FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Information Fusion-Based Model for Lung Nodule Characterization

João António Maricato Malva

PDIS

Master in Informatics and Computation Engineering

Supervisor: Eduardo Rodrigues

November 4, 2024

Information Fusion-Based Model for Lung Nodule Characterization

João António Maricato Malva

Master in Informatics and Computation Engineering

November 4, 2024

Contents

Chapter 1

Introduction

1.1 Context

Lung cancer is the leading cause of cancer-related deaths, often diagnosed in an advanced stage, resulting in a low 5-year survival rate of less than 10%, which occurs in 70% of cases, but if detected in an early stage, it is greater than 90% [3]. In 2022, lung cancer had the highest incidence and mortality rates of all cancers worldwide [4]. In particular, in upper-middle-income countries, there has been a significant increase in lung cancer-related deaths, with a rise of 442,000 deaths, more than 2.5 times the increase in deaths of the combined three other income groups [15].

Efforts to reduce lung cancer mortality by screening have been hampered by the aggressive and diverse nature of the disease [14], for example, CT scans help diagnose lung cancer more precisely and produce a reduction of 20% in mortality. Today, the classification of a pulmonary nodule is dependent on measuring the growth rate of that nodule from multiple CT scans and following it for approximately two years to avoid performing a biopsy, which is an invasive procedure for diagnosis. Another downside of slice-by-slice CT scans in lung cancer detection is that they are challenging for doctors, since the amount of data saved in this medical procedure is time-consuming, expensive, prone to reader bias, and requires a high degree of competency and concentration [16].

As medical data becomes more complex, there is a growing need for models that can effectively integrate and analyze these data to support clinical decision-making [7]. Computer-aided diagnosis (CAD) is increasingly being investigated as an alternative and complementary approach to conventional reading, as it avoids many of these issues. Automated nodule diagnostic systems can save both time and money while avoiding the risks of invasive surgical procedures. The non-invasive CAD system for lung nodule diagnosis is promising and has achieved very high accuracy measures from a single CT scan [16].

The combined gains in medical imaging and deep learning complement new approaches that are accurate and safer ways of recognizing diseases. Deep learning models can overcome projections that show how medical images have been analyzed to locate and determine the type of lung abnormalities that are a common cause of cancer.

1.2 Problem

This thesis addresses the need for a more accurate and reliable diagnostic tool. Existing diagnostic systems, largely based on deep learning models, have certain limitations when it comes to accuracy and generalization in regards towards medical image datasets. This sort of models are often based on deep features from neural networks, which can overshadow superficial features such as texture and shape that are important for accurately classifying nodules.

In addition, the lack of explainability of the model poses a challenge in clinical contexts, which can limit its reliability in making critical medical decisions. The inability to provide interpretable information hinders the adoption of these methods for the diagnosis of lung cancer.

1.3 Hypothesis

Feature extraction is critical for the characterization task, which involves both shallow features (texture and shape descriptors) and deep features learned by deep neural networks (DNNs). The state of the art demonstrates that performance through the application of information fusion techniques could be more efficient in deep learning models when applied in lung nodule characterization [24]. These advancements highlight the need for further research in deep learning and information fusion to prompt early detection and reduce mortality as well as to provide more effective treatment strategies for lung cancer.

We hypothesize that this information fusion-based model approach, with shallow and deep features, will result in a more accurate and reliable model for lung cancer characterization, making it better suited for early detection and precision diagnosis. We seek to overcome the current state-of-the-art limitations in automatic lung cancer diagnostics, offering a solution that not only improves prediction accuracy but also has the potential to assist in clinical decision-making and medical practice.

1.4 Motivation

Promoting the improvement of human life and health through early detection of diseases continues to be a concern throughout the world and is also the main objective of Goal 3 (Ensure healthy lives and promote well-being for all at all ages) of the Sustainable Development Goals (SDGs) of the United Nations [22]. Lung cancer is an enemy of public health care and the development of early and accurate diagnostic tools will help improve survival rates. This research aims to contribute to the goal of promoting health by harnessing modern technologies to address one of the greatest diagnostic issues in oncology today. Through the development of models that support more precise care, this dissertation aligns with the global imperative to promote well-being for all.

1.5 Research Questions

To bring clarity and precision to the hypothesis, we will break it down into three research questions. These questions will guide the investigation, helping us to understand the main points of our hypothesis.

1. Does fusing information from shallow and deep feature extractors bring any improvement (classification performance, generalization, reduction in the number of model parameters) compared to using only a deep feature extractor?

2. How does this approach compare with an approach that only uses a deep feature extractor when varying the dataset? (e.g. training on one set of data and testing on another, using different amounts of data, among others)

3. In what ways can information-fusion-based models contribute to improving the explainability of lung nodule malignancy predictions?

Chapter 2

Literature Review

The main goal of this literature review is to define a strategy for completing the state of research on information fusion-based models for lung nodule characterization, with a focus on identifying fusion techniques currently applied in this field. In particular, it has three main objectives: to gain knowledge by analyzing the different types of techniques that have been used in nodule characterization, to discover the specific methods that have been designed to automate nodule characterization and to evaluate the effectiveness of these methods based on the results obtained in the respective studies. However, in addition to achieving these objectives, the review will help to understand the current state of information fusion methods in this area and will lead to a better understanding of the various approaches.

Through recent studies analyses, we will study the most widely adopted techniques, as well as hybrid approaches that combine various strategies. In addition, methods used for automatic nodule characterization will also be analyzed, with a focus on deep learning architectures adapted to CT scan image analysis. This synthesis aims to establish a comprehensive basis for future studies in the characterization of pulmonary nodules and to guide the development of more effective, interpretable and applicable models in a clinical environment.

2.1 Eligibility Criteria

In order to achieve the relevance and rigor of the selected studies, the criteria for the systematic review were specified. These criteria seek to encompass the entire body of research that has been conducted between shallow and deep feature extractors and information fusion techniques in the characterization of CT scans, mainly related to lung nodules. To expand the scope, a more extensive view of applicable methodologies was included for medical conditions that use CT technologies if the studies presented relevant approaches.

In terms of eligibility, only studies published in the last five years (2019 - 2024) were considered, which ensures that the review reflects recent advances in fusion techniques in the field of medical imaging. Articles were limited to those published in English to maintain consistency

and accessibility. In addition to the characterization of pulmonary nodules, studies aimed at diagnosing other medical conditions revealed by CT scans were also included, as long as they used methodologies that could be used within the scope of this study.

2.2 Search Strategy

In this review, a comprehensive search strategy was formulated to find relevant studies in various reliable databases. The search was carried out mainly in three main academic databases: IEEE Xplore, PubMed and Google Scholar.

In addition to entering keywords, the search strategy was implemented to ensure methodical and specific return of studies. The keywords used in the search involved descriptions such as “lung nodule”, “information fusion” and “CT scan analysis”.

Furthermore, in addition to the database searches, reference chaining was used to expand the list of studies. Specifically, the references of key articles excavated in the initial phase were subjected to close scrutiny. As an example, significant references were extracted from the influential article Fusion of Textural and Visual Information for Medical Image Modality Retrieval Using Deep Learning-Based Feature Engineering [7], which has been helpful in the study of fusion techniques for the analysis of pulmonary nodules. Reviewing the references cited in this document helped to identify other studies relevant to the objectives of the initial work, linked by common topics.

2.3 Screening Process

After the search drew out a set of studies that might be relevant, the title and abstract screening process followed. During this phase, each study’s titles and abstracts were reviewed to weed out studies that are obviously not relevant. This first screening made it possible to discard the articles that didn’t fit the core inclusion criteria, including those that are not associated with CT-based medical imaging or those focusing on papers related to the use of shallow or deep features extractors or information fusion methods.

This strategy was used to cut out irrelevant studies in a fast and easy way as well as to let potentially related papers through to the full-text analysis phase. At this stage, the objective was to screen only titles and abstracts. This enabled the easy inclusion of studies that were worth further investigation in subsequent stages.

2.4 Summary

This literature review lays the foundations for the study and research into lung nodule characterization models based on information fusion. With the implementation of a search strategy, this review presents a current selection of studies focused on the characterization of lung nodules and including other diseases based on CT scans. The structured procedure of using the main databases and linking references ensures that relevant studies are obtained, and the selection is further boosted by the use of the title and abstract screening process, which shortens the selection to studies that are actually dealing with research in terms of deep learning, shallow feature extraction and information fusion techniques.

References

[1] Shubham Agrawal, Aastha Chowdhary, Saurabh Agarwala, Veena Mayya, and Sowmya Kamath S. Content-based medical image retrieval system for lung diseases using deep CNNs. *International Journal of Information Technology*, 14(7):3619–3627, December 2022.

[2] Danilo Avola, Luigi Cinque, Alessio Fagioli, Sebastiano Filetti, Giorgio Grani, and Emanuele Rodolà. Multimodal feature fusion and knowledge-driven learning via experts consult for thyroid nodule classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2527–2534, 2022.

[3] Kirsten M. Bahmer, Andrew J. Schoenfeld, and Denise R. Aberle. Benefits and harms of ct screening for lung cancer: A systematic review. *JAMA*, 307(22):2418–2429, 2011. Accessed: 2024-10-03. Cited on page 1.

[4] International Agency for Research on Cancer. Trachea, bronchus and lung cancer fact sheet. https://gco.iarc.who.int/media/globocan/factsheets/cancers/15-trachea-bronchus-and-lung-fact-sheet.pdf, 2024. Accessed: 2024-10-03. Cited on page 1.

[5] Yu Gu, Jingqian Chi, Jiaqi Liu, Lidong Yang, Baohua Zhang, Dahua Yu, Ying Zhao, and Xiaoqi Lu. A survey of computer-aided diagnosis of lung nodules from ct scans using deep learning. *Computers in Biology and Medicine*, 137:104806, 2021.

[6] Seung Hyup Hyun, Mi Sun Ahn, Young Wha Koh, and Su Jin Lee. A Machine-Learning approach using PET-Based radiomics to predict the histological subtypes of lung cancer. *Clin Nucl Med*, 44(12):956–960, December 2019.

[7] Saeed Iqbal, Adnan N. Qureshi, Musaed Alhussein, Imran Arshad Choudhry, Khursheed Aurangzeb, and Tariq M. Khan. Fusion of textural and visual information for medical image modality retrieval using deep learning-based feature engineering. *IEEE Access*, 11:93238–93253, 2023. Cited on pages 1 and 5.

[8] Jonathan A. Liu, Issac Y. Yang, and Emily B. Tsai. Artificial intelligence (ai) for lung nodules, from the ajr special series on ai applications. *American Journal of Roentgenology*, 219(5):703–712, 2022. PMID: 35544377.

[9] Xindong Liu, Mengnan Wang, and Rukhma Aftab. Study on the prediction method of long-term benign and malignant pulmonary lesions based on LSTM. *Front Bioeng Biotechnol*, 10:791424, March 2022.

[10] Sawsan Mahmoud, Hanan Al-Jubouri, and Tawfeeq Abdulabbas. Chest radiographs images retrieval using deep learning networks. *Bulletin of Electrical Engineering and Informatics*, 11:1358–1369, 06 2022.

[11] Anum Masood, Bin Sheng, Po Yang, Ping Li, Huating Li, Jinman Kim, and David Dagan Feng. Automated decision support system for lung cancer detection and classification via enhanced rfcn with multilayer fusion rpn. *IEEE Transactions on Industrial Informatics*, 16(12):7791–7801, 2020.

[12] Seyed Mohammad Alizadeh, Mohammad Sadegh Helfroush, and Henning Müller. A novel siamese deep hashing model for histopathology image retrieval. *Expert Systems with Applications*, 225:120169, 2023.

[13] Nasrullah Nasrullah, Jun Sang, Mohammad S Alam, Muhammad Mateen, Bin Cai, and Haibo Hu. Automated lung nodule detection and classification using deep learning combined with multiple strategies. *Sensors (Basel)*, 19(17), August 2019.

[14] National Lung Screening Trial Research Team, Denise R Aberle, Amanda M Adams, Christine D Berg, William C Black, Jonathan D Clapp, Richard M Fagerstrom, Ilana F Gareen, Constantine Gatsonis, Pamela M Marcus, and JoRean D Sicks. Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England journal of medicine*, 365(5):395—409, August 2011. Cited on page 1.

[15] World Health Organization. The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2024. Accessed: 2024-10-03. Cited on page 1.

[16] Ahmed Shaffie, Ahmed Soliman, Amr Eledkawy, Victor van Berkel, and Ayman El-Baz. Computer-assisted image processing system for early assessment of lung nodule malignancy. *Cancers*, 14(5), 2022. Cited on page 1.

[17] Ahmed Shaffie, Ahmed Soliman, Luay Fraiwan, Mohammed Ghazal, Fatma Taher, Neal Dunlap, Brian Wang, Victor van Berkel, Robert Keynton, Adel Elmaghraby, and Ayman El-Baz. A generalized deep Learning-Based diagnostic system for early diagnosis of various types of pulmonary nodules. *Technol Cancer Res Treat*, 17:1533033818798800, January 2018.

[18] Gurtaj Singh, Vincenzo Violi, and Marco Fisichella. Federated learning to safeguard patients data: A medical image retrieval case. *Big Data and Cognitive Computing*, 7(1), 2023.

[19] Md Fahimuzzman Sohan and Anas Basalamah. A systematic review on federated learning in medical image analysis. *IEEE Access*, 11:28628–28644, 2023.

[20] Joana Vale Sousa, Pedro Matos, Francisco Silva, Pedro Freitas, Hélder P Oliveira, and Tania Pereira. Single modality vs. multimodality: What works best for lung cancer screening? *Sensors (Basel)*, 23(12), June 2023.

[21] Shanmugapriya Survarachakan, Pravda Jith Ray Prasad, Rabia Naseem, Javier Pérez de Frutos, Rahul Prasanna Kumar, Thomas Langø, Faouzi Alaya Cheikh, Ole Jakob Elle, and Frank Lindseth. Deep learning for image-based liver analysis — a comprehensive review focusing on malignant lesions. *Artificial Intelligence in Medicine*, 130:102331, 2022.

[22] United Nations. The 17 Goals: Sustainable Development Goals, 2015. Accessed: 2024-10-06. Cited on page 2.

[23] Xiyue Wang, Yuexi Du, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Retccl: Clustering-guided contrastive learning for whole-slide image retrieval. *Medical Image Analysis*, 83:102645, 2023.

[24] Yutong Xie, Jianpeng Zhang, Yong Xia, Michael Fulham, and Yanning Zhang. Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest ct. *Information Fusion*, 42:102–110, 2018. Cited on page 2.

[25] Pengshuai Yang, Yupeng Zhai, Lin Li, Hairong Lv, Jigang Wang, Chengzhan Zhu, and Rui Jiang. A deep metric learning approach for histopathological image retrieval. *Methods*, 179:14–25, 2020. Interpretable machine learning in bioinformatics.

[26] Yue Zhao, Zhongyang Wang, Xinyao Liu, Qi Chen, Chuangang Li, Hongshuo Zhao, and Zhiqiong Wang. Pulmonary nodule detection based on multiscale feature fusion. *Comput Math Methods Med*, 2022:8903037, December 2022.