FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

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

**Afonso José Pinheiro Oliveira Esteves Abreu**


Mestrado em Engenharia Informática

Supervisor: Eduardo Rodrigues

March 10, 2025

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

**Afonso José Pinheiro Oliveira Esteves Abreu**

Mestrado em Engenharia Informática

March 10, 2025

# Contents

# Chapter 1

# Introduction

## 1.1 Context

Among all types of cancer, lung cancer is the one who causes the most damage in human health [4] as it is the 6th leading cause of deaths of our species, along with trachea and bronchus [12]. The reason for this remains in the fact that lung cancer is often diagnosed in a late stage of the disease. It happens that in this period, less than 10% of people survive the 5-year survival rate, which reflects the percentage of people still alive after five years of being diagnosed with this disease [23]. This imposes a major challenge in medical treatments as newer methods require earlier detection of lung cancer in order to effectively treat it.

Tissue biopsy has been the main method to identify and characterise lung cancer [5]. 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 [26].

Computer-aided diagnosis (CAD) has gained a better reputation over the time as it can come with deep learning models able to provide a non-invasive tumour characterisation based on CT (computed tomography) scans of the human thorax [22]. 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 tumour 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 Problem

Training deep learning models need a lot of data for them to be reliable and robust [13]. One of the main challenges is that many existing datasets contain a significant amount of unlabelled data due to the difficulty of obtaining annotations. Especially in the case of medical images, the process is often time-consuming and costly, potentially even requiring additional examinations.

Therefore, a framework that enhances predictive performance by effectively utilizing both labelled and unlabelled data is needed.

## 1.3 Hypothesis

We hypothesise that a semi-supervised learning framework can effectively tackle the aforementioned challenges faced by deep learning models. As it takes into account labelled and unlabelled data, newer models could combine both types of datasets in order to have better predictive abilities [20] and help therapeutic decisions furthermore.

## 1.4 Motivation

The aim of better quality of life and healthcare is relevant for this paper, specially when one talks about lung cancer. Due to its complications in the early stages during diagnosis we hope we can develop a method that efficiently predicts the behavior of the disease in order to accurately indicate personalized treatments. The use of semi-supervised learning [2] can revolutionize the way we have been treating lung cancer, starting with non invasive procedures that takes into account a patient's well being.

## 1.5 Research Questions

This paper has the objective of creating a semi-supervised learning model in order to accurately predict the evolution of lung cancer. It makes use of supervised as well as unsupervised learning so it can potentially have more accurate results.

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 labelled to unlabelled data in the dataset(s) that will be selected for this project to achieve the best results in the semi-supervised learning model?

3. How does the number of parameters in the semi-supervised framework compare to that of a supervised learning model?

4. How does the computational time required for training and prediction vary across different combinations of models?

# Chapter 2

# Literature Review

Artificial Intelligence has gained significant relevance and has surely evolved rapidly through the recent years. Although it needs to be more powerful and smart to replace humans in the field of medicine, it can provide several tools to aid professionals in certain tasks which were once considered to be very time-consuming. Analyzing images and further support in medical decisions are one of the advantages we can take from Artificial Intelligence in healthcare, potentially becoming a shift of paradigm in the field.

This chapter aims to explore the recent advancements and challenges of using deep semi-supervised learning to address the shortage of annotated data in cancer characterization. Although labeled images play a crucial role in training machine learning models, these are often costly and impractical, leaving a potential spot for new deep learning models that take into account limited labeled data.

This literature review will also include several studies in the area of AI and deep semi-supervised learning regarding medical imaging. Distinct strategies and approaches used in previous investigations will be presented, each showing the obtained results, which will then be interpreted and carefully analyzed.

## 2.1 Eligibility criteria

For this stage, it was considered a systematic review in order to only get the pertinent studies. Eligibility criteria was established to include deep semi-supervised learning approaches in cancer characterization, which takes into account the limited annotated data. All of the researched studies were in the English language so that we could get better accessibility. Also, the papers should also have been published after the year 2020, ensuring that only the newest methods in this fast-evolving area are approached and therefore interpreted and analyzed. Besides that, it was a necessity that the papers had objective performance outcomes, so one could directly compare different models through several key aspects, such as accuracy and AUC.

To increase the number of papers obtained and also to not specify on the theme too much, other studies that focused on investigating a single technology or using deep learning models combined

with CT scans regarding other diseases were also approached, since they could turn out to be much useful for development of this thesis.

## 2.2 Collection and Search strategy

In order to gather relevant studies, a search strategy was conducted and it aimed to follow the PRISMA 2020 guidelines, guaranteeing diligence in the research of studies for the main problem acknowledged.

As for information sources, these included mainly PubMed, MDPI, IEEE Xplore, and Google Scholar, given their significance in areas such as healthcare and engineering. Search strategies for these databases used combined terms and keywords such as "semi-supervised learning", "cancer characterization" and "CT scans".

Another strategy that was utilized was based on the reading of a document's references. This helped since it could very much redirect to another interesting paper that could not be found in the basic search of the databases referenced above.

## 2.3 Screening and selection process

Regarding the data selection and screening, studies were initially picked if the title or abstract referenced the semi-supervised approach in tumor prediction for any type of cancer with the use of CT scans, always taking into account the lack of labeled data. Then, a broader set of studies that investigated certain technologies of deep learning were also picked. A clear and objective conclusion was also a key point for selection, exposing the various results obtained during the experiments. This technique ensured a somewhat quick selection process since the whole document was only read if it contained sufficient interesting information.

Although no data was extracted during this delivery, some results were carefully looked for in the studies. Such were the used algorithms, combined with their performance levels, accuracy, and AUC metrics, along with their outcomes.

## 2.4 Conclusion

This literature review reinforces the necessity of further research to refine approaches for enhanced interpretability and clinical applicability.

All these topics: searching, collection, screening and selecting were taken into account for the careful gathering of relevant papers and studies that could be a major benefit for this thesis, ensuring a good set of bibliographies that enriches and increases validation.

By building on the insights from these studies, future work can contribute to developing a robust solution that supports early cancer diagnosis and treatment planning.

# Chapter 3

# State of the Art

## 3.1 Introduction

Oncological diseases have for a long time been a major problem in healthcare. Since they are very difficult to treat, the best way to fight cancer is to prevent it and to diagnose it in a very early stage. Artificial Intelligence can play in a big part for this topic by developing models capable of identifying different types of tumors, leading the way to a specialized treatment without being too harsh for the human body, unlike for example the biopsy.

The use of supervised learning needs a large bank of annotated data, which in the present day is a major drawback, because of the expenditure and time-consuming job of expertise labeling. Lately, many investigations had the objective to test the effectiveness of another machine learning method: semi-supervised learning. This approach surpasses the problem of having few annotated images taking also into account images without labeling information to create models capable of doing the same tasks.

This chapter aims to describe the different application methods of semi-supervised learning and how they can gain an advantage over previous models in cancer characterization.

## 3.2 AI-based solutions with Semi-supervised learning

O. Kurasova et al. [7] 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 in 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 firstly trained using the labeled data. After that, the trained model was used to predict pseudo-labels to unlabeled images if the prediction probability PP was above a given threshold t. This process would repeat until no data was remaining or if the condition did not meet 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 a better and stable training and improved classification metrics, compared to a supervised learning model only.

Furthermore, Roberto Philippi and Danilo Silva. [9], 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. The results were calculated using the AUROC values and ranges for the Teacher model went from 0.750 to 0.887 and for the Student model tehy went from 0.822 to 0.893.

Now focusing on the structure of the lung, Phuong Nguyen et al. [11] 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 fisrt 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.

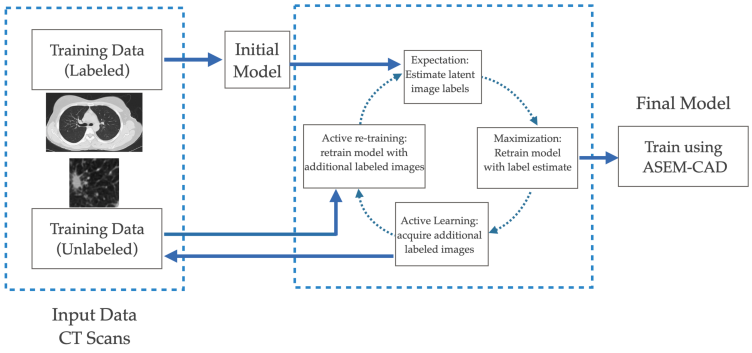


Figure 3.1: ASEM-CAD framework.

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 compared to a fully supervised learning model.

In F. Shi et al. [17] 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 unanotated images with more confidence are added to the training process gradually.

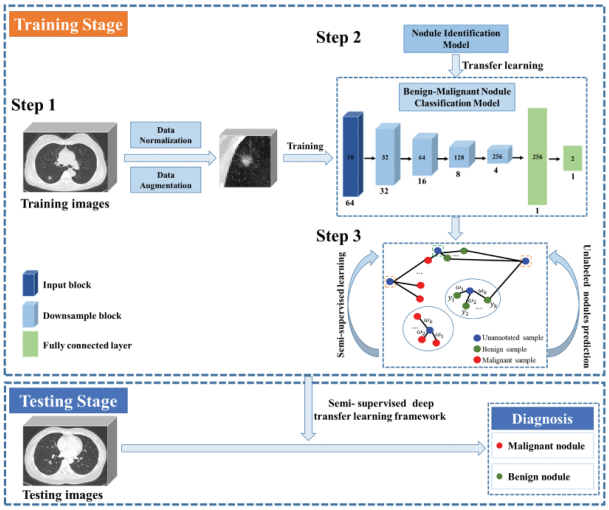


Figure 3.2: STDL framework.

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 for 2% and 2.9% accuracy improvement, respectively.

Similarly G. Wang et al. [24] 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 CNN's and 73.04% in I-CRAWL which also outperformed other semi-supervised methods like CRF-based or mean-teacher approaches.

In a thesis paper presented by a FEUP student with a similar theme to my work, Cláudia Pinheiro [15] 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 and the other, much larger, NLST, for unalabeled 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.

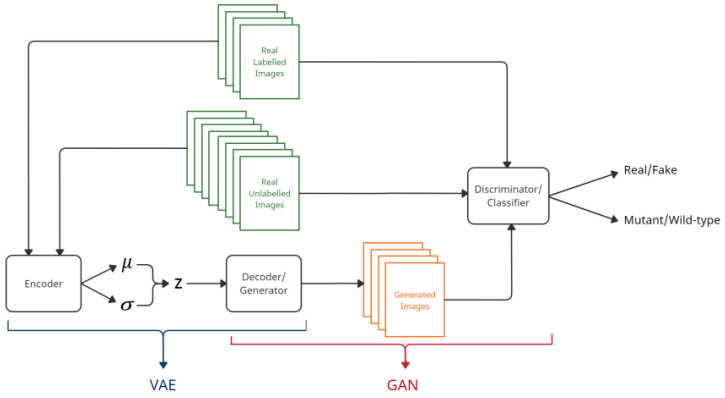


Figure 3.3: Cláudia Pinheiro's proposed method.

Yutong Xie et al. [27] 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.

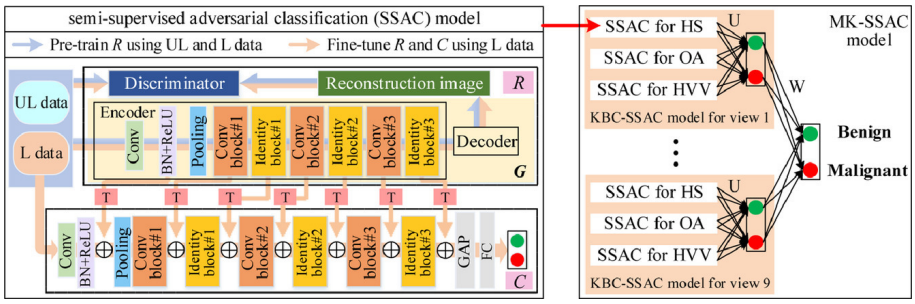


Figure 3.4: SSAC model. "UL": unlabeled data; "L": labeled data; "R": adversarial autoencoder-based reconstruction network; "C": classification network; and "G": generator which contains an encoder and decoder. "T": learnable T layer.

In order to study the relevance of certain features regarding the Epidermal growth factor receptor (EGFR) mutation status, F. Silva et al. [19] used three different regions of interest: the nodule, the lung containing the main nodule and both lungs. For that, it was developed a framework containing 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.

Local nodule analysis resulted in a low classification value of 0.51 of AUC. When considering

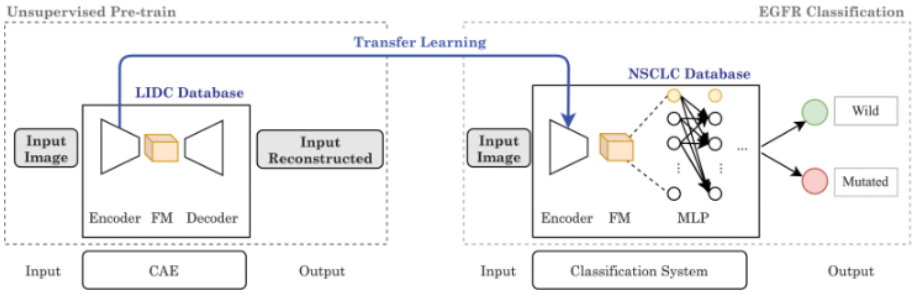


Figure 3.5: Overview of the proposed approach.

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. [8] 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.

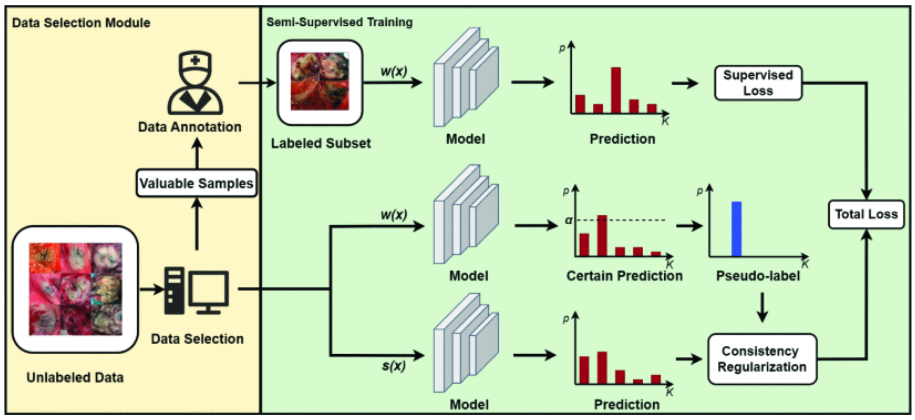


Figure 3.6: DS-FixMatch framework.

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%.

## 3.3 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 was 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 overcome the problem of 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 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.

# References


[1] 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.

[2] 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.

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

[4] 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. Accessed: 2024-10-09.

[5] Z. Hou, Y. Zhan, C. Shen, W. Zhao, K. Wang, S. Yu, S. Gao, and J. Zhu. Signaling pathways driving aberrant splicing in cancer cells. *Cancer Research*, 77(6):1168–1178, 2017.

[6] 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.

[7] 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.

[8] 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.

[9] Roberto Augusto Philippi Martins and Danilo Silva. On teacher-student semi-supervised learning for chest x-ray image classification. *Anais do*, 15, 2021.

[10] 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, et al. Machine learning and feature selection methods for egfr mutation status prediction in lung cancer. *Applied Sciences*, 11(7):3273, 2021.

[11] 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.

[12] World Health Organization. The top 10 causes of death, 2024. Accessed: 2024-10-09.

[13] C. Pinheiro, F. Silva, T. Pereira, and H.P. Oliveira. Semi-supervised approach for egfr mutation prediction on ct images. *Mathematics*, 10(4225), 2022.

[14] Cláudia Pinheiro, Francisco Silva, Tania Pereira, and Hélder P Oliveira. Semi-supervised approach for egfr mutation prediction on ct images. *Mathematics*, 10(22):4225, 2022.

[15] Cláudia Patrícia Ferreira Araújo Pinheiro. Ai-based cancer characterisation using semi-supervised learning algorithms. 2022.

[16] 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.

[17] Feng Shi, Bojiang Chen, Qiqi Cao, Ying Wei, Qing Zhou, Rui Zhang, Yaojie Zhou, Wenjie Yang, Xiang Wang, Rongrong Fan, et al. 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.

[18] 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.

[19] 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.

[20] Zahra Solatidehkordi and Imran Zualkernan. Survey on recent trends in medical image classification using semi-supervised learning. *Applied Sciences*, 12(23):12094, 2022.

[21] V Sudharsanam, VD Vishnusripriya, Yagnavajjula Likhitha, R Thenila, et al. 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.

[22] 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.

[23] T. Baird A.M. et al. The Health Policy Partnership, Albreht. Lung cancer in europe: the way forward, 2022. Accessed: 2024-10-09.

[24] 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.

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

[26] M.R. Wilkins and K.A. Paschke. Clinical practice. cystic fibrosis respiratory infections: optimizing treatment in the era of cftr modulators. *PubMed*, 364(9428):1195–1206, 2011.

[27] 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.

[28] 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.

[29] 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.