Laboratory of Artificial Intelligence and Data Science

Project 1 - Lung Cancer Classification using Computerized Tomography (CT) Data

2025 / 2026

1 Introduction

The objective of the Laboratory of Artificial Intelligence and Data Science (Lab AI & DC) course is to provide students with software development methodologies, AI and DC projects, teamwork and communication through the implementation of projects designed for this purpose. Students should apply the knowledge obtained from previous years courses and research methodologies to solve the problem.

In this first project, students will use images as input data, namely Computerized Tomography (CT) data, from the human torso, for Lung Cancer classification.

2 Context

Lung cancer is at the top of cancer-related mortality numbers worldwide [Wor18, Ame19]. Only 16% of the cases of lung cancer are diagnosed as tumors in the local stage. In these cases, patients have a five-year survival rate greater than 50%; however, when diagnosed in an advanced stage, the chances of a five-year survival decrease to 5%. Thus, achieving an earlier diagnosis is critical to increase the survival rate, and systems capable of providing screening support could play an important role.

As a noninvasive method, computed tomography (CT) images have shown the ability to provide valuable information on tumor status, raising opportunities for the development of computer-aided diagnoses (CAD) systems able to provide an automatic assessment of lung nodules malignancy risk to help the clinical decision. Considering the use of qualitative data, factors such as the high inter-observer variability associated with visual evaluation of relevant characteristics, and the amount of radiological data to be analyzed make the development of completely automatic systems a more attractive approach [AFMH+22]. Radiomics is an emerging field that studies the extraction of mineable data from routinely acquired medical images. The general goal of the Radiomics field is to study the features from the medical images, to improve the healthcare given to patients by creating non-invasive diagnostic tools for early cancer detection.

2.1 Related Works

Several previous works proposed learning-based solutions for lung nodule malignancy classification using the Lung Image Database Consortium image collection (LIDC-IDRI) [AMB+15, AMea11], which is a public dataset of thoracic CT scans with expert annotations, and the most commonly used to develop AI-based solutions for lung cancer. Torres at al. [TDM+23] developed malignancy prediction models in lung cancer using several strategies for fusing multi-channel pyradiomics images. Shen et al. [SZY+15] proposed a hierarchical learning framework to capture the nodule heterogeneity by utilizing a Convolutional neural network (CNN) to extract features and a random forest classifier for the final classification with the highest accuracy of 0.868. Lu et al. [LLZ18] obtained an accuracy of 0.919 using a CNN to extract the features and a support vector machine (SVM) for the final classification. Yutong et al. [XZX+18] developed an algorithm that uses a deep convolutional neural network to automatically learn the feature representation of nodules on a 2D analysis, fuses this information with other more common features (shape, texture), and obtained an AUC of 0.966. A similar approach was developed by Causey et al. [CZM+18] that combines the deep learning CNN features with radiomics features as input in a random forest classifier and obtained an accuracy of 0.990. (These values are to be taken as a reference. It is not expected that you exceed them.)

3 Dataset: LIDC - IDRI

The LIDC-IDRI [AMB$^+$15, AMea11] is a lung cancer screening dataset which comprises thoracic CT scans for a total of 1010 patients, alongside annotated nodules belonging to one of three classes: a) nodule $\geq$ 3 mm; b) nodule $<$ 3 mm or c) non-nodule $\geq$ 3 mm, made during a two-phase annotation process by four experienced radiologists. Regarding data acquisition, slice thickness ranged from 0.6 to 5.0 mm, with X-ray current from 40 to 627 mA (mean: 222.1 mA) at 120-140 kVp.

The Dataset can be downloaded from this website [1]. Besides CT images from the human torso, the dataset includes annotations for the malignancy, position of the nodule/non-nodule and patient clinical information. (See Figure 1).

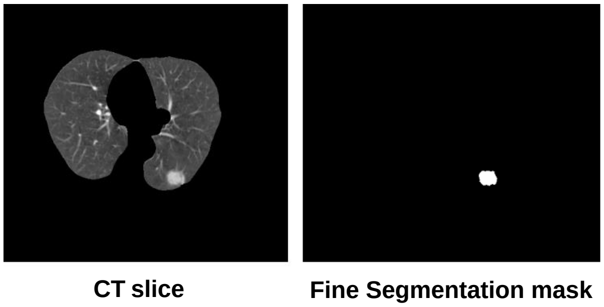

Figure 1: Image of the CT image of a patient and its respective fine segmentation mask from the lidc-idri dataset.

Seven academic centres and eight medical imaging companies collaborated to create this data set. Each subject includes images from a clinical thoracic CT scan and an associated XML file that records the results of a two-phase image annotation process performed by four experienced thoracic radiologists. In the initial blinded-read phase, each radiologist independently reviewed each CT scan and marked lesions belonging to one of three categories ("nodule $>= 3mm$", "nodule $< 3mm$", and "non-nodule $>= 3mm$"). In the subsequent unblinded-read phase, each radiologist independently reviewed their marks along with the anonymized marks of the three other radiologists to render a final opinion. The goal of this process was to identify as completely as possible all lung nodules in each CT scan without requiring forced consensus.

This dataset contains a standardized DICOM representation of the annotations and characterizations collected by the LIDC/IDRI initiative, originally stored in XML. Only the nodules that were deemed to be greater or equal to 3 mm in the largest planar dimensions have been annotated and characterized by expert radiologists performing the annotations. Only those nodules are included in the present dataset.

The conversion was enabled by the `pylidc` library [2] (parsing of XML, volumetric reconstruction of the nodule annotations, clustering of the annotations belonging to the same nodule, calculation of the volume, surface area and largest diameter of the nodules) and the `dcmqi` library [3] (storing of the annotations into DICOM Segmentation objects, and storing of the characterizations and measurements into DICOM Structured Reporting objects). The script used for the conversion is available at [4].

4 Work to develop

- You should prepare a Data Science-based solution to solve the problem proposed: Lung Cancer Classification using Computerized Tomography (CT) Data.
- You should share your solution in `gitlab.up.pt` (share the repository with the Professors[5] by the second practical class);

[1] https://www.cancerimagingarchive.net/collection/lidc-idri/
[2] https://pylidc.github.io/
[3] https://github.com/qiicr/dcmqi
[4] https://github.com/qiicr/lidc2dicom
[5] up384465 and up552793

- The solution should be delivered in Moodle by **October 19, 2025, at 23:59:59**;
- Students must present their work during the practical class on the 21st and 24th of October, 2025. If necessary, in special cases, further examination may be requested.

4.1 Submission of the solution

- Final code solution, as a notebook;
  - you should document your notebook, explaining your decisions and discussion about the results obtained;
- Link to a video summary. This is a team video, but each member should participate in it. This is a very short and to-the-point video (maximum of 5 minutes), summarizing the following:
  - the problem;
  - your solution;
  - the results and the impact you think this has.
- One-page document, including possible ethical and legal implications and the framework for current and future regulation issues.
- Auto-evaluation file provided by Professors.

4.2 Guidelines for the solution

- Assessed data quality and the need for data cleaning. If necessary perform cleaning of the data relevant to the model;
- Perform data pre-processing steps (e.g. range of values (Hounsfield unit [6], 2D vs 3D solution);
- Performed EDA (Exploratory Data Analysis);
- Performs Feature Engineering (e.g. Radiomics [7], Deep Features) and Selection;
- Discusses model/algorithm and technique selection, as well as model/algorithm optimization;
- Chooses performance metrics and performs validation;
- Explores model interpretability and fairness;
- Performs visualization of results;
- Why not consider other datasets to improve the generalization of the model?;
- Shows good programming skills (best practices, code commenting, performance, speed).

Some inspiration can be found in the work of Lee et. al. [GHSH18]. You can use a CRISP-DM-based methodology or other to develop your solution [8].

[6] https://radiopaedia.org/articles/hounsfield-unit/

[7] https://pyradiomics.readthedocs.io/en/latest/

[8] https://www.datascience-pm.com/crisp-dm-2/

4.3 Evaluation Criteria

Your work will be evaluated on the following criteria:

- 15% Product: understanding the needs of the end-user and if your proposal solves that problem;
- 20% Business: understanding if the solution serves the business purpose, its applicability and impact;
- 40% Technical Skills: overall technical evaluation of the solution from a data science point-of-view;
- 15% Soft-Skills: essentially your communication skills;
- 10% Ethical and Legal Considerations: understand if you understand it for this specific area of application.

4.4 Some Tips

Be creative in your solution! Think of how you can use certain approaches in an unusual way for example.

- Consider business constraints: understand the challenge well and identify any business constraints regarding this challenge;
- Mention the constraints you are considering for the solution in the notebook;
- Work as a team: The time is very short, our suggestion is that you distribute tasks well amongst the team;

References

[AFMH+22] Jose Arimateia Batista Araujo-Filho, Maria Mayoral, Natally Horvat, Fernando Santini, Peter Gibbs, and Michelle S Ginsberg. Radiogenomics in personalized management of lung cancer patients: Where are we? *Clinical Imaging*, 2022.

[AMB+15] Samuel G. Armato III, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, and L. P. Clarke. Data From LIDC-IDRI, 2015.

[Ame19] American Cancer Society. Facts & Figures 2019. Technical report, 2019.

[AMea11] Samuel G. Armato, Geoffrey McLennan, and et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 2011.

[CZM+18] Jason L. Causey, Junyu Zhang, Shiqian Ma, Bo Jiang, Jake A. Qualls, David G. Politte, Fred Prior, Shuzhong Zhang, and Xiuzhen Huang. Highly accurate model for prediction of lung nodule malignancy with CT scans. *Scientific Reports*, 2018.

[GHSH18] Lee G, Park H, Bak SH, and Lee HY. Radiomics in Lung Cancer from Basic to Advanced: Current Status and Future Directions. *Korean J Radiol*, 2018.

[LLZ18] Lu Liu, Yapeng Liu, and Hongyuan Zhao. Benign and malignant solitary pulmonary nodules classification based on CNN and SVM. In *ACM International Conference Proceeding Series*, 2018.

[SZY+15] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. Multi-scale convolutional neural networks for lung nodule classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.

[TDM+23] Guillermo Torres, Jan Rodriguez Dueñas, Sonia Baeza Mena, Antoni Rosell Gratacós, Carles Sanchez, and Debora Gil. Prediction of malignancy in lung cancer using several strategies for the fusion of multi-channel pyradiomics images. In *2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 236–240, 2023.

[Wor18] World Health Organisation. Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. *Int. Agency Res. Cancer*, 2018.

[XZX+18] 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*, 2018.