Practical exercise:

Lung tissue classification using Machine and Deep Learning

Daniel Capellán Martín (daniel.capellan@upm.es)
Benito Farina (benito.farina@upm.es)
David Bermejo Peláez (david.bermejo@upm.es)
María Jesús Ledesma Carbayo (mj.ledesma@upm.es)

Biomedical Image Technologies Lab (BIT)
Dpto. Ingeniería Electrónica
Universidad Politécnica de Madrid

Aim

- The goal of this exercise is to familiarize with solving a classification task using both Machine and Deep Learning (ML, DL) approaches and their application in medical image classification.

How To

- The exercises will be developed in Python using **Google Colab**, a free cloud service based on Jupyter Notebooks that provides both CPU and GPU virtual environments for free.
- Access to the Jupyter Notebook. The entire assignment will be developed in Google Colab:

  https://colab.research.google.com/drive/1ST8zKdi7xoQ8IW7QmjRy0ztUG_iG0SVG

Evaluation

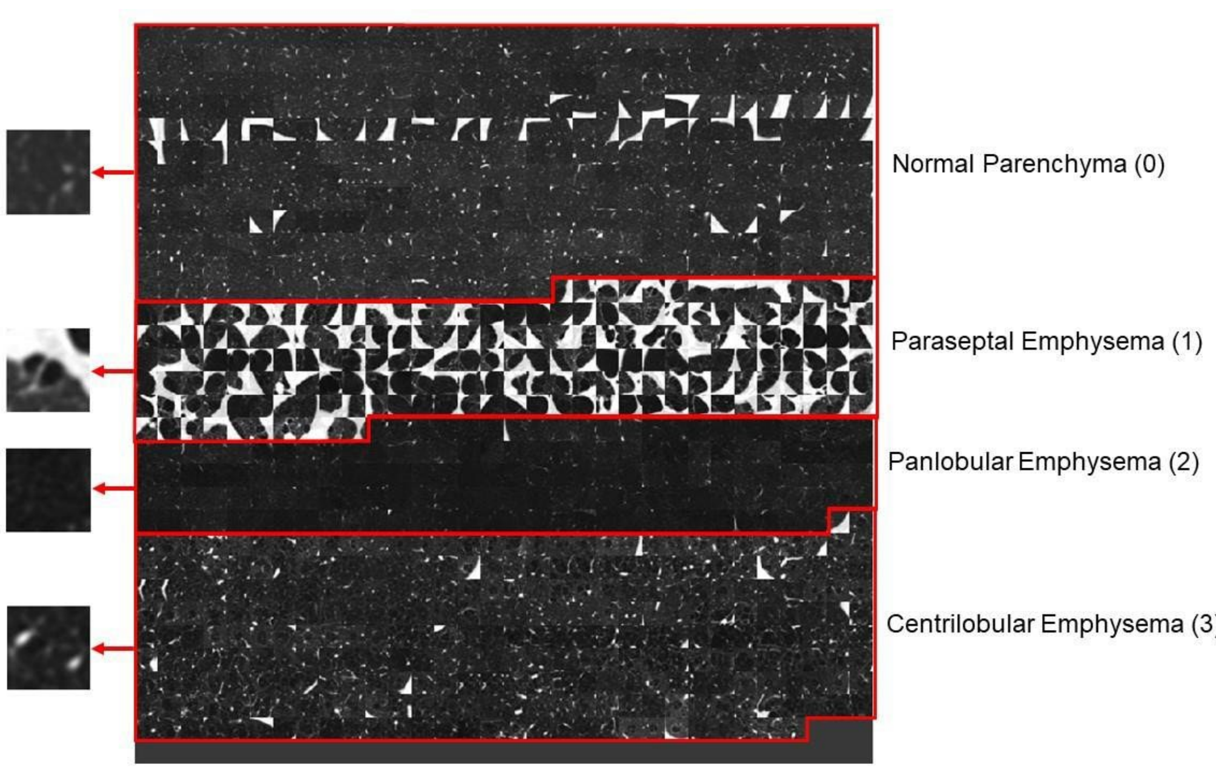
- The exercise will be accomplished **in pairs or groups of three people**.
- The final grade will depend on the following factors:
  - Implementation of all the functions and code requested (Python, Google Colab file).
  - Suitability and justification of the extracted features given the classification problem to be solved.
  - Comparative study between different ML classifiers.
  - The explanation of the Convolutional Network used in the deep learning approach and its training process.
  - Comparative study between ML and DL results.

- The assignment is divided into two different parts (ML, DL), and the students will have to submit **ONLY ONE** file:
  - ML_DL_surname1_surname2.ipynb (To export the file, click on *File/Download .ipynb* in the Google Colab webpage).
  - The Notebook file (.ipynb) **must be uploaded** to the **Moodle** page of the course.

Practical Exercise

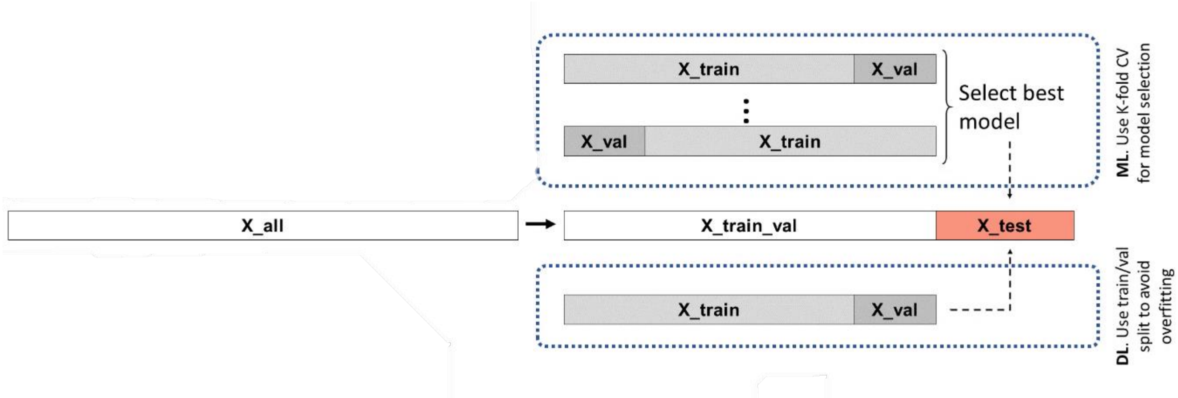
The goal is to classify, using Machine Learning and Deep Learning approaches, different radiological lung tissue patterns. The proposed problem is a multi-label classification problem, considering four different lung tissue patterns: **healthy tissue (HT-0)**, **paraseptal emphysema (PS1)**, **panlobular emphysema (PL-2)** and **centrilobular emphysema (CL-3).**

The dataset for the proposed classification problem has been selected from a subset of 267 CT scans, where experienced pulmonologists have manually selected regions corresponding to these 3 tissue classes —healthy tissue (370), paraseptal emphysema (184), panlobular emphysema (148) and centrilobular emphysema (287)—, providing a total of 989 tissue samples. They are two dimensional (2D) patches of **31x31 pixels**.



Two *.mat* files are provided. One containing the dataset (images), and another containing the corresponding labels.

The students must perform the classification problem proposed using both Machine Learning and Deep Learning techniques. For the ML approach, the students must select the best model using a CV strategy. The best model obtained will be compared to the CNN model **in the same independent test set.**



Machine Learning

The students will have to extract image-based features that represents the data. The students will also have to design several classifiers and select the best one based on a cross validation strategy.

The proposed model performance will be evaluated on an independent test set using evaluation metrics such as global accuracy and confusion matrices.

Deep Learning

The students will have to design and train a Convolutional Neural Network to solve the proposed classification problem. The CNN will be assessed on an independent test set using metrics such as global accuracy and confusion matrices.

We propose the students to implement the following CNN architecture:

$$
\begin{aligned}
INPUT[shape] &\rightarrow CONV[k, n1, act1] \rightarrow BATCHNORM \rightarrow MAXPOOL[p, s] \rightarrow \\
&\rightarrow CONV[k, n2, act1] \rightarrow BATCHNORM \rightarrow MAXPOOL[p, s] \rightarrow \\
&\rightarrow CONV[k, n3, act1] \rightarrow BATCHNORM \rightarrow MAXPOOL[p, s] \rightarrow DROPOUT[prob] \rightarrow \\
&DENSE[n3, act2] \rightarrow FLATTEN \rightarrow DENSE[o, act3]
\end{aligned}
$$

Where:

- **shape***: the shape of the input, based on the size of the images: width (*w*), height (*h*) and number of channels (*c*). Ex: [255,255,3] for RGB image of size 255x255
- **k**: kernel size. k=3
- **n**: number of filters. n1=16; n2=32; n3=64
- **p**: pool size. p=2
- **s**: stride. s=2

- **prob:** dropout probability. prob=0,5
- **act**: activation function. act1= 'relu', act2 = 'linear', act3='softmax'
- **o***: number of output neurons

** These values must be selected by the students according to the classification problem under study.*

Use the following training parameters:

Optimization method	Stochastic Gradient Descent. LR=0.01, Momentum=0, Nesterov=False
Loss function	Categorical crossentropy
Validation metric	Accuracy
Number of epochs	100
Early stopping	Yes. Stop the training if the validation loss has not improved after 15 epochs (patience=15). Restore model weights from the epoch with the best value of the val loss (restore_best_weights=True)
Batch size	100
Validation split	Use 20% of the samples for validation purposes

Additionally, the students will have to provide a brief comparative study based on the results obtained with the best ML obtained model and the ones obtained with the DL approach. This comparative study will be based on the same test set and must be written in the Jupyter Notebook.