Machine Learning

CHALLENGE PRESENTATION

Guillermo Torres, Carles Sanchez, Debora Gil

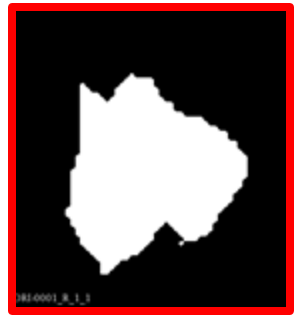
Objectives

Objectives

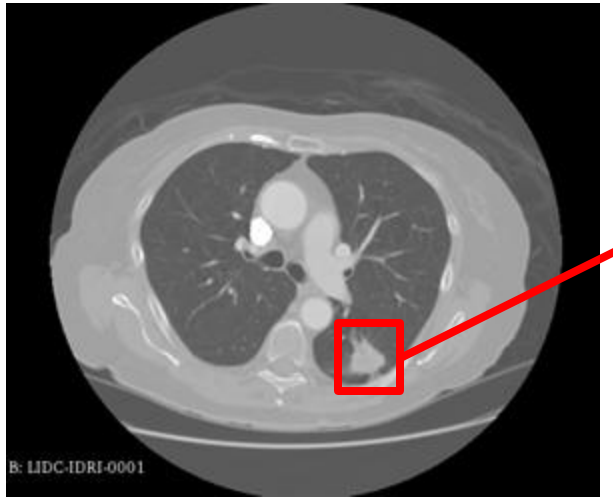
- Develop a machine learning system to characterise the pathology of pulmonary masses (lesions) in CT scans:
  - Segmentation
  - Classification
- Obtain a comprehensive understanding of supervised and unsupervised approaches to machine learning

Sub-Challenges

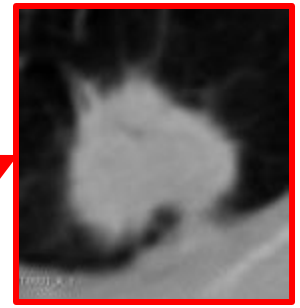
- Analysis of Unsupervised Techniques for Lesion Segmentation:
  - Implementation of a basic pipeline
  - Comparison between different methods
  - Influence of pre and post-processing steps
  - Qualitative and quantitative test analysis



Segmentation



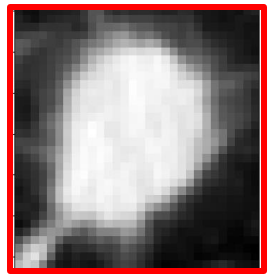
CT



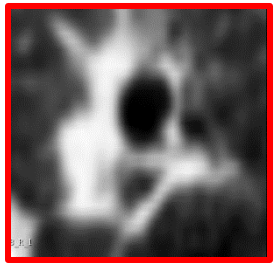
Nodule

Parts of the Challenge

- Analysis of Supervised Techniques for Lesion Classification:
  - Selection of optimal hyper-parameters of supervised techniques
  - Comparison between different feature spaces (radiomic texture descriptors, pre-trained networks)
  - Detection of bias in Models
  - Qualitative and quantitative test analysis



Benign



Malignant

Evaluation Activities

Activities

Milestones. Deliverables (code and report) of partial results during the course for challenge follow-up. Delivered through the Virtual Campus (VC).

Final Report. Code and report including minimums topics of the milestones (highlighted in yellow). Delivered through the Virtual Campus (VC).

Interdisciplinary working groups of 4-5 students.

Challenge2 Mark

The final mark of this part of the subject will be calculated the following way:

Challeng2 Mark = 0.25 * Milestones + 0.75 * Final Report

(Challenge2 Mark >=5 to compute average for Final Mark)

The deadline for submitting the Final Report (including code) is 1-June-2025.

Please note that Milestones should be delivered in intermediate dates (next slide).

Calendar of Milestones Deliverables

MILESTONE	DUE DATE
1. Segmentation	7 May
2. Classification baseline method	16 May
3. Classification DL method	23 May
4. Comparison of Methods & Final Report	1st June

Milestones

Milestones

Milestone 1 (Segmentation): Generation of an Annotated Dataset.

1. Extract VOI (Volume of Interest) from the CTs (intensity and mask).
2. Produce a single annotation for each lesion from the 4 radiologists' annotations using Max-Voting.
3. Make Max-Voting to obtain the "Diagnosis":
   if two or more radiologists have characterized the nodule with a Malignancy score > 3, then Diagnosis=1 (malignant), otherwise Diagnosis=0 (benign).

Deadline: 7 May .
Deliverable: A zip file containing the code and the VOIs and annotations per lesion of the data sample provided.

Milestones

Milestone 1 (Segmentation): Nodule Segmentation. Apply unsupervised techniques to obtain a segmentation of lesions in VOIs:

1. Use a classic standard pipeline over intensity volumes.
2. Use Otsu threholding and different morphological operations.
3. Quantify the performance using fair segmentation metrics.
4. Use kmeans over classic filter banks.
5. Compare between different unsupervised methods.

Deadline: 7 May.

Deliverable: A zip file containing report and code.
Material provided after this deadline: None.

Milestones

Milestone 2 (Classification): Data Exploration

1. Use different unsupervised techniques (eg. hierarchical clustering) and statistical tests to get correlations across radiological descriptions and also detect those annotations more relevant to the diagnosis.

Deadline: 16 May.
Deliverable: A zip file containing code and report for explaining the correlations.
Material provided after this deadline: None.

Milestones

Milestone 2 (Classification): Extraction of Radiomic Features.

1. Extract GLCM texture features using the PyRadiomics library.

Deadline: 16 May.
Deliverable: A zip file containing code for the extraction of GLCM features.
Material provided after this deadline: GLCMs in numpy format (.npz).

Milestones

Milestone 3 (Classification): Feature Extraction using a Pre-trained Convolutional Network.

1. Load a predefined VGG model and modify it to extract the features of the 1st fully connected (FC) layer.

2. Apply techniques for reduction of dimensionality.

Deadline: 23 May.
Deliverable: A zip file containing code for the extraction and selection of VGG features.
Material provided after this deadline: extracted features of the VGG in numpy format (.npz) and indexes of the selected features.

Milestones

Milestone 3 (Classification): Experimental Design and Data Splitting. Use a classifier (like SVM) with default parameters and GLCM features and implement a validation using:

1. K-folds by slice (StratifiedKFold).

2. K-folds grouping by nodule (StratifiedGroupKFold).

Deadline: 23 May.
Deliverable: A zip file containing code for the experimental designs
Material provided after this deadline: code using k-folds by slice.

Milestones

Milestone 3 (Classification): Hyper-parameters Optimization

1. Use a brute force grid search
2. Use a random search using sklearn.
3. Use Optuna.

Deadline: 23 May.
Deliverable: A zip file containing code for the optimization of hyper-parameters
Material provided after this deadline: None

Milestones

Milestone 4 (final report): Compare the performance of the different approaches.

1. Comparison between different feature spaces (radiomic texture descriptors, pre-trained networks)

Deadline: 1 June.
Deliverable: A zip file containing report and code for the Final Report.
Material provided after this deadline: None

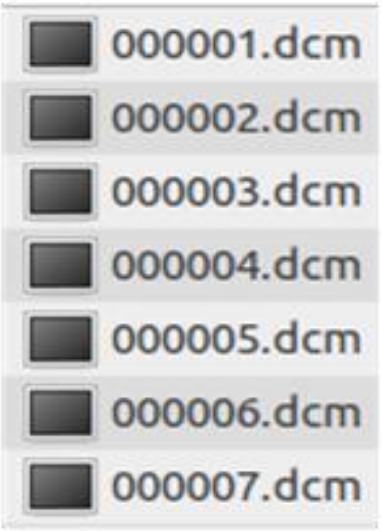
Dataset

LIDC-IDRI Dataset

LIDC-IDRI consists of diagnostic and lung cancer screening thoracic CT scans with marked-up annotated lesions.

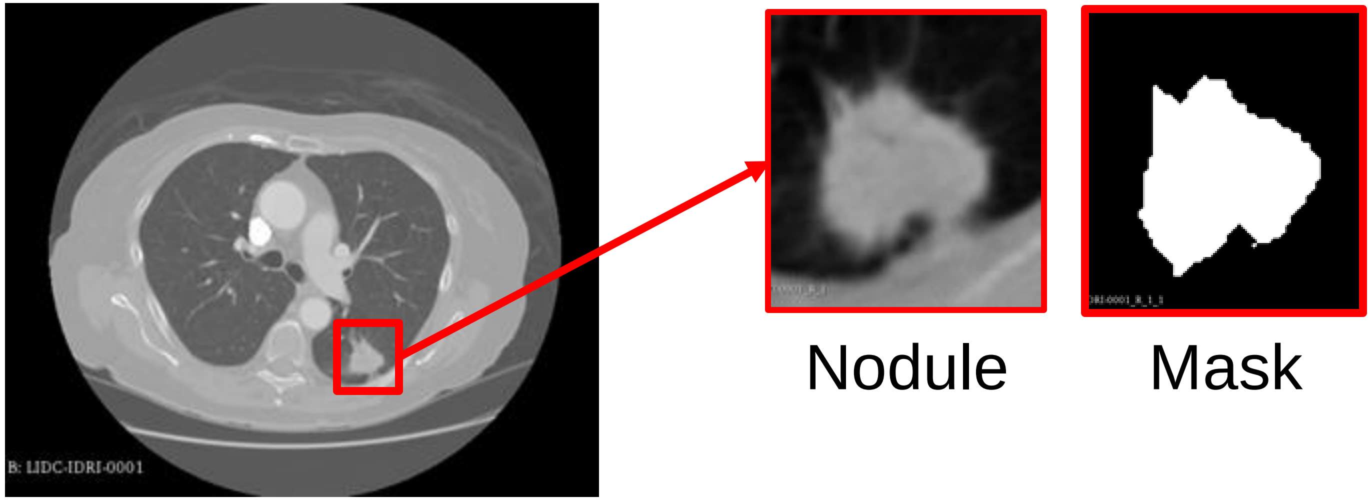
```
<edgeMap>
  <xCoord>363</xCoord>
  <yCoord>338</yCoord>
</edgeMap>
....
<edgeMap>
  <xCoord>362</xCoord>
  <yCoord>339</yCoord>
</edgeMap>
```

XML file for the CT scans



DICOM files for the CT scans

LUNA16 Dataset Description



CT

Nodule

Mask

CT Acquisition Parameters

WindowCenter	WindowWidth	RescaleSlope	RescaleIntercept
-600	1600	1	-1024
-600	1600	1	-1024
40	400	1	-1024
-600	1600	1	-1024

Radiologist's Annotations

patient_id	nodule_id	seriesuid	coordX	coordY	coordZ	diameter_mm
LIDC-IDRI-0001	1	1.3.6.1.4.1.14519.5.2.1.6279.6001.179049373636438705059720603192	56.20840547	86.34341278	-115.8675792	23.35064438
LIDC-IDRI-0001	1	1.3.6.1.4.1.14519.5.2.1.6279.6001.179049373636438705059720603192	56.20840547	86.34341278	-115.8675792	23.35064438
LIDC-IDRI-0001	1	1.3.6.1.4.1.14519.5.2.1.6279.6001.179049373636438705059720603192	56.20840547	86.34341278	-115.8675792	23.35064438
LIDC-IDRI-0001	1	1.3.6.1.4.1.14519.5.2.1.6279.6001.179049373636438705059720603192	56.20840547	86.34341278	-115.8675792	23.35064438

Malignancy	Malignancy_value	anotation_id	Calcification	Calcification_value	InternalStructure	InternalStructure_value
Highly Suspicious	5	84	Absent	6	Soft Tissue	1
Highly Suspicious	5	85	Absent	6	Soft Tissue	1
Highly Suspicious	5	86	Absent	6	Soft Tissue	1
Moderately Suspicious	4	87	Absent	6	Soft Tissue	1

Python Libraries

Python libraries

Virtual Environment

- SimpleITK
- Numpy
- Pandas
- Sklearn

Visualization

- 3D-Slicer
- VolumeCutBrowser

Code provided by instructors