FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

AI-based Cancer Characterization Using Multimodal Data

Pedro Mendes Antunes de Matos

WORKING VERSION

Mestrado em Bioengenharia

Supervisor: Hélder Oliveira, PhD

Second Supervisor: Claúdia Freitas, MD

Third Supervisor: Francisco Silva, MSc

October 28, 2022

AI-based Cancer Characterization Using Multimodal Data

Pedro Mendes Antunes de Matos

Mestrado em Bioengenharia

October 28, 2022

Resumo

Cancro do pulmão é o tipo de cancro com maior taxa de mortalidade no mundo. Um dos fatores mais importantes a contribuir para este alto nível de mortalidade é o diagnóstico tardio desta doença, que frequentemente acontece devido à ausência de sintomas na fase inicial da doença. Apesar da biópsia ser o procedimento diagnóstico mais utilizado para a caracterização de tumores, é também um processo invasivo, que acarreta potenciais riscos para a saúde. A biópsia providencia informação histológica e química importante sobre o tumor, no entanto, não possibilita a caracterização da sua estrutura heterogénea.

Imagens médicas obtidas por procedimentos não-invasivos podem oferecer uma perspetiva mais completa sobre a estrutura tri-dimensional de um tumor. Imagens médicas podem ser também utilizadas para treinar algoritmos preditivos, com diversos estudos a conseguirem atingir resultados promissores na previsão de cancro do pulmão a partir de imagens de Tomografia Computorizada. Uma análise da literatura evidencia um foco nas abordagens baseadas em *deep learning* motivado pelos bons resultados obtidos. Apesar destes modelos conseguirem obter bons resultados, eles diferem da prática médica, na qual, o diagnóstico não é apenas baseado em imagens médicas mas também noutros tipos de informações, como dados clínicos e laboratoriais.

Este trabalho procurou investigar a combinação de diferentes tipos de dados aplicada à previsão de cancro no pulmão. Nesse sentido, utilizaram-se informações clínicas de pacientes e informação extraída de nódulos pulmonares. Uma experiência preliminar foi levada a cabo com o intuito de analisar os benefícios da utilização de um pulmão inteiro para a previsão de cancro de pulmão, tendo sido obtido um valor médio de AUC de 0.594. Optando-se por uma abordagem mais local, efetuou-se a previsão de cancro do pulmão através da análise de nódulos extraídos de exames de tomografia computorizada, tendo sido obtido um valor médio de AUC de 0.730. Foram desenvolvidas três estratégias multimodais, sendo que a melhor destas estratégias originou um resultado com valor médio de AUC de 0.755. Os resultados obtidos demonstram o potencial da utilização de diferentes tipos de dados para o diagnóstico de cancro do pulmão e sugerem que os dados clínicos oferecem informações complementares às características dos nódulos, tendo-se verificado, no entanto, que este fator complementar está altamente dependente da estratégia multimodal escolhida.

Abstract

Lung cancer is the type of cancer with the highest mortality rate in the world. One of the main factors that contributes to this high death rate is late diagnosis, that frequently occurs largely due to lack of early symptoms. Despite biopsy being the most utilized diagnostic method for characterizing tumors, it is an invasive procedure carrying with it health risks. It provides important histological and chemical information on the tumor but fails to yield information on the tumor's heterogeneous structure.

Medical imaging can offer a full perspective on the tumour's three-dimensional structure, while remaining virtually a non-invasive diagnostic method. Medical images can also be utilized for training predictive algorithms, with many works achieving promising results in predicting lung cancer from Computed Tomography (CT) images. Most of research found focuses on image-based deep learning approaches due to the high performances obtained through these methods. While these models can, in fact, perform well in this task, they differ from routine medical diagnosis practice where diagnosis is not based on image data alone, but also in many other different types of data such as clinical and laboratorial information.

This work aims at studying the combination of more than one type of information for predicting lung cancer, specifically, extracted from nodules CT scans and clinical data. A preliminary experiment analysing the value of using a whole lung as imaging data in predicting lung cancer was conducted, having reached a mean AUC of 0.594. Moving to a more local approach, lung cancer was predicted through the analysis of 3D segmented nodules, having the model used reached 0.730 of AUC. Three multimodal strategies, using the extracted nodule volumes and clinical data from the patient, were devised with the best one reaching a mean AUC of 0.755. The results translate the potential of leveraging different types of information when diagnosing lung cancer and suggest that clinical data might offer useful complementary information, additional to the nodule's characteristics. However, this complementary factor was found to be highly dependent on the multimodal approach chosen.

Acknowledgements

Firstly I would like thank Professor Hélder P. Oliveira, Tânia Pereira and Francisco Silva for the tireless support given throughout the development of this dissertation, many times motivating and challenging me to improve my work while always remaining understanding and flexible. I would like to thank my parents and my friends for always supporting me and lifting my spirits throughout this journey.

Pedro Matos

Contents

List of Figures

List of Tables

Abbreviations

3D	Three-Dimensional
AUC	Area Under the Curve
CDC	Centers for Disease Control and Prevention
CNN	Convolutional Neural Network
DL	Deep Learning
XGBoost	Extreme Gradient Boosting
CT	Computed Tomography
HU	Hounsfield Unit
NSCLC	Non-Small Cell Lung Cancer
RF	Random Forest
SCLC	Small Cell Lung Cancer
SVM	Support Vector Machine

Chapter 1

Introduction

1.1 Context

Lung cancer is the second most occurring type of cancer in the world with 2.2 million cases reported in 2020. Representing approximately 1 in 5 cancer deaths (18,0%), lung cancer was the leading cause of cancer death with 1.8 million deaths in the same year [1].

Smoking is by far the single biggest risk factor when it comes to lung cancer. In the US, the Centers for Disease Control and Prevention agency (CDC) estimates that 80% to 90% of lung cancer deaths are directly attributed to smoking habits [2]. When analysing lung cancer incidence in a regional basis, it is apparent that the incidence of this disease in specific world regions accompanies smoking habits trends, as regions with growing smoking rates, such as Eastern Asia and Northern Africa, have been experiencing increases in lung cancer incidence, further highlighting smoking as a major risk factor for lung cancer [1].

Clinically, lung cancer can be divided into three main types of cancer: non-small cell lung cancer (NSCLC) which represents the majority of cancer diagnosis with about 85% of all lung cancer cases, small cell lung cancer (SCLC) and lung carcinoid tumors [3].

The 5-year survival rate for lung cancer is 59% in cases where the disease is localized (within the lungs) and 6% when the disease has metastasized (spread to other organs). However, due to lack of early symptoms, only 17% of cases are diagnosed at an early stage [4].

1.2 Motivation

Besides early detection in lung cancer diagnosis, choosing the more suitable treatment for a specific patient is of major importance in improving the clinical outcome of that patient. Such a decision is dependent on a number of factors, some of whom related to the characteristics of the lung tumor namely, morphology, tissue type, size, structure, among others [5]. The conventional procedure that allows this characterization of the lung tumor is the tissue biopsy, in which a sample of tumor tissue is extracted and analysed. Despite being the most reliable approach for this end, a lung tissue biopsy can be a painful procedure for the patient and can even carry with it risk of complications such as pneumothorax, hemorrhage or infection [6, 7]. In addition, extracting a small tissue sample from a tumor may not provide a full characterization since there can be strong region based heterogeneity within a tumor [8]. In many situations multiple biopsies are required, entailing further risk of complication, and delaying the patient's diagnosis. [7] Also, in certain cases, the location of the tumor can make it impossible for the extraction of tumor tissue, thus, preventing the employment of a biopsy [9]. These issues motivate the need to develop non-invasive methods of characterization of the tumor. Medical imaging can be considered a clear solution for this need since it can provide information on the totality of the tumor, whilst being obtained in non-invasive procedures. Medical imaging can also prompt the development of predictive models that are able to extract meaningful patterns, many of whom beyond human perception, and link them with a given target prediction [10].

In current clinical practice, a CT scan is commonly examined by a radiologist who produces a report describing possible abnormal findings detected as well as an interpretation of those findings [11]. The rapid growth of the medical imaging field has made imaging diagnostic tests more accessible for the population originating large volumes of medical imaging data to be analysed by radiologists, making this a time-consuming process [12]. Furthermore, the interpretation of medical images is prone to human error and has been shown to be variable across different experts [13]. Deep learning based predictive models, in specific, models based on Convolutional Neural Networks (CNN) have been shown to achieve high performances in medical imaging related tasks such as classification [14, 15], segmentation[16, 17], and lesion detection [18, 19], which has motivated researchers on the development of this particular models in the medical imaging field [20]. On the other hand, the majority of research found in the literature on classification of lung nodules has focused on approaches using only information extracted from medical images [14, 15, 21].

Although image-only based approaches have demonstrated potential for accurate automated diagnosis prediction, they differ from the routine clinical practice in which diagnosis is based not only on image data, but also on other available information such as clinical and laboratorial data [11]. In addition, different studies investigating the influence of clinical information on the computed tomography (CT) reporting of radiologists have concluded that providing radiologists with correct clinical information improves the quality of radiology reports [22, 23]. For this reason, there may be some advantages in developing approaches that combine information from different data modalities, exploiting possible complementary relations between them, and provide more contextually relevant perspectives on medical imaging classification tasks.

1.3 Objectives

This dissertation aims to study the combination of different data modalities in lung cancer classification of lung nodules. For this end, this work will feature the development of a model that takes in lung nodule CT volumes, as well as clinical information from the patient in order to produce a malignancy prediction for a given patient. The combination of these types of information is expected to produce a more robust model of lung nodule classification that approximates this computational method to the current clinical practice of diagnosis.

1.4 Contributions

This dissertation presents the following contributions:

- Development of a lung nodule malignancy classification model that uses CT images and is tested on two different datasets.
- Development of a lung nodule malignancy classification model that uses CT images and clinical data as input.
- A study on the benefits of combining different data modalities in the classification of lung nodules.

Furthermore, during the thesis work, an application was developed capable of acquiring nodule 3D regions of interest through a manual segmentation process.

1.5 Document Structure

This document is divided into 7 chapters. The first and current chapter serves as an introduction to the dissertation, explaining the motivation and goals behind it. Chapter 2 provides information on a few clinical concepts important for understanding the work developed, including a general analysis of lung cancer and its impact, and a more technical characterization of CT scans. Chapter 3 exposes some relevant research, on one hand, on the topic of classification of lung nodules using deep learning models and, on the other, on approaches that include deep learning models using different modalities of data for biomedical applications. Chapter 4 details the datasets utilized in this dissertation as well as the necessary steps for transforming the data into a proper input format. Chapter 5 regards the methodology of the approaches describing the experiences conducted. Chapter 6 exposes the results achieved by the experiences conducted and provides an interpretation for the results. In Chapter 7 concluding remarks and future work is discussed.

Chapter 2

Background

2.1 Lung Cancer

Lung cancer originates when a number of abnormal cells divide in an uncontrolled fashion. A lung tumor is formed from the aggregation of these cells. Cancer that first forms itself in the lungs is called primary lung cancer. Secondary lung cancer is formed as a result of a different tumour spreading, in a process referred to as metastasis. [24].

Lung cancer is one of just a few types of cancer commonly referred to as "silent cancers". This is because patients with lung cancer often only manifest symptoms in advanced stages, causing the late diagnosis of this disease, which stands as a major factor that contributed to lung cancer being the leading cause of cancer death in 2020 [1]. In fact, only 17% of lung cancer cases are diagnosed at an early stage. If the disease has metastasized then the 5-year survival rate is only 6%. With an early diagnosis of lung cancer there is a greater chance that the disease is localized within the lungs, in which case the 5-year survival rate increases to 59% [4].

With a timely diagnosis being of such importance to the outcome of the patient [25], the questions arise of how lung cancer can be detected earlier and how all the necessary information for the diagnosis can be obtained promptly.

2.2 CT images

From the many types of medical imaging that exist, the one most closely associated with lung cancer screening and diagnosing is the CT scan. Although chest radiography is also a possibility for detecting lung cancer, CT scans seem to be more suitable for this task due to being able to reveal smaller abnormalities [26].

A CT scan is an imaging procedure that uses X-ray radiation to produce detailed three-dimensional (3D) representations of the structures inside of the human body [27]. The exam is performed with a CT scanner that features a tubular gantry containing X-ray sensitive detectors placed in opposite sides of an X-ray emitter. The X-ray beams pass through the body suffering attenuation (reduction of energy) as a result of the different body tissues absorbing some of the radiation. The level of attenuation can be measured by the detectors and is directly correlated with the density of the tissues through which the X-ray beam passes [27]. The relative radiodensity of a tissue can be expressed through a scale of units called Hounsfield units (HU). Equation 2.1 translates how to calculate the HU values based on the linear attenuation coefficient of water μ_{water}, the linear attenuation coefficient of air μ_{air} and the linear attenuation coefficient of the substance μ [28].

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{2.1}$$

In the Hounsfiled scale, water at standard pressure and temperature is considered the reference standard, having an HU value of 0. A denser tissue will have a higher HU value, translating in a brighter representation in the image. Lung tissue in particular is one of the tissues in the body with less density having a HU value that can range from -500 to -900 HU [28]. In most cases, the full range of the Hounsfield scale is not represented in an image since the limited number of gray shades the human eye can perceive restricts the level to which anatomical structures can be differentiated. Instead, typically a specific range of HU values is expanded to cover the entire gray spectrum in a process commonly referred to as "windowing". This technique allows for better contrast between tissues with approximated HU values [27]. Figure 2.1 highlights the difference between a CT slice employing the full range of the Hounsfield scale and the same image adjusted to a window of HU with center value in -300 HU and width of 1400 HU, a typical window of values employed in lung CT scans.

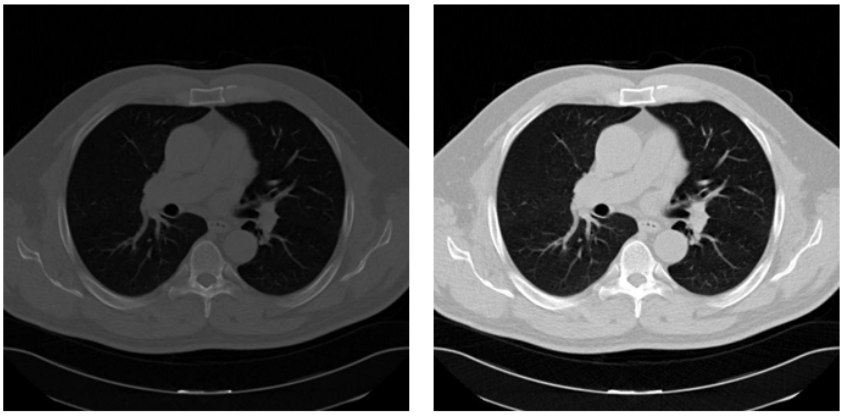


Figure 2.1: Visual difference between a CT slice employing the full HU scale (on the left) and the same slice with the window technique applied (on the right).

2.3 Features

The different types of information necessary for pattern recognition algorithms to be applied are often structurally formatted in features. In this work, two types of features will be utilized:

- **Clinical features** — these originate from general information about a patient such as age and gender but also patient background information like smoking history and family disease history. These types of features are presented in a tabular form in which, typically, each column constitutes a separate feature.
- **Deep imaging features** — these are large amounts of quantitative information extracted from the medical images. In the case of this investigation, all of these features were retrieved from the CT scans through the aid of deep learning algorithms. Despite appearing as unstructured data, these features have been proven to hold information useful for lung cancer predictive tasks.

Chapter 3

Literature Review

This chapter is divided in two sections. The first section contains relevant studies regarding the classification of lung cancer nodules (Section 3.1) and the second section (Section 3.2) features multimodal approaches combining medical imaging and other types of data for various medical predictive tasks.

3.1 Classification of lung nodules

With the rapid progress of the computer vision field in recent years, many different algorithms have been developed for automatic analysis of medical imaging, in many cases achieving very promising results [21, 14, 15]. Deep learning based algorithms have been proven to achieve very competitive performances in image classification tasks and, therefore, the bulk of approaches verified in the literature for classification of lung nodules are dominated by deep learning algorithms [29].

Ozdemir et al. [21] proposed an approach based on a 3D CNN model incorporated in an end-to-end pipeline for detection and diagnosis of lung tumors. In this work, firstly nodule-level assessment was performed and, by maintaining a relation between a given patient and his different nodules, patient-level malignancy classification was also performed, achieved with a multiple instance learning framework. The network's performance was evaluated using the 2017 Data Science Bowl on Kaggle [30] achieving an area under the curve (AUC) of 0.87. To increase the robustness of the approach, two different strategies of model uncertainty were employed, namely deep ensembles, consisting of training different models with different train and testing dataset splits and using Monte Carlo dropout, a technique in which random neurons in the network are dropped both in training and testing, allowing predictive distributions to be calculated from multiple evaluations on the testing data.

Liu et al. [14] proposed an ensemble method consisting of three different types of 3D CNN in order to better make full use of the 3D spatial information in CT images of lung nodules. The three different structures were based on the VGG-Net [31], ResNet [32] and InceptionNet [33]. Furthermore, as shown in Figure 3.1, each network architecture was divided into three sub-networks with different input sizes.

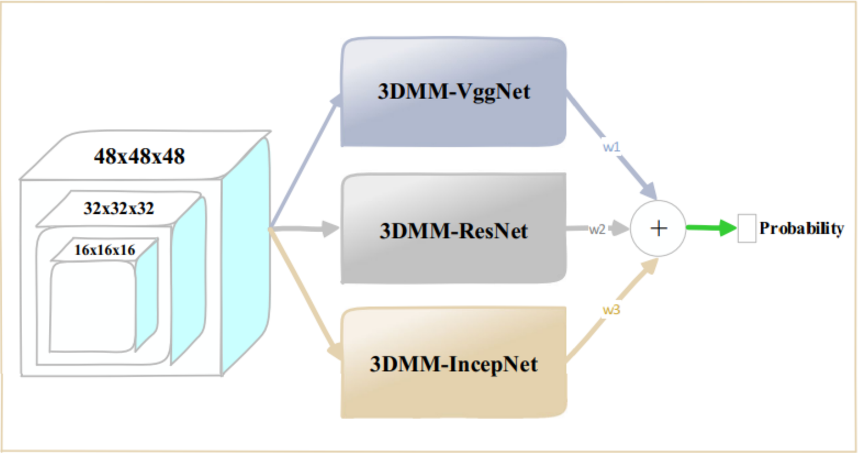


Figure 3.1: General pipeline of the approach. The probability originated by the model is a weighted average of the output probabilities of the three structures. Adapted from [14].

An important step of pre-processing, meant to improve performance in low-contrast nodules, is also mentioned in which image enhancement is performed to the lung nodules by adding another element to the forth dimension of the data (number of channels). This meant that, along with the original nodule volume, an enhanced version of the same nodule is added featuring alterations to the voxels intensities either by manipulating the intensity histograms or by directly transforming them with formula based methods. The ensemble method slightly outperformed each of the used networks individually obtaining 0.939 of AUC using the LIDC-IDRI dataset [34].

Using the same dataset, Dey et al. [15] aimed at improving the optimization process of a 3D CNN by introducing early outputs in a network tasked with lung nodule malignancy prediction. The architecture of the network, observed in Figure 3.2, has a connection structure based on the DenseNet [35] and contains outputs after every convolution and pooling blocks. The feature maps originating the intermediate outputs were also concatenated to the final classification output layer. In the optimization process, a custom loss function was developed taking into account the intermediate outputs of the network. According to this study, obtaining and using intermediate outputs as well as adding connections between layers (achieved through a DenseNet architeture) shorten the distance between input and output, having the potential to improve the optimization of the model. The performance of the proposed model was compared to three other models: two of these models, (one with DenseNet connection layout and one without) did not contain the intermediate outputs; the other one featured such outputs but no DenseNet structure. The work found that the proposed multi-output 3D DenseNet model outperformed the other models having reached 0.9548 of AUC and 0.9040 of accuracy.

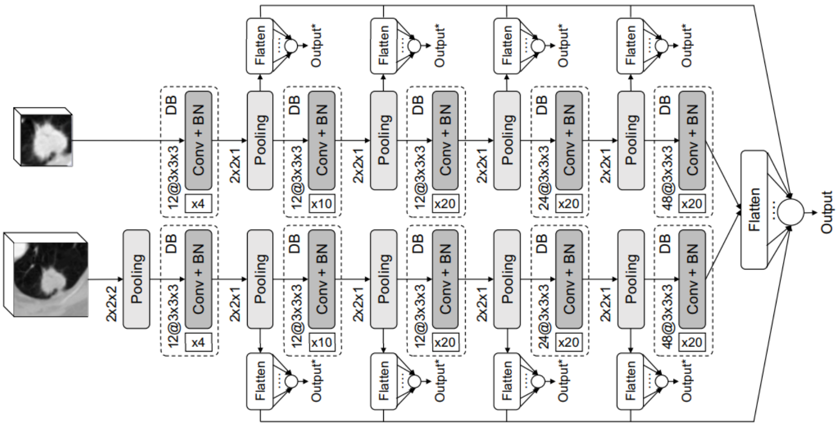


Figure 3.2: Architecture of the proposed network. Adapted from [15].

3.2 Multimodal approaches to medical imaging classification tasks

In a medical setting, a physician will make a medical decision on a given patient based on various and diverse types of information. In the specific case of lung cancer, a physician often has access to not only medical imaging but also clinical information and laboratorial data [11]. Seen as these types of information can hold relevancy in the diagnosis or management of the disease, pattern recognition algorithms can be designed to take advantage of all these types of information.

In the literature, the majority of approaches that can be found for classification of lung nodules analyse only medical imaging and just a few attempt some level of data fusion. One of the reasons that might explain the lack of multimodal approaches in the literature for this end might reside with the fact that most publicly available lung cancer medical imaging datasets do not provide much information besides the images themselves. Nevertheless, datasets related to other medical challenges exist which offer medical images as well as other different types of data, allowing for investigators to develop approaches that are able to combine all of this information and use it for a predictive task.

Huang et al. [36] investigated different types of multimodal data fusion aiming at detecting pulmonary embolism in CT images. The dataset used for the experiments included 1837 Computed Tomography Pulmonary Angiography (CTPA) studies as well as the correspondent patient's electronic medical records containing several different types of information such as demographics, recorded vitals and laboratorial data. For analysing CT images, a 3D CNN previously designed by Huang et al [37], referred to as PENet was utilized consisting of 3D convolutions with skip connections and squeeze-and-excitation blocks. The model was also pretrained with the Kinetics-600 video dataset [38]. In the training process, a sliding window made up of a defined number of CT slices was used, instead of feeding the all CT scan, thereby increasing the proportion of a possible pulmonary embolism to the total input. In the case of the electronic medical records features, an ElasticNet [39] was used, being hypothesized that a neural network would have difficulty learning meaningful data representations due to the sparsity of the data. In the study, 7 multimodal fusion techniques were proposed, being them Early Fusion, Joint All Fusion, Joint Separate Fusion, Late NN Average, Late Elastic Average, Late Separate Average and Late Meta. In early fusion architectures, features are joined at the input level, before being fed to the model, while late fusion involves the combination of prediction probabilities of different models, ultimately reaching a final prediction. The specific architectures of each fusion technique can be seen in Figure 3.3. Besides

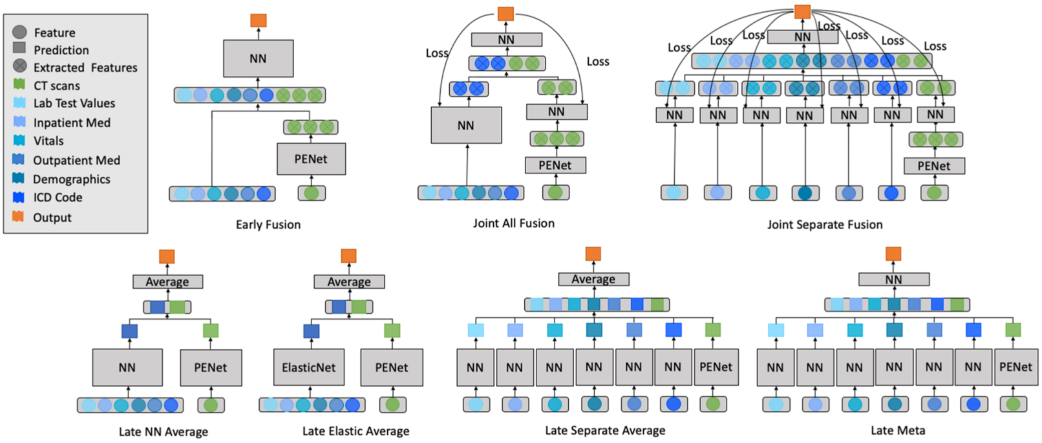


Figure 3.3: Architectures of each fusion technique. Adapted from [36]

the comparison between the different multimodal strategies, 2 single modality models were also used and compared, one with image-only input and another with only electronic medical records data input. The best performing model in the experiments conducted was the Late Elastic Average model achieving the highest test AUC of 0.947.

Wu et al. [40] proposed a multimodal deep learning method for predicting non-small cell lung cancer survival. The proposed approach leveraged CT image analysis with clinical data to produce a survival time estimate. A 3D ResNet [32] is employed for feature extraction of the CT images and a more simple 2 hidden layer neural network is used to extract features from clinical records. The features resulting from extraction of each modality are combined and fed to a shallow neural network comprised of one hidden layer. With the final classification output being a risk score, the evaluating metric used was the Harrell's C-index, obtaining a final value of 0.6580.

Li et al. [41] investigated the effect of combining 3D CNN extracted features with a set of imaging features referred to as handcrafted features (HF) based on intensity, geometric and texture information, in lung nodule malignancy prediction. Three 3D CNN based on the AlexNet [42], VGG-16 [31] and Multi-crop Net [43] were designed to extract the CNN features and, after fusion with the HF, a support vector machine (SVM) [44] was used to classify the lung nodules. In the feature fusion step, the study proposes to combine the HF with output score of the CNN, instead of the more traditional fusion approach of using the CNN representations learned before the final classification layer, stating that the representation learned at the output layer of the CNN is a higher level more abstract than the one learned at the final hidden fully connected layer. After the fusion, an optimal subset of features was selected using sequential forward feature selection (SFS) [45].

Figure 3.4 translates the overall pipeline of the approach.

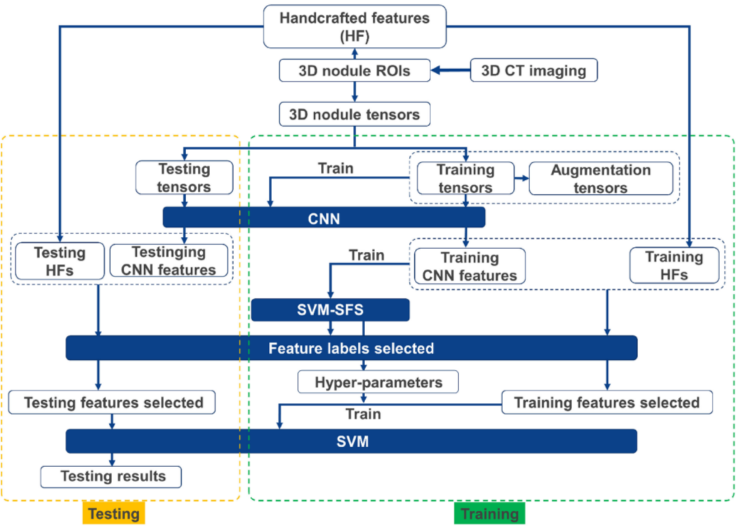


Figure 3.4: The overall pipeline of the approach by Li et al.. Adapted from [41]

For comparison, all the networks were trained and tested on one hand with data fusion, and on the other with just the features resulting from the CNN extraction. The experimental results of the study showed that all the different networks performed better when employing the data fusion approach with the network based on the Multi-crop Net obtaining the steadier results with 0.8260 and 0.9182 for sensitivity and specificity, respectively.

3.3 Summary

Analysing the different studies outlined in this chapter, some conclusions may be drawn. Firstly, the use of deep learning for tackling lung cancer classification tasks has shown promising results. Additionally, the majority of the literature found employs recognizable architectures such as ResNet [32] or VGG-Net [31] due to their high performances in various image classification tasks. The second conclusion that is drawn is that deep learning models that integrate data from different modalities have been shown to perform better than single-modality models in some medical applications, suggesting that multimodal models may be advantageous for introduction in a clinical workflow and motivating research and development of such models. Furthermore, as shown by Huang et al. [36], the data fusion strategy used can have a large influence on the models performance in a specific task and so, different strategies should be explored when developing a multimodal approach.

Chapter 4

Data Overview

The present chapter, firstly, provides an overview of the data utilized in model development and, secondly, describes the processing steps applied to the medical images in order to convert them into a structured format, necessary for the the model's learning process. The chapter is, therefore, divided into two sections: Section 4.1 which offers a data description and Section 4.2 that focuses on data processing.

4.1 Datasets

4.1.1 LIDC-IDRI

The Lung Image Database Consortium image collection (LIDC-IDRI) [34] is composed of lung cancer diagnostic and screening thoracic CT scans. The dataset, garnered by the National Cancer Institute (NCI), is publicly available specifically for enabling the development of lung cancer detection and diagnostic digital tools. The dataset contains 1018 scans available,as well as annotated lesions taken, in a two-phase image annotation process, by four experienced radiologists who aggregated the lesions into 3 particular categories: (1) non-nodule ≥ 3 mm with non-nodules being defined as any other pulmonary lesion that does not possess characteristics consistent with those of a nodule; (2) nodule < 3mm being included all sizeable nodules that are not clearly benign; (3) nodule $\geq$ 3mm containing all nodules with the necessary size regardless of presumed histology. As of result of this annotation process, 7371 nodules were segmented of which 2669 were labelled as nodules $\geq$ 3mm. Included in the annotations of each nodule is an independent assessment of each radiologist on the characteristics of the nodules, one of which being the likelihood of it being a malignant nodule. Additionally, with the segmented lesions are provide nodule masks, that allow for removal of background esterior to the nodule. The LIDC-IDRI is the largest publicly available annotated database on thoracic CT scans [46] and, consequently, many of the research on pattern recognition models applied to lung cancer utilize this database [34, 14, 15].

4.1.2 NLST

The National Lung Screening Trial (NLST) [47] was a medical trial organized by the Lung Screening Study group and the American College of Radiology Imaging Network, aimed at studying the effect on mortality rates of screening for lung cancer with CT scanning when compared to screening with chest radiography, in high-risk individuals. The study was conducted along multiple screening centers and included 53,452 participants divided across two study arms, one where the patients were scanned using CT and another where they were scanned with chest radiography. The patient study arm assignment was a semi-random process since it looked to maintain equal proportionality in gender and age across both groups of patients.

As previously stated, the study intended to screen high risk individuals and, thus, there were eligibility criteria that the participants had to meet for study entry. The participants needed to be age 55 to 74 and to have had a smoking history of at least 30 or more cigarette packs per year inside the prior 15 years to the date of the patient's entrance in the study.

As part of the study, every patient performed three exams (T0, T1 and T2) at one-year intervals, with the first exam (T0) being performed soon after the patient's entry in the study. With each CT scan, radiologists at the screening centers reviewed the images and performed 2 distinct analysis:

- **Isolation read** — in this analysis the radiologist reviewed the CT scan and annotated information on all visible abnormalities, specifying a preliminary screening result for the exam. If either a non-calcified nodule $\geq$ 4 mm or an abnormality consistent with lung cancer were identified, then the exam received a positive screening result. A negative screening result meant that either just minor, clinically irrelevant or no significant abnormalities at all were identified.

- **Comparison read** — following the isolation read, in this analysis, radiologists compared the findings of each exam with the information recorded in previous year exams. Information on previously missed abnormalities was recorded as well as on the evolution of previously identified nodules. With the same screening criteria as the isolation read, the radiologist classified the exam as positive or negative, with this classification being considered the final screening result. Effectively, the comparison read was only feasible in the T1 and T2 exams and so, for the T0, the isolation read's screening result would be the final result of the scan.

Participants with a positive screening result were advised to pursue diagnostic evaluation and, in case of a confirmed diagnosis of lung cancer, information on cancer characteristics was recorded and included in the study's database. A participant with a confirmed diagnosis of lung cancer did not complete any more exams within the study. Therefore, if a participant was diagnosed with lung cancer following a positive screening at the T0 exam, then the T1 and T2 exams did not take place. In the CT arm of the study, there were 720 patients with a confirmed diagnosis of lung cancer within the study years 0, 1 or 2.

Along with the abnormalities and lung cancer annotations, the dataset contains other types of information for each participant including demographic like age, gender and race as well as medically relevant history such as smoking history, passive smoke exposure and family history of lung cancer.

4.2 Data Pre-processing

4.2.1 DICOM conversion to image array

The current standard format of storing and transmitting medical images is Digital Imaging and Communications in Medicine (DICOM), and so, both datasets referred to in the previous section contain their CT scans in this format. Although this format may be advantageous for medical systems, it is necessary to convert it into an image array for this data to be used in a predictive model. For that end, the *pydicom* package, more specifically, the *PixelArray* attribute of this package was used, transforming the various *.dcm* files (corresponding to the CT slices) into the image array. After this, the resulting pixel values of the image array were converted to the Hounsfield Unit scale using two other attributes from the *pydicom* package: $RescaleIntercept$ and $RescaleSlope$.

4.2.2 Resampling

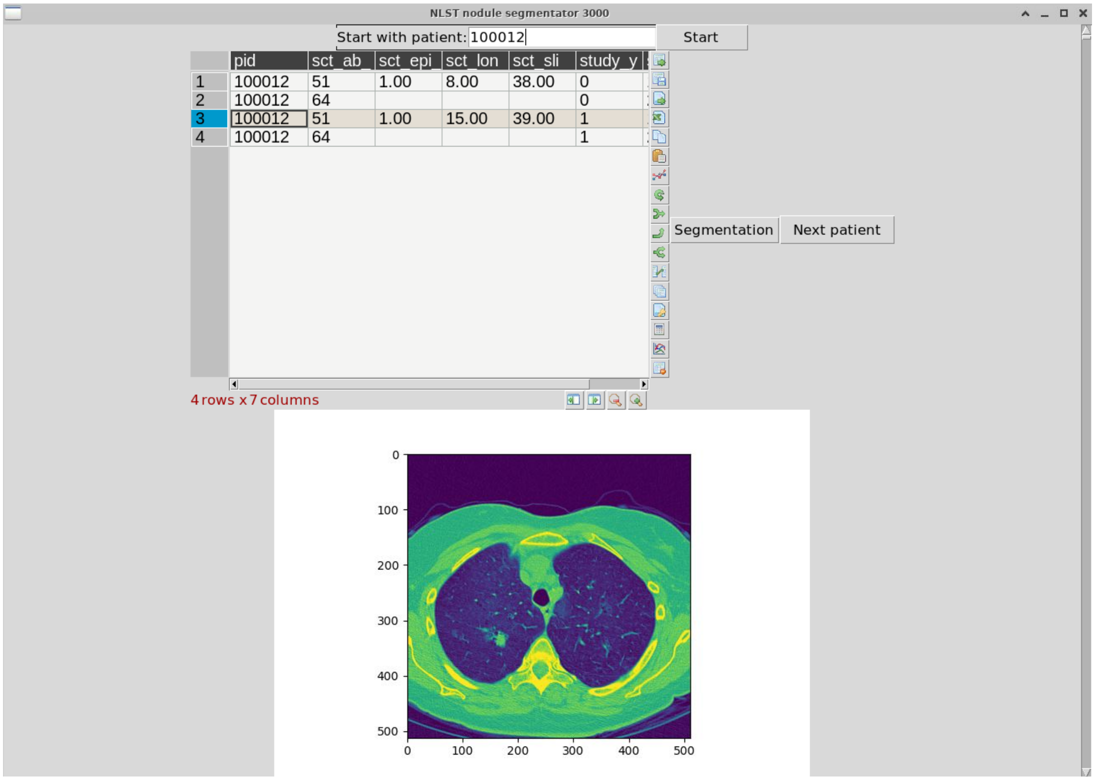
CT scans can be performed under very different scanning protocols which have a direct impact on the CT scan characteristics. The CT scans from the used datasets were obtained from different scanners and as so present dissimilar values of slice thickness and pixel spacing. for example the slice thickness of the CT scans in LIDC-IDRI dataset can range between 0.625 to 3 mm from scan to scan. Applying data with variablity in its characteristics to a pattern recognition algorithm may cause it to learn unwanted relations in the input data. To avoid this, it is common for the dataset to be standardized through a process called resampling, where the spaces between pixels are set to a common value, typically 1 mm. Resampling was achived by firstly calculating a resize factor based on the target CT pixel spacing (in this case 1 mm) as well as the CT original pixel spacing and overall dimensions. The resize factor is then applied to the CT original pixel spacing producing the target 1 mm spacing. The final image is obtained by applying the function *zoom* from the *scipy* package, a method based on spline interpolation.

4.2.3 CT windowing

As described in Section 2.2, adjusting the brightness of the image by applying a grey-level mapping to the CT scan helps to enhance the contrast between structures with similar HU values. Thus, a window was designed with a minimum value of -1000 HU correspondent to air's radiodensity and with maximum value of 400 HU, above which are the radiodensitys of hard tissues such as bone. Additionaly, the CT scans HU values in that window were adjusted into a [0,1] range, through a min-max normalization step.

4.3 Nodule Segmentation

For designing a multimodal lung nodule classification model, a dataset that comprises both nodule annotations and clinical data was required. The NLST dataset offers extensive clinical data but does not have segmented nodules as the LIDC-IDRI dataset. To solve this issue, a system for segmenting 3D nodule volumes was developed. As referred in the previous Section , the NLST database contains a record of all abnormalities identified by radiologists in the CT scans. This records contain the information of the CT slice in which the abnormalities biggest diameter resides. These record were crossed with the records of patients with a confirmed case of lung cancer for segmenting cancer nodules. The number of patients and nodules to be checked made this a very time-consuming process and so, a computational tool was developed to automatize this process to the fullest extent possible. Figure 4.1 shows the interface of application developed.



	pid	sct_ab_	sct_epi_	sct_lon	sct_sli	study_y
1	100012	51	1.00	8.00	38.00	0
2	100012	64				0
3	100012	51	1.00	15.00	39.00	1
4	100012	64				1

4 rows x 7 columns

Figure 4.1: Graphical interface of the application developed for the manual segmentation of nodules from the NLST dataset

The upper mentioned crossing of NLST records would lead to a patient's CT slice where the cancer nodule was most visible. Segmentation was then performed through visual observation of the slice since no nodule coordinates were available in this dataset. This visual observation was the biggest limitation of this process since it can be prone to observational error. In the example shown in Figure 4.1 is fairly visible where the nodule is located, nevertheless, in a few cases, pinpointing the nodule proved to be a difficult task. Therefore, caution was applied in this process and in a great number of cases no nodule was segmented due to uncertainty. Ultimately, 1029 nodules were manually segmented.

Chapter 5

Lung Cancer Classification

The present Chapter details the experiments conducted for classification of lung cancer. The architecture of the model used for this end is described in Section 5.1. The first experiment, depicted in Section 5.2, focused on classifying whole lung volumes as benign or malignant. In the second experiment, covered by Section 6.2, an approach at lung nodule classification is developed based only on deep features extracted from nodules segmented from CT scans. Section 5.4 details the approach at lung nodule classification based on the fusion of deep features with clinical features

5.1 Proposed Model

Since the input data is three-dimensional, the proposed architecture for the model used in this work is the 3D ResNet. Figure 5.1 shows the fundamental architecture of the network's building block, employing its signature shortcut connections. Different ResNet models exist employing the overall shortcut connection principle but differing in the number of layers (e.g. ResNet-18, Resnet-34 and ResNet-50). The motivation behind the use of the ResNet in this work is due to this architecture allowing for the use of deeper networks with inherent regularization and correcting for the vanishing gradient problem. These advantages are important especially when using medical

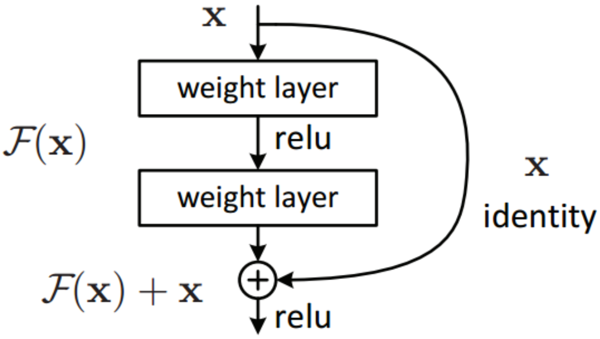


Figure 5.1: Building block of ResNet featuring the signature shortcut connection in which the output of one layer is added to the output of every other layer. Adapted from [32]

imaging datasets since some of them are limited in size. Different variants of the ResNet were also tested to understand the impact of increasingly deep networks.

Taking a look at the 3D ResNet-34 in Figure 5.2, this network is composed by an initial convolutional layer with 64 filters, 7 × 7 × 7 kernel and a stride of 2, followed by batch normalization [48], ReLU activation, and a max pooling layer, with a pool size of 3 and a stride of 2. The output from the max pooling layer is fed into a chain of 16 residual blocks (3 with 64 filters, 4 with 128 filters, 6 with 256 filters, and 3 with 523 filters), each consisting of 2 convolutional layers with batch normalization and ReLU activation. The output from the last residual block is fed to an adaptive average pooling layer, flattened, passed to a fully connected layer, a dropout layer and finnaly relayed to a sigmoid layer producing a probability value in the 0 to 1 range.

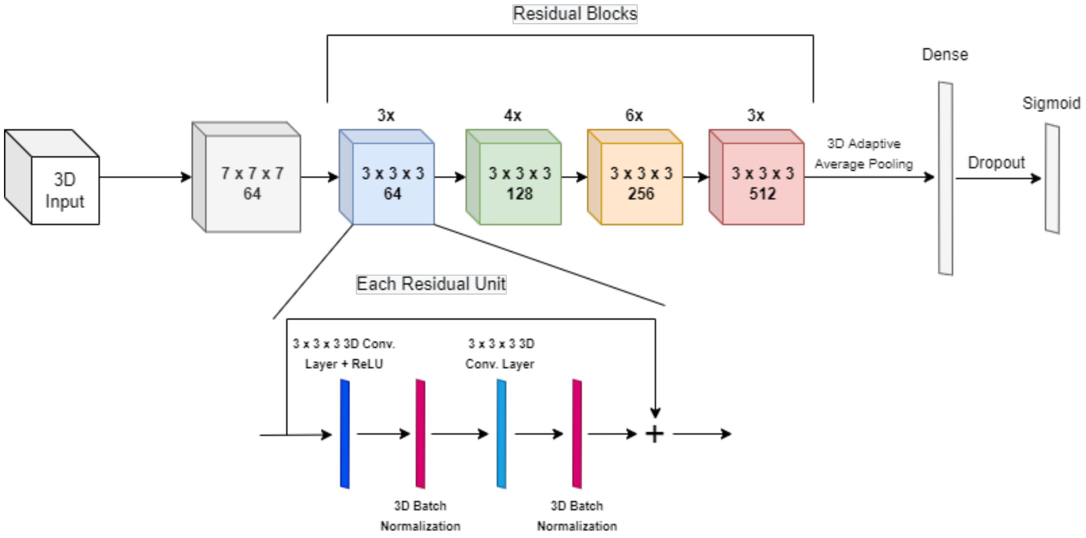


Figure 5.2: Architecture of the 3D ResNet-34.

5.2 Lung classification - a holistic approach

Some studies have demonstrated that information relevant for lung cancer not only resides at the local nodule level but may also be present in areas external to the nodule [49, 50]. With this idea in mind, a first approach at lung cancer classification was devised, aiming at classifying whole lung CT scans as malignant or benign. The intent of this approach was to study if a deep learning model could extract meaningful information from the whole lung CT scan and use it to accurately detect lung cancer.

5.2.1 Data

For this experiment, the NLST database [47] was used, from which 1300 CT scans in total were collected, 700 benign exams and 600 malignant. With the different steps of data pre-processing described in Section 4.2, the final CT scan volume shape had a fixed value of height and width

(512 x 512) but variable depth, meaning different number of CT slices. To achieve input shape uniformity and reduce computational costs, the volumes were reshaped to a final shape of (128 x 128 x 64) voxels by using the same strategy of spline interpolation as referred in Section 4.2.2.

Through the lung cancer records in the NLST database, it was possible to discriminate which lung contained the tumor in patients with a confirmed diagnosis. With this information, in the CT scan volume, the lung was selected for each patient by cropping a portion of the volume containing the affected lung. For patients with no cancer diagnosis, the lung was chosen randomly while keeping the proportionality between left and right lungs selected in both malignant and benign patient pools. This selection was done to further reduce the computational costs and also possibly improve the network's pattern recognition by increasing the ratio of relevant tissue to the total input volume.

5.2.2 Training

Thorough experimentation, the ResNet variant chosen for this task was the 3D ResNet-34. The experiments were conducted targeting the classification of a whole lung, with the model outputing a probability of a given exam being malignant. Binary Cross-Entropy (BCE) was the loss function selected to be minimized through the model's training. Weight initialization was done through the "Kaiming uniform" method [51]. Besides dropout and batch normalization, early stopping based on validation error was employed to avoid overfitting. AdditionallyA set of hyper-parameters seen in Table 5.1 were selected for manual search in order to find the best combination for the model.

Table 5.1: Search space of hyper-parameters used in the manual search for the lung classification model

Hyper-parameter	Search space
Learning Rate	0.1, 0.01, 0.001, 0.0001
Optimizer	SGD
Dropout	0.2, 0.3, 0.4, 0.5
Batch-size	8, 16, 32

Data was randomly split into a training set (70%), validation set (15%) and testing set (15%). Additionally, to ensure that the results are not influenced by a favorable split of the data, 20 random train/test splits were done, with the final results being the average of all data split results.

5.3 Lung Nodule Classification - image only models

Lung nodule classification using only the CT scans data was conducted to establish a baseline for comparison with a multimodal approach at the same task. Since the lung nodules dataset planned to be used for the multimodal approach (NLST dataset) was obtained during the work of this thesis, there is no possibility of comparing the baseline results with other related works. Therefore, a first experiment of lung nodule classification using the LIDC-IDRI dataset was conduted as to validate the deep learning framework created for these experiments. The LIDC-IDRI dataset allows for result comparison since it has been abundantly used in much of the work regarding lung nodule classification [52].

5.3.1 Data

5.3.1.1 LIDC-IDRI

As described in Section 6.2 the LIDC-IDRI data collection comprises a series of annotated lesions separated into different categories. The nodules relevant for inclusion in the model's input data are the ones annotated as nodules $\geq 3mm$. For these nodules, a malignancy assessment is available through the XML annotations of the database. That assessment was recorded in the form of an integer value, ranging from 1 to 5, with each value representing a malignancy degree: (1) Highly Unlikely, (2) Moderately Unlikely, (3) Indeterminate, (4) Moderately Suspicious and (5) Highly Suspicious. Furthermore, more than one radiologist assigned a malignancy value to the same nodule, causing different assessments based on the opinion of each expert. Since this experiment intends for binary classification, the malignancy values available for each nodule were averaged and, if the resulting value was ≤ 2.0 then the nodule was considered benign, while if it was ≥ 4.0, the nodule was considered malignant. Any nodule with a malignancy value in between was considered to have an inconclusive assessment and, therefore, was not used. As a result of these conditions, the pool of available patients in the dataset was reduced to 1093 patients with 787 benign and 306 malignant. The annotated lesions were in the form of cubes centered in the nodule with shape 80 x 80 x 80 voxels. Nodule masks were available in the annotation records of this database and, therefore, applied to these volumes.

5.3.1.2 NLST

The process through which the NLST nodules dataset is obtained is detailed in Section 4.3. This dataset is composed of 1029 nodules, from which 655 benign and 374 malignant. The nodule's volumes are centered in the nodule and have a shape of 50 x 50 x 20 voxels. The segmentation strategy that originated this dataset did not retreive nodule masks and, as so, no nodule maks were applied to the volumes.

5.3.2 Training

Data was randomly split into a training set (80%) and testing set (20%). For better evaluating the model, 5 fold cross-validation was employed in the training data, as shown in Figure 5.3. Additionally, to ensure that the results are not influenced by a favorable split of the data, 5 random train/test splits were done, with the final results being the average of each evaluation metric. Additionally, undersampling of the majority class (in this case bening nodules) was tried out in order to tackle the imbalanced nature of the datasets used.

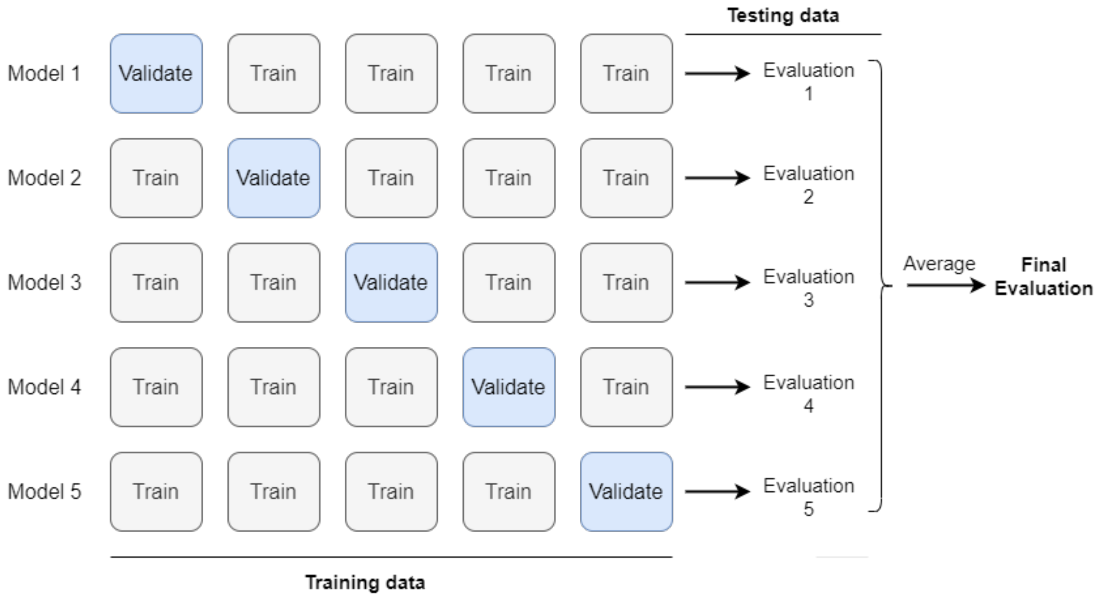

Testing data
Model 1 Validate Train Train Train Train → Evaluation 1
Model 2 Train Validate Train Train Train → Evaluation 2
Model 3 Train Train Validate Train Train → Evaluation 3
Model 4 Train Train Train Validate Train → Evaluation 4
Model 5 Train Train Train Train Validate → Evaluation 5
Average
Final Evaluation
Training data

Figure 5.3: 5-fold Cross-validation training method

The model used for this classification task was the same model (ResNet) referred in Section 5.1, specifically, the 3D ResNet-18. Table 5.2 shows the range of values used to find the best hyper-parameters for the model, through manual search, in both datasets.

Table 5.2: Search space of hyper-parameters used in the manual search for the lung classification model

Hyper-parameter	Search space
Learning Rate	0.01, 0.001, 0.0001, 0.00001
Optimizer	SGD
Dropout	0.2, 0.3, 0.4, 0.5
Batch-size	4, 8, 16, 32

5.4 Lung Nodule Classification - multimodal approach

5.4.1 Biomedical Data Fusion

Data fusion applied to biomedical tasks has the potential to harness complementary information from different data modalities and effectively use it to solve an inference problem. Deep learning models have the ability to extract hierarchical representations of the data, which makes them particularly suitable for developing multimodal approaches [53]. The central issue of developing data fusion models is discovering how best to combine the different modalities to maximize the extraction of data representations, based on complementary information of different types of data.

Data fusion strategies can be distinguished based on the state of the information of different modalities at the fusion moment [54].

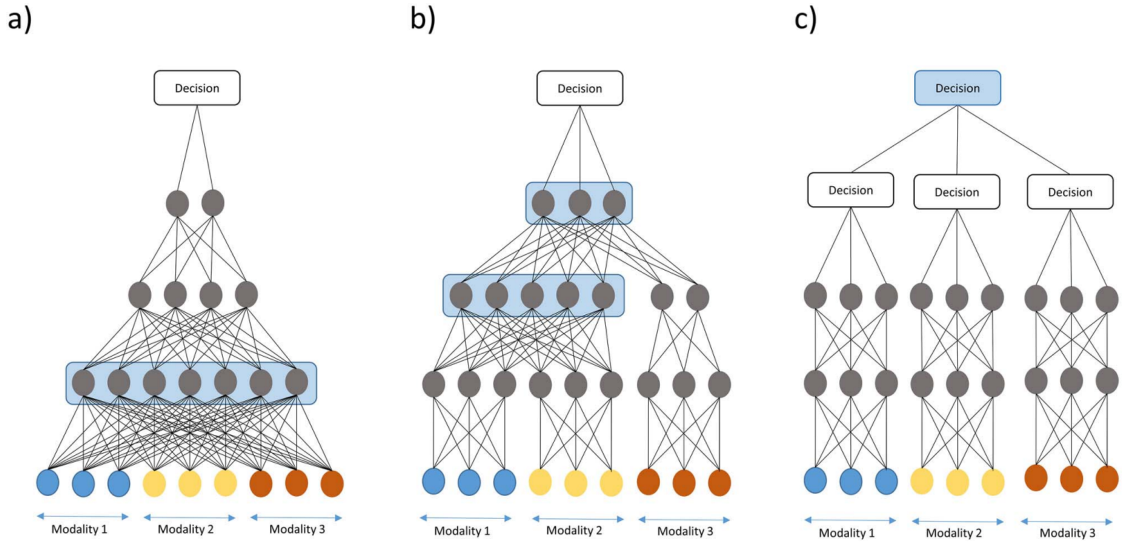


Figure 5.4: Deep-learning based fusion strategies. **(a)** Early fusion with combination of features at the input level. **(b)** Intermediate fusion with combination of data representations **(c)** Late fusion with the combination of single-modality decisions. Adapted from [53]

- **Early fusion** — if the data fusion takes place before the application of a machine-learning algorithm to any of the data modalities, then it is called early fusion. Through this approach, the models can learn cross-modal relationships from low-level features

- **Intermediate fusion** — in this particular strategy, machine learning algorithms transform raw data into higher-level representations. These extracted features can be combined and fed to a model continuing the learning process or outputting a final decision. Through this approach representations of the data can be learned within each modality after which cross-modal relationships between the learned representations can be learned and utilized in the final prediction.

- **Late fusion** — also referred to as decision-level fusion, in this strategy, predictions from the different modalities are aggregated to reach a final prediction. Different methods exist for prediction combination since this type of fusion resembles ensemble classifiers, a widespread and well researched technique.

5.4.2 Data

In this experiments two modalities of data will be used: Nodules extracted from CT scans and clinical data. As the first has been extensively covered already, this section will focus on detailing the clinical data utilized by the model.

The NLST dataset offers a variety of additional patient data ranging from personal medical information (demographics, personal disease history, etc.) to screening and follow-up study details. From all the available information, a group of features were selected based on their perceived importance for the diagnosis of lung cancer. The group of selected features is exposed in Table X. After retrieving the selected features from the dataset, some missing data was verified, specifically in binary features related to personal cancer history and family history of lung cancer. This problem was solved through a data imputation strategy based on inserting a missing value corresponding to the higher represented class in that feature. Furthermore, numeric clinical features were scaled through a min-max normalization technique in order for gradient descent based algorithms (e.g. XGBoost and neural networks) as well as distance-based algorithms (e.g. SVM) to not be affected by different scales during the training process.

Table 5.3: Features selected for clinical data

Demographic	Smoking	Personal cancer History	Family history of lung cancer
Age	Smoked cigar?	Bladder cancer?	Brother with cancer?
Ethnicity	Smoking status at study entrance	Breast cancer?	Child with cancer?
Gender	Smoked pipe?	Cervical cancer?	Father with cancer?
Height	Cigarrete packs per year	Colorectal cancer?	Mother with cancer?
Race	Age at smoking onset	Esophageal cancer?	Sister with cancer?
Weight	Average cigarettes per day	Kidney cancer?	
	Lives with smoker?	Larynx cancer?	
	Works with exposure to smoke?	Lung cancer?	
	Total years of smoking	Nasal cancer?	
		Oral cancer?	
		Pancreaticcancer?	
		Pharynx cancer?	
		Stomach cancer?	
		Thyroid cancer?	
		Transitional cell cancer?	

5.4.3 Data Fusion Strategy

There seems to be no conclusive evidence as to which type of data fusion architecture offers the best results since the variations in performance seem to be very much application dependent [54]. Therefore, it might be advantageous to try different strategies in this work, in order to more thoroughly evaluate the value of a multimodal strategy in the classification of lung nodules.

A preliminary experiment using only clinical data was conducted to evaluate if the information present in the clinical features can be used to predict lung cancer in patients.

In this work, early fusion will not be used since the two modalities utilized (clinical data and nodule cubes extracted from CT scans) are in different formats and, in the case of the nodules, require extraction of deep features. Therefore, only strategies based on intermediate fusion and late fusion were adopted.

Three data fusion strategies, represented in Figure 5.5, were devised to obtain a nodule malignancy classification:

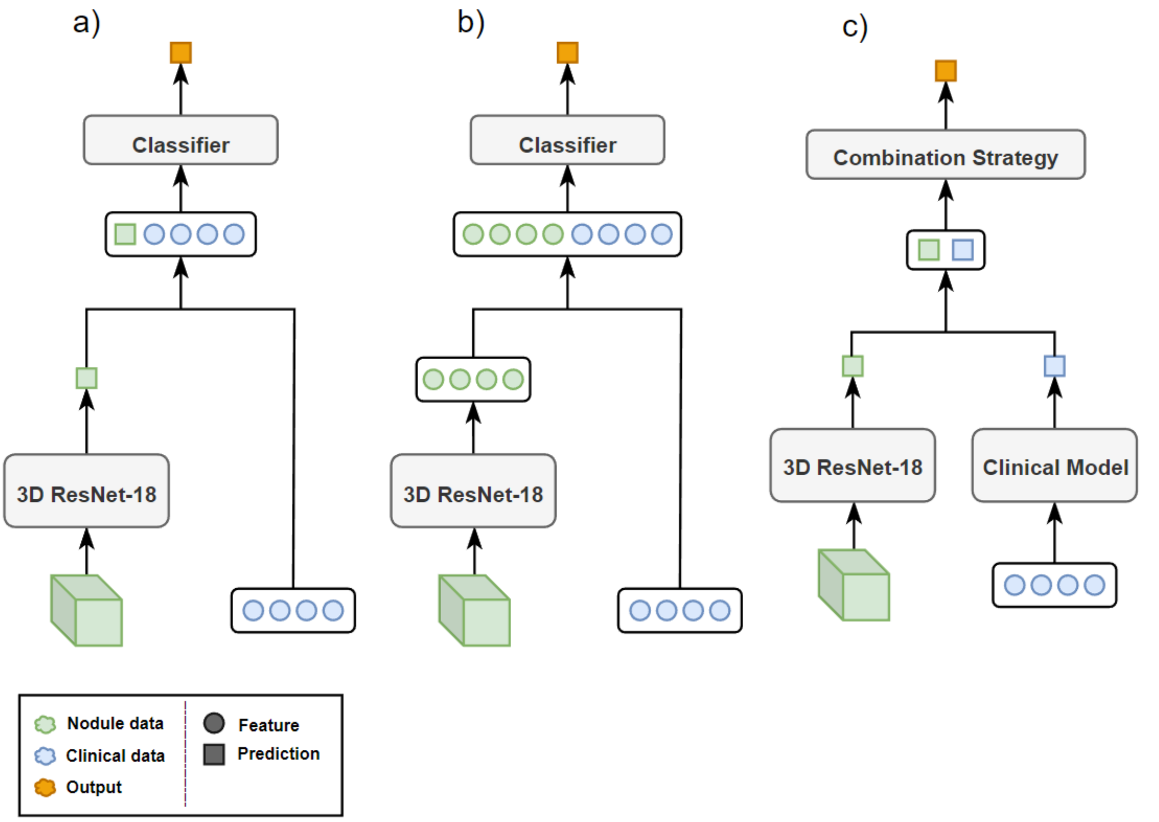


Figure 5.5: Pipelines detailing the multimodal strategies adopted in this work. **(a)** Half Intermediate fusion (HIF). **(b)** Full Intermediate fusion (FIF). **(c)** Late fusion (LF).

- **Half Intermediate fusion (HIF)** — In this strategy, the 3D ResNet outputs a malignancy prediction that is concatenated with the clinical features and fed to a final classifier.

- **Full Intermediate fusion (FIF)** — In this case, the classification layers of the 3D ResNet are removed, with this network outputting a 512 sized vector of deep features. The clinical features are joined with the deep features and fed to a final classifier.
- **Late fusion (LF)** — In this strategy, different models output a prediction for each modality, after which a combination of the two produces the final prediction. The combination strategy utilized was the weighted average of the predictions.

In the intermediate data fusion strategies present in Figure 5.5, the classifier utilized was a fully connected layer followed by a sigmoid layer, outputting a single probabilistic prediciton. The optimal combination of learning rate and optimizer to train this classifier was 0.01 and Adam optimizer respectively.

In the late fusion strategy, the combination strategy will favor the result of the imaging model through a weighted average in which the optimal weight assigned to each prediction will be studied.

5.4.4 Clinical Models - preliminary experiment

Three Machine Learning (ML) models were chosen to be implemented in this experiment: Random Forest (RF), Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). To maximize performance each of the models hyper-paramenters were tuned with Grid Search CV (cross-validation) on the training data. With this technique a search space of hyper-parameters is defined and all possible combinations are evaluated to discover the one that offers the best performance [55].

Table 5.4: Set of hyper-parameter values that were used to find the best clinical models through Grid Search CV.

Algorithm	Hyper-parameters	Search space	Best Values
RF	Number of Estimators	200, 300, 400, 500	400
	Maximum Features	sqrt, log2	sqrt
	Criterion	gini, entropy	entropy
	Maximum Depth	3, 4, 5, 6, 7	4
SVM	C	0.1, 1, 10, 100	1
	Kernel	rbf, poly,	poly
XGBoost	Number of Estimators	10, 100, 1000	10
	Learning Rate	0.001, 0.01, 0.1	0.01
	Subsample	0.6, 0.8, 1	0.8
	Maximum Depth	3, 5, 7, 9	5

Chapter 6

Results and Discussion

Chapter 6 contains the results achieved in the different experiments detailed, as well as an interpretation of those results.

6.1 Lung Cancer Classification - a holistic approach

The hyper-parameters combination that showed the best performance in the test set are depicted in Table 6.1.

Table 6.1: Combination of hyper-parameters that achieved the best results for lung classification.

Hyper-parameter	Best values
ResNet variant	ResNet-34
Learning Rate	0.001
Optimizer	SGD[1]
Dropout	0.3
Batch-size	8

[1] Stochastic Gradient Descent

As a final result, the lung classification model achieved a mean AUC of **0.594** ± **0.038** averaged over 5 train/test random splits.

The results show that the model struggled to recognize meaningful patters in the lung for malignancy detection.

One explanation for the poor performance of the model, in this task, is that the relevant tissue in the lungs, holding the information for malignancy classification, is too small in volume when compared to the total input data. This can make lung cancer hard to detect. Huang et al. [37] faced a similar problem when trying to detect pulmonary embolism in whole CT scans with a deep learning approach. To tackle this issue, the neural network used took in a sliding window with a fixed number of CT slices, instead of the whole CT scan, thus increasing the proportion of the target tissue relative to the total input which ultimately improved results. It might be interesting to explore a similar approach in the future although such an attempt would require data labels for each input of the sliding window, which the dataset used in this work does not provide.

6.2 Lung Nodule Classification - image only models

The combination of hyper-parameters found to produce the best results, for both datasets used, is shown in Table 6.2. The final results presented in Table 6.3 correspond to the computed average of 5 different train/test splits. Undersampling did not produce better results in any of the datasets used for experimentation.

Table 6.2: Search space of hyper-parameters used in the manual search for the lung classification model

Hyper-parameter	Values LIDC-IDRI model	Values NLST model
Learning Rate	0.0001	0.001
Optimizer	SGD[1]	SGD[1]
Dropout	0.3	0.3
Batch-size	8	8

[1] Stochastic Gradient Descent

Table 6.3: Nodule classification results for both datasets

Train and Test Dataset	**AUC (mean ± standard deviation)**
LIDC-IDRI	0.926 ± 0.053
NLST	0.730 ± 0.011

The results found in Table 6.3 suggest that the deep learning approach was successful at recognizing relevant patterns for the nodule classification task. The results also show that nodule classification of the LIDC-IDRI dataset achieved much better results than the NLST dataset, despite these two datasets being similar in the amount of data. A few reasons may explain this disparity in performance. Firstly, unlike the NLST dataset, the LIDC-IDRI dataset annotations contain nodule masks as depicted in Section. These masks can help the model focus the recognition of patterns in more relevant areas, possibly leading to more significant representations of the data being extracted and, consequently, a better performance in the classification task being achieved. One other possible contributor for the lower performance is the fact that the NLST nodule cubes were obtained in a process (mentioned in Section 4.3) prone to human error.

Overfitting was a major concern while training these models. Although, the 3D ResNet-18 is one of the smaller variants of the ResNet architecture, it is still a significantly deep CNN with about

33 million trainable parameters and, coupled with the fact that these networks were trained from scratch, overfitting occurred very early in the training process. Besides the inherent regularization of the ResNet architecture, dropout and early stopping were used to mitigate this phenomenon.

The results obtained by the LIDC-IDRI model are in pair with results obtained in several related works for lung nodule malignancy classification using the same dataset. Dey et al. [15] reached 0.9548 of AUC and 0.9040 of accuracy. Liu et al. [14] reached 0.939 of AUC. Althought the LIDC-IDRI dataset is used frquently for nodule classification tasks, it is important to mention that the malignancy labels obtained for each nodule are the result of averaged subjective opinions of the annotator radiologists, given through visual analysis of the CT scan, as described in Section 5.3.1.1. There is no confirmed pathology diagnosis for any of these nodules and looking through the malignancy scores given to each nodule, there is significant inter-observer variability among radiologists. This lack of diagnosis assurance may result in mislabelled data which can have an unpredictable effect on the performance of the model.

Despite the disparity in performance across the two datasets, the overall goal of this experiment was reached, meaning that, on one hand, the deep learning framework implemented was able to achieve competitive results for a curated dataset (LIDC-IDRI), thus, suggesting that the developed model is capable of recognizing patterns in lung nodules relevant for malignancy prediction. On the other hand, a baseline result for the next experiments was established, allowing for comparison of data fusion strategies with the image-only model developed.

6.3 Clinical Models - preliminary experiment

The set of hyper-parameter values that were used to find the best clinical models can be seen in Table 6.4, as well as the combination of values that achieved the best performance in each of the algorithms chosen.

Table 6.4: Balanced accuracy of the clinical models

Algorithm	**Balanced accuracy (mean ± standard deviation)**
Random Forest	0.595 ± 0.025
SVM	0.602 ± 0.030
XGBoost	0.612 ± 0.029

Table 6.4 presents the results achieved by each algorithm obtained for 30 random train/test splits. Since the prediction of the models was obtained in a binary form, the metric chosen for evaluation of these models was accuracy. Overall the performance of all algorithms was similar with XGBoost having a slight advantage in performance when compared to the other two models.

The results achieved indicate that the utilized data alone is not sufficient to accurately predict the diagnosis of lung cancer. However, one must bear in mind that all the patients included in the NLST dataset have an extensive smoking history, otherwise they could not have entered the study.

With smoking history being highly correlated with lung cancer it is hypothesized that a pool of patients with more balanced smoking habits would offer better results with the clinical models. Also, it might be very difficult to develop models that base a diagnosis decision solely on clinical data. This reinforces the idea of utilizing clinical data as complementary information to be fused with data from an imaging modality.

6.4 Multimodal approaches

6.4.1 Results

The results of this approach can be found in Table 6.5. The combination of weight image factor and clinical model that offered the best performance for this approach was when using Random Forest and 0.7 weight factor for the image model's prediction with a mean AUC of **0.733** ± **0.030**. It is important also to note that the standard deviation of this particular combination of algorithm and weight factor was one of the highest verified in the approach.

Table 6.5: Comparison of LF approach AUC results for different combinations of clinical model and image weight factor.

Image Weight Factor	**RF**	**SVM**	**XGBoost**
0.5	0.708 ± 0.012	0.722 ± 0.004	0.731 ± 0.105
0.6	0.714 ± 0.008	0.728 ± 0.009	0.731 ± 0.061
0.7	**0.733 ± 0.030**	0.729 ± 0.003	0.731 ± 0.003
0.8	0.731 ± 0.008	0.730 ± 0.007	0.730 ± 0.030
0.9	0.731 ± 0.005	0.729 ± 0.005	0.731 ± 0.074

The results of the intermediate data fusion approaches are shown in Table 6.6. The best performing strategy was the FIF approach achieving a mean AUC of **0.755** ± **0.010**.

Table 6.6: AUC results of the Intermediate fusion strategies.

Intermediate Fusion Approach	**AUC (mean ± standard deviation)**
HIF	0.732 ± 0.011
FIF	**0.755 ± 0.010**

6.4.2 Discussion

It is important for a multimodal approach that tackles a biomedical predictive task to include different data fusion strategies, leveraging information from all the available types of data. The justification behind this is the idea that it is difficult to predict the optimal way of harnessing the most useful information out of each modality without effectively trying different methods. This difficulty is exacerbated when combining various modalities for the same prediction. In the case of this work, with only two types of data being combined, (clinical data and nodule imaging data) an expectation of performance can be drawn for each data fusion strategy.

The results of this approach show that no significant improvements were made to the performance reached by the image-only model. One explanation for the lack of improvements of the LF results when compared to the image-only result (AUC = 0.730 ± 0.075) resides with the poor performance of the clinical models, verified in the previous experiment. Furthermore, the predictions of the clinical model were found to be highly concentrated in the [0.35:0.55] range, showing that the model did not have much confidence when making predictions. This might explain the similarity between the results of the three models across the different weight factors since averaging the weakly confident predictions of the clinical model with nodule's model predictions is not expected to have significant influence in the final prediction. The increase in performance when compared to the image-only model was not significant and the standard deviation of this particular combination of algorithm and weight factor was one of the highest verified in the approach.

The overall results of the multimodal approaches showed some increases in performance when compared to the results achieved by the image-only model. The approach that attained the best performance was the FIF strategy achieving a result of **0.755** ± **0.010** averaged over 20 random combinations for train and test patients.

It was expected that the FIF strategy would fare better results than the other strategies proposed. The reasoning behind this expectation is based on the fact that a few studies suggest that correlations between lung cancer histological characteristics and smoking habits of a patient exist [56, 57]. In the FIF approach, the combination of deep features with clinical features allows for the capture of possible underlying biological relationships between the smoking habits of a patient and the abstract representations extracted from the nodule's volumes by the CNN.

The result obtained show that the FIF strategy outperformed the image-only model (AUC = 0.730 ± 0.075) suggesting that a multimodal approach may be capable of improving the results of single-modality approaches.

Chapter 7

Conclusions and Future Work

Lung cancer is the type of cancer with the highest mortality rate in the world. Crucial to improving patient outcomes, and, therefore, reducing the mortality rate of this disease, is not only diagnosing it in an early stage but also develop an effective treatment plan catered to each patient's specific condition. To accomplish this, a proper characterization of the lung tumor is necessary, to which medical imaging presents the distinct advantage of providing 3D information on the nodule's structure that, given its heterogeneity, is highly relevant for medical decision-making. Medical imaging also opens the possibility of developing automated computational tools capable of extracting meaningful patterns from lung tissue images and use them to perform different predictions related to tumor characterization.

This work focused on studying the benefits of leveraging more than one data modality for lung cancer classification using a dataset composed of manually segmented nodule volumes. For that, firstly, a baseline for result comparison was established with an imaging-only model developed, obtaining a mean AUC of 0.730. The multimodal approach devised featured three different strategies aiming at combining information extracted from nodule volumes with patient clinical data. The first of the data fusion strategies was based on late fusion with the result (AUC = 0.733) not showing relevant improvements to the results of the imaging-only model. Two different data fusion strategies based on intermediate fusion were attempted, one of whom based on the combination of clinical features with deep features extracted from the imaging data offered favourable results (AUC = 0.755) outperforming the imaging-only model.

The results obtained in this work suggest that multimodal approaches are deserving of further investigation, with particular emphasis on the study of data fusion strategies that can combine complementary information from different data types in the most effective way for the task at hand. A clear limitation of this work was the lack of diversity of data modalities available, with only two existing for this dataset. Vital signs, medication records and laboratorial data are some of the different types of data that other multimodal works, found in the literature, have used for data combination. A way to further deepen the multimodal study, in this work, would be to extract not only the nodule segmentations but also quantitative data related to the whole 3D volume, based on intensity and gradient distributions, which could potentially offer information relevant for the classification task. Additionally, different data fusion strategies and classifiers could be used to investigate the optimal way of combining the information relevant in each data type.

Bibliography

[1] Hyuna Sung et al. “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[2] Centers for Disease Control and Prevention. *Lung Cancer: What are the risk factors?* URL: https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm. (accessed: 13.04.2022).

[3] Sean Blandin Knight et al. “Progress and prospects of early detection in lung cancer”. In: *Open biology* 7.9 (2017), p. 170070.

[4] All Races, Age-Adjusted Rates, and Age-Specific Rates. “SEER Cancer Statistics Review 1975-2017”. In: (2020).

[5] Meng-Hua Tao. “Epidemiology of lung cancer”. In: *Lung Cancer and Imaging* (2019).

[6] Jonathan Lorenz and Matthew Blum. “Complications of percutaneous chest biopsy”. In: *Seminars in interventional radiology*. Vol. 23. 02. Copyright© 2006 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New ... 2006, pp. 188–193.

[7] Yichen Zhang et al. “Biopsy frequency and complications among lung cancer patients in the United States”. In: *Lung cancer management* 9.4 (2020), LMT40.

[8] Andreas Heindl, Sidra Nawaz, and Yinyin Yuan. “Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology”. In: *Laboratory investigation* 95.4 (2015), pp. 377–384.

[9] José Marrugo-Ramırez, Mònica Mir, and Josep Samitier. “Blood-based cancer biomarkers in liquid biopsy: a promising non-invasive alternative to tissue biopsy”. In: *International journal of molecular sciences* 19.10 (2018), p. 2877.

[10] Rikiya Yamashita et al. “Convolutional neural networks: an overview and application in radiology”. In: *Insights into imaging* 9.4 (2018), pp. 611–629.

[11] Konstantinos Loverdos et al. “Lung nodules: a comprehensive review on current approach and management”. In: *Annals of thoracic medicine* 14.4 (2019), p. 226.

[12] RJM Bruls and RM Kwee. “Workload for radiologists during on-call hours: dramatic increase in the past 15 years”. In: *Insights into imaging* 11.1 (2020), pp. 1–7.

[13] Diego Jaramillo. *Radiologists and Their Noise: Variability in Human Judgment, Fallibility, and Strategies to Improve Accuracy*. 2022.

[14] Hong Liu et al. “Multi-model ensemble learning architecture based on 3D CNN for lung nodule malignancy suspiciousness classification”. In: *Journal of Digital Imaging* 33.5 (2020), pp. 1242–1256.

[15] Raunak Dey, Zhongjie Lu, and Yi Hong. “Diagnostic classification of lung nodules using 3D neural networks”. In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 774–778.

[16] Mohammad Hesam Hesamian et al. “Deep learning techniques for medical image segmentation: achievements and challenges”. In: *Journal of digital imaging* 32.4 (2019), pp. 582–596.

[17] Risheng Wang et al. “Medical image segmentation using deep learning: A survey”. In: *IET Image Processing* 16.5 (2022), pp. 1243–1267.

[18] Hongtao Xie et al. “Automated pulmonary nodule detection in CT images using deep convolutional neural networks”. In: *Pattern Recognition* 85 (2019), pp. 109–119.

[19] Beibei Jiang et al. “Deep learning reconstruction shows better lung nodule detection for ultra–low-dose chest CT”. In: *Radiology* 303.1 (2022), pp. 202–212.

[20] Samir S Yadav and Shivajirao M Jadhav. “Deep convolutional neural network based medical image classification for disease diagnosis”. In: *Journal of Big Data* 6.1 (2019), pp. 1–18.

[21] Onur Ozdemir, Rebecca L Russell, and Andrew A Berlin. “A 3D probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose CT scans”. In: *IEEE transactions on medical imaging* 39.5 (2019), pp. 1419–1429.

[22] Adones Leslie, AJ Jones, and PR Goddard. “The influence of clinical information on the reporting of CT by radiologists.” In: *The British journal of radiology* 73.874 (2000), pp. 1052–1055.

[23] Chelsea Castillo et al. “The effect of clinical information on radiology reporting: A systematic review”. In: *Journal of Medical Radiation Sciences* 68.1 (2021), pp. 60–74.

[24] Lauren G Collins et al. “Lung cancer: diagnosis and management”. In: *American family physician* 75.1 (2007), pp. 56–63.

[25] Anni R Jensen, Jan Mainz, and Jens Overgaard. “Impact of delay on diagnosis and treatment of primary lung cancer”. In: *Acta Oncologica* 41.2 (2002), pp. 147–152.

[26] John Gohagan et al. “Baseline findings of a randomized feasibility trial of lung cancer screening with spiral CT scan vs chest radiograph: the Lung Screening Study of the National Cancer Institute”. In: *Chest* 126.1 (2004), pp. 114–121.

[27] Robert B Daroff and Michael J Aminoff. *Encyclopedia of the neurological sciences*. Academic press, 2014.

[28] King-Hay Yang. *Basic finite element method as applied to injury biomechanics*. Academic Press, 2017.

[29] Diego Riquelme and Moulay A Akhloufi. "Deep learning for lung cancer nodules detection and classification in CT scans". In: *Ai* 1.1 (2020), pp. 28–67.

[30] *Data Science Bowl 2017*. URL: `https://www.kaggle.com/c/data-science-bowl-2017`. (accessed: 13.04.2022).

[31] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[32] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[33] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[34] Samuel G Armato III et al. "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans". In: *Medical physics* 38.2 (2011), pp. 915–931.

[35] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[36] Shih-Cheng Huang et al. "Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection". In: *Scientific reports* 10.1 (2020), pp. 1–9.

[37] Shih-Cheng Huang et al. "PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging". In: *NPJ digital medicine* 3.1 (2020), pp. 1–9.

[38] Joao Carreira et al. "A short note about kinetics-600". In: *arXiv preprint arXiv:1808.01340* (2018).

[39] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

[40] Yujiao Wu et al. "DeepMMSA: A novel multimodal deep learning method for non-small cell lung cancer survival analysis". In: *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2021, pp. 1468–1472.

[41] Shulong Li et al. "Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features". In: *Physics in Medicine & Biology* 64.17 (2019), p. 175012.

[42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[43] Wei Shen et al. "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification". In: *Pattern Recognition* 61 (2017), pp. 663–673.

[44] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[45] A Wayne Whitney. “A direct method of nonparametric measurement selection”. In: *IEEE transactions on computers* 100.9 (1971), pp. 1100–1103.

[46] Lea Marie Pehrson, Michael Bachmann Nielsen, and Carsten Ammitzbøl Lauridsen. “Automatic pulmonary nodule detection applying deep learning or machine learning algorithms to the LIDC-IDRI database: a systematic review”. In: *Diagnostics* 9.1 (2019), p. 29.

[47] National Lung Screening Trial Research Team et al. “The national lung screening trial: overview and study design”. In: *Radiology* 258.1 (2011), p. 243.

[48] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[49] Gil Pinheiro et al. “Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS”. In: *Scientific reports* 10.1 (2020), pp. 1–9.

[50] Francisco Silva et al. “Towards Machine Learning-Aided Lung Cancer Clinical Routines: Approaches and Open Challenges”. In: *Journal of Personalized Medicine* 12.3 (2022), p. 480.

[51] Kaiming He et al. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[52] Shailesh Kumar Thakur, Dhirendra Pratap Singh, and Jaytrilok Choudhary. “Lung cancer identification: a review on detection and classification”. In: *Cancer and Metastasis Reviews* 39.3 (2020), pp. 989–998.

[53] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. “Multimodal deep learning for biomedical data fusion: a review”. In: *Briefings in Bioinformatics* 23.2 (2022), bbab569.

[54] Dhanesh Ramachandram and Graham W Taylor. “Deep multimodal learning: A survey on recent advances and trends”. In: *IEEE signal processing magazine* 34.6 (2017), pp. 96–108.

[55] James Bergstra and Yoshua Bengio. “Random search for hyper-parameter optimization.” In: *Journal of machine learning research* 13.2 (2012).

[56] Michael J Thun et al. “Cigarette smoking and changes in the histopathology of lung cancer”. In: *Journal of the National Cancer Institute* 89.21 (1997), pp. 1580–1586.

[57] Ayesha Bryant and Robert James Cerfolio. “Differences in epidemiology, histology, and survival between cigarette smokers and never-smokers who develop non-small cell lung cancer”. In: *Chest* 132.1 (2007), pp. 185–192.