FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

CT-Based Lung Adenocarcinoma Prediction using 2D and 3D CNN Architectures

Tânia Mendes

Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Hélder Oliveira

August 9, 2023

© Tânia Mendes, 2023

Abstract

Cancer is recognized as one of the most alarming diseases worldwide, with lung cancer representing a significant portion of both its incidence and mortality rates. Adenocarcinoma is the most common subtype of lung cancer, therefore the importance of its early and accurate diagnosis is undeniable since it has a direct impact on the outcome of the patient life.
Computed tomography is a widely used radiological imaging modality that allows the detection of the tumor, even in its early stages. Through the computed tomography scan, it is possible to identify the location, size, and shape of the tumor, but in order to assess its histological subtype a biopsy is the standard procedure adopted. Biopsy consists in collecting a sample from the tumor area to be further analyzed. Being an invasive procedure, the biopsy can have adverse consequences on patients and does not provide the results immediately, due to the need for laboratory analysis of the sample. Deep learning algorithms have proven their ability to recognize patterns in computed tomographies scans and perform the classification accordingly. Therefore, this type of models can be used as a non-invasive method to identify the lung cancer subtype using computed tomography scans.
In the current study, two main approaches were employed: using 2D and 3D models for classification, both based on the ResNet architecture. The 2D models were firstly trained with three different datasets: the Lung-PET-CT-Dx, the NSCLC-Radiomics and a combined dataset, that results from the merging of the other two. Having the datasets defined, the data went on a preprocessing stage, where image normalization and geometric data augmentation techniques were applied. The Lung-PET-CT-Dx trained model exhibited superior overall performance, attaining an AUC value of 0.638 on the test set. Consequently, this model was selected as training set for the subsequent experiments. Then, a comparison was made between employing a single slice and three slices per patient in the training set, with the model that was trained with three slices per patient showing a substantial improvement, resulting in an AUC of 0.717.
Regarding the 3D approach, experiments were conducted to evaluate the performances of the models trained from scratch and those employing transfer learning from MedicalNet, with two, four and six layers being fine-tuned. The ones that achieved best results were the model trained from scratch, which obtained an AUC value of 0.779, and the model with six layers trained, achieving an AUC value of 0.757.
After evaluating the performance of the models and the results, certain points been identified in the methodology that may present areas for improvement. Consequently, in future work, techniques such as cross-validation, alternative approaches to data augmentation, and resampling the slice thickness of the 3D volumes could be considered to potentially enhance the results.

Keyworkds: *Lung Cancer, Adenocarcinoma, CT, Deep Learning, Classification, 2D CNN, 3D CNN*

Resumo

O cancro é reconhecido como uma das doenças mais alarmantes a nível mundial, sendo que o do pulmão representa uma parte significativa das suas taxas de incidência e mortalidade. Adenocarcinoma é o subtipo mais comum de cancro do pulmão, pelo que a importância do seu diagnóstico precoce e preciso é inegável, tendo um impacto direto no resultado da vida do doente.
A tomografia computadorizada é uma modalidade de imagem radiológica amplamente utilizada que permite a deteção do tumor, mesmo em estadios iniciais, sendo também possível identificar a localização, tamanho e forma do tumor. No entatno, para avaliar o seu subtipo histológico, a biópsia é o procedimento padrão adotado. A biopsia consiste na recolha de uma amostra da área do tumor, que é posteriormente analisada. Sendo um procedimento invasivo e que não fornece resultados imediatos, pode ter consequências para os pacientes. Algorítmos de deep learning já provaram a sua capacidade de reconhecer padrões w realizar a classificação em tomografias computadorizadas. Este tipo de modelos contitui um método não invasivo para identificar o subtipo de cancro do pulmão através de tomografias computadorizadas.
No presente estudo, foram adotadas duas abordagens principais: utilizando modelos 2D e 3D para classificação, ambos baseados na arquitetura ResNet. Os modelos 2D foram inicialmente treinados com três conjuntos de dados diferentes: o Lung-PET-CT-Dx, o NSCLC-Radiomics e um conjunto de dados combinado, que resulta da fusão dos outros dois. Tendo os conjuntos de dados definidos, os dados passaram por uma fase de pré-processamento, onde métodos de normalização de imagem e técnicas geométricas de aumento de dados foram aplicadas. O modelo treinado Lung-PET-CT-Dx apresentou um desempenho global superior, atingindo um valor de AUC de 0,638 no conjunto de teste. Consequentemente, este modelo foi selecionado como conjunto de treino para as experiências seguintes. Posteriormente, foi efectuada uma comparação entre a utilização de uma única camada e de três camadas por doente no conjunto de treino, tendo o modelo que foi treinado com três slices por paciente apresentado uma melhoria substancial nos resultados, atingindo uma AUC de 0,717.
Em relação à abordagem 3D, foram realizadas experiências para avaliar o desempenho dos modelos trained from scratch e dos que utilizam transfer learning do MedicalNet, com duas, quatro e seis camadas a serem treinadas. Os que obtiveram melhores resultados foram o modelo trained from scratch, que obteve um valor de AUC de 0,779, e o modelo com seis camadas treinadas, que obteve um valor de AUC de 0,757.
Analisando o desempenho e resultados dos modelos, foram identificados alguns pontos de melhoria na metodologia. Consequentemente, em trabalhos futuros, podem ser consideradas técnicas como a cross-validation, abordagens alternativas para o aumento de dados e a reamostragem da espessura das camadas dos volumes 3D, para potencialmente melhorar os resultados.

Palavras Chave: *Cancro do Pulmão, Adenocarcinoma, TAC, Deep Learning, Classificação, 2D CNN, 3D CNN*

Acknowledgments

I would like to express my deepest gratitude to all those who have crossed paths and supported me along my dissertation journey. This accomplishment would not have been possible without everyone's contribution, support, and encouragement.
First, I would like to thank my supervisor, Hélder Oliveira, my co-supervisors at INESC TEC, Margarida Gouveia and Joana Sousa, and to the remaining of the team, Tânia Pereira and Francisco Silva. Thank you for the dedication in following my work, for the guidance, and constant support throughout this dissertation.
I would like to extend my thanks to my family, specially my sister and parents, and my friends, for their unconditional support, encouragement and understanding throughout this journey.
I feel truly grateful to all those who contributed to the completion of this dissertation, and to those who played a significant role in shaping my academic journey. Thank you all for believing in me and helping me reach this milestone in my academic career.
This work is co-financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project HfPT, with reference 41.

Tânia Mendes

Contents

1 Introduction **1**
1.1 Motivation and Importance . . . 1
1.2 Possible Solutions and Goals . . . 3
1.3 Document Structure . . . 3

2 Literature Review **5**
2.1 Datasets . . . 5
2.1.1 Chest-CT—Scan Images . . . 5
2.1.2 Long and Short Survival in Adenocarcinoma Lung CTs (LUAD-CT-Survival) 6
2.1.3 Quantitative CT Descriptors associate tumor shape complexity and intra-tumor heterogeneity with prognosis in lung adenocarcinoma (LungCT-Diagnosis) . . . 6
2.1.4 CT Volume Samples for Lung Adenocarcinoma Classification . . . 7
2.1.5 A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lung-PET-CT-Dx) . . . 7
2.1.6 NSCLC-Radiomics . . . 8
2.1.7 Datasets Overview . . . 9
2.2 Previous Research Works on Lung Cancer Classification . . . 10
2.2.1 Lung cancer histology classification from CT images based on radiomics and deep learning models . . . 10
2.2.2 An Approach Toward Automatic Classification of Tumor Histopathology of Non-Small Cell Lung Cancer Based on Radiomic Features . . . 12
2.2.3 Classification of Lung Nodules based on Transfer Learning with K-Nearest Neighbor (KNN) . . . 13
2.2.4 Deep Features Based Automated Multimodel System for Classification of Non-Small Cell Lung Cancer . . . 14
2.2.5 Lung Cancer Detection and Classification from Chest-CT Scans Using Machine Learning Techniques . . . 15
2.2.6 Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples . . . 16
2.2.7 Minimum Sample Size Estimate for Classifying Invasive Lung Adenocarcinoma . . . 16
2.2.8 Deep learning for predicting subtype classification and survival of lung adenocarcinoma on computed tomography . . . 17
2.2.9 Approaches and Results Overview . . . 19

3 Methodology **23**
3.1 Dataset Preparation . . . 23

3.1.1 Dataset Selection 23
3.1.2 Data Filtering 24
3.2 Preprocessing 26
3.2.1 Segmentation 26
3.2.2 Image Normalization 27
3.2.3 Data Augmentation 28
3.3 Feature Extraction and Classification 29
3.3.1 3D Approach 29
3.3.2 2D Approach 30
3.4 Training and Evaluation 31
3.5 Methodology Overview 32

4 Results and Discussion **35**
4.1 Study 1 - Comparison between the 2D models' performances when trained with three different datasets 35
4.2 Study 2 - Evaluation of using one slice and three slices per patient for training . . 39
4.3 Study 3 - Comparison between the 3D models when trained from scratch and when transfer learning is employed 40
4.4 Study 4 - Comparison between the performance of the 2D and the 3D models . . 44

5 Conclusion **47**

A Cases Removed from the Datasets **49**
A.1 NSCLC-Radiomics 49
A.2 Lung-PET-CT-Dx 50

References **51**

List of Figures

1.1 CT images of ADC, LCC, and SCC. 1
1.2 CT scanner from Sri Kota Specialist Medical Centre. 2

2.1 Samples of adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and normal cell from Chest-CT dataset. 6
2.2 Samples of adenocarcinoma with corresponding tumor segmentation from LUAD-CT-Survival dataset. 6
2.3 Sample of adenocarcinoma from LungCT-Diagnosis dataset. 7
2.4 Samples of AAH, AIS, and MIA from LUAD-CT-Volume dataset. 7
2.5 Samples of adenocarcinoma, small cell carcinoma, large cell carcinoma, and squamous cell carcinoma, with corresponding bounding boxes, from Lung-PET-CT-Dx dataset. 8
2.6 Samples of adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and not otherwise specified, with corresponding tumor segmentation from NSCLC Radiomics dataset. 8
2.7 Radiomic features approach. 11
2.8 Pre-trained CNN features approach. 11
2.9 CNN features + LSTM approach (left bottom) and fine-tuned CNN approach (right bottom). 12
2.10 Pipeline with SVM, XGBoost, and FCNN for classification of deep features. . . . 14
2.11 Sequence for classifying the features extracted by AlexNet with SVM. 17
2.12 Deep learning model (upper half) and radiomics model (bottom half). 18
2.13 Deep radiomics model. 19

3.1 Generic pipeline describing the constituting steps of the proposed methodology. . 23
3.2 CT images and corresponding segmentation masks from the NSCLC-Radiomics dataset illustrating examples from cases that have been removed based on Criterion 1, Criterion 2, and Criterion 3. 25
3.3 Steps to transform the NSCLC-Radiomics segmentation masks into bounding boxes: starting from the original segmentation mask, then defining the boundaries based on the original mask edges and filling to obtain the final bounding box. 27
3.4 Original (first row) and corresponding preprocessed (second row) CT samples of adenocarcinoma and non-adenocarcinoma from the Lung-PET-CT-Dx dataset, and adenocarcinoma and non-adenocarcinoma from the NSCLC-Radiomics dataset. . 28
3.5 Workflow for 3D approach using ResNet10 architecture from MedicalNet. 30
3.6 Methodology overview with the sequential steps employed in the 2D and 3D approaches. 32

4.1 Loss and AUC evolution in the train and validation sets during the training phase of the 2D approach. The parameters utilized in the four experiments were: batch size of 32, 0.5 dropout, SGD without momentum, 0.0005 learning rate, and 0.01 weight decay. 36
4.2 Confusion matrices on the 2D approach. Each row corresponds respectively to the models trained on the NSCLC-Radiomics, on the Lung-PET-CT-Dx, on the combined and on the multi-slice Lung-PET-CT-Dx. 39
4.3 Loss and AUC evolution in the train and validation sets during the training phase of the 3D approach. Concerning the parameters, the batch size of 32, 0.5 dropout and 0.01 weight decay were common in all experiments. For the TFS model, SGD without momentum was utilized with learning rate of 0.0001, and for the remaining three experiences was the SGD with 0.9 momentum and 0.001 learning rate. 41
4.4 Confusion matrices on the 3D approach. Each row corresponds respectively to the models trained from scratch, with six layers trained, with four layers trained, and with two layers trained. 44

A.1 Block diagram for the NSCLC-Radiomics dataset, including the exclusion criteria and the number of patients removed at each step. 49
A.2 Block diagram for the Lung-PET-CT-Dx dataset, including the exclusion motives and the number of patients and series removed at each step. 50

List of Tables

2.1 Overview of datasets. 10
2.2 Overview of previous researches approaches and corresponding accuracy. The top results achieved in each study are marked in bold. 21

3.1 Distribution per dataset of number of patients and classes, specifically the adenocarcinoma (ADC) class, and the non-ADC class, which encompasses squamous cell carcinoma (SCC), large cell carcinoma (LCC), small cell carcinoma (SCLC) and not otherwise specified (NOS). 26
3.2 Range of values defined for the parameters of the data augmentation transformations. 28
3.3 Experimented values for the training hyperparameters of each model, respectively in the 2D and 3D models trained from scratch (TFS), and in the 3D models in which transfer learning (TL) was employed. 31

4.1 Results of the 2D approach, including the training set, the epoch selected to test the model, corresponding to the maximum validation AUC, the test set and the evaluation metrics: AUC, balanced accuracy, precision, recall, F1-Score, and specificity. The top results achieved for each metric are highlighted in bold. 37
4.2 Results of the 3D approach, including the training set, the number of layers trained, the epoch selected to test the model, corresponding to the maximum validation AUC, the test set and the evaluation metrics: AUC, balanced accuracy, precision, recall, F1-Score, and specificity. The top results achieved for each metric are highlighted in bold. 42

Abbreviations and Symbols

ADC	Adenocarcinoma
AUC	Area Under the Receiver Operating Characteristic Curve
BFFC	Before-Final-FC
CNN	Convolutional Neural Network
CT	Computed Tomography
DCT	Discrete Cosine Transform
FC	Fully Connected
FCNN	Fully Connected Neural Network
GAP	Global Average Pooling
GGNs	Ground Glass Nodules
GLCM	Gray-Level Co-occurence Matrix
GLRLM	Gray-Level Run-Length Matrix
GTV	Gross Tumor Volume
HU	Hounsfield Units (HU)
IA	Invasive Adenocarcinoma
KNN	K-Nearest Neighbors
LBP	Local Binary Pattern
LCC	Large Cell Carcinoma
LSTM	Long Short-term Memory
MRI	Magnetic Resonance Imaging
NOS	Not Otherwise Specified
NSCLC	Non-small Cell Lung Cancer
pAAC	Predominant Acinar Adenocarcinoma
PCA	Principal Component Analysis
pLAC	Predominant Lepidic Adenocarcinoma
pMAC	Predominant Micropappillary Adenocarcinoma
pPAC	Predominant Pappilary Adenocarcinoma
pSAC	Predominant Solid Adenocarcinoma
RF	Random Forest
ROI	Region of Interest
SCC	Squamous Cell Carcinoma
SCLC	Small Cell Lung Cancer
SGD	Stochastic Gradient Descent
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TFS	Trained From Scratch
VC	Vapnik-Chervonenkis

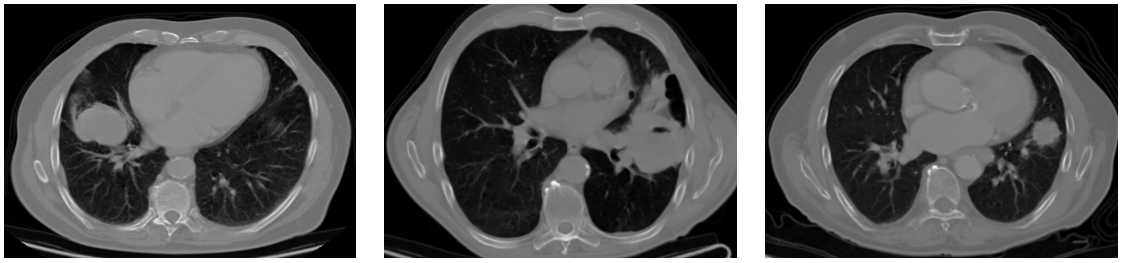
Chapter 1

Introduction

1.1 Motivation and Importance

Lung cancer is the second most common type of cancer, representing the leading cause of cancer-related mortality among both male and female genders. This type of cancer is characterized by the unexpected and uncontrollable division and growth of lung cells, forming a tumor that can be either malignant or benign. In 2023, the estimated daily average of lung cancer deaths is of 350 in the United States, resulting in a projected total of around 130 thousand deaths by the year's end.[1].

Small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) are the two main types of lung cancer, the latter representing 84% of all lung cancer cases. According to the World Health Organization (WHO), NSCLC is divided into four subtypes: adenocarcinoma (ADC), squamous cell carcinoma (SCC), large cell carcinoma (LCC), and not otherwise specified (NOS) [2]. Figure 1.1 shows a CT image of each of the main types of NSCLC (ADC, LCC, and SCC). ADC is the



(a) (b) (c)

Figure 1.1: CT images of ADC (1.1a), LCC (1.1b), and SCC (1.1c) [3].

most common type of lung cancer, accounting for 40% of the total cases, while SCC and LCC correspond approximately to 25%-30% and 5%-10%, respectively [4]. NSCLC can start anywhere in the lung while adenocarcinoma begins to grow in epithelial cells, which are the most common type of cells in the lung. These cells are located in the periphery of the outer part of the lung and produce mucus [5].

Patients with NSCLC can develop a variety of symptoms, being cough, fatigue, shortness of breath, and chest pain some of the most frequent ones. In cases where no symptoms are developed, Computed tomography (CT) scans can still detect the tumor. A CT is an examination of a specific part of the body, in which x-rays are taken in slices from different angles, forming a three-dimensional image [5] and is performed by a CT scanner (see Figure 1.2). CT scans allows

Figure 1.2: CT scanner from Sri Kota Specialist Medical Centre [6].

physicians to detect the location, size and shape of the tumor, but it does not provide accurate information about the tumor's histological subtype and stage. The tumor stage indicates the evolution of the cancer, taking into account its location and whether it has spread to other parts of the body. For NSCLC, there are five stages, from zero to IV, the latter being the most severe and means that the cancer has spread to distant parts of the body [5].

There are several options possible for cancer treatment, but in order to direct the patient to the most appropriate treatment for their particular case, the physician needs to have information about the type of cancer, the stage of the tumor, the patient's lifestyle, and medical history. The main and most well-known types of treatment for patients with NSCLC are surgery, radiation therapy, and chemotherapy. In surgery, the goal is to remove the tumor and the nearby healthy tissue. Radiation therapy consists in destroying the cancer cells using X-ray beams. In chemotherapy, chemical drugs are used to eliminate cancer cells. Although these treatments have proven to be effective, they are extremely invasive and aggressive for the patient, who can suffer from multiple side effects. If the treatment is successful, cancer may go into remission, which can be temporary or permanent. Otherwise, if the cancer cannot be controlled, the patient is considered in a terminal state [5]. In order to assess the lung cancer histological subtype and the tumor stage, the standard procedure is to perform a tumor biopsy. A biopsy consists of taking a sample from the tumor area for further laboratory analysis. Despite its effectiveness, the biopsy is not optimal, as it is an invasive procedure and therefore can be painful for the patients. Furthermore, the complexity of the

biopsy is increased when the tumor resides in anatomical regions that are challenging to access. Another disadvantage of this procedure is that it has to be performed by a medical professional and the tissue sample taken needs to be further analyzed, which means that the results will not be available immediately.

The 5-year survival rate of NSCLC patients is heavily influenced by the tumor stage at diagnosis. If at the time of detection the cancer is in a early stage, the 5-year survival rate is around 68%, whereas if it is in an advanced stage and has already spread to other organs, the 5-year survival rate can go down to 10% to 0% [4]. These statistics emphasize the importance of an early and correct diagnosis in the outcome of the patient's life.

Considering these factors, the detection and identification of the cancer subtype at an early stage is crucial, and therefore the methods currently used need to evolve or new approaches need to be adopted so that the whole process from diagnosis to treatment can become faster and more effective.

1.2 Possible Solutions and Goals

Machine learning models are able to recognize patterns and detect anomalies in CT scans with reasonable efficiency, even in early stages [7, 8]. These models can be used to create computer-aided diagnosis (CAD) systems to support physicians, and can also decrease the need for biopsy, reducing patient suffering and allowing results to be obtained sooner. Keeping in mind the end goal, the models can be trained for several classification tasks such as histological tumor subtype, survival rate, and tumor stage.

Given the high incidence of adenocarcinoma, it is important to forward patients with diagnosed adenocarcinoma to the most appropriate treatment, aiming to increase the probability of survival. Therefore, the use of machine and learning algorithms to predict adenocarcinoma is of utmost importance. The goal of this work was to develop an accurate and reliable model for the lung adenocarcinoma classification, thereby providing physicians with an additional supportive tool to aid in the diagnosis process of this particular subtype of lung cancer.

1.3 Document Structure

This document is organized as follows: Chapter 1 is the current one, where the problem at hand, motivation, goals and medical background were introduced; Chapter 2 presents an analysis of previous research works focused on the use of machine and deep learning in this area, along with potentially useful datasets that are publicly available; Chapter 3 describes the methodology followe for the lung adenocarcinoma classification based on CT images; in Chapter 4 the results obtained in each conducted experience are presented, followed by the discussion in chapter 5; lastly, Chapter 6 is dedicated to the conclusions derived from the results.

Chapter 2

Literature Review

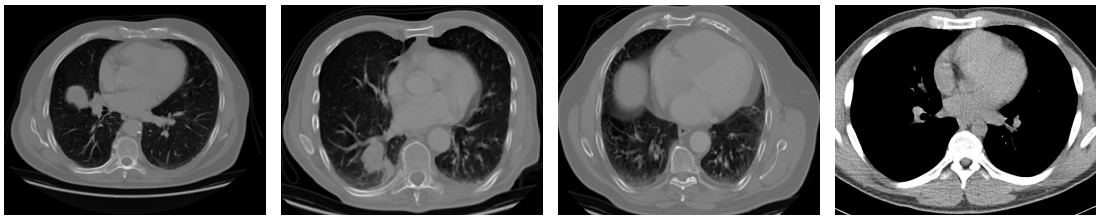
The current point of development and knowledge about the use of machine learning for the classification of lung cancer histology through CT images will be addressed. The most notable techniques and results found from the literature research will be described, as well as the available datasets that can possibly be useful.
This chapter is organized into two sections, as follows: first, a section about datasets, where several publicly available datasets composed of lung CT will be presented; and second, a section focused on previous works, where multiple research works concerning this area will be detailed.

2.1 Datasets

In this section, an analysis of several publicly available datasets of lung CT scans is performed. The datasets are described in terms of type and amount of data, classes represented and corresponding distribution, tumor segmentation and how it was performed, and also in terms of file formats of each data type provided. Samples representing each class of the datasets will be presented together with, if available, corresponding tumor segmentations.

2.1.1 Chest-CT—Scan Images

The Chest-CT dataset [3] is available on Kaggle and contains samples corresponding to the following classes: adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and normal cell. The dataset has 1000 files, either in JPG or PNG format, being adenocarcinoma the most represented class with 338 images, followed by squamous cell carcinoma with 260 images, normal cell with 215 images, and finally large cell carcinoma with 187 images. Samples of each class are presented in Figure 2.1.



(a) (b) (c) (d)

Figure 2.1: Samples of adenocarcinoma (2.1a), large cell carcinoma (2.1b), squamous cell carcinoma (2.1c), and normal cell (2.1d) from Chest-CT dataset [3].

2.1.2 Long and Short Survival in Adenocarcinoma Lung CTs (LUAD-CT-Survival)

The LUAD-CT-Survival dataset [9], available on The Cancer Imaging Archive (TCIA), contains contrast-enhanced CT images acquired at the Moffitt Cancer Center Research Institute. The CT images are in DICOM format and belong to 40 patients with adenocarcinoma verified with a biopsy, with 2 years of monitoring.

The segmentation of the tumors is available in NIfTI (.nii) format and was performed using region-growing with the seeds being selected by radiologists (see Figure 2.2). Information about the patients' survival quartile as well as the images' features are specified in a CSV file.

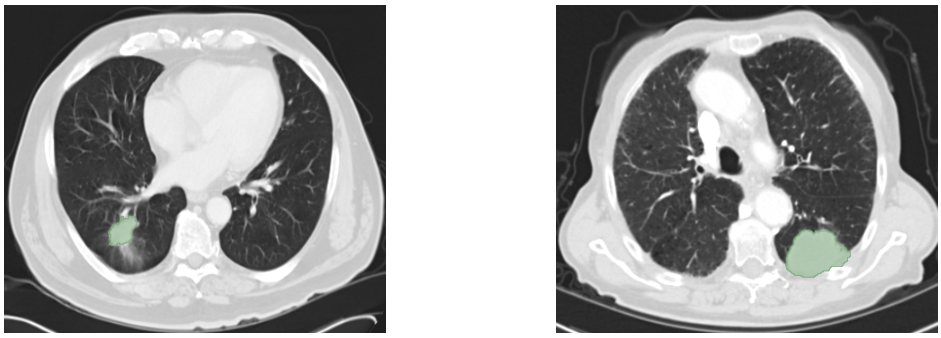


Figure 2.2: Samples of adenocarcinoma with corresponding tumor segmentation from LUAD-CT-Survival dataset [9].

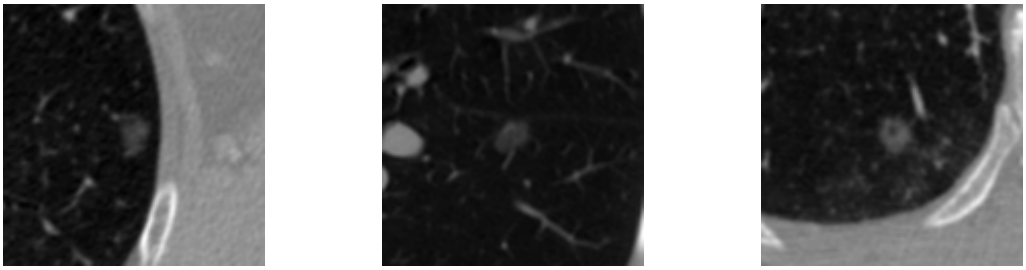
2.1.3 Quantitative CT Descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma (LungCT-Diagnosis)

The LungCT-Diagnosis dataset [10] is also available on TCIA and includes CT scans of pre-surgery and contrast-enhanced adenocarcinoma tumors, with slice thicknesses between 3 and 6mm. The CT images are in DICOM format, having 4682 images from 61 patients. The representative tumor slices are available in a XLS file, containing the patient ID, instance number, and the tumor position coordinates. Clinical data, such as pathology stage and survival status, are also available in a DOC file. Figure 2.3 illustrates a sample from this dataset.

Figure 2.3: Sample of adenocarcinoma from LungCT-Diagnosis dataset [10].

2.1.4 CT Volume Samples for Lung Adenocarcinoma Classification

This dataset can be accessed through Mendeley Data and is composed of 3D CT scans of adenocarcinoma [11]. The images of this dataset are available in npy files and are divided into three labeled folders, each corresponding to a subtype of lung adenocarcinoma. The aggregated set of folders results in 1050 samples, with the following distribution: 21 samples of atypical adenomatous hyperplasia (AAH), 444 samples of adenocarcinoma in situ (AIS), and 585 samples of minimally invasive adenocarcinoma (MIA). The samples of this dataset are patches with 128x128x128 resolution of the CT scan, where the adenocarcinoma tumor is located at the center of the 3D volume, as shown in Figure 2.4. From now on, this dataset will be referred to as LUAD-CT-Volume.



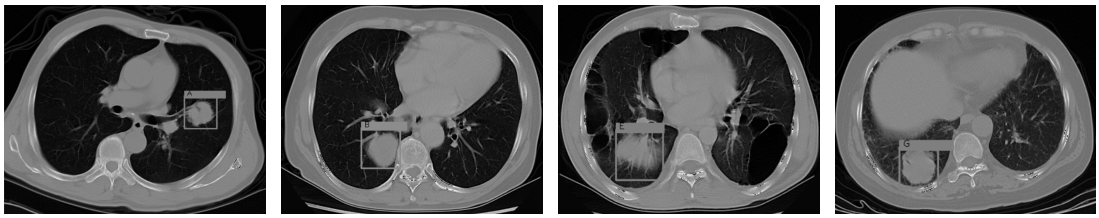
(a) (b) (c)

Figure 2.4: Samples of AAH (2.4a), AIS (2.4b), and MIA (2.4c) from LUAD-CT-Volume dataset[11].

2.1.5 A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lung-PET-CT-Dx)

The Lung-PET-CT-Dx dataset [12] is available on TCIA and contains CT and PET-CT images of 355 lung cancer patients. Adenocarcinoma, small cell carcinoma, large cell carcinoma, and squamous cell carcinoma are the lung cancer subtypes represented in this dataset, confirmed through biopsy, and are identified with the following labels: A, B, E, and G, respectively. Adenocarcinoma is the most represented class in this dataset, with 251 patients, then squamous cell carcinoma with

61, small cell carcinoma with 38 and finally large cell carcinoma with 5 patients. The CT and PET-CT scans are in DICOM format and the clinical data of the patients are also available in XLSX format. The identification of the tumor location was performed by a radiologist using LabelImg [13], a graphical image annotation tool, then a verification was conducted by four other radiologists and the final annotation was saved in XML format. This dataset also provides a python script to visualize the tumor annotations on the CT and PET-CT images, as shown in Figure 2.5.

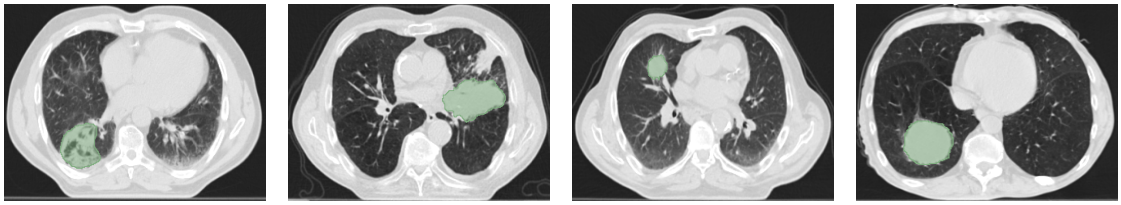


(a) (b) (c) (d)

Figure 2.5: Samples of adenocarcinoma (2.5a), small cell carcinoma (2.5b), large cell carcinoma (2.5c), and squamous cell carcinoma (2.5d), with corresponding bounding boxes, from Lung-PET-CT-Dx dataset[12].

2.1.6 NSCLC-Radiomics

NSCLC-Radiomics dataset [14], also known as Lung1 dataset, is composed of 52,073 samples corresponding to CT scans of 422 NSCLC patients in DICOM format. A radiation oncologist performed the 3D volume segmentations of the GTV, which are also available in DICOM format. This dataset contains 51 patients with adenocarcinoma, 152 with squamous cell carcinoma, 114 with large cell carcinoma, 63 with not otherwise specified, and 42 patients with no histology information provided. This information is provided in a CSV file, along with additional clinical data of the patients such as age, gender, tumor stage, survival time, and dead status. Figure 2.6 presents a sample of each class with respective tumor segmentation.



(a) (b) (c) (d)

Figure 2.6: Samples of adenocarcinoma (2.6a), large cell carcinoma (2.6b), squamous cell carcinoma (2.6c), and not otherwise specified (2.6d), with corresponding tumor segmentation from NSCLC Radiomics dataset [14].

2.1.7 Datasets Overview

The classification of adenocarcinoma through CT images will be the principal objective of the current study, therefore the datasets will be evaluated in terms of their usefulness, outlining some benefits and drawbacks of each. Starting with the Chest-CT dataset, which is composed of 1,000 CT images corresponding to three subtypes of lung cancer and normal cells. This dataset has some disadvantages, including the absence of information about the tumor location, the images not being in DICOM format, which is the most common format for medical image data, and the fact that the CT volume is not available, which could be a constrain if the spacial information between slices is to be taken into account.
LUAD-CT-Survival and LungCT-Diagnosis are considerably small datasets with DICOM samples. For the first, the tumor segmentation through the CT volume is provided while for the second only the tumor location coordinates and the representative slices are given. Since these two datasets consist only of adenocarcinoma samples, they could be used to train models, but preferably combined with other datasets given their small scale. LUAD-CT-Volume is also a dataset composed of adenocarcinoma samples, as CT volume patches and, since the samples are in npy format, it is suitable to be used in python with the *numpy* module, as suggested in the dataset description.
Lung-PET-CT-Dx and NSCLC-Radiomics are two large datasets with samples in DICOM format belonging to four different classes (with adenocarcinoma, squamous cell carcinoma, and large cell carcinoma in common), and with tumor segmentation available. A possible disadvantage of these two datasets is their imbalanced distribution of classes, being adenocarcinoma the most represented one in Lung-PET-CT-Dx and the least represented in NSCLC-Radiomics.
Table 2.1 summarizes the principal characteristics of each dataset including the images files format, tumor segmentation file format, represented classes, and amount of data provided.

Dataset	Format	Tumor Segmentation	Classes	Count	No. of Patients	No. of Samples
Chest-CT [3]	JPEG/PNG	-	ADC SCC LCC Normal Cell	338 260 187 215	1,000	1,000
LUAD-CT-Survival [9]	DICOM	NIfTI	ADC	40	40	3,142
LungCT-Diagnosis [10]	DICOM	-	ADC	61	61	4,682
LUAD-CT-Volume [11]	npy	-	AAH AIS MIA	21 444 585	1,050	-
Lung-PET-CT-Dx [12]	DICOM	XML	ADC SCC SCLC LCC	251 61 38 5	355	251,135
NSCLC-Radiomics [14]	DICOM	DICOM	ADC SCC LCC NOS	51 152 114 63	422	52,073

Table 2.1: Overview of datasets.

2.2 Previous Research Works on Lung Cancer Classification

Previous research works concerning the use of machine learning for mainly lung cancer histology classification through CT images will be addressed. These research works will be thoroughly described in terms of classification task and corresponding classes, dataset used, image pre-processing techniques adopted, feature extraction and selection approaches, classification algorithms employed, and finally results obtained.

2.2.1 Lung cancer histology classification from CT images based on radiomics and deep learning models

Marentakis et al. [7] used a 102 patients subset of the NSCLC Radiomics dataset, with the focus of classifying CT images into two subtypes of NSCLC: adenocarcinoma and squamous cell carcinoma. For this purpose, three major approaches were tested: the first based on radiomic features, the second based on features extracted with convolutional neural networks (CNN), and the third using long short-term memory (LSTM) and CNN.

For the radiomic features approach, four types of features were extracted: first-order statistics, shape, texture, and wavelet features, resulting in 529 features. Since the dataset was short in comparison to the total amount of features extracted, the 40 most important features were selected. Then, the support vector machine (SVM) and k-nearest neighbors (KNN) algorithms were used

to classify the extracted features into adenocarcinoma or squamous cell carcinoma (see Figure 2.7). The accuracy obtained with the SVM classifier was 58% while KNN led to a better result, achieving 67%.

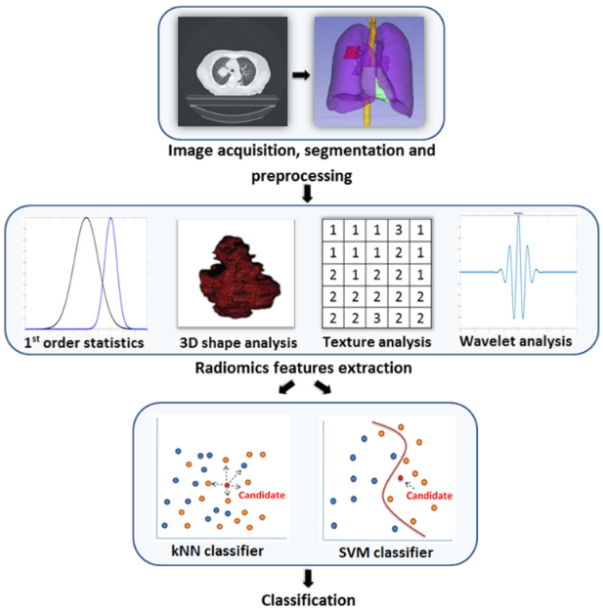


Figure 2.7: Radiomic features approach [7].

In the second approach, based on deep features, four pre-trained models were used: AlexNet, ResNet-101, Inception-v3, and InceptionResNet-v2. The features were extracted from each CT slice through a before-final-fc (BFFC) layer and then two sub-approaches were considered for the classification: the first using the KNN classifier and the second applying fine-tuning to the pre-trained CNN models. In the first sub-approach, max-pooling and mean-pooling were employed to concatenate the features of each CT slice, resulting in a single feature vector per patient. Then the feature vectors were fed to the KNN classifier, as illustrated in figure 2.8. The CNN architecture that achieved better results was the Inception-v3, with an accuracy of 61%.

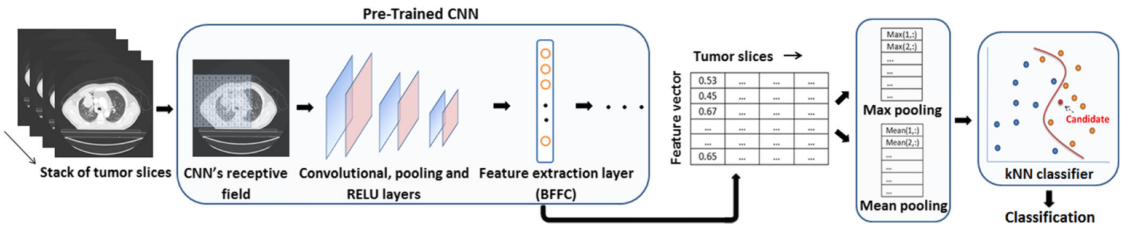


Figure 2.8: Pre-trained CNN features approach [7].

For the second sub-approach, the last layer was modified to have 2 nodes, each corresponding to the probability of each class. Then fine-tuning of the last two fully connected (FC) layers was performed with Adam optimizer. Finally, majority voting and average confidence were calculated for each CT slice to make the prediction, as shown in the right bottom part of Figure 2.9. This sub-approach provided better results, with Inception-v3 increasing the accuracy up to 70%.

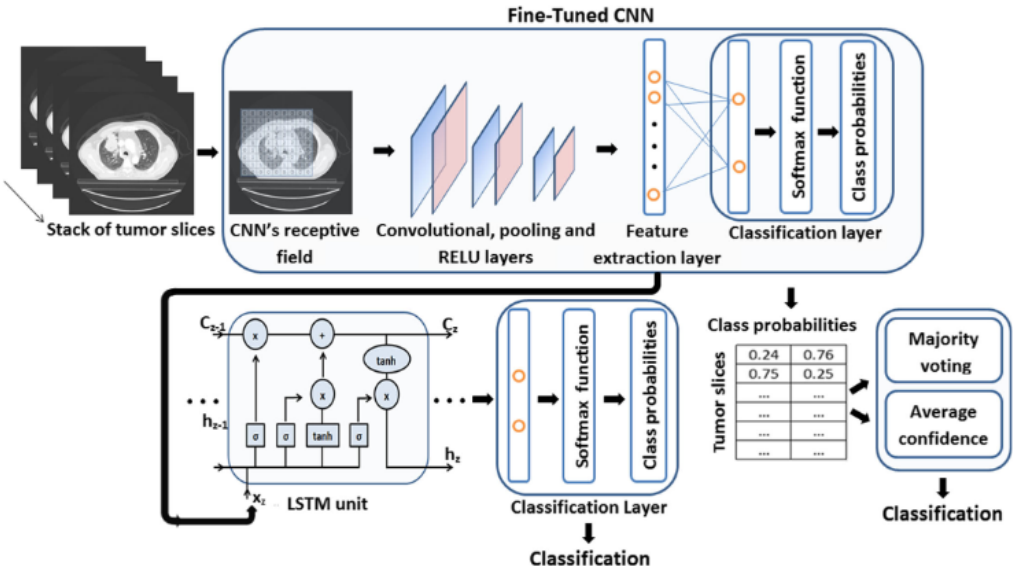


Figure 2.9: CNN features + LSTM approach (left bottom) and fine-tuned CNN approach (right bottom) [7].

For the third approach, the fine-tuned CNN model with the best performance (Inception-v3) was chosen for feature extraction and an LSTM network with a final softmax FC layer was used for classification. The LSTM network allows the classification to be done taking into account the spatial coherency of the features extracted from the CT slices. The left bottom part of Figure 2.9 presents this approach.

Then, a final sub-approach was employed, in which the classification of the CNN + LSTM model was performed by the KNN and SVM classifiers.
To evaluate the performance of each approach, five metrics were considered: accuracy, sensibility, specificity, precision, and F1 score. LSTM + Inception-v3 was the best-performing model across all five metrics, with 74%, 67%, 81%, 73%, and 76%, respectively. Other conclusions from this evaluation are that KNN had a better overall performance than SVM and that the classification of adenocarcinoma can be achieved with higher efficiency than the classification of squamous cell carcinoma for almost all models.

2.2.2 An Approach Toward Automatic Classification of Tumor Histopathology of Non-Small Cell Lung Cancer Based on Radiomic Features

Patil et al. [15] performed the classification of CT images into four classes: adenocarcinoma, squamous cell carcinoma, large cell carcinoma and not otherwise specified. The chosen dataset, which was a subset of the NSCLC Radiomics, included 317 patients, with the following distribution: 40 samples of adenocarcinoma, 110 of squamous cell carcinoma, 108 of large cell carcinoma, and 59 of not otherwise specified. Given the imbalanced proportion between classes, the Synthetic Minority Oversampling Technique (SMOTE) [16] algorithm was applied to the dataset.

The main focus of this research was the comparison between the use of non-radiomic and radiomic features and the impact it has on the performance of the classifiers. The radiomic approach resulted in 431 features extracted per sample, corresponding to four types of features: first-order statistical, shape and size-based, textural, and multiscale wavelet. The first-order statistical features provide information about the gray-level distribution of the tumor while the shape and size-based features specify tumor characteristics such as volume, compactness, and area. The textural features were computed with the gray-level co-occurrence matrix (GLCM) and the gray-level run-length matrix (GLRLM), and finally, the multiscale wavelet allows the data separation into high and low frequency components. The nonradiomic features, on the other hand, were a subset of the above described radiomic features, including information such as volume, diameter, compactness, variance, energy, and entropy.

Then, the extracted features were given as inputs to the SVM classifier, optimized with 10-fold cross-validation and grid search. To evaluate the performance of both approaches, three metrics were employed: accuracy, sensitivity, and specificity. The radiomic features approach led to better results across all metrics, with 88% accuracy, 87% sensitivity, and 89% specificity, whereas the nonradiomic features approach achieved 67%, 70%, and 65%, respectively. Finally, the authors conducted an analysis of feature importance and concluded that the wavelet features were the ones that most contributed to the classification.

2.2.3 Classification of Lung Nodules based on Transfer Learning with K-Nearest Neighbor (KNN)

Saikia et al. [8] proposed a model in which four pre-trained CNN architectures were modified to execute the feature extraction and the KNN algorithm was selected to perform the classification of CT scans into four classes: small cell carcinoma, adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. The CT images belong to a subset of 4,105 images from the Lung-PET-CT-Dx dataset.

Four pre-processing techniques, including resizing, normalization, labeling, and data augmentation, were first applied to the CT images: the images were resized to 256x256 and normalized, then binary labeling was applied by defining a value for each class, and finally, data augmentation was performed with rescale, shear range, zoom range, and horizontal flip.

VGG-16, VGG-19, Inception-v3, and MobileNet-v2 were the CNN architectures chosen to perform the feature extraction. A transfer learning approach was adopted, being the selected architectures pre-trained with the ImageNet dataset. Adam and RMSProp optimizers were used to improve the models' classification accuracy, with RMSProp providing the highest value of accuracy. First, classification was performed directly by the pre-trained Inception-v3 and MobileNet-v2 architectures, achieving accuracy values of 72.33% and 74.48%, accordingly. Then, the four CNN architectures were modified to extract the features, that would then be utilized by KNN to perform the classification into the four classes mentioned above.

The evaluation metrics considered were accuracy, precision, recall, and F1 score. The model that

led to the best performance, across the four metrics, was the VGG-16 architecture with the KNN classification, achieving 87.66% in accuracy, 93.46% in precision, 87.66% in recall, and an F1 score of 90.42%.

2.2.4 Deep Features Based Automated Multimodel System for Classification of Non-Small Cell Lung Cancer

Pandey et al. [17] carried out the classification of CT images into adenocarcinoma and squamous cell carcinoma based on deep features. The CT images belong to a subset with 525 samples of the Chest-CT dataset, including only the ones corresponding to the mentioned histological subtypes. The proposed workflow (Figure 2.10) begins with a data pre-processing step, followed by feature extraction with five CNN architectures: InceptionResNet-v2, Inception-v3, Xception, VGG16, and VGG19, and finally the classification with three algorithms: SVM, XGBoost, and a Fully Connected Neural Network (FCNN).

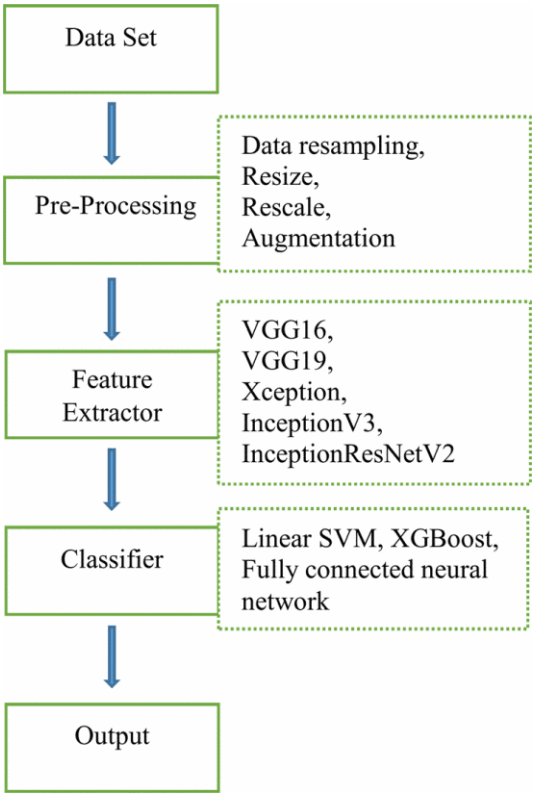


Figure 2.10: Pipeline with SVM, XGBoost, and FCNN for classification of deep features [17].

Two approaches were proposed, resulting in seven models obtained by different combinations of deep feature extractors and classifiers. Since the dataset is relatively short and class representation is not proportional (338 samples of adenocarcinoma and 187 of squamous cell carcinoma), data augmentation and resampling techniques were employed. The resampling of the dataset produced 628 samples, of which 400 constituted the training set. Then, data augmentation methods, such as horizontal and vertical flip, 10-degree rotation, width and height shift, and image shear,

were applied to the training set.
In the first approach, two models were considered: InceptionResNet-v2 architecture was selected for feature extraction in both, while the classification was performed by FCNN and SVM. After extracting the features and in order to reduce the risk of overfitting, the following layers were added: a global average pooling layer, followed by an fully connected layer with 512 neurons, a batch normalization, and a dropout layer of 30%. Then the final layer was defined according to the classifier, either SVM or an FC layer with the sigmoid as activation function. These two models were optimized with the Adam algorithm.
In the second approach, all five CNN architectures were combined with the XGBoost classifier. The data augmentation procedure was not employed and, to improve the models' performance, hyperparameter tuning was done with 10-fold cross-validation and random search.

From the results, the authors concluded that the best-performing model was the combination of InceptionResNet-v2 and SVM, achieving an accuracy of 93.50% on the training set and 93.16% on the validation set. The second best model was the integration of VGG19 and XGBoost, from the second approach, with training and validation accuracy of 92.25% and 91.05%, respectively. It is worth mentioning that although the best training accuracy, 97.25%, was obtained with the model composed of InceptionResNet-v2 and FCNN, it was not considered the best-performing model since its validation accuracy was 85.26%, which is a significant overfitting case.

2.2.5 Lung Cancer Detection and Classification from Chest-CT Scans Using Machine Learning Techniques

Rehman et al. [18] used the Chest-CT dataset to perform the CT images classification into four classes: adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and non-cancerous. SVM and KNN were chosen for the classification task, which was performed considering the texture features selected and extracted with the Local Binary Pattern (LBP) and the Discrete Cosine Transform (DCT) techniques, respectively.

Firstly the CT scans were converted into grayscale, since the LBP can only be applied to images in the referred color mode. The LBP was used to encode the images by assigning a binary value to the central pixel based on the similarity with its three-patch neighborhood pixels. Then, the feature vector was created by normalizing and concatenating the histograms of each region division. The DCT coefficients were computed from the features vector, converting them into frequencies and magnitudes. The high frequencies features were discarded while the low frequency ones were extracted using the zigzag method. Finally, the resulting features were provided as inputs to the SVM and KNN classifiers.

For the performance evaluation, six metrics were calculated: accuracy, sensitivity, specificity, precision, recall, and F1-Score. The SVM produced better results than KNN across all metrics, being worth noting that both classifiers achieved high accuracy values: 93% for SVM and 91% for KNN. Concerning adenocarcinoma, SVM also performed better than KNN, with detection rates of 91.66% and 87%, respectively.

2.2.6 Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples

Ubaldi et al. [19] focused on histology and tumor stage classification as well as exploring the viability of combining datasets to train and test models, especially using large public datasets to train and apply to small and private datasets for testing. The public dataset consists of a subset of 130 patients from NSCLC-Radiomics, whereas the private dataset, identified as L-RT, contains CT images from 47 individuals. Both datasets include clinical data, such as tumor histology and stage, along with Gross Tumor Volume (GTV) segmentation performed by radiologists. Adenocarcinoma, large cell carcinoma, and squamous cell carcinoma were the three classes taken into account for the histology classification, while stages I and II were considered for the stage classification, since they are represented in both datasets.

Radiomic features were extracted using the PyRadiomics [20] plugin of the 3D Slicer software, generating 107 features. Then, the Principal Component Analysis (PCA) algorithm was used to define the components to be retained by hyperparameter optimization and 10% of the features with the highest Mutual Information value were selected. Finally, the classification was carried out by the AdaBoost, KNN, SVM, and Random Forest (RF) classifiers, with hyperparameters optimization by cross-validation.

The Mann-Whitney U test was applied to each feature to determine if the two datasets could be merged and the results revealed that the datasets could be combined for 74 features. Then, two classification approaches were evaluated: the first applied to the Lung1, L-RT, and merged datasets individually, while the second used different datasets for training and testing. Regarding the histology classification, the best performance was obtained with the RF classifier for the merged dataset. However, for the tumor stage classification, the best results were provided by the linear SVM classifier when the Lung1 dataset is used for training and L-RT for testing. Based on these results, the authors concluded that combining small, private datasets with larger, publicly available ones improve the classification reliability of tumor histology and stage using radiomic characteristics.

2.2.7 Minimum Sample Size Estimate for Classifying Invasive Lung Adenocarcinoma

Ma and Yue [21], conducted a research based on the Statistical Learning Theory to determine the appropriate number of samples to perform the classification of invasive adenocarcinoma (IA) and non-invasive adenocarcinoma. Two different perspectives were explored: the first was to assess the impact of dataset size on testing error, and the second was to establish the minimal dataset size required to accomplish specific performance requirements. A private dataset was used for the classification of IA and non-IA, manifested as ground glass nodules (GGNs) on CT scans. For feature extraction, the CNN approach was chosen, applying transfer learning to the AlexNet model with fine-tuning of the last FC layer, then for feature selection, the ANOVA method was applied

to select the ten features with the lowest p-value and finally, the classification was performed by the SVM classifier. Figure 2.11 outlines the adopted procedure.

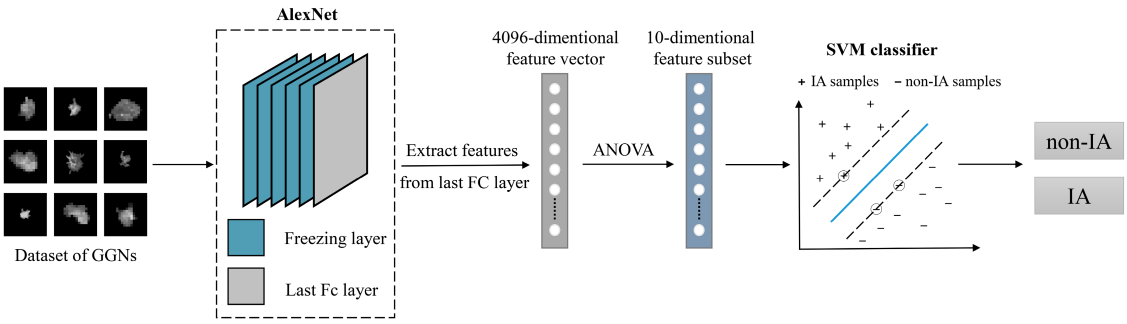


Figure 2.11: Sequence for classifying the features extracted by AlexNet with SVM [21].

The Statistical Learning Theory establishes a relationship between training error (learning error) and testing error (generalization error), which is influenced by factors such as the confidence range, the number of training samples, and the Vapnik–Chervonenkis (VC) dimension, which measures the capacity of a classifier to learn and depends on the number of features and classes.
Four training subsets were created, with 100, 200, 300, and 400 samples respectively, to investigate the influence of the training set size on the classifier's generalization ability and, following training, the model was applied to a fixed test set of 36 samples. The accuracies obtained in the test set were 63.9%, 77.8%, 80.6%, and 86.1%, when the training set size was respectively 100, 200, 300, and 400 samples. From this experiment, the authors concluded that the generalization ability of the model is proportional to the dataset size and that the classifier's VC dimension should be as small as possible.
In the second approach, to determine the number of samples necessary to achieve a given requirement of accuracy, the individual influence (modifying one value at a time) of the number of classes and features selected on the generalization error was explored, and it was concluded that the minimal number of samples required grows with the increase in classes and features. Finally, three binary classification tasks developed by other authors were analyzed and, in the three works, the dataset sizes were greater than the predicted ones, which supports the theoretical results for the minimal limit of samples.

2.2.8 Deep learning for predicting subtype classification and survival of lung adenocarcinoma on computed tomography

Wang et al. [22] compared the use of radiomics, deep learning, and a combination of both for the classification of CT scans into lung adenocarcinoma subtypes and survival. In this research, a private dataset of 1,222 patients from three different institutions was used. Of them, 998 patients' CT scans were provided by the West China Hospital, while the remaining belonged to other two independent medical institutions. This dataset includes samples from four major classes: AAH, AIS, MIA, and IA, being IA divided into four subclasses predominant lepidic adenocarcinoma

(pLAC), predominant acinar adenocarcinoma (pAAC), predominant papillary adenocarcinoma (pPAC), predominant micropapillary adenocarcinoma (pMAC), and predominant solid adenocarcinoma (pSAC). The segmentation of the tumor region of interest (ROI) was performed automatically by a deep residual network, and further assessed by two senior radiologists.

Three approaches were considered for the histology classification task: 2-category classification, 3-category, and 8-category. For the 8-category, all aforementioned classes were considered, then in the 3-category, the regarded classes were non-MAC/SAC, non-pMAC/pSAC, and pMAC+pSAC, and finally, for the 2-category approach, the images were classified into IA and non-IA. The training set was composed of 998 patients from the West China hospital, and to evaluate the performance of the models, two testing sets were considered: an internal test set, with samples from the same medical institution as the training set, and an external test set, with samples from the other two mentioned institutions.

For the deep learning model, a pre-trained 3D ResNet34 architecture from MedicalNet was chosen and some adjustments were employed. First, a global average pooling (GAP) layer was added after the last convolutional layer to increase the generalization ability. Then, in addition to the 512 features from the 3D cropped image taken at the GAP layer, 11 features concerning characteristics such as tumor transparency and tumor proportional size, were defined as inputs to the first FC layer. A dropout of 0.5 was adopted at the last two FC layers to minimize the risk of overfitting. The structure of the deep learning model is represented in Figure 2.12. For the training phase, data augmentation techniques (such as image size and gray value scaling, transposing, rotation, and translation) were applied and the stochastic gradient descent optimizer was utilized. For this model, the best accuracy was obtained with the 2-category classification, achieving a mean value of 85.75% (86.22% on the internal test set and 85.27% on the external).

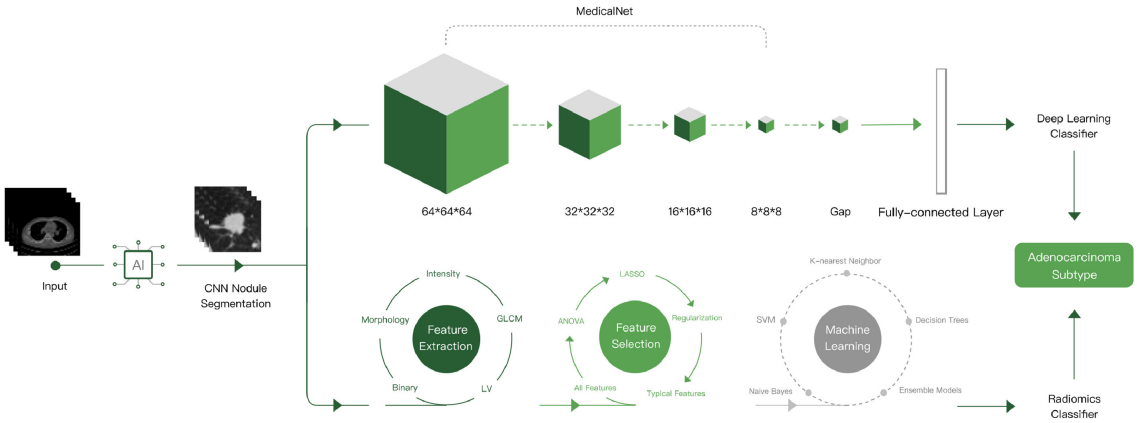


Figure 2.12: Deep learning model (upper half) and radiomics model (bottom half) [22].

In the radiomics model, the feature selection was performed based on the similarity between features. Pearson's correlation coefficient was calculated for the feature pairs and if the value of this coefficient was above 0.9, one feature of the pair would be removed. The classification was performed by a multinomial logistic regression with LASSO regularization and a 5-fold cross

validation was conducted. Figure 2.12 illustrates the workflow of the radiomics model. Radiomics provided better results than the deep learning model, with a mean accuracy of 89% (88.27% and 89.73% for internal and external test sets, respectively) on the 2-category approach.

For the combinatorial deep radiomics model, three groups of features were initially considered: a subset of the 512 features extracted in the deep learning model, selected via the multinomial logistic regression, the radiomic features, and the patients' clinical data. Then, the Akaike information criterion was used for feature selection, being selected the combination of features that led to the lowest value. Finally, classification was performed once again with the multinomial logistic regression algorithm. Figure 2.13 describes the steps followed in the deep radiomics model. The deep radiomics model was the one that led to a slightly better performance than radiomics, with a mean accuracy of 89.2% (87.76% on the internal set and 90.63% on the external).

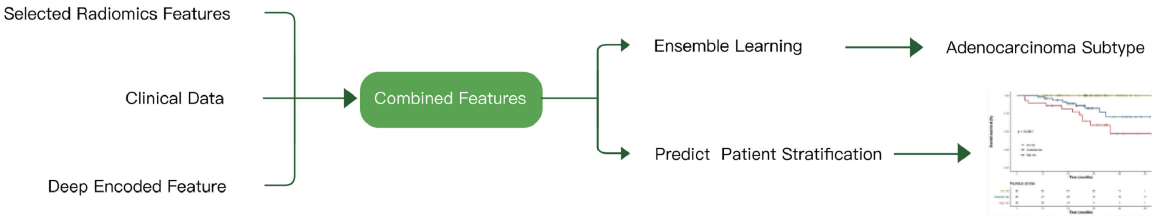


Figure 2.13: Deep radiomics model [22].

From the results, the authors concluded that the increase in the number of classes negatively impacts the performance of the models, since the accuracy drastically drops in the 8-category classification. Additionally, in practically all approaches, the difference between the performance results obtained with the internal and external testing sets was not significant.

2.2.9 Approaches and Results Overview

In this section, a comparison between the models described and corresponding performances will be conducted.

Regarding the type of feature utilized, one of the following two approaches was adopted in the majority of the cases: radiomic feature and deep features, extracted by CNNs. [7] and [22] directly compared the use of radiomic features and deep features. The results from [22] showed that radiomics features led to slightly better overall performance, while in [7] the radiomic features + KNN combination also achieved better results, except for the cases in which the fine-tuned Inception-v3 architecture was selected. Additionally, the combination of radiomics and deep learning was assessed in both studies, but in neither one did it result in appreciable improvements.

The radiomic features approach of [7] was based on [15], but the performance obtained in [7] was significantly inferior, with lower values in terms of accuracy, specificity, and sensitivity (58%, 53%, and 63% for [7] and 88%, 89%, and 87% for [15], respectively). Both proposed models used the NSCLC-Radiomics dataset, SVM (with equal hyperparameters) as the classifier, and are very similar in terms of the type of features extracted. The most notable differences between the

models are that [7] performed two-class histology classification and selected only the 40 most important extracted features, while [15] considered four classes and all 431 features extracted. To perform the two-class classification, [7] used a subset of NSCLC-Radiomics containing 94% of the adenocarcinoma samples from the original dataset and only 36% of the squamous cell carcinoma samples, corresponding to a significant reduction in this last class, which could have led to a loss of important samples, impairing the generalization capacity of the model, and so could be a reason for the model's poor performance when compared to [15].

Considerably good performances were also achieved using the deep features approach [8, 17, 21, 22]. [8] compared the use of pre-trained CNN for both feature extraction and classification with the combination of CNNs as feature extractors and KNN as the classifier, and concluded that the former led to higher accuracy values, with the best performance being achieved by VGG-16 + KNN model. [21] focused on evaluating the influence of the training set size on the model's performance, and showed that the performance can be improved by increasing the number of samples, obtaining an accuracy of 86.1% on the test set for the model trained with 400 samples. [22] studied the impact of the increase in the number of classes on the models' performance, and confirmed that it reduces the classification accuracy, being 86.75% for the two-class deep features approach and 42.61% for eight-class. It is worth noting that [22] selected a pre-trained 3D ResNet-34 architecture from MedicalNet, while the previously referred authors [7, 8, 17, 21] used CNNs pre-trained from ImageNet. Finally, [17] proposed the InceptionResNet-v2 + SVM model for the Chest-CT dataset, which was the one that led to the best performance of all the deep feature models. [18] proposed a model that was neither based on radiomic nor deep features but instead the features were obtained through a three-patch LBP and the DCT, achieving two of the highest accuracy values across all models, with 93% for the SVM classifier and 91% for KNN.

Table 2.2 summarizes the approaches and models suggested by the authors in previous research works, as well as the accuracy values obtained.

Ref.	Dataset	Classes	Feature Extraction	CNN Architecture	Classification	Accuracy
[7]	NSCLC Radiomics	ADC, SCC	Radiomic	-	KNN SVM	0.67 0.58
			CNN (pre-trained)	AlexNet ResNet-101	KNN	0.61
			CNN (pre-trained)	Inception-v3 InceptionResNet-v2	Inception-v3	0.70
			CNN (pre-trained)	Inception-v3	LSTM LSTM + KNN LSTM + SVM	**0.74** 0.69 0.69
[15]	NSCLC Radiomics	ADC, SCC, LCC, NOS	Radiomic Non-Radiomic	-	SVM	**0.88** 0.67
[8]	Lung-PET-CT-Dx	ADC, SCC, LCC, SCLC	CNN (pre-trained)	Inception-v3 MobileNet-v2	Inception-v3 MobileNet-v2	0.72 0.75
			CNN	VGG-16 VGG-19 Inception-v3 MobileNet-v2	KNN	**0.88** 0.87 0.74 0.82
[17]	Chest-CT	ADC, SCC	CNN	InceptionResNet-v2	FCNN SVM	0.85 **0.93**
				VGG-16 VGG-19 Xception Inception-v3 InceptionResNet-v2	XGBoost	0.88 0.91 0.87 0.80 0.81
[18]	Chest-CT	ADC, SCC, LCC, Normal	LBP + DCT	-	SVM KNN	**0.93** 0.91
[21]	Private	IA, non-IA	CNN	AlexNet	SVM	**0.86**
[22]	Private	IA, non-IA	CNN (pre-trained) Radiomics CNN + Radiomics	ResNet-34 - ResNet-34	ResNet-34 Logistic Regression	0.86 0.89 **0.89**

Table 2.2: Overview of previous researches approaches and corresponding accuracy. The top results achieved in each study are marked in bold.

Chapter 3

Methodology

The methodology followed is thoroughly described in this chapter and it was developed based on the frameworks and models from previous related works (Section 2.2), which were modified and adapted accordingly to suit the problem at hand. The first step was the selection of the dataset, followed by its filtering, and after that image preprocessing techniques were applied to the resulting data. Then, the feature extraction and classification were based on deep learning, where two main approaches were employed by using two-dimensional and three-dimensional CNNs. Figure 3.1 illustrates the main steps followed for the prediction of lung adenocarcinoma based on CT images.

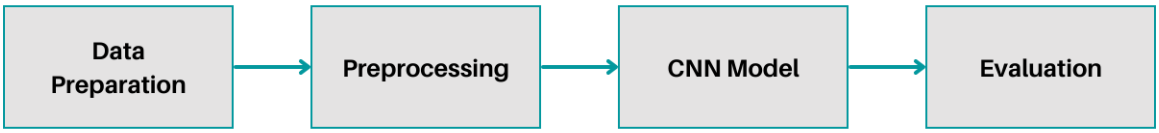


Figure 3.1: Generic pipeline describing the constituting steps of the proposed methodology.

3.1 Dataset Preparation

3.1.1 Dataset Selection

Considering the datasets listed in Section 2.1, priority was given to the ones with DICOM data, since it is one of the most common storage formats in medical imaging and includes additional metadata information about the examinations; then, datasets that did not provide the tumor annotations or that did not include multiple subtypes of lung cancer were excluded. From the described datasets, the Lung-PET-CT-Dx and NSCLC-Radiomics were selected after the filtering.
Classification tasks with fewer classes can achieve higher accuracy values [21, 22] and since the main focus of the current work was the identification of lung adenocarcinoma, the datasets were grouped into two classes: adenocarcinoma and non-adenocarcinoma. After re-labelling, the distribution of the adenocarcinoma class was significantly imbalanced in both datasets, being the most

represented one on the Lung-PET-CT-Dx (with a proportion of 71:29) and the least represented on the NSCLC-Radiomics (with a proportion of 12:82).
Given that both datasets were in DICOM format and that, after the binary re-labeling, both represented equivalent classes, the two datasets were merged in order to increase the amount of data available and to minimize the imbalanced class distribution of the individual datasets. The resulting combined dataset included 777 patients with a class proportion of 39:61 (302 adenocarcinoma cases and 475 non-adenocarcinoma cases).

3.1.2 Data Filtering

Having the datasets selected, the next step was performing data cleaning, where some cases were removed based on the analysis of the metadata and the nodule segmentations available. This process was conducted independently for each dataset, as each one possessed distinct characteristics necessitating the definition of specific criteria.

3.1.2.1 NSCLC-Radiomics

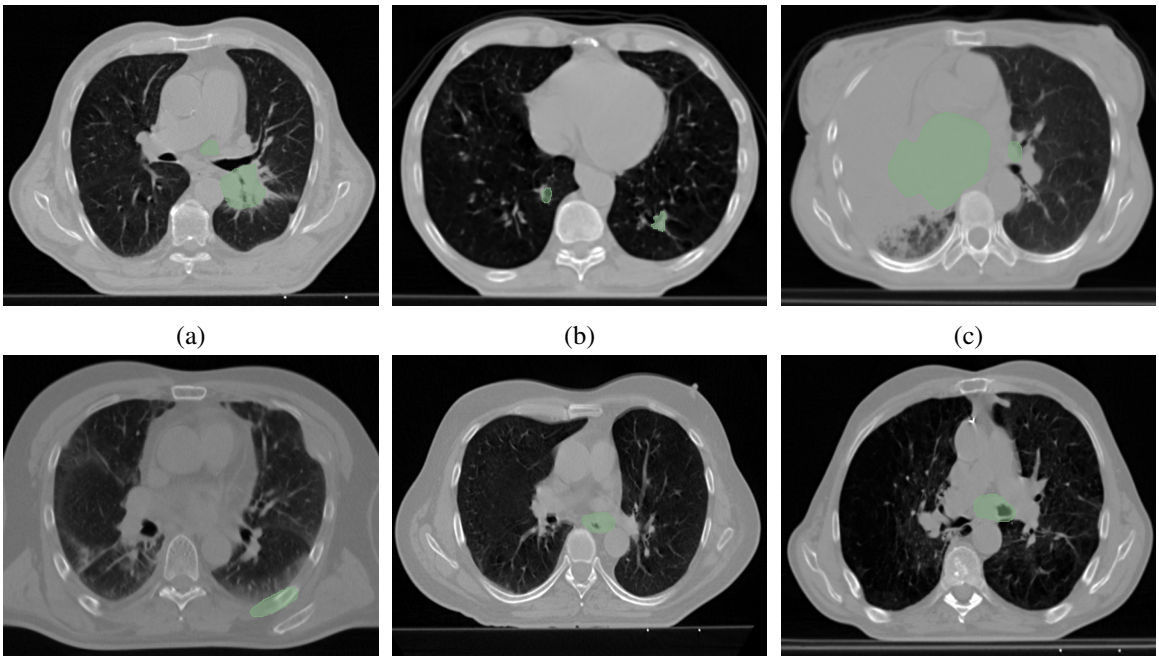
For the NSCLC-Radiomics dataset the process of case elimination was divided into two phases: initially, cases were filtered based on the available metadata, followed by the assessment of the CT scans and their corresponding segmentation masks.
Firstly, based on the metadata, patients with no histology information or with no segmentations available were removed. Cases where the *ImageType* DICOM attribute did not include the "ORIGINAL" flag were also discarded.
In the second phase, based on the segmentation masks, a sequence of criteria was define with the collaboration of medical professionals to uphold the integrity of the process. The list of criteria established for the exclusion of cases was as follows:

- Criterion 1: cases in which the GTV segmentation is not continuous (e.g. presence of multiple or divided nodules).

- Criterion 2: cases in which the GTV represents the segmentation of anatomical structures other than lung.

- Criterion 3: cases in which the nodule is located outside the lung parenchyma.

These cases were removed to ensure the reliability of the dataset, as well as to mitigate the introduction of irrelevant or confounding information, which could potentially have a negative impact on the performance of the models. Samples of excluded cases corresponding to each criterion are demonstrated in Figure 3.2.
Appendix A.1 illustrates the sequential steps performed for this selection, along with the number of patients removed in each one.



(a) (b) (c)

(d) (e) (f)

Figure 3.2: CT images and corresponding segmentation masks from the NSCLC-Radiomics dataset illustrating examples from cases that have been removed based on Criterion 1 (3.2a, 3.2b, 3.2c), Criterion 2 (3.2d), and Criterion 3 (3.2e, 3.2f).

3.1.2.2 Lung-PET-CT-Dx

The Lung-PET-CT-Dx dataset, consisted of a total of 355 patients with multiple series per study, unlike the NSCLC-Radiomics dataset, which included only one version of each examination. Consequently, a sequential set of criteria was implemented to select only one series per study.
The first set of series to be removed were those whose modality was not "CT", followed by the cases where the *ImageType* DICOM attribute, which denotes image characteristics, did not include the flags: "ORIGINAL", "AXIAL", and "PRIMARY", with the first referring to the pixel data, the second to the anatomical plane and the third to the examination characteristics.
Subsequently, multi-frame CT volumes were discarded, along with the series that lacked accessible nodule segmentations and the ones that did not to correspond to the earliest date of each study. The *SliceThickness* and *ConvKernel* DICOM attributes were also considered for the series selection, since the first corresponds to the spacing in millimeters between each consecutive slice of the CT volume and the latter to the convolution kernel used by the scanner for the reconstruction of the CT. For each of these attributes a priority list of values was defined based on the frequency of occurrence of the values in both datasets (Lung-PET-CT-Dx and NSCLC-Radiomics), with the most common values being preferably chosen. The slice thickness is constant throughout the entire NSCLC-Radiomics dataset with a value of 3.0 mm, being this the most common value followed by 1.0 mm and 5.0 mm. Additionally, the defined priority list for the *ConvKernel* attribute was: "B30f", "B31f", "B40f", "B", "LUNG", "STANDARD" and "B70f".

Thereafter, cases that involved contrast-enhanced CT scans were also excluded and finally, for each remaining study with multiple series, the serie with the highest number of slices was maintained.
Similarly to the NSCLC-Radiomics dataset, Appendix A.2 depicts the sequential application of criteria and the corresponding patients and series that were eliminated during the selection process.

3.1.2.3 Resulting Datasets

After the process of data filtering, wherein a careful exclusion of cases was employed for each dataset, a refined set of samples remained for the further stages. Table 3.1 summarizes the number of patients, the samples per class, and the corresponding percentages for each of the three resultant datasets.

Dataset	Patient Count	ADC	%	Non-ADC							
				SCC	%	LCC	%	SCLC	%	NOS	%
Lung-PET-CT-Dx	336	235	70	58	17	5	2	38	11	-	-
NSCLC-Radiomics	344	47	14	137	40	103	30	-	-	57	16
Combined	680	282	41	195	29	108	16	38	6	57	8

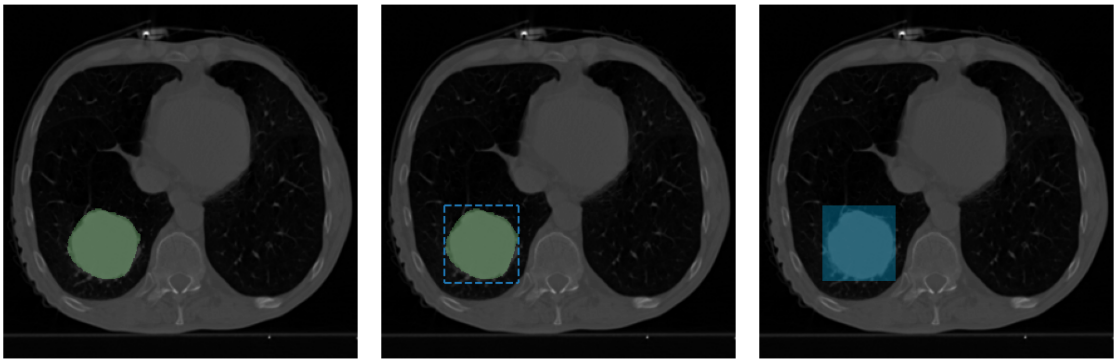
Table 3.1: Distribution per dataset of number of patients and classes, specifically the adenocarcinoma (ADC) class, and the non-ADC class, which encompasses squamous cell carcinoma (SCC), large cell carcinoma (LCC), small cell carcinoma (SCLC) and not otherwise specified (NOS).

3.2 Preprocessing

Once the datasets were stabilized, an image preprocessing stage was employed to improve the quality and relevance of the data and to ensure it complied with the models' requirements. This process was divided into three main steps: segmentation of the CT scans, the application of image processing techniques and finally data augmentation.

3.2.1 Segmentation

The two individual datasets provide the tumor annotations through the CT volume; however, the types of segmentation masks available for each dataset are different: while in NSCLC-Radiomics the masks correspond to the thin contour of the tumor, in the Lung-PET-CT-Dx the segmentations are presented as bounding boxes. The NSCLC-Radiomics dataset masks were converted into bounding boxes for each slice by forming a rectangle around the tumor, with the edges of the original masks serving as the boundaries of the new ones, as shown in Figure 3.3.



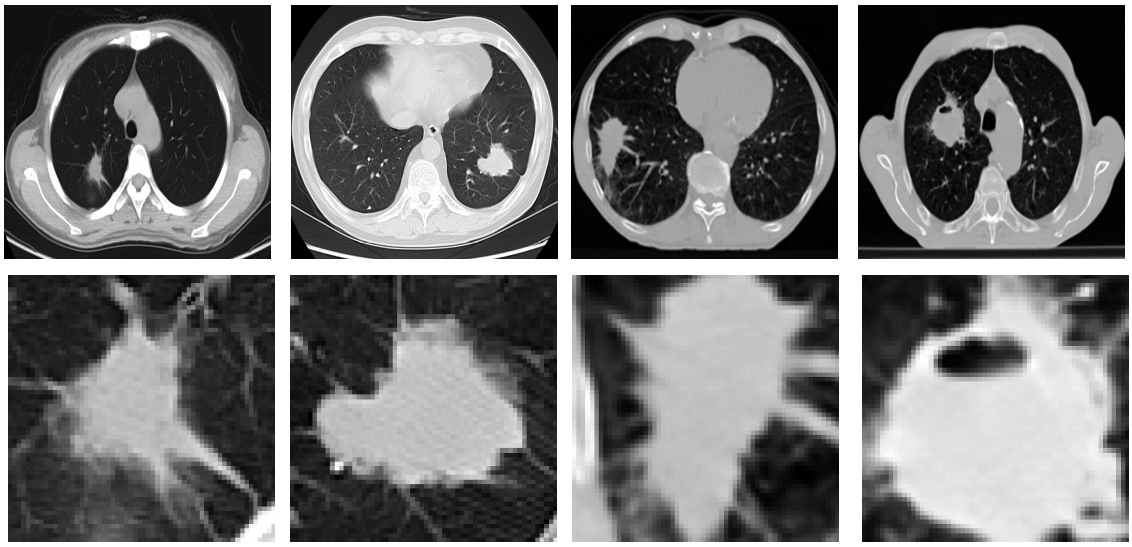
(a) (b) (c)

Figure 3.3: Steps to transform the NSCLC-Radiomics segmentation masks into bounding boxes: starting from the original segmentation mask (3.3a), then defining the boundaries based on the original mask edges (3.3b) and filling to obtain the final bounding box (3.3c).

With the segmentation masks well-defined for all samples, the CT scans were cropped at the region of interest. Since both 2D and 3D approaches were considered, the definition of the cropping boundaries for the CTs was tailored to each approach. In the case of the 2D datasets, the CTs were cropped at the slice level in accordance with each individual segmentation mask. For the 3D datasets the cropping was performed taking into account complete volume: the outermost edges across all slices were used to calculate the maximum length occupied by the tumor and also its center, the slices were then cropped according to the nodule center and the maximum length calculated for each volume. The final volumes were composed by twelve slices, the smallest number of slices in a CT across both datasets, centered on the slice which mask corresponded to the largest area.

3.2.2 Image Normalization

The pixel intensity values of the CT images were converted to Hounsfield Units (HU) using the *RescaleSlope* and *RescaleIntercept* DICOM attributes to ensure that all pixel intensities were on the same scale. These values were clipped to [-1000 HU, +400 HU], which correspond respectively to the air and the hard tissues radiodensity values [23], and then min-max normalization was employed to convert the values into a 0 to 1 scale.
After cropping, the images were of various sizes, so resizing was the next step. To select the value for the new size, and in order to minimize distortion added by the resizing operation, the median value of the distribution of the original size patches was chosen as the target size. The resize was performed by bilinear interpolation and the final dimensions for the 2D and 3D datasets were, respectively, 60×60 and $12 \times 66 \times 66$ pixels, with twelve being the number of slices. A collection of the CT patches obtained after the preprocessing stage is shown in Figure 3.4.



(a) (b) (c) (d)

Figure 3.4: Original (first row) and corresponding preprocessed (second row) CT samples of adenocarcinoma (3.4a) and non-adenocarcinoma (3.4b) from the Lung-PET-CT-Dx dataset, and adenocarcinoma (3.4c) and non-adenocarcinoma (3.4d) from the NSCLC-Radiomics dataset.

3.2.3 Data Augmentation

Each dataset was then divided into three subsets: train, validation and test, with respectively 75%, 15% and 10% of the total number of samples. The subsets are equivalent for each 2D-3D pair of datasets, i.e. the patients that are in the train subset of the 2D Lung-PET-CT-Dx dataset, are the same as those in the 3D Lung-PET-CT-Dx, and the same applies to the validation and test subsets. An effort was made to make the training subsets balanced, but in the case of the individual datasets this was not possible given their extremely imbalanced nature. Therefore, data augmentation techniques, namely horizontal and vertical flip, rotation, shift and shear, were applied to the minority class in each training set. The aforementioned transformations were implemented using the MONAI package [24], whereby the parameters were defined within a predefined range of values as outlined in Table 3.2. Again, the geometric transformations were equivalent for each 2D-3D pair of datasets, so no volumetric transformations were employed, all modifications were applied at the slice level even on the 3D datasets.

Transformations	**Values**
Flip	-
Rotation	[-20, 20]°
Shift	[-15, 15]%
Shear	[-20, 20]%

Table 3.2: Range of values defined for the parameters of the data augmentation transformations.

After ensuring that the class distribution in the training sets was balanced, the geometric data augmentation techniques were used once again, this time on both classes, to increase the number of samples available to 1300 in each training set, and also adding variability to the training sets.
Additionally, for the 2D approach, three new datasets were generated in which, instead of selecting only one slice per patient, the three slices with the largest tumor area (determined by the corresponding segmentation masks) were selected. In this case, geometric data augmentation techniques were applied just in order to balance the class distribution. This procedure was adopted as a means of augmenting the available sample size, while reducing the generation of artificial data. By incorporating multiple slices, the diversity of the datasets was increased, while minimizing the potential risk of modifying relevant structures that could significantly impact the classification task.

3.3 Feature Extraction and Classification

Feature extraction is the process of retrieving specific information from the provided data, in this case from the CT images, generating feature vectors that will serve as inputs to the classifiers. Two widely used types of features for lung cancer classification based on CT images are radiomic features [7, 15, 22] and deep features [7, 8, 17, 21, 22] with the latter being a growing trend in the medical field over recent years [25].
Deep learning, specifically CNNs, was selected as the foundational framework for this study. This choice was motivated by the fact that CNNs have demonstrated promising results and achieved state-of-the-art performance in numerous medical imaging applications. Moreover, these models are capable of capturing relevant and intricate patterns within the images, enabling them to automatically extract high-level features without the necessity of labor-intensive manual feature engineering.
In order to explore the alternatives available, both 2D and 3D approaches were employed. While 2D CNNs process individual 2D image slices independently, 3D CNNs consider the volumetric nature of the CT images by simultaneously examining information across multiple adjacent slices. The implementation details and specificities of each of these approaches will be explained in the subsequent subsections.

3.3.1 3D Approach

One of the major drawbacks of CNNs is the complexity of these models, and therefore the necessity of substantial amounts of data to achieve optimal performances. In the medical field, the availability of public large-scale annotated datasets is scarce due to the process of manual annotation by medical experts being costly and time-consuming, therefore transfer learning is commonly employed as a solution to address this challenge [26]. The majority of the works described in the literature review [7, 8, 17, 21] adopted ImageNet's pre-trained models, while [22] used a 3D ResNet34 pre-trained from MedicalNet.

MedicalNet [27] makes available eight 3D ResNet architectures trained with a large-scale multi-center dataset composed of CT and magnetic resonance imaging (MRI) scans of several organs and pathologies. Given that the weights of these architectures were pre-trained specifically on medical images, it was anticipated that they would be more appropriate for medical imaging features and, consequently, a 3D ResNet10 architecture sourced from MedicalNet was chosen for the 3D experiments. It is important to notice that the Med3D networks were devised and trained for segmentation tasks, leading to the original architecture comprising a ResNet encoder module along with a decoder component, therefore, in order to adapt network to the lung adenocarcinoma classification, slight modifications were employed. First, a GAP layer was entered after the last convolutional layer, as proposed by [22], and then, to accomplish the binary classification task, a fully connected layer with one output neuron and sigmoid activation function was added into the architecture. Figure 3.5 illustrates the pipeline followed for the 3D approach.

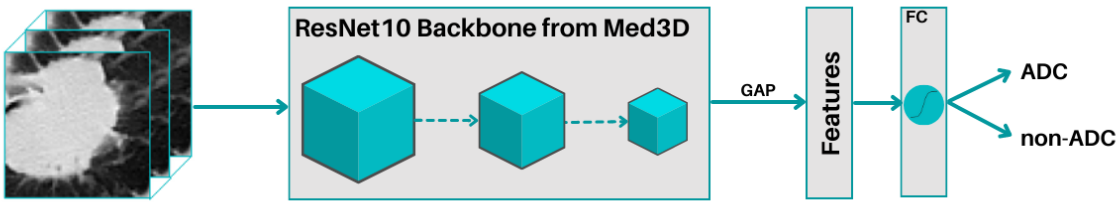


Figure 3.5: Workflow for 3D approach using ResNet10 architecture from MedicalNet [27].

To assess the influence of transfer learning in this particular scenario, the models underwent two phases of training. Initially, the models were trained from scratch (TFS) to establish a baseline performance. Subsequently, with the aim to enhance performance and mitigate the risk of overfitting, a series of fine-tuning experiments was conducted: initially, all convolutional layers were frozen except for the last two, allowing only the FC layer and the aforementioned last two layers to be trained and then the following convolutional layers were progressively unfrozen in a gradual manner.

3.3.2 2D Approach

In order to ensure a equitable comparison with the 3D approach, a ResNet10 model was also employed for the 2D approach. For this purpose, the implementation available in the torchvision package from PyTorch was utilized, which is based on the architecture proposed in [28]. Similar to the adjustments made in the 3D approach, certain modifications were necessary in the implementation. Firstly, the number of channels in the initial convolutional layer was adjusted from three to one, considering that the CT images are represented in grayscale. Additionally, a dropout layer was introduced before the fully connected layer, which consists of a single neuron with a sigmoid activation function, at the final stage of the network.
Transfer learning was not employed in the 2D approach because of two main reasons. Firstly,

the conducted literature review revealed an absence of 2D pre-trained models specifically tailored for the medical domain and, although some researchers utilized ImageNet pre-trained weights in their studies, given the inherent dissimilarities between CT images and ImageNet images along with the limited size of the selected datasets, it was anticipated that using ImageNet pre-trained models would be more demanding than beneficial. Secondly, the 2D models have significantly lower complexity compared to the 3D ones: the 2D-ResNet10 model comprises nearly 5 million trainable parameters, while the 3D-ResNet10 model has approximately 14 million parameters, consequently, the expectation was that training the 2D models from scratch would proceed more smoothly than in the 3D approach.

3.4 Training and Evaluation

As mentioned above, all datasets were divided into train, validation and test subsets, with percentages of 75%, 15% and 10%, respectively.
During the training phase of all experiments, the loss function utilized was the binary cross entropy, along with stochastic gradient descent (SGD) as the optimizer. To optimize the model's performance, numerous combinations of hyperparameters were tuned, including batch size, dropout rate, learning rate, momentum, and weight decay. The range of values explored for each parameter is presented in Table 3.3. During the training phase, the evolution of the train and validation performances was evaluated using the loss function and the Area Under the Receiver Operating Characteristic Curve (AUC). The fine-tuning of hyperparameters was guided by the validation performance, aiming to attain an optimized model configuration. Upon completion of the training phase, the model corresponding to the epoch with the maximum validation AUC was selected for testing.

Parameters	**Models**	
	2D and 3D TFS	3D with TL
Batch Size	{16, 32}	{32}
Dropout	{0, 0.5}	{0, 0.5}
Learning Rate	{0.0001, 0.0005, 0.001}	{0.0005, 0.001}
Momentum	{0.0, 0.8, 0.9}	{0.8, 0.9}
Weight Decay	{0, 0.001, 0.01}	{0, 0.01}

Table 3.3: Experimented values for the training hyperparameters of each model, respectively in the 2D and 3D models trained from scratch (TFS), and in the 3D models in which transfer learning (TL) was employed.

The models' generalization capability holds particular significance in the medical domain, given that the goal is to achieve a robust performance across diverse datasets, due to the inherent heteroginity present in the data originating from multiple medical centers. Therefore, to evaluate the models' ability to generalize, independent testing was conducted on both individual test sets

(Lung-PET-CT-Dx and NSCLC-Radiomics).
In order to ensure a supported assessment of the models, a set of evaluation metrics were calculated on the test sets, specifically the AUC, balanced accuracy, precision, recall, F1-Score, and specificity. Additionally, the computation of confusion matrices was undertaken, providing a visual depiction of the results. A conscientious evaluation is particularly crucial in the field of medical image analysis, where precise and accurate assessment of the models' performance is of utmost importance to assure a reliable diagnosis.

3.5 Methodology Overview

In summary, two primary deep learning approaches were employed: a 2D CNN approach and a 3D CNN approach, both based on the ResNet architecture. For each approach three base datasets were prepared, specifically the NSCLC-Radiomics, the Lung-PET-CT-Dx, and the combined dataset (NSCLC-Radiomics + Lung-PET-CT-Dx). Subsequently, the images were cropped at the tumor level, and common image processing techniques were applied to the resulting patches. Standard data augmentation was then employed, and finally the resultant 2D datasets were fed into the respective CNN models.
Afterwards, based on the best-performing 2D dataset, three more experiments were conducted. Firstly, for the 2D approach, a different type of data augmentation was applied: instead of using a single slice per patient, a multi-slice version of the selected dataset was created, consisting of three slices per patient.
Secondly, the corresponding 3D version of the dataset that yielded better results in the 2D approach, was used to train the 3D model from scratch. Finally, transfer learning from the MedicalNet model was employed, starting by freezing all layers except for the FC layer and the last two convolutional layers, and then a gradually unfreezing two layers at a time. Figure 3.6 provides an overview of the entire sequence of steps followed in the pipeline.

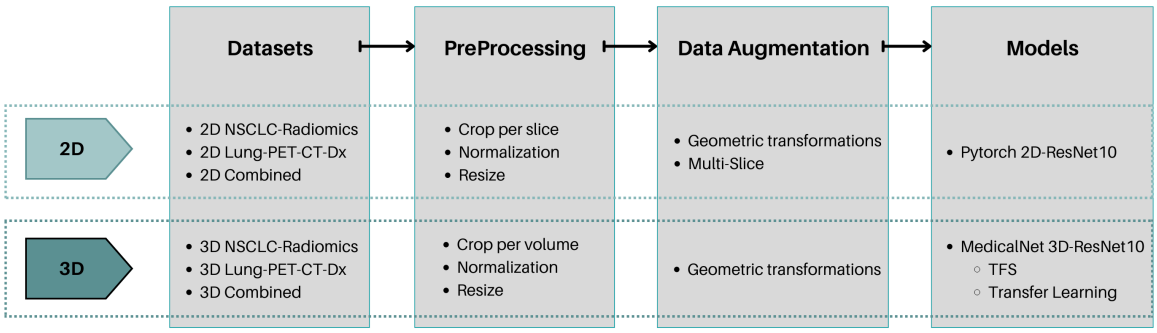


Figure 3.6: Methodology overview with the sequential steps employed in the 2D and 3D approaches.

In total, four comparative experiments were conducted, with different objectives:

- **Study 1**: comparison of the 2D models' performances when trained with three different datasets (Lung-PET-CT-Dx, NSCLC-Radiomics and combined).
- **Study 2**: evaluation of the impact that using a 2D multi-slice dataset for training.
- **Study 3**: comparison between training the 3D models from scratch and using transfer learning from MedicalNet.
- **Study 4**: comparison between the 2D and 3D best-performing models.

In studies 2, 3 and 4, the training datasets were selected based on the optimal performance achieved in study 1, which served as standard for comparing the resulting outcomes. Finally, the performance of the models was assessed by examining the outcomes obtained from the test sets, employing conventional classification evaluation metrics.

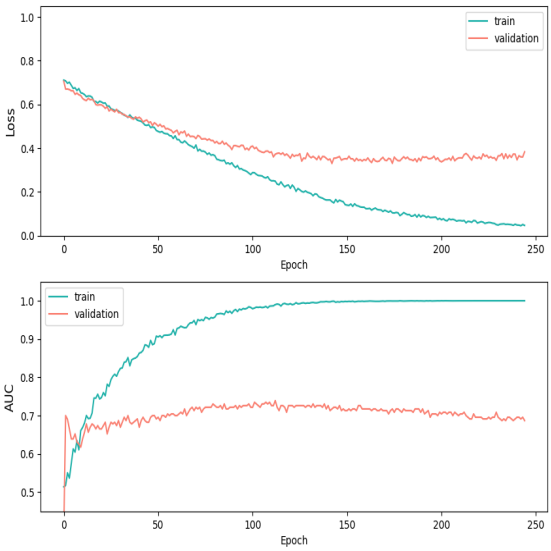
Chapter 4

Results and Discussion

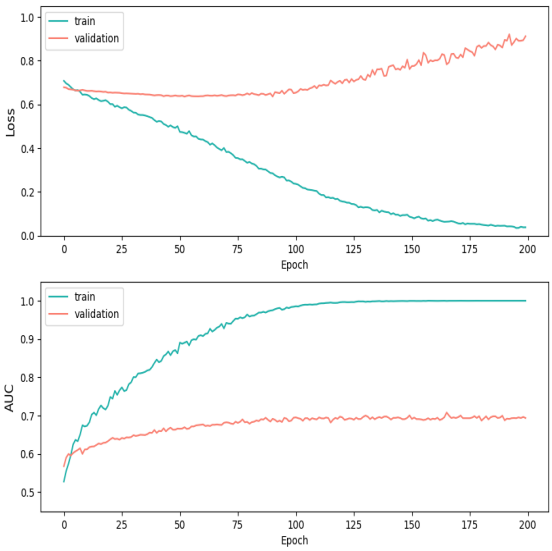
In his chapter an interpretation and critical analysis of the results will be conducted. It is divided into four sections, each one corresponding to one of the studies, being the first two related to the 2D approach, the third related to the 3D approach and the final comprises the comparison between the 2D and 3D approaches.

4.1 Study 1 - Comparison between the 2D models' performances when trained with three different datasets

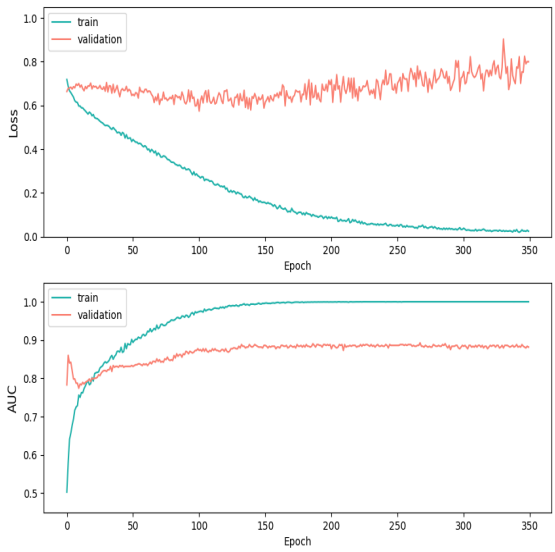
Starting with the evaluation of the learning curves (Figures 4.1a-4.1c), it is evident that all models are experiencing overfitting, since while the training loss is continuously decreasing, the validation loss only decreases until a certain epoch and then starts increasing.
Regarding the validation AUC, the model trained with the Lung-PET-CT-Dx dataset was the one with highest value once the training stabilized, achieving maximum of 0.893, followed by the NSCLC-Radiomics trained model that obtained a maximum value of 0.739, and lastly the combined dataset reached 0.708. One possible explanation for these results is the presence of intra-class variability within the datasets. Specifically, the Lung-PET-CT-Dx dataset primarily consists of samples from patients diagnosed with ADC, comprising approximately 70% of the dataset. The remaining 30% includes cases of SCC, LCC, and SCLC, with LCC accounting for just 2% of the dataset. Conversely, ADC is the least represented subtype in the NSCLC-Radiomics dataset, comprising only 14% of the cases. The remaining 86% of the dataset encompasses three distinct subtypes, all of which have substantial portions and were aggregated into the negative class. Consequently, the intra-class diversity within the negative class of the NSCLC-Radiomics dataset is presumed to be higher than that of the Lung-PET-CT-Dx dataset. Taking into account the limited size of the datasets, this increased heterogeneity within the negative class could have a detrimental effect on the performance of the model.
This effect is even more accentuated in the combined dataset, as it inherits the intra-class heterogeneity from the other two datasets and includes data from two different centers.



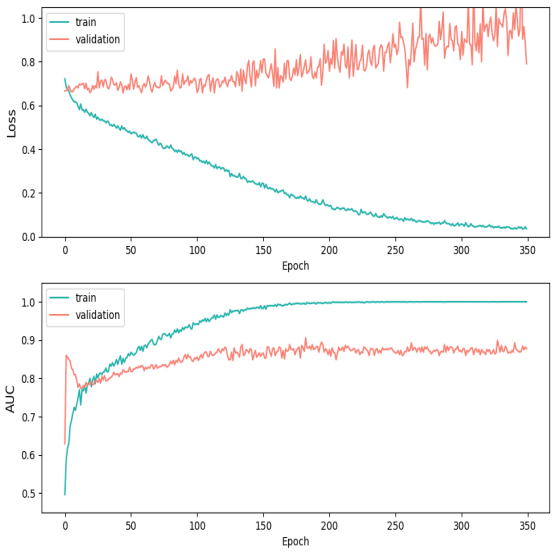
(a) NSCLC-Radiomics



(b) Combined



(c) Lung-PET-CT-Dx



(d) Multi-Slice Lung-PET-CT-Dx

Figure 4.1: Loss and AUC evolution in the train and validation sets during the training phase of the 2D approach. The parameters utilized in the four experiments were: batch size of 32, 0.5 dropout, SGD without momentum, 0.0005 learning rate, and 0.01 weight decay.

Consequently, the CT scans have inevitable dissimilarities resulting from distinct acquisition procedures, this additional intra-variability can impose a constrain on capacity of the model to recognize and learn patterns from the data, thus justifying the lower AUC value in the combined dataset validation.

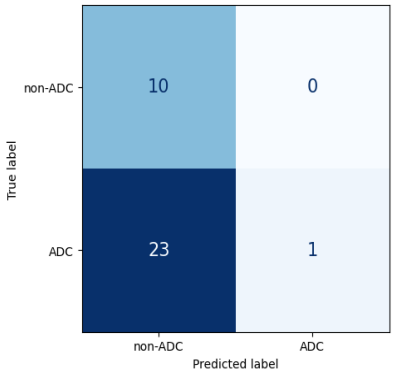
Considering the test sets, a notable decline in the AUC was observed across all cases (see Table 4.1), as anticipated due to overfitting.

Approach	Training Set	Test Epoch	Test Set	AUC	Balanced Accuracy	Precision	Recall	F1-Score	Specificity
2D	NSCLC-Radiomics	112	Lung-PET-CT-Dx	0.367	0.521	**1.000**	0.042	0.080	**1.000**
			NSCLC-Radiomics	0.583	0.542	0.111	0.333	0.167	0.750
	Combined	165	Lung-PET-CT-Dx	0.579	0.500	0.706	0.500	0.585	0.500
			NSCLC-Radiomics	0.615	0.599	0.118	**0.667**	0.200	0.531
	Lung-PET-CT-Dx	267	Lung-PET-CT-Dx	0.638	0.633	0.800	**0.667**	**0.727**	0.600
			NSCLC-Radiomics	0.490	0.448	0.067	0.333	0.111	0.563
	Lung-PET-CT-Dx (Multi-Slice)	182	Lung-PET-CT-Dx	**0.717**	**0.671**	0.867	0.542	0.667	0.800
			NSCLC-Radiomics	0.531	0.432	0.063	0.333	0.105	0.531

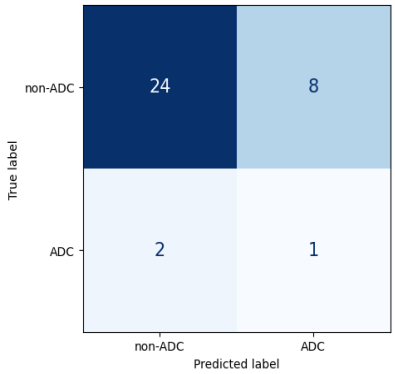
Table 4.1: Results of the 2D approach, including the training set, the epoch selected to test the model, corresponding to the maximum validation AUC, the test set and the evaluation metrics: AUC, balanced accuracy, precision, recall, F1-Score, and specificity. The top results achieved for each metric are highlighted in bold.

Regarding each individual model on their respective internal and external test set, the performance dropped in the external test sets, meaning that both models have poor generalization capability. The one trained with the NSCLC-Radiomics exhibited inferior overall performance on both test sets: however, it is worth noting that this particular model was the one that demonstrated the highest specificity for each test set, indicating its ability to identify true negative cases. This in addition to its poor performance for both test sets in terms of recall, which measures the ability to detect true positive cases, suggests that the model has a tendency to classify samples as belonging to the negative class (non-ADC). This poses significant risks, particularly in the medical domain where the misclassification of a positive case can have severe consequences, including erroneous treatment planning and substantial time costs.

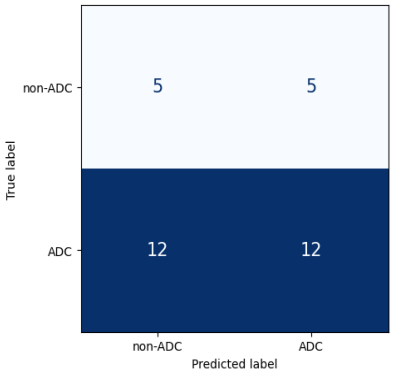
The model trained on the combined dataset was the one that produced the most similar results between the different test sets, as expected since it was trained with both corresponding training sets. Although the AUC obtained for the NSCLC-Radiomics test set is slightly higher than the one of the Lung-PET-CT-Dx test set, the respective confusion matrices (Figure 4.2) reveal that in the Lung-PET-CT-Dx test, 50% of the samples were misclassified, whereas in the NSCLC-Radiomics test set, approximately 46% were misclassified. Given that the total number of samples between the test sets differs by only one (34 samples in the Lung-PET-CT-Dx and 35 samples in the NSCLC-Radiomics), this suggests that each individual training set contributes proportionally to the definition of the model's trainable parameters, which subsequently divides the performance between the two test sets.



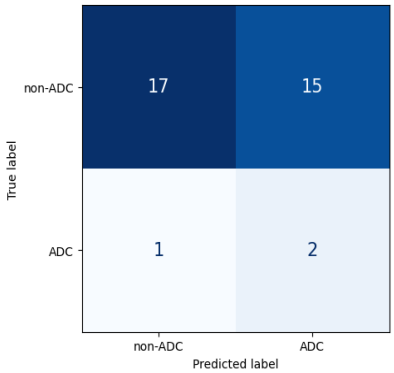
(a) Lung-PET-CT-Dx test set



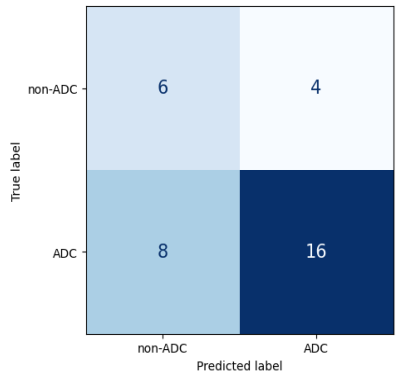
(b) NSCLC-Radiomics test set



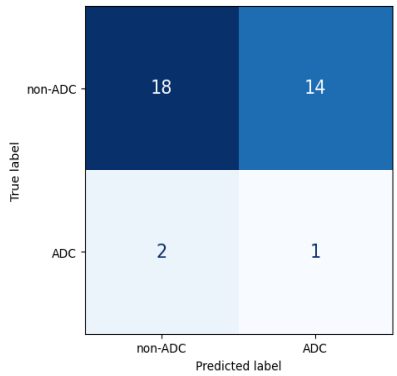
(c) Lung-PET-CT-Dx test set



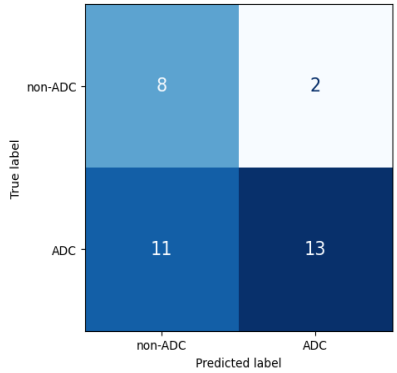
(d) NSCLC-Radiomics test set



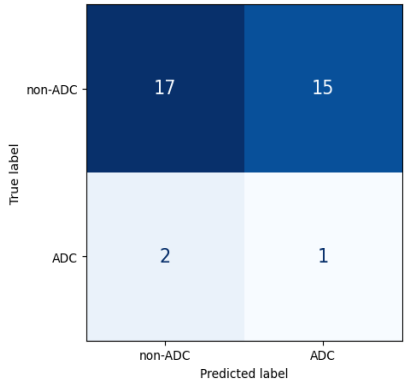
(e) Lung-PET-CT-Dx test set



(f) NSCLC-Radiomics test set



(g) Lung-PET-CT-Dx test set



(h) NSCLC-Radiomics test set

Figure 4.2: Confusion matrices on the 2D approach. Each row corresponds respectively to the models trained on the NSCLC-Radiomics (4.2a, 4.2b), on the combined (4.2c, 4.2d), on the Lung-PET-CT-Dx (4.2e, 4.2f), and on the multi-slice Lung-PET-CT-Dx (4.2g, 4.2h).

The model trained and tested with the Lung-PET-CT-Dx was the one that yield better overall results, achieving the highest value for AUC, balanced accuracy, recall, and F1-Score, only being outperformed on precision and specificity by NSCLC-Radiomics datasets. Given that and also taking into account the learning curves, the Lung-PET-CT-Dx trained model was considered the best-performing and employed in the subsequent studies.

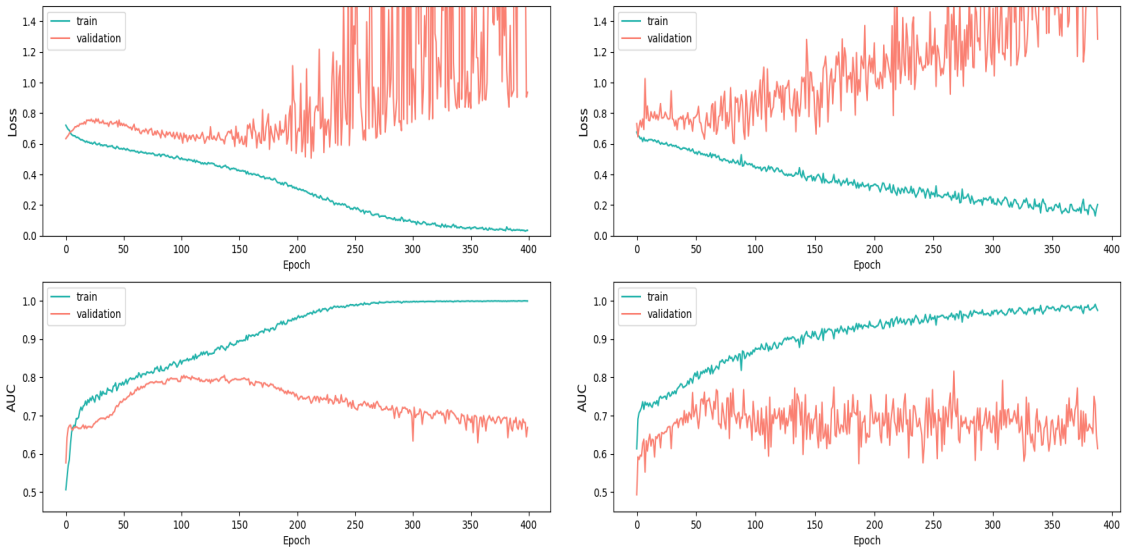
4.2 Study 2 - Evaluation of using one slice and three slices per patient for training

Comparing the learning curves of the 1-slice Lung-PET-CT-Dx trained model and the 3-slices Lung-PET-CT-Dx trained model (Figures 4.1c and 4.1d), they exhibit similar behavior, however, there are some differences to consider. Firstly, the multi-slice model requires a few more epochs to stabilize the training loss and AUC, and additionally both the validation loss and AUC of the multi-slice model displays more oscillations compared to the 1-slice model. These two differences suggest that the inclusion of multiple slices per patient introduces additional complexity to the data and, as a result, the model has taken a bit longer to converge and the validation loss and AUC

evolution presents slightly more fluctuations than in the 1-slice model. Despite these differences, the maximum validation AUC of the multi-slice model is 0.906, showing a marginal increase over the 0.893 from the 1-slice model.
Considering the test performances on Table 4.1, the multi-slice trained model had better overall results on the Lung-PET-CT-Dx test set, only being outperformed by the 1-slice model in terms of recall and F1-Score. On the other hand, the improvement of the generalization ability was not significant, as evidenced by the increase in the AUC from 0.490 to 0.531, the unchanged recall value, and the minimal decreases in the remaining four metrics on the NSCLC-Radiomics test set. However, the improvement in the overall performance suggest that it is preferred to utilize more than one slice per patient over relying on geometric data augmentation techniques to increase the number of samples in the training set. The disadvantage of using geometric data augmentation is that it involves artificially modifying the existing data, with the possibility of inadvertently alter anatomical structures of the tumors that are important for the classifier. Furthermore, the employed data augmentation methods generate new versions of the same slice, increasing the likelihood of redundant information being added to the data. These modifications could introduce artificial patterns or distortions in the data, leading to a less reliable model performance. On the other hand, by incorporating multiple slices per patient, the number of samples is increased without modifying the real structures of the nodules. Moreover, the inclusion of the additional slices offers distinct perspectives of the nodules, providing valuable information for the model, as evidenced by the improvement of the performance.

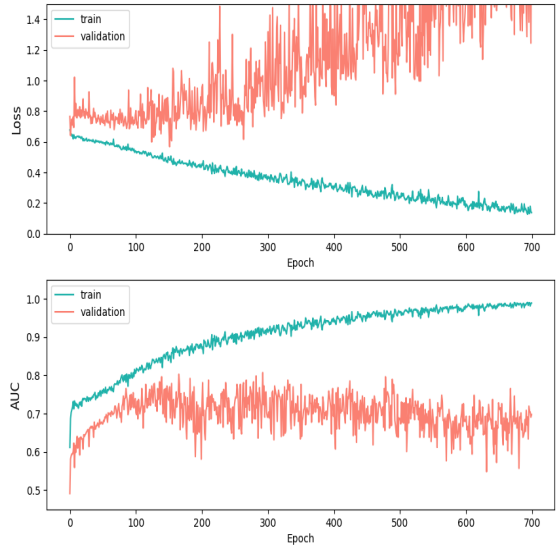
4.3 Study 3 - Comparison between the 3D models when trained from scratch and when transfer learning is employed

Upon analyzing the learning curves (Figures 4.3a-4.3d), the models with fewer frozen layers show earlier signs of overfitting, aligning with expectations as a larger number of trainable layers enables a swifter adaptation of the network's parameters to the training data. Additionally, in the case of the models that involved transfer learning, the training and validation curves for both loss and AUC exhibit considerable oscillation compared to the model trained from scratch. This may stem from potential disparities between the preprocessing techniques employed in the current study and those applied to the data used during the pretraining of the model's weights. Factors such as variations in the normalization method, resampling procedures, and original image dimensions may contribute to this disparity.

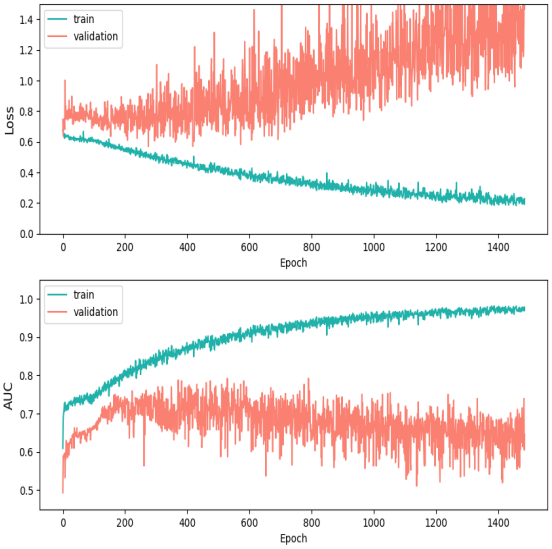


(a) Trained from scratch

(b) Six layers trained



(c) Four layers trained



(d) Two layers trained

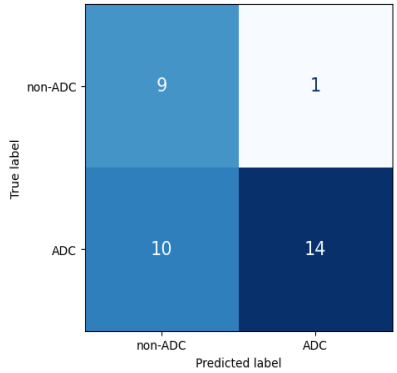
Figure 4.3: Loss and AUC evolution in the train and validation sets during the training phase of the 3D approach. Concerning the parameters, the batch size of 32, 0.5 dropout and 0.01 weight decay were common in all experiments. For the TFS model, SGD without momentum was utilized with learning rate of 0.0001, and for the remaining three experiences was the SGD with 0.9 momentum and 0.001 learning rate.

Observing the performances on the test sets (Table 4.2), it becomes apparent that all models had a poor performance on the NSCLC-Radiomics test set. First, it is important to consider the variability in the inter-slice spacing between the training and the test dataset. This could be a reason for the unsatisfactory results on the NSCLC-Radiomics test set, since it is composed only by CT volumes with 3mm of slice thickness, in contrast to the Lung-PET-CT-Dx, that presents volumes with slice thickness of 1mm, 3mm, and 5mm.

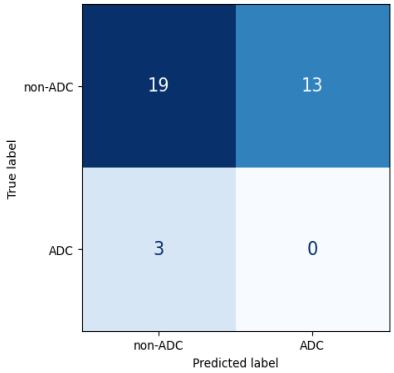
Approach	Training Set	Layers Trained	Test Epoch	Test Set	AUC	Balanced Accuracy	Precision	Recall	F1-Score	Specificity
3D	Lung-PET-CT-Dx	TFS	102	Lung-PET-CT-Dx	**0.779**	0.742	0.933	0.583	0.718	**0.900**
				NSCLC-Radiomics	0.406	0.297	0.000	0.000	0.000	0.594
		2 layers	529	Lung-PET-CT-Dx	0.658	0.613	0.789	0.625	0.698	0.600
				NSCLC-Radiomics	0.229	0.297	0.000	0.000	0.000	0.594
		4 layers	292	Lung-PET-CT-Dx	0.671	0.621	0.813	0.542	0.650	0.700
				NSCLC-Radiomics	0.385	0.313	0.000	0.000	0.000	0.625
		6 layers	267	Lung-PET-CT-Dx	0.757	**0.783**	**0.941**	**0.667**	**0.780**	**0.900**
				NSCLC-Radiomics	0.240	0.328	0.000	0.000	0.000	0.656

Table 4.2: Results of the 3D approach, including the training set, the number of layers trained, the epoch selected to test the model, corresponding to the maximum validation AUC, the test set and the evaluation metrics: AUC, balanced accuracy, precision, recall, F1-Score, and specificity. The top results achieved for each metric are highlighted in bold.

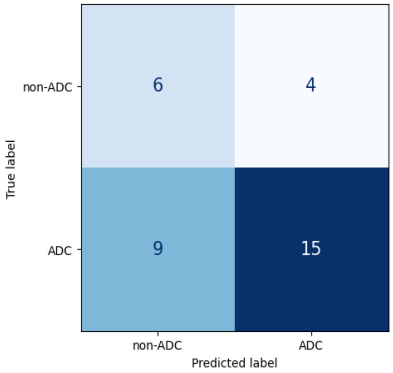
An analysis of the slice thickness distribution on the training set revealed that only 2.38% of the samples had 3mm slice thickness. This can cause a detrimental effect on the performance on the NSCLC-Radiomics test set since the 3D architectures treat the samples as volumes, and may not generalize optimally when confronted with varying values of slice spacing, deviating from the most common ones on the training set. Another fact that is noteworthy is that the NSCLC-Radiomics test set only had three samples of the positive class (ADC), which were all misclassified by the models, as evidenced by the confusion matrices in Figure 4.4.



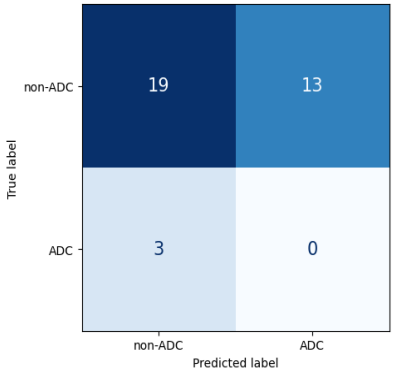
(a) Lung-PET-CT-Dx test set



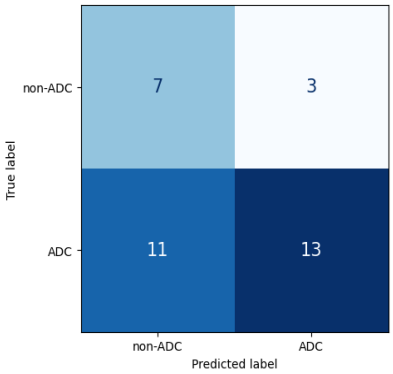
(b) NSCLC-Radiomics test set



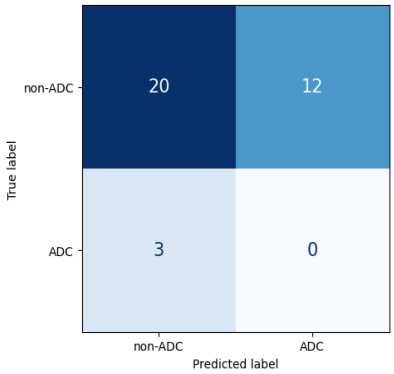
(c) Lung-PET-CT-Dx test set



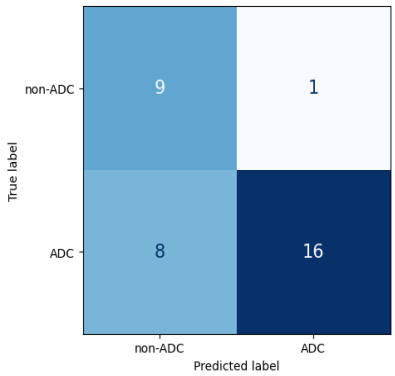
(d) NSCLC-Radiomics test set



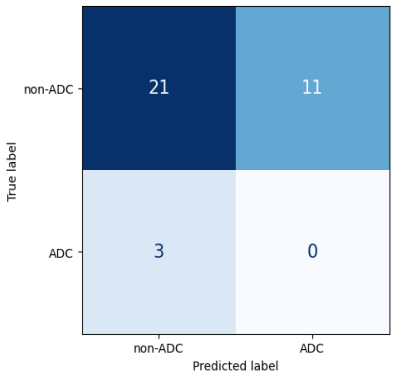
(e) Lung-PET-CT-Dx test set



(f) NSCLC-Radiomics test set



(g) Lung-PET-CT-Dx test set



(h) NSCLC-Radiomics test set

Figure 4.4: Confusion matrices on the 3D approach. Each row corresponds respectively to the models trained from scratch (4.4a, 4.4b), with two layers trained (4.4c, 4.4d), with four layers trained (4.4e, 4.4f), and with six layers trained (4.4g, 4.4h).

Consequently, the precision, recall, and F1-Score values resulted in zero. This, in addition to the subpar performances of the 2D models on this particular test set, its reduced size, and the intra-class variability of the NSCLC-Radiomics dataset, may suggest that the test set it is not representative of the data.

Based on the performance evaluation conducted on the Lung-PET-CT-Dx test set, it was confirmed that there is an increase in the AUC as the number of trained layers in the model increases, achieving its highest value on the model trained from scratch. Moreover, there is a significant improvement in the overall performance of the models as they transition from having four unfreezed layers to having six unfreezed layers, that is when more than half of the layers in the network are trained. This observation suggests that the most generic MedicalNet weights, that are those associated with the earlier layers, are most appropriate for this specific classification task. One possible explanation is that these weights were initially trained for segmentation tasks including various organs in the human body, thereby making the weights of the final layers more suitable for this type of applications.

Both the TFS model and the model trained with six layers demonstrated similar performance, with the six-layers trained model achieving slightly better results in terms of balanced accuracy, precision, recall, and F1-Score. However, considering that these differences were marginal and the fact that the TFS model achieved a higher AUC and had a smoother training progress, the TFS model was selected to be the best-performing 3D model.

4.4 Study 4 - Comparison between the performance of the 2D and the 3D models

This final section aims to provide a comparative analysis between the top-performing models in both the 2D and 3D approaches, specifically the 2D model trained with the multi-slice version of the Lung-PET-CT-Dx and the 3D model trained from scratch.

From the learning curves of the 3D model trained from scratch, it is noticeable that the performance of the validation set starts to decline abruptly compared to the 2D case, being characterized by a significant escalation of the loss with heavy oscillations and, concurrently, by a decrease in the AUC. This behavior can be justified given the high susceptibility of the 3D architectures to the overfitting phenomenon, which is significantly more pronounced compared to their 2D counterparts. The augmented complexity of 3D architectures is evidenced by their substantially larger number of trainable parameters, nearly three times that of 2D architectures.
Looking at the performances on the NSCLC-Radiomics test set, although still not with a good performance, the multi-slice trained model achieved highest values in the metrics. This can mean that the inherent differences between the Lung-PET-CT-Dx and the NSCLC-Radiomics datasets are even more evident when the samples are analysed as volumetric units.
On the Lung-PET-CT-Dx test set, the 3D model outperformed the 2D one in every evaluation metric. This is supported by the fact that the 3D model takes into account the spatial context of the volume, capturing spatial relationships among consecutive slices. Consequently, the 3D models are able to learn patterns and structures along the volume that can not be fully captured by the 2D models.

Chapter 5

Conclusion

The proposed methodology was based on deep learning, aligning with the majority of the studies in the literature. Both 2D and 3D CNNs were employed, enabling a direct comparative analysis of these architectural variants, an aspect that was absent in the literature review conducted. Furthermore, a noteworthy aspect of the current work was the evaluation of the effects of varying the number of slices per patient in the 2D approach, encompassing both single-slice and three-slice training sets, in order to evaluate the influence of augmenting the number of samples while resorting to fewer artificially generated data. It is also important to emphasize that in the literature review conducted no studies were found that performed the binary classification between adenocarcinoma and non-adenocarcinoma, with the negative class encompassing multiple subtypes of lung cancer.

The first study conducted for the 2D approach confirmed the importance that choosing a suitable dataset has on model performance. From the second study, it was possible to conclude that it is preferable to use multiple slices per patient rather than applying multiple geometric data augmentation techniques to increase the number of training samples. In the 3D approach, a comparison was made between training the model from scratch or using transfer learning, varying the number of layers that were fine tuned. The TFS model led to better results, followed by the model with six layers trained. Finally, a comparison was made between the performance of the 2D and 3D models, with the 3D model providing better overall results on the internal test set, achieving an AUC of 0.779.

In future works, some enhancements can be implemented in the methodology to attain optimal outcomes. Primarily, the utilization of cross-validation should be incorporated to mitigate the risk of evaluating models on unrepresentative subsets, ensuring more reliable results and improving the robustness of the models. The data augmentation process should also be review, and given the findings in study 2, the employment of fewer transformations, for example just to balance the classes on the test set, could be experimented. This way, the possibility of introducing misleading information and redundancy to the training set is minimized, at a cost of having less training data. Additionally, the employment of resampling techniques to ensure uniformity in the slice thickness across the datasets should be considered in order to improve the performance of the 3D models.

Appendix A

Cases Removed from the Datasets

A.1 NSCLC-Radiomics

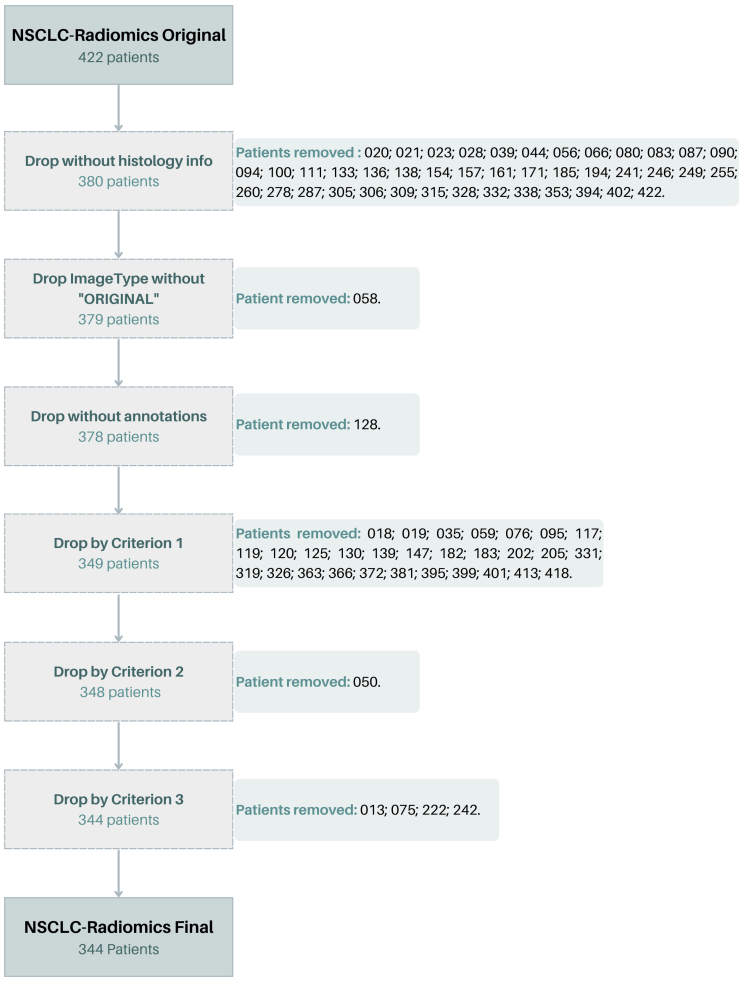


Figure A.1: Block diagram for the NSCLC-Radiomics dataset, including the exclusion criteria and the number of patients removed at each step.

A.2 Lung-PET-CT-Dx

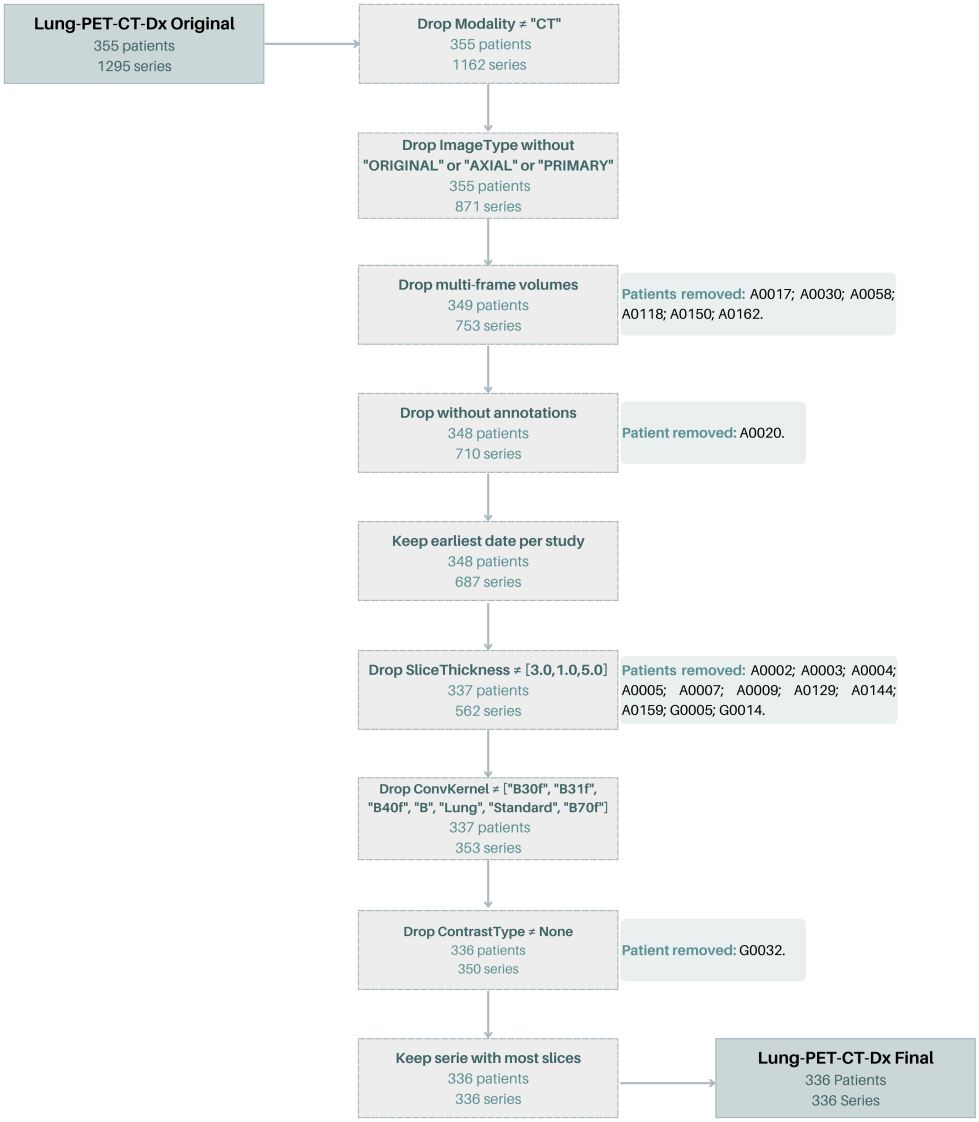


Figure A.2: Block diagram for the Lung-PET-CT-Dx dataset, including the exclusion motives and the number of patients and series removed at each step.

References

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

[2] T. D. William, E. Brambilla, A. G. Nicholson, Y. Yatabe, J. H. M. Austin, M. B. Beasley, L. R. Chirieac, S. Dacic, E. Duhig, D. B. Flieder, K. Geisinger, F. R. Hirsch, Y. Ishikawa, K. M. Kerr, M. Noguchi, G. Pelosi, C. A. Powell, M. S. Tsao, and I. Wistuba. "The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances since the 2004 Classification". *Journal of Thoracic Oncology*, 10(9):1243–1260, 2015. doi:10.1097/JTO.0000000000000630.

[3] H. Mohamed and Omarhanyy. "Chest CT-Scan images Dataset | Kaggle". (accessed Dec. 16, 2022). URL: https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images.

[4] N. Duma, R. Santana-Davila, and J. R. Molina. "Non–Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment". *Mayo Clinic Proceedings*, 94(8):1623–1640, 2019. doi:10.1016/j.mayocp.2019.01.013.

[5] "Lung Cancer - Non-Small Cell | Cancer.Net". (accessed Nov. 07, 2022). URL: https://www.cancer.net/cancer-types/lung-cancer-non-small-cell.

[6] "CT Scan In Klang - Sri Kota Specialist Medical Centre". (accessed Jan. 05, 2023). URL: https://www.srikotamedical.com/2021/08/05/ct-scan-in-klang/.

[7] P. Marentakis, P. Karaiskos, V. Kouloulias, N. Kelekis, S. Argentos, N. Oikonomopoulos, and C. Loukas. "Lung cancer histology classification from CT images based on radiomics and deep learning models". *Medical Biological Engineering Computing*, 59:215–226, 1 2021. doi:10.1007/s11517-020-02302-w.

[8] T. Saikia, M. Hansdah, K. K. Singh, and M. K. Bajpai. "Classification of Lung Nodules based on Transfer Learning with K-Nearest Neighbor (KNN)". In *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–6, 2022. doi:10.1109/IST55454.2022.9827668.

[9] D. Goldgof, L. Hall, S.H. Hawkins, M.B. Schabath, O. Stringfield, A. Garcia, Y. Balagurunathan, J. Kim, S. Eschrich, A.E. Berglund, R. Gatenby, and R.J. Gillies. "Long and Short Survival in Adenocarcinoma Lung CTs". *The Cancer Imaging Archive*, 2017. doi:10.7937/K9/TCIA.2017.0tv7b9x1.

[10] O. Grove, A.E. Berglund, M.B. Schabath, H.J.W.K. Aerts, A. Dekker, H. Wang, E.R. Velazquez, P. Lambin, Y. Gu, Y. Balagurunathan, E. Eikman, R.A. Gatenby, S. Eschrich, and

R.J. Gillies. "Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma". *The Cancer Imaging Archive*, 2015. doi:10.7937/K9/TCIA.2015.A6V7JIWX.

[11] Y. Feng. "CT Volume Samples for Lung Adenocarcinoma Classification". *Mendeley Data*, 2, 2020. doi:10.17632/R3TBSGTPZG.2.

[12] P. Li, S. Wang, T. Li, J. Lu, Y. HuangFu, and D. Wang. "A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis". *The Cancer Imaging Archive*, 2020. doi:10.7937/TCIA.2020.NNC2-0461.

[13] "LabelImg - PyPI". (accessed Dec. 28, 2022). URL: https://pypi.org/project/labelImg/.

[14] H. J. W. L. Aerts, L. Wee, E. Rios Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin. "NSCLC-Radiomics". *The Cancer Imaging Archive*, 2019. doi:10.7937/K9/TCIA.2015.PF0M9REI.

[15] R. Patil, G. Mahadevaiah, and A. Dekker. "An Approach Toward Automatic Classification of Tumor Histopathology of Non–Small Cell Lung Cancer Based on Radiomic Features". *Tomography*, 2(4):374–377, 2016. doi:10.18383/j.tom.2016.00244.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". *Journal of Artificial Intelligence Research*, 16:321–357, jun 2002. doi:10.1613/jair.953.

[17] A. Pandey and A. Kumar. "Deep Features Based Automated Multimodel System for Classification of Non-Small Cell Lung Cancer". In *2022 IEEE Delhi Section Conference (DELCON)*, pages 1–7, 2022. doi:10.1109/DELCON54057.2022.9753643.

[18] A. Rehman, M. Kashif, I. Abunadi, and N. Ayesha. "Lung Cancer Detection and Classification from Chest CT Scans Using Machine Learning Techniques". In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, pages 101–104, 2021. doi:10.1109/CAIDA51941.2021.9425269.

[19] L. Ubaldi, V. Valenti, R. F. Borgese, G. Collura, M. E. Fantacci, G. Ferrera, G. Iacoviello, B. F. Abbate, F. Laruina, A. Tripoli, A. Retico, and M. Marrale. "Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples". *Physica Medica: European Journal of Medical Physics*, 90:13–22, 10 2021. doi:10.1016/J.EJMP.2021.08.015.

[20] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts. "Computational Radiomics System to Decode the Radiographic Phenotype". *Cancer Research*, 77(21):e104–e107, 10 2017. doi:10.1158/0008-5472.CAN-17-0339.

[21] C. Ma and S. Yue. "Minimum Sample Size Estimate for Classifying Invasive Lung Adenocarcinoma". *Applied Sciences*, 12(17), 2022. doi:10.3390/app12178469.

[22] C. Wang, J. Shao, J. Lv, Y. Cao, C. Zhu, J. Li, W. Shen, L. Shi, D. Liu, and W. Li. "Deep learning for predicting subtype classification and survival of lung adenocarcinoma on computed tomography". *Translational Oncology*, 14(8):101141, 2021. doi:https://doi.org/10.1016/j.tranon.2021.101141.

[23] F. Silva, T. Pereira, J. Frade, J. Mendes, C. Freitas, V. Hespanhol, J. L. Costa, A. Cunha, and H. P. Oliveira. "Pre-Training Autoencoder for Lung Nodule Malignancy Assessment Using CT Images". *Applied Sciences*, 10:7837, 11 2020. doi:10.3390/app10217837.

[24] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, V. Nath, Y. He, Z. Xu, A. Hatamizadeh, A. Myronenko, W. Zhu, Y. Liu, M. Zheng, Y. Tang, I. Yang, M. Zephyr, B. Hashemian, S. Alle, M. Z. Darestani, C. Budd, M. Modat, T. Vercauteren, G. Wang, Y. Li, Y. Hu, Y. Fu, B. Gorman, H. Johnson, B. Gaux, B. S. Erdal, V. Gupta, A. Diaz-Pinto, A. Dourson, L. Maier-Hein, P. F. Jaeger, M. Baumgartner, J. Kalpathy-Cramer, M. Flores, J. Kirby, L. A. D. Cooper, H. R. Roth, D. Xu, D. Bericat, R. Floca, S. K. Zhou, H. Shuaib, K. Farahani, K. H. Maier-Hein, S. Aylward, P. Dogra, S. Ourselin, and A. Feng. "MONAI: An open-source framework for deep learning in healthcare". 11 2022. doi:https://doi.org/10.48550/arXiv.2211.02701.

[25] J. Egger, C. Gsaxner, A. Pepe, K. L. Pomykala, F. Jonske, M. Kurz, J. Li, and J. Kleesiek. "Medical deep learning—A systematic meta-review". *Computer Methods and Programs in Biomedicine*, 221, 6 2022. doi:10.1016/j.cmpb.2022.106874.

[26] Lulu Wang. "Deep Learning Techniques to Diagnose Lung Cancer". *Cancers*, 14:5569, 11 2022. doi:10.3390/cancers14225569.

[27] S. Chen, K. Ma, and Y. Zheng. "Med3D: Transfer Learning for 3D Medical Image Analysis". 4 2019. doi:10.48550/ARXIV.1904.00625.

[28] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". *CoRR*, 2015. doi:https://doi.org/10.48550/arXiv.1512.03385.