Development of a Lung Cancer Diagnosis Support System

Nelson Faria and Vítor Carvalho
2Ai, School of Technology
Polytechnic Institute of Cávado and Ave
Barcelos, Portugal
e-mail: a14805@alunos.ipca.pt; vcarvalho@ipca.pt

Sofia Campelos
Institute of Molecular Pathology and Immunology
University of Porto
Porto, Portugal
e-mail: acampelos@ipatimup.pt

***Abstract*— Lung cancer is the leading type of cancer death worldwide, and a correct diagnosis in an early stage gives more possibilities for treatment. Whole Slide Images generated from glass slides can be analysed using Artificial Intelligence technologies to help pathologists. In this study, an overview of lung cancer is made, exploring the methodologies used to improve the histopathological diagnosis of lung cancer. These methods are composed of Detection and Classification phases. To detect the neoplastic cells, the Whole Slide Image is split into patches, and a convolutional neural network is applied to identify the tumour regions and generate a heatmap to highlight them. 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.**

Keywords- lung cancer; digital pathology; artificial intelligence; convolutional neural networks; whole slide images.

I. INTRODUCTION

On a global scale, lung cancer has been the type of neoplasia that causes the highest number of deaths and is the second most common in terms of new cases [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 the sooner it is detected, the greater 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. 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, which is the process of scanning microscopic glass slides to produce digital slides, pathologists are able 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 making the diagnosis, and improving the accuracy of the results. This article is composed of 4 sections: Section 2 begins with a brief overview of lung cancer's physiopathology before focusing on lung biopsy. Then, image processing techniques applied in pictures of biopsy tissue samples will be investigated, followed by an examination of the use of artificial intelligence in lung cancer detection. In Section 3, it is described the proposed approach to a lung cancer detection and classification system. Finally, Section 4 presents the main conclusions and the next steps.

ss

II. LITERATURE REVIEW

This section presents the etiology, classification, and methodologies for lung cancer detection. In addition, artificial intelligence techniques that aim to assist pathologists in the lung cancer diagnosis process are analysed.

A. Lung Cancer: Etiology, Classification and Detection Methodologies

According to data collected by the World Health Organization, the number of deaths from lung cancer reaches 1.8 million and 2.21 million new cases in 2020, which, when compared to other types of cancer, makes this neoplasia responsible for the highest number of deaths in the world [7]. The age group with the highest lung cancer rate is above the 50 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, and they are split into four types: Adenocarcinoma, Squamous Cell Carcinoma, Large Cell Carcinoma and Small Cell Lung Carcinoma. The most common is Adenocarcinoma which affects about 80 % 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

 ISBN: 978-1-61208-987-4

microscope, or through WSIs that were obtained by digitising the glass slides [1].

B. *Artificial Intelligence techniques applied to Lung Cancer*

The increased amount of data generated by clinical systems, as well as the computational capacity, enable the development of computational systems capable of facilitating the diagnosis process, improving the accuracy and moving faster to the final diagnosis.

According to literature [1][8][9], the analysis of lung cancer in WSIs can be split into two moments: Detection and Classification. Wang et al. [8] 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 [9] 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][9]. Another study from Li et al. [10] used 256 x 256 pixels 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 et al. [11], they also used the different CNN types where each Whole Slide Image (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][11]. To overcome the limitation that states the need for annotations by a pathologist to get a better result, Chen et al. [12] designed a technique that, instead of getting tiles from the WSI, the WSI is given as input without being split, which resulted in an accuracy of nearly 93% for lung cancer detection [1][12].

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 et al. [8] selected features associated with survival outcomes and used a univariate Cox proportional hazard model with a penalty to prevent overfitting, Yu et al. [11] employed Naïve Bayes classifiers, Support Vector Machines (SVMs) 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 APPROACH

Following the reviewed authors' approaches and as shown in Figure 2, the system to be developed must be capable of executing two phases: 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 contain 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 indicate 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. CONCLUSION

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 in medical systems, such as the recently implemented breast cancer diagnostic support system. However, since it is a developing area, there are limited amounts of image datasets available that allow training the algorithms with the desired performance to the point of being reliable 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 to optimise the diagnostic process. Further phases will include, among other things, a thorough examination of the AI approaches used and the creation of algorithms to assess the presence or absence of tumors in the photos being analysed.

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. This paper was funded by national funds (PIDDAC), through the FCT – *Fundação para a Ciência e Tecnologia* and FCT/MCTES under the scope of the projects UIDB/05549/2020 and UIDP/05549/2020.

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, Eti- ology, 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 - Pulmonary Disorders," *Merck Manuals Professional Edition*, 2020. [Online] [retrieved: June, 2022]. Available: https://www.merckmanuals.com/professional/pulmonary-disorders/tumors-of-the-lungs/lung-carcinoma#v923730

[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. [Online] [retrieved: June, 2022]. Available: 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.

[9] N. Coudray *et al.*, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nat. Med.*, vol. 24, no. 10, pp. 1559–1567, Oct. 2018, doi: 10.1038/s41591-018-0177-5.

[10] Z. Li *et al.*, "Computer-aided diagnosis of lung carcinoma using deep learning - a pilot study," Mar. 2018, [Online]. [retrieved: June, 2022]. Available: http://arxiv.org/abs/1803.05471.

[11] K. H. Yu *et al.*, "Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks," *J. Am. Med. Informatics Assoc.*, vol. 27, no. 5, pp. 757–769, May 2020, doi: 10.1093/jamia/ocz230.

[12] C. L. Chen *et al.*, "An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning," *Nat. Commun.*, vol. 12, no. 1, pp. 1–13, Dec. 2021, doi: 10.1038/s41467-021-21467-y.

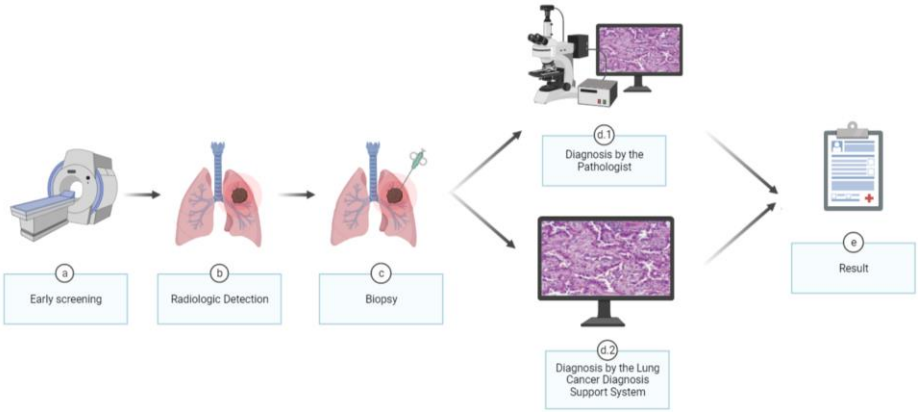

Figure 1. Actual Approach for Lung Cancer Diagnosis Support System.

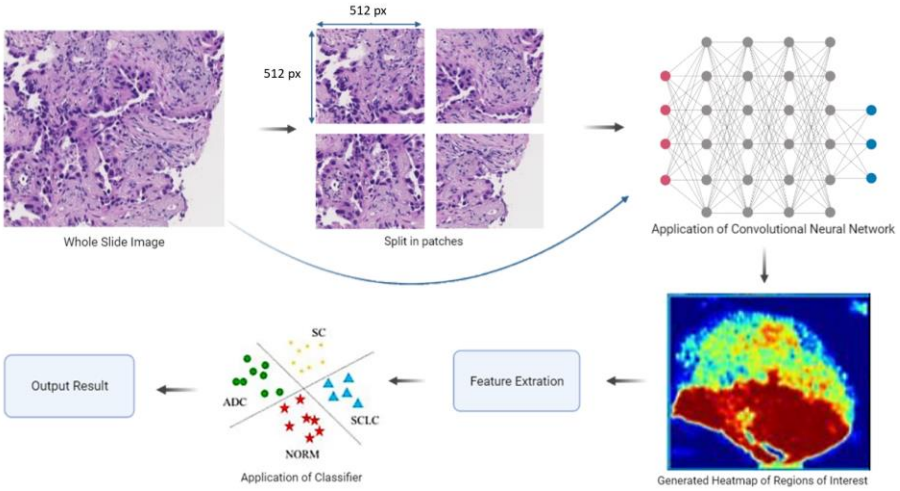

Figure 2. Diagram of the proposed approach to detect and classify the Lung Cancer according to a given WSI