Physica Medica 90 (2021) 13–22

Contents lists available at ScienceDirect

Physica Medica

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

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 [g], G. Iacoviello [f], B.F. Abbate [f], F. Laruina [a,b], A. Tripoli [c], A. Retico [b,*], M. Marrale [d,e]

[a] Physics Department, University of Pisa, Pisa, Italy
[b] National Institute for Nuclear Physics (INFN), Pisa Division, Pisa, Italy
[c] REM Radiation Therapy Center, Viagrande (CT), I-95029 Catania, Italy
[d] Physics and Chemistry Department "Emilio Segrè", University of Palermo, Palermo, Italy
[e] National Institute for Nuclear Physics (INFN), Catania Division, Catania, Italy
[f] Medical Physics Department, ARNAS-Civico Hospital, Palermo, Italy
[g] Radiation Oncology, ARNAS-Civico Hospital, Palermo, Italy

ARTICLE INFO

Keywords:
Radiomics
Machine learning
Cross validation
Non-small cell lung cancer

ABSTRACT

Predictive models based on radiomics and machine-learning (ML) need large and annotated datasets for training, often difficult to collect. We designed an operative pipeline for model training to exploit data already available to the scientific community. The aim of this work was to explore the capability of radiomic features in predicting tumor histology and stage in patients with non-small cell lung cancer (NSCLC).

We analyzed the radiotherapy planning thoracic CT scans of a proprietary sample of 47 subjects (L-RT) and integrated this dataset with a publicly available set of 130 patients from the MAASTRO NSCLC collection (Lung1). We implemented intra- and inter-sample cross-validation strategies (CV) for evaluating the ML predictive model performances with not so large datasets.

We carried out two classification tasks: histology classification (3 classes) and overall stage classification (two classes: stage I and II). In the first task, the best performance was obtained by a Random Forest classifier, once the analysis has been restricted to stage I and II tumors of the Lung1 and L-RT merged dataset (AUC = 0.72 ± 0.11). For the overall stage classification, the best results were obtained when training on Lung1 and testing of L-RT dataset (AUC = 0.72 ± 0.04 for Random Forest and AUC = 0.84 ± 0.03 for linear-kernel Support Vector Machine).

According to the classification task to be accomplished and to the heterogeneity of the available dataset(s), different CV strategies have to be explored and compared to make a robust assessment of the potential of a predictive model based on radiomics and ML.

Introduction

Radiomics is an emerging field of research in the context of medical image analysis [1,2]. It is based on the extraction and analysis of quantitative imaging features from medical images to exploit them in clinical decision support. The primary hypothesis is that radiological images are much more than just anatomical representations [3]. Modern diagnostic techniques generate a huge amount of information in the form of numerical data that escape mere visual observation and that the human mind cannot process. In daily clinical practice, medical images are in general only visually assessed by medical experts, not fully exploiting their quantitative potential. In this way, a lot of possibly meaningful information, which is not appreciable by the human eye, is lost. Radiomics aims to use this information to help clinicians in different tasks, such as making diagnosis, predicting prognosis and therapeutic response [3]. Descriptive quantities extracted from images are defined as radiomic features.

Machine learning (ML) and deep learning (DL) algorithms are currently successfully applied in many different fields, due to their capability to make predictions [4]. Usually, these techniques need large data samples for appropriate model training. Therefore, they are particularly suited to deal with the so-called big data, which indicate a massive volume of data that is too large or complex to be effectively analyzed using traditional software.

* Corresponding author at: National Institute for Nuclear Physics (INFN), Pisa Division, Largo Bruno Pontecorvo 3, 56127 Pisa, Italy.

https://doi.org/10.1016/j.ejmp.2021.08.015
Received 3 April 2021; Received in revised form 21 August 2021; Accepted 28 August 2021
Available online 11 September 2021

Since the beginning of the digital era, the amount of data produced has considerably increased, even in the field of medicine [5]. At the same time, the interest in the application of ML and DL approaches in medical data analysis has grown enormously [6–9]. A specific difficulty that is encountered in developing ML and DL analysis tools for medical applications is that the training processes often require large amounts of annotated data. Manual data annotation is a time-consuming task that prevents the creation of sufficiently large samples in most cases. In addition to the annotation problem, it is not easy to collect large homogenous datasets in the field of medical imaging. In particular, the intrinsic heterogeneity of retrospective data accumulated in daily clinical practice creates a trade-off between the quality and the size of the datasets [3]. In most radiomic works, the dataset sizes range from a few dozens to a few hundreds of patients [10]. Despite the relatively small sizes of these datasets, the quantitative features data extracted from images are generally analyzed with machine learning or deep learning techniques.

In this study we propose a strategy to develop ML models to predict the tumor histology and stage by using radiomic features extracted from thoracic CT scans of patients with Non-Small Cell Lung Cancer (NSCLC). Lung cancer is the leading cause of cancer-related mortality worldwide with a median age at diagnosis of 70 years [11], and 85% of cases are represented by NSCLC [12]. Historically, surgical resection with curative intent is considered the cornerstone of treatment for early-stage NSCLC which, while accounting for only 25–30% of lung cancer, theoretically, offers the highest possibility of modifying the outcome of NSCLC [13,14]. For early-stage NSCLC, the Tumor-Node-Metastasis (TNM) stage is traditionally considered the most important post-operative prognostic factor [15,16]. However, the broad spectrum of survival times that exist even after complete resection of stage NSCLC demonstrates the mandatory need for personalized medicine [17]. The improvement in survival estimation has mostly been achieved as a result of advances in biological and genomic technologies that have allowed for the implementation of biological or genetic signatures associated with survival [18,19]. However, the inability to obtain complete information on heterogeneous tumors remains a limitation of these invasive methods [20,21]. Similarly, patients with clinically suspected NSCLC, especially elderly ones [22], may have medical comorbidities that increase biopsy risks, making them more likely to receive stereotactic body radiation therapy (SBRT) without a biopsy [21].

By using radiomics and advanced analysis systems, it is possible to extract many quantitative descriptors from routinely acquired CT studies. Radiomics allows the non-invasive identification of predictive signature of tumor heterogeneity [23–25], which, in clinical practice, can be used for tumor detection, subtype classification and therapeutic response assessment. This approach could be important for patients treated with radiotherapy to stratify patients at risk of relapsing disease [26]. In particular, this is important for the patients with early-stage NSCLC treated with SBRT of whom 13–23% develop distant metastases and 4–14% experience relapse after treatment [27–32]. Therefore, patients with the highest risk of relapse post-SBRT could be candidates to receive additional or escalated treatment to prevent it. Radiomics could potentially identify these high-risk patients.

The target of this work is the development of a radiomics and ML-based method for classifying histology and stage on patients with NSCLC that have undergone SBRT treatments at the A.R.N.A.S. Civico Hospital of Palermo. These patients treated with radiation therapy have lesions classified as belonging to early (I and II) TNM stages.

One of the key points of this study regards the possibility of carrying out a robust assessment of ML model performance when dealing with small samples. We emphasized the importance of adopting cross-validation (CV) strategies to this purpose. More specifically, we implemented from scratch a nested CV strategy which is considered as the most rigorous method for training, optimizing and evaluating ML models [7].

Although the potential impact of radiomics and ML-based decision support systems on lung cancer characterization and personalized treatments is very high, the development and validation of predictive models is challenged by the difficulty of collecting large, annotated samples for model training. We investigated in this study the feasibility of using publicly accessible larger datasets to train decision-making systems and to transfer the knowledge gained to less populated private data samples.

This paper is organized as follows: the two data samples used in this analysis are first described and their composition in terms of tumor histology and stage are compared; the analysis workflow, both regarding the radiomic feature extraction, their selection and the implementation of a number of ML algorithms, is thus presented, paying special attention to the need to set up CV strategies when dealing with data samples of limited size, and nested CV loops for hyperparameter optimization; the results are thus reported for the most widely used ML algorithms.

Materials and methods

Datasets

Private data collection: L-RT dataset

A dataset of thoracic CT scans and radiotherapy structures has been collected at the A.R.N.A.S. Civico University Hospital of Palermo. It will be referred to in this study as the L-RT dataset. It consists of 47 patients with non-small cell lung cancer (NSCLC). Informed consent was obtained from all participants included in the present study. All procedures were performed in agreement with the 1964 Helsinki declaration. For each subject, the thoracic CT scan acquired for radiotherapy planning and the segmentation of the Gross Tumor Volume (GTV) region of interest, drawn by radiation oncologists, have been considered in this study.

All the patients in this dataset underwent staging with bronchoscopy and lung and upper abdomen CT (2.5 mm slice thickness), without intravenous contrast. Diagnostic information, such as the tumor histology and the tumor stage or overall stage (OS) is also available (see Table 1). The histology labels of the L-RT dataset correspond to adenocarcinoma, large cell carcinoma and squamous cell carcinoma, whereas their stages are limited to the I and II stage according to the most recent TNM classification at the time of analysis [33].

The characteristics of the enrolled patients are summarized in the Supplementary Materials. Regarding the dose prescription, using a risk-adapted strategy [34], different dose schedules were used based on tumor location: 60–70 Gray (Gy) in 8–10 fractions for peripheral lesions, 50–60 Gy in 10 fractions for central and ultra-central lesions. The size and location of the tumor also had an impact in the fraction selection process [35,36], leading to the choice of schedules with BED10 < 100 Gy in those patients with more complex characteristics [37,38]. The study population consisted of 36 males and 11 females, with a median age of 73 years (range, 44–91 years). All patients completed SBRT without interruption. The most used fractionation schedule was 7.5 Gy $\times$ 8

Table 1
Number of instances per histology and per overall stage of the Lung1 and the L-RT datasets.

Histology	**Lung1**	**L-RT**
Adenocarcinoma	16	20
Large Cell Carcinoma	60	4
Squamous Cell Carcinoma	54	10
Not Available	–	13
Total number of subjects	**130**	**47**
Overall Stage	**Lung1**	**L-RT**
I	27	42
II	13	5
IIIa	37	–
IIIb	53	–
Total number of stage I-II/IIIa-IIIb tumors	**40/90**	**47/0**

fractions (BED10 105 Gy) [39] administered in 22 patients and 6 Gy × 10 fractions (BED10 96 Gy) administered in 14 patients. A total dose with BED10> = 100 Gy was delivered in 24 treatments. The overall median time on treatment was 15 days (range 10–24 days). SBRT was administered on consecutive days in 33 patients and on alternate days in 14 patients.

Public data collection: TCIA lung 1 dataset

A subsample of 130 subjects has been selected from the public data collection Lung1 [24,40] of patients with NSCLC, which contains data of 422 subjects and it is available via The Cancer Imaging Archive (TCIA) (https://www.cancerimagingarchive.net/). The Lung1 dataset is composed of PET and CT images, and the radiotherapy structures for each subject. A spiral CT (3 mm slice thickness), with and without intravenous contrast, performed covering the complete thoracic region, is reported for each subject. The subset of the 130 subjects of the Lung1 sample considered in this work has been selected considering only the patients whose GTVs were fully contained in the lung area and with overall size in the same range of those available in the private L-RT collection. Thus, only subjects of Lung1 for which the GTV segmentation was carried out with a similar segmentation criterion, as judged by a radiation oncologist, to the subjects of the L-RT dataset were included in the selection. An example of images available in the chosen subset of Lung1 is shown in Fig. 1.

This subset of the Lung1 dataset will be referred to in this paper as Lung1 for short. The complete list of subjects used in this study is reported in the Supplementary Materials. Analogously to the L-RT dataset, it contains diagnostic information such as the tumor histology (adenocarcinoma, large cell carcinoma and squamous cell carcinoma) and the tumor overall stage (I, II, IIIa and IIIb), as shown in Table 1.

As shown in Table 1, the L-RT dataset does not contain instances of tumors with overall stage IIIa and IIIb, which are the two more advanced stages, that are, instead, the largest classes represented in the Lung1 dataset.

Regarding the histology, for 13 subjects of the L-RT dataset this information is not available, and these subjects were not considered for the histology classification analyses but only in the overall stage classification task. Furthermore, it can be noticed that the most represented histology in the L-RT dataset (adenocarcinoma) is the least frequent one in the Lung1 dataset.

The radiomic workflow

A typical radiomic workflow, based on radiomic feature extraction and machine learning classification has been followed in this study. As schematized in Fig. 2, it is composed of several analysis steps, starting from the medical image acquisition, to end with the prediction of either the tumor histology or its stage from imaging descriptive features. Each step of this analysis workflow is detailed in the paragraphs below.

Segmentation

For each subject of both the L-RT and the Lung1 datasets, the segmentation of the Gross Tumor Volume (GTV) region of interest (ROI), drawn within the treatment planning by a radiotherapist, was available. Thus, these tumor masks were considered for the computation of radiomic features.

Feature extraction

Descriptive quantities extracted from images are defined as radiomic features. They are in general composed by size- and shape-based features, descriptors of the image intensity histogram and descriptors of the relationships between image pixels (or voxels). Radiomic features can be extracted from different kinds of medical images, such as computed tomography (CT), positron emission tomography (PET) or magnetic resonance imaging (MRI) [41]. In our analysis, radiomic features were extracted from original CT images without any specific preprocessing, using the PyRadiomics [42] plugin of 3D Slicer [43,44], which is a software package for viewing and post-processing medical images. It allows the computation of radiomic features on CT scans within the chosen regions of interest (i.e. the GTV in our analysis). The same procedure was used for both datasets: 107 radiomic features were extracted. A brief description and a complete list of these features is provided in the Supplementary Materials.

Feature selection and Machine-learning algorithms

We build a feature processing and analysis pipeline that is intended to define the set of features to be handled by the machine learning algorithms. We create an analysis pipeline [45] by combining different analysis modules: each step takes as input the output of the previous step. The pipeline is composed of: 1) a preprocessing step (i.e. feature scaling), 2) a dimensionality reduction step, 3) a machine learning algorithm. The hyperparameter space of the pipeline is composed by the product of all the possible choices for each step. The point of the

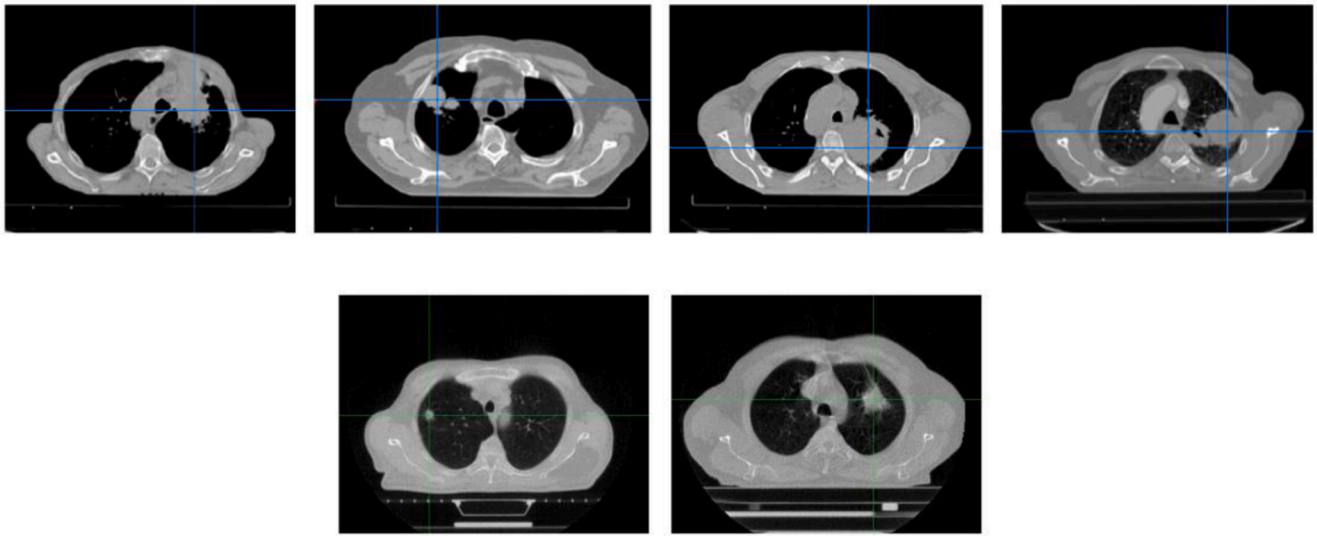

Fig. 1. Top: Sample images of the Lung1 dataset. From left to right the stage I, II, IIIa and IIIb of squamous cell carcinomas. Bottom: Sample images of the Lung-RT dataset. From left to right the stage I and II. The lung tumor position is indicated by the crossing lines.

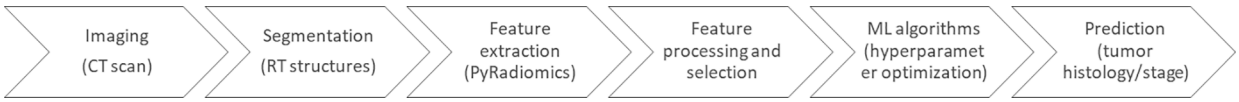

Fig. 2. Main steps characterizing the radiomic and ML approach, from the image acquisition to the prediction of either tumor stage or histology.

hyperparameter space in which the pipeline obtains the best performances, according to a given metric, has to be found. In other words, we have to tune the pipeline finding the best set of hyperparameters. Let us define the 'best estimator' as the pipeline tuned with the best set of hyperparameters. A pipeline for each chosen classifier was built.

The features scaling step refers to the operation of transforming the range of each single feature. It is an important step in the machine learning workflow. In fact, many machine learning algorithms, generally, do not perform well when the input features are in very different ranges. Using unscaled data, the performances of several algorithms can be degraded and the convergence may be slowed down or prevented [46].

Moreover, in many machine learning applications one has to deal with a large number of features for each instance in our dataset. Working in high-dimensional space may introduce some issues: the training is slower and there is a higher risk of overfitting. The latter occurs when a model performs very well on training data but does not generalize well on test data. A symptom of overfitting is that the performances on the test set are much lower than those on the training set. Generally, overfitting is due to the fact that the model is too complex given the underlying data, i.e. the model has too many parameters to learn from the data. The dimensionality reduction step allows to prevent overfitting by reducing the number of features considered [47]. The dimensionality reduction algorithms used in this study are: Principal Component Analysis (PCA) and Mutual Information (MI). Regarding the PCA, we tuned the number of principal components to retain by using the hyperparameter optimization strategy illustrated in section 1.2.4. The only hyperparameter to set is the variance explained by the principal components. The optimization algorithm chooses the percentage of explained variance which leads to the best performance among the three values of 0.85, 0.90, 0.95. Concerning the MI, we selected the 10% of the features with the highest MI value. Additional details are reported in the Supplementary Materials.

The classifiers considered in our study are among the most widely used ones: AdaBoost, Nearest Neighbors, SVM (with linear and radial-basis-function kernel) and Random Forest. A technical description on how each of them works can be found in textbooks [48] and in the scikit-learn documentation [49]. We report a brief description of the Random Forest classifier used in our analysis. The Random Forest classifier is a bootstrap aggregating ensemble model composed by decision tree classifiers.

In this work we consider the optimization of several hyperparameters, among which the most important are: the number of trees in the forest, the maximum depth of the tree, the criterion that measures the quality of each split of data, the minimum number of samples necessary to split an internal node and the minimum number of samples needed to be a leaf node.

The machine learning analysis was implemented using the Python package *scikit-learn* (v.0.23.2).

Hyperparameter optimization within nested Cross-Validation

The nested Cross-Validation (nested CV) is a procedure in which CV is used simultaneously for hyperparameter optimization and for performance assessment. It is composed of an inner CV loop nested in an outer CV loop. The inner CV loop performs hyperparameter tuning, while the outer CV loop evaluates the performances [50]. In our applications, we set the number of folds of both the outer and the inner CV loop equal to 5. The mean and the standard deviation of the 5 scores obtained by each best estimator on its own test set are calculated.

In this study, the exploration of the hyperparameter space is carried out with a Grid Search Cross-Validation (GSCV), which implements an exhaustive search over the space of hyperparameters given as input. The metric used is the area under the ROC curve (AUC). A description of the complete hyperparameter space considered for each algorithm is provided in the Supplementary Materials.

Strategies for integrating public and private datasets

We used in this study a private dataset of thoracic CT scans used for radiotherapy planning (L-RT) and a publicly accessible MAASTRO NSCLC dataset (Lung1), as detailed in Sec. 1.1. As the smaller private data sample may hardly be fully representative of the underlying population and thus suitable to train a robust decisional system, we propose the integration of the private sample with the publicly available one. A series of conditions should be verified to combine the two different datasets. Moreover, dealing with small datasets, which are generally characterized by more features per subject than subjects in the dataset, poses several challenges during the ML model training, e.g. those related to the optimization of the hyperparameters and to the risk of model overfitting. We implemented and compared three different CV strategies for training and testing the ML-based predictive models: 1) k-fold CV within each of the two datasets (within-sample train and test, WS-TT); 2) k-fold CV within the merged dataset (merged-sample train and test, MS-TT); 3) training a predictive model on one sample and testing it on the other one (cross-sample train and test, CS-TT), specifically expecting better performance when training a predictive model on the publicly available larger Lung1 dataset and then testing it on the smaller proprietary L-RT dataset.

Evaluation of similarity between the datasets and merging

The hypothesis of merging the Lung1 dataset and L-RT dataset was explored. To verify if the two datasets can be merged, the Mann Whitney U test is performed. It is a non-parametric version of the Student *t*-test that can be used to investigate whether two independent samples were selected from populations having the same distribution or not. In fact, we wanted to verify if the hyperparameter search is more stable, and the performances improve by increasing the amount of data. For the Mann-Whitney U test the Python package *SciPy* (v.1.5.2) is used. We choose the two-sided Mann-Whitney U test.

As discussed, the two datasets have some differences regarding the overall stages. Therefore, the Mann-Whitney U test was applied by selecting those subjects of the Lung1 dataset (40 subjects) that have the same overall stage (I and II) of the subjects of the L-RT dataset. For each feature, the Mann-Whitney U test is performed. The dataset obtained by merging the two datasets will be referred to as TOTAL-L, and the within sample CV evaluation scheme on this merged sample will be referred to as MS-TT. Because of the dataset merging and the consequent increase in the amount of data available, we expect the hyperparameter optimization procedure to be more stable and the performance to improve.

Cross validation between public and private datasets

Another interesting aspect that can be investigated on the Lung1 and the L-RT datasets concerns how well a pipeline trained on one of the two datasets performs on the other one, which is referred in this paper as cross-sample train and test (CS-TT). This situation simulates a real application in which an algorithm trained, for instance, on available public data, is used to make predictions on the data collected in a particular clinical structure.

The optimization of the hyperparameters is performed on the training set using a Grid Search Cross-Validation (GSCV) with k = 5 and the AUC metric. Then, the optimal pipeline found was tested on the other dataset. This process was repeated 10 times in GSCV, by randomly assigning the example to either the train or the test sets. The mean and the standard deviation of the scores obtained were calculated.

Intra-dataset and between-dataset cross validation to predict tumor histology and stage

The histology and overall classification tasks are addressed by considering an intra-dataset (both WS-TT and MS-TT) and a between-dataset (CS-TT) cross validation strategy. However, the overall stage classification is performed by considering only stage I and II, which are the only ones reported in the L-RT dataset.

Regarding the intra-dataset CV strategy for the classification tasks, the nested CV approach for the optimization and evaluation of the pipelines, described in 1.2.4, is applied by considering Lung1, L-RT and TOTAL-L. Moreover, only for the histology classification task, the same analysis is performed on a subset of TOTAL-L obtained by considering only the patterns that have overall stage I and II and known histology (74 out of 164 subjects). The CS-TT CV strategy is applied by testing on L-RT the pipelines trained on the Lung1 dataset and *vice versa*, as described in 1.3.2. The results are reported and discussed in section 2.

Results

The results obtained in intra-dataset (both WS-TT and MS-TT) and inter-dataset (CS-TT) cross validation are reported in this section. The results are obtained by considering the nested CV optimization and evaluation strategy for the analysis pipelines composed by the dimensionality reduction and the classification algorithms.

As described, the Mann-Whitney U test is performed to investigate whether the dataset L-RT and Lung1 can be merged. The results suggested that, for 74 over 107 features (applying the Bonferroni correction), the two samples could be considered as drawn from populations with the same distribution. Given this, the two datasets are merged to form the TOTAL-L dataset.

Histology classification

In Table 2, the results obtained in the histology classification task are reported. In particular, the mean value of the AUC and the relative standard deviation obtained with the Random Forest classifier is presented. From these results we can notice that the Random Forest classifier outperforms the chance level (CL) only in the MS-TT case, i.e., if we consider the intra-dataset CV on the Total-L dataset (164 subjects, 107 features) and on the subset of Total-L dataset (74 subjects, 107 features) obtained by considering only the patterns that have overall stage I and II.

Table 2
The mean value of the AUC and the relative standard deviation obtained in Histology classification (3 classes). On the diagonal the performance obtained in the intra-dataset CV (both WS-TT and MS-TT) are reported, whereas the out-of-diagonal performance reported in the first 2 rows and columns of the matrix are referred to the CS-TT strategy. C.L. indicates chance level performances.

Random Forest		**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

In Fig. 3 and in Table A (Supplementary Materials) a comparison between the results obtained by all the classifiers considered for the histology classification task by using the intra-dataset cross validation strategy is shown. From the Table A (Supplementary Materials) we can notice that the performances reported in the second and third column are compatible with a random classifier, in fact the AUC is consistent with 0.5 within the error. Moreover, by comparing these two columns, we can state that there is no performance improvement even merging the two datasets, despite the number of instances has increased.

The results in the first column in Table A are obtained by considering 74 out of 164 subjects of Total-L, discarding those with overall stage IIIa and IIIb. Comparing these results with those in the second and third column, we can notice that the performances obtained in the case where stage IIIa and IIIb are omitted are equal to those obtained with the whole sample within the error.

However, the results of the Welch's unequal variances *t*-test, performed on the scores obtained within the nested CV, indicate that for the Random Forest classifier the null hypothesis of equal averages can be rejected with $p = 0.04$. Therefore, for the Random Forest classifier we can state that the results obtained in the case where stage IIIa and IIIb are omitted are slightly better. Moreover, in this latter case, the performances of the Random Forest classifier (AUC = 0.72 ± 0.11) are significantly better than a random classifier.

The improvement of performances, despite the lower number of instances considered by omitting the stage IIIa and IIIb subjects, is most likely due to the greater homogeneity of the sample, where only stage I and stage II tumors of the three histology categories are represented.

Another important aspect to consider is the presence of overfitting. In fact, despite the use of dimensionality reduction algorithms to prevent it, the results obtained are still affected by overfitting as can be seen in Fig. 4. In fact, the heatmaps reported in Fig. 4 highlight that the performances obtained, in this case with the Random Forest classifier, in the training phase are considerably higher than the ones obtained in the test phase. Moreover, it can be seen in the heatmap that the region characterized by the highest performances in the test phase is totally included in the range of hyperparameter chosen and reported in the supplementary materials.

Identification of the most relevant features

We extracted a rank of the most important features by the Random Forest classifier considering the histology classification task on TOTAL-L with only the patients with stage I and II, the results are shown in Fig. 5. The importance of a feature is measured by the reduction of the Gini index (or entropy) brought by that feature. The higher the reduction, the more important the feature (Fig. 6).

We associated a score from 107 to 1 to each feature based on its importance: the most important feature has the maximum score. As the final score we consider the sum of the scores obtained for each Random Forest trained and optimized via nested CV. This approach was only implemented in the case in which dimensionality reduction by PCA is not chosen in the optimization strategy. The results shown that in this case the most important features are (as named in PyRadiomics): SmallAreaLowGrayLevelEmphasis, ShortRunHighGrayLevelEmphasis, SmallAreaEmphasis, MeshVolume, SizeZoneNonUniformityNormalized.

Overall stage classification

The results obtained in the overall stage classification task for the Random Forest classifier and the SVM linear are shown in Table 3 and Table 4, respectively. In these assessments we considered only instances with overall stage I and II. We can notice that the Random Forest classifier outperforms the chance level (CL) only if we consider the CS-TT CV strategy. The SVM linear classifier is above the CL only if the pipeline is trained on Lung1 and tested on L-RT or the MS-TT case, where the intra-dataset CV strategy is applied by considering the TOTAL-L dataset.

In Fig. 5 and in Table B (Supplementary Materials) a comparison

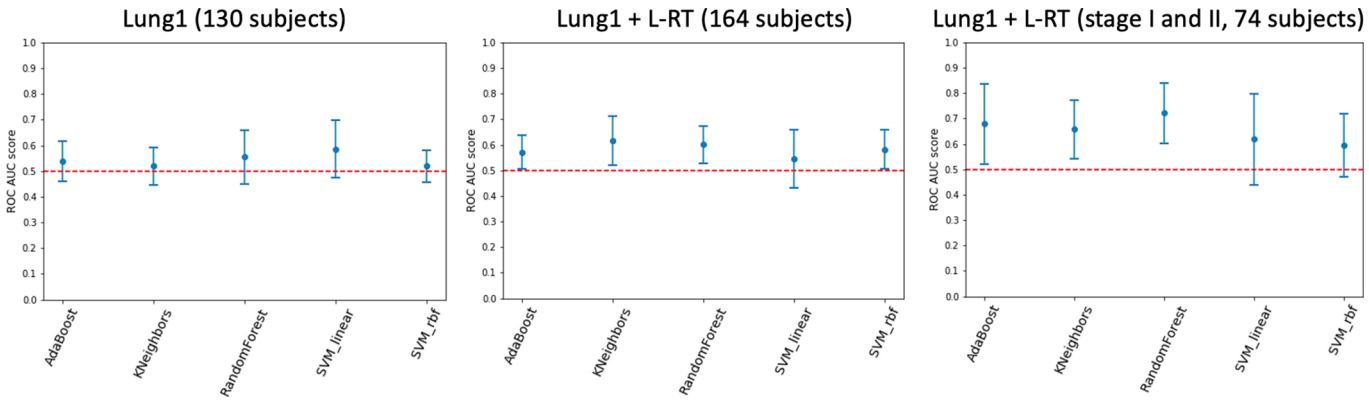

Fig. 3. The mean value of the AUC and the relative standard deviation obtained in Histology classification (3 classes) of Lung1 dataset (130 subjects, 107 features), of Total-L dataset (164 subjects, 107 features) and of the subset of Total-L dataset (74 subjects, 107 features) are shown. The results are obtained with a nested CV.

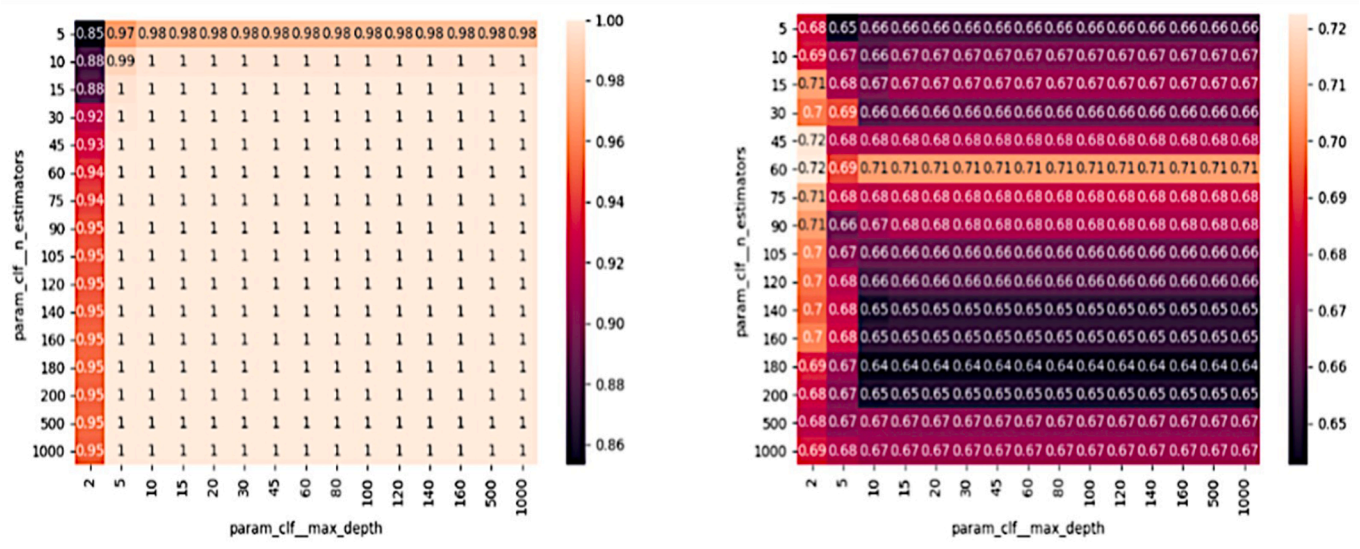

Fig. 4. Heatmaps showing the performances obtained on the training phase (on the left) and on the test phase (on the right) for the random Forest Classifier. In particular, the changes in performances obtained by varying the number of estimators and the max depth of each estimator are reported.

between the results obtained by all the classifiers considered for the overall stage classification task by using the CS-TT cross validation strategy is shown.

From the Table B (Supplementary Materials), we can notice that the performances obtained are equivalent within the uncertainties for all the classifiers except for the KNeighbors and the SVM linear. Considering the other classifiers, the results of the Welch's unequal variances *t*-test, performed on the scores obtained by using the 10 best estimators found, indicate that for the AdaBoost and the SVM-RBF, the null hypothesis of equal averages can be rejected ($p = 0.04$ for the AdaBoost, $p = 0.01$ for the SVM-RBF). Therefore, for all the classifiers except for the Random Forest and the SVM, we can state that the results obtained when training on Lung1 and testing on L-RT are better than those obtained when training on L-RT and testing on Lung1. This difference may be attributable to the different proportions between the two classes in Lung1 and L-RT. In fact, as already stated, the composition of Lung1 and L-RT regarding the overall stage, is that Lung1 is composed of 27 subjects with stage I out of 40, while L-RT is composed of 42 subjects with stage I out of 47. Therefore, algorithms trained on L-RT have only 5 instances from which to learn the features of subjects with overall stage II. By contrast, in Lung1 both the classes are well represented.

Discussion

The workflow that goes from the radiomic features extraction to the development of predictive models based on machine learning techniques was analyzed in this study in the specific case of dealing with small data samples. We focused on the possibility to predict the tumor stage and the tumor histology by considering the radiomic features extracted from the thoracic CT scans of patients with non-small cell lung cancer. This task is addressed by considering the public Lung1 and the proprietary L-RT datasets.

The results obtained regarding the histology classification of NSCLC, performed by considering all labeled tumor stages, are at the chance level. Nevertheless, taking into consideration only the subjects with overall stages I and II, an improvement of performances is observed. In particular, the best performances are reached by the Random Forest classifiers (AUC $= 0.72 \pm 0.11$). Thus, if we restrict our analysis to stage I and stage II tumors, and thus reduce the heterogeneity of cases within the sample, radiomic features extracted from thoracic CT can predict the three different types of histology of NSCLC with a discrete performance.

The results achieved in the overall stage classification of NSCLC, by using Lung1 as training set and L-RT as test-set, are considerably above the random guess. In particular, the best performances are obtained by considering the linear-kernel SVM classifiers (AUC $= 0.84 \pm 0.03$). Thus,

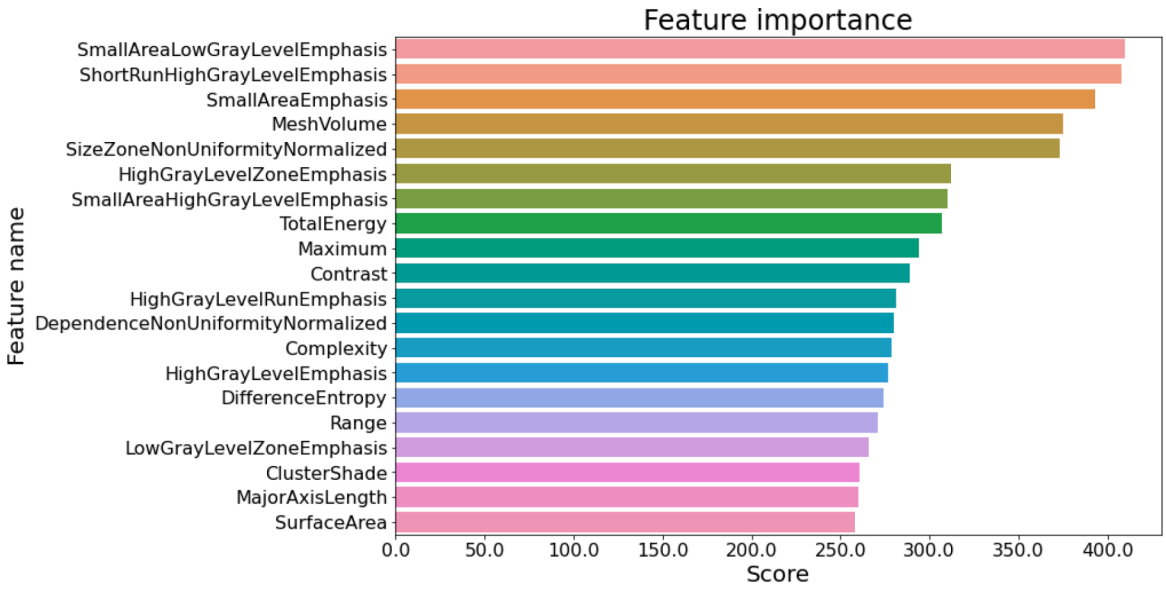

Fig. 5. Barplot showing the most important features selected by the Random Forest classifier considering the histology classification task on TOTAL-L with only patients with stage I and II.

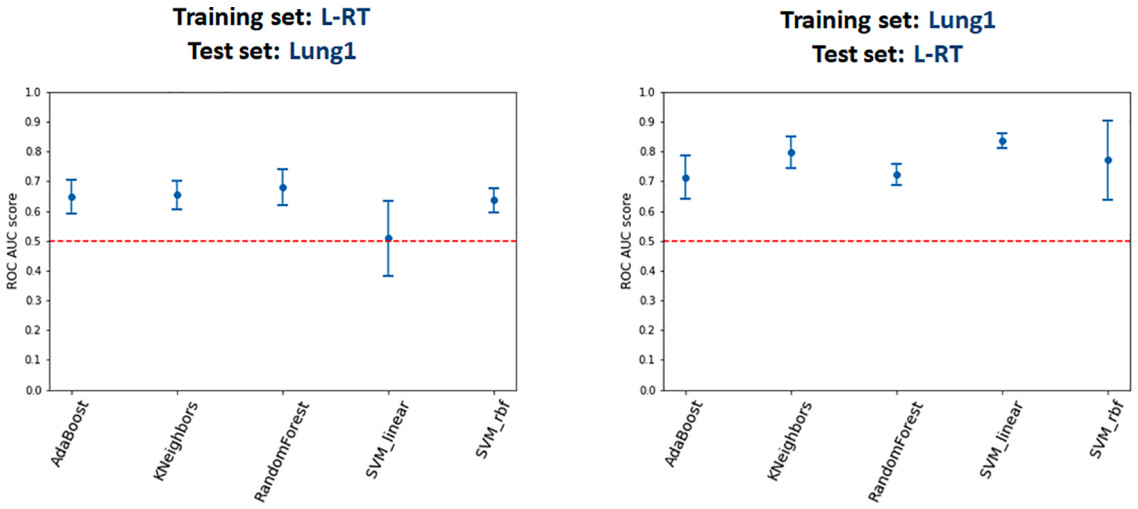

Fig. 6. The mean value of the AUC and the relative standard deviation obtained in Overall stage classification (2 classes) training on Lung1 (40 subjects, 107 features) testing on L-RT (47 subjects, 107 features) and *vice versa* are shown. The results are obtained by averaging the scores obtained on L-RT by the 10 best estimators found with a 5-fold GSCV on Lung1.

Table 3
The mean value of the AUC and the relative standard deviation obtained in Overall stage classification (2 classes) for the Random Forest classifier. On the diagonal the performance obtained in the intra-dataset CV (both WS-TT and MS-TT) are reported, whereas the out-of-diagonal performance reported in the first 2 rows and columns of the matrix are referred to the CS-TT strategy. C.L. indicates chance level performances.

Random Forest		TEST SET		
OS classification (only I and II)		L-RT	Lung1	Total-L
TRAIN SET	L-RT	C.L.	0.68 ± 0.06	//
	Lung1	0.72 ± 0.04	C.L.	//
	Total-L	//	//	C.L.

Table 4
The mean value of the AUC and the relative standard deviation obtained in Overall stage classification (2 classes) for the SVM linear classifier. On the diagonal the performance obtained in the intra-dataset CV (both WS-TT and MS-TT) are reported, whereas the out-of-diagonal performance reported in the first 2 rows and columns of the matrix are referred to the CS-TT strategy. "C.L." indicates chance level performances.

SVM linear		TEST SET		
OS classification (only I and II)		L-RT	Lung1	Total-L
TRAIN SET	L-RT	C.L.	C.L.	//
	Lung1	0.84 ± 0.03	C.L.	//
	Total-L	//	//	0.78 ± 0.13

regarding the stage prediction, and training the models on the Lung1 dataset and testing on the L-RT dataset, it is possible to predict the overall stage of NSCLC with good performance.

These promising results imply that the approach developed in this work could provide support for tumor analysis in terms of stage identification and histology classification for NSCLC cases.

In literature there are many examples of the application of radiomics in lung cancer study. In the work by Aerts et al. [24], the stability of

radiomic features extracted from CT images, and their prognostic power are investigated by considering different datasets of lung and head-neck cancer. They selected the single best performing radiomic feature from each group of features in order to remove redundancy within the radiomic information. In this way they create a more stable prognostic radiomic signature. In the study by Patil et al. [51], a dataset composed of 317 non-small cell lung cancer subjects is considered. The authors investigated the predictive power of features extracted from thoracic CT in a histology classification task. The sensitivity, specificity and accuracy obtained in determining the histology using radiomic features are 87%, 89% and 88%, respectively. Despite a direct comparison with our results is not feasible due to the different sample size and composition, the results reported by Patil et al. [51] highlight how radiomic features have high predictive power in tumor histology classification, suggesting the possibility to allow diagnosis without tissue biopsy in high-risk patients or in which histological characterization is not possible, due to the patient's comorbidities. In these cases, in our center, positive metabolic imaging was considered a substitute for the diagnosis of malignancy, according to the literature [52]. Histology also could be considered in the GTV delineation and in the choice of treatment to customize the radiation dose for stereotactic ablative body radiotherapy [53]. For example, higher prescription doses could be considered for squamous cell carcinomas, within normal tissue tolerances, where a lower BED translates to a worse outcome [54,55]. We also suppose that radiomics could help differentiate between post-radiotherapy benign changes and residual tumor tissues [56] or predict which patients are more prone to develop treatment-related adverse events [57]. Finally, radiomics could be a powerful tool in providing other biological and genomics tumor characterization, translating in a most accurate treatment [58].

Despite the encouraging results that we obtained, the ML-based decision system we have developed is not yet ready to be applied in clinical workflows. Before application-specific ML algorithms can be successfully translated into clinics, a number of challenges must indeed be overcome, including the harmonization of different data samples [7,59], the reliability and reproducibility of the results [7,10,60] the interpretability of the models and the results [59,61], and the compliance with current regulations [62].

Apart from the regulatory issues, we addressed in this study the main challenges listed above that become even more severe when dealing with small datasets, where we typically have more features per subject than subjects in the dataset.

Multisite data harmonization

Harmonizing multi-site data is important because the quality of radiomic features, their association with data and therefore the performance of models created using these features are related to image properties and quality, which can be different according to the acquisition site. Images collected in clinical routine work reveal a wide variation of acquisition parameters. Regarding CT imaging, different reconstruction algorithms and parameters are used in clinical practice. This great variability affects the values of the image, and consequently the values of the radiomic features. In this way, it might happen that, comparing features extracted from images acquired using different acquisition protocols, their differences could be due to different image properties rather than to different biological properties of tissues [1]. Another critical issue in the radiomic process is the segmentation step. Features are extracted from segmented volumes and, since many tumors show unclear borders, the segmentation task can be very challenging and have a strong impact on final predictions [1].

In this study, we did not implement pre-processing algorithms to harmonize raw CT images prior to feature extraction, as we evaluated that the acquisition characteristics of the CT images were quite similar between the two data sets (slice thickness of 2.5–3 mm and tube voltage of 120–140 kVp) for both data samples. To investigate whether the two independent samples of extracted radiomic features were selected from populations having the same distribution, we implemented the Mann Whitney U test, whose Bonferroni-corrected results demonstrated that for most features the null hypothesis was true.

Reliability and reproducibility of the results

Interesting review papers reported on the repeatability and reproducibility of the results of medical data analyses [10,60], with particular reference to radiomic features [10]. In the latter work, the authors analyzed 41 studies (35 human studies and 6 phantom studies) focusing on the assessment of the repeatability and reproducibility of radiomic features. Studies with different types of cancer (NSCLC, lung cancer, oropharyngeal cancer, esophageal cancer, rectal cancer, breast cancer, cervical cancer, and various solid tumors) and different imaging modalities (CT, Cone Beam CT, positron emission tomography, and magnetic resonance) were considered. In this review the authors considered different factors that could influence the repeatability and reproducibility of radiomic features, for example: the image acquisition settings, the image reconstruction algorithm, the digital image preprocessing and the software used to extract radiomic features. The authors found that the First Order Statistic Features are more reproducible than the Size- and Shape-based Features and the Higher Order Statistics features.

Especially when data samples are limited in size and a large amount of features are analyzed with ML algorithms, beside the problem of model overfitting, there is the problem of the stability of the results. The problem of model overfitting, which can easily occur with small and high-dimensional datasets, limits the generalization capability of a ML model, whereas the performance instability prevents the identification of a unique set of optimized hyperparameters for ML algorithms. Both issues lead to a low reliability and reproducibility of ML results.

In this study, the hyperparameter optimization of the pipelines, performed through an exhaustive search, turned out to be unstable. Therefore, we could not find a single best optimization of the algorithms and the results were not very stable. To overcome this issue, the performances were evaluated through a nested CV scheme, which allows to obtain unbiased assessments, and thus reliable and reproducible results within their uncertainty.

In the work of Patil et al. [51], in which the histology of NSCLC starting from radiomic features is studied, the authors did not implement a nested CV strategy but they used a simple CV that could be subject to biases. In fact, the hyperparameter optimization implemented by a non-nested CV scheme involves the use of the same data in optimizing the model hyperparameters and in evaluating the model performance; this situation could produce over-optimistic results [7].

In general, if the available dataset is not sufficiently large, it is not recommended to divide it into a training and test set. In fact, if the test set is too small, we may have to deal with large statistical fluctuations in the estimation of the test performances. As a result, it can be difficult to compare different algorithms on a given task. In this situation, it is convenient to evaluate the performances through a cross validation strategy that allows us to consider all the instances of the dataset. These procedures generally increase the computational cost, as they are based on many repetitions of the training and testing phases.

Interpretability and explainability of results

It is widely recognized that ML-based algorithms in order to be receivable by the clinical community and usefully implemented in clinical workflows should provide human intelligible results and rely on explainable methodology [59,61]. Complex models generally provide better performance at the expenses of their interpretability. In the field of lung cancer diagnosis, for instance, a recent work by Astraki et al. [63] proposed the implementation of Random Forest classifier to distinguish benign from malignant nodules. The descriptive features they computed on both nodules and background parenchymal tissue were extracted by a convolutional neural network. This approach, which

showed very high discrimination performance (AUC > 90%), does not allow a direct identification of which image features are responsible for the classification results. In the work of Patil et al. [51], which is closer to ours in terms of objectives, the authors implemented an SVM classifier to classify the histology of NSCLC from radiomic features. In this case, the use of a conceptually easy to explain ML algorithm, such as the SVM, allowed authors to rank the features by importance and discover the most discriminative ones.

In our work, we emphasized that one advantage of using a Random Forest classifier is that it allows the identification of the most important features that contributed to the classification. This characteristic is a peculiarity of some traditional ML classifiers, including those based on linear models (e.g. linear-kernel Support Vector Machines, Linear Regression, Linear Discriminant Analysis) and decision trees (e.g. Random Forest). The possibility to rank the features and to identify those more relevant for a given classification task automatically allows an interpretation of the results achieved, which is crucial for clinicians to trust ML-based decisional systems.

Conclusions

In conclusion, radiomics and ML approaches are becoming widespread within the Medical Physics community. However, the availability of small, annotated data samples may limit the impact of this rapidly growing field of research. We demonstrated in this study that small data cohorts can be integrated with public data samples to achieve more reliable performance in radiomics and ML studies. Particular attention must be paid to evaluating the compatibility between the two cohorts before merging them, and to carry out an unbiased evaluation of the classification performance by means of a cross-validation strategy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work has been carried out within the Artificial Intelligence in Medicine (AIM) project funded by INFN (CSN5, 2019-2021), https://www.pi.infn.it/aim.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2021.08.015.

References

[1] Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp 2018;2(1). https://doi.org/10.1186/s41747-018-0068-z.
[2] Avanzo M, Wei L, Stancanello J, Vallières M, Rao A, Morin O, et al. Machine and deep learning methods for radiomics. Med Phys 2020;47(5). https://doi.org/10.1002/mp.v47.510.1002/mp.13678.
[3] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. Radiology 2016;278(2):563–77. https://doi.org/10.1148/radiol.2015151169.
[4] Hrnjica B, Danandeh Mehr A. Optimized genetic programming applications 2018: 310. https://doi.org/10.4018/978-1-5225-6005-0.
[5] Kortesniemi M, Tsapaki V, Trianni A, Russo P, Maas Ad, Källman H-E, et al. The European Federation of Organisations for Medical Physics (EFOMP) White Paper: Big data and deep learning in medical imaging and in relation to medical physics profession. Phys Med 2018;56:90–3. https://doi.org/10.1016/j.ejmp.2018.11.005.
[6] Diaz O, Guidi G, Ivashchenko O, Colgan N, Zanca F. Artificial intelligence in the medical physics community: an international survey. Phys Med 2021;81:141–6. https://doi.org/10.1016/j.ejmp.2020.11.037.
[7] Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, et al. AI applications to medical images: From machine learning to deep learning. Phys Med 2021;83:9–24. https://doi.org/10.1016/j.ejmp.2021.02.006.
[8] Avanzo M, Porzio M, Lorenzon L, Milan L, Sghedoni R, Russo G, et al. Artificial intelligence applications in medical imaging: a review of the medical physics research in Italy. Phys Med 2021;83:221–41. https://doi.org/10.1016/j.ejmp.2021.04.010.
[9] Manco L, Maffei N, Strolin S, Vichi S, Bottazzi L, Strigari L. Basic of machine learning and deep learning in imaging for medical physicists. Phys Med 2021;83:194–205. https://doi.org/10.1016/j.ejmp.2021.03.026.
[10] Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys 2018;102(4):1143–58. https://doi.org/10.1016/j.ijrobp.2018.05.053.
[11] Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 2019;144(8):1941–53. https://doi.org/10.1002/ijc.v144.810.1002/ijc.31937.
[12] Brawley OW. Avoidable cancer deaths globally. CA Cancer J Clin 2011;61(2):67–8. https://doi.org/10.3322/caac.v61:210.3322/caac.20108.
[13] Chansky K, Sculier J-P, Crowley JJ, Giroux D, Van Meerbeeck J, Goldstraw P. The international association for the study of lung cancer staging project: Prognostic factors and pathologic TNM stage in surgically managed non-small cell lung cancer. J Thorac Oncol 2009;4(7):792–801. https://doi.org/10.1097/JTO.0b013e3181a7716e.
[14] Scott WJ, Howington J, Feigenberg S, Movsas B, Pisters K. Treatment of non-small cell lung cancer stage I and stage II: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007;132:234S–242S. https://doi.org/10.1378/chest.07-1378.
[15] Crinò L, Weder W, van Meerbeeck J, Felip E. Early stage and locally advanced (non-metastatic) non-small-cell lung cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann Oncol 2010;21:103–15. https://doi.org/10.1093/annonc/mdq207.
[16] Vansteenkiste J, Crinò L, Dooms C, Douillard JY, Faivre-Finn C, Lim E, et al. 2nd ESMO consensus conference on lung cancer: early-stage non-small-cell lung cancer consensus on diagnosis, treatment and follow-up. Ann Oncol 2014;25(8):1462–74. https://doi.org/10.1093/annonc/mdu089.
[17] Ost D, Goldberg J, Rolnitzky L, Rom WN. Survival after surgery in stage IA and IB non-small cell lung cancer. Am J Respir Crit Care Med 2008;177(5):516–23. https://doi.org/10.1164/rccm.200706-815OC.
[18] Halabi S, Lin C-Y, Kelly WK, Fizazi KS, Moul JW, Kaplan EB, et al. Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. J Clin Oncol 2014;32(7):671–7. https://doi.org/10.1200/JCO.2013.52.3696.
[19] Zhang J-X, Song Wu, Chen Z-H, Wei J-H, Liao Y-J, Lei J, et al. Prognostic and predictive value of a microRNA signature in stage II colon cancer: A microRNA expression analysis. Lancet Oncol 2013;14(13):1295–306. https://doi.org/10.1016/S1470-2045(13)70491-1.
[20] Tran B, Dancey JE, Kamel-Reid S, McPherson JD, Bedard PL, Brown AMK, et al. Cancer genomics: technology, discovery, and translation. J Clin Oncol 2012;30(6):647–60. https://doi.org/10.1200/JCO.2011.39.2316.
[21] Hofman V, Ilie M, Long E, Lassalle S, Butori C, Bence C, et al. Immunohistochemistry and personalised medicine in lung oncology: advantages and limitations. Bull Cancer 2014;101:958–65. https://doi.org/10.1684/bdc.2014.2041.
[22] Cuccia F, Mortellaro G, Mazzola R, Donofrio A, Valenti V, Tripoli A, et al. Prognostic value of two geriatric screening tools in a cohort of older patients with early stage Non-Small Cell Lung Cancer treated with hypofractionated stereotactic radiotherapy. J Geriatr Oncol 2020;11(3):475–81. https://doi.org/10.1016/j.jgo.2019.05.002.
[23] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48(4):441–6. https://doi.org/10.1016/j.ejca.2011.11.036.
[24] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5(1). https://doi.org/10.1038/ncomms5006.
[25] Yip SSF, Aerts HJWL. Applications and limitations of radiomics. Phys Med Biol 2016;61(13):R150–66. https://doi.org/10.1088/0031-9155/61/13/R150.
[26] Muren LP, Thwaites DI. The on-going quest for treatment precision and conformality in radiotherapy. Radiother Oncol 2013;109(3):337–41. https://doi.org/10.1016/j.radonc.2013.11.007.
[27] Baumann P, Nyman J, Hoyer M, Wennberg B, Gagliardi G, Lax I, et al. Outcome in a prospective phase II trial of medically inoperable stage I non-small-cell lung cancer patients treated with stereotactic body radiotherapy. J Clin Oncol 2009;27(20):3290–6. https://doi.org/10.1200/JCO.2008.21.5681.
[28] Chi A, Liao Z, Nguyen NP, Xu J, Stea B, Komaki R. Systemic review of the patterns of failure following stereotactic body radiation therapy in early-stage non-small-cell lung cancer: clinical implications. Radiother Oncol 2010;94(1):1–11. https://doi.org/10.1016/j.radonc.2009.12.008.
[29] Rusthoven KE, Pugh TJ. Stereotactic body radiation therapy for inoperable lung cancer. JAMA - J Am Med Assoc 2010;303:2354–5. https://doi.org/10.1001/jama.2010.777.
[30] Onishi H, Araki T, Shirato H, Nagata Y, Hiraoka M, Gomi K, et al. Stereotactic hypofractionated high-dose irradiation for stage I nonsmall cell lung carcinoma: clinical outcomes in 245 subjects in a Japanese multiinstitutional study. Cancer 2004;101(7):1623–31. https://doi.org/10.1002/(ISSN)1097-014210.1002/cncr.v101:710.1002/cncr.20539.

[31] Ganeshan B, Abaleke S, Young RCD, Chatwin CR, Miles KA. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. Cancer Imaging 2010;10(1):137–43. https://doi.org/10.1102/1470-7330.2010.0021.
[32] Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. Front Oncol 2016;6. https://doi.org/10.3389/fonc.2016.00071.
[33] Edge SB, Compton CC. The american joint committee on cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol 2010;17:1471–4. https://doi.org/10.1245/s10434-010-0985-4.
[34] NCCN Guidelines Version 7.2019 Non-small cell lung cancer. 2019.
[35] Arcangeli S, Agolli L, Portalone L, Migliorino MR, Lopergolo MG, Monaco A, et al. Patterns of CT lung injury and toxicity after stereotactic radiotherapy delivered with helical tomotherapy in early stage medically inoperable NSCLC. Br J Radiol 2015;88(1048):20140728. https://doi.org/10.1259/bjr.20140728.
[36] Franks KN, Jain P, Snee MP. Stereotactic ablative body radiotherapy for lung cancer. Clin Oncol 2015;27(5):280–9. https://doi.org/10.1016/j.clon.2015.01.006.
[37] Chang JY, Li Q-Q, Xu Q-Y, Allen PK, Rebueno N, Gomez DR, et al. Stereotactic ablative radiation therapy for centrally located early stage or isolated parenchymal recurrences of non-small cell lung cancer: how to fly in a "No Fly Zone". Int J Radiat Oncol 2014;88(5):1120–8. https://doi.org/10.1016/j.ijrobp.2014.01.022.
[38] Senthi S, Haasbeek CJA, Slotman BJ, Senan S. Outcomes of stereotactic ablative radiotherapy for central lung tumours: a systematic review. Radiother Oncol 2013;106(3):276–82. https://doi.org/10.1016/j.radonc.2013.01.004.
[39] Figlia V, Mazzola R, Cuccia F, Alongi F, Mortellaro G, Cespuglio D, et al. Hypofractionated stereotactic radiation therapy for lung malignancies by means of helical tomotherapy: report of feasibility by a single-center experience. Radiol Medica 2018;123(6):406–14. https://doi.org/10.1007/s11547-018-0858-7.
[40] NSCLC-Radiomics - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki n.d. https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics (accessed February 12, 2021).
[41] Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging 2012;30(9):1234–48. https://doi.org/10.1016/j.mri.2012.06.010.
[42] van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.
[43] Kikinis R, Pieper SD, Vosburgh KG. 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. Intraoperative Imaging Image-Guided Ther., Springer New York; 2014, p. 277–89. https://doi.org/10.1007/978-1-4614-7657-3_19.
[44] Pieper S, Halle M, Kikinis R. 3D Slicer. 2004 2nd IEEE Int. Symp. Biomed. Imaging Macro to Nano, vol. 1, 2004, p. 632–5. https://doi.org/10.1109/isbi.2004.1398617.
[45] sklearn.pipeline.Pipeline — scikit-learn 0.24.1 documentation n.d. https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html (accessed February 12, 2021).
[46] Aurélien Géron. Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems. 2019.
[47] Raschka S, Mirjalili V. Python Machine learning. Machine learning and Deep Learning with Python, scikit-learn, and tensorflow. 2017.
[48] Bishop C. Pattern Recognition and Machine Learning. 2006.
[49] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[50] Nested cross validation explained - Weina Jin, MD n.d. https://weina.me/nested-cross-validation/ (accessed February 12, 2021).
[51] Patil R, Mahadevaiah G, Dekker A. An approach toward automatic classification of tumor histopathology of non-small cell lung cancer based on radiomic features. Tomography 2016;2:374–7. https://doi.org/10.18383/j.tom.2016.00244.
[52] Louie Alexander V, Senan Suresh, Patel Pretesh, Ferket Bart S, Lagerwaard Frank J, Rodrigues George B, et al. When is a biopsy-proven diagnosis necessary before stereotactic ablative radiotherapy for lung cancer? A decision analysis. Chest 2014;146(4):1021–8. https://doi.org/10.1378/chest.13-2924.
[53] Shiue Kevin, Cerra-Franco Alberto, Shapiro Ronald, Estabrook Neil, Mannina Edward M, Deig Christopher R, et al. Histology, tumor volume, and radiation dose predict outcomes in NSCLC patients after stereotactic ablative radiotherapy. J Thorac Oncol 2018;13(10):1549–59. https://doi.org/10.1016/j.jtho.2018.06.007.
[54] Woody Neil M, Stephans Kevin L, Andrews Martin, Zhuang Tingliang, Gopal Priyanka, Xia Ping, et al. A histologic basis for the efficacy of SBRT to the lung. J Thorac Oncol 2017;12(3):510–9. https://doi.org/10.1016/j.jtho.2016.11.002.
[55] Baine Michael J, Verma Vivek, Schonewolf Caitlin A, Lin Chi, Simone Charles B. Histology significantly affects recurrence and survival following SBRT for early stage non-small cell lung cancer. Lung Cancer 2018;118:20–6. https://doi.org/10.1016/j.lungcan.2018.01.021.
[56] Vadalà RE, Santacaterina A, Sindoni A, Platania A, Arcudi A, Ferini G, et al. Stereotactic body radiotherapy in non-operable lung cancer patients. Clin Transl Oncol 2016;18(11):1158–9. https://doi.org/10.1007/s12094-016-1552-7.
[57] Ferini G, Pergolizzi S. A ten-year-long update on radiation proctitis among prostate cancer patients treated with curative external beam radiotherapy. Vivo (Brooklyn) 2021;35:1379–91. https://doi.org/10.21873/invivo.12390.
[58] Nardone Valerio, Tini Paolo, Pastina Pierpaolo, Botta Cirino, Reginelli Alfonso, Carbone Salvatore, et al. Radiomics predicts survival of patients with advanced non-small cell lung cancer undergoing PD-1 blockade using Nivolumab. Oncol Lett 2020. https://doi.org/10.3892/ol10.3892/ol.2019.11220.
[59] Papadimitroulas Panagiotis, Brocki Lennart, Christopher Chung Neo, Marchadour Wistan, Vermet Franck, Gaubert Laurent, et al. Artificial intelligence: deep learning in oncological radiomics and challenges of interpretability and data harmonization. Phys Med 2021;83:108–21. https://doi.org/10.1016/j.ejmp.2021.03.009.
[60] Balagurunathan Y, Mitchell R, El Naqa I. Requirements and reliability of AI in the medical context. Phys Med 2021;83:72–8. https://doi.org/10.1016/j.ejmp.2021.02.024.
[61] Vellido Alfredo. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl 2020;32(24):18069–83. https://doi.org/10.1007/s00521-019-04051-w.
[62] Beckers R, Kwade Z, Zanca F. The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics. Phys Med 2021;83:1–8. https://doi.org/10.1016/j.ejmp.2021.02.011.
[63] Astaraki M, Zakko Y, Toma Dasu I, Smedby Ö, Wang C. Benign-malignant pulmonary nodule classification in low-dose CT with convolutional features. Phys Med 2021;83:146–53. https://doi.org/10.1016/j.ejmp.2021.03.013.