UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Physics “Giuseppe Occhialini”

Inter-University Master’s Degree Programme
Artificial Intelligence for Science and Technology

Lung Nodule Detection and Explainability in CT Scans

Andrea Borghesi

REGISTRATION NR 916202

SUPERVISOR Prof. Pietro Govoni

CO-SUPERVISOR Dr. Simone Gennai

Academic Year 2024/2025

Contents

Chapter 1

Introduction

1.1 Problem Statement and Motivation

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, with early detection being crucial for improving patient outcomes and survival rates. Computed Tomography (CT) screening programs, such as the National Lung Screening Trial (NLST), have demonstrated the potential to reduce lung cancer mortality through early identification of pulmonary nodules [1, 3]. However, the manual interpretation of CT scans is a time-intensive process that places significant burden on radiologists, while also being subject to inter-observer variability and potential oversight of small or subtle lesions.

Radiologists are often required to review hundreds of slices per patient, sometimes across dozens of patients in a single day. Under such conditions, cognitive fatigue can accumulate, potentially leading to decreased diagnostic precision, delayed reading times, or even missed findings, especially for low-contrast or small nodules that may appear on only a few slices [13, 14]. This challenge is further amplified in high-volume screening programs, where maintaining consistent accuracy over long hours is difficult even for experienced professionals.

Artificial intelligence (AI) based detection tools have the potential to alleviate this strain by automatically flagging suspicious regions of interest, enabling radiologists to focus attention on the most relevant slices. Such systems do not replace human expertise but can act as a second reader, improving sensitivity to subtle findings, reducing oversight caused by fatigue, and providing explainable visual cues that support decision-making and increase trust in automated recommendations [6]. On this note, it is important to mention that the use of AI in radiology is not always well-accepted, as it has been shown by Liu et al. [10] that radiologists with a higher workload and lower AI-acceptance are more likely to experience burnout. Regardless, the integration of AI tools into clinical workflows has been shown to enhance diagnostic accuracy and efficiency, ultimately leading to better patient care and outcomes [7, 8]

Nevertheless, existing AI solutions are often limited by computational demands and lack transparency in their predictions, motivating the need for efficient and explainable detection methods specifically tailored to lung nodule identification in CT scans. This need is further reinforced by the recent European Union Artificial Intelligence Act, which establishes a regulatory framework that emphasizes transparency, accountability, and human oversight for AI systems—particularly in high-risk domains such as healthcare [5].

1.2 Research Challenges

Designing AI-based tools for lung nodule detection in CT scans involves navigating a complex landscape of technical and practical challenges that significantly limit the applicability of existing solutions in real-world clinical environments.

The most immediate challenge stems from the computational demands of state-of-the-art approaches. While three-dimensional convolutional neural networks can theoretically exploit the full volumetric context of CT scans to improve detection accuracy, their practical implementation reveals significant limitations. These models typically require substantial GPU memory, often exceeding 16GB, and demand extensive training times that can extend to days or weeks, even on high-end hardware such as A100 GPUs with 40GB of memory [16]. While inference times for individual patient scans may be manageable, the development and iterative improvement of such models becomes prohibitively expensive and time-consuming. This computational burden became particularly acute in our research setting where access to high-end hardware was limited and unreliable. Working with a T4 GPU (16GB), the memory constraints made 3D approaches largely infeasible for our experiments. While we had occasional access to an A100 through shared research infrastructure, the instability of access, environment limitations, and restricted interaction modes (Jupyter-only) severely hindered our exploratory research process. A few preliminary experiments to understand model behavior and optimal hyperparameters consumed weeks of our available computational time, making the iterative development essential for novel research practically impossible. Such computational requirements create a fundamental mismatch with research development processes, where rapid experimentation and iteration are crucial for advancing the field. This computational constraint naturally leads to consideration of more efficient 2D approaches, which process individual CT slices rather than entire volumes. This approach significantly reduces memory requirements and training times, allowing for more agile development cycles, at the cost of losing correlations between slices that may be crucial for accurate nodule detection. Regardless, a 2D approach allows for better accessbility to healthcare professionals, as it can be run on standard laptops or workstations without the need for high-end GPUs.

Equally critical, yet often overlooked in the technical literature, is the challenge of explainability in object detection systems. The medical domain presents unique requirements for AI interpretability, driven by regulatory demands, clinical decision-making needs, and the fundamental requirement for physician trust and acceptance. However, most existing explainability methods were developed for image classification tasks [12, 2, 4, 9], where the goal is to explain a single prediction score for an entire image. Object detection fundamentally differs in that it produces structured outputs consisting of multiple bounding boxes, each with associated class probabilities and confidence scores. This structural difference creates significant technical obstacles for applying standard explainability techniques.

1.3 Research Questions and Objectives

The challenges outlined in the previous section give rise to three primary research questions that guide this thesis:

RQ1: How can 2D object detection approaches achieve clinically relevant performance for lung nodule detection within computational constraints? This question addresses the fundamental trade-off between computational efficiency and detection performance. While 3D approaches theoretically offer superior performance by exploiting volumetric context, their computational demands make them impractical for many research and deployment scenarios. We investigate whether 2D slice-based detection can achieve acceptable performance levels while operating within the memory and processing constraints of standard hardware configurations.

RQ2: How can explainability techniques be adapted for the structured outputs of object detection models? Standard explainability techniques are designed for classification tasks with single prediction scores, but object detection produces complex structured outputs with multiple bounding boxes, confidence scores, and non-differentiable post-processing steps. This question explores how gradient-free Class Activation Map (CAM) methods can be modified to generate meaningful explanations for object detection predictions, addressing the technical challenges posed by structured outputs and non-differentiable operations.

To address these research questions, this thesis pursues the following specific objectives:

O1: Develop a computationally efficient 2D object detection pipeline for lung nodule detection Implement and compare multiple object detection architectures and their variants. Design preprocessing pipelines optimized for 2D slice-based analysis, including slice selection algorithms to maximize information content Achieve detection performance that demonstrates clinical viability while operating within $<$6GB GPU memory constraints

O2: Adapt CAM methods for object detection tasks Research and adapt CAM methods that can handle structured object detection outputs Implement comparison algorithms that can assess similarity between structured predictions for explainability evaluation Implement end-to-end explainable object detection pipeline that provides both predictions and corresponding explanation maps

O3: Demonstrate integration of detection and classification with explainability Extend the object detection pipeline with a classification head for detected regions Apply explainability methods to both detection and classification components Evaluate the clinical utility of multi-stage explainable predictions

O4: Provide comprehensive experimental validation Evaluate the complete system on the established medical imaging datasets NLST [11], DLCSD24 [15]. Compare performance against relevant baselines and alternative approaches Analyze computational efficiency and practical deployment considerations

1.4 Thesis Structure

Chapter 2

Object Detection

2.1 Fundamentals of Object Detection

What's object detection? Describe the task itself, how it differs from classification and other computer vision tasks. What are the usual approaches in the deep learning context (one-stage vs two-stage, anchor-based vs anchor-free, etc.)? What are the most common architectures and their characteristics? highlight problems related to their structured outputs and the inherited challenges for gradient backpropagation

2.1.1 RetinaNet

a one-stage anchor-based architecture, essentially a toned-down Faster R-CNN, this is why i'd describe it first Describe the its architecture and components, address that it is born to address the class imbalance problem with the Focal Loss introduction, and its intended for medical imaging applications

2.1.2 Faster R-CNN

two-stage anchor-based architecture, the most common one and the one that is used in most of the literature. Describe its components, RPN, RoI pooling etc., it is the culmination of the R-CNN family of architecutres.

2.1.3 YOLOv8

one-stage anchor-free architecture (from v8 onwards), the most common one-stage architecture in literature, mostly for its inference speeds. its objectives were to reach real-time inference speeds while maintaining an acceptable accuracy.

2.1.4 DETR

a one-stage anchor-free architecture, the first to introduce the transformer architecture in object detection, it is a fully end-to-end architecture that does not rely on anchors or region proposals, but rather directly predicts the bounding boxes and class labels in a single pass. Unfortunately, it is not suitable for our use case due to its large needs of data.

2.1.5 Evaluation Metrics: Average Precision and Average Recall

Explain the need to formally define evaluation metrics as COCO's definitions are shaky, seems like everyone's using them but no one ever explains them.

2.2 Explainability in Deep Learning

Quick introduction of explainability in deep learning, highlighting how models are essentially black boxes and the need to understand, although partially, their inner workings. This need is even more pronounced in the medical field, where explainability is a requirement for clinical acceptance and regulatory compliance. Highlight how explainability methods, especially the ones that we are going to cover, are desinged to provide *insights* into the model's decision-making process, it does not make it fully explainable, but rather makes it a gray-box model, which is still better than a black box. Trainsition to CAMs for computer vision tasks

2.2.1 Class Activation Maps (CAM)

What are CAMs, how they work, and their (usual) limitations. Describe GradCAM as a gradient-based method and then explain why it is problematic for object detection tasks due to their sensitivity to the structured outputs and the non-differentiable post-processing steps.

2.2.2 Adaptations for Object Detection

Discuss gradient-free CAM methods ScoreCAM, SS-CAM and EigenCAM, highlighting how they can be adapted to work with object detection outputs. there's a bit of math involved here as we got to pinpoint what fails in their original implementations and how we can fix it.

Chapter 3

Data and Preprocessing

3.1 Datasets: NLST and DLCSD24

Briefly describe the datasets and their characteristics. mention that one does not have benign annotations, while the other does, nonetheless the first one can be yet used to train a detection model (as long as it detects nodules, regardless of their malignancy). regardless, we might also use the first one as a pretrain for the second one.

3.2 Preprocessing Pipeline

Resampling, and HU clipping, explaining why we're clipping and why those ranges and not others, this is supported by the HU table.

3.3 Slice Selection Algorithms

Explain the need for slice selection algorithms in this 2D setting, as some might not contain relevan information (nodule mass). explain the statistical approach, the sliding window approach and finally the simple thresholding approach.

3.4 Data Augmentation Techniques

Discuss the data augmentation techniques used to enhance model robustness, such as random rotations, flips, and intensity variations. Explain how these techniques help mitigate overfitting and improve generalization to unseen data.

Chapter 4

Methodology

4.1 Object Detection Architectures and Backbones

Provide an overview of the object detection used (they're already described in the background chapter), and mention how we also experiment with different backbones.

4.2 Object Detection Pipeline

Describe the overall object detection pipeline, including the preprocessing steps, the training process, and the evaluation metrics used. I might include here curriculum learning

4.3 Classification Head Extension

Discuss the classification network, how we extracted a classification dataset from the detection dataset using the bounding boxes.

4.4 Explainability Methods

Describe how are we using the adapted CAM methods, which layers we're targeting and an example of expected output. Explain how we're going to compare the explainability methods using the segmentation game.

Chapter 5

Results

5.1 Object Detection Performance Analysis

Present the results of the object detectors, making sure that each setup is clearly defined: architecture, backbone, training hyperparameters, eventual pretraining, etc. For each setup we present the performance using AP and AR metrics, their inference times, and the GPU memory usage.

5.2 Explainability Evaluation

Discuss the performance of the explainability methods, comparing them based on the segmentation game, performing some statistical analysis on the results.

Chapter 6

Conclusions

Summary of the work, and discuss how the research question and objectives were addressed in this work.

6.1 Future Directions

Potential future directions, such as exploring lightweight attention based methods, segmentation-based approaches, and comparing the adapted CAM methods with other explainability techniques ad-hoc designed for object detection tasks.

Bibliography

[1] Denise R. Aberle, Alden M. Adams, Christine D. Berg, William C. Black, Jonathan D. Clapp, Roger M. Fagerstrom, Ilana F. Gareen, Constantine Gatsonis, Pamela M. Marcus, and Jennifer D. Sicks. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

[2] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018.

[3] Harry J de Koning, Carlijn M van Der Aalst, Pim A de Jong, Ernst T Scholten, Kristiaan Nackaerts, Marjolein A Heuvelmans, Jan-Willem J Lammers, Carla Weenink, Uraujh Yousaf-Khan, Nanda Horeweg, et al. Reduced lung-cancer mortality with volume ct screening in a randomized trial. *New England journal of medicine*, 382(6):503–513, 2020.

[4] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks, 2021.

[5] European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence. Official Journal of the European Union (OJ L 2024/1689), adopted 13 June 2024, published 12 July 2024, entered into force 1 August 2024, 2024. High-risk AI systems in healthcare are regulated under this act.

[6] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020.

[7] Ali Guermazi, Chadi Tannoury, Andrew J Kompel, Akira M Murakami, Alexis Ducarouge, André Gillibert, Xinning Li, Antoine Tournier, Youmna Lahoud, Mohamed Jarraya, et al. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology*, 302(3):627–636, 2022.

[8] Elizabeth Huynh, Ahmed Hosny, Christian Guthier, Danielle S Bitterman, Steven F Petit, Daphne A Haas-Kogan, Benjamin Kann, Hugo JWL Aerts, and Raymond H Mak. Artificial intelligence in radiation oncology. *Nature Reviews Clinical Oncology*, 17(12):771–781, 2020.

[9] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[10] Hui Liu, Ning Ding, Xinying Li, Yunli Chen, Hao Sun, Yuanyuan Huang, Chen Liu, Pengpeng Ye, Zhengyu Jin, Heling Bao, et al. Artificial intelligence and radiologist burnout. *JAMA network open*, 7(11):e2448714–e2448714, 2024.

[11] National Cancer Institute. National lung screening trial (nlst) - low-dose ct imaging dataset. `https://wiki.cancerimagingarchive.net/display/NLST`, 2002. Data collection: 2002–2009. Accessed via The Cancer Imaging Archive (TCIA), subset of NLST. Available under CC BY 4.0 License.

[12] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019.

[13] Nadia Stec, Danielle Arje, Alan R Moody, Elizabeth A Krupinski, and Pascal N Tyrrell. A systematic review of fatigue in radiology: is it a problem? *American Journal of Roentgenology*, 210(4):799–806, 2018.

[14] Sian Taylor-Phillips and Chris Stinton. Fatigue in radiology: a fertile area for future research. *The British journal of radiology*, 92(1099):20190043, 2019.

[15] Fakrul Islam Tushar, Avivah Wang, Lavsen Dahal, Michael R. Harowicz, Kyle J. Lafata, Tina D. Tailor, and Joseph Y. Lo. Ai in lung health: Benchmarking detection and diagnostic models across multiple ct scan datasets, 2025.

[16] Hui Wu, Matrix Yao, Albert Hu, Gaofeng Sun, Xiaokun Yu, and Jian Tang. A systematic analysis for state-of-the-art 3d lung nodule proposals generation, 2018.

Acknowledgements