *29*, 476–487. https://doi.org/10.1007/s10278-016-9859-z

Kanwal, B., Biswas, S., Seminara, R. S., Jeet, C., & Arora, C. J. (2018). Immunotherapy in advanced non–small cell lung cancer patients: Ushering chemotherapy through the checkpoint inhibitors? *Cureus*, *10*(9). https://doi.org/10.7759/cureus.3254

Kastner, J., Hossain, R., Jeudy, J., Dako, F., Mehta, V., Dalal, S., Dharaiya, E., & White, C. (2021). Lung-RADS version 1.0 versus Lung-RADS version 1.1: Comparison of categories using nodules from the national lung screening trial. *Radiology*, *300*(1), 199–206. https://doi.org/10.1148/radiol.2021203704

Kato, S., Li, B., Adashek, J. J., Cha, S. W., Bianchi-Frias, D., Qian, D., Kim, L., So, T. W., Mitchell, M., Kamei, N., et al. (2022). Serial changes in liquid biopsy-derived variant allele frequency predict immune checkpoint inhibitor responsiveness in the pan-cancer setting. *Oncoimmunology*, *11*(1), 2052410. https://doi.org/10.1080/2162402X.2022.2052410

Khorrami, M., Prasanna, P., Gupta, A., Patil, P., Velu, P. D., Thawani, R., Corredor, G., Alilou, M., Bera, K., Fu, P., et al. (2020). Changes in ct radiomic features associated with lymphocyte distribution predict overall survival and response to immunotherapy in non–small cell lung cancer. *Cancer immunology research*, *8*(1), 108–119. https://doi.org/10.1158/2326-6066.CIR-19-0476

Korpanty, G. J., Graham, D. M., Vincent, M. D., & Leighl, N. B. (2014). Biomarkers that currently affect clinical practice in lung cancer: Egfr, alk, met, ros-1, and kras. *Frontiers in Oncology*, *4*, 204. https://doi.org/10.3389/fonc.2014.00204

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., Forster, K., Aerts, H. J., Dekker, A., Fenstermacher, D., et al. (2012). Radiomics: The process and the challenges. *Magnetic Resonance Imaging*, *30*(9), 1234–1248. https://doi.org/10.1016/j.mri.2012.06.010

Lambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., De Jong, E. E., Van Timmeren, J., Sanduleanu, S., Larue, R. T., Even, A. J., Jochems, A., et al. (2017). Radiomics: The bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, *14*(12), 749–762. https://doi.org/10.1038/nrclinonc.2017.141

Lampeter, W. A. (1985). Computer-aided detection of pulmonary nodules. *Computer Assisted Radiology/Computergestützte Radiologie: Proceedings of the International Symposium/Vorträge des Internationalen Symposiums*, 502–506.

Larue, R. T., van Timmeren, J. E., de Jong, E. E., Feliciani, G., Leijenaar, R. T., Schreurs, W. M., Sosef, M. N., Raat, F. H., van der Zande, F. H., Das, M., et al. (2017). Influence of gray level discretization on radiomic feature stability for different ct scanners, tube currents and slice thicknesses: A comprehensive phantom study. *Acta oncologica*, *56*(11), 1544–1553. https://doi.org/10.1080/0284186X.2017.1351624

Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 255–268. https://doi.org/10.2307/2532051

Lee, G., Lee, H. Y., Park, H., Schiebler, M. L., van Beek, E. J., Ohno, Y., Seo, J. B., & Leung, A. (2017). Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art. *European Journal of Radiology*, *86*, 297–307. https://doi.org/10.1016/j.ejrad.2016.09.005

Li, F., Sone, S., Abe, H., MacMahon, H., & Doi, K. (2004). Malignant versus benign nodules at CT screening for lung cancer: Comparison of thin-section CT findings. *Radiology*, *233*(3), 793–798. https://doi.org/10.1148/radiol.2333031018

Li, Z., Li, H., Ralescu, A. L., Dillman, J. R., Parikh, N. A., & He, L. (2023). A novel collaborative self-supervised learning method for radiomic data. *NeuroImage*, *277*, 120229. https://doi.org/10.1016/j.neuroimage.2023.120229

Ligero, M., Jordi-Ollero, O., Bernatowicz, K., Garcia-Ruiz, A., Delgado-Muñoz, E., Leiva, D., Mast, R., Suarez, C., Sala-Llonch, R., Calvo, N., et al. (2021). Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *European Radiology*, *31*, 1460–1470. https://doi.org/10.1007/s00330-020-07174-0

Liu, Y., Balagurunathan, Y., Atwater, T., Antic, S., Li, Q., Walker, R. C., Smith, G. T., Massion, P. P., Schabath, M. B., & Gillies, R. J. (2017). Radiological image traits predictive of cancer status in pulmonary nodules. *Clinical Cancer Research*, *23*(6), 1442–1449. https://doi.org/10.1158/1078-0432.CCR-15-3102

Liu, Y., Wu, M., Zhang, Y., Luo, Y., He, S., Wang, Y., Chen, F., Liu, Y., Yang, Q., Li, Y., et al. (2021). Imaging biomarkers to predict and evaluate the effectiveness of immunotherapy in advanced non–small-cell lung cancer. *Frontiers in Oncology*, *11*, 657615. https://doi.org/10.3389/fonc.2021.657615

Lu, H. (2021). Computer-aided diagnosis research of a lung tumor based on a deep convolutional neural network and global features. *BioMed Research International*, *2021*. https://doi.org/10.1155/2021/5513746

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems: Proceedings of the 31st International Conference on Neural Information Processing Systems*, *30*, 4768–4777. https://doi.org/10.48550/arXiv.1705.07874

Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1412–1421). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1166

Mali, S. A., Ibrahim, A., Woodruff, H. C., Andrearczyk, V., Müller, H., Primakov, S., Salahuddin, Z., Chatterjee, A., & Lambin, P. (2021). Making radiomics more reproducible across scanner and imaging protocol variations: A review of harmonization methods. *Journal of Personalized Medicine*, *11*(9), 842. https://doi.org/10.3390/jpm11090842

Mantia, C. M., & Buchbinder, E. I. (2019). Immunotherapy toxicity. *Hematology/Oncology Clinics*, *33*(2), 275–290. https://doi.org/10.1016/j.hoc.2018.12.008

Martini, K., Chassagnon, G., Frauenfelder, T., & Revel, M.-P. (2021). Ongoing challenges in implementation of lung cancer screening. *Translational Lung Cancer Research*, *10*(5), 2347. https://doi.org/10.21037/tlcr-2021-1

Massion, P. P., & Walker, R. C. (2014). Indeterminate pulmonary nodules: Risk for having or for developing lung cancer? *Cancer Prevention Research*, *7*(12), 1173–1178. https://doi.org/10.1158/1940-6207.CAPR-14-0364

McKee, B. J., Regis, S. M., McKee, A. B., Flacke, S., & Wald, C. (2016). Performance of ACR Lung-RADS in a clinical CT lung screening program. *Journal of the American College of Radiology*, *13*(2), R25–R29. https://doi.org/10.1016/j.jacr.2015.12.009

Mercieca, S., Belderbos, J. S., & van Herk, M. (2021). Challenges in the target volume definition of lung cancer radiotherapy. *Translational Lung Cancer Research*, *10*(4), 1983. https://doi.org/10.21037/tlcr-20-627

Mikhael, P. G., Wohlwend, J., Yala, A., Karstens, L., Xiang, J., Takigami, A. K., Bourgouin, P. P., Chan, P., Mrah, S., Amayri, W., et al. (2023). Sybil: A validated deep learning model to

predict future lung cancer risk from a single low-dose chest computed tomography. *Journal of Clinical Oncology*, *41*(12), 2191–2200. https://doi.org/10.1200/JCO.22.01345

Mu, W., Jiang, L., Shi, Y., Tunali, I., Gray, J. E., Katsoulakis, E., Tian, J., Gillies, R. J., & Schabath, M. B. (2021). Non-invasive measurement of PD-L1 status and prediction of immunotherapy response using deep learning of pet/ct images. *Journal for immunotherapy of cancer*, *9*(6). https://doi.org/10.1136/jitc-2020-002118

Munir, K., Elahi, H., Ayub, A., Frezza, F., & Rizzi, A. (2019). Cancer diagnosis using deep learning: A bibliographic review. *Cancers (Basel)*, *11*(9), 1235. https://doi.org/10.3390/cancers11091235

Mushti, S. L., Mulkey, F., & Sridhara, R. (2018). Evaluation of overall response rate and progression-free survival as potential surrogate endpoints for overall survival in immunotherapy trials. *Clinical Cancer Research*, *24*(10), 2268–2275. https://doi.org/10.1158/1078-0432.CCR-17-1902

Nanavaty, P., Alvarez, M. S., & Alberts, W. M. (2014). Lung cancer screening: Advantages, controversies, and applications. *Cancer Control*, *21*(1), 9–14. https://doi.org/10.1177/107327481402100102

Nardone, V., Tini, P., Pastina, P., Botta, C., Reginelli, A., Carbone, S. F., Giannicola, R., Calabrese, G., Tebala, C., Guida, C., et al. (2020). Radiomics predicts survival of patients with advanced non–small cell lung cancer undergoing pd-1 blockade using nivolumab. *Oncology Letters*, *19*(2), 1559–1566. https://doi.org/10.3892/ol.2019.11220

Nie, R.-C., Chen, F.-P., Yuan, S.-Q., Luo, Y.-S., Chen, S., Chen, Y.-M., Chen, X.-J., Chen, Y.-B., Li, Y.-F., & Zhou, Z.-W. (2019). Evaluation of objective response, disease control and progression-free survival as surrogate end-points for overall survival in anti–programmed death-1 and anti–programmed death ligand 1 trials. *European Journal of Cancer*, *106*, 1–11. https://doi.org/10.1016/j.ejca.2018.10.011

Orlhac, F., Eertink, J. J., Cottereau, A.-S., Zijlstra, J. M., Thieblemont, C., Meignan, M., Boellaard, R., & Buvat, I. (2022). A guide to combat harmonization of imaging biomarkers in multicenter studies. *Journal of Nuclear Medicine*, *63*(2), 172–179. https://doi.org/10.2967/jnumed.121.262464

Pastorino, U., Sverzellati, N., Sestini, S., Silva, M., Sabia, F., Boeri, M., Cantarutti, A., Sozzi, G., Corrao, G., & Marchianò, A. (2019). Ten-year results of the multicentric italian lung detection trial demonstrate the safety and efficacy of biennial lung cancer screening. *European Journal of Cancer*, *118*, 142–148. https://doi.org/10.1016/j.ejca.2019.06.009

Patel, S. A., & Weiss, J. (2020). Advances in the treatment of non–small cell lung cancer: Immunotherapy. *Clinics in Chest Medicine*, *41*(2), 237–247. https://doi.org/10.1016/j.ccm.2020.02.010

Patz, E. F., Pinsky, P., Gatsonis, C., Sicks, J. D., Kramer, B. S., Tammemägi, M. C., Chiles, C., Black, W. C., Aberle, D. R., Team, N. O. M. W., et al. (2014). Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Internal Medicine*, *174*(2), 269–274. https://doi.org/10.1001/jamainternmed.2013.12738

Paul, R., Hall, L., Goldgof, D., Schabath, M., & Gillies, R. (2018). Predicting nodule malignancy using a cnn ensemble approach. *2018 international joint conference on neural networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2018.8489345

Paz-Ares, L., Ciuleanu, T.-E., Cobo, M., Schenker, M., Zurawski, B., Menezes, J., Richardet, E., Bennouna, J., Felip, E., Juan-Vidal, O., et al. (2021). First-line nivolumab plus ipilimumab

combined with two cycles of chemotherapy in patients with non–small-cell lung cancer (checkmate 9LA): An international, randomised, open-label, phase 3 trial. *The Lancet Oncology*, *22*(2), 198–211. https://doi.org/10.1016/S1470-2045(20)30641-0

Paz-Ares, L., Luft, A., Vicente, D., Tafreshi, A., Gümüş, M., Mazières, J., Hermes, B., Çay Şenler, F., Csőszi, T., Fülöp, A., et al. (2018). Pembrolizumab plus chemotherapy for squamous non–small-cell lung cancer. *New England journal of medicine*, *379*(21), 2040–2051. https://doi.org/10.1056/NEJMoa1810865

Plautz, T. E., Zheng, C., Noid, G., & Li, X. A. (2019). Time stability of delta-radiomics features and the impact on patient analysis in longitudinal ct images. *Medical Physics*, *46*(4), 1663–1676. https://doi.org/10.1002/mp.13395

Punekar, S. R., Shum, E., Grello, C. M., Lau, S. C., & Velcheti, V. (2022). Immunotherapy in non-small cell lung cancer: Past, present, and future directions. *Frontiers in Oncology*, *12*, 877594. https://doi.org/10.3389/fonc.2022.877594

Purandare, N. C., & Rangarajan, V. (2015). Imaging of lung cancer: Implications on staging and management. *Indian Journal of Radiology and Imaging*, *25*(02), 109–120. https://doi.org/10.4103/0971-3026.155831

Refaee, T., Salahuddin, Z., Widaatalla, Y., Primakov, S., Woodruff, H. C., Hustinx, R., Mottaghy, F. M., Ibrahim, A., & Lambin, P. (2022). Ct reconstruction kernels and the effect of pre-and post-processing on the reproducibility of handcrafted radiomic features. *Journal of Personalized Medicine*, *12*(4), 553. https://doi.org/10.3390/jpm12040553

Ruano-Ravina, A., Heleno, B., & Fernández-Villar, A. (2015). Lung cancer screening with low-dose CT (LDCT), or when a public health intervention is beyond the patient's benefit. *Journal of Epidemiology and Community Health*, *69*(2), 99–100. https://doi.org/10.1136/jech-2014-204293

Saad, M. B., Hong, L., Aminu, M., Vokes, N. I., Chen, P., Salehjahromi, M., Qin, K., Sujit, S. J., Lu, X., Young, E., et al. (2023). Predicting benefit from immune checkpoint inhibitors in patients with non–small-cell lung cancer by ct-based ensemble deep learning: A retrospective study. *The Lancet Digital Health*. https://doi.org/10.1016/S2589-7500(23)00082-1

Schiller, J. H., Harrington, D., Belani, C. P., Langer, C., Sandler, A., Krook, J., Zhu, J., & Johnson, D. H. (2002). Comparison of four chemotherapy regimens for advanced non–small-cell lung cancer. *New England Journal of Medicine*, *346*(2), 92–98. https://doi.org/10.1056/NEJMoa011954

Schöckel, L., Jost, G., Seidensticker, P., Lengsfeld, P., Palkowitsch, P., & Pietsch, H. (2020). Developments in x-ray contrast media and the potential impact on computed tomography. *Investigative Radiology*, *55*(9), 592–597. https://doi.org/10.1097/RLI.0000000000000696

Sequist, L. V., Waltman, B. A., Dias-Santagata, D., Digumarthy, S., Turke, A. B., Fidias, P., Bergethon, K., Shaw, A. T., Gettinger, S., Cosper, A. K., et al. (2011). Genotypic and histological evolution of lung cancers acquiring resistance to egfr inhibitors. *Science Translational Medicine*, *3*(75), 75ra26–75ra26. https://doi.org/10.1126/scitranslmed.3002003

Seymour, L., Bogaerts, J., Perrone, A., Ford, R., Schwartz, L. H., Mandrekar, S., Lin, N. U., Litière, S., Dancey, J., Chen, A., et al. (2017). iRECIST: Guidelines for response criteria for use in trials testing immunotherapeutics. *The Lancet Oncology*, *18*(3), e143–e152. https://doi.org/10.1016/S1470-2045(17)30074-8

Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(1), 7–30. https://doi.org/10.3322/caac.21590

Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, *73*(1), 17–48. https://doi.org/10.3322/caac.21763

Singh, A., Horng, H., Chitalia, R., Roshkovan, L., Katz, S. I., Noël, P., Shinohara, R. T., & Kontos, D. (2022). Resampling and harmonization for mitigation of heterogeneity in image parameters of baseline scans. *Scientific Reports*, *12*(1), 21505. https://doi.org/10.1038/s41598-022-26083-4

Singh, A., Horng, H., Roshkovan, L., Weeks, J. K., Hershman, M., Noël, P., Luna, J. M., Cohen, E. A., Pantalone, L., Shinohara, R. T., et al. (2022). Development of a robust radiomic biomarker of progression-free survival in advanced non-small cell lung cancer patients treated with first-line immunotherapy. *Scientific Reports*, *12*(1), 9993. https://doi.org/10.1038/s41598-022-14160-7

Sinoquet, L., Jacot, W., Quantin, X., & Alix-Panabières, C. (2023). Liquid biopsy and immuno-oncology for advanced non-small cell lung cancer. *Clinical Chemistry*, *69*(1), 23–40. https://doi.org/10.1093/clinchem/hvac166

Spyratos, D., Zarogoulidis, P., Porpodis, K., Tsakiridis, K., Machairiotis, N., Katsikogiannis, N., Kougioumtzi, I., Dryllis, G., Kallianos, A., Rapti, A., et al. (2013). Occupational exposure and lung cancer. *Journal of thoracic disease*, *5*(Suppl 4), S440. https://doi.org/10.3978/j.issn.2072-1439.2013.07.09

Steliga, M. A., & Dresler, C. M. (2011). Epidemiology of lung cancer: Smoking, secondhand smoke, and genetics. *Surgical Oncology Clinics*, *20*(4), 605–618. https://doi.org/10.1016/j.soc.2011.07.003

Stojšić, J. (2018). Precise diagnosis of histological type of lung carcinoma: The first step in personalized therapy. In *Lung cancer-strategies for diagnosis and treatment*. IntechOpen. https://doi.org/10.5772/intechopen.75316

Su, R., van der Sluijs, M., Cornelissen, S. A., Lycklama, G., Hofmeijer, J., Majoie, C. B., van Doormaal, P. J., Van Es, A. C., Ruijters, D., Niessen, W. J., et al. (2022). Spatio-temporal deep learning for automatic detection of intracranial vessel perforation in digital subtraction angiography during endovascular thrombectomy. *Medical Image Analysis*, *77*, 102377. https://doi.org/10.1016/j.media.2022.102377

Sun, R., Limkin, E. J., Vakalopoulou, M., Dercle, L., Champiat, S., Han, S. R., Verlingue, L., Brandao, D., Lancia, A., Ammari, S., et al. (2018). A radiomics approach to assess tumour-infiltrating cd8 cells and response to anti-pd-1 or anti-pd-l1 immunotherapy: An imaging biomarker, retrospective multicohort study. *The Lancet Oncology*, *19*(9), 1180–1191. https://doi.org/10.1016/S1470-2045(18)30413-3

Suresh, K., Naidoo, J., Lin, C. T., & Danoff, S. (2018). Immune checkpoint immunotherapy for non–small cell lung cancer: Benefits and pulmonary toxicities. *Chest*, *154*(6), 1416–1423. https://doi.org/10.1016/j.chest.2018.08.1048

Tang, C., Hobbs, B., Amer, A., Li, X., Behrens, C., Canales, J. R., Cuentas, E. P., Villalobos, P., Fried, D., Chang, J. Y., et al. (2018). Development of an immune-pathology informed radiomics model for non–small cell lung cancer. *Scientific Reports*, *8*(1), 1922. https://doi.org/10.1038/s41598-018-20471-5

Team, N. L. S. T. R. (2011a). The national lung screening trial: Overview and study design. *Radiology*, *258*(1), 243–253. https://doi.org/10.1148/radiol.10091808

Team, N. L. S. T. R. (2011b). Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England Journal of Medicine*, *365*(5), 395–409. https://doi.org/10.1056/NEJMoa1102873

Thawani, R., McLane, M., Beig, N., Ghose, S., Prasanna, P., Velcheti, V., & Madabhushi, A. (2018). Radiomics and radiogenomics in lung cancer: A review for the clinician. *Lung Cancer*, *115*, 34–41. https://doi.org/10.1016/j.lungcan.2017.10.015

Tian, P., He, B., Mu, W., Liu, K., Liu, L., Zeng, H., Liu, Y., Jiang, L., Zhou, P., Huang, Z., et al. (2021). Assessing PD-L1 expression in non–small cell lung cancer and predicting responses to immune checkpoint inhibitors using deep learning on computed tomography images. *Theranostics*, *11*(5), 2098. https://doi.org/10.7150/thno.48027

Trebeschi, S., Bodalal, Z., Boellaard, T. N., Tareco Bucho, T. M., Drago, S. G., Kurilova, I., Calin-Vainak, A. M., Delli Pizzi, A., Muller, M., Hummelink, K., et al. (2021). Prognostic value of deep learning-mediated treatment monitoring in lung cancer patients receiving immunotherapy. *Frontiers in Oncology*, *11*, 609054. https://doi.org/10.3389/fonc.2021.609054

Trebeschi, S., Drago, S., Birkbak, N., Kurilova, I., Călin, A., Pizzi, A. D., Lalezari, F., Lambregts, D., Rohaan, M., Parmar, C., et al. (2019). Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Annals of Oncology*, *30*(6), 998–1004. https://doi.org/10.1093/annonc/mdz108

Tunali, I., Gray, J. E., Qi, J., Abdalah, M., Jeong, D. K., Guvenis, A., Gillies, R. J., & Schabath, M. B. (2019). Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: An early report. *Lung Cancer*, *129*, 75–79. https://doi.org/10.1016/j.lungcan.2019.01.010

Usuda, K., Saito, Y., Sagawa, M., Sato, M., Kanma, K., Takahashi, S., Endo, C., Chen, Y., Sakurada, A., & Fujimura, S. (1994). Tumor doubling time and prognostic assessment of patients with primary lung cancer. *Cancer*, *74*(8), 2239–2244. https://doi.org/10.1002/1097-0142(19941015)74:8<2239::AID-CNCR2820740806>3.0.CO;2-P

Valero, C., Lee, M., Hoen, D., Weiss, K., Kelly, D. W., Adusumilli, P. S., Paik, P. K., Plitas, G., Ladanyi, M., Postow, M. A., et al. (2021). Pretreatment neutrophil-to-lymphocyte ratio and mutational burden as biomarkers of tumor response to immune checkpoint inhibitors. *Nature Communications*, *12*(1), 729. https://doi.org/10.1038/s41467-021-20935-9

Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G., Fillion-Robin, J.-C., Pieper, S., & Aerts, H. J. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, *77*(21), e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

Van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H., & Baessler, B. (2020). Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging*, *11*(1), 91. https://doi.org/10.1186/s13244-020-00887-2

Vanguri, R. S., Luo, J., Aukerman, A. T., Egger, J. V., Fong, C. J., Horvat, N., Pagano, A., Araujo-Filho, J. d. A. B., Geneslaw, L., Rizvi, H., et al. (2022). Multimodal integration of radiology, pathology and genomics for prediction of response to PD-L1 blockade in patients with non–small cell lung cancer. *Nature Cancer*, *3*(10), 1151–1164. https://doi.org/10.1038/s43018-022-00416-8

Veasey, B. P., Broadhead, J., Dahle, M., Seow, A., & Amini, A. A. (2020). Lung nodule malignancy prediction from longitudinal CT scans with siamese convolutional attention networks. *IEEE*

Open Journal of Engineering in Medicine and Biology, *1*, 257–264. https://doi.org/10.1109/OJEMB.2020.3023614

Vegas-Sánchez-Ferrero, G., Díaz, A. A., Ash, S. Y., Baraghoshi, D., Strand, M., Crapo, J. D., Silverman, E. K., Humphries, S. M., Washko, G. R., Lynch, D. A., et al. (2024). Quantification of emphysema progression at ct using simultaneous volume, noise, and bias lung density correction. *Radiology*, *310*(1), e231632. https://doi.org/10.1148/radiol.231632

Vegas-Sánchez-Ferrero, G., Ledesma-Carbayo, M. J., Washko, G. R., & Estépar, R. S. J. (2017). Statistical characterization of noise for spatial standardization of ct scans: Enabling comparison with multiple kernels and doses. *Medical Image Analysis*, *40*, 44–59. https://doi.org/10.1016/j.media.2017.06.001

Vegas-Sánchez-Ferrero, G., Ledesma-Carbayo, M. J., Washko, G. R., & Estépar, R. S. J. (2018). Autocalibration method for non-stationary ct bias correction. *Medical Image Analysis*, *44*, 115–125. https://doi.org/10.1016/j.media.2017.12.004

Vegas-Sánchez-Ferrero, G., Ledesma-Carbayo, M. J., Washko, G. R., & San José Estépar, R. (2019). Harmonization of chest ct scans for different doses and reconstruction methods. *Medical Physics*, *46*(7), 3117–3132. https://doi.org/10.1002/mp.13578

Vlahos, I., Stefanidis, K., Sheard, S., Nair, A., Sayer, C., & Moser, J. (2018). Lung cancer screening: Nodule identification and characterization. *Translational Lung Cancer Research*, *7*(3), 288. https://doi.org/10.21037/tlcr.2018.05.02

Vonder, M., Dorrius, M. D., & Vliegenthart, R. (2021). Latest CT technologies in lung cancer screening: Protocols and radiation dose reduction. *Translational Lung Cancer Research*, *10*(2), 1154. https://doi.org/10.21037/tlcr-20-808

Walter, J. E., Heuvelmans, M. A., de Jong, P. A., Vliegenthart, R., van Ooijen, P. M., Peters, R. B., Ten Haaf, K., Yousaf-Khan, U., van der Aalst, C. M., de Bock, G. H., et al. (2016). Occurrence and lung cancer probability of new solid nodules at incidence screening with low-dose CT: Analysis of data from the randomised, controlled nelson trial. *The Lancet Oncology*, *17*(7), 907–916. https://doi.org/10.1016/S1470-2045(16)30069-9

Wang, X., Mao, K., Wang, L., Yang, P., Lu, D., & He, P. (2019). An appraisal of lung nodules automatic classification algorithms for CT images. *Sensors*, *19*(1), 194. https://doi.org/10.3390/s19010194

Wang, X., Jiang, Y., Chen, H., Zhang, T., Han, Z., Chen, C., Yuan, Q., Xiong, W., Wang, W., Li, G., et al. (2023). Cancer immunotherapy response prediction from multi-modal clinical and image data using semi-supervised deep learning. *Radiotherapy and Oncology*, *186*, 109793. https://doi.org/10.1016/j.radonc.2023.109793

Wang, X., Niu, X., An, N., Sun, Y., & Chen, Z. (2021). Comparative efficacy and safety of immunotherapy alone and in combination with chemotherapy for advanced non-small cell lung cancer. *Frontiers in Oncology*, *11*, 611012. https://doi.org/10.3389/fonc.2021.611012

Wasserthal, J., Breit, H.-C., Meyer, M. T., Pradella, M., Hinck, D., Sauter, A. W., Heye, T., Boll, D. T., Cyriac, J., Yang, S., et al. (2023). Totalsegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, *5*(5). https://doi.org/10.1148/ryai.230024

Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., Mak, R. H., & Aerts, H. J. (2019a). Deep learning predicts lung cancer treatment response from serial medical imaging. *Clinical Cancer Research*, *25*(11), 3266–3275. https://doi.org/10.1158/1078-0432.CCR-18-2495

Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., Mak, R. H., & Aerts, H. J. (2019b). Deep learning predicts lung cancer treatment response from serial medical imaging. *Clinical Cancer Research*, *25*(11), 3266–3275. https://doi.org/10.1158/1078-0432.CCR-18-2495

Yanagawa, M., Johkoh, T., Noguchi, M., Morii, E., Shintani, Y., Okumura, M., Hata, A., Fujiwara, M., Honda, O., & Tomiyama, N. (2017). Radiological prediction of tumor invasiveness of lung adenocarcinoma on thin-section ct. *Medicine*, *96*(11), e6331. https://doi.org/10.1097/MD.0000000000006331

Yang, Y., Yang, J., Shen, L., Chen, J., Xia, L., Ni, B., Ge, L., Wang, Y., & Lu, S. (2021). A multi-omics-based serial deep learning approach to predict clinical outcomes of single-agent anti-pd-1/pd-l1 immunotherapy in advanced stage non–small-cell lung cancer. *American journal of translational research*, *13*(2), 743.

Zhang, L., Lu, L., Wang, X., Zhu, R. M., Bagheri, M., Summers, R. M., & Yao, J. (2019). Spatio-temporal convolutional LSTMs for tumor growth prediction by learning 4D longitudinal patient data. *IEEE Transactions on Medical Imaging*, *39*(4), 1114–1126. https://doi.org/10.1109/TMI.2019.2943841

Zhang, R., Tian, P., Qiu, Z., Liang, Y., & Li, W. (2020). The growth feature and its diagnostic value for benign and malignant pulmonary nodules met in routine clinical practice. *Journal of Thoracic Disease*, *12*(5), 2019. https://doi.org/10.21037/jtd-19-3591

Zhao, B., Tan, Y., Tsai, W.-Y., Qi, J., Xie, C., Lu, L., & Schwartz, L. H. (2016). Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Scientific Reports*, *6*(1), 23428. https://doi.org/10.1038/srep23428

Zhao, Y., de Bock, G. H., Vliegenthart, R., van Klaveren, R. J., Wang, Y., Bogoni, L., de Jong, P. A., Mali, W. P., van Ooijen, P. M., & Oudkerk, M. (2012). Performance of computer-aided detection of pulmonary nodules in low-dose CT: Comparison with double reading by nodule volume. *European Radiology*, *22*, 2076–2084. https://doi.org/10.1007/s00330-012-2437-y

Zhu, W., Liu, C., Fan, W., & Xie, X. (2018). Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. *2018 IEEE winter conference on applications of computer vision (WACV)*, 673–681. https://doi.org/10.1109/WACV.2018.00079

Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., et al. (2020). The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, *295*(2), 328–338. https://doi.org/10.1148/radiol.2020191145

Appendix A

Manufacturers	Models
Siemens	SOMOTOM (Force, Volume Zoom, Drive, Definition, Definition Falsh, X.cite)
	Sensation 64
	Biograph64
	Emotion 16
Philips	Brilliance 64
Toshiba	Aquilion
	Aquilion Lightning
GE Medical Systems	Revolution EVO
	LightSpeed VCT
	Optima CT660
	Discovery MI
	Revolution HD

Table A1: CT scanner manufacturers and models. List of the manufacturers and specific models of CT scanners used for image acquisition in the study.

Model	Features	N test	AUC [95% CI]	ACC [95% CI]	SENS [95% CI]	SPEC [95% CI]	PREC [95% CI]	bACC [95% CI]
RF-baseline	Clinical data	43	0.667 [0.485,0.833]	0.651 [0.512,0.791]	0.833 [0.667,0.962]	0.421 [0.2,0.65]	0.645 [0.48,0.812]	0.627 [0.488,0.774]
RF-baseline	DF-imm	43	0.588 [0.409,0.767]	0.558 [0.419,0.698]	0.833 [0.679,0.96]	0.211 [0.05,0.417]	0.571 [0.406,0.735]	0.522 [0.403,0.638]
RF-longitudinal	Clinical data	32	0.586 [0.413,0.753]	0.594 [0.406,0.750]	0.467 [0.200,0.733]	0.706 [0.467,0.909]	0.583 [0.300,0.867]	0.586 [0.417,0.75]
RF-longitudinal	DF-imm	32	0.727 [0.576,0.875]	0.719 [0.562,0.875]	0.867 [0.667,1.0]	0.588 [0.333,0.833]	0.727 [0.575,0.875]	0.650 [0.429,0.857]
Ensemble RF-baseline	DF-imm Clinical data	43	0.678 [0.513,0.836]	0.605 [0.442,0.744]	**0.875** [0.731,1.0]	0.263 [0.071,0.467]	0.600 [0.436,0.758]	0.569 [0.448,0.684]
Ensemble RF-longitudinal	DF-imm Clinical data	32	**0.824** [0.658,0.953]	**0.750** [0.594,0.906]	0.733 [0.500,0.938]	**0.765** [0.533,0.947]	**0.733** [0.471,0.933]	**0.749** [0.594,0.897]

Table A2: Response prediction performance comparison between longitudinal and ensemble models in the independent test set for endpoint PFS6 by evaluating AUC, ACC, SENS, SPEC, PREC and bACC, respectively. For each metric, the 95% confidence interval has been shown and the highest value has been highlighted in bold.

Model	Features	N test	AUC [95% CI]	ACC [95% CI]	SENS [95% CI]	SPEC [95% CI]	PREC [95% CI]	bACC [95% CI]
RF-baseline	Clinical data	43	0.563 [0.392,0.735]	0.581 [0.442,0.721]	0.793 [0.636,0.929]	0.143 [0.0,0.357]	0.657 [0.5,0.811]	0.468 [0.352,0.591]
RF-baseline	DF-imm	43	0.541 [0.359,0.724]	0.628 [0.488,0.767]	0.759 [0.6,0.903]	0.357 [0.118,0.6]	0.710 [0.533,0.867]	0.558 [0.405,0.711]
RF-longitudinal	Clinical data	32	0.573 [0.4,0.742]	0.531 [0.375,0.688]	0.579 [0.353,0.789]	0.462 [0.188,0.727]	0.611 [0.4,0.842]	0.52 [0.333,0.697]
RF-longitudinal	DF-imm	32	0.717 [0.558,0.865]	0.750 [0.594,0.875]	0.895 [0.737,1.0]	0.538 [0.273,0.833]	0.739 [0.55,0.913]	0.717 [0.562,0.869]
Ensemble RF-baseline	DF-imm Clinical data	43	0.560 [0.377,0.731]	0.581 [0.442,0.721]	0.793 [0.643,0.933]	0.143 [0.0,0.364]	0.657 [0.487,0.811]	0.468 [0.36,0.59]
Ensemble RF-longitudinal	DF-imm Clinical data	32	**0.753** [0.549,0.931]	**0.813** [0.656,0.938]	**0.947** [0.826,1.0]	**0.615** [0.357,0.889]	**0.783** [0.609,0.95]	**0.781** [0.631,0.923]

Table A3: Response prediction performance comparison between longitudinal and ensemble models in the independent test set for endpoint PFS9 by evaluating AUC, ACC, SENS, SPEC, PREC and bACC, respectively. For each metric, the 95% confidence interval has been shown and the highest value has been highlighted in bold.

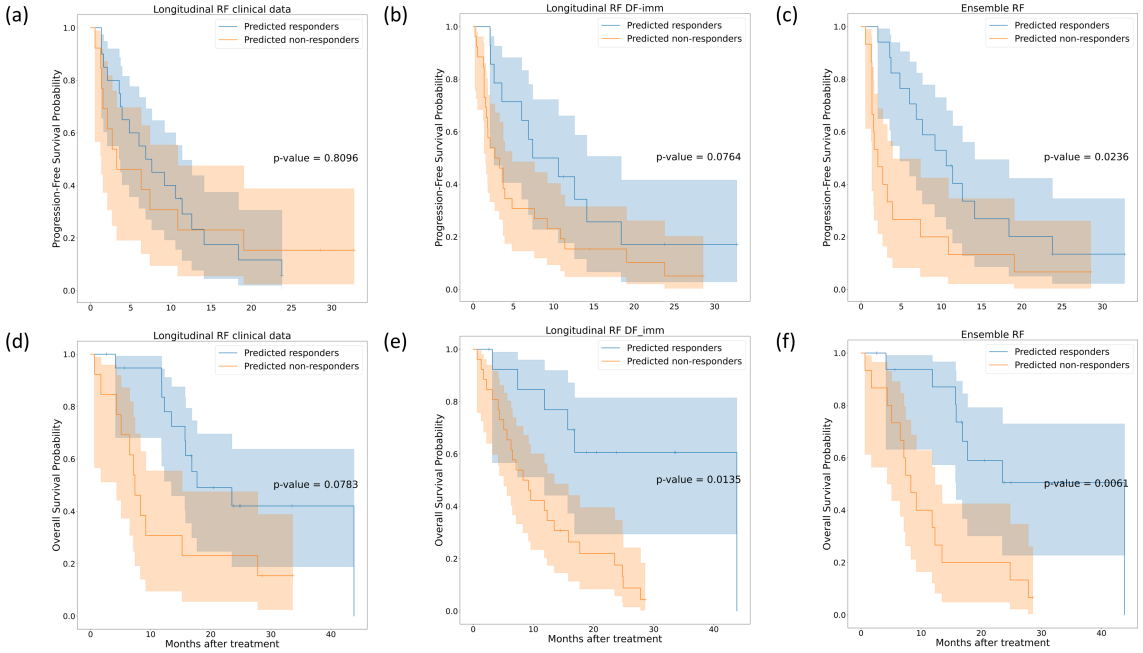


Figure A1: Kaplan-Meier survival curves on the independent test cohort for longitudinal RF based on clinical data ((a) and (d)), longitudinal RF with deep features ((b) and (e)) and ensemble RF ((c) and (f)) trained for endpoint PFS6, according to risk groups based on each models' predictions. The first row represents the progression-free survival Kaplan-Meier curves, while the second row represents the overall survival Kaplan-Meier curves.

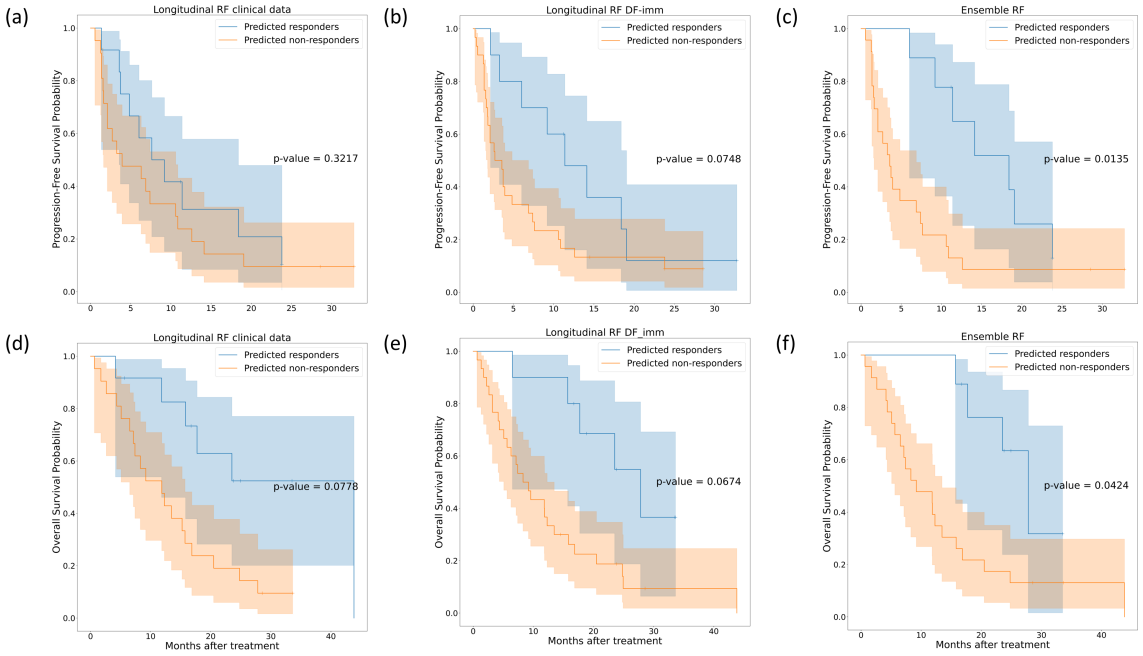


Figure A2: Kaplan-Meier survival curves on the independent test cohort for longitudinal RF based on clinical data ((a) and (d)), longitudinal RF with deep features ((b) and (e)) and ensemble RF ((c) and (f)) trained for endpoint PFS9, according to risk groups based on each models' predictions. The first row represents the progression-free survival Kaplan-Meier curves, while the second row represents the overall survival Kaplan-Meier curves.

Appendix B

		Missing	Overall	Non-responders	Responders	P-Value (adjusted)
n patients	{}		155	74	81	
Treatment Response, n (%)	Non-responders	0	74 (47.7)	74 (100.0)		<0.001
	Responders		81 (52.3)		81 (100.0)	
Treatment, n (%)	Combined Immunological Agents	0	20 (12.9)	13 (17.6)	7 (8.6)	1.000
	Immuno+Chemotherapy		38 (24.5)	13 (17.6)	25 (30.9)	
	Immunotherapy+Radiotherapy		16 (10.3)	10 (13.5)	6 (7.4)	
	Monotherapy		80 (51.6)	37 (50.0)	43 (53.1)	
	Other		1 (0.6)	1 (1.4)		
Stage, n (%)	III	6	11 (7.4)	3 (4.2)	8 (10.4)	1.000
	IV		138 (92.6)	69 (95.8)	69 (89.6)	
Gender, n (%)	Female	0	50 (32.3)	25 (33.8)	25 (30.9)	1.000
	Male		105 (67.7)	49 (66.2)	56 (69.1)	
Age, mean (SD)		0	65.2 (9.5)	64.2 (8.9)	66.1 (10.0)	1.000
PFS (months), mean (SD)		0	11.7 (15.5)	2.8 (1.4)	19.9 (17.9)	<0.001
OS (months), mean (SD)		0	17.8 (17.2)	8.1 (8.3)	26.7 (18.5)	<0.001
Status, n (%)	Alive	0	68 (43.9)	21 (28.4)	47 (58.0)	0.008
	Dead		87 (56.1)	53 (71.6)	34 (42.0)	
Progression, n (%)	No	0	30 (19.4)	3 (4.1)	27 (33.3)	<0.001
	Yes		125 (80.6)	71 (95.9)	54 (66.7)	
Progression type, n (%)	Clinical	0	11 (7.1)	7 (9.5)	4 (4.9)	0.043
	Death		9 (5.8)	3 (4.1)	6 (7.4)	
	Other		33 (21.3)	6 (8.1)	27 (33.3)	
	Radiological		94 (60.6)	54 (73.0)	40 (49.4)	
	Toxicity		8 (5.2)	4 (5.4)	4 (4.9)	
IPA, mean (SD)		12	45.3 (31.5)	47.2 (34.0)	43.6 (29.2)	1.000
Smoking habit, n (%)	Current smoker	1	35 (22.7)	17 (23.3)	18 (22.2)	1.000
	Former smoker		108 (70.1)	51 (69.9)	57 (70.4)	
	Non-smoker		11 (7.1)	5 (6.8)	6 (7.4)	
Tumor histology, n (%)	Adenocarcinoma	2	111 (72.5)	54 (74.0)	57 (71.2)	1.000
	Squamous cell carcinoma		36 (23.5)	15 (20.5)	21 (26.2)	
	SCLC		6 (3.9)	4 (5.5)	2 (2.5)	
PDL1, mean (SD)		73	0.4 (0.4)	0.4 (0.4)	0.4 (0.4)	1.000
Surgery, n (%)	No	29	101 (80.2)	49 (84.5)	52 (76.5)	1.000
	Yes		25 (19.8)	9 (15.5)	16 (23.5)	
Pre-treatment ECOG, mean (SD)		30	0.6 (0.5)	0.7 (0.5)	0.5 (0.5)	1.000
Weight, mean (SD)		8	70.2 (13.2)	67.5 (11.9)	72.5 (14.0)	0.432
Height, mean (SD)		12	168.0 (7.9)	168.3 (8.1)	167.7 (7.7)	1.000
n metastatic sites, n (%)	0.0	94	16 (26.2)	7 (23.3)	9 (29.0)	1.000
	1.0		19 (31.1)	6 (20.0)	13 (41.9)	
	2.0		13 (21.3)	8 (26.7)	5 (16.1)	
	3.0		7 (11.5)	3 (10.0)	4 (12.9)	
	4.0		5 (8.2)	5 (16.7)		
	5.0		1 (1.6)	1 (3.3)		
Steroids, n (%)	No	17	77 (55.8)	35 (55.6)	42 (56.0)	1.000
	Yes		61 (44.2)	28 (44.4)	33 (44.0)	
Antibiotics, n (%)	No	26	102 (79.1)	50 (80.6)	52 (77.6)	1.000
	Yes		27 (20.9)	12 (19.4)	15 (22.4)	
COPD, n (%)	No	69	40 (46.5)	17 (47.2)	23 (46.0)	1.000
	Yes		46 (53.5)	19 (52.8)	27 (54.0)	

Table B1: Demographic and clinical characteristics of the patients in the internal cohort. Adjusted p-values (using Bonferroni correction) were calculated for comparisons between responders and non-responders using two-sample T-tests for continuous variables and Chi-square tests for categorical variables. SD represents the standard deviation, and Q1 and Q3 represent the first and third quarterlies, respectively.

		Missing	Overall	Non-responders	Responders	P-Value (adjusted)
n patients			39	26	13	
Treatment Response, n (%)	Non-responders	0	26 (66.7)	26 (100.0)		<0.001
	Responders		13 (33.3)		13 (100.0)	
Treatment, n (%)	Monotherapy	0	39 (100.0)	26 (100.0)	13 (100.0)	1.000
Stage, n (%)	IV	0	37 (94.9)	26 (100.0)	11 (84.6)	1.000
	III		2 (5.1)		2 (15.4)	
Gender, n (%)	Female	0	6 (15.4)	5 (19.2)	1 (7.7)	1.000
	Male		33 (84.6)	21 (80.8)	12 (92.3)	
Age, mean (SD)		0	65.8 (10.4)	64.4 (10.4)	68.8 (10.4)	1.000
PFS (months), mean (SD)		0	5.8 (6.1)	2.4 (1.3)	12.8 (6.0)	0.001
OS (months), mean (SD)		0	12.4 (9.1)	8.8 (6.1)	19.7 (10.1)	0.052
Status, n (%)	Alive	0	5 (12.8)	1 (3.8)	4 (30.8)	0.760
	Dead		34 (87.2)	25 (96.2)	9 (69.2)	
Progression, n (%)	Yes	0	38 (97.4)	26 (100.0)	12 (92.3)	1.000
	No		1 (2.6)		1 (7.7)	
Progression type, n (%)	Other	0	39 (100.0)	26 (100.0)	13 (100.0)	1.000
IPA, mean (SD)		0	46.4 (0.0)	46.4 (0.0)	46.4 (0.0)	nan
Smoking habit, n (%)	Current smoker	0	9 (23.1)	7 (26.9)	2 (15.4)	1.000
	Former smoker		29 (74.4)	18 (69.2)	11 (84.6)	
	Non-smoker		1 (2.6)	1 (3.8)		
Tumor histology, n (%)	Adenocarcinoma	0	16 (41.0)	11 (42.3)	5 (38.5)	1.000
	Squamous cell carcinoma		18 (46.2)	10 (38.5)	8 (61.5)	
	SCLC		5 (12.8)	5 (19.2)		
PDL1, mean (SD)		0	0.1 (0.2)	0.0 (0.0)	0.1 (0.4)	1.000
Surgery, n (%)	No	0	39 (100.0)	26 (100.0)	13 (100.0)	1.000
Pre-treatment ECOG, mean (SD)		0	1.0 (0.4)	1.1 (0.3)	0.7 (0.5)	0.230
Weight, mean (SD)		0	73.6 (12.5)	72.1 (11.7)	76.6 (13.9)	1.000
Height, mean (SD)		0	166.8 (6.7)	167.2 (7.3)	166.0 (5.7)	1.000
n metastatic sites, n (%)	0	0	27 (69.2)	15 (57.7)	12 (92.3)	1.000
	1		8 (20.5)	7 (26.9)	1 (7.7)	
	2		1 (2.6)	1 (3.8)		
			3 (7.7)	3 (11.5)		
Steroids, n (%)	No	0	36 (92.3)	25 (96.2)	11 (84.6)	1.000
	Yes		3 (7.7)	1 (3.8)	2 (15.4)	
Antibiotics, n (%)	No	0	32 (82.1)	21 (80.8)	11 (84.6)	1.000
	Yes		7 (17.9)	5 (19.2)	2 (15.4)	
COPD, n (%)	No	0	29 (74.4)	21 (80.8)	8 (61.5)	1.000
	Yes		10 (25.6)	5 (19.2)	5 (38.5)	

Table B2: Demographic and clinical characteristics of the patients in the external independent cohort. Adjusted p-values (using Bonferroni correction) were calculated for comparisons between responders and non-responders using two-sample T-tests for continuous variables and Chi-square tests for categorical variables. SD represents the standard deviation, and Q1 and Q3 represent the first and third quarterlies, respectively.

		Missing	Overall	Internal Train	Internal Test	External Test	P-Value (adjusted)
n patients			197	111	44	42	
Treatment Response, n (%)	Non-responders	0	102 (51.8)	52 (46.8)	22 (50.0)	28 (66.7)	1.000
	Responders		95 (48.2)	59 (53.2)	22 (50.0)	14 (33.3)	
Treatment, n (%)	Monotherapy	0	122 (61.9)	56 (50.5)	24 (54.5)	42 (100.0)	<0.001
	Combined Immunological Agents		20 (10.2)	17 (15.3)	3 (6.8)		
	Immuno+Chemotherapy		38 (19.3)	27 (24.3)	11 (25.0)		
	Immunotherapy+Radiotherapy		16 (8.1)	10 (9.0)	6 (13.6)		
	Other		1 (0.5)	1 (0.9)			
Stage, n (%)	III	6	13 (6.8)	8 (7.4)	3 (7.3)	2 (4.8)	1.000
	IV		178 (93.2)	100 (92.6)	38 (92.7)	40 (95.2)	
Gender, n (%)	Female	0	57 (28.9)	38 (34.2)	12 (27.3)	7 (16.7)	1.000
	Male		140 (71.1)	73 (65.8)	32 (72.7)	35 (83.3)	
Age, mean (SD)		0	65.4 (9.7)	65.8 (8.8)	63.6 (11.1)	66.0 (10.4)	1.000
PFS (months), mean (SD)		0	10.4 (14.3)	11.6 (13.9)	12.1 (19.3)	5.7 (5.9)	0.954
OS (months), mean (SD)		0	16.6 (16.0)	17.6 (16.5)	18.4 (19.2)	11.8 (9.1)	1.000
Status, n (%)	Alive	0	73 (37.1)	47 (42.3)	21 (47.7)	5 (11.9)	0.012
	Dead		124 (62.9)	64 (57.7)	23 (52.3)	37 (88.1)	
Progression, n (%)	No	0	31 (15.7)	23 (20.7)	7 (15.9)	1 (2.4)	0.419
	Yes		166 (84.3)	88 (79.3)	37 (84.1)	41 (97.6)	
IPA, mean (SD)		12	45.5 (27.7)	47.9 (32.5)	39.2 (28.5)	46.4 (0.0)	1.000
Smoking habit, n (%)	Current smoker	1	45 (23.0)	23 (20.9)	12 (27.3)	10 (23.8)	1.000
	Former smoker		139 (70.9)	81 (73.6)	27 (61.4)	31 (73.8)	
	Non-smoker		12 (6.1)	6 (5.5)	5 (11.4)	1 (2.4)	
Tumor histology, n (%)	Adenocarcinoma	2	128 (65.6)	77 (70.6)	34 (77.3)	17 (40.5)	0.054
	Squamous cell carcinoma		56 (28.7)	27 (24.8)	9 (20.5)	20 (47.6)	
	SCLC		11 (5.6)	5 (4.6)	1 (2.3)	5 (11.9)	
PDL1, mean (SD)		73	0.3 (0.4)	0.4 (0.4)	0.4 (0.4)	0.1 (0.2)	<0.001
Surgery, n (%)	No	29	143 (85.1)	72 (78.3)	29 (85.3)	42 (100.0)	0.092
	Yes		25 (14.9)	20 (21.7)	5 (14.7)		
Pre-treatment ECOG, mean (SD)		30	0.7 (0.5)	0.6 (0.5)	0.6 (0.5)	1.0 (0.4)	0.002
Weight, mean (SD)		8	70.9 (13.0)	71.3 (13.4)	67.1 (12.2)	73.5 (12.1)	1.000
Height, mean (SD)		12	167.7 (7.6)	167.9 (8.0)	168.2 (7.6)	167.0 (6.6)	1.000
Steroids, n (%)	No	17	116 (64.4)	59 (57.8)	18 (50.0)	39 (92.9)	0.001
	Yes		64 (35.6)	43 (42.2)	18 (50.0)	3 (7.1)	
Antibiotics, n (%)	No	26	135 (78.9)	75 (81.5)	27 (73.0)	33 (78.6)	1.000
	Yes		36 (21.1)	17 (18.5)	10 (27.0)	9 (21.4)	
COPD, n (%)	No	69	71 (55.5)	32 (47.8)	8 (42.1)	31 (73.8)	0.258
	Yes		57 (44.5)	35 (52.2)	11 (57.9)	11 (26.2)	

Table B3: Demographic and clinical characteristics of the patients in the training and internal and external tests. Adjusted p-values (using Bonferroni correction) were calculated for comparisons between the internal train, test sets and external test set using two-sample T-tests for continuous variables and Chi-square tests for categorical variables. SD represents the standard deviation, and Q1 and Q3 represent the first and third quarterlies, respectively.

Manufacturers	Models
Siemens	SOMOTOM (Force, Volume Zoom, Drive, Definition, Definition Falsh, X.cite)
	Sensation (16, 64)
	Biograph64
	Emotion 16
Philips	Brilliance (10,16, 64 Big Bore)
Toshiba	Aquilion
	Aquilion Lightning
GE Medical Systems	Revolution EVO
	LightSpeed VCT
	Optima CT660
	Discovery MI
	Revolution HD

Table B4: CT scanner manufacturers and models. List of the manufacturers and specific models of CT scanners used for image acquisition for both internal and external cohorts.

		Missing	Overall	Non-responders	Responders	P-Value (adjusted)
n images			368	169	199	
Manufacturer, n (%)	GE Medical Systems	0	6 (1.6)	2 (1.2)	4 (2.0)	1.000
	Philips		92 (25.0)	45 (26.6)	47 (23.6)	
	Siemens		242 (65.8)	110 (65.1)	132 (66.3)	
	Toshiba		28 (7.6)	12 (7.1)	16 (8.0)	
Exposure (mAs), mean (SD)		0	192.2 (65.2)	190.8 (61.9)	193.4 (68.0)	1.000
kVp (kV), mean (SD)		0	113.6 (11.6)	113.4 (11.8)	113.7 (11.4)	1.000
Pixel Spacing (mm), mean (SD)		0	0.7 (0.1)	0.7 (0.1)	0.7 (0.1)	1.000
Slice Thickness (mm), mean (SD)		0	2.2 (1.4)	2.1 (1.4)	2.2 (1.4)	1.000
Exposure Time (ms), mean (SD)		0	496.0 (55.8)	488.6 (61.2)	502.4 (50.0)	0.240
stdNoise, mean (SD)		0	26.7 (4.8)	26.5 (4.5)	26.8 (5.2)	1.000

Table B5: CT image acquisition and reconstruction parameters for images from responders and non-responders across the internal cohort. Adjusted p-values (using Bonferroni correction) were calculated using two-sample T-tests for continuous variables and Chi-square tests for categorical variables. SD represents the standard deviation.

		Missing	Overall	Non-responders	Responders	P-Value (adjusted)
n images			94	58	36	
Manufacturer, n (%)	Philips	0	48 (51.1)	31 (53.4)	17 (47.2)	1.000
	Siemens		45 (47.9)	27 (46.6)	18 (50.0)	
	GE Medical Systems		1 (1.1)		1 (2.8)	
Exposure (mAs), mean (SD)		0	120.6 (77.2)	126.1 (83.0)	111.8 (66.8)	1.000
kVp (kV), mean (SD)		0	137.9 (6.7)	137.6 (6.3)	138.3 (7.4)	1.000
Pixel Spacing (mm), mean (SD)		0	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)	1.000
Slice Thickness (mm), mean (SD)		0	3.0 (0.2)	3.0 (0.0)	2.9 (0.4)	1.000
Exposure Time (ms), mean (SD)		28	779.0 (162.1)	788.0 (201.5)	765.2 (68.1)	1.000
stdNoise, mean (SD)		0	24.3 (4.2)	24.3 (4.4)	24.3 (4.1)	1.000

Table B6: CT image acquisition and reconstruction parameters for images from responders and non-responders across external cohort. Adjusted p-values (using Bonferroni correction) were calculated using two-sample T-tests for continuous variables and Chi-square tests for categorical variables. SD represents the standard deviation.

		Missing	Overall	External Test	Internal Test	P-Value (adjusted)
n			196	94	102	
Manufacturer, n (%)	GE Medical Systems	0	2 (1.0)	1 (1.1)	1 (1.0)	0.002
	Philips		73 (37.2)	48 (51.1)	25 (24.5)	
	Siemens		113 (57.7)	45 (47.9)	68 (66.7)	
	Toshiba		8 (4.1)		8 (7.8)	
Exposure (mAs), mean (SD)		0	156.8 (78.7)	120.6 (77.2)	190.1 (64.2)	<0.001
kVp (kV), mean (SD)		0	125.3 (15.3)	137.9 (6.7)	113.7 (11.4)	<0.001
Pixel Spacing (mm), mean (SD)		0	0.8 (0.1)	0.8 (0.1)	0.7 (0.1)	<0.001
Slice Thickness (mm), mean (SD)		0	2.6 (1.1)	3.0 (0.2)	2.2 (1.4)	<0.001
Exposure Time (ms), mean (SD)		28	605.0 (182.2)	779.0 (162.1)	492.4 (73.4)	<0.001
stdNoise, mean (SD)		0	25.6 (4.6)	24.3 (4.2)	26.8 (4.6)	0.001

Table B7: CT image acquisition and reconstruction parameters for images from the internal and external test sets. Adjusted p-values (using Bonferroni correction) were calculated using two-sample T-tests for continuous variables and Chi-square tests for categorical variables. SD represents the standard deviation.

Data type	Harmonization Type	Batch effect	AUC [95% CI]	ACC [95% CI]	SENS [95% CI]	SPEC [95% CI]	PREC [95% CI]	bACC [95% CI]
Original	No harmo	-	0.695 [0.529, 0.848]	0.674 [0.535, 0.814]	0.667 [0.450, 0.864]	0.682 [0.480, 0.875]	0.667 [0.450, 0.857]	0.674 [0.533, 0.813]
Stabilized	No harmo	-	0.641 [0.463, 0.801]	0.535 [0.372, 0.674]	0.571 [0.350, 0.789]	0.500 [0.286, 0.714]	0.522 [0.300, 0.720]	0.536 [0.377, 0.690]
Harmonized	No harmo	-	0.671 [0.495, 0.844]	**0.721** [0.581, 0.860]	**0.762** [0.571, 0.933]	0.682 [0.476, 0.867]	0.696 [0.500, 0.875]	**0.722** [0.581, 0.857]
Original	ComBat	kVp	0.503 [0.336, 0.677]	0.581 [0.442, 0.721]	0.571 [0.368, 0.783]	0.591 [0.381, 0.792]	0.571 [0.360, 0.773]	0.581 [0.430, 0.726]
Stabilized	ComBat	kVp	0.663 [0.495, 0.827]	0.674 [0.535, 0.814]	**0.762** [0.571, 0.933]	0.591 [0.381, 0.800]	0.640 [0.440, 0.818]	0.676 [0.536, 0.815]
Harmonized	ComBat	kVp	0.680 [0.509, 0.843]	0.651 [0.512, 0.791]	0.571 [0.350, 0.789]	0.727 [0.524, 0.905]	0.667 [0.421, 0.882]	0.649 [0.508, 0.793]
Original	ComBat	Manufacturer	**0.712** [0.550, 0.857]	0.674 [0.535, 0.814]	0.571 [0.360, 0.783]	**0.773** [0.593, 0.941]	**0.706** [0.462, 0.917]	0.672 [0.534, 0.810]
Stabilized	ComBat	Manufacturer	0.659 [0.490, 0.827]	0.674 [0.535, 0.814]	**0.762** [0.571, 0.933]	0.591 [0.381, 0.800]	0.640 [0.440, 0.818]	0.676 [0.536, 0.815]
Harmonized	ComBat	Manufacturer	0.607 [0.426, 0.793]	0.651 [0.512, 0.791]	0.524 [0.304, 0.750]	**0.773** [0.591, 0.947]	0.688 [0.444, 0.923]	0.648 [0.508, 0.789]
Original	ComBat	Slice Thickness	0.663 [0.491, 0.825]	0.581 [0.442, 0.721]	0.619 [0.409, 0.826]	0.545 [0.333, 0.762]	0.565 [0.350, 0.762]	0.582 [0.434, 0.726]
Stabilized	ComBat	Slice Thickness	0.640 [0.463, 0.807]	0.674 [0.535, 0.814]	0.714 [0.500, 0.900]	0.636 [0.429, 0.842]	0.652 [0.440, 0.842]	0.675 [0.533, 0.817]
Harmonized	ComBat	Slice Thickness	0.701 [0.528, 0.857]	0.674 [0.535, 0.814]	0.619 [0.409, 0.840]	0.727 [0.538, 0.909]	0.684 [0.450, 0.889]	0.673 [0.531, 0.814]
Original	ComBat	StdNoise	0.708 [0.549, 0.861]	**0.721** [0.581, 0.860]	**0.762** [0.571, 0.933]	0.682 [0.480, 0.875]	0.696 [0.500, 0.880]	**0.722** [0.590, 0.860]
Stabilized	ComBat	StdNoise	0.598 [0.424, 0.764]	0.558 [0.395, 0.698]	0.667 [0.444, 0.857]	0.455 [0.238, 0.667]	0.538 [0.333, 0.724]	0.561 [0.409, 0.708]
Harmonized	ComBat	StdNoise	0.701 [0.537, 0.859]	0.628 [0.488, 0.767]	0.524 [0.304, 0.731]	0.727 [0.529, 0.905]	0.647 [0.400, 0.875]	0.626 [0.483, 0.766]

Table B8: Comparison of response prediction performance among longitudinal models in the internal test set using various image and ComBat harmonization techniques, assessed by evaluating AUC, ACC, SENS, SPEC, PREC and bACC, respectively. For each metric, the 95% confidence interval is shown and the highest value is highlighted in bold.

Data type	Harmonization Type	Batch effect	AUC [95% CI]	ACC [95% CI]	SENS [95% CI]	SPEC [95% CI]	PREC [95% CI]	bACC [95% CI]
Original	No harmo	-	0.692 [0.476, 0.875]	0.590 [0.436, 0.744]	0.538 [0.346, 0.733]	0.692 [0.417, 0.923]	0.778 [0.562, 0.947]	0.615 [0.442, 0.769]
Stabilized	No harmo	-	0.673 [0.489, 0.840]	0.564 [0.410, 0.718]	0.500 [0.308, 0.692]	0.692 [0.412, 0.923]	0.765 [0.533, 0.947]	0.596 [0.424, 0.760]
Harmonized	No harmo	-	0.701 [0.518, 0.859]	0.641 [0.487, 0.795]	0.577 [0.400, 0.762]	**0.769** [0.500, 1.000]	0.833 [0.643, 1.000]	0.673 [0.512, 0.822]
Original	ComBat	kVp	0.385 [0.209, 0.573]	0.513 [0.359, 0.667]	0.654 [0.458, 0.833]	0.231 [0.000, 0.467]	0.630 [0.444, 0.800]	0.442 [0.296, 0.599]
Stabilized	ComBat	kVp	0.655 [0.469, 0.833]	0.667 [0.513, 0.795]	0.654 [0.462, 0.833]	0.692 [0.417, 0.923]	0.810 [0.619, 0.957]	0.673 [0.501, 0.828]
Harmonized	ComBat	kVp	0.604 [0.412, 0.784]	0.564 [0.410, 0.718]	0.500 [0.308, 0.700]	0.692 [0.400, 0.929]	0.765 [0.533, 0.944]	0.596 [0.425, 0.757]
Original	ComBat	Manufacturer	0.765 [0.592, 0.912]	0.667 [0.513, 0.821]	0.615 [0.429, 0.808]	**0.769** [0.500, 1.000]	**0.842** [0.650, 1.000]	0.692 [0.523, 0.834]
Stabilized	ComBat	Manufacturer	0.737 [0.551, 0.900]	0.744 [0.590, 0.872]	0.769 [0.593, 0.923]	0.692 [0.417, 0.923]	0.833 [0.667, 0.960]	0.731 [0.563, 0.878]
Harmonized	ComBat	Manufacturer	0.701 [0.507, 0.873]	0.667 [0.513, 0.821]	0.769 [0.600, 0.920]	0.462 [0.182, 0.750]	0.741 [0.565, 0.900]	0.615 [0.447, 0.783]
Original	ComBat	Slice Thickness	0.710 [0.513, 0.892]	0.692 [0.538, 0.821]	0.769 [0.600, 0.920]	0.538 [0.273, 0.812]	0.769 [0.593, 0.923]	0.654 [0.493, 0.806]
Stabilized	ComBat	Slice Thickness	0.729 [0.562, 0.877]	0.641 [0.487, 0.795]	0.654 [0.478, 0.833]	0.615 [0.333, 0.875]	0.773 [0.588, 0.941]	0.635 [0.474, 0.794]
Harmonized	ComBat	Slice Thickness	0.754 [0.586, 0.894]	0.667 [0.513, 0.821]	0.615 [0.429, 0.800]	**0.769** [0.500, 1.000]	0.842 [0.667, 1.000]	0.692 [0.532, 0.840]
Original	ComBat	StdNoise	**0.790** [0.603, 0.951]	**0.744** [0.590, 0.872]	**0.808** [0.650, 0.952]	0.615 [0.333, 0.889]	0.808 [0.643, 0.955]	**0.712** [0.556, 0.861]
Stabilized	ComBat	StdNoise	0.632 [0.450, 0.802]	0.564 [0.410, 0.718]	0.577 [0.385, 0.769]	0.538 [0.250, 0.818]	0.714 [0.500, 0.900]	0.558 [0.383, 0.731]
Harmonized	ComBat	StdNoise	0.698 [0.513, 0.855]	0.667 [0.513, 0.795]	0.615 [0.423, 0.808]	**0.769** [0.500, 1.000]	0.842 [0.647, 1.000]	0.692 [0.529, 0.833]

Table B9: Comparison of response prediction performance among longitudinal models in the external test set using various image and ComBat harmonization techniques, assessed by evaluating AUC, ACC, SENS, SPEC, PREC and bACC, respectively. For each metric, the 95% confidence interval is shown and the highest value is highlighted in bold.