UAB
Universitat Autònoma de Barcelona

Evaluation of the Reproducibility of Radiomic Intelligent Systems for Early Detection of Lung Cancer

Guillermo Eduardo Torres

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons: https://creativecommons.org/licenses/?lang=ca

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons: https://creativecommons.org/licenses/?lang=es

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license: https://creativecommons.org/licenses/?lang=en

Universitat Autònoma
de Barcelona

Evaluation of the Reproducibility of Radiomic Intelligent Systems for Early Detection of Lung Cancer

A dissertation submitted by **Guillermo Eduardo Torres** to the Universitat Autònoma de Barcelona in fulfilment of the degree of **Doctor of Philosophy** in the Departament de Ciències de la Computació.

Bellaterra, January 6, 2024

Director	**Dra. Débora Gil Resina** Centre de Visió per Computador(CVC) Departament de Ciències de la Computació Universitat Autònoma de Barcelona (UAB)
Co-Director	**Dr. Carles Sanchez Ramos** Centre de Visió per Computador(CVC) Departament de Ciències de la Computació Universitat Autònoma de Barcelona (UAB)
Thesis committee	**Dr. Xavier Baró Solé** Universitat Oberta de Catalunya (UOC) Estudis d'Informàtica, Multimèdia i Telecommunicació **Dra. Aura Hernandez Sabaté** Centre de Visió per Computador(CVC) Departament de Ciències de la Computació Universitat Autònoma de Barcelona (UAB) **Dr. Jose Ibeas** Instituto de Invistigació e Innovació Parc Tauli (I3PT). Parc Taulí Hospital Universitari Universitat Autònoma de Barcelona (UAB)

This document was typeset by the author using LaTeX2ε.

The research described in this book was carried out at the Centre de Visió per Computador, Universitat Autònoma de Barcelona.

Copyright © 2024 by **Guillermo Eduardo Torres**. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the author.

ISBN xxx-xx-xxxxxx-x-x

Printed by Ediciones Gráficas Rey, S.L.

To my mom, dad, and family, your teachings on resilience have been my unwavering anchor through life's challenges.

Acknowledgements

I am sincerely grateful to my advisors, Debora Gil and Carles Sanchez, for their unwavering support and mentorship throughout the entire journey of this thesis. Their exceptional expertise and insightful guidance have played a pivotal role in shaping the direction and quality of this research, and I truly appreciate their patience.

My heartfelt thanks go to my family for their unwavering encouragement and understanding. Their consistent support has been the driving force that motivated me to overcome challenges and ultimately reach the finish line of this academic endeavor.

I extend special thanks to the Computer Vision Center (CVC) and Universitat Autònoma de Barcelona (UAB) for their financial support through the scholarship-PIF and the invaluable human support that has significantly contributed to the success of this thesis. In this regard, I want to highlight the exceptional contribution of Joan Farré Vila.

I express immense gratitude to Aura Hernandez Sabaté, Oriol Ramos Terrades, Alejandro Párraga, and his family, José Elías Yauri Vidalón, Joan Masoliver, Raquel Gomez, and every member of the Interactive and Augmented Modelling (IAM) research group, as well as the students and staff at CVC.

I also take this opportunity to express gratitude to Javier Balladini, Eduardo Grosclaude, and my eternal brother of life, Francisco Guillermo Lopez-Luro. Their continuous support and encouragement have been instrumental in inspiring me to embrace new challenges in both life and studies.

Finally, I express my sincere gratitude to all those who have been instrumental in this journey. I thank Analía Zúniga and Eduardo Torres, my guiding lights. Mom and Dad, your lessons on resilience have been my unwavering anchor through life's challenges. To my siblings, gratitude for instilling the virtues of forgiveness, perseverance, and the significance of education in my path. This dedication stands as a testament to your enduring influence on my life. To my children and partner, who have encouraged me and provided unwavering support in every way, enabling me to write these words today. Your encouragement and backing have been crucial in every sense.

Thank you all for your invaluable contributions and unwavering support.

Abstract

Currently, there is a growing trend in cancer cases, with lung cancer leading in cancer-related deaths and ranking second in new cases, just behind breast cancer. Upon lung cancer detection, patients enter a follow-up circuit within the healthcare system, with the frequency depending on the case, for instance, ranging from check-ups every 3, 6 months, or annually. Early detection of lung cancer is crucial, increasing survival chances, reducing patient anxiety, and alleviating the demand for healthcare resources.

To address research gaps, we created a reliable dataset with cases diagnosed histologically through biopsy, promoting transparency while respecting data confidentiality. Numerous studies using machine learning and deep learning report promising performances in lung cancer research. However, commonly used public datasets lack biopsy diagnoses and rely on visual classification by health experts. This constraint motivated us to create a dataset diagnosed through biopsy, adhering to globally accepted acquisition protocols. We also developed an infrastructure that facilitates multi-center data collection. Our dataset is publicly available, fostering research progress while while ensuring data confidentiality.

We explored strategies to generate representation spaces characterizing lung nodules from computed tomography scans, addressing challenges such as small sample size and data imbalance through dimensionality reduction and feature selection. Deep learning faces challenges in biomedical applications, particularly in screening benign nodules, due to limited annotated data and class imbalance, leading to overfitting.

To address these challenges, we developed a framework to explore the impact of representation spaces through three levels of data splitting in experimental design. It provides insights into model performance, generalization capabilities, and ensures robust evaluation and reproducibility. Additionally, we conducted a statistical analysis of the impact of scanner acquisition parameters.

The experimental results allow us to analyze outcomes at different levels of generalization using cross-validation, varying the experimental unit by slice or nodule and relating various visual representation spaces and found hyperparameters.

Keywords – Lung Cancer, Early Lung Cancer Diagnosis, Features Embedding, Hyperparameter Optimization, Meta Learning, Machine Learning, Deep Learning, Computer Vision, Radiomics, Representation Spaces.

Resumen

Actualmente, hay una creciente tendencia en casos de cáncer, siendo el cáncer de pulmón el líder en muertes relacionadas por cáncer y ocupando el segundo lugar en nuevos casos, justo detrás del cáncer de mama. Tras la detección del cáncer de pulmón, los pacientes ingresan a un circuito de seguimiento dentro del sistema de salud, con una frecuencia que depende de cada caso, por ejemplo, con revisiones cada 3, 6 meses o anuales. La detección temprana del cáncer de pulmón es crucial, aumentando las probabilidades de supervivencia, reduciendo la ansiedad del paciente y aliviando la demanda de los recursos del sistema de salud.

Para abordar las brechas en la investigación, creamos una base de datos confiable con casos diagnosticados histológicamente mediante biopsia, promoviendo la transparencia y respetando la confidencialidad de los datos. Numerosos estudios que utilizan aprendizaje automático y aprendizaje profundo informan de rendimientos prometedores en la investigación del cáncer de pulmón. Sin embargo, las base de datos públicos comúnmente utilizados carecen de diagnósticos por biopsia y dependen de la clasificación visual hecha por expertos en salud. Esta limitación nos motivó a crear una base de datos con casos diagnosticados mediante biopsia, siguiendo un protocolo de adquisición aceptados globalmente. También desarrollamos una infraestructura que facilita la recopilación de datos de múltiples centros. Nuestra base de datos está públicamente disponible, fomentando el progreso en la investigación mientras garantiza la confidencialidad de los datos.

Exploramos estrategias para generar espacios de representación que caracterizan los nódulos pulmonares de las tomografías computarizadas, abordando desafíos como el pequeño tamaño de muestra y el desequilibrio de datos mediante la reducción de dimensionalidad y la selección de características. El aprendizaje profundo enfrenta desafíos en aplicaciones biomédicas, especialmente en la detección de nódulos benignos, debido a la falta de datos anotados y al desequilibrio de clases, lo que lleva al sobreajuste.

Para abordar estos desafíos, desarrollamos un marco para explorar el impacto de los espacios de representación a través de tres niveles de división de datos en el diseño experimental. Proporciona información sobre el rendimiento del modelo, las capacidades de generalización y garantiza una evaluación y reproducibilidad robustas. Además, realizamos un análisis estadístico del impacto de los parámetros de adquisición del escáner.

Los resultados experimentales nos permiten analizar los resultados a diferentes

niveles de generalización mediante validación cruzada, variando la unidad experimental por corte o nódulo y relacionando diversos espacios de representación visual y parámetros encontrados.

Palabras Clave – Cáncer de Pulmón, Detección Precóz de Cancer de Pulmón, Características Embebidas, Optimización de Hiperparámetros, Metaaprendizaje, Aprendizaje automático, Aprendizaje profundo, Visión por computadora, Radiómica, Espacios de representación.

Resum

Actualment, hi ha una creixent tendència en casos de càncer, sent el càncer de pulmó el líder en morts relacionades per càncer i ocupant el segon lloc en nous casos, just darrere del càncer de mama. Després de la detecció del càncer de pulmó, els pacients ingressen a un circuit de seguiment dins del sistema de salut, amb una freqüència que depèn de cada cas, per exemple, amb revisions cada 3, 6 mesos o anuals. La detecció precoç del càncer de pulmó és crucial, augmentant les probabilitats de supervivència, reduint l'ansietat del pacient i alleujant la demanda dels recursos del sistema de salut.

Per a abordar les bretxes en la recerca, creem una base de dades de confiança amb casos diagnosticats *histológicamente mitjançant biòpsia, promovent la transparència i respectant la confidencialitat de les dades. Nombrosos estudis que utilitzen aprenentatge automàtic i aprenentatge profund informen de rendiments prometedors en la recerca del càncer de pulmó. No obstant això, les base de dades públics comunament utilitzats manquen de diagnòstics per biòpsia i depenen de la classificació visual feta per experts en salut. Aquesta limitació ens va motivar a crear una base de dades amb casos diagnosticats mitjançant biòpsia, seguint un protocol d'adquisició acceptats globalment. També desenvolupem una infraestructura que facilita la recopilació de dades de múltiples centres. La nostra base de dades està públicament disponible, fomentant el progrés en la recerca mentre garanteix la confidencialitat de les dades.

Explorem estratègies per a generar espais de representació que caracteritzen els nòduls pulmonars de les tomografies computades, abordant desafiaments com la petita grandària de mostra i el desequilibri de dades mitjançant la reducció de dimensionalitat i la selecció de característiques. L'aprenentatge profund enfronta desafiaments en aplicacions biomèdiques, especialment en la detecció de nòduls benignes, a causa de la falta de dades anotades i al desequilibri de classes, la qual cosa porta al *sobreajuste.

Per a abordar aquests desafiaments, desenvolupem un marc per a explorar l'impacte dels espais de representació a través de tres nivells de divisió de dades en el disseny experimental. Proporciona informació sobre el rendiment del model, les capacitats de generalització i garanteix una avaluació i *reproducibilidad robustes. A més, realitzem una anàlisi estadística de l'impacte dels paràmetres d'adquisició de l'escàner.

Els resultats experimentals ens permeten analitzar els resultats a diferents nivells de generalització mitjançant validació creuada, variant la unitat experimental per cort o nòdul i relacionant diversos espais de representació visual i paràmetres trobats.

Paraules Clau – Càncer de Pulmó, Detecció Precóz de Cancer de Pulmó, Características Embegudes, Optimització de Hiperparámetros, Metaaprendizaje, Aprenentatge automàtic, Aprenentatge profund, Visió per computadora, Radiómica, Espais de representació.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Goal

Cancer is a major public health problem worldwide [76]. According to the World Health Organization's International Agency for Research on Cancer (IARC), reported in GLOBOCAN 2020 [80, 1] that there were around 19.3 million new cases of cancer and nearly 10.0 million cancer-related deaths worldwide. On a global scale, lung cancer is the top cause of cancer-related deaths, causing 1.7 million deaths (18.4%) and ranking second in new cases with 2.2 million cases (11.4%). In any case excluding non-melanoma skin cancers and including people of all ages and both sexes.

Looking at the European Union (EU), data from 2020 in EU-27 countries, provided by Eurostat (European statistics) [62, 25, 2], shows that lung cancer makes up 11.9% of all new cancer diagnoses and 20.4% of cancer-related deaths. This makes lung cancer the fourth most common cancer (after prostate, breast, and colorectal cancers) and the leading cause of cancer death.

Lung cancer is divided into Small Cell Carcinoma (SCC) and Non-Small Cell Carcinoma (NSCC) based on histological and morphological characteristics of the cancer cells [64, 65]. This classification is crucial for treatment decisions, as SCC and NSCC exhibit different behaviors, treatment responses, and prognoses.

Small Cell Carcinoma (SCC) [37, 72] is characterized by small, round cells with a high nucleus-to-cytoplasm ratio. It is highly aggressive and tends to grow rapidly. SCC often metastasizes early and extensively, making it less amenable to surgical intervention. Typically, SCC is centrally located in the lung, near the hilum.

On the other hand, Non-Small Cell Carcinoma (NSCC) [64, 54, 50] includes various subtypes, such as adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. These subtypes have distinct cellular characteristics. NSCC, especially adenocarcinoma (ADC) and squamous cell carcinoma (SCC), tends to grow more slowly

compared to small cell carcinoma. Large cell carcinoma behaves more similarly to small cell carcinoma in terms of aggressiveness. NSCC can occur in different areas of the lung, including the peripheral lung tissue.

The decision to classify lung cancer into SCC and NSCC [69, 71] is pivotal for several reasons. Treatment approaches differ significantly; small cell carcinoma is often treated with chemotherapy, responding well to systemic treatments. Non-small cell carcinomas, depending on the stage and subtype, may be treated with a combination of surgery, radiation therapy, and chemotherapy.

Moreover, the prognosis varies between SCC and NSCC. Small cell carcinoma [60, 9] generally has a poorer prognosis due to its aggressive nature and early metastasis. Non-small cell carcinomas [42, 11], especially when detected at an earlier stage, may have better treatment outcomes.

The clinical management of lung cancer is also influenced by this distinction. It helps in determining the appropriate diagnostic and staging procedures, as well as guiding the selection of targeted therapies. In summary, the division into SCC and NSCC is based on histological and clinical characteristics, playing a crucial role in treatment decisions and prognostic assessments for individuals with lung cancer.

The financial impact is significant, with cancer costs in the EU reaching almost €200 billion in 2018 [41]. These costs cover both direct spending in the healthcare system and other costs related to early death, illness, and informal care.

The National Lung Screening Trial (NLST) [82] and Dutch-Belgian Randomized Lung Cancer Screening Trial (NELSON) [102] have shown that lung cancer screening (LCS) with computed tomography of low dose (CTLD) reduces mortality by 20-25%. However, the average of false positive rate of the radiological diagnosis obtained by visual inspection of scans was 23% of the nodules detected. This inaccuracy meant long follow-up of patients with repetitive computed tomography (CT) or performing an invasive procedure like a biopsy or surgery, which accounted to be futile in 73% of the cases. A reduction of false positives would increase the efficiency of screening for early detection of lung cancer.

The largest screening program in Europe, the NELSON study, introduced volumetry of the nodules in consecutive CT, which meant a significant reduction of the average of false positive rate to 13%. This suggests that the application of radiomics [96] (a recent discipline that extracts a large number of image features correlating to treatment outcome), could represent a critical shift in the reduction of the false positive rate and an improvement of early diagnosis of lung cancer.

Radiomics uses sophisticated image analysis and machine learning tools to obtain quantitative image-based features (signatures) that correlate to final diagnosis and treatment outcome [49]. Radiomics involves the extraction of a large number of quantitative features from medical images, such as computed tomography (CT) scans, magnetic resonance imaging (MRI), or positron emission tomography (PET). These features capture the heterogeneity and characteristics of lung tumors at a microscopic

level, enabling more precise diagnosis and treatment planning.

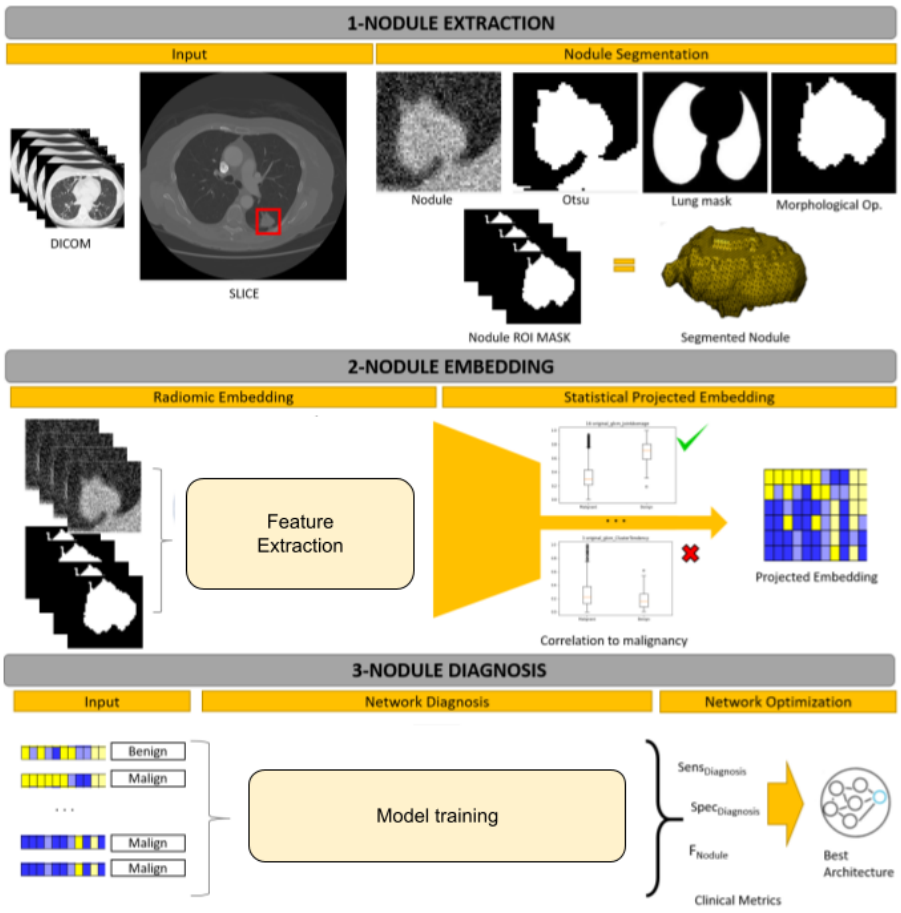


Figure 1.1: Overview of a framework for machine learning models for lung diagnosis.

1. **Nodule extraction**: This phase entails using 3D bounding boxes to extract nodules from CT scans, representing each nodule as a volume of interest (VOI).
2. **Nodule Embedding**: The process of embedding nodules into a representation space for malignancy characterization involves the computation of a Radiomic embedding [88] using either 2D slices from the Volume of Interest (VOI) or the whole 3D VOI.
3. **Nodule Diagnosis**: Distinct classifiers are trained using the features obtained from the preceding embedding phase. In case of a 2D embedding, the output of the classifiers is aggregated using a max-voting approach on the slice predictions in order to obtain the nodule diagnosis.

This PhD Thesis focuses on some aspects of the last two steps, which State-of-Art we review in the next Section.

1.2 State of art and challenges

In a pilot study [58] the authors retrospectively extracted 150 quantitative image features and performed a random forest classification, which finally obtained a significantly better predictive value than volumetry alone (AUC= 0.87 vs 0.74). More recently, Peikert et al. [67] built a radiomic classifier based upon eight quantitative radiologic features selected by the least absolute shrinkage and selection operator (LASSO) method from 726 indeterminate nodules of the LCST. These 8 features include variables capturing location, size, shape descriptors and texture analysis. In this retrospective study, the optimism-corrected AUC for these 8 features was 0.939 with a sensitivity and specificity of, respectively, 90% and 85%.

An alternative to classic radiomics is the use of machine learning methods that extract image features using well known methods such as, Gabor, Local Binary Patterns (LBP), or SIFT descriptor to represent a nodule. Then machine learning techniques (e.g. Support Vector Machine (SVM) and Random Forest) are used to define a classification of nodules in this representation space according to their diagnosis [101, 51]. This methods achieve better diagnostic power than radiomic methods with AUC equal to 0.97, sensitivity equal to 96% with 95% of specificity for [101].

Recently, and motivated by its performance in other areas of application, researchers have began to classify Pumonary Nodule (PN) by using CNNs (Convolutional Neural Network). The early work of Shen et al. proposed to use a multi-crop CNN [75] to make the model robust to scales of nodules while keeping 2D input images. Results showed an overall accuracy (including malign and benign cases) of 87%. However, the authors did not report sensitivity for malignancy detection and specificity for discarding benign nodules and, thus, its true clinical value is uncertain.

Since nodules are 3D structures, some works have addressed the problem using 3D CNNs. Yan et al. [99] explored 3D CNNs for pulmonary nodule classification in comparison to a slice-level 2D CNN and a nodule-level 2D CNN analysis. The 3D approach was the best performer with a 87% of overall accuracy and similar specificity and sensitivity at the cost of a significantly higher demand of computational resources and annotated data. Zhu et al. [103] used 3D deep dual path networks (DPNs) a 3D Faster Regions with Convolutional Neural Net (R-CNN) designed for nodule detection with 3D dual path blocks and a U-net-like encoder-decoder structure to effectively learn nodule features. Despite the complex architecture used, this approach could only achieve a 81% of sensitivity and specificity was not reported. Jiang et al. [45] sequentially deployed a contextual attention module and a spatial attention module to 3D DPN to increase the representation ability. A main novelty of this work is that it ensembles different model variants to improve the prediction robustness. Results show an increase of sensitivity to 90% while keeping a specificity similar to [99].

GLCM (Gray-Level Co-occurrence Matrix) texture features, have demonstrated effectiveness in cancer diagnosis across various medical imaging modalities [44]. In a recent study [87], researchers proposed a hybrid approach that combined GLCM textural features with a neural network for nodule characterization in CT scans. To ensure reproducibility with limited training data, an embedding technique based on the statistical significance of radiomic features was used. This embedded representation served as the input for a neural network, with its architecture and hyperparameters optimized using custom-defined metrics. The best performing model achieved a sensitivity of 100% and specificity of 83% (with an AUC of 0.94) for malignancy detection when evaluated on an independent patient set. This innovative approach shows promise in improving the accuracy and reliability of lung cancer screening by integrating radiomic features and deep learning techniques, offering potential solutions to the challenges posed by false positives in current screening methods.

The output of a classic CNN are features that have no meaning from radiological point of view. In this way, introducing classic radiomic features in the models will be helpful for radiologist in the interpretability of the results by means of most correlated features to malignancy of tumours. It is worth to mention that radiomic features can describe tumour heterogeneity [14], which is a parameter related to malignancy and well known from radiologist.

Like most clinical applications, any radiomic system must deal with Small Sample Size (SSS) and minority classes in possibly unbalanced settings. Most matching learning methods are ill-posed under such conditions and might drop their performance [26]. Dimensionality reduction and automatic feature selection tools have been crucial to mitigate the curse of dimensionality and SSS inherent to classification.

Principal Component Analysis (PCA) is an unsupervised method that uses a linear orthogonal transformation to project features into a dimensionally reduced set of uncorrelated variables called principal components. This technique transforms the data in a reduced dimension but does not perform feature selection. The main problem of this technique is the loss of interpretation of the variables [39]. Partial Least Square - Discriminant Analysis (PLS-DA) PLS-DA is a supervised classification method based on Partial Least Squares Regression and Linear Discriminant Analysis. In this case the technique performs both dimensionality reduction and classification. The dimensionality reduction is similar to PCA but the new components are created by projecting the variables and the outcome into a new space based on linear regression models. The variables of the new reduced subspace called latent variables can predict the outcome and, unlike PCA, there is an interpretation of the projected variables given that the importance of the original variables in the new subspace is quantified [53].

Kernel Trick (KT) has been widely used to extend the above linear methods to the nonlinear case. Kernel methods have aroused great interest in the last decade since they are universal nonlinear approximations and facilitate solving complex problems where the samples are not linearly separable as is the case of many machine learning and pattern recognition application. Kernel methods use nonlinear mapping to project samples from the original space to a feature space where the samples are ex-

pected to be easily separable using linear approaches [32].

A variety of subspace-based kernel methods have been proposed, including Kernel Principal Component Analysis (KPCA) and Kernel Discriminant Analysis (KDA) [100]. A Kernel-Independent Component Analysis (KICA) [57] by using the KT and the Infomax algorithm has also been proposed for enhancing classification, but its application is limited to classes statistically independent. KDA-based approaches are better suited for supervised classification applications since a similar supervision process is performed during the dimensionality reduction, but they require solving an expensive optimization problem [16]. Efficient KDA approaches have been proposed as a solution such as the Kernel Discriminant Analysis via QR decomposition (KDAQR) [97], based on the QR decomposition to replace the costly eigen decomposition of the kernel matrix, and the Kernel Discriminant Analysis by using Spectral Regression (KDASR) [16] which combines spectral graph analysis and regularised regression. Finally, a Discriminative Common Vector with Kernel (KDCV) was originally proposed by Cevikalp et al. [19, 20] and extended (Kernel Generalized Discriminative Common Vectors, KGDCV) to manage large dimensional data in [32].

Methods for dimension reduction in classification problems rely on the probabilistic distribution of samples and, thus, might not be the best suited for SSS in unbalanced settings. Furthermore, the features projected in the reduced spaces are computed following probabilistic considerations and are not easy to be clinically interpreted. In the context of clinical applications (especially in radiomics for personalized medicine), feature selection methods are a preferred choice. Several methods for feature selection are applied in the field of predictive models for personalized medicine.

Random forest selects features according to the change in the classification error. Although it is accurate in case of highly uncorrelated data, in radiomic multi-view problems variables are highly correlated and their selection usually leads to a correlated subset of variables that can produce over-fitting [55]. Besides, the selection of the subsets is random and there is a lot of variability when applying the technique.

To avoid correlation and over-fitting, Minimum Redundancy Maximum Relevance (mRMR) algorithm [70] selects features according to the Mutual Information (MI) between the set of features and the class variable outcome. Features are selected by a threshold on MI which must be carefully adjusted to avoid inclusion of redundant variables or the elimination of clinical relevant ones.

The least absolute shrinkage and selection operator (LASSO) [84] uses a logistic regression model with a penalty term to select features according to their significance in class variable prediction. This method is quite popular for the definition of radiomic signatures [78] and malignancy classification [35]. However, there are some limitations like the low repeatability and reproducibility of textural features in the clinical setting and the limitation of the method to properly modelling SSS unbalanced problems. We consider this could be corrected by the introduction of uncertainty measures into predictive models and the filtering of most unstable data in the training stage.

None of the above methods considers feature reproducibility for their selection.

In the process of clinical data collection, there are several factors prone to introduce variability in multi-view features. Among others, the main ones are medical scans acquisition parameters with an impact on intensity-based values and inter-observer variability in manual annotations required to identify tumors with an impact on shape and volumetric descriptors [4]. Such sources of variability introduce an uncertainty in models that should be considered to issue more reliable reproducible predictions, while avoiding over-fitting.

A recent work [78] adds scan acquisition parameters as fixed factors in a regression model for the development of a radiomic signature that predicts immunotherapy response. However, being unable to select which radiomic features were most affected, the signature reproducibility was low due to over fitting. In [4] it is reported that method reproducibility increases if features are selected based on their stability and reproducibility. While uncertainty modelling is central in statistical analysis (like confidence intervals estimation for computation of hypothesis test significance), in machine learning it has not been addressed until recent years. Latest methods [63] based on fully connected convolutional networks use dropout as boosting method to define a measure of uncertainty in semantic classifiers output that it is used as post-segmentation filtering. However, up to our knowledge a selection of features based on reproducibility and uncertainty remains unexplored.

In [56], we conducted a study with data from Vall d'Hebron Oncology Institute (VHIO) to analyze the reproducibility of radiomics features against different image acquisition conditions and inter-observer variability in lesion identification. The reproducibility study was based on the correlation of feature values obtained from data collected using different conditions and settings. In this study, features were selected as reproducible if they had high inter-class correlation coefficient (ICC) for all sources of variability. The performance of the selected features was compared to state-of-art methods on a different public data base. Results obtained for the classification of lesion malignancy show the better performance of our selection based on reproducibility.

A main limitation of [56] is that it should be replicated for each CT device and new representation space in order to select the most reproducible features. This is not feasible in clinical practice, given that it requires the repeated acquisition of scans of the same lesion with different parameters. This suggests an alternative selection, as well as, studying the set of optimal ranges and values for each scan using single acquisitions.

Regardless of the approach, the classifier has several hyperparameters defining its structure (e.g. neural network architecture) and training process. Such parameters have a strong impact on the performance of the method and they should be optimized using some meta-learning strategy.

Meta-learning, or learning to learn, is the science of systematically observing how different machine learning approaches perform on a given learning task, and then learning from this experience, or meta-data, to learn which approach is the optimal

one for the task [94]. This not only accelerates and improves the design of machine learning or neural architectures pipelines, it also allows us to replace hand-designed algorithms with new learning data-based approaches. One of the most important meta-learning sub fields is automated hyper-parameter optimization (HPO). HPO finds a tuple of hyper-parameters that yields an optimal model which minimizes a predefined loss function on given independent data [23]. HPO has several outcomes: 1) reduce the human effort applying at hand machine learning configurations; 2) improve the performance of machine learning algorithms; and 3) improve the reproducibility and fairness of scientific studies.

In particular, if only a single approach is considered, HPO techniques can be applied to optimize its hyper-parameters. In the case of a neural network, the set of hyper-parameters define its architecture, as well as, its training, which includes back-propagation configuration (like learning rate) and the definition of the loss function weights. Any strategy for HPO should include three main steps: search space, search strategy, and performance estimation.

First, the parameter-search-space (statically constructed by the user) is defined by a tuple of hyper-parameters and its possible discrete value ranges where, the space dimensionality is the number of hyper-parameters, and its search area is bounded by the value ranges. Each tuple of hyper-parameters values define a candidate network to be the optimal one [5][7]. Many approaches have been proposed in order to select a candidate network or apply a search strategy.

Second, search strategies can be grouped into two baseline methodologies: exhaustive and evolutionary searching. In one hand, exhaustive or brute-force search (like grid search or random search [13]) consists of systematically enumerating all possible candidates for the solution and checking whether each candidate satisfies the objective. On the other hand, evolutionary search (like Bayesian [95] or evolutionary genetic [8] algorithms) chooses the new candidates based on the knowledge achieved by previous ones. [6] reports a comparison between these strategies.

Third, regardless of the search strategy used, a critical point is how to define the metrics evaluating the performance and generalization power of the different candidate networks. The simplest approach is to apply a single-objective function over a tuple of hyper-parameters and returns the associated loss [23]. Then k-fold or cross-validation applied to the training set [43] or evaluation to a hold out [22] are often used to estimate this generalization performance [13]. Nowadays the usual approach is to apply a nested cross-validation procedure where hyper-parameter selection is performed in the inner cross-validation, while the outer cross-validation computes an unbiased estimate of the expected accuracy of the algorithm [93].

A main challenge in the application of deep learning to biomedical problems is the limited amount of good quality data with annotations, which is a must for training new models with complex architectures. Besides, in the case of benign nodule screening, this is aggravated with the fact that the problem is highly unbalanced with benign cases being the minority class. Under such experimental settings, models are often over-

fitted [29] results are non-reproducible [26, 29] and most times [99, 103, 45, 99, 46] do not outperform conventional machine learning approaches [101]. Another pitfall, especially for deep methods is models should also be easily interpreted from a clinical point of view to allow the analysis of the clinical factors that have an impact on the clinical decision [61].

Numerous studies use the LIDC-IDRI dataset; however, [74] provides a decisive critique, exposing fundamental flaws linked to its reliance on subjective radiologist assessments for labeling, lacking a robust foundation in pathology reports. This critique questions the credibility of benign and malignant classifications, emphasizing the potential for misdiagnosis in the absence of pathological examination, thereby casting doubt on the dataset's reliability. The critique extends to the dataset's inherent limitations, notably its even distribution of benign and malignant nodules, a scenario that may not reflect real-world occurrences. This uniform distribution could impact the efficacy of machine learning models trained on the data. Additionally, the exclusion of diagnostically challenging samples, especially those with malignancy scores of 3, limits the model's adaptability to nuanced or difficult scenarios in clinical settings, as a malignancy score of 3 signifies diagnostic uncertainty. The paper strongly advocates for a shift towards datasets like LIDP, which meticulously integrates pathology information, offering a more comprehensive and clinically relevant foundation for training and evaluating machine learning models. This critique serves as a rallying call in academia, urging a reconsideration of reliance on LIDC-IDRI and an active embrace of datasets aligning more closely with rigorous clinical standards.

In another study ([48]), the critique continues, highlighting LIDC-IDRI's vulnerability to interobserver variability in nodule annotations. This raises serious concerns regarding the dataset's reliability for training robust lung cancer detection algorithms. The inconsistent characterization of lung nodules by different radiologists introduces challenges in standardization, potentially compromising the accuracy of machine learning models trained on this data. This variability undermines the dataset's credibility as a gold standard for algorithm development in lung cancer diagnosis. Additionally, the study underscores limitations in image quality and resolution due to inconsistencies in the parameters used during image acquisition. Critics argue that the dataset lacks comprehensive representation across demographics and clinical scenarios, thereby limiting its applicability to real-world settings. Despite its historical significance, these critiques serve as a call for researchers to exercise caution and explore supplementary datasets to address the inherent limitations of LIDC-IDRI.

In [74], cases have been collected following an acquisition protocol, taking into account the weaknesses of the LIDC-IDRI dataset. However, the authors have not made their LIDP dataset public, making it impossible for external use. On the other hand, [48] features a federated data architecture with distributed processing, emphasizing security to ensure that data always remains within each region or zone. It is designed for validating pre-trained models. Nevertheless, the challenge lies in the inability to interact between the model and the data during runtime to analyze the model's performance, and the output is restricted to a non-interactive text log. While it's true that

the data in each zone is protected, the model to be validated in this structure is accessible to those maintaining the structure. In other words, the model doesn't have the same level of protection as the data. Currently, this framework is definitively unsuitable for model training. Moreover, if validation is pursued, there is a requirement to furnish the Python code of the model along with its weights.

1.3 Goal and contributions

The goal of this thesis is to improve the early diagnosis of lung cancer. In order to achieve this and following clinical practice, we need an accurate characterization of the nodules, which play a key role in the process. In this context, we contribute to machine learning systems for diagnosis of lung cancer in, both, system's pipeline and acquisition of a dataset for training systems for early diagnosis of lung cancer.

Our main contributions to a system for diagnosis of lung cancer are:

- **Visual Representation Spaces**: we introduce various representation spaces for the characterization of lesions visual appearance in CT scans. These representation spaces include classic radiomics texture features, deep features extracted from the intensity VOIs and a novel combination of deep and radiomic features, which we call deep radiomics. We also present two statistical strategies for the selection of the most meaningful features.
- **Framework for Reproducible HyperParameter Optimization**: we introduce strategies for reproducible optimization and validation of models, focusing on two key aspects. Firstly, a structured examination of data splitting levels is introduced, incorporating Nodule k-folds, Leave-1-Nodule-Out, and Slice k-folds. These strategies provide valuable insights into model performance and generalization capabilities, ensuring a robust evaluation marked by a high degree of generalization and reproducibility. Secondly, we formulate hyperparameter optimization as a multi-objective optimization problem. It employs the Non-Dominated Sorting Genetic Algorithm (NSGA-II) integrated into a Nested Cross-Validation framework, addressing the search space, search strategy, and performance evaluation. This comprehensive approach enhances our understanding of the model optimization process.

Regarding the acquisition of a database for early diagnosis, we contribute in the following aspects:

- **RadioLung Dataset**: we present an own collected dataset for early lung cancer diagnosis, including imaging and clinical data. Our approach involves the development of a precise imaging acquisition protocol. This protocol utilizes Multi-Detector Row CT Scanners with high-resolution features and incorporates a low radiation dose strategy to prioritize patient safety. Notably, our protocol

meticulously considers patient factors during both image capture and reconstruction. Aligned with globally recognized standards in the radiology community, the acquisition protocol is designed to detect lung cancer nodules in their early stages. Subsequently, each identified nodule undergoes histopathological diagnosis through biopsy samples.

- **Impact of CT Acquisition Parameters** : The study presented in [56] showed the impact of device parameters in the reproducibility of radiomic systems. We present the first statistical study on the impact that these parameters have on the performance of methods. We use generalized mixed models [15] to estimate significant difference in the Odds Ratio of failure for different methods and set ranges of acquisition parameters ensuring reproducible results.

This thesis follows a structured progression. In Chapter 1, we introduce Lung Cancer Screening (LCS) and outline our research goal: developing a 1-shot algorithm for simultaneous malignancy detection and histological diagnosis. Moving to Chapter 2, we delve into the representation spaces designed for diagnosis of lung cancer. Chapter 3 explores hyperparameter optimization, utilizing well-defined search spaces and multi-objective functions for thorough performance evaluation. In Chapter 4, we introduce fundamental CT scan concepts and meticulously detail the RadioLung dataset, covering aspects like patient recruitment, imaging protocols, ethical considerations, and comprehensive data management. Chapter 5 details experiments on lung cancer screening, specifying experimental setups for the optimization process and comparing results with existing approaches. Finally, Chapter 6 presents our conclusions and suggests future research directions.

Chapter 2

Representation Spaces for Diagnosis of Lung Cancer

In this chapter we present different embedding strategies for the computation of visual representation of spaces of nodules describing their appearance in CT scans. Nodule embedding has two main steps: 1) extraction of visual features and 2) selection of most discriminant features defining the input for the classifier.

For the extraction of visual features from CT scans we propose 3 different approaches:

1. **Radiomic**. This approach generates a representation space utilizing GLCM textural features. The process involves normalizing the volume of interest (VOI) and extracting GLCM textural features slice by slice. The intensity gray level images are transformed into co-occurrence matrices, from which the GLCM textural features are derived.

2. **Deep Intensity**. Following the normalization of the VOI, the nodule slices are directly fed into a deep pre-trained network, serving as a feature extractor from an internal layer of the network architecture.

3. **Deep Radiomic**. This method combines the aforementioned approaches. Initially, a set of GLCM nodules is extracted from the VOI. Each of these nodules is then individually processed through a deep pre-trained network, extracting features from one of its internal layers to create an deep radiomic embedding.

Once a representation space is generated, we apply t-test that compare the means of the features within the generated representation space. The goal is to identify features that show a statistically significant difference between cases associated with malignancy according the nodule diagnosis. The t-test provides a quantifiable measure of how well a particular feature can discriminate between these two groups (benign and

malignant nodules), helping to select the most correlated features with the malignancy and thus improve the classification tasks.

2.1 Nodule embedding

The feature embedding aims to create representation spaces that holds meaningful and discriminative features of the nodules. These features can be subsequently analyzed and utilized for various classification tasks.

The following sections use nodules extracted from the CT scans, as described in Figure 1.1. Clinicians were asked to adjust the VOI to the lesion. However, previous studies [10, 17] underscore the significance of incorporating both the intranodular region (inside the nodule) and the perinodular region (the area surrounding the nodule) for accurate classification of benign and malignant nodules. To address this, and considering subsequent processing requirements, the size of the 3D bounding box is enlarged to ensure the inclusion of both intranodular and perinodular regions. A crucial consideration in nodule extraction is maintaining the same voxel size, direction, and origin as the original CT scan. Any misinformation in the metadata of the VOI can introduce image distortion or alter the order of data representation.

2.1.1 Intensity-based Representation Spaces

2.1.1.1 Radiomic Space

This representation space is generated utilizing GLCM textural features [40]. Its effectiveness in cancer diagnosis has been demonstrated across a diverse range of medical imaging modalities [85, 44, 52, 68, 81]. To calculate the GLCM features, the node masks are required. In order to segment the nodule, we applied Otsu thresholding to the ROI volume. Since the segmentation of peripheral nodules can include non-pulmonary tissue, the binarized volumes were masked with a segmentation of lungs. The final nodule segmentation was the largest connected component of the masked volumes. The segmentation of lungs was computed using thresholding and morphological operations [38]. Specifically, CT lungs were selected as the larger connected component of the voxels with intensity between 950 to -300 Hounsfield Units, followed by a closing with a structuring element of size 5.

Before normalizing the nodules, it is essential to consider that CT imaging quantifies tissue density using Hounsfield units (HU) [3], a standardized scale. HU quantifies how X-ray beams are attenuated as they pass through different tissues. The calculation involves a linear transformation using specific acquisition parameters, such as slope and intercept, to convert pixel values to HU values. The formula for HU calculation is as follows:

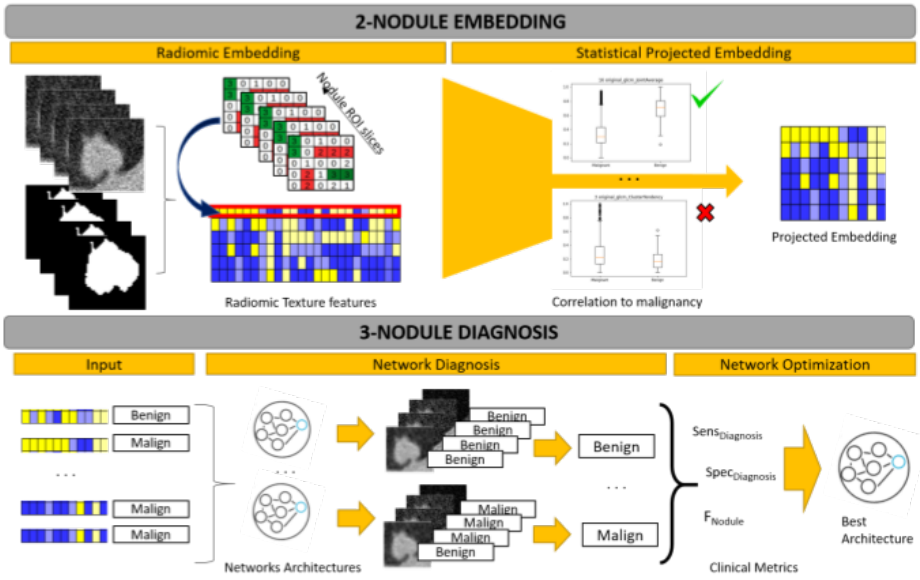


Figure 2.1: Radiomic Embedding Workflow.

$$HU = Pixel\_value \times Slope + Intercept \tag{2.1}$$

This transformation allows for standardized and consistent interpretation of tissue density across different CT scans. The HU scale establishes specific reference points: air is assigned a value of -1000 HU, water is 0 HU, and dense materials like metal (such as steel or silver) can reach values as high as 4000 HU.

Now, for normalizing the VOI to a common intensity range $[0, MaxIntensity]$, the following formula is employed:

$$Pixel\_value = \frac{(HU - Intercept)/Slope}{\max(HU)} * MaxIntensity \tag{2.2}$$

$$= \frac{(HU - Intercept)/Slope}{4000} * MaxIntensity \tag{2.3}$$

where $\max(HU) = 4000$, is the highest HU value, and for $Pixel\_value$ the intensities of the volume. Despite the fact that pixel values are stored as unsigned 16 bits in DICOM, with a range of 0 to 65535, a study conducted by [81] has demonstrated that utilizing only 24, 32 or 64 gray levels is sufficient to extract Gray-Level Co-occurrence

Matrix (GLCM) textural features capable of distinguishing between benign and malignant nodules.

To generate the GLCM features, image intensity is discretized using the histogram of the original volume intensity into n discrete bins. The width of these histogram bins determines the level of granularity at which the GLCM features describe the textural patterns. Smaller bin widths provide a finer level of detail, while larger bin widths result in more generalized information. Once the gray values are discretized, the GLCM is constructed by examining the spatial relationships between pixels within the neighborhood. Specifically, for each pixel, the occurrence of gray-level pairs and their spatial relationships with neighboring pixels are recorded in the co-occurrence matrix. Based on the GLCM, a variety of statistical measures (including contrast, correlation, energy, homogeneity, and many others) are computed to extract textural information. Bin width, namely Δ, is given by:

$$\Delta = \frac{max(Pixel\_value) - min(Pixel\_value)}{Nbins} \tag{2.4}$$

for $Nbins$ the number of histogram bins. In [68, 81] the authors showed the importance of, both, intensity ranges and number of bins. It is reported that a fixed bin count between 30 and 130 bins has good reproducibility and performance.

The GLCM features are extracted by traversing the nodule axially, slice-by-slice, from 2D images, as depicted in Figure 2.2. Afterward, these features are concatenated and analyzed using a t-student test to identify the most strongly correlated features with lesion malignancy. These selected features are intended for later use as inputs to feed a classifier. An overview of workflow followed for the Radiomic Embedding is illustrated in the Figure 2.1.

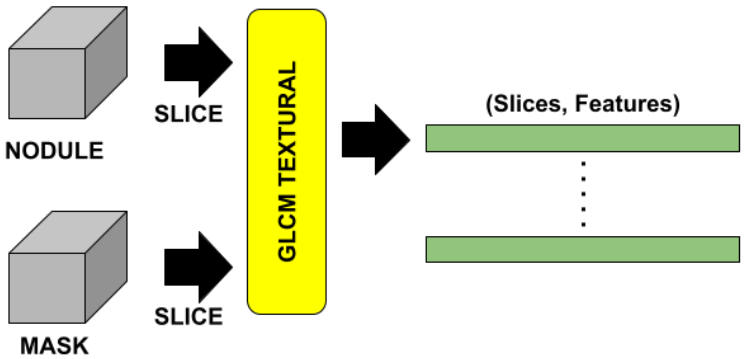


Figure 2.2: Each slice yields a total of GLCM textural features, covering both nodule and mask slices, with (Slices, Features) referring to the number of slices within the nodule and the features extracted from the slices.

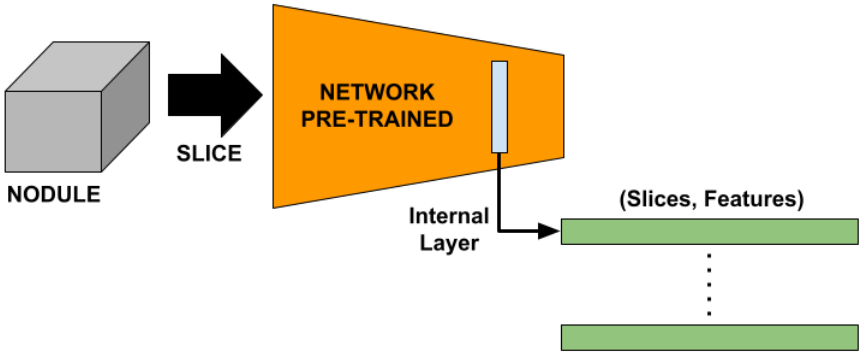


Figure 2.3: The nodule slices are fed through a pre-trained VGG16 network, where features are extracted from the fully connected layer named FC6, resulting in 4096 features per slice.

2.1.1.2 Deep Space

This space representation is generated by extracting deep features from the Gray Level-Intensity (abbreviated as intensity) of the nodules. Numerous studies, such as those highlighted in [86] have demonstrated the efficacy of a pre-trained network in cancer diagnosis. To normalize the nodules, we use the following equation:

$$z = \frac{x - \bar{x}}{s} \tag{2.5}$$

here, x represents the intensity of the nodule, $\bar{x}$ denotes the sample mean of the nodule, and s is the standard deviation of the nodule.

The pre-trained network serves as a robust feature extractor, leveraging its learned weights, exemplified, for instance, in the ImageNet image database [30]. We produce deep embedding features by traversing the nodule slice by slice. When necessary to conform to the network's input shape, each individual slice is replicated N times to match the input layer's channel count. An overview is presented in Figure 2.3.

For example, the generation of deep features using the VGG architecture [77] is illustrated in Figure 2.4. In this scenario, the VGG architecture encompasses 13 convolutional layers, 5 max-pooling layers with a filter size of 2×2, and 2 fully-connected layers. The linear output layer employs the softmax activation function. ReLU activation is applied to all convolutional layers, while dropout regularization is incorporated into the fully connected layers. The deep features for the intensity images are derived from the FC6 layer, generating a vector size of 4096. This layer is the first fully connected layer in the VGG16 model, positioned after the convolutional layers.

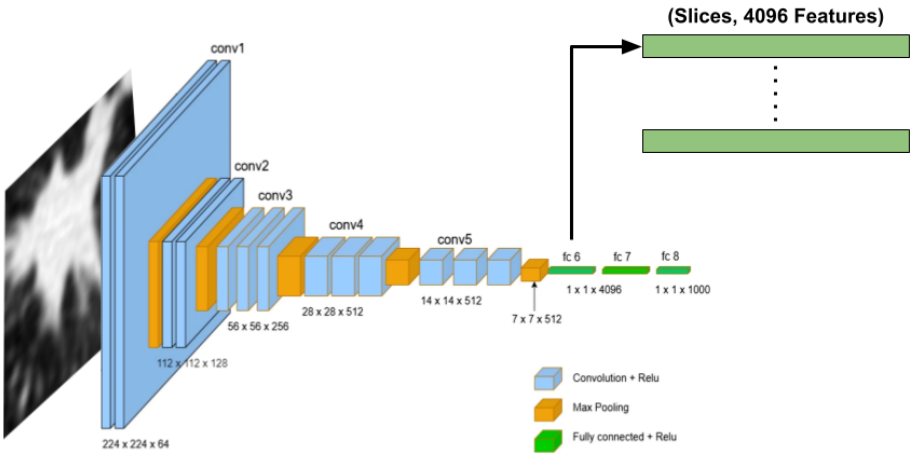


Figure 2.4: Figure depicting the VGG architecture proposed in [77], responsible for receiving slices and generating a 4096-feature vector per slice from the FC6 layer.

2.1.2 Deep Radiomic Representation Spaces

We construct this representation space using GLCM Textural Features, from which we extract deep features using the pre-trained model.

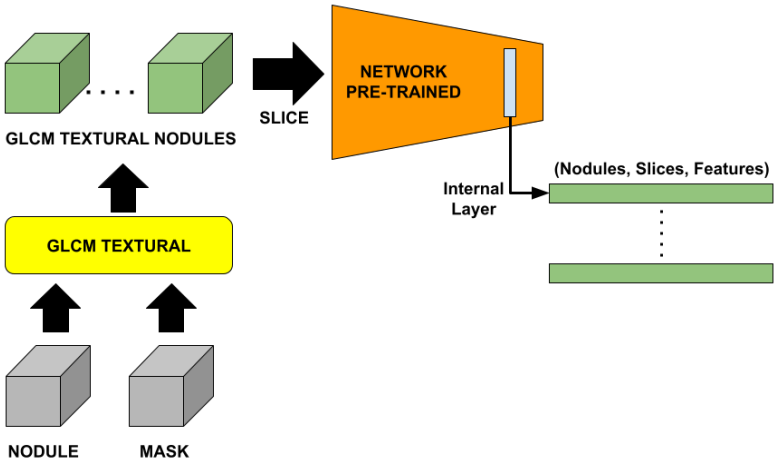


Figure 2.5: GLCM textural nodules are extracted from each nodule and mask. Subsequently, these GLCM nodules are fed into the pre-trained network slice by slice, contributing to the generation of a representation space with dimensions (Nodules, Slices, Features).

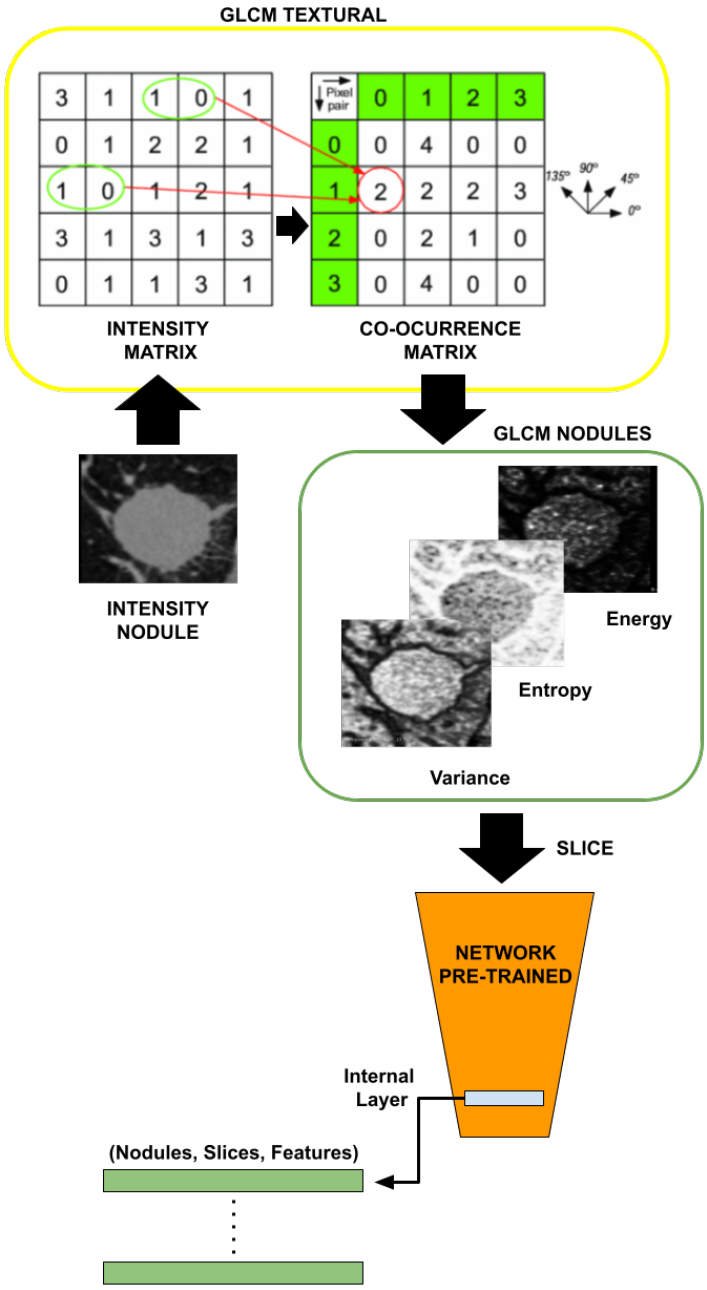


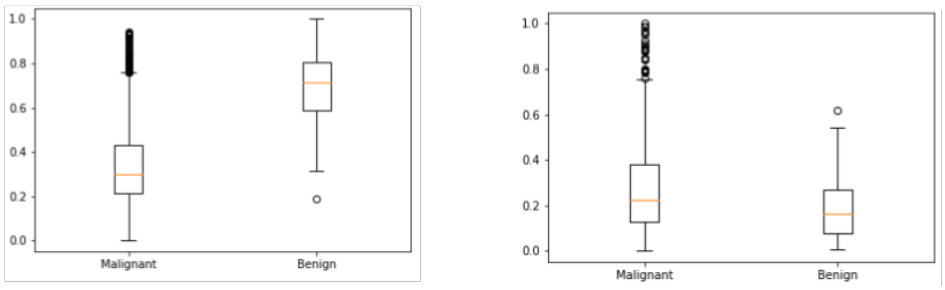
Figure 2.6: Image illustrating GLCM texture extraction, transforming an intensity nodule into GLCM nodules, crucial for generating deep radiomic features via a pre-trained network.

For the normalization of nodules, we employ the defined equations 2.2 and 2.4 outlined in Section 2.1.1.1. An overview of the generation of this representation space is depicted in Figure 2.5. The computation of GLCM textural features [87] involves creating a fictitious mask for each nodule, where all voxel values are set to one. This mask signifies that all voxels within the nodule's Volume of Interest (VOI) are considered when computing GLCM features. Subsequently, a set of GLCM nodules is generated from each nodule and its corresponding mask. Each GLCM nodule is then processed through the pre-trained network slice by slice, resulting in a vector of dimensions $(Nodule, Slice, Features)$. Here, $Nodule$ signifies the quantity of GLCM nodules, $Slice$ denotes the consistent number of slices within the GLCM nodule (equivalent to the original nodule's slice count), and $Features$ represents the features extracted from the intermediate layer of the network.

The detailed process of GLCM texture extraction and GLCM nodules is illustrated in Figure 2.6. Statistical descriptors are computed from a gray-level co-occurrence matrix (GLCM), as describe in Section 2.1.1.1. The intensity nodule is initially represented as an intensity matrix, undergoing transformation into a co-occurrence matrix. This matrix, influenced by the frequency of pixel pairs with specific gray-level values and spatial relationships within a defined neighborhood, facilitates the derivation of a set of GLCM nodules. These GLCM nodules are then fed into the pre-trained network to generate the deep radiomic features.

2.2 Feature Selection

As an alternative to the method [56], we propose to use the distribution of each feature and its correlation to malignancy. Figure 2.7 illustrates the expected distribution of a relevant and non relevant features. For relevant features, boxplots should have minimum overlap, and in particular they should have different means and positive standard deviation.



(a) Joint Average. (b) Cluster Tendency.

Figure 2.7: Boxplots of the distribution of values for a relevant (left) and non-relevant (right) features.

In order to select features with different average values, we use a t-test for each feature. The null hypothesis tests whether the average of the malign and benign cases are equal. Therefore, the p-value of the t-test can be used to rank the features based on their significance in correlating with nodule malignancy. In to particular, select significant features for small p-values rejecting the null hypothesis.

In the case of the radiomic GLCM features, features with a p-value < 0.05 were selected as relevant. This criterion selected 19 of the 24 GLCM features that are identified with a tick in Table 2.1. This subset includes the features selected in [56] according to reproducibility of results and according to the experiments conducted in [87], the whole set has higher performance than the features selected in [56]. In the case of deep features, they are ranked using a p-value obtained from a t-test, measuring the difference in averages between malignant and benign slices. The top N features with the lowest p-values are selected as input for the classifier.

Table 2.1: GLCM textural features chosen through a t-test for Radiomic space.

GLCM Textural Features	**T-test Selection**
Autocorrelation	✓
Cluster Prominence	✓
Cluster Shade	✓
Cluster Tendency	✓
Contrast	X
Correlation	✓
Difference Average	X
Difference Entropy	✓
Difference Variance	X
Inverse Difference	✓
Inverse Difference Moment	✓
Inverse Difference Moment Normalized	X
Informational Measure of Correlation 1	✓
Informational Measure of Correlation 2	✓
Inverse Difference Normalized	X
Inverse Variance	✓
Joint Average	✓
Joint Energy	✓
Joint Entropy	✓
Maximum Probability	✓
Maximal Correlation Coefficient	✓
Sum Average	✓
Sum Entropy	✓
Sum Squares	✓

Chapter 3

Strategies for a Reproducible Optimization and Validation of Models

In the quest to assess the impact of the representation spaces, we introduce a structured examination through three distinct levels of data splitting in our experimental design. These strategies provide valuable insights into the model's performance and generalization capabilities, ensuring a robust evaluation with a high degree of generalization and reproducibility.

The subsequent sections deep into hyperparameter optimization strategies, providing insights into the meta-learning process as a multi-objective optimization problem. The search space, search strategy, and performance evaluation are discussed in detail, paving the way for a comprehensive understanding of the model optimization process.

We conceptualize hyperparameter meta-learning as a multi-objective optimization challenge within the realm of network architectures. This expansive space is parameterized by a set of hyperparameters that collectively define both the architecture and the intricacies of the training process, as described in Section 3.1. The values of these hyperparameters are meticulously optimized through the utilization of a Non-Dominated Sorting Genetic Algorithm (NSGA-II), as explained in Section 3.2, seamlessly integrated into a Nested Cross-Validation (NCV) framework. This integration facilitates the computation of performance metrics that serve as the defining objectives.

The objective functions, in this context, take the form of statistical summaries, specifically the average (μ) and standard deviation (σ), derived from the losses incurred during a k-fold splitting of the training data. In subsequent sub-sections, we delve into the intricacies of the search space, elucidate the search strategy, and expound upon the nuances of the performance evaluation process.

3.1 Search Space

The search space is the comprehensive set of potential candidate solutions that an optimization algorithm explores to find the optimal solution for a given problem. It encompasses all conceivable combinations of hyperparameter values. The definition of a well-structured search space is pivotal for optimization algorithms, significantly influencing the efficiency and effectiveness of the search process.

In our case, the search space comprises various architectures of a specified network model, thus being parameterized by the hyperparameters governing the network's architecture. While some of these hyperparameters may be model-specific, they typically encompass fundamental elements such as the number of layers, neurons per layer, and kernel size (in the case of convolutional models). Beyond the network architecture, certain parameters associated with backpropagation training also impact performance and, thus, they could also be optimized. These can include the learning rate used in backpropagation and the dropout rate within the network. In instances involving multi-task problems, the loss is often determined by a weighted average of individual task losses. Consequently, these weights become another facet subject to optimization to achieve optimal performance.

Hence, the comprehensive search space, denoted as Θ, in our network meta-learning paradigm encompasses not only architectural specifics but also extends to crucial training parameters, forming a holistic optimization landscape defined by the following categories:

$$\Theta = (\text{Architecture Parameters}; \text{Training Parameters}; \text{Loss Parameters}) \tag{3.1}$$

3.2 Search Strategy

In our endeavor, we have opted for the utilization of the Non-Dominated Sorting Genetic Algorithm (NSGA-II) [28] as our preferred optimization algorithm. The primary aim of this exploration is to identify the optimal solution or a collection of Pareto-optimal solutions, with a particular emphasis on the domain of multi-objective optimization.

The NSGA-II is an evolutionary algorithm (EA) based on Genetic Algorithm (GA) designed for solving Multi-objective Optimization Problems (MOOPs). NSGA-II [91] operates on the following main principles:

- *Non-dominated sorting*: It ranks the population members into different Pareto fronts based on their non-domination level.
- *Elite preserving operator*: This operator directly transfers the non-dominated solutions of the current generation into the next generation until other solutions

dominate them.

- *Crowding distance*: It measures the density of the solutions around each solution, representing the average distance of two solutions on either side of a solution along each of the objectives.

- *Selection operator*: This operator selects the population of the next generation. It uses the non-dominant rank and crowding distance to define the next selection criteria. If two solutions have different ranks, the solution with the better (lower in case of minimization) rank is preferred. If both solutions belong to the same rank, the solution with a higher crowding distance is preferred.

In broad terms, the NSGA-II algorithm conducts non-dominated sorting on the combination of the parent and offspring population, assigning them a rank (Pareto front) based on an ascending level of non-domination. Subsequently, a new population is filled according to the front ranking. If a front is taken partially, individuals with less crowding distance are given preference. The algorithm then creates an offspring population from this new population using selection, crossover, and mutation operators. This process continues until the stopping criteria are satisfied. The overall complexity of NSGA-II is $O(MN^2)$, where M is the number of objectives, and N is the population size, primarily determined by the non-dominated sorting part of the algorithm.

As a multi-objective optimization algorithm, NSGA-II endeavors to discern the trade-offs inherent in conflicting objectives, generating a set of optimal solutions known as the Pareto front or non-dominated solutions.

In the following Section we describe the different statistical performance metrics for he definition of our multi-objective problem.

3.3 Performance Evaluation

For performance evaluation, we employ a multi-level data splitting approach. Initially, approximately 20% of nodules are extracted from the full dataset to form the hold-out set. The remaining data are then used to create the first-level outer-folds. Within each of these outer-folds, a second-level splitting occurs to create inner-folds, resulting in a nested cross-validation (NCV) scheme [89]. Our optimization process is designed to operate within these inner-folds to identify the best hyperparameter values, which are subsequently employed in the outer-folds for model selection. The overall data splitting and optimization process is illustrated in Figure 3.1 and elaborated upon in the following.

The NCV scheme, as demonstrated in [18], significantly mitigates bias and yields an error estimate closely aligned with results obtained from an independent testing set.

Algorithm 1 Data Splitting for Nested Cross-Validation

Require: K_1 is the number of outer folds
Require: K_2 is the number of inner folds
Require: D, dataset with features X and output y

1: **procedure** DATASPLITTING(K_1, K_2, D)
2: **for** $i = 1$ ***to*** K_1 **do**
3: Split D into D_i^{train}, D_i^{test} for the *i'th* split
4: **for** $j = 1$ ***to*** K_2 **do**
5: Split D_i^{train} into D_j^{train}, D_j^{test} for the *j'th* split
6: **return** K_1-outer-folds, K_2-inner-folds

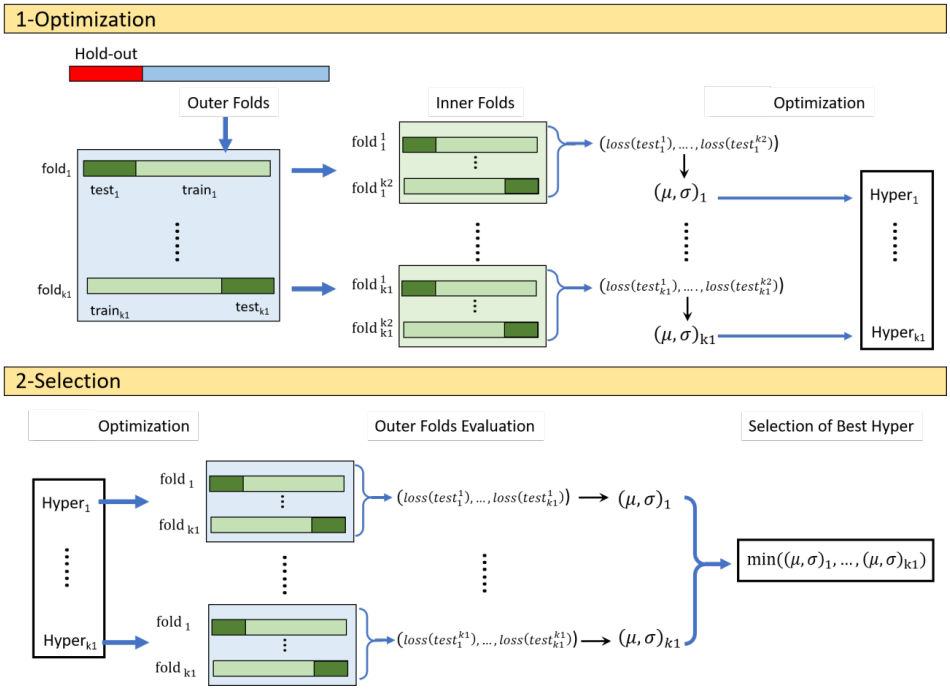


Figure 3.1: Optimization of hyperparameters in a nested cross validation scheme.

Initially, NCV partitions the data into K_1-outer-folds and K_2-inner-folds, as illustrated in Algorithm 1. The K_2-inner-folds are dedicated to identifying optimal hyperparameters, while the K_1-outer-folds are used to evaluate these optimal hyperparameters and perform model selection. It's worth noting that, in line 4, each training set of an outer-fold generates K_2-inner-folds, creating a one-to-many relationship.

For each outer-fold, a new hyperparameter optimization process begins using its associated inner-folds. Within each inner-fold (j), the network is trained using the

D_j^{train} set, and the loss of the trained network is evaluated on D_j^{test}. Consequently, for each network configuration, we obtain K_2 losses evaluated on the inner-fold test sets. Two statistical summaries of these losses across the K_2 inner-folds define our multi-objective optimization problem. Our multi-objective function are defined as follows:

$$\begin{aligned} f_1(\Theta) &= \frac{1}{K_2}\sum_{j=1}^{K_2} \text{loss}(D_j^{test}) \\ f_2(\Theta) &= \sqrt{\frac{\sum_{j=1}^{K_2} |\text{loss}(D_j^{test}) - \mu|^2}{K_2 - 1}} \end{aligned} \tag{3.2}$$

Here, f_1 and f_2 represent the average (μ) and standard deviation (σ) of the loss values generated from the test sets of the K_2-inner-folds.

Algorithm 2 Optimization in the inner folds.

Require: N is the population size
Require: G is the number of generations
Require: $Trials$ is the number of trials
Require: $SearchSpace$ contains the values range of all hyperparamters
Require: K_2 is the number of folds
Require: (D_j^{train}, D_j^{test}) k$_2$-inner-folds with 1 <= i <= K_2
Require: M is an architecture to be instantiated

1: **procedure** INNERFOLDSPROCESSING($K, D_i^{train}, D_i^{test}, M, p$)
2: Nsga2 ← NSGA2(N, G) ▷ Instantiate the NSGA2 Algorithm
3: **for** $t = 1$ ***to*** Trials **do**
4: p ← Nsga2.getSample($SearchSpace$) ▷ hyperparamter values sampling
5: μ_t, σ_t ← CrossValidationProcessing ($K_2, D_j^{train}, D_j^{test}, M, p$)
6: Nsga2.report(μ_t, σ_t)
7: p^* ← Nsga2.getBest() ▷ From the Pareto front, select the optimal solution p^* where both μ and σ are minimized
8: **return** p^*

Once all the K_2-inner-folds have been processed, we obtain a set of optimal solutions $P := (p_j^*)_{j=1}^{K_2}$ (asterisk), which undergo further selection based on the performance of networks trained on the K_1-outer-folds. The performance metrics for this selection process are, as before, the average (μ) and standard deviation (σ) of the loss values generated from the test sets of the K_1-outer-folds. The configuration in P with the lowest upper bound $\mu + \sigma$ is selected as the best among all individuals.

The entire methodology for NCV optimization is detailed in Algorithms 1-3. Algorithm 1 illustrates the NCV splitting of data into K_1-outer-folds and K_2-inner-folds.

The inner-fold optimization is implemented using Algorithm 2. Algorithm 2 processes the K_2-inner-folds and returns a set of optimal solutions p^* (asterisk) named P. The Trials represent the number of individuals to be sampled by the NSGA-II and

evaluated in the K_2-inner-folds. Once all the K_2-inner-folds have been processed, we have a set of optimal solutions p^* (asterisk), named P, required in Algorithm 3. Algorithm 3 evaluates the found individuals to make a selection of the best model and thus determine the best among all individuals, $p^\dagger$ (dagger).

Algorithm 3 Model selection across outer folds.

Require: K_1 is the number of outer-folds
Require: $(D_i^{train}, D_i^{test})$ outer-folds with $1 <= i <= K_1$
Require: M is an architecture to be instantiated
Require: P contains all the bests hyperparameters p^* found by the Algorithm3
1: **procedure** OUTERFOLDSPROCESSING($K, D_i^{train}, D_i^{test}, M, p$)
2: **for each** p^* ***in*** P **do**
3: $\mu, \sigma \leftarrow$ CrossValidationProcessing $(K_1, D_i^{train}, D_i^{test}, M, p^*)$
4: $p^\dagger \leftarrow$ getBest(μ, σ) ▷ Select the optimal solution $p^\star$ where both μ and σ are minimized.
5: **return** $p^\dagger$

Algorithm 4 is utilized by Algorithms 2 and 3. It involves the processing of either inner-folds or outer-folds indistinctly. It instantiates a model M with the hyperparameter values contained in p, conducts the training of the model, tests it, and computes the loss. It later calculates μ and σ, which are used to estimate the optimality of the individuals in the case of NSGA-II (Algorithm 2) or to make model selection and determine which individual is the best of all (Algorithm 3).

Algorithm 4 Processing of a Cross-Validation

Require: K is the number of folds
Require: $(D_i^{train}, D_i^{test})$ with $1 <= i <= K$
Require: M is an architecture to be instantiated
Require: p contains all the hyperparameter values
1: **procedure** CROSSVALIDATIONPROCESSING($K, D_i^{train}, D_i^{test}, M, p$)
2: $M_p \leftarrow M(p)$ ▷ Instantiate the model M with p
3: **for** $i = 1$ ***to*** K **do**
4: Train M_p on D_i^{train} with hyperparameters p
5: $\hat{y}_i \leftarrow$ Predict $M_p(D_i^{test})$
6: $te_i \leftarrow$ compute test error $Loss(y, \hat{y}_i)$
7: Compute the μ and σ from test errors te_i
8: **return** μ, σ

We partitioned the dataset into three levels of generalization to scrutinize the impact of the novel representation space:

- Nodule k-folds: In this methodology, we employ k-fold cross-validation to evaluate our model's performance in predicting unseen data. The dataset is partitioned based on nodules, treating the nodule as the fundamental unit of data

splitting. One subset is singled out as the test data, while the remaining subsets are dedicated to training. This process iterates k times, with each fold utilizing distinct nodules for testing. Following the training of a model for each fold, the diagnosis score is computed by averaging the performance across the k-folds. This approach provides dual levels of measurement: individual fold performance and an overarching evaluation of the model's proficiency across all folds. As a result, we can discern statistical variations both within each fold and across the entirety of folds.

- Leave-1-Nodule-Out: This strategy is a specific implementation of k-fold cross-validation, with k set equal to the maximum number of nodules in the dataset. The nodule is designated as the experimental unit for data splitting. Accordingly, subsets of nodules consist of one nodule assigned to the test data, while the remaining nodules form the training data. Since the test set encompasses only one nodule, this approach yields a single-level measurement, offering an overall assessment of the model's performance across all folds. Consequently, statistical variations can be analyzed comprehensively across all folds.

- Slice k-folds: In this approach, we leverage k-fold cross-validation, employing the slice as the experimental unit. This means that slices from the same nodule can be present in both the training and test data. The process is reiterated k times, with each fold encompassing different slices for testing. The diagnosis score is computed as the average across the k-folds. Analogous to the nodule k-folds method, this approach captures statistical variations at each fold and across all folds.

It is imperative to emphasize the distinction in experimental units among these approaches. In the case of nodule k-fold and leave-1-nodule-out strategies, the experimental unit is the nodule itself. Consequently, slices belonging to a particular nodule are exclusively allocated to either the training or test set, but not both. In contrast, the slice k-fold approach adopts the individual slice as the experimental unit, permitting slices from the same nodule to coexist in both sets. These diverse strategies offer valuable insights into the model's performance and generalization capacities, ensuring a robust evaluation marked by a high degree of generalization and reproducibility.

3.4 Use Case: Interpretability of Radiomic Features

In order to illustrate the benefits of the proposed strategy, we have applied it to the optimization of the hyper-parameters of a network mapping radiomic visual features to radiological annotations for better clinical interpretation of abstract features describing the visual appearance of medical scans. Our approach uses transformers to build an attention map from visual features to radiological annotations related to lesion malignancy. Recently, transformer architectures using the self-attention mechanism [90]

have emerged to be successful in natural language processing (NLP) with its capability of capturing global dependencies across sentence words. A main application is the translation/transformation of an input sequence into an output sequence in a different language [31]. Given that input and output languages are encoded as dictionaries, transformers can compute mappings transforming any two sequences of indexes.

The BERT transformer [31] used for the experiments is a sequence-to-sequence model with a binary cross-entropy loss that follows the architecture described in [90]. The encoder has a trainable embedding layer that maps the sequence of NV visual words to an Euclidean space of $emsize$ dimensions. The embedded sequence words are the input to $nlayers$ transformer identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism with $nhead$, and the second is a fully connected feed-forward layer with $nhid$ neurons, Relu activation and dropout to mitigate over-fitting.

The concatenation of the output of the last layer is the input to the decoder that computes the mapping between abstract and radiological features. The decoder is an ensemble of a fully connected layer with sigmoid activation. The set of hyperparameters of the whole system are the network hyperparameters defining its architecture ($emsize$, $nhead$, $nhid$, $nlayers$) and the training parameters, the dropout of the network and the learning rate, lr:

$$\begin{aligned}\Theta = &(\text{Architecture Parameters; Training Parameters}) = \\ &(emsize, nhead, nhid, nlayers; dropout, lr)\end{aligned} \tag{3.3}$$

Since BERT is trained using a weighted binary cross entropy to account for account across annotations, in this case there are not hyperparameters associated to the definition of the loss.

The input and output of a transformer are sequences of words represented as an index to a input and output corpus containing the collection of all input and output words. In our case, the input is a NV-dimensional vector of radiomic visual features and, thus, they have to be coded as indexes in order to be the input for the transformer.

The input vector of visual features, $\mathbf{v} = (v_0, \ldots, v_{NV-1}) \in [a_0, b_0] \times \cdots \times [a_{NV-1}, b_{NV-1}] \in \mathbb{R}^{NV}$ is transformed to a sequence of NV visual words, $\mathbf{w^v} = [w_0^v, \ldots, w_{NV-1}^v]$, using a discretization of each visual coordinate. For each coordinate, its range, $[a_j, b_j]$, is discretized into nV uniform bins and v_j is assigned to the index of the bin:

$$v_j \mapsto floor((v_j - a_j)/(b_j - a_j) * (nV - 1)) \in \{0, \ldots, nV - 1\} \tag{3.4}$$

for $floor(\cdot)$ the integer part of a real number.

The above transformation maps any coordinate to the same corpus of nV words indexed by $\{0, \ldots, nV - 1\}$. In order to assign each coordinate to a different set of words,

we shift its index by $j * nV$ positions to define w_j^v:

$$w_j^v = floor((v_j - a_j)/(b_j - a_j) * (nV - 1)) + \\ + j * nV \in \{j * nV, \ldots, (j+1) * nV - 1\} \tag{3.5}$$

This manner, **v** is mapped to a corpus of $(NV - 1) * nV$ visual words indexed by $\{0, \ldots, (NV - 1) * nV - 1\}$. We note that our transformation assumes that the visual features are always in the same order, which is a reasonable assumption automatically fulfilled in the case of being the output of a network.

Table 3.1: Hyperparameter search space for our model architecture.

Hyper-parameter	**nhead**	**nhid**	**emsize**	**dropout**	**nlayers**	**lr**
Search Space	[2, 10]	[8, 300]	[2,15]	[0.2, 0.8]	[1, 3]	$[1 \times 10^{-4}, 1]$

Table 3.2: Outer-Folds Ranking

Fold Number	**CI inferior**	**CI superior**
BERT-3	0.0297	0.0334
BERT-1	0.0307	0.0338
BERT-5	0.0308	0.0339
BERT-2	0.0304	0.0340
BERT-4	0.0318	0.0342
BERT-0	0.0320	0.0349
BERT-9	0.0377	0.0414
BERT-7	0.0372	0.0417
BERT-6	0.0321	0.0427
BERT-8	0.0267	0.0564

We have used a subset of 584 patients with 1110 nodules of the LUNA16 public data base [73] annotated by four radiologists. The radiological features are subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation and texture as well as malignancy. Calcification, internal structure, sphericity and subtlety are qualitative descriptors. The remaining ones are considered quantitative attributes. Multiple annotations are summarized as the most frequent label for categorical attributes and the average for quantitative ones. The input visual features are 19 GLCM Pyradiomics [88] features selected according to their correlation to lesion malignancy as described in [87][56]. Texture features were computed in 2D slices, which provides a total number of 6751 samples.

Table 3.3: Hold-Out Ranking

Fold Number	*Loss*
BERT-5	0.0344
BERT-1	0.0344
BERT-0	0.0346
BERT-4	0.0345
BERT-2	0.0347
BERT-3	0.0347
BERT-6	0.0382
BERT-7	0.0383
BERT-9	0.0386
BERT-8	0.0414

The hyperparameters values have been optimized using the Optuna HPO framework [5] integrated in the Nested Cross-Validation (NCV) scheme described in Section 3.3. In order to verify the generalization power of this step, 20% of the lesions (222 nodules) were randomly selected to be held out as an independent test set and the remaining 888 lesions were used for NCV optimization. Our NCV splits these lesions into 10-outer-folds and 10-inner-folds. Each outer fold has 800 lesions in the training set and 88 lesions in the test set. And each inner fold has 720 lesions in the training set and 80 lesions in the test set. Optuna [5] was run using as sampler the fast and elitist multi-objective NSGA-II (Nondominated Sorting Genetic Algorithm II) algorithm [12, 92] and a threshold as pruner.

The search space for BERT hyperparameters is shown in Table 3.1. The optimal hyper-parameters selected by Optuna for each outer-fold inner optimization are identified by this outer-fold number. In order to asses the reproducibility of the NCV optimization using statistical metrics, the BERT architectures defined by this set of hyperparameters selected at the inner folds were train and tested in both, the outer-folds and the holdout set. For the outer-folds, a different model was trained on the whole set of outer-folds and tested on the holdout test. Our hypothesis is that the best/worst performers of the outer-folds, should also be top/bottom performers on the hold-out set.

Table 3.2 reports intervals of the binary cross entropy loss at 95% confidence for the outer-fold test sets and the fold number in the first column. Results are sorted in ascending order, so first rows correspond to best performers. Table 3.3 reports the loss for the holdout set also sorted in ascending order.

We observe that the set of four worse configurations are the same (module a permutation) in both sets. In particular, the worst is always configuration BERT-8. Regard-

ing top performers, the top five are almost the same, with $4/5$ coincidences. The most prominent exception is configuration BERT-3, which is the best for the outer-fold but it is not (it is in 6th position) in the top five of the holdout set. However, the difference in performance across the top one and the 6th is of the order of $10^{(-4)}$, while it increases to $4 \times 10^{(-3)}$ for the 4 worse. This suggest a further investigation of the outer-fold rating and selection and considering some statistical tests to detect significant differences and discard those configurations performing significantly worse.

Chapter 4

RadioLung DataSet

We have established an infrastructure to support multi-center clinical data collection for lung cancer research. This initiative includes an online repository that allows the collection of diverse clinical data, such as histopathology, molecular analysis, and CT or PET scan images. Additionally, we provide details of the technology used to support this infrastructure.

Following globally accepted standards in radiology for early lung cancer detection, we elaborated on an Acquisition Protocol that outlines key parameters in CT scans, patient posture considerations, and low radiation dose for patient safety. Inclusion and exclusion criteria, along with biopsy procedures, are employed to classify nodules between benign and malignant cases, including adenocarcinoma and squamous cell carcinoma subtypes. Ethical considerations, patient consent, and data anonymization are prioritized in this research, aligning with foundational principles outlined in the Declaration of Helsinki.

The dataset, aimed at fostering collaborative research, is publicly accessible via `http://iam.cvc.uab.es/portfolio/radiolung-database`. Overall, these initiatives contribute significantly to advancing lung cancer research by enhancing data accessibility, privacy, and transparency in clinical data handling.

4.1 Annotated CT-Scans

4.1.1 Acquisition Protocol

In the context of medical imaging, such as CT scans, a voxel represents an intensity value that is proportional to the signal intensity of the corresponding volume of tissue, as shown in Figure 4.1. The intensity value is usually measured in Hounsfield Units (HU) and is derived from the x-ray attenuation coefficient of the tissue. This value

helps in distinguishing between different types of tissues, such as air, water, bone, and soft tissue. The smaller the size of the voxel, the better the quality of the image and the more accurate the representation of the underlying anatomy.

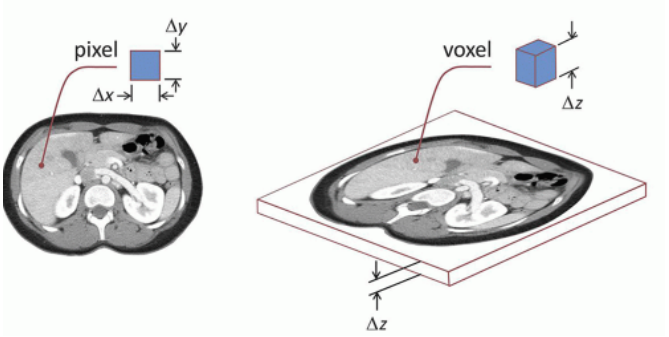


Figure 4.1: This image shows an axial cut of a slice from a CT scan. On the left side depicts a magnified view of a single pixel. On the right side, the same pixel is extended in a third dimension, creating a voxel, which is the smallest unit of a 3D image. (Original source: `https://radiologykey.com/computed-tomography-15`. Accessed date: 1 October 2023).

Based on clinical requirements and the unique characteristics of the patient, the acquisition parameters of a CT scan can be adjusted to yield images with diverse levels of resolution and contrast. They are stored alongside the CT images when utilizing the DICOM format. The following outlines acquisition parameters:

- **Slice thickness (mm)**: the thickness of the cross-sectional slices in the imaged body, denoted as Δz in Figure 4.1, is comparable to the collimator width, which is a factor influencing X-ray beam thickness. The adjustment of the collimator width directly influences slice thickness, thereby determining the precision of fine details visualized in each image slice.

- **Slice spacing (mm)**: also known as interslice or gap, is the distance between two adjacent slices. It represents the gap between two consecutive images along the z-axis. A smaller slice spacing can provide more detailed and continuous images.

- **Pixel spacing (mm)**: is the physical distance between two pixels in the x and y directions of the image, in sequence depicted as Δx and Δy in the Figure 4.1.

- **Convolution kernel**: is a mathematical filter applied during image reconstruction. Different kernels can enhance or suppress certain features in the images. For example, a sharp kernel might be used to emphasize bone structures, while a smoother kernel might be used for soft tissue visualization.

- **kVp (voltage)**: this is the peak voltage of the X-ray beam used during image acquisition. It influences the contrast and penetration of X-rays through tissues.

Higher kVp values are often used for imaging dense structures like bones, while lower values are suitable for soft tissue imaging.

- **Rescale slope and rescale intercept**: are a scaling factor and an offset, respectively, applied for mapping raw pixel values to Hounsfield Units (HU), representing physical density values. They are used in a linear mapping expressed in Equation 2.1.

Our imaging protocol utilizes Multi-Detector Row CT (MDCT) scanners [59, 79, 66] with a section collimation of <=1mm, enabling high-resolution imaging and exceptional sensitivity through a minimum of 16 data acquisition channels. This setup facilitates the detection of subtle abnormalities within the targeted anatomical region.

CT scans are executed with meticulously defined parameters to optimize the imaging process. These parameters serve as the gold standard, ensuring sufficient scan resolution and quality for the radiological evaluation of malignancy [47, 98]. Operating at a voltage (KVP) range of 100-140 kVp (with a recommended setting of 120 kVp), the current (mA) is automatically modulated based on patient size, ranging from 100 to 350 mA. This personalized approach tailors radiation exposure to the unique anatomical characteristics of each patient, achieving a delicate balance between image quality and safety. The pitch, set at 1:0, and a gantry rotation time of <= 0.5 seconds are carefully chosen to minimize motion artifacts during image acquisition, aiming to keep total scan times under 15 seconds for patient comfort and to reduce the likelihood of motion-related distortions in the images.

Patient safety is prioritized, and the imaging protocol adopts a low radiation dose strategy [34], targeting an effective dose of <= 1.5 mSv. This is achieved through a combination of reduced tube current (mA) and careful management of the gantry rotation time. The CT dose index volume (CTDIvol) is capped at <= 3.0 mGy (32cm) for a standard-sized patient (170 cm, 70 Kg, BMI = 24). This commitment underscores the practice's dedication to maintaining the highest standards of safety while obtaining diagnostically relevant images.

The imaging process extends beyond parameter settings to include the patient's posture during image acquisition. Images are acquired during a single inspiratory breath-hold, with the patient in the supine position and arms raised overhead. This specific positioning minimizes motion artifacts and enhances the clarity of the captured images. Notably, no intravenous or oral contrast agents are introduced during the imaging procedure, streamlining the process and reducing the complexity of the scan.

Following the acquisition phase, a meticulous approach is taken to image reconstruction, a critical aspect of delivering accurate diagnostic information. The reconstruction process adopts a section thickness of at most 2 mm or less (recommended 1 mm) and spacing of <= 1 mm, ensuring the creation of detailed and high-resolution images. Two distinct image reconstruction algorithms are employed: a high spatial frequency algorithm designed for lung parenchyma and an intermediate spatial fre-

quency algorithm tailored for mediastinal structures. The latter, specifically the mediastinal reconstruction algorithm, plays a crucial role in reducing noise in images acquired with lower radiation doses, underscoring the commitment to optimal image quality even in scenarios with reduced radiation.

As a final step, the captured images are systematically archived in a hospital-based Picture Archiving and Communication System (PACS) server. These images are not merely stored but are fully annotated, providing a comprehensive record for future reference. Duplicates of these images are also stored in local repositories, ensuring accessibility for clinical use when needed. This robust archival process reflects a dedication to maintaining a complete and organized repository of patient imaging data for ongoing clinical management and analysis.

In essence, the detailed imaging protocol outlined here represents a holistic and patient-centric approach to medical imaging, combining advanced technology, meticulous parameter settings, safety considerations, and a comprehensive data management strategy. This approach aligns with the overarching goal of providing the highest quality of care to patients while contributing to advancements in medical diagnostics and treatment.

4.1.2 Nodule Annotation

We present a comprehensive preprocessing procedure tailored to prepare and standardize the data in the lung cancer database. This essential preparation is pivotal for enabling advanced image analysis tasks and fostering research in the healthcare domain.

To render data from the lung cancer database usable, a thorough preprocessing step becomes imperative. This multi-faceted preprocessing encompasses several key components:

1. **Image Format Conversion**: We start with the initial CT scans stored in the DICOM format, a common standard in healthcare. To optimize compatibility with advanced image analysis tasks, we transform the initial DICOM images into the NIFTI format. NIFTI, known for its lightweight file size and user-friendly structure, offers a versatile and standardized representation for medical imaging data.

2. **Quality Assurance**: Ensuring the quality and compatibility of CT scan data is paramount. We examine the scan acquisition parameters to confirm their compliance with required standards (refer to Chapter 4). If these criteria are not met, the CT scan is removed from further analysis. This step guarantees that the CT scans meet the necessary resolution quality.

3. **Nodule Localization**: the precise location of nodules in the CT scan is determined by one o more experience radiologists through manual o semi-automated assessment. Depending on the database used, the location of nodules can be

represented as follows: 3D bounding boxes, nodule mask, or center and maximum diameter. For the last two cases, we derive the 3D bounding box. In any case, it is necessary to consider the coordinate system of the CT scans to obtain a well-fitted 3D bounding box.

An important issue that arises when dealing with medical images and applications is the conversion between coordinate systems, as it involves the interaction of the following three systems:

- The World Coordinate System, that is a Cartesian coordinate system that defines the position and orientation of the medical image in a global reference frame. It provides a common spatial reference for different images and facilitates their alignment and integration. The world coordinate system typically uses three axes (x, y, z) to represent the three-dimensional space.
- The Anatomical Coordinate System (Patient Coordinate System), also known as the patient coordinate system, describe standard anatomical position of the human. It consists of three orthogonal planes:
  - Axial Plane: is parallel to the ground and separates the head (**S**uperior) from the feet (**I**nferior). It is also referred to as the transverse plane or the horizontal plane.
  - Coronal Plane: is perpendicular to the ground and divides the body into front (Anterior) and back (Posterior) portion. It is sometimes called the frontal plane.
  - Sagittal Plane: divides the body into Left and Right sides. It is commonly used to refer to the plane that separates the body into equal halves.

  Based on these planes, all axes have their positive direction. For example, the negative Superior axis is represented by the Inferior axis. However, it's crucial to understand that different medical applications may adopt different definitions of this 3D basis. The most common bases used are:

  1. LPS (Left, Posterior, Superior): This basis is employed in DICOM images and by the ITK toolkit. It defines the orientation of the three axes as follows: the positive direction is towards the Left, the Posterior, and the Superior directions, respectively.
  2. RAS (Right, Anterior, Superior): Similar to LPS, the RAS basis is used in applications like 3D Slicer. However, it differs by flipping the orientation of the first two axes. In the RAS basis, the positive direction is towards the Right, the Anterior, and the Superior directions, respectively.
- The Image Coordinate System describes how an image was acquired with respect to the anatomy. It consists of three axes (i, j, k) representing the right, bottom, and backward directions. The origin denotes the position of the first acquired

voxel (0, 0, 0) in the anatomical coordinate system, while spacing indicates the distance between voxels along each axis.

To map the world coordinate system (x, y, z) in millimeters to the voxel coordinate system (i, j, k), we rely on the 3D affine matrix (which is a sequence of transformations including shear, reflection, rotation, scaling, and translation) and the patient-oriented Reference Coordinate System (RCS) stored in the DICOM format [1]. By combining these components, we can accurately assign physical positions to ever pixel in an image. This mapping can be achieved using the following equation:

$$\begin{bmatrix} P_x \\ P_y \\ P_z \\ 1 \end{bmatrix} = \begin{bmatrix} X_x\Delta_i & Y_x\Delta_j & 0 & S_x \\ X_y\Delta_i & Y_y\Delta_j & 0 & S_y \\ X_z\Delta_i & Y_z\Delta_j & 0 & S_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i \\ j \\ 0 \\ 1 \end{bmatrix} = \mathbf{M} \begin{bmatrix} i \\ j \\ 0 \\ 1 \end{bmatrix}$$

Then, to map millimeter coordinates to voxels, it is crucial to calculate the inverse of the 3D affine transformation. This inverse matrix enables us to determine the corresponding pixel coordinates (i, j) for the given millimeter coordinates.

$$\begin{bmatrix} i \\ j \\ 0 \\ 1 \end{bmatrix} = \mathbf{M}^{-1} \begin{bmatrix} P_x \\ P_y \\ P_z \\ 1 \end{bmatrix}$$

Where:

- P_{xyz} : The world coordinate system (x, y, z) expressed in mm w.r.t RCS (mm).
- S_{xyz} : The x, y, and z coordinates of the upper left hand corner (center of the first voxel transmitted) of the image w.r.t RCS are expressed in mm. These coordinates are obtained from the DICOM's Image Position Patient attribute.
- X_{xyz}, Y_{xyz} : The direction cosines of the first row and the first column with respect to the patient's orientation expressed in unit vectors. These values are obtained from the DICOM's Image Orientation Patient attribute.
- i, j : Column index and row index to the image plane respectively (index).
- Δ_i, Δ_j : Column pixel resolution and row pixel resolution expressed in mm. These values are obtained from the DICOM's Pixel Spacing attribute.

From the DICOM's Slice Location attribute, we can obtain the relative position of the image plane expressed in mm. By ordering all the 2D slices appropriately, we stack

[1] It is worth noting that this information is also preserved in the NIFTI format.

them in the normal imaging axis and we can obtain the k coordinate which corresponds to the stacked slices in the image plane axis to complete the voxel coordinate (i, j, k). By utilizing theses voxel coordinates, we can compute a well-fitted 3D bounding box.

A respiratory medicine physician with seven years of expertise utilized 3D-Slicer (version 4.11.20200930) to precisely annotate of the Volume of Interest (VOI) for each nodule, as illustrated in Figure 4.2. This free, open-source, and multi-platform software, widely employed in medical and biomedical imaging research, facilitated precise delineation of the minimal nodule space. The physician's task involved defining VOIs that optimally encapsulated each nodule. Table 4.4 provides comprehensive information about our database, including demographic details and statistical data such as the minimum, maximum, and slice count for each nodule type and sex.

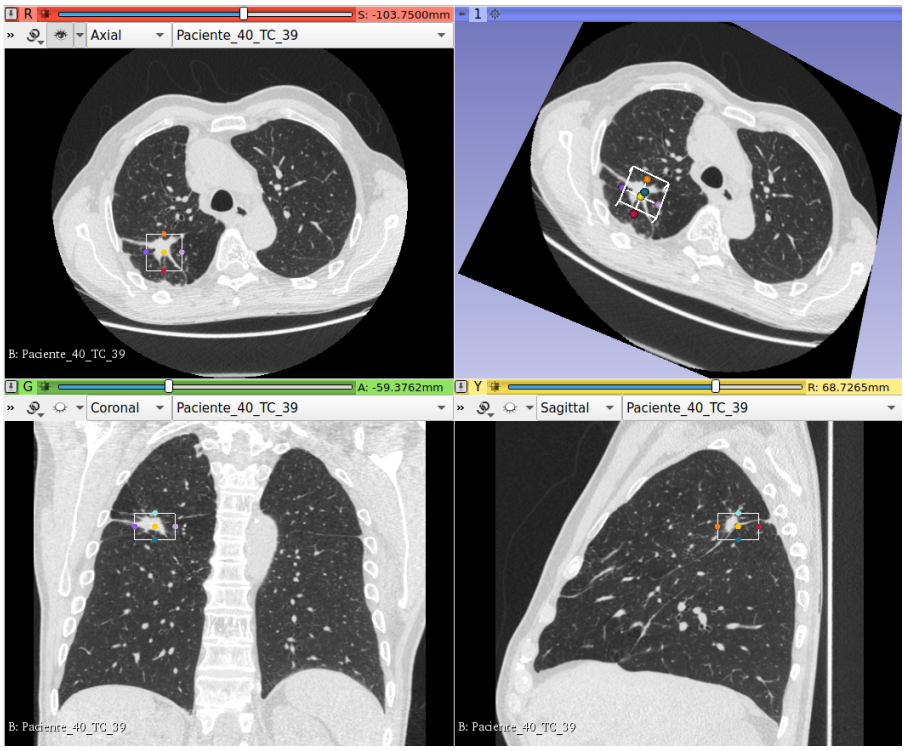


Figure 4.2: CT visualization in axial, coronal, and sagittal cuts was employed to achieve precise delineation of the nodule's VOI using the 3D-Slicer software. The upper-right image presents a three-dimensional representation of the VOI.

4.2 Clinical Data

In order to have a robust database, a part from DICOM CT-Scans, Radiolung database include relevant clinical data such as: clinical variables, image variables, nodule histopathology as well as molecular analysis. Each variable with its description are reported next.

- **Epidemiological History:**
  - **BMI (Body Mass Index):** Used to calculate the weight status of a person. It is calculated using the formula weight(kg)/height^2(m).
  - **Smoking:** Tobacco is one of the main risk factors for developing lung cancer. Risk varies based on total accumulated exposure (pack-years) and time since quitting.
  - **Pack-Years:** A formula used to calculate the cumulative exposure to tobacco over the patient's lifetime. An accumulated exposure of > 20 pack-years is associated with an increased risk of developing lung cancer.
  - **Air Pollution:** Prolonged exposure to inhaled irritants can increase the risk of lung diseases, impacting physiological reserve and interacting with other pathologies.
  - **Family History of Cancer:** Associated with an increased risk of developing one's own cancer due to certain genetic alterations.
  - **Personal History of Cancer:** Increases the risk of developing a new neoplasm and may recur in the form of lung metastasis.
- **Pulmonary History:**
  - **COPD (Chronic Obstructive Pulmonary Disease):** Presence of a progressive obstruction to airflow associated with an abnormal inflammatory response. Population with COPD has a high risk of developing lung cancer.
  - **Asthma:** Reversible inflammation of the airways. Possible risk factor for lung cancer.
  - **Bronchiectasis:** Irreversible pathological dilation and destruction of the large bronchi.
  - **Tuberculosis (TB):** Known association between TB and an increased risk of developing lung cancer.
  - **Pneumonia:** Lung cancer can manifest as repeated episodes of pneumonia.
- **Pulmonary Function Tests:**
  - **FCV (CVF: Forced Vital Capacity):** Total volume of air expelled during a forced expiration. The percentage indicates where the patient stands in relation to their theoretical value, adjusted for age, weight, height, sex, and race.

  - **FEV1 (Forced Expiratory Volume in 1 second):** Maximum expiratory volume exhaled in the first second. The percentage indicates where the patient stands in relation to their theoretical value, adjusted for age, weight, height, sex, and race.
  - **FEV1/FCV:** Percentage relationship between FEV1 and FVC.
  - **DLCO (Diffusing Capacity for Carbon Monoxide):** Measure of the gas transfer conductance from inspired air to the blood. In summary, it measures the gas diffusion capacity of lung tissue.

- **Tumor Variables:**
  - **Type (Stage 1, 2, 3):** An abnormal growth or tumor can be benign or malignant.
  - **Lobe (Stage 1, 2, 3):** Lungs are divided into lobes, each with defined margins. Relevant for incidence and surgical considerations.
  - **Location (Stage 1, 2):** Lung neoplasms can occur anywhere in the lungs. Distance from the hilum or pleura can be indicative of a specific type of tumor.
  - **X-ray (Stage 1, 2):** X-rays detect changes in lung tissue density, providing an additional diagnostic indicator.
  - **Tumor Size (Stage 1, 2, 3):** One of the values used for staging a neoplasm, determining subsequent treatment.
  - **Differentiation (Stage 1, 2, 3):** Assigned by a pathologist based on how cells appear under a microscope. Useful for prognosis.
  - **Necrosis (Stage 1, 2, 3):** Tissue death, usually due to insufficient blood supply. In cancer, necrosis indicates tumor aggressiveness.
  - **Vascular Infiltration (Stage 1, 2, 3):** If it has penetrated the walls of blood vessels and lymphatics, it is an indicator of a more advanced tumor stage and a higher probability of metastasis.
  - **Histological Diagnosis (Stage 2, 3):** The diagnosis assigned by a pathologist based on observed characteristics under a microscope.
  - **TNM (Stage 2, 3) (Malignant Tumor Classification):** A system used to stage a tumor (determine its extent). Examines tumor size, lymph node involvement, and the presence or absence of metastasis.

- **PET Scan Parameters:**
  - **Radiofármaco:** Fluorodeoxyglucose. A radiopharmaceutical injected before PET, emitting positrons detected by PET.
  - **Nodular Uptake (Stage 1, 2, 3):** If there is higher uptake in the nodule, it indicates more metabolic activity and, consequently, more cell division, a typical characteristic of cancerous processes.

  - **Nodule SUV (Stage 1, 2, 3):** The higher the number, the more radiopharmaceutical uptake is detected.
  - **Lymphadenopathy Uptake (Stage 1, 2, 3):** Increased metabolic activity in lymph nodes. It may be due to involvement in the cancerous process or any nearby infectious process.
  - **Uptake in Other Locations (Stage 1, 2, 3):** Increased radiopharmaceutical uptake in other locations.

- **CT Scan Parameters:**
  - **Slice Thickness:** CT scans display images that are like slices of the person in the three axes. Thinner slices provide more precision in diagnosing and staging the nodule.
  - **Nodule Diameter:** The maximum diameter of the nodule on CT helps correlate with the maximum diameter observed under the microscope after surgery.
  - **Nodule Shape (Stage 1, 2, 3):** Typically benign characteristics include an oval or lobulated shape, well-defined, while typically malignant shapes are irregular and spiculated.
  - **Nodule Density (Stage 1, 2, 3):** Ground Glass Density, Partly Solid, Solid.
  - **Emphysema (Stage 1, 2, 3):** Abnormal permanent dilation of air spaces accompanied by the destruction of the wall of the lung alveolus.

- **Surgery Parameters:**
  - **Surgery Type:** Lungs are divided into lobes, each with defined margins. Surgery may involve segmentectomy, lobectomy, or pneumonectomy.
  - **Lymph Node Stations (Stage 1, 2, 3):** Lymph node involvement observed under the microscope after surgery.
  - **Resection Margins (Stage 1, 2, 3):** Examined during surgery to detect cancer cells at the margins of the extracted specimen.
  - **Lymphatic and Vascular Invasion (Stage 1, 2, 3):** Microscopic examination to determine if the tumor has invaded vessel tissue.
  - **STAS (Stage 1, 2, 3):** Spread through air spaces. Dissemination implies higher tumor aggressiveness.
  - **Surgical Complications:** Prognostic markers indicate a patient's risk of surgical complications, and surgical complications are a prognostic factor for the patient's recovery.

4.3 Online digital repository for Multicentric data Collection

To facilitate efficient collection, storage, and sharing of multi center clinical data, a digital repository online available has been implemented. To achieve this, a private website accessible only with a password is implemented to upload all relevant clinical data (including clinical variables, nodule histopathology and molecular analysis), acquisition characteristics and parameters, as well as image data acquired by CT and PET-scan. A database in MySQL [33] has been also created to allow efficient storage, access and data modification. The Entity-Relationship model [21] and logic design [83] of the database for restoration and modification clinical data is shown in figure4.3 and 4.4 respectively.

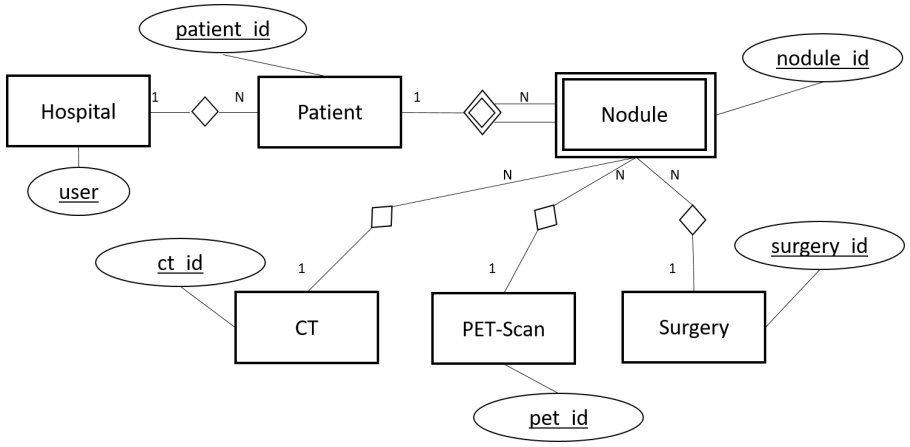


Figure 4.3: Entity-Relationship model of the online digital repository for multicentric data collection.

Concerning the front-end interface, we have developed a platform based on PhP [24], HTML5-CSS [36] and mySQL database for easy data gathering and interactive management. According to data collection requirements, web forms have been customized to include PhP instructions managing data storage and security in terms of data privacy and visualization for each of the hospitals.

The different views of the data gathering webpage are reported in the following screenshots in figures 4.5, 4.6,4.7,4.8,4.9,4.10.

4.4 Radiolung Dataset Description

To be eligible for recruitment, patients were required to undergo CT-chest examinations for pulmonary nodules and meet well-defined inclusion and exclusion criteria:

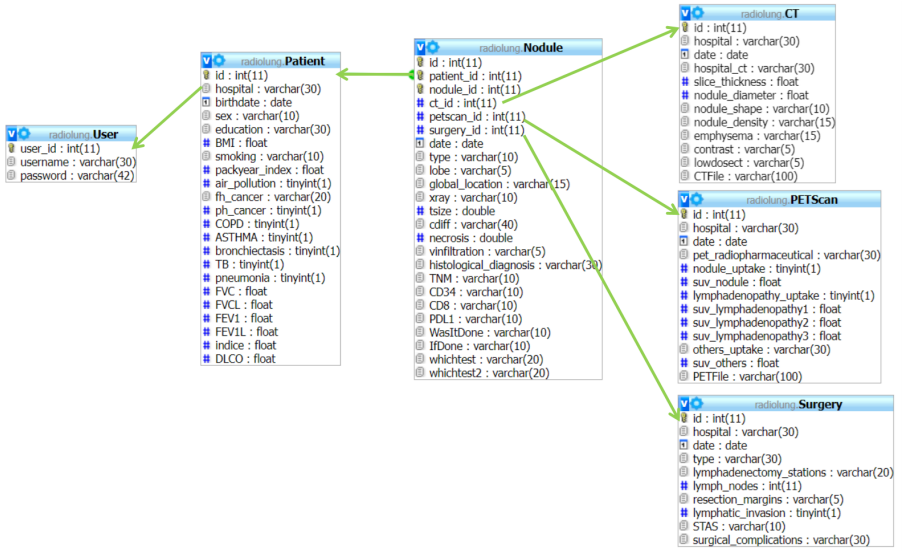


Figure 4.4: Relational model of the online digital repository for multicentric data collection.

Figure 4.5: Initial page asking for the personal login.

- **Inclusion criteria**: stipulated nodules with a diameter ranging from 8 to 30 mm and a final diagnosis of non-small cell lung carcinoma or a non-malignant tumor.
- **Exclusion criteria**: encompassed individuals previously diagnosed with lung cancer, those with uncured extra-pulmonary cancer (excluding non-melanoma skin cancer), pregnant individuals, those who received chemotherapy or cytotoxic drugs in the last 6 months, and those declining to sign the consent.

It is crucial to underscore that pulmonary nodules were systematically classified through **biopsy procedures** in each case, ensuring a thorough and precise assessment.

This study adheres to the foundational ethical principles outlined in the Declaration of Helsinki - Fortaleza/Brazil, 2013. In every instance, informed consent is diligently sought, and both images and clinical data are handled with utmost anonymity to

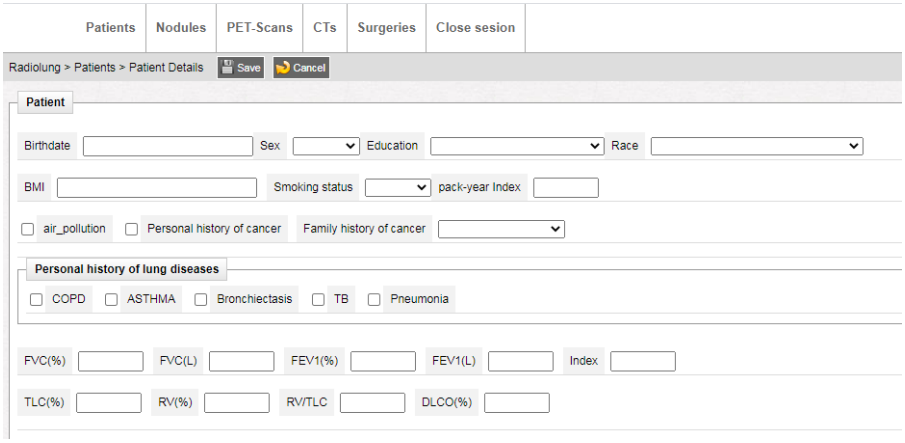


Figure 4.6: Patient data inserting and modification.

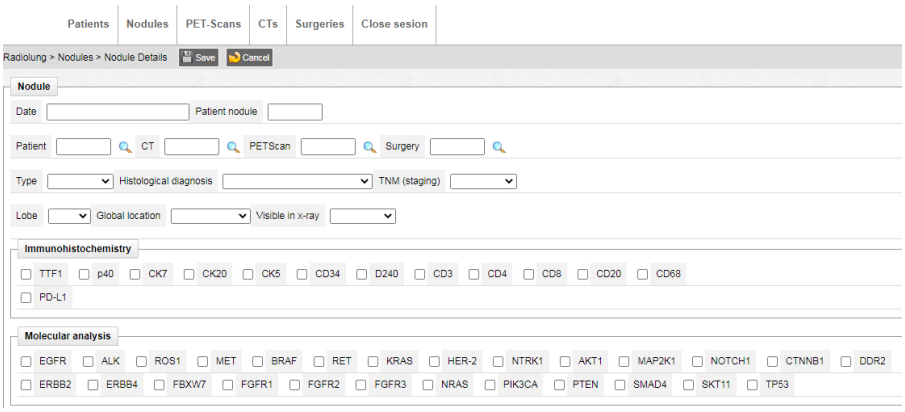


Figure 4.7: CT data inserting and modification.

protect patient confidentiality. Prior to participant recruitment, the research received approval from the ethics committee at HUGTiP (CEIC H. Germans Trias i Pujol: PI-19-169).

To facilitate data sharing between a hospital and an external institution, such as the CVC at UAB, which operate independently from the healthcare system, it was imperative to establish a confidentiality agreement beforehand. This enables us to integrate essential information for both clinical data and diagnostic assessments in the evaluation of lung cancer.

- **Slice thickness:** CT scans display images that resemble cuts of the person in all three axes. Thinner slices provide more precision in diagnosing and staging the nodule.

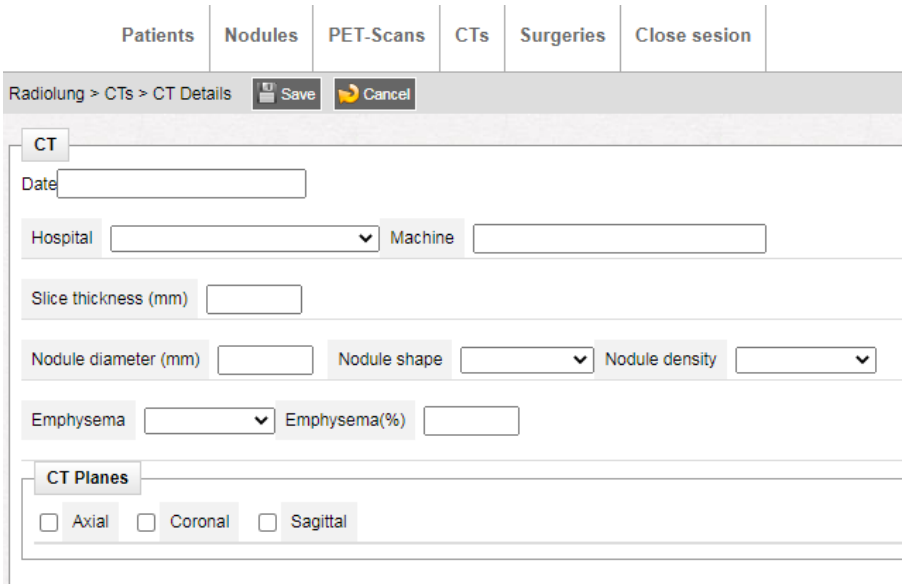


Figure 4.8: CT data inserting and modification.

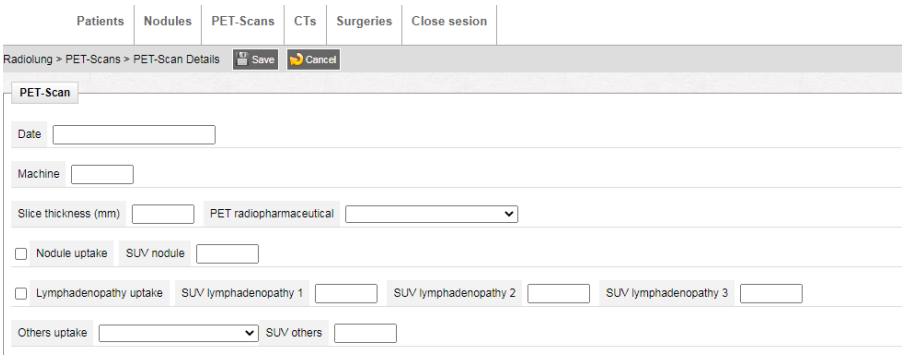


Figure 4.9: PET data inserting and modification.

- **Nodule diameter:** The maximum diameter of the nodule by CT helps correlate with the maximum diameter observed under the microscope after surgery.
- **Nodule shape (phase 1, 2, 3):** Typically benign characteristics include an oval or lobulated, well-defined shape, while typically malignant ones have irregular, spiculated forms.
- **Nodule density (phase 1, 2, 3):**
  - **Ground glass density:** An area in the lungs with increased density while preserving bronchial and vascular margins.

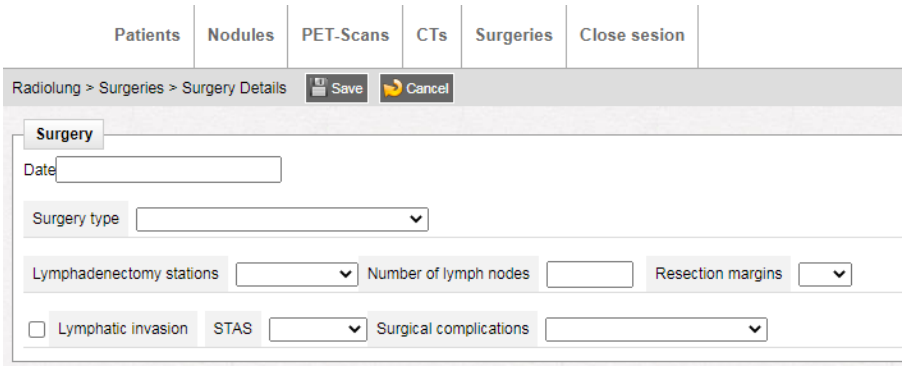


Figure 4.10: Surgery data inserting and modification.

  - **Partially solid:** Part solid and part ground glass. Associated with malignancy in 63-92
  - **Solid:** Typically well-defined, without preservation of bronchial and vascular margins. Associated with malignancy in a lower percentage than partially solid nodules.

- **Emphysema (phase 1, 2, 3):** Abnormal permanent dilation of air spaces, accompanied by the destruction of the wall of the pulmonary alveolus. It is the most important risk factor for lung cancer in COPD. The hospital is committed to anonymizing the data before transmitting it in the DICOM format.

Patients were recruited at Germans Trias i Pujol University Hospital (HUGTiP) in Barcelona, Spain, for a prospective cohort study conducted between December 2019 and September 2023. During this period, comprehensive data collection, including images and clinical/demographic information, was carried out. A cohort of 90 recruited patients underwent focused CT-chest examinations targeting pulmonary nodules, adhering to precisely defined inclusion and exclusion criteria outlined earlier in this section.

CT scans were performed using GE Medical Systems, Philips, and Siemens CT scanners, following the acquisition parameters aligned with the established data acquisition protocol detailed in Section 4.1.1. Refer to Table 4.1 for detailed acquisition settings corresponding to each manufacturer. The scanners imaged a total of 95 nodules. Out of these, 23 were identified as benign, while 72 were classified as malignant. Subsequent analysis revealed that 57 of the malignant nodules were adenocarcinoma, and 15 were squamous cell carcinomas. A visual representation of nodules with various diagnoses is provided in Figure 4.11. For a comprehensive breakdown of nodule diagnoses by scanner, please consult Table 4.2. Additionally, refer to Table 4.3 for a detailed breakdown of nodule histology based on the scanner used.

The dataset is accessible to the public through the following link: `http://iam.cv`

Table 4.1: Specifications of acquisition parameters for each scanner manufacturer.

Description\Manufacturer	**GE Medical Systems**	**Philips**	**Siemens**
Model	BrightSpeed Discovery ST LightSpeed VCT Revolution CT	Brilliance 16 GeminiGXL 16 Incisive CT TruFlight Select	Somatom Drive
Convolution Kernel	LUNG SOFT STANDARD	B YA YB YC	Bf42f Bl57f
Slice Thickness	0.62-1.25	1-2	0.6-1.5
Slice Spacing	0.62	0.5-1	1
Pixel Spacing XY	0.56-0.87	0.35-92	0.63-0.98
kVp	100-120	120	100-140

Table 4.2: Nodules, both benign and malignant, captured by each scanner.

Nodule	**GE Medical Systems**	**Philips**	**Siemens**	**Total**
Benign	3	9	11	23
Malign	25	36	11	72
Total	28	35	22	95

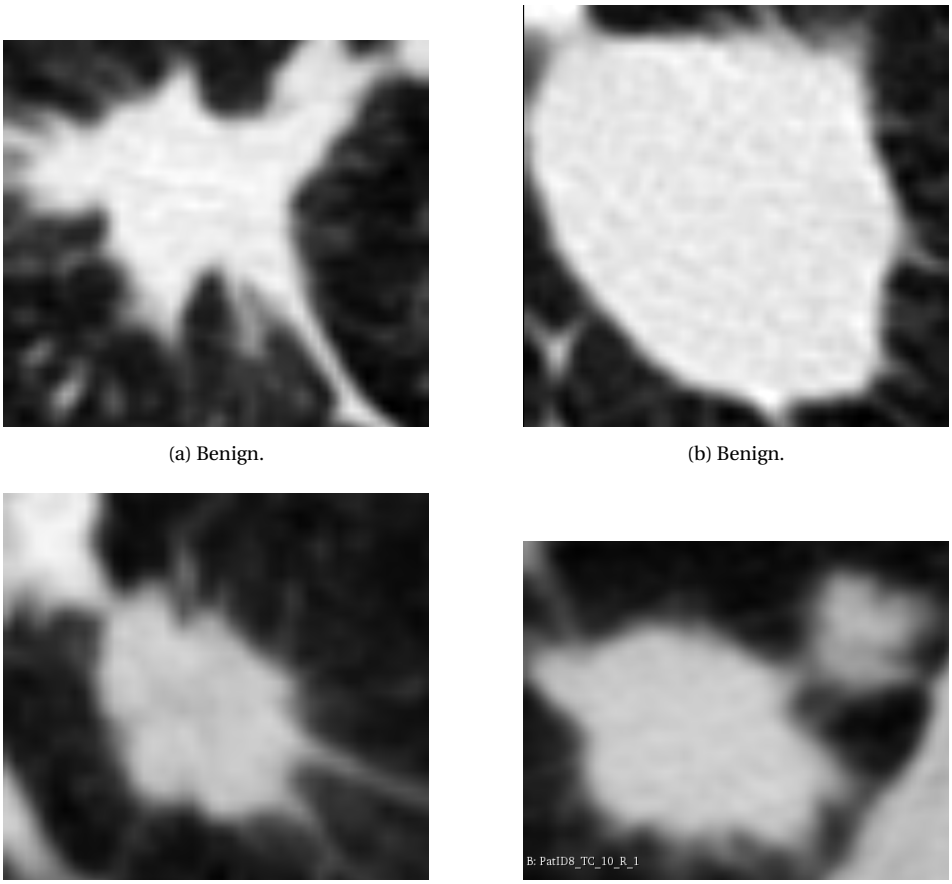
`c.uab.es/portfolio/radiolung-database`.

Table 4.3: Distribution of malignant nodule subtypes among the CT scanner manufacturers.

Nodule	**GE Medical Systems**	**Philips**	**Siemens**	**Total**
Malign				
Adenocarcinoma	19	30	8	57
Squamous Cell Cancer	6	6	3	15
Total	25	36	11	72

	Description	**Male**	**Female**	**Total**
Demographic Population	Patients	67	28	95
	Age Avg ± Std	71.2±6.34	64.24±11.53	67.72±8.94
	Benign PN	12	11	23
	Malign PN	42	30	72
Nodule Characterization	Benign Slices Min/Max/Avg	6/80/47	18/42/30	6/80/37
	Malign Slices Min/Max/Avg	8/93/43	12/81/41	8/93/39

Table 4.4: Demographic population and nodule characterization.



(a) Benign.

(b) Benign.

(c) Adenocarcinoma.

(d) Squamous cell carcinoma.

Figure 4.11: Axial cuts of pulmonary nodules with diverse diagnoses and imaging sources. (a) and (b) showcase benign nodules imaged from the GE Medical System and Philips scanners, respectively. Moving to malignant nodules, (c) represents an adenocarcinoma imaged from the Siemens scanner, while (d) shows a squamous cell carcinoma imaged from the Philips scanner.

Chapter 5

Experiments and Results

In order to validate our methods, we have conducted the following different experiments:

1. **Selection of the Optimal Representation Space.** The optimal representation space for the benign and malignant classification has been selected using the strategy described in Chapter 2.
2. **Comparison to SoA.** To assess the advantages of the proposed strategy, the best representation space selected in the first experiment was compared to state of the art methods.
3. **Impact of Acquisition Parameters.** The impact on performance of CT acquisition parameters has been assessed for the top performers of the previous experiment in order to determine critical values for a clinical acquisition protocol.

In the next Sections, we report the experimental set-up and results obtained for each of the experiments.

5.1 Optimal Representation Space

We use PyRadiomics [88] (version 3.01) specifically to extract GLCM features for both Radiomic Embedding and Deep Radiomic Embedding, as detailed in chapter 2. PyRadiomics stands out as an open-source Python package designed for extracting radiomic features from medical imaging volumes. This versatile tool encompasses shape features, first-order features, and textural features, including those derived from Gray Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM), and Gray Level Dependency Matrix (GLDM), providing a comprehensive characterization of various aspects of the lesion.

Table 5.1: Distribution of the RadioLung dataset across holdout and training sets.

Pulmonary Nodule	**Holdout Set**	**Training Set**	**Total**
Total number of Nodules	25	59	84
Benign	4	8	12
Malignant	21	51	72
Adenocarcinoma	15	42	57
Squamous Cell Cancer	6	9	15

Based on findings from our earlier investigation [87], we identified the optimal parameters for two distinct aspects: normalizing the gray level intensity of the VOI with $MaxIntensity = 24$ and establishing the number of histogram bins with $Nbins = 128$, as detailed in Equations 2.2 and 2.4, respectively. The setting $MaxIntensity = 24$ defines the range for normalizing the gray level intensity within the VOI to 24 gray levels. Simultaneously, the choice of $Nbins = 128$ determines the level of granularity with which GLCM features encapsulate the textural patterns.

In the various representation spaces outlined in Section 2.1, features are extracted by traversing the nodule slice-by-slice in an axial manner from 2D images. Subsequently, these features are concatenated and subjected to a t-student test, comparing average values between malignant and benign slices. The goal is to discern the correlation between these features and lesion malignancy.

Specifically, for Radiomic Embedding, our study [87] identified the 19 most relevant GLCM features, enumerated in Table 2.1. In the case of VGG Embedding and VGG Radiomic Embedding (where features are flattened), the top 500 features with the lowest p-values are selected. These chosen features, as detailed in Table 5.2, serve as input for training a classifier.

Table 5.2: Number of selected features for different nodule embeddings.

Nodule Embedding	**Selected Features**
Radiomic Embedding	19
VGG Embedding	500
VGG Radiomic Embedding	500

In accordance with our optimization methodology outlined in Chapter 3, we approach network optimization by formulating hyperparameters as a multi-objective optimization problem within the space of network architectures. The Non-Dominated Sorting Genetic Algorithm II (NSGA-II) serves as the *search strategy* for addressing this multi-objective problem. The *search space* encompasses various models parameterized by hyperparameters defining the network architecture, including the number of layers, number of neurons, activation function, weight initialization, as well as optimizer, weight decay, and learning rate, all of which significantly impact performance.

Table 5.3 displays the search space, where 'input_features' denotes the number of features in the input layer. In particular, the Scaler hyperparameter leverages functions from the scikit-learn package, with 'Standard' corresponding to StandardScaler, 'MinMax' to MinMaxScaler, 'Robust' to RobustScaler, 'Quantile' to QuantileTransformer, 'MaxAbs' to MaxAbsScaler, and 'Power Transformer' to PowerTransformer functions. The remaining hyperparameters, along with the neural network, were incorporated using the PyTorch package.

Table 5.3: Specification of the search space.

Hyperparameter	Search Space
Number of Layers	[3, 5]
Number of Neurons	[3, input_features]
Activation Function	ReLU, ReLU6, LeakyReLU, Sigmoid, Tanh
Weight Initialization	Normal, Xavier, Kaiming, Orthogonal
Optimizer	SGD, Adam, RMSprop
Weight Decay	$[1 \times 10^{-5}, 1]$
Learning Rate	$[1 \times 10^{-6}, 1]$
Scaler	None, Standar, MinMax, Robust, Quantile Transformer, MaxAbs, Power Transformer
Batch Size	[32, input_features]

The optimization of hyperparameter values is achieved by leveraging the Optuna framework [5]. This involves dynamically constructing the search space at runtime and employing the NSGA-II algorithm as the sampler to generate candidate solutions. The initial population size of NSGA-II is set to 500. Thus, during the first iteration, 500 initial random solutions are generated for 'generation 0,' used to produce the first offspring set, 'generation 1,' and subsequent generations. Offspring solutions evolve through crossovers and mutations of the parent solutions, as detailed in Section 3.2. Additionally, a pruning threshold is applied to prematurely terminate less promising trials during the training process. The sampling algorithm identifies optimal configurations based on the multi-objective function, precisely formulated in Equation 3.2.

To evaluate the performance of different hyperparameter configurations, we implemented Nested Cross-Validation (NCV) scheme on the training set. Our multi-objective problem is defined by the μ (mean) and σ (standard deviation) of the loss function, evaluated on the test folds. The evaluation follows specific steps detailed in Algorithms 1 to 4. Furthermore, to mitigate data imbalances during training, we employ a weighted cross-entropy loss function, formulated as:

$$\text{loss} = \frac{\sum_{i=1}^{N} weight[class[i]] loss(i, class[i])}{\sum_{i=1}^{N} weight[class[i]]} \tag{5.1}$$

where $\text{loss}(i, class[i])$ is the cross-entropy loss for the i-th class computed from the

classifier prediction x and the true class as:

$$\text{loss}(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_{j=1}^{N}\exp(x[j])}\right) \tag{5.2}$$

and the weight $weight[class[i]]$ is given by the inverse of the class frequency.

The search space encompasses various neural network architectures dynamically defined at runtime based on hyperparameters sampled by NSGA2. These hyperparameters, including the number of fully connected layers, hidden neurons and activation function (see Table 5.3), play a pivotal role in shaping and defining the model.

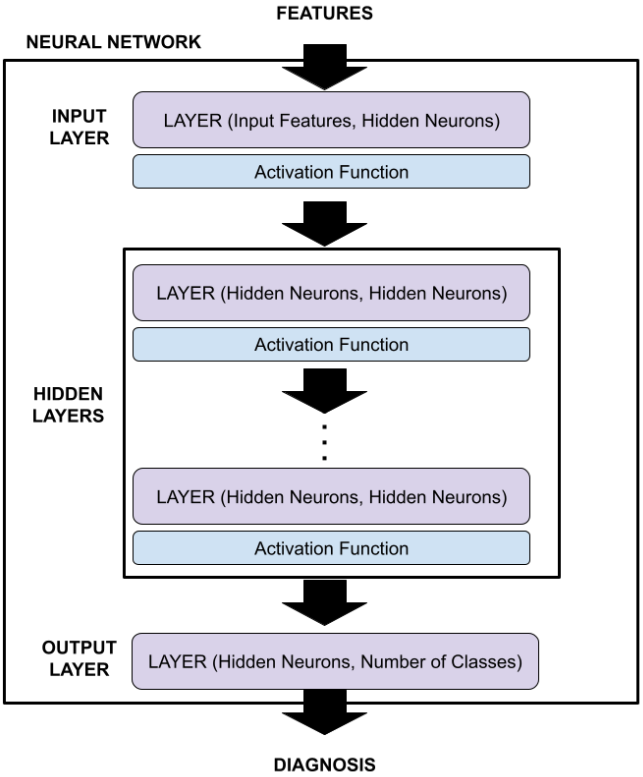


Figure 5.1: Dynamically configured neural network architecture established at runtime based on the hyperparameters selected through the NSGA2 optimization algorithm.

For the first experiment we use the RadioLung dataset described in chapter 4. Randomly, 25 PN have been set aside for the holdout set, while the remaining 59 PN constitute the training set. Within the holdout set, there are 4 benign and 21 malignant PN, with 15 classified as adenocarcinoma and 6 as squamous cell cancer. In the training

set, there are 8 benign and 51 malignant PN, further categorized as 42 adenocarcinoma and 9 squamous cell cancer. The distribution of PN in the RadioLung dataset is summarized in Table 5.1. Notably, the data highlights a considerable imbalance, not only between benign and malignant cases but also within the malignant category, specifically between adenocarcinoma and squamous cell cancer.

This section presents a comprehensive evaluation of models derived from diverse data domain, encompassing data splitting and integrated performance metrics within an optimization process. The data domain comprises distinct representation spaces obtained from Radiomic Embedding, VGG Embeddings, MobileNet Embedding and VGG Radiomic Embedding, as introduced in Chapter 2. Notably, VGG Radiomic Embedding is combined using Concatenation and Average fusion of features.

To assess the impact of these representation spaces, the data is stratified into three levels of generalization. Each level uses a unique experimental sampling unit, either nodule or slice, to partition the data into training and test folds. Further details on the specific methods, including Nodule k-folds, Leave-1-Nodule-Out, and Slice k-folds, are elaborated in chapter 3.

For data splitting, we adopted a 5-fold approach at both the nodule and slice levels, utilizing the Python StratifiedGroupKFold function to maintain consistent class proportions in both the training and test sets. Additionally, 25 nodules from the dataset were randomly selected as an independent set (Holdout) of test patients. This selection aimed to evaluate the reproducibility of the ranges computed in a slice split. Quality metrics such as precision, recall, and the F1-score were employed to assess model performance. The results obtained for the optimal configurations are summarized (mean ± standard deviation) in Table 5.7 and 5.5. Furthermore, details regarding the optimal hyperparameters are presented in Tables 5.8 and 5.9.

We notice that the Intensity domain has the lowest score among all domains. When using slice folds for splitting, both VGG Radiomic Concatenation and VGG Radiomic Average domains exhibit high recall for both benign and malignant nodules. The recall range for VGG Radiomic Concatenation is (1, 1) for malignant cases and (0.84, 1) for benign cases. However, when splitting at the nodule level, the VGG Radiomic Average domain experiences a significant drop in benign recall, almost reaching 0. On the other hand, for the VGG Radiomic Concatenation domain, while the malignancy recall score falls within the range of (0.88, 1), the recall range for benign cases is (0.37, 1). It is worth noting that the high standard deviation (around 30%) indicates considerable variability across folds for the VGG Radiomic Concatenation domain. MobileNet Embedding outcomes in the Nodule 5-folds setting, demonstrating higher scores precision, recall, and F1-Score at both slice and nodule levels. While some values are not available for this model, it exhibits competitive performance across metrics, establishing itself as the top-performing model in this context.

Analyzing Tables 5.8 and 5.9, we observe that optimal hyperparameters for VGG Radiomic Embedding with Concatenation or Average are discovered by NSGA2 in the 4th and 5th generations of offsprings, with an exception for Slice 5-folds of Concate-

nation found in generation 0, randomly. Notably, the configuration with the fewest hidden neurons corresponds to Radiomic Embedding, likely due to the reduced input dimensionality of its 19 features. The prevailing optimizer is SGD, and the predominant activation function is ReLU6, offering additional saturation to control the exploding gradient problem by constraining outputs between 0 and 6. The frequently utilized scalers encompass StandardScaler, which is apt for situations where features exhibit varying scales or adhere to a normal distribution. Additionally, MaxAbsScaler proves beneficial when dealing with data containing a mixture of positive and negative values, ensuring the preservation of sign information.

Table 5.4: Cross Validation Statistical Summary. Intensity Representation Spaces.

Data Domain	Data Split	Diagnosis	Metrics	Precision	Recall	F1-Score
Radiomic-Embedding	Nodule 5-folds	Malign	Slice	0.85 (± 0.09)	0.67 (± 0.16)	0.74 (± 0.10)
		Benign		0.29 (± 0.24)	0.44 (± 0.21)	0.31 (± 0.17)
		Malign	Nodule	0.77 (± 0.10)	0.64 (± 0.14)	0.69 (± 0.07)
		Benign		0.28 (± 0.22)	0.34 (± 0.25)	0.28 (± 0.18)
	L1O	Malign	Slice	0.64	0.80	0.71
		Benign		0.12	0.20	0.04
		Malign	Nodule	0.60	0.80	0.68
		Benign		0.04	0.20	0.07
	Slice 5-folds	Malign	Slice	0.95 (± 0.01)	0.78 (± 0.02)	0.86 (± 0.01)
		Benign		0.42 (± 0.03)	0.79 (± 0.04)	0.55 (± 0.03)
		Malign	Nodule	0.97 (± 0.03)	0.85 (± 0.04)	0.90 (± 0.02)
		Benign		0.57 (± 0.09)	0.88 (± 0.10)	0.68 (± 0.06)
VGG-Embedding	Nodule 5-folds	Malign	Slice	0.90 (± 0.10)	0.72 (± 0.09)	0.80 (± 0.09)
		Benign		0.34 (± 0.15)	0.72 (± 0.16)	0.43 (± 0.14)
		Malign	Nodule	0.93 (± 0.10)	0.74 (± 0.05)	0.82 (± 0.05)
		Benign		0.51 (± 0.12)	0.88 (± 0.15)	0.62 (± 0.08)
	L1O	Malign	Slice	0.90)	0.57	0.69
		Benign		0.32	0.75	0.45
		Malign	Nodule	0.98	0.55	0.69
		Benign		0.41	0.95	0.56
	Slice 5-folds	Malign	Slice	0.99 (± 0.01)	0.96 (± 0.01)	0.97 (± 0.01)
		Benign		0.82 (± 0.05)	0.93 (± 0.03)	0.87 (± 0.03)
		Malign	Nodule	1.00 (± 0.00)	1.00 (± 0.00)	1.00 (± 0.00)
		Benign		1.00 (± 0.00)	1.00 (± 0.00)	1.00 (± 0.00)
MobileNetV2-Embedding	Nodule 5-folds	Malign	Slice	0.72 (± 0.13)	0.77 (± 0.11)	0.75 (± 0.12)
		Benign		0.63 (± 0.14)	0.66 (± 0.17)	0.64 (± 0.16)
		Malign	Nodule	0.82 (± 0.17)	0.78 (± 0.21)	0.80 (± 0.19)
		Benign		0.67 (± 0.16)	0.68 (± 0.19)	0.67 (± 0.17)
	L1O	Malign	Slice	-	-	-
		Benign		-	-	-
		Malign	Nodule	-	-	-
		Benign		-	-	-
	Slice 5-folds	Malign	Slice	-	-	-
		Benign		-	-	-
		Malign	Nodule	-	-	-
		Benign		-	-	-

Table 5.5: Cross Validation Statistical Summary. Deep Radiomic Representation Spaces.

Data Domain	Data Split	Diagnosis	Metrics	Precision	Recall	F1-Score
VGG-Radiomic Concatenation	Nodule 5-folds	Malign	Slice	0.92 (± 0.06)	0.75 (± 0.17)	0.81 (± 0.12)
		Benign		0.41 (± 0.24)	0.73 (± 0.12)	0.47 (± 0.17)
		Malign	Nodule	0.90 (± 0.06)	0.70 (± 0.20)	0.76 (± 0.11)
		Benign		0.46 (± 0.28)	0.72 (± 0.16)	0.51 (± 0.17)
	L1O	Malign	Slice	0.90	0.72	0.74
		Benign		0.41	0.55	0.41
		Malign	Nodule	0.89	0.74	0.76
		Benign		0.39	0.57	0.42
	Slice 5-folds	Malign	Slice	0.99 (± 0.00)	0.98 (± 0.01)	0.98 (± 0.01)
		Benign		0.89 (± 0.04)	0.95 (± 0.02)	0.92 (± 0.03)
		Malign	Nodule	0.99 (± 0.01)	0.99 (± 0.01)	0.99 (± 0.01)
		Benign		0.97 (± 0.05)	0.98 (± 0.04)	0.97 (± 0.03)
VGG-Radiomic Average	Nodule 5-folds	Malign	Slice	0.89 (± 0.10)	0.75 (± 0.12)	0.80 (± 0.09)
		Benign		0.33 (± 0.12)	0.64 (± 0.21)	0.40 (± 0.08)
		Malign	Nodule	0.88 (± 0.10)	0.75 (± 0.14)	0.79 (± 0.07)
		Benign		0.46 (± 0.16)	0.67 (± 0.18)	0.52 (± 0.10)
	L1O	Malign	Slice	0.47	0.58	0.52
		Benign		0.08	0.40	0.14
		Malign	Nodule	0.44	0.58	0.50
		Benign		0.08	0.40	0.13
	Slice 5-folds	Malign	Slice	1.00 (± 0.00)	0.99 (± 0.00)	0.99 (± 0.00)
		Benign		0.96 (± 0.03)	0.97 (± 0.01)	0.97 (± 0.02)
		Malign	Nodule	1.00 (± 0.00)	1.00 (± 0.00)	1.00 (± 0.00)
		Benign		1.00 (± 0.00)	1.00 (± 0.00)	1.00 (± 0.00)

Table 5.6: Holdout Statistical Summary. Intensity Representation Spaces.

Data Domain	HPO Data Split	Diagnosis	Metrics	Precision	Recall	F1-Score
Radiomic	Nodule 5-folds	Malign	Slice	0.71	0.47	0.57
		Benign		0.19	0.40	0.26
		Malign	Nodule	0.60	0.43	0.50
		Benign		0.11	0.20	0.14
	Slice 5-folds	Malign	Slice	0.72	0.51	0.59
		Benign		0.19	0.37	0.25
		Malign	Nodule	0.69	0.64	0.67
		Benign		0.17	0.20	0.18
VGG Intensity	Nodule 5-folds	Malign	Slice	0.00	0.00	0.00
		Benign		0.23	1.0	0.37
		Malign	Nodule	0.00	0.00	0.00
		Benign		0.26	1.00	0.42
	Slice 5-folds	Malign	Slice	0.76	0.70	0.73
		Benign		0.20	0.26	0.23
		Malign	Nodule	0.71	0.71	0.71
		Benign		0.20	0.20	0.20
MobileNet Intensity	Nodule 5-folds	Malign	Slice	0.86	0.61	0.72
		Benign		0.38	0.70	0.49
		Malign	Nodule	0.91	0.67	0.77
		Benign		0.44	0.80	0.57
	Slice 5-folds	Malign	Slice	-	-	-
		Benign		-	-	-
		Malign	Nodule	-	-	-
		Benign		-	-	-

Table 5.7: Holdout Statistical Summary. Deep Radiomic Representation Spaces.

Data Domain	HPO Data Split	Diagnosis	Metrics	Precision	Recall	F1-Score
VGG Radiomic Concat.	Nodule 5-folds	Malign	Slice	0.77	1.00	0.87
		Benign		0.00	0.00	0.00
		Malign	Nodule	0.74	1.00	0.85
		Benign		0.00	0.00	0.00
	Slice 5-folds	Malign	Slice	0.78	0.71	0.74
		Benign		0.24	0.30	0.27
		Malign	Nodule	0.69	0.64	0.67
		Benign		0.17	0.20	0.18
VGG Radiomic Average	Nodule 5-folds	Malign	Slice	0.76	0.55	0.64
		Benign		0.21	0.41	0.28
		Malign	Nodule	0.73	0.57	0.64
		Benign		0.25	0.40	0.31
	Slice 5-folds	Malign	Slice	0.67	0.16	0.26
		Benign		0.20	0.73	0.32
		Malign	Nodule	1.00	0.14	0.25
		Benign		0.29	1.00	0.45

Table 5.8: Optimized hyperparameters for Radiomic, MobileNet and VGG Embeddings.

Data Domain	Data Split	Scaler	Activation Function	Hidden Neurons	Weights Init.	Batch Size	Optimizer	Weight Decay	Learning Rate	NSGA2 Generation
Radiomic	Nodule 5-folds	Quantile	ReLU6	3	Kaiming	2153	SGD	0.0028	0.1219	1
	L1O	Quantile	Tanh	9	Xavier	1027	Adam	0.6976	0.7915	2
	Slice 5-folds	Standard	ReLU	12	Kaiming	386	SGD	0.012	0.0284	3
VGG Intensity	Nodule 5-folds	MaxAbs	Leaky ReLU	489	Orthogonal	2256	SGD	0.1664	0.3517	3
	L1O	Standard	Tanh	64	Orthogonal	220	Adam	0.1571	0.0459	4
	Slice 5-folds	Standard	ReLU	415	Normal	988	SGD	0.0545	0.6403	3
MobileNet Intensity	Nodule 5-folds	Scaler	ReLU	420	Normal	352	SGD	0.00764	0.01372	3
	L1O	Standard	Tanh	64	Orthogonal	420	SGD	0.0711	0.0352	5
	Slice 5-folds	Standard	ReLU	385	Normal	529	SGD	0.0355	0.0013	4

Table 5.9: Optimized hyperparameters for VGG Radiomic Embedding with Concatenation and Average fusion of features.

Data Domain	Data Split	Scaler	Activation Function	Hidden Neurons	Weights Init.	Batch Size	Optimizer	Weight Decay	Learning Rate	NSGA2 Generation
VGG Radiomic Concatenation	Nodule 5-folds	MinMax	ReLU6	430	Normal	2297	Adam	0.2093	0.0207	5
	L1O	Quantile	Tanh	114	Normal	1603	Adam	0.4717	0.0816	3
	Slice 5-folds	Standard	ReLU6	366	Kaiming	1422	SGD	0.0384	0.634	0
VGG Radiomic Average	Nodule 5-folds	MaxAbs	ReLU6	98	Xavier	571	SGD	0.7375	0.0036	4
	L1O	MaxAbs	ReLU6	303	Normal	1687	RMSprop	0.1199	0.4541	4
	Slice 5-folds	Standard	ReLU6	166	Kaiming	432	SGD	0.0189	0.3003	5

5.2 Comparison to SoA

We have compared our best model selected in 5.1 with state of the art methods which include the three type of approaches: radiomics [27], machine learning [101] and deep CNN [75, 99, 103, 45, 46]. In order to compare to the results reported for each of state of the art methods, we have computed the following metrics from true positive, TP, true negative, TN, false negative, FN, and false positive, FP diagnosis at nodule level:

$$Sensitivity = 100 \cdot \frac{TP}{TP + FN} \tag{5.3}$$

Sensitivity measures the percentage of correctly diagnosed malignant nodules.

$$Specificity = 100 \cdot \frac{TN}{TN + FP} \tag{5.4}$$

Specificity measures the percentage of benign nodules correctly identified.

$$Accuracy = 100 \cdot \frac{TP + TN}{Number\ of\ Nodules} \tag{5.5}$$

for *Number of Nodules* denoting the total amount of nodules. The accuracy measures the percentage of correctly diagnosed nodules (both malign and benign nodules) among the total number of nodules in the dataset.

$$F1\ \ Score = 100 \cdot \frac{2 \cdot Prec \cdot Rec}{Prec + Rec} \tag{5.6}$$

for Rec, $Prec$ denoting, respectively, the precision and recall at diagnosis level:

$$Rec = 100 \cdot \frac{TP}{TP + FN} \qquad Prec = 100 \cdot \frac{TP}{TP + FP} \tag{5.7}$$

The metric (5.6) measures the trade-off between recall and precision, and in general, a higher F1-score means a better performance. We also computed the receiver operating characteristic (ROC) curves and the area under the curve (AUC).

Table 5.10 shows the metrics for state of the art methods grouped according to type of approach and our method with best performance in boldface. It reports the metrics obtained by Model3 in our test set together with the results obtained by the selected state of the art in their datasets and reported in their works. We also report the number of parameters of each method as indicator of its complexity and computational and data cost for training. Our method outperforms in Accuracy, Sensitivity and F1 Score. In computer-aided diagnose, sensitivity is significant because correctly finding out patients with malignant nodules is crucial. Besides, the highest F1 Score implies that our method achieves the best trade-off between precision and recall. Our method has a splendid compromise between the performance of the system and the number of trainable parameters. A remarkable point compared to Deep CNN approaches, is that,

our method needs strongly less samples to train the model, which is a must in medical imaging.

Table 5.10: Results of our method compared to the state of the art with malignant nodules as positive cases.

Approaches	Accuracy	Sensitivity	Specificity	F1 Score	AUC	Param. (M)
Radiomics						
Peikert et al. [27]	–	90.40	85.50	–	0.939	**<0.29**
Machine Learning						
Zhang et al. [101]	96.09	96.84	**95.34**	–	**0.979**	**<0.29**
Deep CNN						
Multicrop [75]	87.14	77.00	93.00	–	0.930	–
Nodule-level 2D [99]	87.30	88.50	86.00	87.23	0.937	–
Vanilla 3D [99]	87.40	89.40	85.20	87.25	0.947	–
DeepLung [103]	90.44	81.42	–	–	–	141.57
AE-DPN [45]	90.24	92.04	88.94	90.45	0.933	678.69
NASLung [46]	90.77	85.37	95.04	89.04	–	16.84
Hybrid						
Our	**96.30**	**100**	83.33	**97.67**	0.940	0.29

5.3 Impact of Acquisition Parameters

The impact of acquisition parameters has been assessed using the Odds ratio (OR). This score calculates the relationship between a variable and the likelihood of an event occurring. In our case, OR can be interpret as identifying influential acquisition parameter by assessing the relationship between each parameter and the method outcome (failure). That is, the risk of error under a given acquisition parameter. For each parameter, its OR has been estimated using a logistic regression model using a generalized linear model under a binomial distribution and logit link:

$$model_f 1 <- glm(failure\ parameter, data = dades, familiy = binomial()) \tag{5.8}$$

the exponential of the coefficient estimated for the parameter gives the OR for the parameter. For each model, we report OR, its 95% CI, p values for significance (with significant values indicated in bold face) and the number of samples. For all statistical analysis a p-value < 0.05 was considered significant. Statistical analysis were conducted using R version 4.3.2.

Table 5.11 reports the failure ratios for each of the acquisition parameters prone to impact on method's performance. For the categorical parameters (manufacturer and KVP), we report the number and percentage of failures and successful predictions. For the continuous parameters (X Ray, number of slices, slice thickness and spiral pitch factor) we report the median and inter-quartile ranges (IQR). The nodules acquired with GE device have a greater rate of success than the other devices. Regarding acquisition parameters, scans acquired with a KVP different than 120 have a smaller rate of success.

Table 5.11: Global Prediction failures, n (%); Median (IQR).

Characteristic	N	Correct, $N = 69$	Failure, $N = 18$
Manufacturer	87		
GE MEDICAL SYSTEMS		23 (85%)	4 (15%)
Philips		32 (78%)	9 (22%)
SIEMENS		14 (74%)	5 (26%)
KVP	87		
120		58 (82%)	13 (18%)
Another voltage		11 (69%)	5 (31%)
X Ray Tube Current	87	243.00 (164.00, 299.00)	247.50 (193.75, 379.50)
Number of slices	87	358.00 (311.00, 497.00)	451.50 (322.75, 468.00)
Slice thickness	87	1.50 (0.63, 2.00)	1.50 (1.31, 2.00)
Spiral Pitch Factor	48	9.84 (9.84, 13.75)	13.50 (9.88, 15.00)

Table 5.12: Prediction failures for each acquisition parameter. [1]OR = Odds Ratio, CI = Confidence Interval

Predictors	OR[1]	95% CI[1]	p-value	N
Manufacturer			0.607	87
GE MEDICAL SYSTEMS	—	—	—	—
Philips	1.62	0.46-6.56	—	—
SIEMENS	2.05	0.47-9.57	—	—
KVP			0.267	87
120	—	—	—	—
Another voltage	2.03	0.56-6.68	—	—
X Ray Tube Current (per 50 units)	1.11	0.96-1.27	0.145	87
Number of slices (per 50 units)	1.01	0.81-1.23	0.933	87
Slice thickness	1.73	0.81-4.16	0.152	87
Spiral Pitch Factor (per 0.1 units)	1.22	0.97-1.58	0.098	48

Table 5.12 reports the statistical summary of the glm models adjusted for each pa-

rameter. As for the possible acquisition parameters that could explain the miss classification of the predicted diagnosis, we do not have significant results, probably due to the low number of samples. Even though, it seems that questions about the manufacturer, voltage, X-ray or slice thickness could be relevant. The failure rate with a GE device is substantially lower than with Siemens or Philips, and volts different from 120 almost double the predicted failure rate.

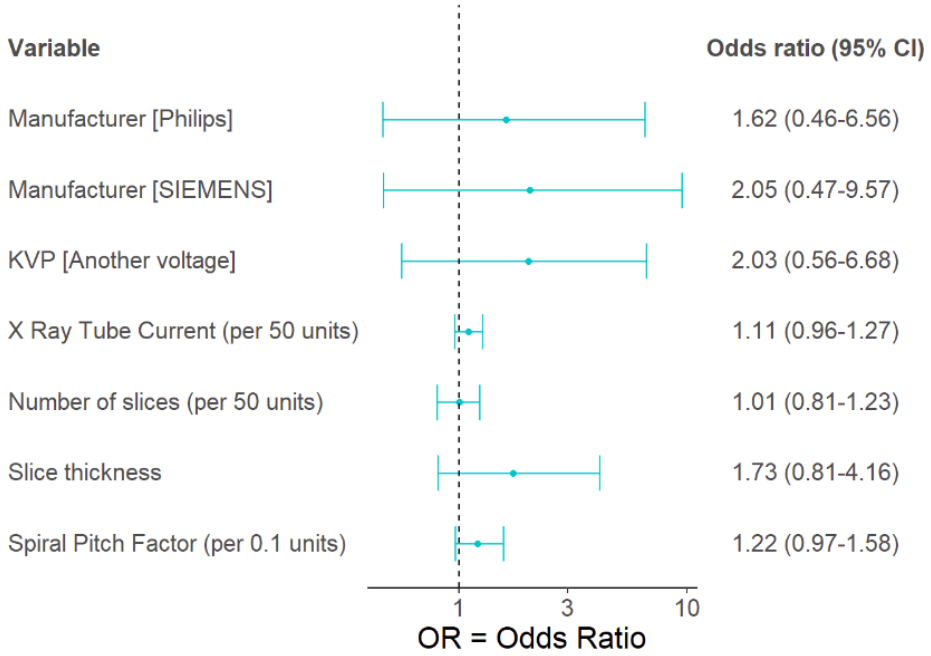


Figure 5.2: Forest plot of univariate odds ratio results - Failures

Figure 5.2 plots the 95% CI for each of the factors to visually check any deviation from OR=1, which indicates no influence of the parameter. All factors, except the number of slices show a deviation.

Chapter 6

Conclusions and future work

The goal of this thesis is to improve the early diagnosis of lung cancer. In order to achieve this and following clinical practice, an accurate characterization of the nodules has been done. In this context, we contribute to machine learning systems for diagnosis of lung cancer in, both, system's pipeline and acquisition of a dataset for training systems for early diagnosis of lung cancer. Intelligent artificial methods applied to medical imaging have to face two key drawbacks. The available small amount of labelled data and the obligation that methods must ensure good rates avoiding false positives. In order to overcome with these two main challenges, we have proposed an hybrid method that combines an embedded radiomic texture features to characterize nodules and an optimized feed-forward network for nodule diagnosis. The nodule embedding step is based on selecting those radiomic features that significantly correlate to malignancy ensuring reproducibility with minimal training data. The fully connected network architecture and hyperparameters are optimized using own-defined metrics of the diagnostic power to ensure maximum clinical outcome.

These are the main conclusions for each of the contributions of the thesis regarding a system for diagnosis of lung cancer:

- **Visual Representation Spaces**. We have presented an hybrid method based on classic radiomic features combined with a network with an architecture optimized for the malignancy diagnosis of a pulmonary nodules using a novel strategy based on multi-objective optimization. Our optimized approach achieves competitive (being best for some metrics) results for identification of malignancy using a highly unbalanced small size number of cases. These intermediate results show that radiomics are able to approximate the malignancy diagnosis of a pulmonary nodule and encourage further research including a higher number of cases and the optimization of convolutional architectures.
- **Framework for Reproducible HyperParameter Optimization**. We have presented a strategy for the optimization of network hyper-parameters using a multi ob-

jective Non-dominated Sorting Genetic Algorithm combined with a nested cross validation to optimize statistical metrics of the performance of networks. In order to illustrate the benefits of the proposed strategy, we have apply it to an application use case of a network mapping radiomic visual features to radiological annotations for better clinical interpretation of abstract features describing the visual appearance of medical scans. Results obtained indicate the generalization power of the proposed optimization strategy. However, we have to notice that a model configuration exception has been find out. In this way, further investigation should be addressed in order to filter this sort of outliers. A proposed solution, could be to consider some statistical tests to detect significant differences and discard those configurations performing significantly worse.

- **RadioLung Dataset**. We presented an own collected dataset for early lung cancer diagnosis, including imaging and clinical data. Our approach involves the development of a precise imaging acquisition protocol. This protocol utilizes Multi-Detector Row CT Scanners with high-resolution features and incorporates a low radiation dose strategy to prioritize patient safety. Notably, our protocol meticulously considers patient factors during both image capture and reconstruction. Aligned with globally recognized standards in the radiology community, the acquisition protocol is designed to detect lung cancer nodules in their early stages. Subsequently, each identified nodule undergoes histopathological diagnosis through biopsy samples.

- **Impact of CT Acquisition Parameters.** We have conducted a statistical analysis of the impact of CT scan acquisition parameters on the performance of methods. In particular, the impact of the manufacturer, KVP voltage, slice thickness, X Ray, number of slices,and spiral pitch factor has been analyzed using logistic regression models to estimate OR and detect significant differences. Descriptive statistics show that scans acquired with GE devices and KVP=120 have a greater rate of success. Although we do not have significant results, probably due to the low number of samples, all parameters except the number of slices seem to have an impact on performance.

6.1 Future research lines

In our future endeavors, we are keen on delving deeper into histological discrimination, specifically in developing models capable of distinguishing between adenocarcinoma and squamous cell cancer. To attain this goal, our focus will be on exploring diverse methods of contrastive learning. This approach involves learning representations by contrasting positive samples (similar) against negative samples (dissimilar). By doing so, we can significantly increase the number of samples, enhancing the model's ability to discern features crucial for distinguishing lung cancer nodules. To further optimize this process, we may need to incorporate positional information for the samples. Rather than comparing all samples indiscriminately against each

other, considering their positions can refine the categorization process. Additionally, employing contrastive learning has the potential to enhance the model's robustness against variations in nodules, such as differences in size, shape, density, and even variations in scan acquisition parameters.

The RadioLung Dataset continues to grow as new cases are continuously added. In our ongoing efforts, we aim to conduct a comprehensive analysis of the impact of various parameters with the goal of establishing minimum requirements for clinical practice.

Moreover, we plan to integrate pre-trained transformers as generators of representation spaces, harnessing their capacity to capture intricate patterns and features. This integration will enable us to explore additional representation spaces by incorporating diverse perspectives, including various cuts of the nodules, such as coronal and sagittal views. Furthermore, we intend to extend our investigation to include volumetric data, providing a more holistic understanding of lung cancer characteristics. These advancements in data analysis and representation will contribute to refining the dataset and improving the accuracy and robustness of our methods in the context of lung cancer diagnosis

List of Publications

6.2 Journals

1. (Abstract) **G. Torres**, D. Gil, A.Rosell, S.Mena, C.Sanchez."Virtual Radiomics Biopsy for the Histological Diagnosis of Pulmonary Nodules - Intermediate Results of the RadioLung Project.". IJCARS 2023. (JCR: 3.42, Q2, 66/136, Medical imaging / SJR: 1, Q1, Engineering, Biomedical).

2. (Abstract) A Rosell Gratacos, S Baeza, S Garcia-Reina, JL Mate, I Guasch, I Nogueira, I Garcia-Olivé, **G. Torres**, C Sànchez-Ramos, D Gil (2022). Radiomics to increase the effectiveness of lung cancer screening programs. Radiolung preliminary results. European Respiratory Journal 60 (suppl 66). (JCR: 10.56, Q1, 3/59, Respiratory System / SJR: 3.50, Q1, Pulmonary and Respiratory Medicine).

3. (Abstract) A Rosell Gratacos, S Baeza, S Garcia-Reina, JL Mate, I Guasch, I Nogueira, I Garcia-Olivé, **G. Torres**, C Sànchez-Ramos, D Gil (2022). EP01.05-001 Radiomics to Increase the Effectiveness of Lung Cancer Screening Programs. Radiolung Preliminary Results. Journal of Thoracic Oncology 17 (9), S182. (JCR: 10.56, Q1, 3/59, Respiratory System / SJR: 2.969, Q1, Pulmonary and Respiratory Medicine).

4. Baeza, S., Gil, D., Garcia-Olivé, I., Salcedo-Pujantell, M., Deportós, J., Sanchez, C., **Torres G.**, Morogas G., Rosell, A. (2022). A novel intelligent radiomic analysis of perfusion SPECT/CT images to optimize pulmonary embolism diagnosis in COVID-19 patients. EJNMMI physics, 9(1), 1-17. (JCR: 4.65, Q2, 41/136, Medical imaging / SJR: 1.032, Q1, Biomedical Engineering).

5. **Torres, G.**, Baeza, S., Sanchez, C., Guasch, I., Rosell, A., Gil, D. (2022). An Intelligent Radiomic Approach for Lung Cancer Screening. Applied Sciences, 12(3), 1568. (JCR: 2.83, Q2, 39/92, Engineering / SJR: 0.507, Q2, Engineering).

6. (Abstract) **G.Torres**, D.Gil. "A multi-shape loss function with adaptive class balancing for the segmentation of lung structures". IJCAR, 15 (1), S154-55 (JCR: 2.924, Q2, 61/134, RADIOLOGY, NUCLEAR MEDICINE & MEDICAL IMAGING).

6.3 Proceedings

1. **G. Torres**, J. Rodriguez , A.Rosell, S.Mena, C.Sanchez, D. Gil."Prediction of Malignancy in Lung Cancer using several strategies for the fusion of Multi-Channel Pyradiomics Images". SYNASC-2023, Not published yet

2. **G. Torres**, C. Sanchez and D. Gil."Learning networks hyper-parameter using multi-objective optimization of statistical performance metrics". In 2022 24rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), pp. 233-238.

3. M. Ligero, **G. Torres**, C. Sanchez, et al. "Selection of Radiomics Features based on their Reproducibility". EMBC Proc. IEEE Proc, pp. 403-08. 2019.

6.4 Presentation in International Conferences

1. D. Gil, **G. Torres**, C.Sanchez." Transforming radiomic features into radiological words". ISBI 2023. April 18-23, 2023 Cartagena de Indias, Colombia, Poster.

2. **G. Torres**, D. Gil, A.Rosell, S.Mena, C.Sanchez."A radiomic biopsy for virtual histology of pulmonary nodules". ISBI 2023. April 18-23, 2023 Cartagena de Indias, Colombia, Poster.

3. **G. Torres**, D. Gil, A.Rosell, S.Mena, C.Sanchez."An Intelligent Radiomic Approach for Lung Cancer Screening". Deep Learning Barcelona Symposium 2023. Poster and oral.

4. **G. Torres**, D. Gil, C.Sanchez."Learning networks hyper-parameter using multi-objective optimization of statistical performance metrics". Deep Learning Barcelona Symposium 2023. Poster and oral.

5. **G. Torres**, D. Gil, A.Rosell, S.Mena, C.Sanchez."Virtual Radiomics Biopsy for the Histological Diagnosis of Pulmonary Nodules - Intermediate Results of the RadioLung Project.". CARS 2023. June 20–23, Munich, Germany. Oral.

6. **G. Torres**, C. Sanchez and D. Gil."Learning networks hyper-parameter using multi-objective optimization of statistical performance metrics". 24rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), September 12-15. Linz, Austria. Oral and Poster. Oral and Poster.

7. A. Rosell Gratacos, S. Baeza, S. Garcia-Reina, J. L. Mate, I. Guasch, I. Nogueira, I. Garcia-Olivé, **G. Torres**, C. Sànchez-Ramos, D. Gil. "Radiomics to increase the effectiveness of lung cancer screening programs. Radiolung preliminary results". ERS (European Respiratory Society (ERS) International Congress) - September 2022. Oral.

8. A. Rosell Gratacos, S. Baeza, S. Garcia-Reina, J. L. Mate, I. Guasch, I. Nogueira, I. Garcia-Olivé, **G. Torres**, C. Sànchez-Ramos, D. Gil. "Radiomics to increase the effectiveness of lung cancer screening programs. Radiolung preliminary results". WCLC (IASLC World Conference on Lung Cancer) - August 2022. Poster.

9. A. Rosell Gratacos, S. Baeza, S. Garcia-Reina, J. L. Mate, I. Guasch, I. Nogueira, I. Garcia-Olivé, **G. Torres**, C. Sànchez-Ramos, D. Gil. "Aplicació de la radiòmica en el diagnòstic del nòdul pulmonar. Resultat spreliminars del projecte Radiolung.". SOCAP (XXXIX Diada Pneumològica, Societat Catalana de Pneumologia) - May 2022. Oral.

10. A. Rosell Gratacos, S. Baeza, S. Garcia-Reina, J. L. Mate, I. Guasch, I. Nogueira, I. Garcia-Olivé, **G. Torres**, C. Sànchez-Ramos, D. Gil. "Ponencia proyecto RADIOLUNG". SEPAR (Sociedad española de neumología) - Jun 2022. Oral.

11. D. Gil, S. Baeza, C. Sanchez, **G. Torres**, I. Garcia-Olive, G. Moragas, J. Deportós, M. Salcedo, A. Rosell. (2021, October). "Intelligent Radiomic Analysis of Q-SPECT/CT images to optimize pulmonary embolism diagnosis in COVID-19 patients". 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 11-17 Oct. 2021. Montreal, Canada. Oral.

12. M. Ligero, **G. Torres**, C. Sanchez, et al. "Selection of Radiomics Features based on their Reproducibility". EMBC, July 23– 27, 2019 Berlin, Germany. Oral.

6.5 Public Repositories

1. The **RadioLung Database** is accessible to the public at the following URL: `http://iam.cvc.uab.es/portfolio/radiolung-database`.

2. Our dataset will be published in Cancer Imaging Archive at `https://wiki.cancerimagingarchive.net`.

6.6 Awards

1. Application of radiomics in the diagnosis of lung nodules. Preliminary results from the Radiolung project. SOCAP (XXXIX Pneumological Day, Catalan Society of Pneumology) – Best communication - May 2022.

Bibliography

[1] Cancer today. https://gco.iarc.fr/today/online-analysis-pie [Accessed on October 4, 2023].

[2] Eurostat - statistics explained. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Causes_of_death_statistics#Major_causes_of_death_in_the_EU_in_2020 [Accessed on September 20, 2023].

[3] Hounsfield unit. https://www.ncbi.nlm.nih.gov/books/NBK547721/.

[4] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5:4006, 2014.

[5] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, et al. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2019.

[6] Hussain Alibrahim and Simone A Ludwig. Hyperparameter optimization: comparing genetic algorithm against grid search and bayesian optimization. In *2021 IEEE Congress on Evolutionary Computation (CEC)*, pages 1551–1559. IEEE, 2021.

[7] Răzvan Andonie. Hyperparameter optimization in learning systems. *Journal of Membrane Computing*, 1(4):279–291, 2019.

[8] Nurshazlyn Mohd Aszemi and PDD Dominic. Hyperparameter optimization in convolutional neural network using genetic algorithms. *International Journal of Advanced Computer Science and Applications*, 10(6), 2019.

[9] Rana Bahij, Stefan Starup Jeppesen, Karen Ege Olsen, Ulrich Halekoh, Karin Holmskov, and Olfred Hansen. Outcome of treatment in patients with small cell lung cancer in poor performance status. *Acta Oncologica*, 58(11):1612–1617, 2019.

[10] Niha Beig, Mohammadhadi Khorrami, Mehdi Alilou, Prateek Prasanna, Nathaniel Braman, Mahdi Orooji, Sagar Rakshit, Kaustav Bera, Prabhakar Rajiah, Jennifer Ginsberg, et al. Perinodular and intranodular radiomic features on lung ct images distinguish adenocarcinomas from granulomas. *Radiology*, 290(3):783–792, 2019.

[11] Thierry Berghmans, Marianne Paesmans, and Jean-Paul Sculier. Prognostic factors in stage iii non-small cell lung cancer: a review of conventional, metabolic and new biological variables. *Therapeutic advances in medical oncology*, 3(3):127–138, 2011.

[12] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[13] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[14] Kinga Bernatowicz, Francesco Grussu, Marta Ligero, Alonso Garcia, Eric Delgado, and Raquel Perez-Lopez. Robust imaging habitat computation using voxel-wise radiomics features. *Scientific reports*, 11(1):1–8, 2021.

[15] James G. Booth. Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood by Y. Lee, J. A. Nelder, and Y. Pawitan. *Biometrics*, 63(4):1296–1297, 12 2007.

[16] Deng Cai, Xiaofei He, and Jiawei Han. Speed up kernel discriminant analysis. *The VLDB Journal—The International Journal on Very Large Data Bases*, 20(1):21–33, 2011.

[17] José Lucas Leite Calheiros, Lucas Benevides Viana de Amorim, Lucas Lins de Lima, Ailton Felix de Lima Filho, José Raniery Ferreira Júnior, and Marcelo Costa de Oliveira. The effects of perinodular features on solid lung nodule classification. *Journal of Digital Imaging*, pages 1–13, 2021.

[18] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

[19] Hakan Cevikalp, Marian Neamtu, and Atalay Barkana. The kernel common vector method: A novel nonlinear subspace classifier for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(4):937–951, 2007.

[20] Hakan Cevikalp, Marian Neamtu, and Mitch Wilkes. Discriminative common vector method with kernels. *IEEE Transactions on Neural Networks*, 17(6):1550–1565, 2006.

[21] Peter Pin-Shan Chen. The entity-relationship model—toward a unified view of data. *ACM transactions on database systems (TODS)*, 1(1):9–36, 1976.

[22] Davide Chicco. Ten quick tips for machine learning in computational biology. *BioData mining*, 10(1):1–17, 2017.

[23] Marc Claesen and Bart De Moor. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127*, 2015.

[24] Ángel Cobo. *PHP y MySQL: Tecnología para el desarrollo de aplicaciones web*. Ediciones Díaz de Santos, 2005.

[25] Edviges Coelho. *Eurostat Database*, pages 1–2. Springer International Publishing, Cham, 2020.

[26] Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *International conference on medical image computing and computer-assisted intervention*, pages 529–536. Springer, 2018.

[27] HJ de Koning, CM van der Aalst, PA de Jong, et al. Screening met een thoracale lage-dosis-ct-scan vermindert de sterfte na 10 jaar door longkanker bij mannelijke actieve of ex-rokers. *N Engl J Med*, 382:503–13, 2020.

[28] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

[29] Frauke Degenhardt, Stephan Seifert, and Silke Szymczak. Evaluation of variable selection methods for random forests and omics data sets. *Briefings in bioinformatics*, 20(2):492–503, 2019.

[30] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[32] Katerine Diaz-Chito, Jesús Martínez del Rincón, Aura Hernández-Sabaté, Marçal Rusiñol, and Francesc J Ferri. Fast Kernel Generalized Discriminative Common Vectors for Feature Extraction. *Journal of Mathematical Imaging and Vision*, 60(4):512–524, 2018.

[33] Paul DuBois. *MySQL*. New riders publishing, 1999.

[34] Ali Mohamed Abdelrazig Elmahdi. *Optimization of Radiation Dose in Multislices CT scan Examinations*. PhD thesis, sudan university of science and technology, 2011.

[35] Peikert T et al. Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the National Lung Screening Trial. *PLOS ONE*, 13 (10), 2018.

[36] Ben Frain. *Responsive web design with HTML5 and CSS3*. Packt Publishing Ltd, 2012.

[37] Fernando Franco, Enric Carcereny, Maria Guirado, Ana L Ortega, Rafael López-Castro, Delvys Rodríguez-Abreu, Rosario García-Campelo, Edel Del Barco, Oscar Juan, Francisco Aparisi, et al. Epidemiology, treatment, and survival in small cell lung cancer in spain: Data from the thoracic tumor registry. *PloS one*, 16(6):e0251761, 2021.

[38] Debora Gil, Carles Sanchez, Agnes Borras, Marta Diez-Ferrer, and Antoni Rosell. Segmentation of distal airways using structural analysis. *Plos one*, 14(12):e0226006, 2019.

[39] Piotr S Gromski, Howbeer Muhamadali, David I Ellis, Yun Xu, Elon Correa, Michael L Turner, and Royston Goodacre. A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. *Analytica chimica acta*, 879:10–23, 2015.

[40] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973.

[41] Thomas Hofmarcher, Peter Lindgren, Nils Wilking, and Bengt Jönsson. The cost of cancer in europe 2018. *European Journal of Cancer*, 129:41–49, 2020.

[42] Yoonki Hong, Sunmin Park, and Myoung Kyu Lee. The prognosis of non-small cell lung cancer patients according to endobronchial metastatic lesion. *Scientific Reports*, 12(1):13588, 2022.

[43] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

[44] Chih-Ling Huang, Meng-Jia Lian, Yi-Hsuan Wu, Wei-Ming Chen, and Wen-Tai Chiu. Identification of human ovarian adenocarcinoma cells with cisplatin-resistance by feature extraction of gray level co-occurrence matrix using optical images. *Diagnostics*, 10(6):389, 2020.

[45] Hanliang Jiang, Fei Gao, Xingxin Xu, Fei Huang, and Suguo Zhu. Attentive and ensemble 3d dual path networks for pulmonary nodules classification. *Neurocomputing*, 398:422–430, 2020.

[46] Hanliang Jiang, Fuhao Shen, Fei Gao, and Weidong Han. Learning efficient, explainable and discriminative representations for pulmonary nodules classification. *Pattern Recognition*, 113:107825, 2021.

[47] Young Jae Kim, Hyun-Ju Lee, Kwang Gi Kim, and Seung Hyun Lee. The effect of ct scan parameters on the measurement of ct radiomic features: a lung nodule phantom study. *Computational and mathematical methods in medicine*, 2019, 2019.

[48] Stephen Lam, Murry W Wynes, Casey Connolly, Kazuto Ashizawa, Sukhinder Atkar-Khattra, Chandra P Belani, Domenic DiNatale, Claudia I Henschke, Bruno Hochhegger, Claudio Jacomelli, et al. The iaslc early lung imaging confederation (elic) open-source deep learning and quantitative measurement initiative. *Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer*, pages S1556–0864, 2023.

[49] Philippe Lambin, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben THM Larue, Aniek JG Even, Arthur Jochems, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12):749–762, 2017.

[50] Hyoun Wook Lee, Seung Yeon Ha, and Mee Sook Roh. Non-small cell carcinoma-not otherwise specified on cytology specimens in patients with solitary pulmonary lesion: Primary lung cancer or metastatic cancer? *Journal of Cytology*, 38(1):8, 2021.

[51] Shu Ling Alycia Lee, Abbas Z Kouzani, and Eric J Hu. Random forest based lung nodule classification aided by clustering. *Computerized medical imaging and graphics*, 34(7):535–542, 2010.

[52] Ralph TH Leijenaar, Georgi Nalbantov, Sara Carvalho, Wouter Jc Van Elmpt, Esther GC Troost, Ronald Boellaard, Hugo JWL Aerts, Robert J Gillies, and Philippe Lambin. The effect of suv discretization in quantitative fdg-pet radiomics: the need for standardized methodology in tumor texture analysis. *Scientific reports*, 5(1):1–10, 2015.

[53] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: Principal component analysis, 2017.

[54] Bi-Qing Li, Jin You, Tao Huang, and Yu-Dong Cai. Classification of non-small cell lung cancer based on copy number alterations. *PLoS One*, 9(2):e88300, 2014.

[55] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, 19(2):325–340, 2016.

[56] Marta Ligero, Guillermo Torres, Carles Sanchez, Katerine Diaz-Chito, Raquel Perez, and Debora Gil. Selection of radiomics features based on their reproducibility. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 403–408. IEEE, 2019.

[57] Qingshan Liu, Jian Cheng, Hanqing Lu, and Songde Ma. Modeling face appearance with nonlinear independent component analysis. In *null*, page 761. IEEE, 2004.

[58] Ying Liu, Jongphil Kim, Yoganand Balagurunathan, et al. Prediction of pathological nodal involvement by ct-based radiomic features of the primary tumor in patients with clinically node-negative peripheral lung adenocarcinomas. *Medical physics*, 45(6):2518–2526, 2018.

[59] AS Lowe and CL Kay. Recent developments in ct: a review of the clinical applications and advantages of multidetector computed tomography. *Imaging*, 18(2):62–67, 2006.

[60] Xiangjuan Ma, Ziran Zhang, Xiaoling Chen, Jie Zhang, Jun Nie, Ling Da, Weiheng Hu, Guangming Tian, Di Wu, Jindi Han, et al. Prognostic factor analysis of patients with small cell lung cancer: real-world data from 988 patients. *Thoracic Cancer*, 12(12):1841–1850, 2021.

[61] Gabriel Maicas, Andrew P Bradley, Jacinto C Nascimento, Ian Reid, and Gustavo Carneiro. Training medical image analysis systems like radiologists. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 546–554. Springer, 2018.

[62] BETTIO Manola, RANDI Giorgia, NEGRAO DE CARVALHO Raquel, MARTOS JIMENEZ Maria Del Carmen, DYBA Tadeusz Artur, NICHOLSON Nicholas, FLEGO Manuela, NEAMTIU Luciana, GIUSTI Francesco, ASLANOVSKI Davor, et al. Lung cancer burden in eu-27. 2021.

[63] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–663. Springer, 2018.

[64] Andrew G Nicholson, Ming S Tsao, Mary Beth Beasley, Alain C Borczuk, Elisabeth Brambilla, Wendy A Cooper, Sanja Dacic, Deepali Jain, Keith M Kerr, Sylvie Lantuejoul, et al. The 2021 who classification of lung tumors: impact of advances since 2015. *Journal of Thoracic Oncology*, 17(3):362–387, 2022.

[65] National Health Commission of the People et al. National guidelines for diagnosis and treatment of lung cancer 2022 in china (english version). *Chinese Journal of Cancer Research*, 34(3):176, 2022.

[66] Yoshiharu Ohno, Hisanobu Koyama, Astushi Kono, Mari Terada, Hiroyasu Inokawa, Sumiaki Matsumoto, and Kazuro Sugimura. Influence of detector collimation and beam pitch for identification and image quality of ground-glass attenuation and nodules on 16-and 64-detector row ct systems: experimental study using chest phantom. *European journal of radiology*, 64(3):406–413, 2007.

[67] Tobias Peikert, Fenghai Duan, Srinivasan Rajagopalan, Ronald A Karwoski, Ryan Clay, Richard A Robb, Ziling Qin, JoRean Sicks, Brian J Bartholmai, and Fabien Maldonado. Correction: Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the national lung screening trial. *PloS one*, 13(10):e0205311, 2018.

[68] Marc Pomeroy, Hongbing Lu, Perry J Pickhardt, and Zhengrong Liang. Histogram-based adaptive gray level scaling for texture feature classification of colorectal polyps. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105752A. International Society for Optics and Photonics, 2018.

[69] Manca Povsic, Ashley Enstone, Robin Wyn, Klaudia Kornalska, John R Penrod, and Yong Yuan. Real-world effectiveness and tolerability of small-cell lung cancer (sclc) treatments: a systematic literature review (slr). *PloS one*, 14(7):e0219622, 2019.

[70] M Radovic, M Ghalwash, N Filipovic, and Z Obradovic. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, pages 18–19, 2017.

[71] Konstantinos Rounis, Dimitrios Makrakis, Chara Papadaki, Alexia Monastirioti, Lambros Vamvakas, Konstantinos Kalbakis, Krystallia Gourlia, Iordanis Xanthopoulos, Ioannis Tsamardinos, Dimitrios Mavroudis, et al. Correction: Prediction of outcome in patients with non-small cell lung cancer treated with second line pd-1/pdl-1 inhibitors based on clinical parameters: Results from a prospective, single institution study. *Plos one*, 18(11):e0294382, 2023.

[72] Charles M Rudin, Elisabeth Brambilla, Corinne Faivre-Finn, and Julien Sage. Small-cell lung cancer. *Nature Reviews Disease Primers*, 7(1):3, 2021.

[73] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

[74] Yanbo Shao, Minghao Wang, Juanyun Mai, Xinliang Fu, Mei Li, Jiayin Zheng, Zhaoqi Diao, Airu Yin, Yulong Chen, Jianyu Xiao, et al. Lidp: A lung image dataset with pathological information for lung cancer screening. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 770–779. Springer, 2022.

[75] Wei Shen, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, 61:663–673, 2017.

[76] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, Ahmedin Jemal, et al. Cancer statistics, 2021. *Ca Cancer J Clin*, 71(1):7–33, 2021.

[77] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[78] Roger Sun, Elaine Johanna Limkin, Maria Vakalopoulou, Laurent Dercle, Stéphane Champiat, Shan Rong Han, Loïc Verlingue, David Brandao, Andrea Lancia, Samy Ammari, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *The Lancet Oncology*, 19(9):1180–1191, 2018.

[79] Baskaran Sundaram, Aamer R Chughtai, and Ella A Kazerooni. Multidetector high-resolution computed tomography of the lungs: protocols and applications. *Journal of thoracic imaging*, 25(2):125–141, 2010.

[80] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[81] Jiaxing Tan, Yongfeng Gao, Zhengrong Liang, Weiguo Cao, Marc J Pomeroy, Yumei Huo, Lihong Li, Matthew A Barish, Almas F Abbasi, and Perry J Pickhardt. 3d-glcm cnn: A 3-dimensional gray-level co-occurrence matrix-based cnn model for polyp classification via ct colonography. *IEEE transactions on medical imaging*, 39(6):2013–2024, 2019.

[82] National Lung Screening Trial Research Team. The national lung screening trial: overview and study design. *Radiology*, 258(1):243–253, 2011.

[83] Toby J Teorey, Sam S Lightstone, Tom Nadeau, and HV Jagadish. *Database modeling and design: logical design*. Elsevier, 2011.

[84] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–88, 1996.

[85] Florent Tixier, Catherine Cheze Le Rest, Mathieu Hatt, Nidal Albarghach, Olivier Pradier, Jean-Philippe Metges, Laurent Corcos, and Dimitris Visvikis. Intratumor heterogeneity characterized by textural features on baseline 18f-fdg pet images predicts response to concomitant radiochemotherapy in esophageal cancer. *Journal of Nuclear Medicine*, 52(3):369–378, 2011.

[86] Selene Tomassini, Nicola Falcionelli, Paolo Sernani, Laura Burattini, and Aldo Franco Dragoni. Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: A survey. *Computers in Biology and Medicine*, 146:105691, 2022.

[87] Guillermo Torres, Sonia Baeza, Carles Sanchez, et al. An intelligent radiomic approach for lung cancer screening. *Applied Sciences*, 2022.

[88] Joost JM van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[89] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):1–8, 2006.

[90] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[91] Shanu Verma, Millie Pant, and Vaclav Snasel. A comprehensive review on nsga-ii for multi-objective combinatorial optimization problems. *Ieee Access*, 9:57757–57791, 2021.

[92] Shanu Verma, Millie Pant, and Vaclav Snasel. A comprehensive review on nsga-ii for multi-objective combinatorial optimization problems. *IEEE Access*, 9:57757–57791, 2021.

[93] Jacques Wainer and Gavin Cawley. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, 182:115222, 2021.

[94] Jonathan Waring, Charlotta Lindvall, and Renato Umeton. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial intelligence in medicine*, 104:101822, 2020.

[95] Jia Wu, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, and Si-Hao Deng. Hyperparameter optimization for machine learning models based on bayesian optimization. *Journal of Electronic Science and Technology*, 17(1):26–40, 2019.

[96] Hongtao Xie, Dongbao Yang, Nannan Sun, Zhineng Chen, and Yongdong Zhang. Automated pulmonary nodule detection in ct images using deep convolutional neural networks. *Pattern Recognition*, 85:109–119, 2019.

[97] Tao Xiong, Jieping Ye, Qi Li, Ravi Janardan, and Vladimir Cherkassky. Efficient kernel discriminant analysis via QR decomposition. In *Advances in neural information processing systems*, pages 1529–1536, 2005.

[98] Yan Xu, Lin Lu, Shawn H Sun, Wei Lian, Hao Yang, Lawrence H Schwartz, Zhenghan Yang, Binsheng Zhao, et al. Effect of ct image acquisition parameters on diagnostic performance of radiomics in predicting malignancy of pulmonary nodules of different sizes. *European Radiology*, pages 1–11, 2021.

[99] Xingjian Yan, Jianing Pang, Hang Qi, Yixin Zhu, Chunxue Bai, Xin Geng, Mina Liu, Demetri Terzopoulos, and Xiaowei Ding. Classification of lung nodule malignancy risk on computed tomography images using convolutional neural network: A comparison between 2d and 3d strategies. In *Asian Conference on Computer Vision*, pages 91–101. Springer, 2016.

[100] Ming-Hsuan Yang. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In *fgr*, page 0215. IEEE, 2002.

[101] Fan Zhang, Yang Song, Weidong Cai, Min-Zhao Lee, Yun Zhou, Heng Huang, Shimin Shan, Michael J Fulham, and Dagan D Feng. Lung nodule classification with multilevel patch-based context analysis. *IEEE Transactions on Biomedical Engineering*, 61(4):1155–1166, 2013.

[102] Ying Ru Zhao, Xueqian Xie, Harry J de Koning, Willem P Mali, Rozemarijn Vliegenthart, and Matthijs Oudkerk. Nelson lung cancer screening study. *Cancer Imaging*, 11(1A):S79, 2011.

[103] Wentao Zhu, Chaochun Liu, Wei Fan, and Xiaohui Xie. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 673–681. IEEE, 2018.