Development of a Lung Cancer Diagnosis Support System

Nelson Faria
2Ai, School of Technology
Polytechnic Institute of Cávado and Ave
Barcelos, Portugal
a14805@alunos.ipca.pt

Sofia Campelos
IPATIMUP
University of Porto
Porto, Portugal
acampelos@ipatimup.pt

Vítor Carvalho
2Ai, School of Technology
Polytechnic Institute of Cávado and Ave
Barcelos, Portugal
vcarvalho@ipca.pt

***Abstract*—Lung cancer is the leading cause of death worldwide and, a correct diagnosis in an early-stage gives more possibilities of treatment. Whole-slide images generated from glass slides can be analysed using Artificial Intelligence technologies to help pathologists. In this study, it is given an overview of the lung cancer, exploring the methodologies done until the present. These methods are divided into Detection and Classification phases. To detect the neoplastic cells, the WSI is split into patches, and then a convolutional neural network is applied to identify and generate a heatmap highlighting the tumour regions. Then, the features are extracted from the cancerous regions and submitted in a classifier to determine the histologic type of tumour present in each patch. In addition, it is proposed a possible solution based on the literature review that could be used as an aid in the pathological diagnosis of lung cancer.**

***Index Terms*—lung cancer, digital pathology, artificial intelligence, convolutional neural networks, whole slide imaging**

I. Introduction

On a global scale, lung cancer has been the main cause of death and the second most common in terms of new causes [1]. The principal cause of this cancer is smoking, and the presence of a tumour nodule is detected by performing a radiologic detection methodology like chest X-ray or computed tomography [2], [3]. After that, a confirmatory pathologic diagnosis is usually made on small biopsy and cytology samples [4], [5]. In fact, the detection of lung cancer in an early stage is extremely important because as sooner as it is detected as greater are the chances of effective treatment and survival. However, in more than 50 % of the new cases, the tumour has already metastasized to different parts of the body where the reasons of late detection could be the lack of symptoms at early-stage, and incorrect diagnosis of the symptoms such as cough and wheezing [6]. With the emergence of whole slide imaging (Process to scan microscopic glass slides to produce digital slides), pathologists are allowed to examine the whole slide images (WSIs) on a computer which makes possible the integration of software to assist pathologists (Figure 1). In this way, this project aims to develop a system that could benefit patients and pathologists by making lung cancer diagnosis simpler, decreasing the time spent by pathologists executing the diagnosis, and improving the accuracy of the results.

II. Literature Review

A. Lung Cancer: Etiology, Classification and Detection Methodologies

According to the data collected by World Health Organization, the lung cancer numbers for deaths reaches the 1.8 million and 2.21 million for new cases in 2020, which means that this neoplasm is responsible for the highest number of deaths worldwide [7]. The most incident age group is above the fifty years old and, even though, the lung cancer appears in the lungs, it can metastasise to other organs in the human body [1]. Also, the principal cause of this type of cancer is smoking, however, other risk factors have been identified such as previous respiratory diseases, exposure to occupational carcinogens (arsenic, asbestos, chromium, nickel, and radon), polycyclic aromatic hydrocarbons, human immunodeficiency, virus infection, and alcohol consumption [1]–[3].

Carcinomas are the most common type of lung cancer being split into four types: Adenocarcinoma, Squamous Cell Carcinoma, Large Cell Carcinoma and Small Cell Lung Carcinoma. The most common is Adenocarcinoma which affects about eighty percent of cases. Currently, there are different methods to detect lung cancer such as chest x-ray, computed tomography, and magnetic resonance imaging, but when the initial detection method is in the field of radiology, a confirmation pathological diagnosis is followed by a transthoracic needle biopsy. The glass slides acquired in the biopsy will be analysed by the pathologists in a brightfield microscope, or through whole slide images that were obtained by digitising them [1].

B. Artificial Intelligence techniques applied to Lung Cancer

The increase amount of data generated by clinical systems, as well as the computational capacity, enables the development of computational systems capable of facilitating the diagnosis process, improve the accuracy and move faster to the final diagnosis.

According to the literature, the analysis of lung cancer in whole slide images can be split into two moments: Detection and Classification. Wang and his team presented a model using Inception V3 that was capable of analyse adenocarcinoma WSIs, classify and give a prognostic value with an accuracy of 89.8 %. By using the sliding window method, the model searches the presence of tumour in patches of 300 x 300 pixels. In 2018, Coudray also used Convolutional Neural Networks (CNNs) with Inception V3 for their model, but the sliding window mechanism was used for 512 x 512 pixels patches which resulted in an accuracy of 87 % [1]. Another study from Li and his team used 256 x 256 pixel patches and cropped them with a stride of 196 pixels with the aim to ensure sufficient overlapping between adjacent patches. The samples were applied into different types of CNNs, where the higher accuracy was given by AlexNet when it was trained from scratch with 97 % of accuracy and by ResNet when the pre-trained strategy was used (93 %) [1]. According to a study published in 2019 by Yu and his team, they also used the different CNN types where each WSI was split into tiles with 1000 x 1000 pixels with a 50 % overlap. The best resulting accuracy achieved from the evaluated models was 93.5 % [1]. To overcome the limitation that states the need for annotations by a pathologist to get a better result, Chen and his team designed a technique that, instead of getting just tiles of the WSI, the WSI is given as input without being split. The obtained accuracy for was nearly 93 % [1].

The detection task is followed by the classification task and provides the generated heatmap and, by applying morphological operations as erosion and dilation in these maps, the distribution and shape are features that can be extracted for further analysis and classification [8]. These features are used as inputs for models that will classify the lung cancer type present (e.g.: Adenocarcinoma, squamous cell carcinoma, etc.). While Wang and colleagues selected features associated with survival outcomes and used a univariate Cox proportional hazard model with a penalty to prevent overfitting. Yu and team employed Naïve Bayes classifiers, Support Vector Machines (SVM) with Gaussian, linear and polynomial kernels, and Breiman's random forest that received the features as input and gave the predicted lung cancer types as output, obtaining as higher accuracy 85 % [1].

III. Proposed Solution

Following the reviewed author's approach, the system to be developed must be capable of execute two phases, being them, the Detection and Classification of the tumour present in a WSI. In the detection phase, the system should load the WSI and retain all the WSI's tiles that contains tissue samples according to a given size (512 x 512 pixels). Then, it will execute transformations in the tiles such as Augmentation and Rotation to obtain more samples for training. After that, the tiles will be used in a Deep Learning Neural Network to detect the tumour using features that indicates its presence like colour and area of the cells. At the end of the detection steps, the system will compile the tiles back to the WSI format, generate a heatmap and give the prediction. In the classification phase, it will extract characteristics like perimeter and texture of the tumour from the regions of interest, apply a classifier (ex.: SVMs) that will distribute them by lung cancer subtypes, and, finally, show the output prediction to the user.

IV. Final Remarks

The development of systems for optimisation of the lung cancer diagnosis is an area that is growing and many studies of methods to improve the current process are emerging. Furthermore, the use of deep learning is showing good results, as it is already validated in the case of the breast cancer where it has been implemented a similar process recently. However, as a limitation for lung cancer and since this is a recent area, there are few image datasets that allow training the algorithms with the desired performance to the point of being reliable its implementation in clinical use. In this paper, a possible solution was proposed as a high-level approach based on the available studies, with the aim of assisting the pathologist in the first morphological approach to the lesion in order to optimise the diagnostic process.

Acknowledgment

The authors would like to thank to the National Lung Screening Trial, The Cancer Genome Atlas and the Genomic Data Commons Data Portal for the lung cancer datasets availability and to FCT – *Fundação para a Ciência e Tecnologia* and FCT/MCTES in the scope of the project UIDB/05549/2020 for funding.

References

[1] N. Faria, S. Campelos, and V. Carvalho, "Cancer Detec - Lung Cancer Diagnosis Support System: first insights," in Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOINFORMATICS, 2022, pp. 83–90.

[2] B. C. Bade and C. S. Dela Cruz, "Lung Cancer 2020: Epidemiology, Etiology, and Prevention," Clinics in Chest Medicine, vol. 41, no. 1. W.B. Saunders, pp. 1–24, Mar. 01, 2020, doi: 10.1016/j.ccm.2019.10.001..

[3] N. Duma, R. Santana-Davila, and J. R. Molina, "Non–Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment," Mayo Clin. Proc., vol. 94, no. 8, pp. 1623–1640, Aug. 2019, doi: 10.1016/j.mayocp.2019.01.013.

[4] R. L. Keith, "Lung Carcinoma," 2020.

[5] A. El-Baz et al., "Computer-aided diagnosis systems for lung cancer: Challenges and methodologies," International Journal of Biomedical Imaging, vol. 2013. 2013, doi: 10.1155/2013/942353.

[6] C. Goebel, C. L. Louden, R. McKenna, O. Onugha, A. Wachtel, and T. Long, "Diagnosis of Non-small Cell Lung Cancer for Early Stage Asymptomatic Patients," Cancer Genomics and Proteomics, vol. 16, no. 4, pp. 229–244, 2019, doi: 10.21873/cgp.20128.

[7] World Health Organization, "Cancer," 2021. https://www.who.int/news-room/fact-sheets/detail/cancer.

[8] S. Wang et al., "Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome," Sci. Rep., vol. 8, no. 1, pp. 1–9, Dec. 2018, doi: 10.1038/s41598-018-27707-4.

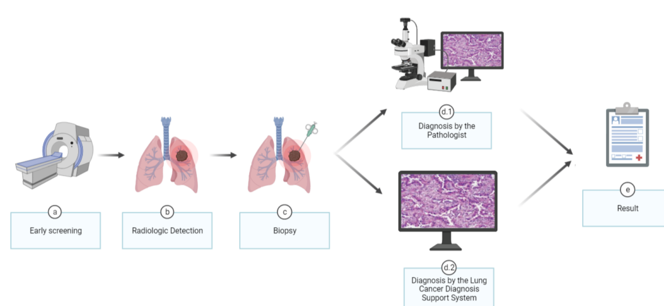

Fig. 1. Actual Approach for Lung Cancer Diagnosis Support System