FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Overcoming scarce annotations through deep semi-supervised learning in cancer characterisation

Afonso José Pinheiro Oliveira Esteves Abreu

Master's in Informatics and Computing Engineering

Supervisor: Hélder Oliveira

Co-Supervisor: Tânia Pereira

Co-Supervisor: Eduardo de Matos Rodrigues

October 1, 2025

Overcoming scarce annotations through deep semi-supervised learning in cancer characterisation

Afonso José Pinheiro Oliveira Esteves Abreu

Master's in Informatics and Computing Engineering

Approved in oral examination by the committee:

President: Prof. Rui Camacho
Referee: Prof. Joel Arrais
Referee: Prof. Tânia Perreira

October 1, 2025

Resumo

O cancro do pulmão continua a ser uma das causas mais comuns de mortalidade em todo o mundo. Isso deve-se principalmente ao diagnóstico tardio da doença, acompanhado por baixas taxas de sobrevivência. Atualmente, o padrão de ouro para a caracterização da malignidade tumoral é a biópsia. No entanto, ela apresenta várias limitações, como o facto de ser invasiva e de ter riscos clínicos significativos. Além disso, os sistemas de diagnóstico assistido por computador (CAD) baseados em aprendizagem profunda que utilizam tomografias computadorizadas (TCs)para auxiliar na tomada de decisões clínicas surgiram como opções não invasivas que oferecem uma alternativa mais promissora.

Na maioria dos casos, o sucesso dos modelos de aprendizagem profunda depende de bases de dados muito grandes de imagens médicas anotadas. A recolha de tal quantidade de informação é demorada, cara e também trabalhosa, o que pode ser uma barreira crítica à adoção generalizada de sistemas CAD. O desafio dos dados não rotulados em repositórios de imagens médicas pode ser resolvido com a aprendizagem semi-supervisionada, uma vez que permite a inclusão de dados não rotulados no processo de treino.

Esta dissertação aplica a estrutura FixMatch, um método semi-supervisionado que combina pseudo-rotulagem com regularização de consistência, para classificar a malignidade dos nódulos pulmonares. A avaliação da estrutura em dados de TC, tanto rotulados do conjunto de dados LIDC-IDRI como não rotulados do conjunto de dados Luna25, apresentou uma análise de explicabilidade com Gradient-weighted Class Activation Mapping (Grad-CAM) para comparar a interpretabilidade dos modelos totalmente supervisionados e semi-supervisionados.

Os resultados mostraram que a incorporação de dados não rotulados melhorou o desempenho em relação a uma linha de base puramente supervisionada, com um aumento de 2% no AUROC ao usar uma proporção de 5:1 de dados não rotulados para rotulados. No entanto, a análise de explicabilidade não promoveu conclusões sustentáveis no contexto clínico, apesar das observações de mapas de ativação mais localizados do modelo semi-supervisionado.

Este trabalho contribui para o esforço de melhorar a deteção precoce do cancro do pulmão e a expansão do acesso aos cuidados de saúde. A integração e a experimentação de uma estrutura semi-supervisionada ajudam no objetivo de desenvolvimento contínuo de ferramentas de IA que podem ter um papel importante na área médica.

Abstract

Lung cancer continues to be one of the most common causes of mortality across the globe. This is primarily owing to the late-stage diagnosis of the illness, accompanied by low survival rates. Currently, the gold standard for tumour malignancy characterization is tissue biopsy. However, it comes with various limitations, such as being invasive, potentially having significant clinical risks. Furthermore, Computer-Aided Diagnosis (CAD) systems based on deep learning that utilize Computed Tomography scans to assist in clinical decision-making have emerged as noninvasive options that offer a more promising alternative.

In most cases, deep learning models' success relies on the very large annotated databases of medical images. Gathering such an amount of information is time-consuming, expensive, and also labour-intensive, which can be a critical barrier to the widespread adoption of CAD systems. The challenge of unlabeled data in medical imaging repositories can be solved with semi-supervised learning, as it allows the inclusion of unlabeled data into the training process.

This dissertation applies the FixMatch framework, a semi-supervised method that combines pseudo-labeling and consistency regularization, to classify lung nodules' malignancy. The framework evaluation on CT data, both labeled from the LIDC-IDRI dataset and unlabeled from the Luna25 dataset, featured an explainability analysis with Gradient-weighted Class Activation Mapping (Grad-CAM) to compare the fully-supervised and semi-supervised models' interpretability.

Results showed that incorporating unlabeled data improved performance over a purely supervised baseline, with an increase of 2% in AUROC when using a 5:1 ratio of unlabeled to labeled data. Nevertheless, the explainability analysis didn't promote sustainable conclusions in the clinical context, despite observations of more localized activation maps from the semi-supervised model.

This work contributes to the effort to improve the early detection of lung cancer and the expansion of healthcare access. Integrating and experimenting with a semi-supervised framework helps with the goal of continuous development of AI tools that could have a major role in the medical area.

UN Sustainable Development Goals

The United Nations Agenda for Sustainable Development proposes the most critical social, economic, and environmental issues with a singular framework for all countries. In this framework, the Sustainable Development Goals (SDGs) delineate solid objectives for innovation and equity, and promote health services globally. This dissertation focuses on SDG 3 (Good Health and Well-Being) and SDG 9 (Industry, Innovation, and Infrastructure) most prominently.

The research making progress towards SDG 3 focuses on automated methods for early lung cancer detection through non-invasive and AI-based techniques, therefore improving health outcomes and increasing access to vital healthcare technologies. It also works towards SDG 9 by driving innovation in artificial intelligence through the application of semi-supervised learning and explainable AI, utilizing publicly available datasets.

The specific Sustainable Development Goals mentioned have the following names:

SDG 3 Ensure healthy lives and promote well-being for all at all ages

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation

SDG	Target	Contribution	Performance Indicators and Metrics
3	**3.4** "By 2030, reduce by one-third premature mortality from non-communicable diseases through prevention, treatment, and promotion of mental health and well-being."	The proposed framework improves early detection of lung cancer through semi-supervised CAD, offering a less invasive alternative to biopsies and supporting timely treatment.	Improvement in performance using deep learning metrics
	3.8 "Achieve universal health coverage, including access to quality healthcare services, medicines, and technologies."	By reducing the reliance on large annotated datasets, this work enables larger adoption of CAD systems in resource-limited healthcare settings.	Percentage of institutions adopting CAD with fewer annotations; Reduction in annotation costs measured in expert hours saved.
	3.b "Support the research and development of vaccines and medicines for communicable and non-communicable diseases, ensuring equitable access."	The research advances AI-based diagnostic tools, using semi-supervised learning, encouraging innovation in cancer malignancy characterization, and expanding the use of open datasets.	Use of the technology in clinical contexts; Datasets employed.
9	**9.5** "Enhance scientific research, upgrade the technological capabilities of industrial sectors, and encourage innovation."	This work contributes to scientific experimentation by applying the FixMatch framework in several ablation studies.	Number of ablation studies conducted.

Acknowledgements

It's a weird feeling knowing these five years are coming to an end. On one side, it is a healthy relief, with a bit of a taste of nostalgia from the initial years in this faculty. On the other side, I can also feel reluctant about what might come next. All of this, and it seems that only an instant had passed.

My first words are destined to my family, who were the ones to support me in enrolling in this course, and always knew about my capabilities. Without you, who educated me to become the person I am today, the path I have taken could never have been achieved.

To Marta, the person whom fate has chosen to live with me for the rest of my life, and with a special note to her family, which is now part of mine. Not enough words can describe how much you cared, and how much time you spent just to give me strength and to help me accomplish my personal goals, this dissertation being one of them. You are my best support, in good and bad times, always with something nice to say. I hope you know how special and important your presence is for me.

To all my close friends, thank you for being part of my second home, for all the help you gave me, and for all the times you knew I needed a break; those are important too! I can't imagine how these years would end up if you weren't by my side; you made my life happier.

Lastly, a big thank you to my supervisor team, Hélder, Tânia, and especially Eduardo, for all the time you spent clearing all my doubts, which I know were a lot and required a handful amount of patience. Thank you for making this dissertation possible.

I can say I feel accomplished for finishing this work, and lucky for having the right people around me. May this be just one beautiful chapter in the rest that is yet to come.

Afonso Abreu

"We are all in the gutter, but some of us are looking at the stars."

Oscar Wilde, Lady Windermere's Fan

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
AUROC	Area Under the Receiver Operating Characteristic Curve
BCE	Binary Cross-Entropy
BB	Bounding Box
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CT	Computed Tomography
CUDA	Compute Unified Device Architecture
DL	Deep Learning
EGFR	Epidermal Growth Factor Receptor
GAN	Generative Adversarial Network
Grad-CAM	Gradient-weighted Class Activation Mapping
HU	Hounsfield Unit
LIDC-IDRI	Lung Image Database Consortium and Image Database Resource Initiative
LNM	Lung Nodule Malignancy
LNVA	Lung Nodule Visual Attributes
NLST	National Lung Screening Trial
NSCLC	Non-Small Cell Lung Cancer
PET	Positron Emission Tomography
ReLU	Rectified Linear Unit
ROI	Region of Interest
SCLC	Small Cell Lung Cancer
SLURM	Simple Linux Utility for Resource Management
SSL	Semi-Supervised Learning

ViT	Vision Transformer
VAE	Variational Autoencoder

Chapter 1

Introduction

1.1 Context

Among all types of cancer, lung cancer is the one that causes the most damage in human health [11] as it is the 6th leading cause of death of our species, along with trachea and bronchus [28]. The reason for this remains in the fact that lung cancer is often diagnosed in a late stage of the disease. It is notable that, during this period, less than 10% of people survive the 5-year survival rate, a metric that reflects the percentage of people still alive after five years of being diagnosed with this disease [46]. This imposes a major challenge in medical treatments as newer methods require earlier detection of lung cancer to effectively treat it.

Tissue biopsy has been one of the the main methods to identify and characterise lung cancer [48]. However, since this method requires a piece of tissue to be taken for tests, it is considered an invasive procedure, potentially leading to complications such as pneumothorax, infections, hemorrhage, and damage to the tissue [52].

Computer-aided diagnosis (CAD) has gained a better reputation over time as it can come with deep learning models able to provide a non-invasive tumour characterisation based on computed tomography (CT) scans of the human thorax [45]. This would counteract the effects of the biopsy, giving also more information about the disease more accurately.

The use of deep learning could be a major improvement in decision-making support. Being able to deeply understand the characteristics of a tumor, combined with the use of a non invasive procedure can help clinicians to develop targeted therapies which would then be more effective and harmless.

1.2 Motivation

An early diagnosis of lung cancer has proven to dramatically increase a patient's chance of living after that period, increasing the 5-year survival rate metric mentioned above.

The malignancy status of a lung tumor is traditionally detected with molecular tests, which come from tissue that was extracted from the body during the biopsy procedure. The problem comes from the fact that this technique is considered invasive for the patient, as it can lead to physical complications or sometimes be even impossible to conduct.

Recent advancements have introduced new, less invasive techniques such as Computer-aided diagnosis (CAD). It is focused on deep learning models, capable of learning from large amounts of data, typically derived from CT scans. These images are widely utilized in oncology, especially in lung cancer, as they can highlight crucial features from the thorax and therefore, the tumor itself, sometimes not visible to the human eye. By designing systems capable of predicting the malignancy status of several mutations, it is possible to counteract the invasive nature of traditional procedures in cancer characterization. This achievement can be a step toward the utilization of CAD systems in Medicine, potentially evolving to the prediction of cancer mutation types. It would help deliver targeted therapies to patients while suppressing even further the need for the invasive biopsy procedure.

Recent studies about the matter suffered from one common limitation, referring to the small databases containing CT scans accompanied by molecular results for mutation status. The reason for this comes from the fact that, to provide annotations for CT scans, it is necessary to have the analysis of professionals in the area, which can easily become very time-consuming and, at the same time, have high costs attached. Other larger databases with much more information also exist; however, they often do not contain the intended labels.

All these limitations urge the need for a framework capable of taking advantage of both types of databases. Semi-supervised learning algorithms, which combine labeled and unlabeled data, may be the answer to combat this problem and to achieve a high predictive model capable of assessing the malignancy status of several lung tumors.

1.3 Objectives

The primary objective of this thesis is to develop and evaluate a semi-supervised learning framework able to correctly predict the malignancy status of pulmonary nodules, making use of CT images. Specifically, the goal is to distinguish between benign and malignant nodules while also exploiting the larger collection of unlabeled datasets to improve classification performance in low-label regimes.

To achieve these objectives, a FixMatch algorithm was chosen for the approach. This pipeline aims to demonstrate the feasibility of reducing the annotation weight in developing deep learning frameworks, contributing to scalable, non-invasive AI tools in early lung cancer screening.

1.4 Research questions

For the development of this dissertation, four research questions were broken down to clarify our hypothesis and to stratify an objective path, in order to create solid conclusions about the investigated topic.

1. **To what extent does the combination of supervised and unsupervised learning within a semi-supervised framework improve model performance for lung cancer malignancy classification, compared to using only supervised learning?**
2. **What is the optimal ratio of labeled to unlabeled data in the dataset(s) that maximizes the performance of the semi-supervised learning algorithm?**
3. **How do ablation study results vary with changes in conditions such as bounding-box definitions, pseudo-label thresholds, different CNNs, and different loss functions?**
4. **To what extent can explainability techniques like Grad-CAM reveal differences in the decision-making process between models trained with supervised learning and those trained using a semi-supervised algorithm for lung cancer malignancy classification?**

1.5 Project Structure

The organization of this document comprises six different chapters. The first one serves as an introduction to the theme of this thesis. It contains the context, the motivation for the development of this project, as well as the objectives that are aimed to be achieved. The second chapter is a detailed State of the Art that reviews existing literature on deep semi-supervised learning methods for medical CT scans, focused on cancer characterization. After that, a Background chapter describes a clinical framework, covering the lung, its diseases related to cancer, and current diagnostic techniques. It also gives a general overview of deep learning paradigms relevant for this work. Chapter four describes all the materials used in the project. From the utilized datasets to external environments and libraries that supported the creation of the pipeline. Chapter five details the core methodology and the description of the experimental design. Results and their discussion come next, providing a strong analysis and evaluation of the proposed semi-supervised approach. Lastly, a chapter of conclusions explains the achievements of this thesis and its contributions, as well as some notes about future work related to this theme.

Chapter 2

State of the Art

2.1 Introduction

The subject of this thesis isn't new to the field of Medicine. The exponential growth of AI has led many researchers to try and develop ways to combat the most debated diseases, giving a special attention to cancer, using many forms of deep learning.

This chapter reviews numerous research papers available on scientific databases that investigate the application of semi-supervised learning for classifying medical CT images. Analyzing these articles provides insights into data preparation, algorithm selection, and evaluation methods, allowing us to assess the benefits semi-supervised learning brought to this field's scope.

With this valuable information, we can take one step forward in designing our methodology, learning what went well and what could be improved in others' research, so that we can develop a cohesive and well-structured solution.

2.2 AI-based solutions with Semi-supervised learning

O. Kurasova et al. [18] developed a technique to detect pancreatic cancer while having a limited number of annotated images. The images were taken from several datasets (CT scans from Vilnius University Hospital Santaros Klinikos, the Memorial Sloan Kettering Cancer Center dataset, and the TCIA dataset), and in order to achieve high classification accuracy, the preprocessing phase consisted of cropping the original labeled images into a region of interest area which were then cropped again into smaller patches. These patches were later labeled the same as the original image. The algorithm used was a CNN (Convolutional Neural Network)-based model, which was

first trained using the labeled data. After that, the trained model was used to predict pseudo-labels for unlabeled images if the prediction probability PP was above a given threshold t. This process would repeat until no data remained or if the condition was not met anymore. Results were measured using the F1 score, and the highest achieved was 0.9 using the Santaros dataset (CT scans from Vilnius University Hospital Santaros Klinikos). Results also showed that adding the pseudo-labeled images resulted in better and stable training and improved classification metrics, compared to a supervised learning model only.

Furthermore, Roberto Philippi and Danilo Silva [24], regarding the classification of chest X-ray images, utilized a teacher-student pipeline, where two models are used in a multistep training algorithm so that the unlabeled data could take its role. In this framework, the unlabeled data is given pseudo labels by the teacher model, previously trained using the labeled data, which is then processed by the student model. The objective of this approach was to measure how much the data without labeling information could improve the performance compared to a fully supervised algorithm. Only high-confidence prediction values were taken into account when labeling the images in order to decrease noisy labels. The dataset used was the ChestX-ray14, and various percentages of labeled data were experimented with. The results were calculated using the AUROC measure and using a range of 2% to 20% of labeled data. The values went from 0.750 to 0.887 for the Teacher model and from 0.822 to 0.893 for the Student model. The largest increase in the AUROC value occurred when the percentage of labeled data went from 2% to 5%.

Now focusing on the structure of the lung, Phuong Nguyen et al. [26] presented an active, semi-supervised algorithm called ASEM-CAD (Active Semi-supervised Expectation Maximization for Computer-aided diagnosis). It starts by training the model using only labeled data evaluated by experts. After creating the first model, it assigns labels to the non-labeled images, which are then used for training. The expectation-maximization algorithm is used to estimate the maximum likelihood of labels that weren't observed, given the current model.

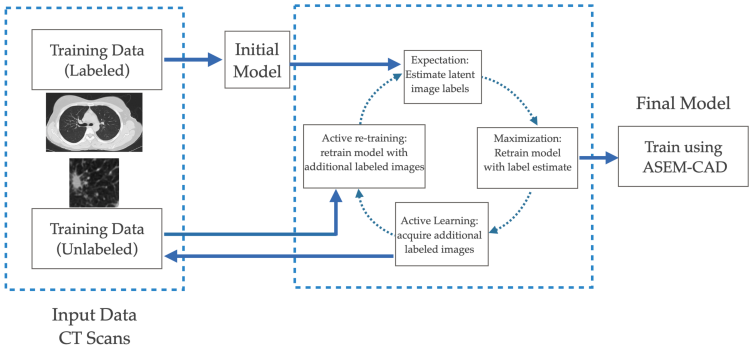

Figure 2.1: ASEM-CAD framework.

It used three public CT scans datasets: the National Lung Screening Trial (NLST), the Lung Image Database Consortium (LIDC), and Kaggle Data Science Bowl 2017 for lung cancer classification using CT scans. The results were evaluated using the AUC measurement, and the numbers were 0.94 (Kaggle), 0.95 (NLST), and 0.88 (LIDC), using significantly fewer labeled images (52%

to 59%) compared to a fully supervised learning model.

In F. Shi et al. [37] it is proposed a semi-supervised deep transfer learning (SDTL) framework to identify benign-malignant pulmonary nodules. It works by using a pre-trained nodule identification model and by adopting a semi-supervised learning method with iterations. Feature similarity is calculated between labeled and unlabeled data, and the unannotated images with more confidence are added to the training process gradually.

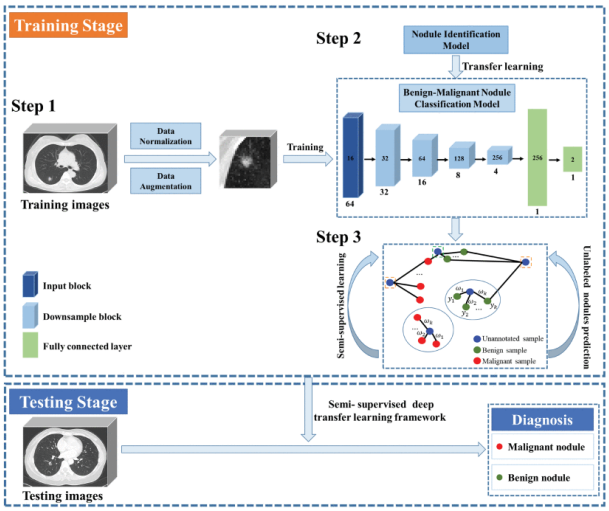

Figure 2.2: STDL framework.

Results showed that this framework achieves an accuracy of 88.3% and an AUC of 91.0% in the main dataset and an accuracy of 74.5% and an AUC of 79.5% in an independent dataset for testing. Furthermore, it was observed that the transfer learning technique and the use of semi-supervised learning contributed to 2% and 2.9% accuracy improvement, respectively.

Similarly, G. Wang et al. [50] proposed a novel convolutional neural network called PF-NET combined with a semi-supervised learning method using I-CRAWL (Iterative Confidence-based Refinement And Weighting of pseudo Labels). The PF-NET uses 2D and 3D convolutions, and the I-CRAWL utilizes pixel-level uncertainty-based confidence to get more accurate pseudo labels. Experiments were made with scans of Rhesus Macaques, and results showed a Dice score of 70.36% in PF-NET, outperforming other 2D, 3D, and hybrid CNNs. In I-CRAWL it was observed a gain between 0.6% to 3% was observed with the use of 10% -50% of annotated data, which also outperformed other semi-supervised methods like CRF-based or mean-teacher approaches.

In a paper made by Cláudia Pinheiro [31] exploited the power of adversarial training and used a combination of a Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) to incorporate labeled and unlabeled images. Due to the difficulty of training GANs, regarding convergence and stability, the idea of a regular adversarial network with random vectors as a starting point was discarded. Instead, it was used shared network of GAN and VAE, where the decoder of the VAE acts as the GAN generator.

This work made use of 3 datasets, NSCLC-Radiogenomics and UHCSJ, for labeled images

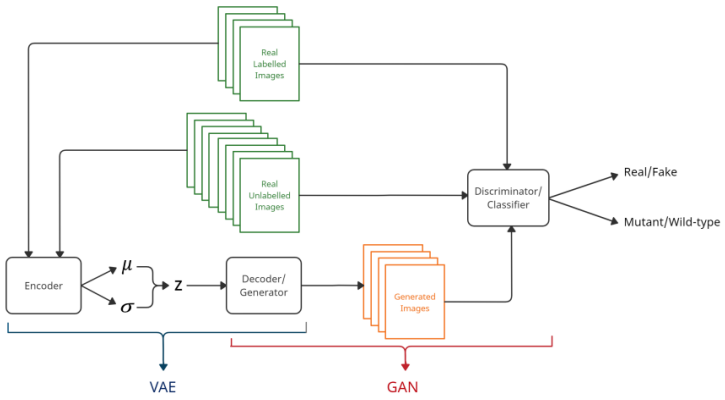

Figure 2.3: Combination of Variational Autoencoder and Generative Adversarial Network, the method by *Cláudia Pinheiro*.

and the other, much larger, NLST, for unlabeled images. The best results were achieved using a weighted loss function and got an AUC value of 0.7011, using 14% of labeled data. The semi-supervised learning method improved the discrimination ability by nearly 7% over a fully supervised model.

Yutong Xie et al. [53] proposed another semi-supervised adversarial classification (SSAC) model. It consists on an unsupervised reconstruction network, a supervised classification network, and learnable transition layers, adapting image representation ability from the first to the latter. The paper also develops an extended MK-SSAC (Multi-View Knowledge-Based SSAC) model by deploying 27 submodels, each assessing a nodule's overall appearance, shape heterogeneity, and voxel heterogeneity. It also operates across nine planar views, three orthogonal and six diagonal. The MK-SSAC model was evaluated in the LIDC-IDRI dataset and achieved an accuracy of 92.53% and an AUC of 95.81%. Using the unlabeled data, this model was also able to outperform the fully-supervised one.

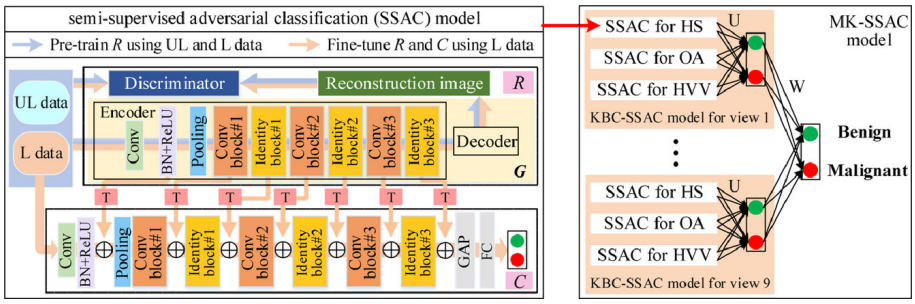

Figure 2.4: SSAC model. "UL": unlabeled data; "L": labeled data; "R": adversarial autoencoder-based reconstruction network; "C": classification network; and "G": generator, which contains an encoder and decoder. "T": learnable T layer.

To study the relevance of certain features regarding the Epidermal growth factor receptor (EGFR) mutation status, F. Silva et al. [39] used three different regions of interest: the nodule, the lung containing the main nodule, and both lungs. For that, a framework was developed con-

taining a CAE using unsupervised learning utilizing the LIDC-IDRI dataset. Taking advantage of Transfer Learning, the necessary knowledge gained would then be used by a multi-layer perceptron (MLP) for the classification task of the tumor, utilizing the NSCLC-Radiogenomics dataset.

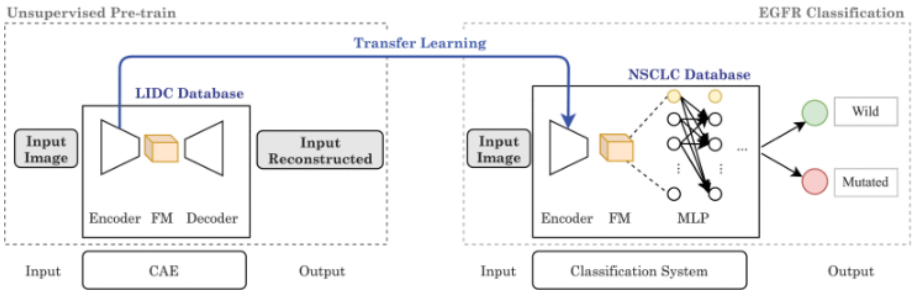

Figure 2.5: Overview of the proposed approach by *F.Silva et al*.

Local nodule analysis resulted in a low classification value of 0.51 of AUC. When considering an extended region of interest (lung containing the nodule), the value increased to 0.68, still not outperforming the State of the Art.

Finally, J. Lao et al. [19] also developed a semi-supervised learning framework, DS-FixMatch, which combines selective labeling and semi-supervised training. First, it uses an unsupervised algorithm that selects unlabeled images that best represent the distribution of the dataset, avoiding samples that are influenced by the intraoperative environment, which could affect the labeling. That subset of images is then sent to a human expert for labeling, which is then computed using supervised training. For the remaining unlabeled images, DS-FixMatch comprises two semi-supervised approaches: consistency regularization and pseudo-labeling.

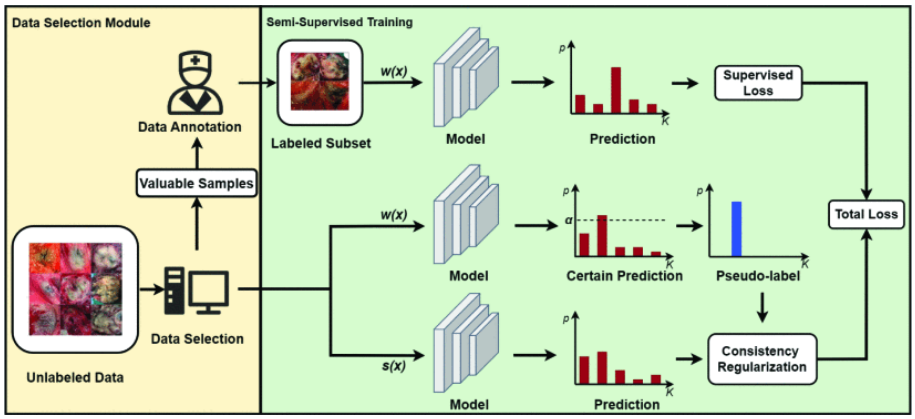

Figure 2.6: DS-FixMatch framework.

DS-FixMatch, with a labeling ratio of 20%, obtained an accuracy of 85.09%, an F1-score of 89.22% and a AUC value of 81.71%. It was also concluded that the use of semi-supervised training for that ratio of annotated data contributed for 2.05% and 3.62% improvement in terms of accuracy and AUC.

2.3 Conclusions

This chapter presented some papers containing solutions to the problem of few annotated data when developing models for cancer characterization. The use of semi-supervised learning could leverage that problem by making use of medical images without labels, creating other efficient methods to detect cancer, classification of pulmonary nodules, and segmentation of tumors.

The study of several investigations in the area and the analysis of the State of the Art were key to understanding the several techniques that are mostly utilized and what kind of results those procedures were able to achieve. Semi-supervised learning methods showed improved classification accuracy in several evaluation parameters when compared to fully supervised models. Also, the efficient use of unlabeled data was crucial to overcoming the problem of a lack of annotated data. Methods like pseudo-labeling, teacher-student pipelines, and expectation-maximization algorithms were particularly effective in creating valuable pseudo-labeled datasets. Expanding the dataset using SSL also exhibited stable performance with improved generalization. This approach demonstrated great utility in many areas of the medical spectrum, such as CT scans, chest X-rays, and even adversarial frameworks combining VAE and GAN models. Semi-supervised methods often include techniques such as transfer learning, adversarial training, and multi-view analysis, all of which greatly improve the effectiveness of unlabeled data in the pipeline.

The improvement seen by implementing semi-supervised learning methods validates the efficacy of this idea. The ability to take into account unlabeled data not only minimizes the cost and labor associated with manual annotations but also creates better AI-based diagnostic tools. While results seem to be promising, there are still refinements to make in semi-supervised techniques in order to be utilized in real-world scenarios.

Chapter 3

Background

3.1 Clinical Framework

3.1.1 The lung

The lung is an organ in the human body that is part of the respiratory system. It is responsible for the gas exchange with the environment, specifically the oxygen that enters the bloodstream and reaches all the cells in the body, and the carbon dioxide, a product of several chemical reactions that is ultimately removed. [47]

The lungs are located in the chest cavity and are cone-shaped, with the left lung being smaller than the right one. The lungs are also divided into lobes: the left has 2 lobes due to the presence of the heart, and the right lung has three different lobes.

The internal structure of the lungs is highly ramified, starting with the bronchi and ending in the alveoli, which are small air sacs filled with capillaries where the gas exchange happens. The high density of blood vessels and the nature of lung tissue make it more susceptible to nodular formations, which can be either benign or malignant, the latter being associated with lung cancer.

3.1.2 Lung cancer

Lung cancer is still one of the main causes of death and is the deadliest of all types of cancer. The main reason for that to happen is usually due to the late diagnosis of the disease. By that time, the cancer is often in an advanced stage of development and frequently accompanied by metastases.

Clinically, lung cancer is divided into two main histological groups:

Non-Small Cell Lung Cancer (NSCLC):

This is the most common lung cancer type. About 80% to 85% of lung cancers are NSCLC, which subdivides itself into three groups: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. These diseases appear grouped because their treatments and prognoses are often similar, despite originating from different types of lung cells. [1, 2]

- Adenocarcinoma: a disease that starts in the outer part of the lung cells, responsible for the synthesis of mucus. It is the most common subtype of NSCLC and is more likely to appear in the younger generation in comparison with other types of lung cancer.
- Squamous cell carcinoma: this type of lung cancer is often linked to smoking history. It starts in the squamous cells inside the airways, more specifically in the bronchus.
- Large cell carcinoma: may occur in any part of the lung and is characterized by its tendency to grow and spread rapidly, which can make the treatment harder.

Small Cell Lung Cancer (SCLC):

Accounts for 10% to 15% of all cases of lung cancer. It is normally associated with smoking and is known to grow quickly, even spreading beyond the lungs before the first diagnosis. Due to these characteristics, chemotherapy and radiation therapy are effective treatment options for this disease. However, there is a high risk of cancer recurrence. [1, 2]

Other types of lung tumors do not represent the majority of the cases. These are lung carcinoid tumors, which account for roughly 5% of lung cancer, or adenoid cystic carcinomas, lymphomas, and sarcomas, as well as benign tumors. However, these kinds of tumors are treated differently from the first two that were presented. [1]

Despite NSCLC being less aggressive than SCLC and having a slower growth rate, the time for its diagnosis usually happens in a late stage, which often indicates that the tumor has already spread to other parts of the body. This decreases the survival rate in an exponential way, which begs the need for treatments such as target therapies that act on specific molecular targets.

3.1.3 Computed Tomography

Computed tomography is a medical imaging technique widely utilized in the evaluation of thoracic diseases, which includes the screening, diagnosis, and monitoring of lung cancer. It has been one of the most used techniques due to its simplicity in the procedure and its high spatial resolution.

CT works by acquiring multiple X-ray projections around the patient's body, which are then processed by computational algorithms. The results consist of axial images, called slices, representing cross-sections of the human body. These slices can be further reconstructed to create 3D images of an organ of interest, like the lungs. [56]

The X-ray detectors collect the signals from the different tissues of the body. The attenuation of X-rays depends on the density and composition of the tissues they pass through, allowing the distinction of structures with unequal characteristics. The density values are measured in

Hounsfield units (HU), where, by convention, the air (at standard pressure and temperature) has a value of -1000 HU, water has a value of 0 HU, and bones, for example, have values superior to 1000 HU [8]. CT images are displayed in grayscale: less dense tissues appear darker, while denser tissues appear in lighter shades.

Although CT is the commonly used technique in lung evaluation and nodule detection, there are still other methods that are applicable. These can be the conventional X-ray, MRI (magnetic resonance imaging), PET (positron emission tomography), and Thoracic ultrasounds, each one utilized in specific clinical situations and depending on the type of tissue being analyzed. [17]

3.1.4 Biopsy

The identification of cancer malignancy plays a crucial role in guiding therapeutic decisions in oncology. Determining the tumor stage enables clinicians to select the most appropriate treatment protocols.

Today, tissue biopsy is considered the standard diagnostic procedure to obtain the biological matter necessary for molecular tests, especially when the tissue is easily available. [34] It works by collecting part of the tumor that will later be analyzed in the lab to determine a tumor's malignancy and relevant genetic mutations. However, despite its efficacy, regular biopsy has some limitations and some clinical risks. It is an invasive procedure, often painful for the patient, and carries the risk of postoperative complications, such as infections [52]. The situation is further complicated by the frequent need for multiple biopsies to obtain an accurate diagnosis, due to variations in tumor cell characteristics. When the tumor is located in regions that are difficult to access, a biopsy may not be the most suitable option. The same applies to patients with certain morbidities. All of this adds to the high laboratorial and logistical costs of this procedure for healthcare services.

Several less invasive alternatives have emerged to combat these problems, such as liquid biopsies, cytology, and computerized imaging combined with artificial intelligence. Liquid biopsy takes liquid samples, normally from saliva, blood, or urine, that can contain DNA from tumor cells, making this procedure safer, quicker, and easier. It can be used to identify early stages of cancer, as well as monitor treatment effectiveness, checking for its possible recurrence. [35] However, this operation is also costly and presents a low level of sensitivity, a consequence of the low concentration of tumor-derived components. Cytology samples involve examining cells from tumor tissues or fluids to determine a diagnosis. It distinguishes itself from a biopsy because of the quantity of cells needed to be analyzed [15]. However, this approach presents the same challenge as liquid biopsy: the low quantity of tumor material, which may be insufficient for a precise diagnosis. Computerized imaging associated with AI, also the theme of this thesis, is an emergent approach that aims to infer molecular information from image data, like CT scans. It takes advantage of AI algorithms to draw medical conclusions, reducing the need for invasive methods [45].

3.1.5 Summary

Understanding lung cancer and its potential progression is crucial for developing effective prevention and treatment strategies. Biopsy plays an important role in collecting tissue samples for further analysis of lung cancer malignancy and genetic mutations, helping with diagnosis and the right treatment. However, due to its invasive nature and procedural risks, alternative methods have been developed. AI and DL models, with their rapid growth in recent years, can leverage several medical images, like CT images from lung cancer patients, a standard modality for thoracic evaluation. By learning from this data, it has become more possible to extract important lung features and, therefore, make medically supported decisions, while preserving the patients' well-being.

3.2 Deep Learning Framework

Deep learning, a subset of machine learning, has made significant breakthroughs in areas such as natural language processing and image recognition. Its advantage comes from the fact that these models are composed of neural networks with multiple layers, which are capable of learning representations of data with many levels of abstraction. [20]

In the context of medical imaging, DL models, particularly convolutional neural networks (CNNs), have shown great potential in tasks such as detection, segmentation, and classification of structures and pathological findings. However, the amount of annotated data greatly influences the performance of these models. Depending on data availability, three learning paradigms can be distinguished: supervised learning, unsupervised learning, and semi-supervised learning.

3.2.1 Supervised Learning

Supervised learning is a paradigm where the model is trained on a dataset of input-output pairs, where inputs are typically images and outputs are their corresponding labels. Often called classification or inductive learning in machine learning, the goal is for the algorithm to minimize a certain loss function, so that it can accurately predict the output on unseen data. There are two main types of supervised learning: classification, where data is categorized into predefined classes; and regression, where the objective is to predict a numerical value that is continuous in time. The effectiveness of supervised learning depends heavily on the quantity of the labeled data, the class distribution balance, and the representativeness of the training samples. [22]

3.2.2 Unsupervised Learning

Unsupervised learning refers to the class of methods that operate solely on unlabeled data. Unlike supervised learning, it doesn't rely on predefined categories or classes during the training phase. This means that the objective of unsupervised learning is to discover hidden patterns and relationships within the data, beyond what would be considered pure unstructured noise [13]. Two simple and classic examples of unsupervised learning are clustering and dimensionality reduction, which

group similar samples together and project high-dimensional data into lower-dimensional spaces, respectively.

3.2.3 Semi-supervised Learning

In many real-world applications, especially in the medical domain, the cost of obtaining labeled data is prohibitively high. On the other hand, vast amounts of unlabeled data are often available. This asymmetry has driven the development of semi-supervised learning (SSL), a paradigm that takes into account a small set of labeled and a large set of unlabeled data to improve model generalization while reducing reliance on manual annotation. [43] Semi-supervised learning methods have also been applied in scenarios where there is no significant lack of labeled data: if the unlabeled data provides sufficient information relevant for prediction, it can be utilized for a better classification performance. [49]

SSL methods generally fall into three major categories: [49]

- **Pseudo-labeling**: Trains a model iteratively, promoting confident predictions on unlabeled data, which is added to the training set.
- **Consistency-based methods**: Encourage the model to produce consistent outputs under perturbations of the same input, enforcing stable predictions.
- **Graph-based or manifold methods**: Exploit data geometry and neighborhood relationships to propagate label information.

These techniques are used in different combinations in the creation of DL models, built to be capable of surpassing the predictive ability of normal supervised and unsupervised learning methods.

Chapter 4

Materials

This chapter provides a comprehensive overview of all the resources and preparatory steps required to support the development of this project. It includes a description of the datasets used, the preprocessing steps and techniques applied to the imaging data, and the external computational infrastructure and software framework that enabled the experimentation of the proposed pipeline. All these materials were crucial in the foundation and development of this work.

4.1 Datasets used

LIDC - IDRI

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset is one of the most comprehensive publicly available collections of lung CT images with marked-up annotated lesions. It was initiated by the National Cancer Institute (NCI), advanced by the Foundation for the National Institutes of Health (FNIH), and accompanied by the Food and Drug Administration (FDA). This dataset contains 1018 different cases, gathered by seven academic centers and eight imaging companies from the United States. The annotation process consisted of two phases and was performed by four experienced thoracic radiologists. The initial blinded-read phase involved each radiologist reviewing each CT scan and classifying the lesions into three different categories: "nodule >= 3mm", "nodule < 3mm", and "non-nodule >= 3mm". In the second, unblinded-read phase, each radiologist reviewed their analysis as well as the other

professionals' in an anonymized manner, to gather a consensual opinion [4, 5]. This approach was meant to maximize the identification of lung nodules in different CT scans without clashes in reviews.

Luna25

Luna25 is a recently released public dataset, and it offers approximately 4000 CT scans and 6000 nodules, combining the malignant nodules with the benign ones, the latter being on a much larger scale. [30] It was derived from the National Lung Screening Trial (NLST), a randomized controlled trial conducted by the Lung Screening Study group (LSS) and the American College of Radiology Imaging Network (ACRIN) in the United States. The objective was to assess the differences in lung cancer mortality rates by comparing screening with low-dose helical computed tomography with screening using chest radiography. Approximately 54000 participants at heavy risk of lung cancer due to a history of heavy smoking enrolled between 2002 and 2004, making it one of the largest chest CT datasets publicly available [27].

4.2 Data Preparation and Preprocessing

LIDC - IDRI

The raw data from the LIDC-IDRI dataset required some form of preparation and preprocessing to ensure the CT images were suitable for the training of DL models. In this case, this preparation/preprocessing pipeline was based on the procedures that were described in the Efficient-ProtoCaps framework [9], then adapted to the semi-supervised architecture employed in this thesis. It comprises two main phases, those being related to the labels and the CT images.

The steps taken in the preparation phase included:

- **Label preparation:** For both LNVAs (Lung Nodule Visual Attributes) and LNM (Lung Nodule Malignancy), the latter with a score ranging from 1 to 5 (Highly Unlikely, Moderately Unlikely, Indeterminate, Moderately Suspicious, and Highly Suspicious, respectively), the mean scores attributed by up to four radiologists were calculated;
- **Data Filtering:** Nodules with a malignancy mean score equal to three, meaning they are indeterminate, were excluded. Nodules that were annotated by fewer than three radiologists were also excluded.

Regarding the preprocessing algorithm, the library [1]*pylidc* was used to read the 3D array volumes, the annotations, and the lung nodule bounding box coordinates. It included the following image transformations:

- **Hounsfield Units (HU) clipping:** Values corresponding to less than -1000, which represent the atmospheric air at standard pressure and temperature, were limited to values of

[1] Library website: https://pylidc.github.io/

-1000. Values corresponding to more than 400, corresponding to dense tissue, were limited to values of 400.

- **Min-Max normalization:** HU values were rescaled from [-1000, 400] to [0, 1] using a linear transformation.
- **Resolution adjustment:** Slice thickness and pixel spacing were adjusted from 1.91 $\pm$ 0.73mm and 0.68 $\pm$ 0.008 mm, respectively, to 1.0mm.
- **Bounding Box extraction:** It was extracted two types of lung nodule bounding boxes. The first was a 2D variant, corresponding to a standard anatomical plane, centered on the nodule. The second variant was a 2.5D representation, comprising of three orthogonal slices also centered on the nodule.

Luna25

Regarding the Luna25 dataset, various steps of preprocessing were also made to meet the changes employed in the LIDC dataset and to effectively train the utilized deep learning models. For these reasons, the used methods were the same as described in the subsection before, which follow the procedures described in the Efficient-Proto-Caps framework [9].

4.3 Data labelling

Hence the complexity of the multi-label classification task, we reformulated it as a binary classification problem. Each lung nodule was assigned a label of 0 or 1, depending a benign or malignant case, respectively. This labelling was based on the Lung Nodule Malignancy (LNM) mean score, where nodules with a score below 3 were considered benign, and those above 3 were considered malignant. This binarized dataset was utilized as ground truth throughout the experiments.

4.4 Slurm Job Management

To efficiently manage computational resources and execute large-scale experiments, this work relied on SLURM (Simple Linux Utility for Resource Management). It is an open-source job scheduler for high-performance computing environments.

It was designed to allocate resources to CPUs and GPUs across a cluster of machines, and to manage job queuing, execution, monitoring, and logging. It is often used in business environments because of its scalability to deal with clusters of various sizes, flexibility of adaptation, and efficiency in the execution of resources.

For a project involving training deep learning models with different configurations, SLURM is crucial to speed up the process, not needing to use resources from one's personal computer for long periods of time. Also, it helps reduce the debugging time, since it saves log files captured while running.

4.5 Utilized Libraries

During the development of this project and training of deep learning models, many libraries were utilized; however, two of them had a higher scale of significance: CUDA, PyTorch, and PyTorch Lightning. These libraries were essential in the model implementation, GPU acceleration, and for scalable experiment management.

4.5.1 CUDA

CUDA (Compute Unified Device Architecture) is a parallel computing platform developed by NVIDIA that allows developers to make use of the computational power of GPUs. It is a collection of libraries designed to accelerate various computationally intensive tasks commonly found in convolutional neural networks. In this project, CUDA contributed to reducing the training time of deep learning models.

4.5.2 Pytorch

PyTorch is an open-source framework used to build deep learning models, commonly utilized in applications for image recognition and language processing. In this project, PyTorch was the core for model development, supporting neural network architectures, custom loss functions, and tensor manipulation, facilitating rapid prototyping and debugging.

4.5.3 Pytorch Lightning

PyTorch Lightning is a lightweight, open-source framework that aims to simplify the development and training of deep learning models using PyTorch. It works as a wrapper of PyTorch, abstracting from the complexity of deep learning pipelines and focusing on their logic, making the code cleaner and more readable.

4.6 Datasets and Environment Details

The materials and resources presented in this chapter were fundamental for the development of this project. In Table 4.1 it is presented the final images considered for the datasets utilized. In Table 4.2, the version of the project libraries is shown.

Table 4.1: Summary of the datasets used in this project.

Dataset	Number of Images	With Labels Associated
LIDC-IDRI	850	Yes
Luna25	4437	No

Table 4.2: Version of used libraries

Library	Version
CUDA	12.1
PyTorch	2.1.1
PyTorch Lightning	2.1.2

Chapter 5

Architecture and Methods

This chapter aims to describe the methodological framework adopted for the creation of a semi-supervised learning pipeline designed to characterize lung CT scans in the malignancy aspect. It will include topics such as the explanation of the implemented method and the experimental setup, the latter containing the choice of hyperparameters, the training protocol, and the strategy for dataset partitioning and evaluation. The architecture and methodology presented in this chapter will serve as the base for the results and analysis later discussed in this dissertation.

5.1 Method Overview

5.1.1 FixMatch

FixMatch is a semi-supervised algorithm that combines two main strategies: pseudo-labeling and consistency regularization. Its novelty comes from the fact that it uses these two ingredients combined with the separate use of weak and strong augmentation when performing consistency regularization.

The algorithm works in the following way: [40]

Let $X = \{(x_b, p_b) : b \in (1, ..., B)\}$ be a batch of B labeled images, where x_b and p_b are training examples and their one-hot labels, respectively. Let $U = \{u_b : b \in (1, ..., \mu B)\}$ be a batch of μB unalabeled images where μ determines the realtive sizes of X and U. Let also $p_m(y \mid x)$ be the predicted class distribution produced by the model for input x. The cross entropy between two

probability distributions p and q is represented by $H(p,q)$, and the two types of augmentations in this algorithm, strong and weak, are denoted by $\mathscr{A}(\cdot)$ and $\alpha(\cdot)$, respectively.

For each training batch, labeled images and unlabeled images are extracted. The labeled images are used to train a given model, in an identical way to that of a classic supervised learning method, where the cross-entropy loss is calculated on weakly augmented labeled images:

$$\mathscr{L}_{\mathrm{s}} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{H}(p_b, p_m(y \mid \alpha(x_b))) \tag{5.1}$$

For the unlabeled batch of images, a weak augmentation is applied to each one, and a prediction is obtained. When the model assigns a probability to any class that is above a defined threshold, that prediction is converted to a one-hot pseudo-label, and a strong augmentation is applied to that same image. A model prediction is computed on the strongly augmented image to further train the model by matching it with the pseudo-label via a cross-entropy loss:

$$\mathscr{L}_{\mathrm{u}} = \frac{1}{\mu B}\sum_{b=1}^{\mu B} 1\left(\max(q_b) \geq \tau\right) \mathrm{H}\left(\hat{q}_b, p_m\left(y \mid \mathscr{A}(u_b)\right)\right) \tag{5.2}$$

where $q_b = p_m(y \mid \alpha(u_b))$, $\hat{q}_b = arg\ max(q_b)$ represents a pseudo-label, and τ is a scalar hyperparameter defining the threshold above which we should include the pseudo-label.

This way, we can conclude that the loss minimized by FixMatch is simply:

$$\mathscr{L}_s + \lambda_u \mathscr{L}_u \tag{5.3}$$

, where λ_u is a fixed scalar hyperparameter that represents the relative weight of the unlabeled loss.

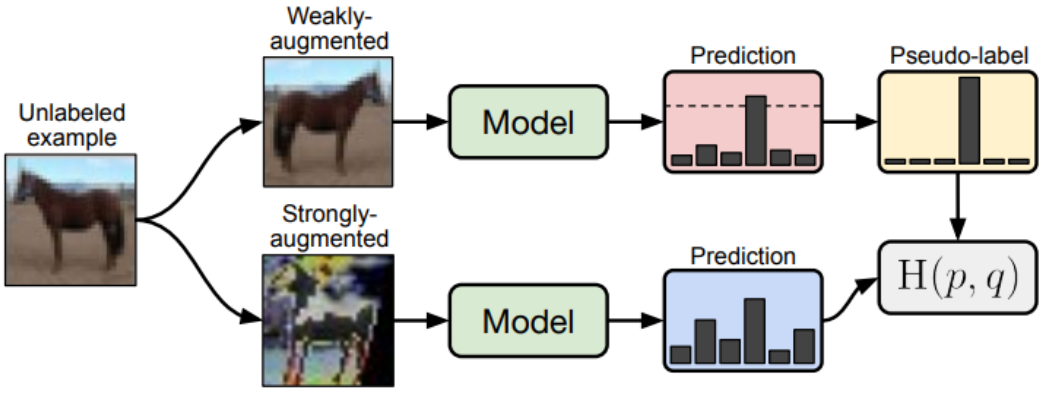

Figure 5.1: Diagram of FixMatch from [40]. It begins by weakly augmenting an unlabeled image that is fed to the model, which makes a prediction. If that prediction is above a predetermined threshold, a one-hot pseudo label is created. Then, the same image goes through a strong augmentation and is fed again to the model to make another prediction. The model is trained by matching this last prediction with the pseudo-label assigned before.

5.1.2 Image Augmentations

In the FixMatch framework implemented in this dissertation, data augmentation plays a critical role in enforcing consistency regularization between weakly and strongly augmented views of the same image. The augmentation pipeline was designed to generate diverse perturbations while preserving the diagnostic features relevant to lung nodule classification. Two distinct augmentation strategies were utilized: weak augmentation and strong augmentation.

Weak Augmentation

This augmentation was applied to produce minimally altered versions of the input data, primarily to serve as a stable reference for pseudo-labelling. The transformations included:

- Random horizontal flip and random vertical flip, each applied with a defined probability;
- Random translation, allowing small positional shifts within a fraction of the image size, with pixel filling to maintain dimensions.

Strong Augmentation

Strong Augmentation was applied to create heavily perturbed versions of the same input, encouraging the model to maintain prediction consistency under substantial image variations. The transformations included:

- Elastic deformation to simulate non-rigid anatomical variations;
- Random translation to introduce spatial shifts;
- Shearing transformations along different axes;
- Random rotation within a specified degree range;
- Gaussian blur to mimic image acquisition noise;
- Color jittering for brightness and contrast adjustments;
- Random erasing (cutout) applied to occlude random regions.

5.2 Convolutional Neural Networks

The creation and development of Convolutional Neural Networks, or CNNs, was established in 1980 by Kunihiko Fukushima and his "Neocognitron" work. [12] It took inspiration from neuroscience in the early 1960s, when David Hubel and Torsten Wiesel worked on the visual cortex of cats and concluded that individual cells responded to specific patterns of light in different regions of the visual field. [7]

CNNs are designed primarily for image-based inputs, with architectures optimized to process and extract features from this specific type of data. Overall, they are comprised of three different types of hidden layers. These include convolutional, pooling, and fully-connected layers. [29, 21]

The convolutional layers are a key component of CNNs. They are responsible for automatically extracting local characteristics of an image, such as borders, textures, or structural patterns. It works by applying a group of filters with adjustable values during training, called kernels. These kernels are small matrices that slide along the depth of the input and compute the scalar product between each value in the kernel and image patches. The result of the operation is 2D activation maps, or feature maps, which are utilized to capture specific features or patterns in the input. Convolutional layers can also significantly reduce model complexity by optimizing the input representation through three key hyperparameters: depth, stride, and zero-padding. The depth of the output denotes the number of filters (or kernels) employed in a convolutional layer. Increasing the number of filters enables the network to identify and recognize a greater variety of patterns; however, this also increases the computational cost and the number of parameters. Conversely, employing too few filters may limit the model's capacity to capture and represent complex patterns in the data. The stride specifies the number of pixels by which the filter shifts at each step when traversing the input image. A stride of one indicates that the kernel advances by a single pixel at a time, producing substantial overlap and yielding a relatively large output. Using a larger stride means that certain pixels will be skipped when sliding the kernel, reducing the overlap and the output complexity. Zero padding works by adding extra pixels, each with a value of 0, around the borders of the input image. This helps preserve the output size, especially when using a stride of 1, and the information at the edges of the input, which might otherwise be lost.

Pooling layers have the objective of progressively reducing the spatial dimensionality of activation maps generated by the convolutional layers, hence following them in the network architecture. This reduction has two major advantages, namely the decrease in the number of parameters, which helps to avoid overfitting, and the lowering of the model's computational complexity. The pooling operation is applied separately to each activation map, being the usual version called max pooling, where it retains the maximum value inside a small region of the image. Another type of pooling, known as average pooling, instead computes the average value inside that same region.

The fully connected layers follow the convolutional and pooling layers and precede the output layer. They are characterized by containing fully connected neurons from the two adjacent layers, which take the pre-learned features and learn complex combinations, using them to generate final predictions in classification tasks.

5.3 Model Architectures

5.3.1 ResNet

ResNet (Residual Network) is a convolutional neural network architecture designed by He et al. [14] to address the limitations of training very deep models, which, contrary to expectations, didn't always perform better due to the difficulty of learning identity mappings through multiple non-linear transformations.

The solution was the introduction concept of residual learning. It works by using skip connections, which allow the input of a layer to bypass one or more layers and be added directly to their output. This structure allows the network to learn the residual mapping, with the function $H(x) = F(x) + x$, representing the difference between the input and the desired output, instead of learning the desired output mapping directly, $H(x)$. This results in an output of the form $H(x) = F(x) + x$, known as a residual block, enabling the network to preserve identity mappings when needed and simplifying optimization. This approach helps mitigate the vanishing gradient problem, where gradients during backpropagation in deeper networks can become extremely small, making learning in earlier layers a challenge.

The baseline models, such as ResNet-18 or ResNet-34, use simple residual blocks, while deeper ResNet-50, ResNet-101, or ResNet-152 adopt bottleneck blocks to maintain performance with fewer parameters. In this dissertation, ResNet-18 is used due to its balance between computational efficiency and expressive power.

5.3.2 EfficientNet

EfficientNet is a family of architectures of CNNs developed by Tan and Le [44] for improving efficiency and performance of computer vision tasks. Its novelty came from the fact that it utilized compound scaling, a parameter used to evenly scale the width, depth, and resolution of the network using a compound coefficient ϕ, contrasting standard practice, which adjusts these factors separately. The architecture was able to optimize the FLOPS (floating-point operations per second) level in terms of effectiveness and precision [36].

EfficientNet B0 represents the baseline model, obtained through a small grid search to determine the optimal values for the three dimensions, and it is the one utilized in this dissertation. There are, however, other variants, from EfficientNet-B1 to EfficientNet-B7, which are scaled versions of B0 using larger values of ϕ, offering higher accuracy at increased computational cost.

5.3.3 ConvNext

ConvNeXt is a relatively new CNN architecture developed by Liu et al. [23] to modernize CNNs and put them on the same standard as Vision Transformers (ViTs), which have been dominating benchmarks of computer vision in the last years. Instead of abandoning convolutions, ConvNeXt tries to make Transformer-inspired adjustments to achieve state-of-the-art performance.

Unlike classic CNNs like ResNet that begin with a 7 x 7 stem convolution followed by max pooling, ConvNeXt starts with a 4 x 4 convolution with a stride of 4, so that patches don't overlap. Besides this, the compute ratio of the ResNet Blocks used per stage was also changed. These mechanism changes mimic the mechanism utilized in Vision Transformers.

On the micro level, some changes were also made to the ResNet blocks. The traditional 3 × 3 convolutions were replaced with depthwise convolutions, and the bottleneck structure was inverted, placing the widest layer in the center. The kernel size was expanded to 7 × 7 (depthwise), and the depthwise convolution was moved to the block's beginning. Activation functions were

modernized by replacing ReLU with GELU (Gaussian Error Linear Unit), and their number was reduced to just one per block. Similarly, Batch Normalization was replaced with Layer Normalization, and redundant normalization layers were removed. Finally, dedicated downsampling layers were introduced between stages, each including layer norm and a 2×2 convolution with stride 2.

These incremental changes held significant accuracy gains with minimal increases in computational cost, which enabled ConvNeXt to match or surpass transformer-based models while preserving the inductive biases of CNNs.

5.4 Evaluation

5.4.1 Model selection

During the training process, three distinct criteria based on validation set metrics were defined for model checkpoint selection: the checkpoint corresponding to the highest AUROC, the checkpoint achieving the lowest loss, and the checkpoint associated with the final training epoch. For the purposes of reporting results, we present exclusively those obtained from the checkpoint corresponding to the highest AUROC on the validation set, as this criterion consistently yielded superior preliminary performance relative to the alternatives.

5.4.2 Model evaluation

The performance evaluation of the employed models in this dissertation was carried out based on deeply utilized metrics in binary classification: *AUROC (Area Under the Receiver Operating Characteristic Curve)*, *accuracy*, *precision*, *sensitivity*, and *specificity*.

The *AUROC* metric measures the global capacity of a model in distinguishing between positive classes (malignant nodules) and negative (benign nodules), independently of the decision threshold value.

Accuracy represents the proportion of correct predictions in relation to the total number of samples. Despite being a global metric, in the clinical context, this value can sometimes be misleading in scenarios where the data has imbalanced classes. However, it brings a general vision of the model's performance on all types of cases.

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{5.4}$$

Precision measures the proportion of cases identified as malignant among those that are malignant in reality. A high precision value translates to a low value of false positive cases. This could prevent cancer-free patients from being subjected to unnecessary treatments that could potentially be invasive.

$$\frac{TP}{TP+FP} \tag{5.5}$$

The *sensitivity* metric measures the capacity of the model in identifying positive cases correctly. In this context, *sensitivity* is critical, because a high value means that fewer cases of cancer are ignored (false negatives), increasing the chances of early detection and effective treatment.

$$\frac{TP}{TP+FN} \tag{5.6}$$

specificity is related to the negative cases predicted correctly. Like the *precision* value, a high *specificity* ensures avoiding false alarms.

$$\frac{TN}{TN+FP} \tag{5.7}$$

To prevent the model from being biased only for a particular train-test split, the evaluation followed a 5-fold cross-validation scheme, in which, for each iteration, 80% of the data were utilized for training and validation (72% for training and 8% for validation), and 20% remaining data was reserved exclusively for testing purposes. This method also prevents overfitting and helps the model get a more reliable generalization capacity for unseen data.

5.5 Experiments

5.5.1 Experiment setup

Training employed the Adam optimizer with default values for β_1 of 0.9 and β_2 of 0.999, and all models were initialized with ImageNet pre-trained weights. Class weights were also implemented in the training set distribution to address class imbalance. Computations were executed on CUDA-enabled GPUs, and to provide a flexible environment with large-scale experimentation, this framework was implemented using PyTorch and PyTorch Lightning.

Table 5.1: Values used for hyperparameter optimisation

Fixed hyperparameters	**Values**
Optimizer	Adam
Weight Initialization	ImageNet pre-trained weights
Framework	PyTorch & PyTorch Lightning
Hardware	CUDA-enabled GPU

5.5.2 Hyperparameter Optimization

To assess different types of studies in this work and to have them on equal standards, a configuration of parameters was pre-determined. That was done by adopting a systematic grid search strategy to explore the hyperparameter space and identify optimal training options. The search combined five different learning rates of 0.005, 0.002, 0.001, 0.0001, and 0.00001 with dataset-specific batch sizes of 32 and 64 for LIDC-IDRI, and 128 and 256 for Luna25. The number of epochs was also varied: 100, 150, 200, 250, and 300, applying an early stopping criterion when the

validation metric failed to improve for 5 consecutive epochs. The model utilized was EfficientNet B0, alongside the loss function binary cross-entropy.

Table 5.2: Values used for hyperparameter optimisation

Hyperparameter	Values
Learning Rates	0.00001, 0.0001, 0.001, 0.002, 0.005
LIDC-IDRI batch sizes	32, 64
Luna25 batch sizes	128, 256
Training epochs	100, 150, 200, 250, 300

5.5.3 Baseline

Following the hyperparameter configuration phase, a baseline model was defined to serve as the primary reference for performance comparisons in subsequent experiments. This baseline made use of the EfficientNet-B0 model, along with the Adam optimizer and binary cross-entropy loss function. The selected hyperparameters, which got the best preliminary results, included a learning rate of 0.001, a batch size of 32 for the LIDC-IDRI dataset and 128 for the LUNA25 dataset, and a total of 300 training epochs. Given the adoption of a FixMatch framework, a pseudo-label confidence threshold of 0.95 was applied. For the baseline configuration, no unlabelled data was incorporated, as the primary objective was to establish a controlled benchmark for assessing the impact of adding unlabelled samples in later experiments. Regarding data dimensionality, a bounding box of 32×32 pixels was used in a 2D representation. These experiments were repeated with three different random seeds for reproducibility and robustness efforts.

Table 5.3: Baseline parameters

Hyperparameter	Values
Learning Rate	0.001
LIDC-IDRI batch sizes	32
Luna25 batch sizes	128
Training epochs	300
Pseudo-label treshold	0.95
Bounding Box	32 x 32
Dimensionality	2D
% of unlabelled data	0%
Model	EfficientNet B0
Robustness	three different seeds
Loss Function	Binary cross-entropy

5.5.4 Ablation Studies

To investigate the contribution of individual components to the overall model performance, a series of ablation studies was conducted. These experiments followed a controlled approach, in which the characteristics of the baseline configuration were maintained, modifying only a single parameter at a time.

The ablation experiments explored the following modifications:

Proportion of unlabelled data

The percentage of unlabelled data incorporated into the semi-supervised training was varied from $0N$ to $5N$, where N corresponds to the number of labelled samples. This experiment represents the primary research focus of this dissertation, which aimed to quantify the impact of unlabelled data on model performance.

Input dimensionality and resolution

The bounding box size and dimensionality were altered to assess the effect of spatial context:

- 2D representation with 64 x 64 bounding box;
- 2.5D representation with 32 x 32 bounding box;
- 2.5 D representation with 64 x 64 bounding box.

Loss Function

The baseline BCE loss was replaced with alternative formulations:

- Focal loss with γ values of 1,2, and 3. These values address class imbalance by focusing more on hard-to-classify samples.
- BCE + Dice loss with η values of 0.25, 0.5, 1, 2, and 5. These values represent the proportion of the Dice value compared to the BCE value and aim to jointly optimise classification accuracy and region overlap.

Model Architecture

Two alternative architectures were evaluated to compare representational capacity and performance: ResNet18 and ConvNeXt.

Pesudo-Label confidence threshold

The confidence threshold for FixMatch pseudo-labelling was adjusted to values of 0.99, 0.90, 0.85, and 0.80 to study its effect on the quality of generated labels.

5.5.5 Explainability

To provide interpretability of the model predictions, we utilized Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM is a visualization technique used in CNNs to explain the regions of an input image that most strongly influence the network's decision. It produces a localization heatmap that can be overlaid on the original image, thereby giving insights into the spatial areas most relevant for classification.

In this work, Grad-CAM was applied to the ResNet-18 architecture in layer 3, before the final stage (layer 4), with two experimental configurations: one considering fully supervised training, where only the labeled data was used, and one considering semi-supervised training, where the FixMatch framework utilized an additional set of unlabeled images. In the latter, the size of the unlabeled dataset was five times the number of labeled samples ($5N$, with N corresponding to the total number of labeled samples.

Using Grad-CAM in these two scenarios allowed us to compare the interpretability of the supervised and semi-supervised approaches. In particular, it helped us assess whether the inclusion of unlabeled data influenced the spatial focus of the model, potentially highlighting regions of interest for classification.

Chapter 6

Results and Discussion

This chapter presents and discusses the experimental results obtained throughout this dissertation. The primary objective of this analysis is to evaluate the performance of the proposed architecture under different training conditions, to identify the factors that had the most influence on the detection of malignant pulmonary nodules.

The presentation of results follows a progressive structure. Starting with the baseline results, which served as the main reference point for all subsequent comparisons, and coming next with the outcomes of the ablation studies are presented, showing the differences in individual parameters and their specific impact. Particular emphasis is placed on the role of incorporating different proportions of unlabeled data into the semi-supervised framework, which forms the core of this study and provides insights into how such data enhances the model's generalization ability. Finally, the results of applying Grad-CAM are shown to get a visual analysis of how the model computes different parts of the image.

6.1 Baseline Performance

The baseline experiment was designed to establish a consistent reference point for all subsequent analyses. Across five-fold cross-validation, the baseline metrics achieved an AUC of 81%, an accuracy of 83%, and a precision of 87%. The model showed a sensitivity of 71%, indicating its ability to correctly identify malignant nodules, while maintaining a specificity of 91%, reflecting reliable detection of benign cases.

Table 6.1: Baseline model performance on the test set across seeds. The highest value in each column is highlighted in bold

Configuration	AUC	Accuracy	Precision	Sensitivity	Specificity
Baseline seed	0.81 ± 0.02	**0.83 ± 0.02**	**0.87 ± 0.06**	0.71 ± 0.05	**0.91 ± 0.04**
Seed 1	**0.82 ± 0.02**	0.82 ± 0.02	0.82 ± 0.06	**0.78 ± 0.07**	0.86 ± 0.06
Seed 2	**0.82 ± 0.01**	0.82 ± 0.01	0.82 ± 0.03	0.77 ± 0.05	0.86 ± 0.04
Seed 3	0.81 ± 0.03	0.82 ± 0.03	0.80 ± 0.03	0.78 ± 0.04	0.85 ± 0.03

When comparing the baseline to the other experimental seeds, we could observe that the AUROC and accuracy values didn't fluctuate a lot. On the other hand, precision and specificity saw an increase of about 5% to 7%, while sensitivity values dropped as low as 7 pp.

These results demonstrated that the values on the baseline were not biased by the utilized seed, showing comparable outcomes in different paradigms and thus reinforcing the idea of robustness of the FixMatch framework.

6.2 Model Comparisons

To address the influence of a model architecture on performance, two models were tested: ResNet18 and ConvNeXt Tiny. They were compared to the baseline configuration, which utilizes EfficientNet B0.

Across the experiments, the baseline with EfficientNet-B0 consistently achieved the strongest and most stable performance. It obtained an AUROC value of around 81%, with an accuracy (83%) and high precision (87%). It maintained a relatively low sensitivity (71%) while preserving specificity close to 90%.

ResNet18 produced slightly weaker results overall. AUC values ranged from 80%, with accuracy and precision also trailing those of EfficientNet-B0. While sensitivity was competitive and higher by 1%, the model displayed less consistency across the metrics. This indicates that ResNet18, despite the advantage of its residual blocks, lacked the same robustness and generalization ability observed in EfficientNet-B0 with its compound scaling feature.

ConvNeXt Tiny, despite being a more recent architecture and with a higher computational cost, did not achieve superior results in this context. Its AUC values ranged between 78%, with sensitivity lower than both EfficientNet-B0 and ResNet18. This reduced sensitivity suggests limitations in detecting malignant nodules, which is a critical weakness in the medical setting. The findings indicate that architectural advances designed for natural image benchmarks do not necessarily translate into gains for specialized medical imaging tasks.

Table 6.2: Performance results on the test set across models. The highest value in each column is highlighted in bold

Configuration	AUC	Accuracy	Precision	Sensitivity	Specificity
EfficientNet-B0 (Baseline)	**0.81 ± 0.02**	**0.83 ± 0.02**	**0.87 ± 0.06**	0.71 ± 0.05	**0.91 ± 0.04**
ResNet18	0.80 ± 0.01	0.81 ± 0.01	0.83 ± 0.02	**0.72 ± 0.02**	0.88 ± 0.02
ConvNeXt Tiny	0.78 ± 0.03	0.79 ± 0.03	0.82 ± 0.05	0.68 ± 0.03	0.88 ± 0.04

The results show that lighter-weight architectures are more prone to successively predict the lung nodules' malignant status on the utilized datasets, with better results across almost all metrics. A possible reason for the lower results observed in the ConvNeXt model may be due to the small Bounding Box (BB) size, which was probably not sufficient to extract important features for classification purposes.

6.3 Dimensionality and Resolution Impact

We evaluated the effect of input dimensionality and spatial resolution by varying the bounding box and slice layout while keeping the remaining settings fixed. The baseline used a 2D 32×32 crop. We then tested 2D 64×64, 2.5D 32×32, and 2.5D 64×64.

For the 64 x 64 BB and 2D dimension setting, the results were mainly worse compared to the baseline, with AUROC being about 80% and with precision losing 7 pp. Notably, the sensitivity metric had a 3% increase. In short, the increase in resolution did not effectively change performance, with a slight tendency to improve sensitivity by 3 pp (percentual point), at the cost of negative shifts elsewhere.

For both bounding boxes in the 2.5 dimensionality, we could observe that stacking neighboring slices did not translate into metric gains. The best AUC values did not go over the 81% mark, and the other metrics also didn't surpass the baseline results, except for the small increase in specificity in the 32 x 32 BB and in sensitivity in the 64 x 64 BB.

Table 6.3: Performance results on the test set across dimensionalities and checkpoints. The highest value in each column is highlighted in bold.

Configuration	AUC	Accuracy	Precision	Sensitivity	Specificity
32 x 32 2D (baseline)	**0.81 ± 0.02**	**0.83 ± 0.02**	**0.87 ± 0.06**	0.71 ± 0.05	0.91 ± 0.04
64 x 64 2D	0.80 ± 0.03	0.81 ±0.03	0.80 ± 0.04	**0.74 ± 0.02**	0.85 ± 0.03
32 x 32 2.5D	**0.81 ± 0.02**	0.82 ± 0.02	**0.87 ± 0.04**	0.70 ± 0.03	**0.92 ± 0.03**
64 x 64 2.5D	0.80 ± 0.02	0.81 ± 0.02	0.83 ± 0.03	0.72 ± 0.06	0.88 ± 0.03

The lower results of the 64 x 64 BB can be due to the increase in noise and area of benign tissue, which can dilute discriminative signals from small nodules. As for the 2.5D representations, the added input complexity could be a determining factor for the model under-utilizing the extra content and not giving better results overall.

6.4 Different Loss Functions

We next investigated the influence of different loss functions on model performance. The Baseline used a weighted BCE. Two alternative formulations were tested: BCE + Dice Loss, where the weighting parameter η was set to 0.25, 0.5, 1, 2, and 5; and Focal Loss, with the focusing parameter γ varied between 1,2, and 3.

The combination of BCE and Dice demonstrated stable and often improved results compared to the baseline. For the extreme values of $\eta = 0.25$ and $\eta = 5$, the AUROC was higher by 2 and 1 pp, and the sensitivity had also increased by 4 and 3 pp, respectively. The accuracy when $\eta = 0.25$ also increased by 1%. The other metrics didn't appear to have meaningful changes. However, we could observe a starting decline as we reached Dice values similar to those of the BCE. This parabola-type curve had its lowest results when the η weight was the same as the BCE, specifically when $\eta = 1$. Here, we could observe a decrease in all metrics compared to the baseline, with the most significant drop being a value of 3% in the accuracy and precision metrics.

The results suggest that the combination of BCE and Dice Loss was most beneficial when the Dice term was either given a relatively low or high weight. A low η preserved the strong pixel-wise discrimination of BCE while still adding a regional overlap constraint, characteristic of the Dice loss. In contrast, a high η shifted the optimization toward global shape consistency, which can help in stabilizing predictions for small or ambiguous nodules. Intermediate values may have diluted the strengths of both terms, leading to less consistent improvements.

Table 6.4: Performance results on the test set for models trained with BCE + Dice loss across loss weighting parameters. The highest value in each column is highlighted in bold.

Configuration	AUC	Accuracy	Precision	Sensitivity	Specificity
$\eta = 0$ (baseline)	0.81 ± 0.02	0.83 ± 0.02	**0.87 ± 0.06**	0.71 ± 0.05	**0.91 ± 0.04**
$\eta = 0.25$	**0.83 ± 0.02**	**0.84 ± 0.02**	0.86 ± 0.04	**0.75 ± 0.02**	0.90 ± 0.03
$\eta = 0.5$	0.81 ± 0.02	0.82 ± 0.02	0.86 ± 0.04	0.73 ± 0.06	0.90 ± 0.03
$\eta = 1$	0.79 ± 0.03	0.80 ± 0.03	0.83 ± 0.04	0.69 ± 0.05	0.89 ± 0.03
$\eta = 2$	0.81 ± 0.01	0.82 ± 0.01	0.86 ± 0.04	0.73 ± 0.03	0.90 ± 0.04
$\eta = 5$	0.82 ± 0.01	0.83 ± 0.01	0.85 ± 0.03	0.74 ± 0.04	0.90 ± 0.03

The Focal Loss results showed that it did not provide systematic improvements over the baseline. With $\gamma = 1$, performance was slightly worse than BCE + Dice with $\eta = 1$, which was also the worst of its category, as shown before, only improving in the specificity metric compared to the baseline. When utilizing higher parameters of $\gamma = 2$ and $\gamma = 3$, performance degraded substantially, in some cases producing AUC values as low as 51–58%, accompanied by large drops in sensitivity and unstable results across folds. This suggests that the increased emphasis on hard-to-classify samples led to over-penalization of uncertain cases, amplifying noise and destabilizing training.

Table 6.5: Performance results on the test set for models trained with Focal loss across loss weighting parameters. The highest value in each column is highlighted in bold.

Configuration	AUC	Accuracy	Precision	Sensitivity	Specificity
$\gamma = 0$ (baseline)	**0.81 ± 0.02**	**0.83 ± 0.02**	**0.87 ± 0.06**	**0.71 ± 0.05**	0.91 ± 0.04
$\gamma = 1$	0.78 ± 0.02	0.80 ± 0.02	0.86 ± 0.03	0.65 ± 0.04	**0.92 ± 0.02**
$\gamma = 2$	0.58 ± 0.03	0.60 ± 0.03	0.57 ± 0.06	0.40 ± 0.11	0.75 ± 0.11
$\gamma = 3$	0.51 ± 0.04	0.49 ± 0.06	0.48 ± 0.09	0.60 ± 0.27	0.41 ± 0.29

6.5 Pseudo-Label Threshold Influence

The impact of different pseudo-label confidence thresholds was also tested, with values set to 0.99, 0.95 (baseline), 0.90, 0.85, and 0.80. This parameter regulates the level of certainty required before assigning a pseudo-label to an unlabeled sample. Higher thresholds ensure that only highly confident predictions are used for training, whereas lower thresholds admit a larger number of pseudo-labeled samples, potentially increasing training capacity but also introducing more noise.

Across all thresholds, AUC remained relatively stable, typically between 80 to 82%, with accuracy ranging from 81 to 83%, precision from 84 to 87%, and specificity from 89 to 91%. The most notable variability was observed in sensitivity, which fluctuated between 70 to 75%.

At the highest threshold (0.99), performance was somewhat constrained: AUC values were approximately 81%, with other metrics also slightly falling compared to the baseline, besides sensitivity, which saw a gain of 2%. This conservative strategy severely restricted the number of pseudo-labels, limiting the contribution of unlabeled data and thereby diminishing the potential benefit of the semi-supervised approach.

At moderately lower thresholds (0.90 and 0.85), performance remained competitive, with AUC around 82% and accuracy 82 to 83%. Sensitivity was generally stable at 73 to 75%.

At the lowest threshold (0.80), performance degraded. AUC values fell to approximately 80%, with sensitivity dropping to 70%, despite precision and specificity remaining high. This indicates that the larger volume of pseudo-labels admitted at this threshold had a detrimental impact of increased label noise.

Table 6.6: Performance results on the test set across pseudo-label thresholds. The highest value in each column is highlighted in bold.

Configuration	AUC	Accuracy	Precision	Sensitivity	Specificity
threshold = 0.99	0.81 ± 0.03	0.82 ± 0.02	0.84 ± 0.04	0.73 ± 0.06	0.89 ± 0.03
threshold = 0.95 (baseline)	0.81 ± 0.02	**0.83 ± 0.02**	**0.87 ± 0.06**	0.71 ± 0.05	**0.91 ± 0.04**
threshold = 0.90	**0.82 ± 0.00**	0.82 ± 0.01	0.84 ± 0.05	**0.75 ± 0.05**	0.89 ± 0.05
threshold = 0.85	**0.82 ± 0.04**	**0.83 ± 0.03**	0.86 ± 0.03	0.73 ± 0.08	**0.91 ± 0.03**
threshold = 0.80	0.80 ± 0.03	0.81 ± 0.02	0.86 ± 0.03	0.70 ± 0.07	**0.91 ± 0.03**

With these results, we could conclude that the 0.85 to 0.95 range provided the most favourable balance, high enough to control label noise, while also permissive to make use of the available unlabeled data. Very high thresholds, such as 0.99, were overly restrictive, while low thresholds, such as 0.80, admitted too much noise, leading to reduced sensitivity without compensatory gains in AUC. Overall, it is suggested that 0.85 was the most effective configuration, despite not having the best precision and sensitivity values, with a good trade-off for this semi-supervised environment.

6.6 Proportions of Unlabelled Data

The central experiment of this dissertation evaluated the impact of incorporating different proportions of unlabeled data within the FixMatch semi-supervised learning framework. The proportion of unlabeled samples was varied from 0N (baseline, fully supervised) to 5N, where N represents the number of labeled samples.

With the inclusion of unlabeled data equal to N, performance remained equally competitive with the baseline, losing 1% on accuracy and specificity, and 2% in precision, but gaining 2% on sensitivity.

When we stepped the N value to 2 and 3, the AUC value started slowly increasing, and we could observe an improvement in sensitivity values, which were up to 3%. As for the rest of the metrics, the values were practically as good as the baseline.

At 4N, results slightly dropped when compared to smaller proportions of unlabelled data regarding AUC, accuracy, and sensitivity, only remaining similar to the N configuration values.

At the highest proportion tested, 5N, the model achieved AUC values of 83% and sensitivity values of 77%, which were 2% and 6% better than the baseline, respectively, despite losing 3 pp in the precision and specificity metrics. The other variables also maintained stable results. These findings at this configuration were somewhat unexpected, due to the poor results at 4N. However, it indicates that under favourable conditions, large volumes of unlabeled data can improve the model's ability to identify malignant cases.

Table 6.7: Performance results on the test set across unlabeled data proportions. The highest value in each column is highlighted in bold.

Configuration	AUC	Accuracy	Precision	Sensitivity	Specificity
0N (baseline)	0.81 ± 0.02	**0.83 ± 0.02**	**0.87 ± 0.06**	0.71 ± 0.05	**0.91 ± 0.04**
N	0.81 ± 0.01	0.82 ± 0.01	0.85 ± 0.03	0.73 ± 0.03	0.90 ± 0.03
2N	0.82 ± 0.02	**0.83 ± 0.02**	0.86 ± 0.04	0.74 ± 0.05	**0.91 ± 0.03**
3N	0.82 ± 0.01	**0.83 ± 0.01**	0.86 ± 0.03	0.74 ± 0.03	0.90 ± 0.02
4N	0.81 ± 0.02	0.82 ± 0.02	0.86 ± 0.02	0.72 ± 0.05	0.90 ± 0.02
5N	**0.83 ± 0.02**	**0.83 ± 0.02**	0.84 ± 0.05	**0.77 ± 0.04**	0.88 ± 0.04

The main study confirms that incorporating unlabeled data has a positive but bounded impact on performance. Gains were evident at intermediate proportions ($1N - 3N$), but specifically at very high proportions ($5N$) with slightly improved AUC and sensitivity, which, in the clinical context, have a higher significance. Importantly, no severe degradation was observed, indicating that FixMatch can effectively make use of unlabeled data while controlling label noise.

6.7 Model Explainability

A Grad-CAM analysis was applied to five representative cases to visualize how the baseline and the semi-supervised model would distinguish the most relevant spatial regions for the classification task.

The analysis revealed distinct differences between the two configurations. The fully supervised model generally produced more scattered heatmaps, highlighting multiple regions across the image. This could indicate that it integrates contextual information beyond the immediate lesion. However, it also suggests a lack of specificity in localizing the actual lung nodule.

In contrast, the semi-supervised model demonstrated heatmaps that were more spatially concentrated, with the activation often focused on a single and well-defined area of the image. This could seem like an improvement over the supervised model, but looking at the focal points, we could observe that they frequently failed to coincide with the region of interest, which is the lung nodule itself.

These observations imply that the results of Grad-CAM are inconclusive. Although the semi-supervised model achieved a more localized focus, this did not consistently correspond to clinically relevant structures.

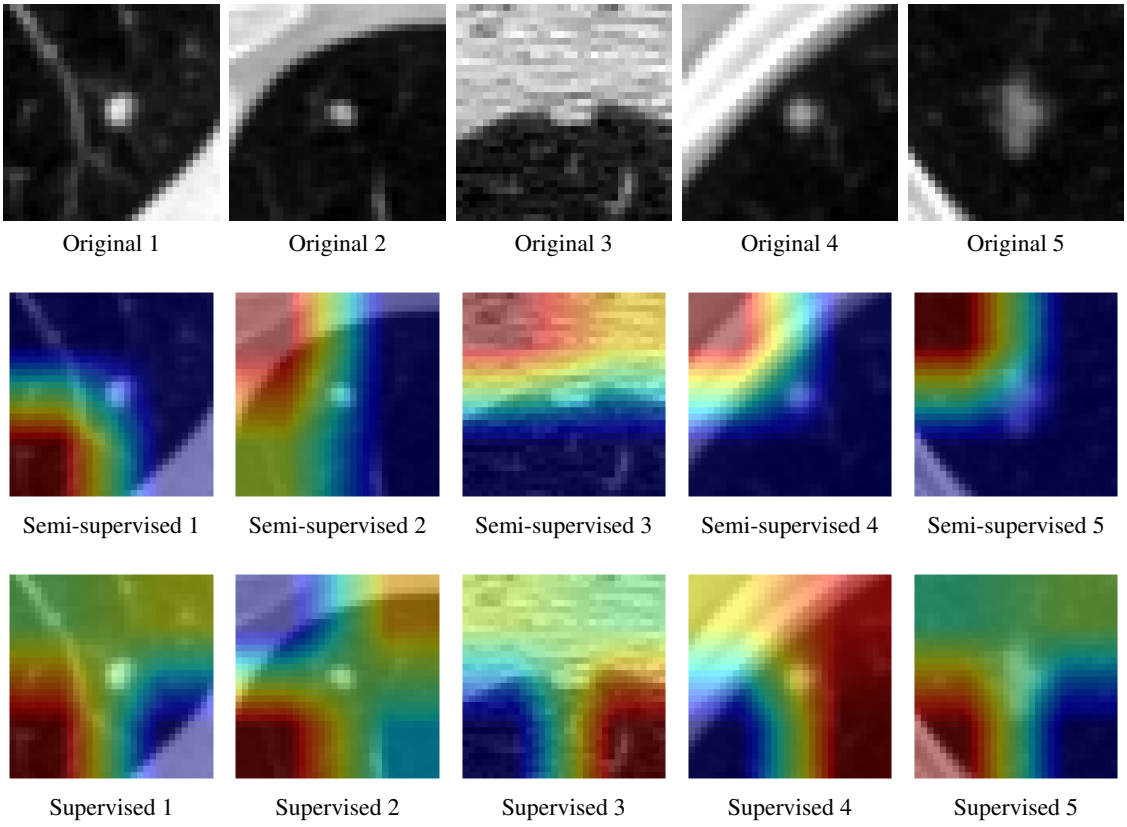

Figure 6.1: Grad-CAM visualizations: originals (top row), semi-supervised model (middle row), and supervised model (bottom row).

Chapter 7

Conclusions

7.1 Research Questions

In the introduction, four research questions were formulated to guide the objectives and scope of this dissertation:

1. **To what extent does the combination of supervised and unsupervised learning within a semi-supervised framework improve model performance for lung cancer malignancy classification, compared to using only supervised learning?**

   The findings of this dissertation demonstrate that incorporating unlabeled data through a semi-supervised framework, such as FixMatch, can lead to measurable but moderate improvements over a purely supervised baseline. Across the experiments that were tested, AUROC and sensitivity metrics managed to obtain slight increases compared to the baseline of 2% and 6%, respectively, when unlabeled data was introduced, despite having precision and specificity slightly lowered. These positive results, especially in the AUROC and the sensitivity metrics, which are critical in a clinical setting, indicate that the semi-supervised approach enhances the model's ability to correctly identify malignant cases, with a small variance in specificity.

2. **What is the optimal ratio of labeled to unlabeled data in the dataset(s) that maximizes the performance of the semi-supervised learning algorithm?**

   The experiments demonstrated that incorporating unlabeled data improved performance up to a point, with the effect depending on the ratio considered. At $1N$, performance remained

almost unchanged compared to the supervised baseline, while $2N$ and $3N$ provided modest improvements in AUROC and sensitivity. The other metrics remained relatively stable. These ratios thus represented the most consistent and balanced configurations. At $4N$, performance declined slightly, particularly in sensitivity, indicating diminishing returns, and positioning itself close to the baseline and N configuration. Interestingly, at $5N$, the model achieved its highest AUROC (83%) and sensitivity (77%), although precision and specificity dropped slightly relative to the best intermediate and baseline settings.

Taken together, these results suggest that the optimal ratio depends on the evaluation priority. From a methodological standpoint, $2N-3N$ offered the most stable trade-off across metrics. However, from a clinical perspective, where sensitivity is paramount, $5N$ can be considered the most favorable configuration, as it maximizes the detection of malignant cases without severely compromising other metrics.

3. **How do ablation study results vary with changes in conditions such as bounding-box definitions, pseudo-label thresholds, different CNNs, and different loss functions?**

   The ablation studies demonstrated that individual design choices had distinct and sometimes contrasting effects on model performance. Changes in bounding-box size and dimensionality showed limited benefit: larger 2D BBs and simple 2.5D representations did not systematically improve results, and in some cases introduced variability, suggesting that the additional context was influenced by a larger amount of noise, a more complex input, and possibly a slight slice misalignment. Regarding model architecture, EfficientNet-B0 consistently provided the most stable and accurate performance, while ResNet18 was moderately competitive, and ConvNeXt underperformed in this medical imaging setting, suggesting that lighter-weight models are more prone to correctly classify lung nodules' malignancy.

   Concerning pseudo-label thresholds, the experiments confirmed the importance of balance. Thresholds that were too strict limited the contribution of unlabeled data, while overly permissive thresholds introduced label noise. Intermediate thresholds (0.85–0.95) offered the most reliable trade-off between utilizing additional data and controlling that noise. Finally, the evaluation of loss functions revealed that Focal Loss was not well-suited to this classification task, whereas combining BCE with extreme Dice Loss values (either too low or too high) produced improved outcomes by complementing pixel-level discrimination with regional overlap optimization.

4. **To what extent can explainability techniques like Grad-CAM reveal differences in the decision-making process between models trained with supervised learning and those trained using a semi-supervised algorithm for lung cancer malignancy classification?**

   The Grad-CAM experience provided some insights, but they appeared too limited for understanding the reasoning behind the supervised and semi-supervised models for classifying lung cancer malignancy. The visualizations showed that the fully supervised model used a more diffuse strategy for feature extraction, paying attention to many scattered regions in

the image. In comparison, the semi-supervised model's activation maps were more focused with areas of greater intensity. However, this alignment of focus and attention did not appear clinically relevant. The semi-supervised model often neglected to put its attention on the lung nodule, which should have been the key area of focus.

Therefore, although some simulations showed differences between both configurations, with the supervised model's focus on many regions, and the semi-supervised model with a sharper focus, those were not enough to suggest any increased explainability or reliability that could be used in clinical contexts.

7.2 Hypothesis

The initial objectives of this dissertation were largely supported by the experimental data, though confirmation was uneven across different dimensions. Primarily, the assumption that integration of unlabeled examples through a semi-supervised strategy would surpass the performance of a purely supervised model was substantiated. The FixMatch framework delivered measurable, although modest, gains that were repeatable across folds: sensitivity and AUROC were the principal beneficiaries, while accuracy, precision, and specificity remained virtually stable.

The second assumption, that certain architectural and operational parameters, including input dimensionality, bounding-box specification, backbone design, loss architecture, and the set point for pseudo-label acceptance, would influence the size of the observed gain, was corroborated by systematic ablation. The EfficientNet-B0 backbone proved optimal, the fusion of binary-cross-entropy and Dice loss yielded the most stable convergence, and a balanced pseudo-label threshold maximized the information without permitting unwanted noise across the datasets. Conversely, variations in bounding-box heuristics and the introduction of 2.5D representations produced no consistent improvements, thus modestly refuting the hypothesis that richer spatial representations would necessarily provide a performance gain.

Finally, the explainability experiences utilizing Grad-CAM revealed that supervised and semi-supervised models diverged in their attention patterns, with the first displaying scattered activations and the latter more concentrated focal points. However, these points often didn't coincide with the regions of interest, the lung nodules, limiting their clinical interpretability. Ultimately, while Grad-CAM highlighted some differences in model behavior, it did not provide clear evidence that semi-supervised training improved transparency or reliability in decision-making. This could indicate the need for complementary interpretability methods and expert validation.

7.3 Contributions

This dissertation makes several contributions to the field of computer-aided lung cancer malignancy classification using deep learning and semi-supervised methods:

- **The empirical evaluation of semi-supervised learning techniques in medical imaging:**

In this research, the efficacy of FixMatch was analyzed for the classification of lung nodule malignancy using a combination of labeled and unlabeled data. The findings confirmed the hypothesis that the presence of unlabeled data slightly improves performance in comparison to a purely supervised framework, especially in terms of sensitivity and AUC, which are clinically critical metrics.

- **Focus on systematic ablation studies to test defined hypotheses:**

  To assess the impact of bounding-box definitions, input dimensions, pseudo-label thresholds, model architectures, and loss functions, a broad set of ablation experiments was conducted. These experiments improved the understanding of design choices that affected the model performance meaningfully and those that had little or counterproductive impact.

- **Improved Explainability in AI-based diagnostics:**

  The architecture was also scrutinized through explainability lenses using Grad-CAM, which illustrates the most salient image areas that affect the model's outputs. This kind of transparency supports the credibility of semi-automated clinical tools.

- **Alignment with the United Nations Sustainable Development Goals:**

  The work aligns with SDG 3 and SDG 9, representing health and well-being, and Industry and Innovation, respectively, by advancing non-invasive AI methods for early lung cancer detection and exploring innovation in semi-supervised learning for medical imaging.

7.4 Future Work

Several directions for future research emerge from the work done in this dissertation:

- **Integration of molecular and genetic data:**

  Exploring whether the proposed framework can predict or incorporate the mutation status of different lung cancers (EGFR, KRAS, etc.) would provide a deeper link between imaging phenotypes and tumor biology. This would extend the clinical relevance of the model and also allow for precision oncology applications.

- **Generalization to 3D data:**

  Even in this work, where 2D and 2.5D representations took the main role, there can be further work done by considering full 3D volumetric inputs. True 3D architectures have the potential to capture more spatial context and to identify faint patterns of malignancy.

- **Integration of fusion methods in a semi-supervised framework:**

  Extending FixMatch to feature-level or decision-level fusion across inputs may further improve generalization. This could potentially allow the semi-supervised approach to use complementary information more effectively.

- **Experimentation on other datasets:**

  The generalizability capacity of the framework could be improved by utilizing more diverse and possibly larger sets of data in the pipeline, beyond LIDC-IDRI and Luna25.

- **Validation against real-world clinical scenarios:**

  Arguably, the most critical step towards this work's translation is external verification in real-world clinical practice. Testing the framework with several hospital datasets would demonstrate the practical utility of the proposed methods.

References

[1] American Cancer Society. What Is Lung Cancer?, 2024.

[2] American Lung Association. Types of Lung Cancer, 2024.

[3] Ioannis D Apostolopoulos, Nikolaos D Papathanasiou, and George S Panayiotakis. Classification of lung nodule malignancy in computed tomography imaging utilising generative adversarial networks and semi-supervised transfer learning. *Biocybernetics and Biomedical Engineering*, 41(4):1243–1257, 2021. Publisher: Elsevier.

[4] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. Van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. Van Casteele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke. Data From LIDC-IDRI, 2015. Published: The Cancer Imaging Archive.

[5] Samuel G. Armato III, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Edwin J. R. van Beek, David Yankelevitz, Alberto M. Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, Gary E. Laderach, Daniel Max, Richard C. Pais, David P.-Y. Qing, Rachael Y. Roberts, Amanda R. Smith, Adam Starkey, Poonam Batra, Philip Caligiuri, Ali Farooqi, Gregory W. Gladish, C. Matilda Jude, Reginald F. Munden, Iva Petkovska, Leslie E. Quint, Lawrence H. Schwartz, Baskaran Sundaram, Lori E. Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallam, Michael D. Heath, Michael H. Kuhn, Ekta Dharaiya, Richard Burns, David S. Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, Barbara Y. Croft, and Laurence P. Clarke. The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics*, 38(2):915–931, 2011.

[6] Yi Lin Luyang Luo Hao Chen Cheng Jin, Zhengrui Guo. Label-Efficient Deep Learning in Medical Image Analysis: Challenges and Future Directions. *Medical Image Analysis*, 2023. Backup Publisher: The Hong Kong University of Science and Technology Publisher: Elsevier.

[7] James L. Crowley. Convolutional Neural Networks. In Mohamed Chetouani, Virginia Dignum, Paul Lukowicz, and Carles Sierra, editors, *Human-Centered Artificial Intelligence: Advanced Lectures*, pages 67–80. Springer International Publishing, Cham, 2023.

[8] Tami D. DenOtter and Johanna Schubert. *Hounsfield Unit*. StatPearls Publishing, Treasure Island (FL), 2025.

[9] E. M. Rodrigues, M. Gouveia, H. P. Oliveira, and T. Pereira. Efficient-Proto-Caps: A Parameter-Efficient and Interpretable Capsule Network for Lung Nodule Characterization. *IEEE Access*, 13:56616–56630, 2025.

[10] Jan-Niklas Eckardt, Martin Bornhäuser, Karsten Wendt, and Jan Moritz Middeke. Semi-supervised learning in cancer diagnostics. *Frontiers in oncology*, 12:960984, 2022. Publisher: Frontiers Media SA.

[11] J. Ferlay, M. Ervik, F. Lam, M. Laversanne, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray. Global Cancer Observatory: Cancer Today, Fact Sheet on All Cancers, 2024.

[12] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980.

[13] Zoubin Ghahramani. Unsupervised Learning. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, pages 72–112. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[15] M. O. Idowu and C. N. Powers. Lung cancer cytology: potential pitfalls and mimics - a review. *International Journal of Clinical and Experimental Pathology*, 3(4):367–385, March 2010.

[16] Rushi Jiao, Yichi Zhang, Le Ding, Bingsen Xue, Jicong Zhang, Rong Cai, and Cheng Jin. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, page 107840, 2023. Publisher: Elsevier.

[17] Abdul M. Khan, Navjot Dullet, and Maryam S. Esfahani. Computed Tomography of the Thorax. *Treasure Island (FL): StatPearls Publishing*, 2022.

[18] Olga Kurasova, Viktor Medvedev, Aušra Šubonienė, Gintautas Dzemyda, Aistė Gulla, Artūras Samuilis, Džiugas Jagminas, and Kęstutis Strupas. Semi-Supervised Learning with Pseudo-Labeling for Pancreatic Cancer Detection on CT Scans. In *2023 18th Iberian conference on information systems and technologies (CISTI)*, pages 1–6. IEEE, 2023.

[19] Jinping Lao, Hongwei Lin, Haiyu Zhou, Chengchuang Lin, Zhaoliang Zheng, Gansen Zhao, and Hua Tang. Detection of High-Low Risk Lung Tumors Using Semi-Supervised and Selective Labeling Techniques. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[21] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999–7019, 2022.

[22] Bing Liu. Supervised Learning. In *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, pages 63–132. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, June 2022.

[24] Roberto Augusto Philippi Martins and Danilo Silva. On Teacher-Student Semi-Supervised Learning for Chest X-ray Image Classification. *Anais do*, 15, 2021.

[25] Joana Morgado, Tania Pereira, Francisco Silva, Cláudia Freitas, Eduardo Negrão, Beatriz Flor de Lima, Miguel Correia da Silva, António J Madureira, Isabel Ramos, Venceslau Hespanhol, and others. Machine learning and feature selection methods for egfr mutation status prediction in lung cancer. *Applied Sciences*, 11(7):3273, 2021. Publisher: MDPI.

[26] Phuong Nguyen, Ankita Rathod, David Chapman, Smriti Prathapan, Sumeet Menon, Michael Morris, and Yelena Yesha. Active semi-supervised learning via Bayesian experimental design for lung cancer classification using low dose computed tomography scans. *Applied Sciences*, 13(6):3752, 2023. Publisher: MDPI.

[27] null null. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

[28] World Health Organization. The Top 10 Causes of Death, 2024.

[29] Keiron O'Shea and Ryan Nash. An Introduction to Convolutional Neural Networks, 2015. _eprint: 1511.08458.

[30] Jeroen Peeters, Colin Jacobs, Matthijs Oudkerk, Matthieu Schreurs, Bram van Ginneken, and Eva van Rikxoort. LUNA25: A Benchmark for Lung Nodule Detection in Chest CT. *arXiv preprint arXiv:2405.04605*, 2024.

[31] C. Pinheiro, F. Silva, T. Pereira, and H.P. Oliveira. Semi-Supervised Approach for EGFR Mutation Prediction on CT Images. *Mathematics*, 10(4225), 2022. Publisher: MDPI.

[32] C. Pinheiro, F. Silva, T. Pereira, and H.P. Oliveira. Semi-Supervised Approach for EGFR Mutation Prediction on CT Images. *Mathematics*, 10(4225), 2022.

[33] Gil Pinheiro, Tania Pereira, Catarina Dias, Cláudia Freitas, Venceslau Hespanhol, José Luis Costa, António Cunha, and Hélder P Oliveira. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. *Scientific reports*, 10(1):3625, 2020. Publisher: Nature Publishing Group UK London.

[34] D. Pérez-Callejo, A. Romero, M. Provencio, and M. Torrente. Liquid biopsy based biomarkers in non-small cell lung cancer for diagnosis and treatment monitoring. *Translational Lung Cancer Research*, 5(5):455–465, October 2016.

[35] Marco Russano, Andrea Napolitano, Giulia Ribelli, Michele Iuliani, Sonia Simonetti, Fabrizio Citarella, Francesco Pantano, Emanuela Dell'Aquila, Cecilia Anesi, Nicola Silvestris, Antonella Argentiero, Antonio Giovanni Solimando, Bruno Vincenzi, Giuseppe Tonini, and Daniele Santini. Liquid biopsy and tumor heterogeneity in metastatic solid tumors: the potentiality of blood samples. *Journal of Experimental & Clinical Cancer Research*, 39(1):95, May 2020.

[36] A. Shamila Ebenezer, S. Deepa Kanmani, Mahima Sivakumar, and S. Jeba Priya. Effect of image transformation on EfficientNet model for COVID-19 CT image classification. *International Conference on Advances in Materials Science*, 51:2512–2519, January 2022.

[37] Feng Shi, Bojiang Chen, Qiqi Cao, Ying Wei, Qing Zhou, Rui Zhang, Yaojie Zhou, Wenjie Yang, Xiang Wang, Rongrong Fan, and others. Semi-supervised deep transfer learning for benign-malignant diagnosis of pulmonary nodules in chest CT images. *IEEE Transactions on medical imaging*, 41(4):771–781, 2021. Publisher: IEEE.

[38] Francisco Silva, Tania Pereira, Joana Morgado, António Cunha, and Hélder P Oliveira. The impact of interstitial diseases patterns on lung CT segmentation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2856–2859. IEEE, 2021.

[39] Francisco Silva, Tania Pereira, Joana Morgado, Julieta Frade, José Mendes, Cláudia Freitas, Eduardo Negrão, Beatriz Flor De Lima, Miguel Correia Da Silva, António J. Madureira, Isabel Ramos, Venceslau Hespanhol, José Luís Costa, António Cunha, and Hélder P. Oliveira. EGFR Assessment in Lung Cancer CT Images: Analysis of Local and Holistic Regions of Interest Using Deep Unsupervised Transfer Learning. *IEEE Access*, 9:58667–58676, 2021.

[40] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc., 2020.

[41] Zahra Solatidehkordi and Imran Zualkernan. Survey on Recent Trends in Medical Image Classification Using Semi-Supervised Learning. *Applied Sciences*, 12(23):12094, 2022.

[42] V Sudharsanam, VD Vishnusripriya, Yagnavajjula Likhitha, R Thenila, and others. Detection of COVID-19 on Lung CT Images Using Semi Supervised Learning. In *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, pages 973–977. IEEE, 2021.

[43] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N. Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.

[44] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[45] M.A. Thanoon, M.A. Zulkifley, M.A.A. Mohd Zainuri, and S.R. Abdani. A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images. *Diagnostics*, 13(16):2617, 2023. Publisher: MDPI.

[46] T. Baird A.M. et al. The Health Policy Partnership, Albreht. Lung cancer in Europe: the way forward, 2022.

[47] Joseph F. Tomashefski and Carol F. Farver. Anatomy and Histology of the Lung. In Joseph F. Tomashefski, Philip T. Cagle, Carol F. Farver, and Armando E. Fraire, editors, *Dail and Hammar's Pulmonary Pathology: Volume I: Nonneoplastic Lung Disease*, pages 20–48. Springer New York, New York, NY, 2008.

[48] A. Tuzi, E. Bolzacchini, M. B. Suter, A. Giaquinto, A. Passaro, S. Gobba, I. Vallini, and G. Pinotti. Biopsy and re-biopsy in lung cancer: the oncologist requests and the role of endobronchial ultrasound transbronchial needle aspiration. *Journal of Thoracic Disease*, 9(Suppl 5):S405–S409, 2017.

[49] Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, February 2020.

[50] Guotai Wang, Shuwei Zhai, Giovanni Lasio, Baoshe Zhang, Byong Yi, Shifeng Chen, Thomas J Macvittie, Dimitris Metaxas, Jinghao Zhou, and Shaoting Zhang. Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung CT scans with multi-scale guided dense attention. *IEEE transactions on medical imaging*, 41(3):531–542, 2021. Publisher: IEEE.

[51] Lulu Wang. Deep learning techniques to diagnose lung cancer. *Cancers*, 14(22):5569, 2022. Publisher: MDPI.

[52] C. C. Wu, M. M. Maher, and J. A. Shepard. Complications of CT-guided percutaneous needle biopsy of the chest: prevention and management. *AJR. American Journal of Roentgenology*, 196(6):W678–W682, June 2011.

[53] Yutong Xie, Jianpeng Zhang, and Yong Xia. Semi-supervised adversarial model for benign–malignant lung nodule classification on chest CT. *Medical image analysis*, 57:237–248, 2019. Publisher: Elsevier.

[54] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2022. Publisher: IEEE.

[55] Guojin Zhang, Yuntai Cao, Jing Zhang, Jialiang Ren, Zhiyong Zhao, Xiaodi Zhang, Shenglin Li, Liangna Deng, and Junlin Zhou. Predicting EGFR mutation status in lung adenocarcinoma: development and validation of a computed tomography-based radiomics signature. *American journal of cancer research*, 11(2):546, 2021. Publisher: e-Century Publishing Corporation.

[56] Li Zhou, Ashley Fong, and Anthony J. Viera. Computed Tomography (CT). *StatPearls [Internet]*, 2023. Publisher: StatPearls Publishing.