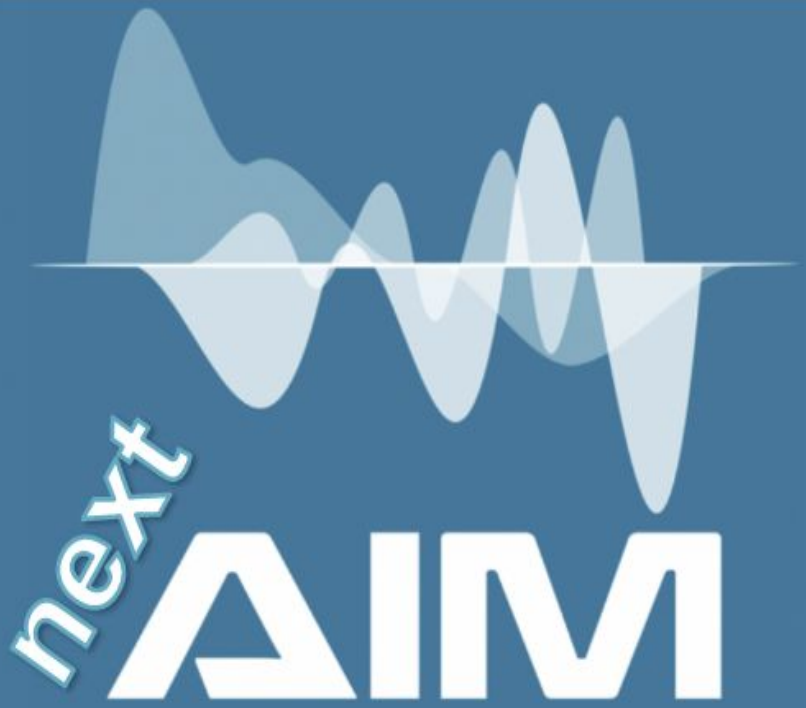
Artificial Intelligence in Medicine



Analysis of small datasets in radiomics and Machine Learning

Leonardo Ubaldi

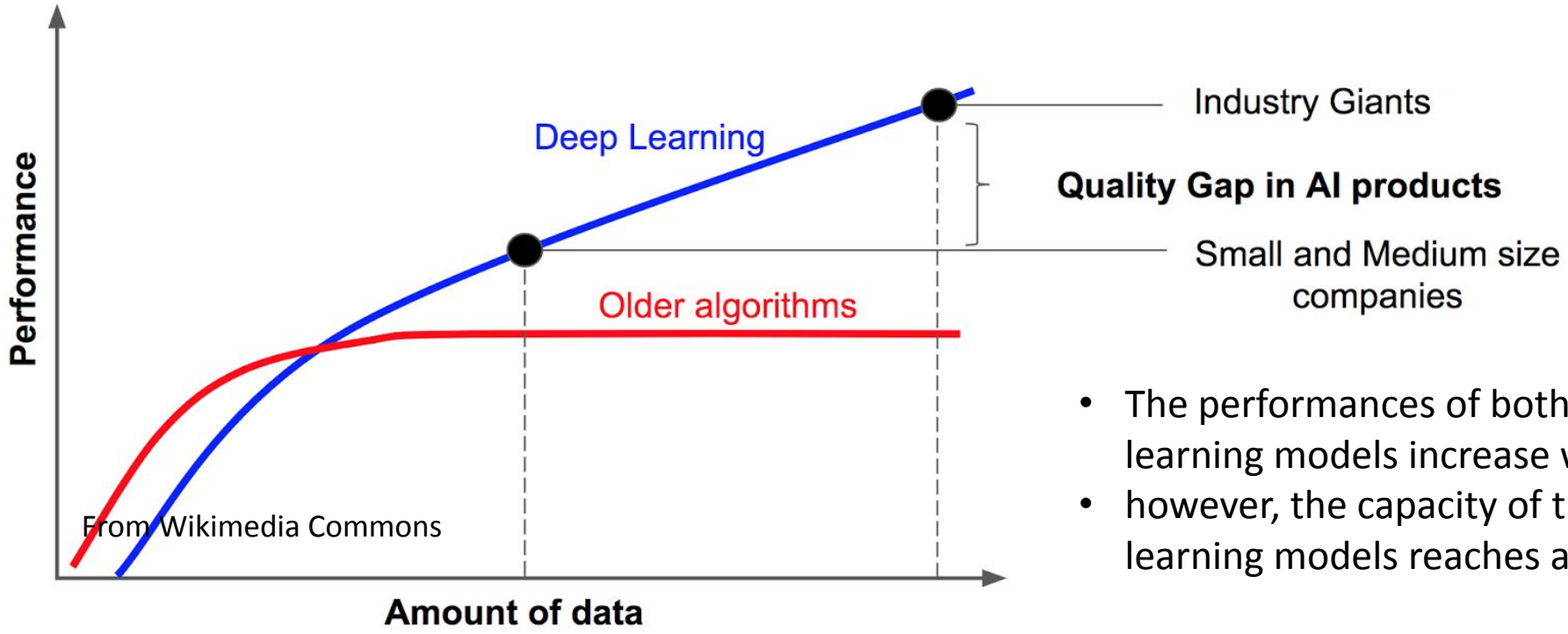
INFN

Outline

- Machine Learning and Deep Learning
- Challenges when dealing with samples of limited size:
  - Efficient optimization
  - Evaluation of robustness and reliability of trained models
- Examples from medical imaging data analysis
- Conclusions

Performance of the algorithms vs. sample size

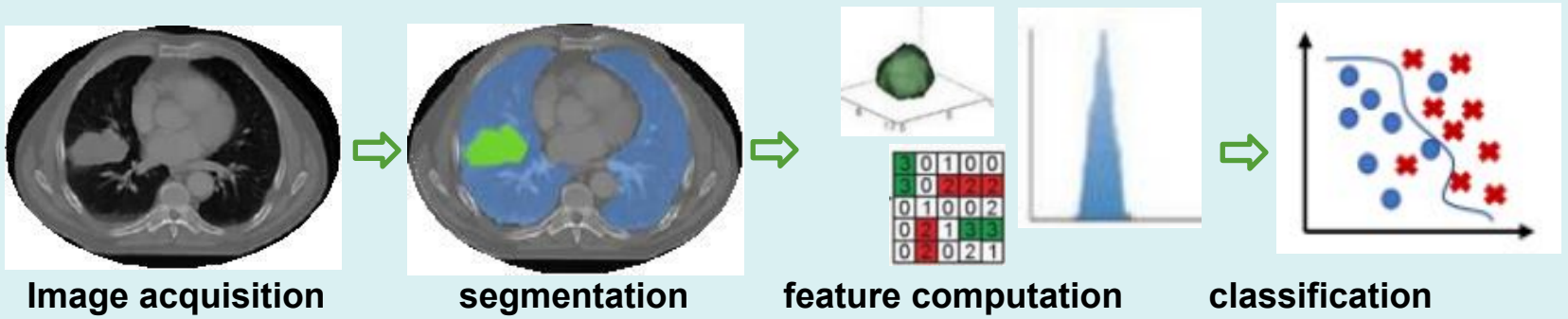
- Traditional machine learning models can perform even better than deep learning ones for small sample sizes
- Deep learning models definitely outperform traditional ones in case of large and meaningful data samples



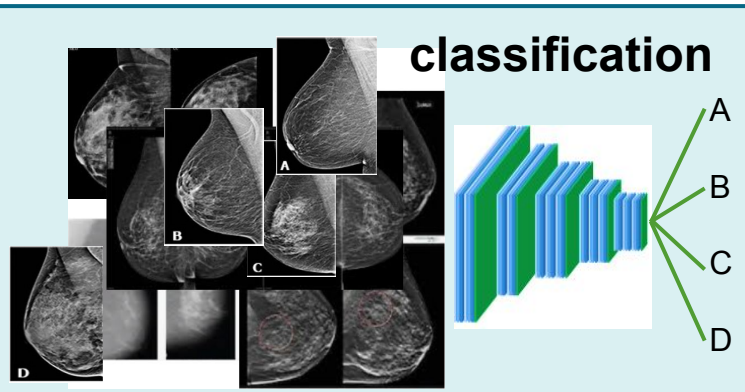
- The performances of both traditional and deep learning models increase with data sample size
- however, the capacity of traditional machine learning models reaches a limit at a certain point

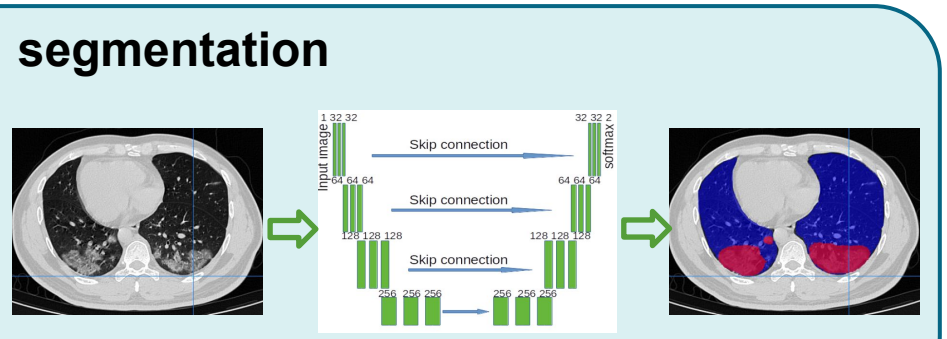
Typical ML-based approaches used in medical imaging data analysis

a) Radiomics + Machine Learning



b) Deep Learning





Difficulties in gathering large annotated samples in the medical field

- Data annotation by human experts is an extremely time-consuming task, which may require the collection of additional information stored in other data sources, expertise in segmenting meaningful regions in images, or specific knowledge to assign class labels.

- Gathering data and annotations from many sources increases the heterogeneity of the sample, which therefore requires to be harmonized.

- Trade-off between the quality and the size of the datasets
- In radiomics the dataset sizes range from a few dozens to a few hundreds of instances.

Challenges when dealing with data samples of limited size

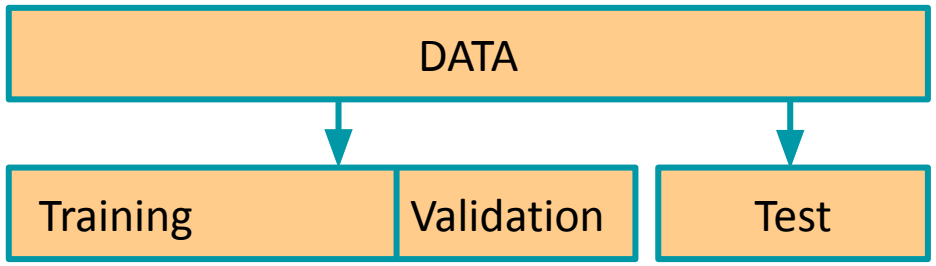
- In machine learning process, there is a trade-off between underfitting and overfitting

- We have to chose a model with a complexity degree suitable to fit our data
- In case of limited sample sized we often risk to use too complex model (overfit)

- Instability →
  - Hyper-parameter optimization
  - Performance evaluation

Data partition schemes

Hold out



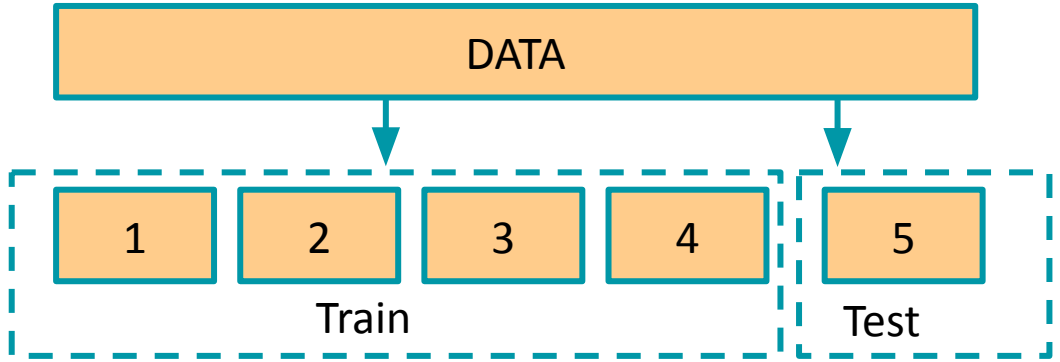
Data can be split in a Training (Training + Validation) and Test sets, both (hopefully) representative of the whole population.

Typical split portions are:

- 80% in train; 20% in test
- 70% in train; 30% in test

The average of five-ten repetitions with random splits provide test performance with standard error

Cross-validation



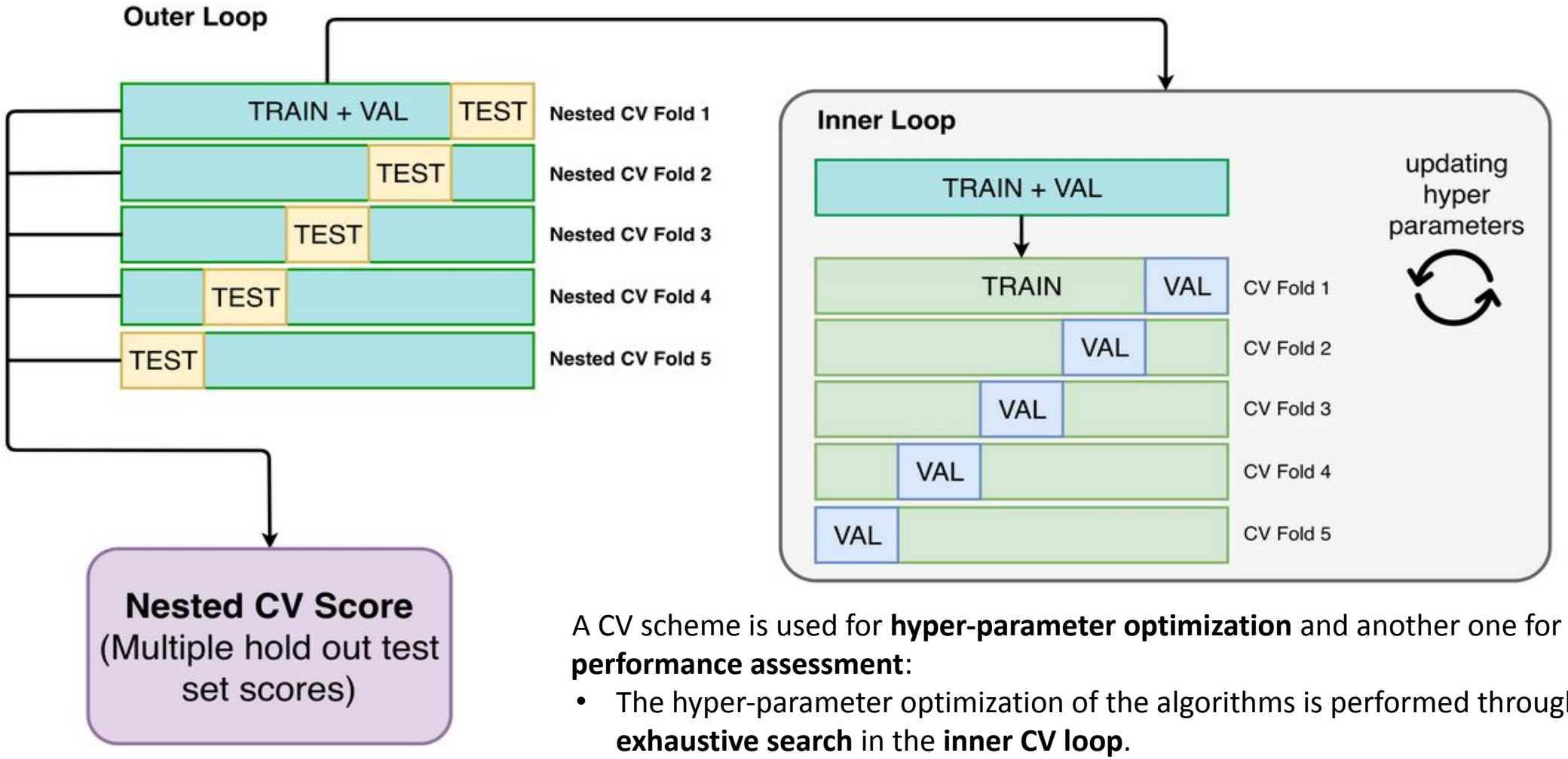
K-fold cross validation (CV):

- Data is partitioned into K subsamples: one is retained as test data while remaining (K − 1) subsamples are used as training data (training).
- CV process is repeated K times (the folds), with each of the K subsamples used exactly once as test data.
- The K results from the folds then can be averaged (or otherwise combined) to produce a single estimation.

Leave-one-out CV:

- K-fold CV with K=Numbers of samples, thus each fold has only one example. It is used in case the dataset is extremely limited in size.

Nested CV

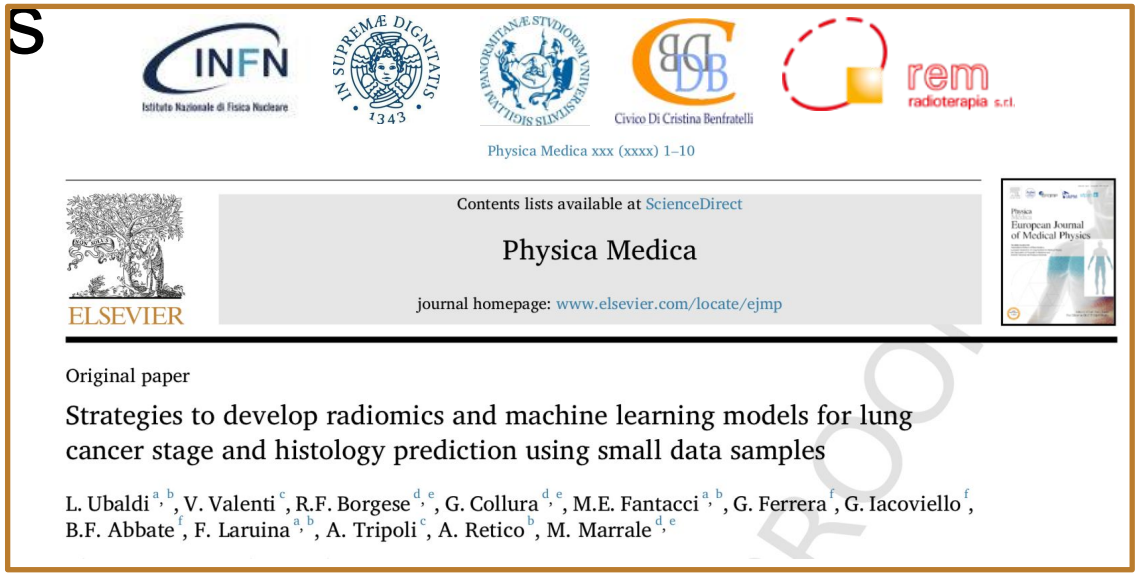


A CV scheme is used for **hyper-parameter optimization** and another one for **performance assessment**:

- The hyper-parameter optimization of the algorithms is performed through an **exhaustive search** in the **inner CV loop**.
- The **performances** are evaluated in the **outer CV loop.**

Practical example: Radiomics and Machine Learning models for lung cancer stage and histology prediction using small data samples

- Goal: To determine the stage and histology is crucial for tumor treatment.
- Imaging-based classification via radiomic features would avoid biopsy, reducing also the risk of biopsy sampling error, as the whole lesion volume is considered.



INFN Istituto Nazionale di Fisica Nucleare

IN SUPREMÆ DIGNITATIS 1343

Civico Di Cristina Benfratelli

rem radioterapia s.r.l.

Physica Medica xxx (xxxx) 1–10

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

ELSEVIER

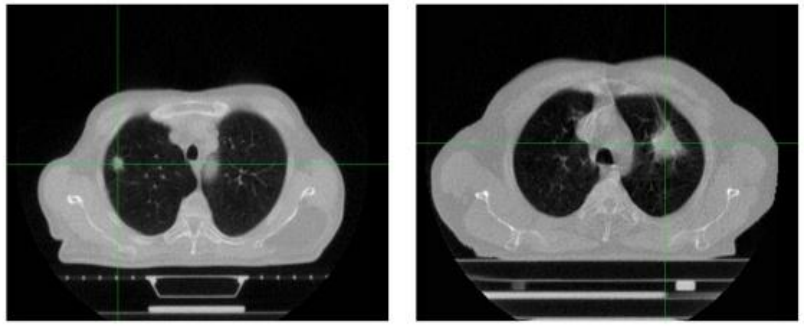
Original paper

Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples

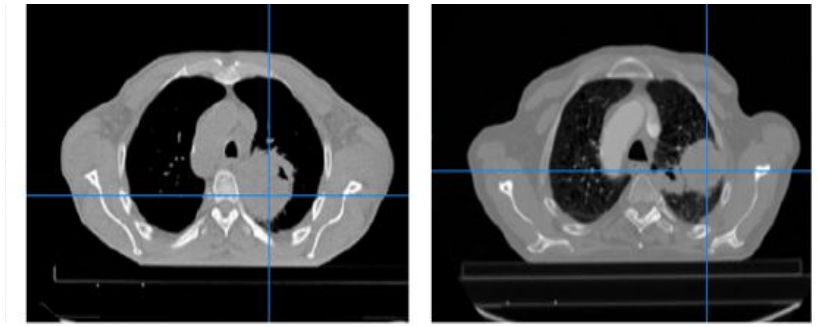
L. Ubaldi [a, b], V. Valenti [c], R.F. Borgese [d, e], G. Collura [d, e], M.E. Fantacci [a, b], G. Ferrera [f], G. Iacoviello [f], B.F. Abbate [f], F. Laruina [a, b], A. Tripoli [c], A. Retico [b], M. Marrale [d, e]

[Ubaldi, L., Valenti, V., Borgese, R. F., Collura, G., Fantacci, M. E., Ferrera, G., … Marrale, M. (2021). Strategies to develop Radiomics and Machine learning models for lung cancer stage and histology prediction using small data samples, *Physica Medica, in press*]

Available datasets: L-RT (proprietary) and Lung1 (public)



L-RT proprietary data sample collected at the A.R.N.A.S. Civico University Hospital of Palermo (IT):
47 CT scans of patients with non-small cell lung cancer (NSCLC)



Lung1 public data sample available on TCIA, https://www.cancerimagingarchive.net/
130 CT scans of patients with NSCLC

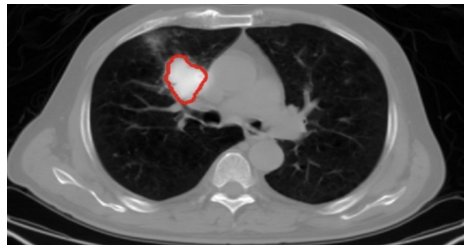
Histology and overall stage distributions

Histology	L-RT	Lung1
Adenocarcinoma	20	16
Large Cell Carcinoma	4	60
Squamous Cell Carcinoma	10	54
Not Available	13	-
Total number of subjects	**47**	**130**

Overall Stage	L-RT	Lung1
I	42	27
II	5	13
IIIa	-	37
IIIb	-	53
Total number of stage I-II/IIIa-IIIb	**47/0**	**40/90**

Radiomic features and ML classification

1)



Lesion segmentation (manually drawn Radio Therapy structures GTV)

2)

	A	B	C	D	E
1	Sphericity	LeastAxisLength	Elongation	SurfaceVolumeRatio	Histology
2	0.694819919246	38.50646144207245	0.78664764976375645	0.163675082636	squamous cell carcinoma
3	0.707875921926	19.443729272738086	0.78845025630697596	0.344411910261	large cell
4	0.579328926872	10.464535362578046	0.37911586182340895	0.48100129891	large cell
5	0.601815100141	26.590179110243287	0.79178437526078194	0.285624057955	adenocarcinoma
6	0.774430121688	19.004433653318628	0.77866838289624618	0.318737279996	squamous cell carcinoma
7	0.626051916544	15.259689499690175	0.78523110726095935	0.392934978679	squamous cell carcinoma
8	0.730329166351	11.41495074124853	0.76690923921714749	0.526202512138	squamous cell carcinoma
9	0.772637658529	16.313901641323049	0.66667694051360216	0.364683076645	adenocarcinoma
10	0.688525901706	28.35568214330625	0.6935222891217816	0.215918348833	adenocarcinoma

Radiomic feature computation

3)

Feature scaling

Transforms the range of the input features. Some ML algorithms do not perform well when the input features have very different ranges.

Dimensionality reduction

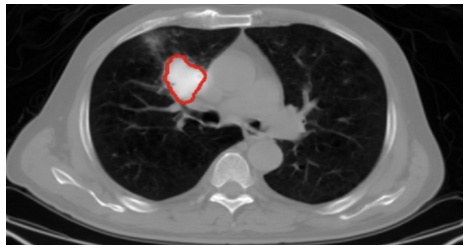
Reduces the number of considered features. In high-dimensional space the training is slower and there is a higher risk of overfitting.

Machine Learning algorithm

Classifiers

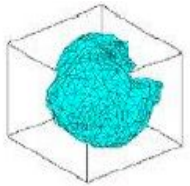
- AdaBoost
- Random Forest
- Nearest Neighbors
- Support Vector Machine (SVM)

Radiomic features



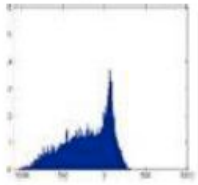
	A	B	C	D	E
1	Sphericity	LeastAxisLength	Elongation	SurfaceVolumeRatio	Histology
2	0.694819919246	38.50646144207245	0.78664764976375645	0.163675082636	squamous cell carcinoma
3	0.707875921926	19.443729272738086	0.78845025630697596	0.344411910261	large cell
4	0.579328926872	10.464535362578046	0.37911586182340895	0.48100129891	large cell
5	0.601815100141	26.590179110243287	0.79178437526078194	0.285624057955	adenocarcinoma
6	0.774430121688	19.004433653318628	0.77866838289624618	0.318737279996	squamous cell carcinoma
7	0.626051916544	15.259689499690175	0.78523110726095935	0.392934978679	squamous cell carcinoma
8	0.730329166351	11.41495074124853	0.76690923921714749	0.526202512138	squamous cell carcinoma
9	0.772637658529	16.313901641323049	0.66667694051360216	0.364683076645	adenocarcinoma
10	0.688525901706	28.35568214330625	0.6935222891217816	0.215918348833	adenocarcinoma

107 radiomic features were extracted within the Gross Tumor Volume (GTV)



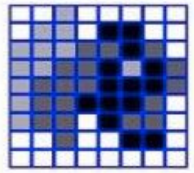
17 Size- and Shape-based Features

Features that describe the 2D or 3D size and shape of the ROI.



23 First Order Statistics Features

Features computed from the histogram that represents the occurrences of voxel values within the ROI.



67 Higher Order Statistics Features

Features that describe the inter-relationships between two or more voxels of the image.

pyradiomics

python

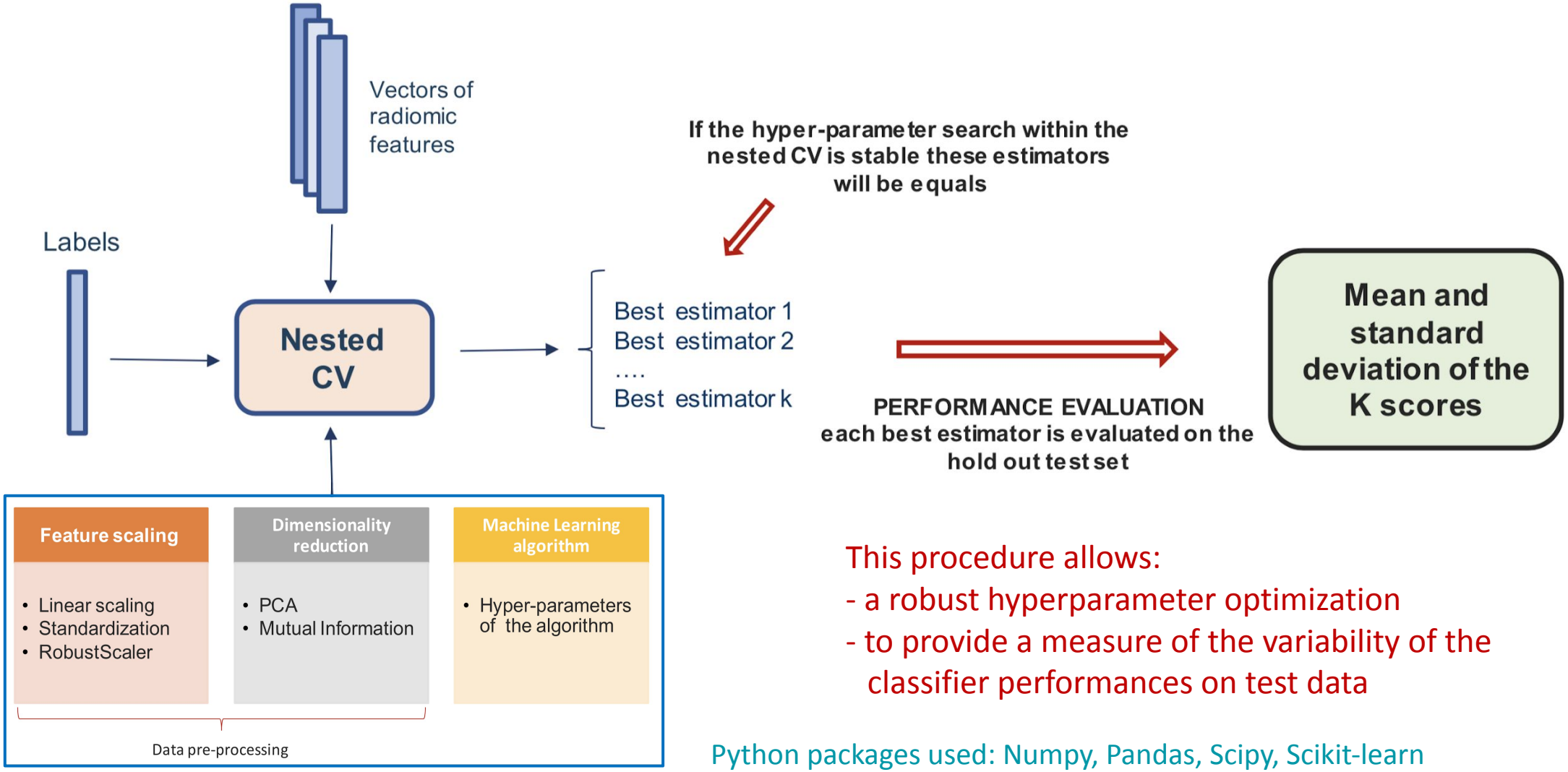
+

RADIOMICS

Radiomic features are computed according to the standardized definitions provided by the Image Biomarker Standardization Initiative (IBSI).

[Zwanenburg A. *et al*. Radiology 2020;295: 328–38. https://doi.org/10.1148/radiol.2020191145.]

Nested-CV scheme for pipeline optimization

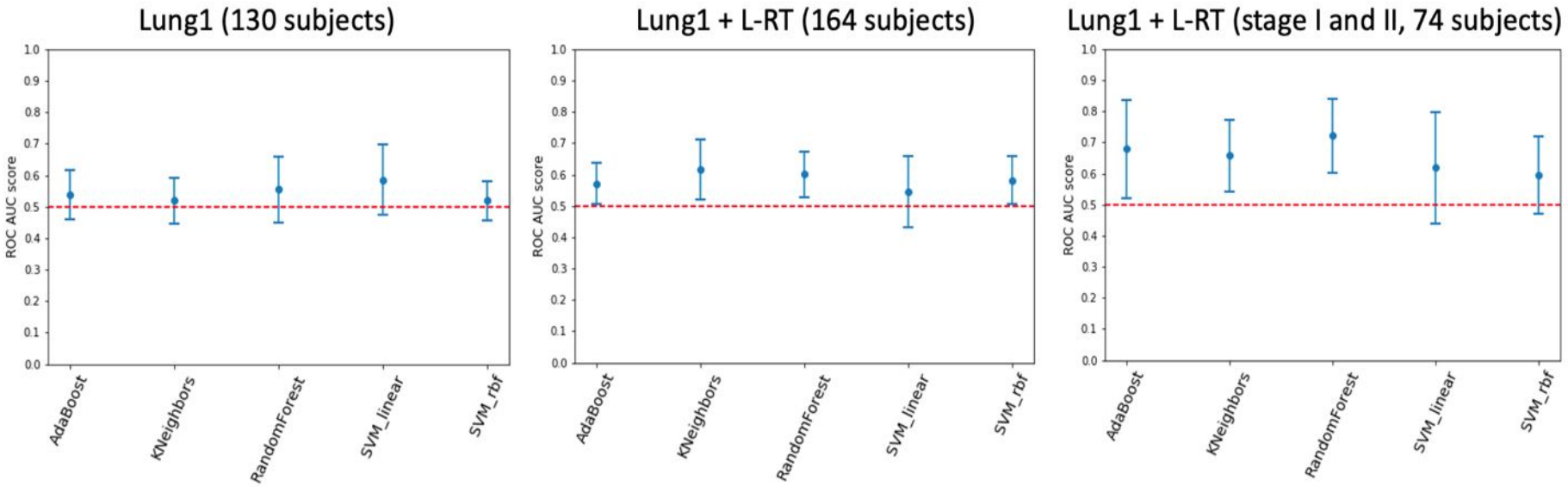


This procedure allows:
- a robust hyperparameter optimization
- to provide a measure of the variability of the classifier performances on test data

Python packages used: Numpy, Pandas, Scipy, Scikit-learn

Results for histology classification

Random Forest		AUC on the TEST SET			
Histology classification		L-RT	Lung1	Total-L	Total-L (only OS I and II)
TRAIN SET	L-RT	C.L.	C.L.	//	//
	Lung1	C.L.	C.L.	//	//
	Total-L	//	//	0.60 ± 0.07	//
	Total-L (only OS I and II)	//	//	//	0.72 ± 0.11



Analysis pipeline optimization with rigorous nested-CV

- L-RT and Lung1 are separately too small and heterogeneous to provide results above the chance level.
- On the merged sample (Total-L), the classification performance is slighty above the chance level.
- It increases for reduced heterogeneity of the sample (restriction to OS I and II)

The variability of the performances on the test sets is high, due to the small sample sizes

Conclusions

- Drawing conclusions from the analysis of data samples of limited size with Radiomics, Machine Learning and Deep Learning approaches is quite common in the field of medical imaging
- Specific technical issues should be addressed in these cases, to ensure to have carried out:
  - efficient training and optimization with limited samples
  - rigorous evaluation of the robustness and reliability of the results
- As general guidelines:
  - Choose the simplest possible model to fit the data
  - Use nested CV for hyperparameter optimization and performance evaluation

Thank you for your kind attention!

For further information and references contact me:

leonardo.ubaldi@unifi.it

Pisa, 17/02/2022