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 [2] as it is the 6th leading cause of deaths of our species, along with trachea and bronchus [4]. 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 [8]. 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 [3]. 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 [9].

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 [7]. 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 [5]. 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 [6] 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 [1] 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?

References

[1] 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.

[2] 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.

[3] 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.

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

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

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

[7] 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.

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

[9] 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.