NEUROENGINEERING 2022-2023
PW 1 – Lung Cancer Detection

DATA STRUCTURE

Two datasets are considered in this project, both derived from the COSMOS dataset.

The dataset "FPreductionDataset_v3.h5" will be used for **Task #1** and **Task #2.** These tasks will be focused on false positive reduction. It includes N = 7161 region proposals associated with the True positive class (TP) or the False Positive (FP) class. Considering that different regions are derived from the same patient CT scan (81 patients), point "a" of all the tasks (**Specific Tasks** section) is fundamental to avoid dependencies among training and validation datasets.

The following fields can be found inside the dataset:

- "images": vector of dimension Nx26x40x40, which corresponds to the volumetric regions centered in the region proposal coordinate;
- "anonID": vector of dimension Nx1 which contains the identification number of the subject;
- "roiID": vector of dimension Nx1, which contains the identification number of the region proposal;
- "target": binary vector of dimension Nx1 which labels each region as TP (value =1) or FP (value=0).
- "ExamDate": vector of dimension Nx1, which contains the date of the CT scan acquisition.

The dataset "MalignancyClassification.h5" will be used for **Task #3** and **Task #4,** which will be focused on the classification of pulmonary lesions as benign or malignant. It includes N = 2658 pulmonary lesions coming from 389 different patients. Different lesions can therefore be associated with the same patient similarly to region proposals in case of false positive reduction. It justifies point a) also for tasks #3 and #4.

The following fields can be found inside the dataset:

- "images": vector of dimension Nx26x40x40, which corresponds to the volumetric regions centered in the lesion;
- "anonID": vector of dimension Nx1 which contains the identification number of the patient;
- "LesionNumber": vector of dimension Nx1 which includes the identification number of the lesion;
- "Malignancy": binary vector of dimension Nx1, which labels each lesion as malignant (value =1) or benign (value=0).
- "radiomics_features": vector of dimension Nx190, which contains a set of radiomics features previously selected through a hierarchical clustering approach. Names of features are given through the file "radiomics_features_names.pickle".
- "Diameter_mm" and "Volume_mm3": additional binary vectors which indicate the dimension of lesions in terms of diameter and volume expressed in millimeters and cubic millimeters, respectively.

MODELS TRAINING APPROACH

Thanks to the availability of labeled data, all the models will be trained through a supervised approach.

SPECIFIC TASKS

Task 1 (5 people)

a) Considering the "FPreductionDataset_v3.h5" dataset, define training and validation sets i) with the same proportion of TP and FP image patches. When multiple patches are associated with the same subject, (ii) assign them to the same set (training or validation).
b) Implement a 3D CNN for false positive reduction, as reported by **Dou et al. 2017**. Two or three 3D-CNN with similar architecture will be therefore defined and their output combined as explained by the authors.
c) Considering a balanced number of true positives and false positives, try to evaluate the network with different parameters like filter dimensions, optimizers, number of layers/filters, etc.
d) Apply the trained model on the validation set and evaluate the classification performance by computing the AUC, F1-score, precision, and recall.
e) Compare the classification performance of the ensembled model with respect to consider the architectures singularly.
f) Investigate the possibility to train simultaneously the two/three architectures considered with respect to combine the probabilities given as output.
g) On the optimized model (best model according to tests applied at point "c"), evaluate the effect of applying different resampling strategies (oversampling of minority class without data augmentation, undersampling of the majority class, undersampling/oversampling combination of majority/minority class, weighted loss).

Task 2 (5 people)

a) Considering the "FPreductionDataset_v3.h5" dataset, define training and validation sets (i) with same proportion of TP and FP image patches. When multiple patches are associated to the same subject, (ii) assign them to the same set (training or validation).
b) Implement a 3D CNN for false positive reduction, as reported by **Ding et al. 2017**.
c) Considering a balanced number of true positives and false positives, try to evaluate the network with different parameters like filter dimensions, optimizers, number of layers/filters, etc.
d) On the optimized model (best model according to tests applied at point "c"), evaluate the effect of applying at least two different resampling strategies (oversampling of the minority class without data augmentation, undersampling of the majority class, undersampling/oversampling combination of majority/minority class, weighted loss).
e) On the optimized model (best model according to tests applied at point "c"), investigate graphically if a relation exists between correctly or misclassified TP and their dimension.
f) Investigate the effect of adding skip connections to the network or investigate the 2D version of the implemented network.

Task 3 (5 people)

a) Considering the "MalignancyClassification.h5" dataset, define training and validation sets (i) with the same proportion of Malignant/benign lesion image patches. When multiple patches are associated with the same subject, (ii) assign them to the same set (training or validation).
b) Implement a 2D version of CNN for malignancy classification, as reported by **Dou et al. 2017**. Two or three 2D-CNN with similar architecture will be therefore defined and their output combined as explained by the authors. (For each 2D ROI, consider the central slice)
c) Considering a balanced number of TP and FP, try to evaluate the network with different parameters like filter dimensions, optimizers, number of layers/filters etc.
d) Apply the trained model on the validation set and evaluate the classification performance by computing the AUC, F1-score, precision, and recall.

e) Compare the classification performance of the ensembled model with respect to consider the architectures singularly.
f) On the optimized model (best model according to tests applied at point “c”), evaluate the effect of combining an additional classification model based on radiomics features.
g) Investigate the possibility to train the architecture with the best performance between the two/three architectures considered, with an additional branch that takes as input radiomics features.

Task 4 (5 people)

a) Considering the “MalignancyClassification.h5” dataset, define training and validation sets (i) with the same proportion of Malignant/benign lesion image patches. When multiple patches are associated with the same subject, (ii) assign them to the same set (training or validation).
b) Implement a 2D CNN for malignancy classification considering the architecture proposed by **Ding et al. 2017**. (For each 2D ROI, consider the central slice)
c) Considering a balanced number of true positives and false positives, try to evaluate the network with different parameters like filter dimensions, optimizers, number of layers/filters, etc.
d) On the optimized model (best model according to tests applied at point c)), evaluate the effect of combining an additional classification model based on radiomics features. The two models should be combined with the same modality proposed by Dou et al. (2017).
e) Investigate the possibility to add a second branch that takes radiomics features as input to avoid the combination of probabilities applied at point “d”.