applied sciences

MDPI

Article

A Machine Learning and Radiomics Approach in Lung Cancer for Predicting Histological Subtype

Antonio Brunetti [1,2,*], Nicola Altini [1], Domenico Buongiorno [1,2], Emilio Garolla [3], Fabio Corallo [3], Matteo Gravina [3], Vitoantonio Bevilacqua [1,2] and Berardino Prencipe [1]

[1] Department of Electrical and Information Engineering, Polytechnic University of Bari, Via Orabona 4, 70126 Bari, Italy; nicola.altini@poliba.it (N.A.); domenico.buongiorno@poliba.it (D.B.); vitoantonio.bevilacqua@poliba.it (V.B.); berardino.prencipe@poliba.it (B.P.)
[2] Apulian Bioengineering SRL, Via delle Violette 14, 70026 Modugno, Italy
[3] Department of Medical and Surgical Sciences, University of Foggia, Viale Pinto 1, 71122 Foggia, Italy; emiliogarolla@gmail.com (E.G.); fabiocor@hotmail.it (F.C.); matteogravina@inwind.it (M.G.)
* Correspondence: antonio.brunetti@poliba.it

Abstract: Lung cancer is one of the deadliest diseases worldwide. Computed Tomography (CT) images are a powerful tool for investigating the structure and texture of lung nodules. For a long time, trained radiologists have performed the grading and staging of cancer severity by relying on radiographic images. Recently, radiomics has been changing the traditional workflow for lung cancer staging by providing the technical and methodological means to analytically quantify lesions so that more accurate predictions could be performed while reducing the time required from each specialist to perform such tasks. In this work, we implemented a pipeline for identifying a radiomic signature composed of a reduced number of features to discriminate between adenocarcinomas and other cancer types. In addition, we also investigated the reproducibility of this radiomic study analysing the performances of the classification models on external validation data. In detail, we first considered two publicly available datasets, namely D1 and D2, composed of $n = 262$ and $n = 89$ samples, respectively. Ten significant features, according to univariate AUC evaluated on D1, were retained. Mann–Whitney *U* tests recognised three of these features to have a statistically different distribution, with a p-value < 0.05. Then, we collected $n = 51$ CT images from patients with lung nodules at the *Azienda Ospedaliero—Universitaria "Policlinico Riuniti"* in Foggia. Resident radiologists manually annotated the lung lesions in images to allow the subsequent analysis of the malignancy regions. We designed a pipeline for feature extraction from the Volumes of Interest in order to generate a third dataset, i.e., D3. Several experiments have been performed showing that the selected radiomic signature not only allowed the discrimination of lung adenocarcinoma from other cancer types independently from the input dataset used for training the models, but also allowed reaching good classification performances also on external validation data; in fact, the radiomic signature computed on D1 and evaluated on the local cohort allowed reaching an AUC of 0.70 ($p < 0.001$) for the task of predicting the histological subtype.

Keywords: radiomics; lung carcinoma; histological subtype; machine learning

Citation: Brunetti, A.; Altini, N.; Buongiorno, D.; Garolla, E.; Corallo, F.; Gravina, M.; Bevilacqua, V.; Prencipe, B. A Machine Learning and Radiomics Approach in Lung Cancer for Predicting Histological Subtype. *Appl. Sci.* **2022**, *12*, 5829. https://doi.org/10.3390/app12125829

Academic Editors: Cecilia Di Ruberto, Andrea Loddo, Lorenzo Putzu, Alessandro Stefano and Albert Comelli

Received: 6 May 2022
Accepted: 6 June 2022
Published: 8 June 2022

1. Introduction

The Global Cancer Observatory estimated that lung cancer was the leading cause of cancer death in the world's population [1]. Phenotyping lung cancer, or cancer in general, has been demonstrated to be crucial for clinical practice and medical research; in fact, in recent years, a profound effort in designing and employing computational solutions for phenotyping pathologies has been made [2–4].

Nowadays, such characterisation of pathologies is necessary to move toward the so-called Precision Medicine. It is a modern paradigm for diagnosing or treating cancers, or diseases in general, based on the identification and characterisation of pathology-specific

characteristics, taking into account the individual variability, i.e., the genetic factors of each subject, their lifestyle and living environment. Made possible also by technical advances in computational sciences, Precision Medicine is definitely revolutionising the healthcare domain. This approach to cancer handling, in fact, aims to precisely target diseases, including lung carcinoma, with a per-subject approach [5].

Medical imaging methodologies, including Computed Tomography (CT), Positron Emission Tomography (PET), or Magnetic Resonance (MR), are crucial for performing clinical tasks related to cancer diagnosis, treatment, or follow up [6–8]. Medical imaging has already been demonstrated to be essential in the Precision Medicine framework thanks to its capability to improve the knowledge about the clinical phenomenon under consideration (regardless of its nature) [9]. In addition, a considerable amount of recent literature has already shown the capabilities of functional imaging methods to build an in-vivo representation of tumour processes from a biological perspective [10]. This characteristic made imaging methods good candidates for allowing the identification of biosignatures for phenotyping the tumour.

Radiomics is an emerging method based on algorithms for data characterization, which allows the extraction of a large number of features from medical images. By exploiting the information carried out by such features, radiomics approaches aim to uncover and quantitatively describe tumoral patterns and characteristics, otherwise not observable through traditional algorithms for image analysis [11–15]. This new perspective for extracting phenotypic information from imaging, which is already performed for biomedical signals, such as electromyography or electroencephalography [16,17], may thus provide valuable information for setting up personalised approaches for therapies.

Several authors carried out radiomics-based studies in the context of characterising solid cancers [18,19] or lesions [20] in the lung region. It also allowed the design and implementation of several applications in oncology, such as solid cancer, glioblastoma [21], hepatocellular carcinoma [22] and breast cancer [23] classification, among many others.

While the description of images by a large number of features may help in the comprehension of the underlying phenomena, it may also lead to several drawbacks. In fact, due to the high dimensionality of radiomic features, dimensionality reduction and clustering techniques may be needed to improve the classification and generalisation capabilities of automatic systems for supporting decisions [24].

Parmar et al. performed analyses to extract clusters of radiomic features and prognostic signatures specific for lung and head and neck (H&N) cancers [25]. In their work, the authors performed consensus clustering before classifying tumour phenotypes, revealing eleven stable clusters of radiomic features for lung tumour classification, also showing associations with clinical parameters.

Besides the high dimensionality of data, classification approaches based on radiomics suffer from other data-related problems, including the "big-p, little-n" problem, feature redundancy, and class unbalancing [24]. Zhang et al., trying to address such problems, compared many feature selection techniques and predictive models to improve the radiomics-based prognosis of patients with non-small cell lung cancer (NSCLC) [26]. With respect to the problem of feature redundancy, their analysis showed that Random Forests (RF) were the optimal predictive model, whereas Principal Component Analysis (PCA) was the optimal feature selection method. The Synthetic Minority Over-sampling (SMOTE) technique was employed to mitigate the problem class unbalancing, significantly increasing the predictive accuracy.

In this work, we designed and implemented a quantitative approach based on a radiomics pipeline to classify lung nodules between adenocarcinoma (LUAD) cases and other histological classes from unenhanced CT images. The overall radiomics pipeline consists of the following stages, as reported in Figure 1. First, we collected CT images of patients with lung cancer in a cohort from the *Azienda Ospedaliero—Universitaria "Policlinico Riuniti"* in Foggia, Italy. Images were manually segmented by expert radiologists to identify and extract the Regions of Interest (ROIs) to process. Radiomic features were then extracted from the ROIs. To do

this, we used `PyRadiomics`, an open-source python package for the extraction of radiomics features from medical images [27], which also has the advantage of increasing reproducibility among different studies, thanks to the adoption of the Imaging Biomarker Standardization Initiative (IBSI) [13], to define and compute the features. Statistical analyses for selecting features based on their discriminative power were accomplished, and predictive models were set up on the extracted feature data to make the decision on the histological subtype. Finally, external data were considered to validate the predictive model.

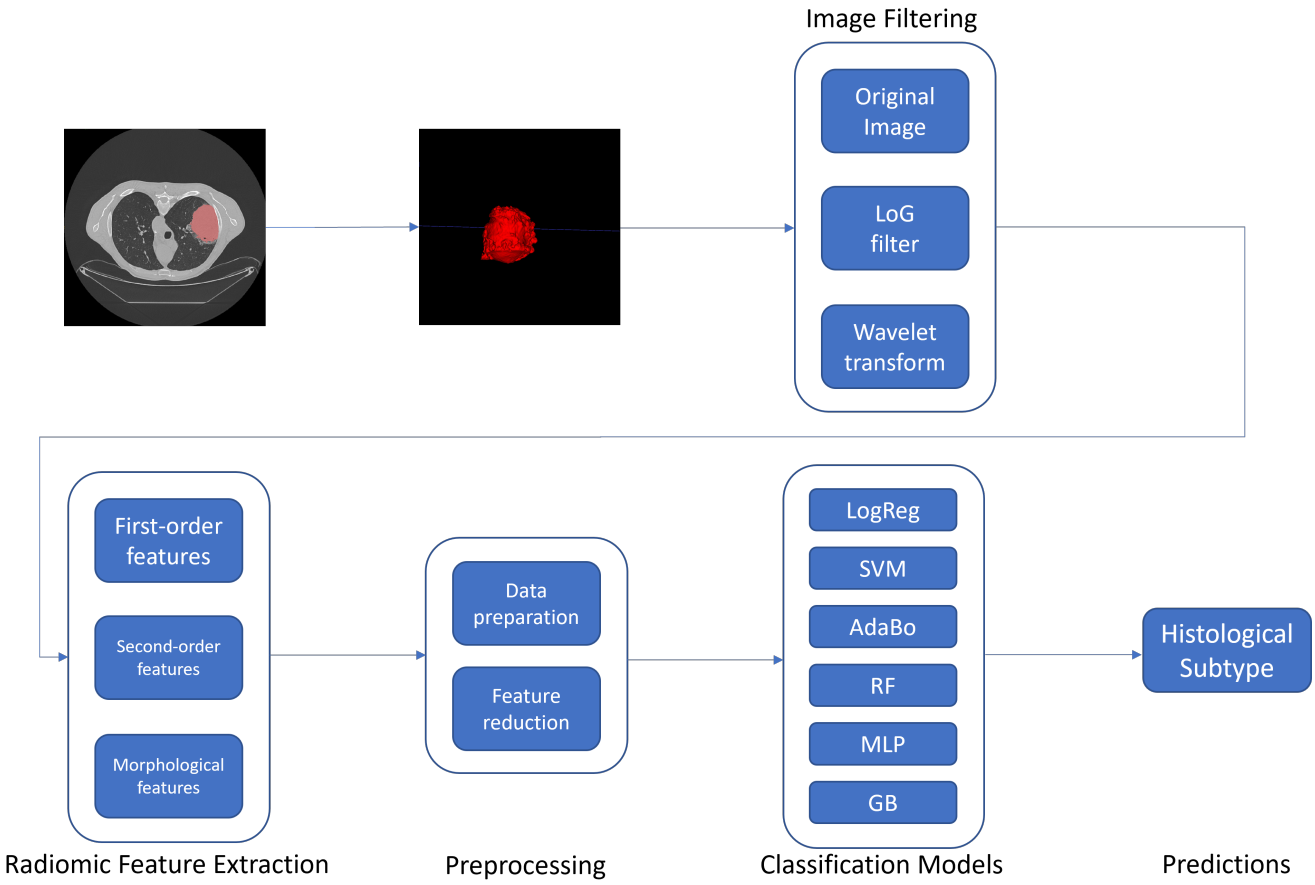

Figure 1. Radiomics workflow. The complete workflow includes the segmentation of the lung region, filtering to enhance the image, extraction of radiomics features, classification models and statistical analysis.

Concerning the task of lung cancer phenotyping, several works could be found in the literature. For example, Ferreira-Junior et al. studied which quantitative features coming from contrast-enhanced CT scans would be more helpful in performing associations with histopathological data, such as the histological subtype classification [28]. However, contrast-enhanced imaging analyses include, albeit minimal, risk factors also related to the injection of the contrast medium. To overcome this, in this work, we analyse radiomic features on unenhanced CT scans for making decisions.

Linning et al. also included unenhanced CT scans in their work [29]. To demonstrate the validity of their approach, the authors performed a ten-fold cross-validation. However, validating the models on external and independent datasets is crucial in radiomics studies [30]. Instead, a similar work, which validated predictive models considering external data, was the one by Wu et al. [31]. The authors, in fact, developed a Naïve Bayes classifier and validated it with external datasets. However, in their work, Wu et al. considered images acquired in the Netherlands only. In our work, instead, external validation is performed using image data obtained in different countries.

To do this, we included in our analysis data from two cohorts employed in the work by Grossmann et al., who made it publicly available in terms of clinical data, radiomics features and genetic characteristics [32]. Grossmann et al. discovered a relationship between imaging features, immune response, survival and inflammation. They found that imaging features have predictive value for specific pathways; also, they concluded that a combination of clinical information, radiomics and genetic biomarkers improve the prognostic predictive performance, showing the complementary value of these data [32]. In this work, we considered only the radiomics features.

The rest of this paper is organised as follows: Section 2 describes materials, i.e., the characteristics of the considered CT images. Section 3 describes the methodology adopted in this work, including feature extraction, a feature reduction strategy, and the experiments' description for training the optimal classifiers. Section 4 discusses the experimental results. Finally, Section 5 draws the final remarks about the conducted study and delineates ideas for future works.

2. Materials

To implement and validate the radiomics approach detailed in Section 3, we used three different datasets of radiomic features. We first considered the two datasets analysed in a previous study by Grossman et al. and which have been made publicly available [32]. They consist of two independent cohorts of North American and European patients with lung cancer. These datasets have been considered in order to assess the validity of the radiomic approach for phenotyping lung cancer. Specifically, Dataset1 (D1) contains data from 262 patients treated within the Thoracic Oncology Program at the *H. Lee Moffitt Cancer Center*, Tampa, FL, USA. The histological analysis was available for 224 patients; 129 subjects (57.6%) had adenocarcinoma, whereas the others (42.4%) suffered from cancer forms other than adenocarcinoma (i.e., 61 patients had squamous carcinoma, whereas the remaining were not further categorized). Dataset2 (D2) includes data from 89 patients treated at the *MAASTRO Clinic* in the Netherlands. The analysis for revealing the histological cancer subtype was available for 87 patients; it showed that 42 subjects (48.3%) had adenocarcinoma, whereas 45 subjects (51.7%) experienced other cancer types, of which 33 (37.9%) were squamous carcinoma.

Lastly, we collected 51 CT scans from patients with lung nodules provided by the *Azienda Ospedaliera—Universitaria "Policlinico Riuniti"* in Foggia. This cohort included 29 patients (56.9%) affected by adenocarcinoma and 22 patients (43.1%) having other cancers, including squamous carcinoma, large cell lung carcinoma, chronic benign inflammation, hepatocarcinoma metastasis, intestinal adenocarcinoma metastasis and endocrine small cell carcinoma). This cohort included only unenhanced CT scans. This choice was made to verify whether the discrimination of lung cancer was also possible from this kind of image, also considering that, in this way, the patient is exposed to a lower amount of radiation. All patients were 18 to 85 years old and accidentally discovered solitary pulmonary nodules 10 to 50 mm in size diagnosed with Computed Tomography. The reports of the histological examinations were collected and the malignancy of the nodules was documented with the relative genetic panel. Subjects with incomplete clinical data and an absence of histological examination in nodules with highly suspected CT criteria of malignancy were excluded.

These CT images were segmented manually by a radiology resident from *Azienda Ospedaliera—Universitaria "Policlinico Riuniti"* in Foggia using the ITK-Snap Software [33], and radiomic features have been subsequently extracted following the pipeline detailed in Section 3.

Figure 2 shows nine slices randomly selected from the CT scans included in D3, with the relative masks. Dataset1 and Dataset2, instead, have been shared by Grossman et al. in terms of radiomic features [32], thus no image processing steps were required. Table 1, instead, summarises the characteristics of the datasets in terms of sample size, imaging modality and type of available data.

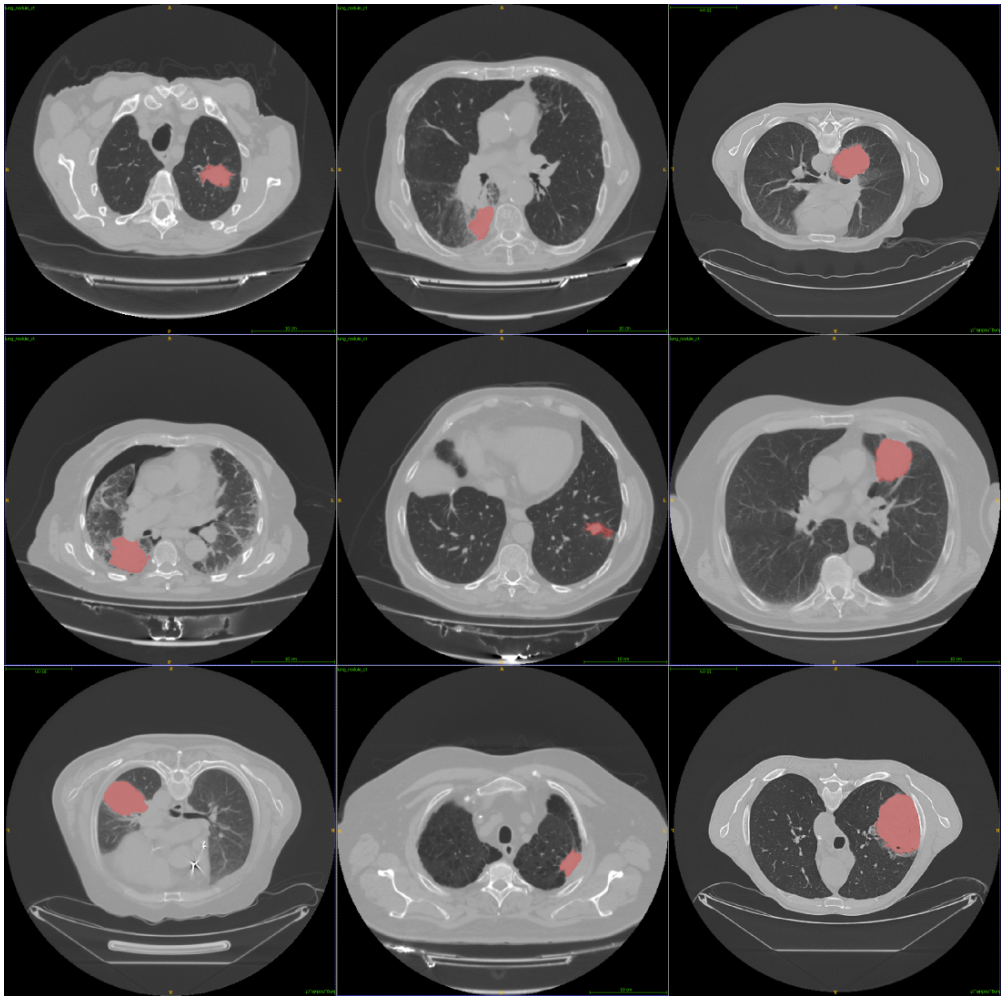

Figure 2. Sample Images with masks from the dataset of the local cohort.

Table 1. Dataset Information.

Dataset	Acronym	Sample	Image Modality	Type of Data
Dataset1 [32]	D1	262	CT scans (89% contrast-enhanced)	Radiomic features, clinical data, genomic features
Dataset2 [32]	D2	89	CT scans (71% contrast-enhanced)	Radiomic features, clinical data, genomic features
Dataset3	D3	51	Unenhanced CT scans	Radiomic features, hystological type

3. Methods

The workflow designed and implemented in this work included three steps, namely features extraction, features selection and classification. Specifically, the features extraction step was performed only for creating Dataset3, since Dataset1 and Dataset2 already included features. Based on Dataset1, univariate statistical methods were implemented in order to select a reduced number of features that allowed the discrimination between lung adenocarcinoma and other cancer types. Then, classification models were evaluated in different training and validation conditions in order to classify LUAD and other cancer types, and investigate the reproducibility of this study on external validation data. The following paragraphs describe in detail the radiomic features constituting the three datasets, the features selection procedure and the classification approaches.

3.1. Radiomic Features

Radiomic features are quantitative descriptors extracted from images using several algorithms. They express different levels of complexity, from local characteristics to global ones. Gillies et al. distinguished radiomic features into two main categories, i.e., "semantics" and "agnostics" [34]. Semantics are those features commonly used by radiologists to describe ROIs visually. Agnostic features, instead, are quantitative descriptors obtained by mathematical operations on the data, which do not necessarily have a meaning in terms of imaging characteristics. These descriptors include first-, second-, or higher-order statistical indicators and shape features.

Radiomic features can be divided into different classes, i.e., intensity-, morphological- and textural-based characteristics. The intensity-based features, also known as first-order features, describe the distribution of the intensity values in the ROIs based on the computation of the intensities histogram; such features do not take into consideration the spatial relationship between pixels, or voxels in the case of 3D Volumes of Interest (VOIs). Morphological features, instead, describe the geometric characteristics of the region. Textural-based features, also known as second-order statistics, describe the spatial composition of the intensity levels; they are defined starting from several data structures, such as the Gray Level Co-occurrence Matrix (GLCM) [35], the Gray Level Size Zone Matrix (GLSZM) [36,37], the Gray Level Run Length Matrix (GLRLM) [38], the Neighboring Gray Tone Difference Matrix (NGTDM) [39] and the Gray Level Dependence Matrix (GLDM) [40].

In this work, we used only agnostic features; in particular, we considered only 3D features extracted from VOIs. After having been segmented manually in CT images, VOIs of the tumoral area were reconstructed as a 3D volume before the features extraction step. To extract the radiomic features, we exploited the open-source Python framework *PyRadiomics*, which implements methods for extracting radiomic features in compliance with the definition present in IBSI [13]. The image processing and features extraction steps were performed in accordance with the analogous procedures described in Grossman et al. [32]. We extracted features from both the original VOIs and the filtered VOIs. In particular, we considered both the image after having applied a wavelet transform, where eight decompositions were obtained with Coiflets from 3D volumes, and the image after filtering with the Laplacian of Gaussian (LoG). The sigma parameter for the LoG filter varied from 0.5 to 5, with ticks spaced by 0.5 each.

3.2. Feature Selection

Before setting up the classification stage, data were preliminarily processed. In the first phase, we applied z-score normalisation. Then, we considered techniques for reducing the dataset dimensionality.

We exploited the algorithm presented in Bevilacqua et al. [20] to eliminate the features of D1 mostly correlated among them. This algorithm iterates over each couple of features and discards those with a correlation value higher than a threshold, set to 0.5. The procedure allowed us to retain 29 uncorrelated features. Figure 3 shows the correlation matrices before and after the aforementioned features reduction phase. In particular, it should be noted that clusters of correlated features, evident in Figure 3a, are not present anymore in Figure 3b.

In order to determine the discriminating relevance of each feature, we performed on D1 a cross-validation procedure with a univariate logistic regressor, assessing the mean AUC obtained. We then selected the features which satisfied the following equation: $mean(x_i) - \frac{1}{2} \cdot std(x_i) > 0.5, i = 1, ..., m$, where m is the number of features. Figure 4 shows the results of the univariate logistic regression analysis. The choice of employing 1/2 standard deviations as features inclusion criterion allowed us to not discard too many of them in the univariate analysis, since they could result in being helpful in the subsequent multivariate model for the classification task. This analysis allowed us to retain 10 features in D1. The same features were also retained from D2 and D3.

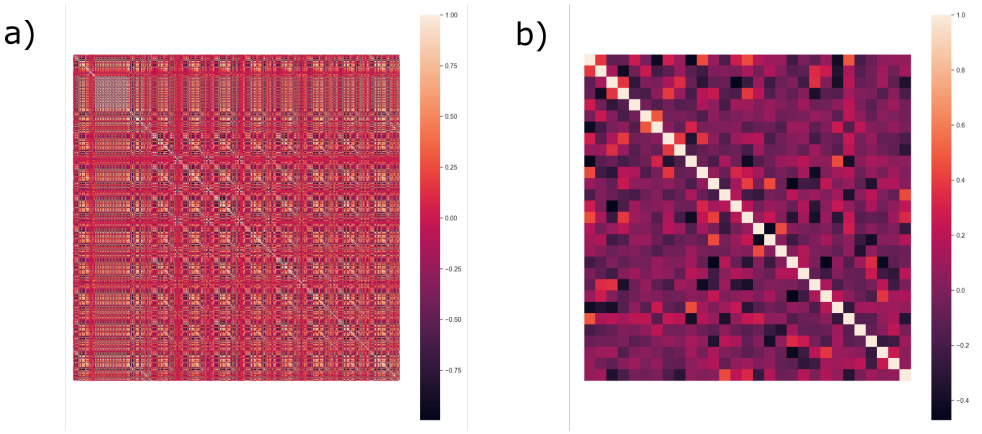

Figure 3. Correlation matrices on the features extracted from D1. (**a**) Correlation matrix before features reduction ($k = 495$). (**b**) Correlation matrix after feature reduction ($k = 29$).

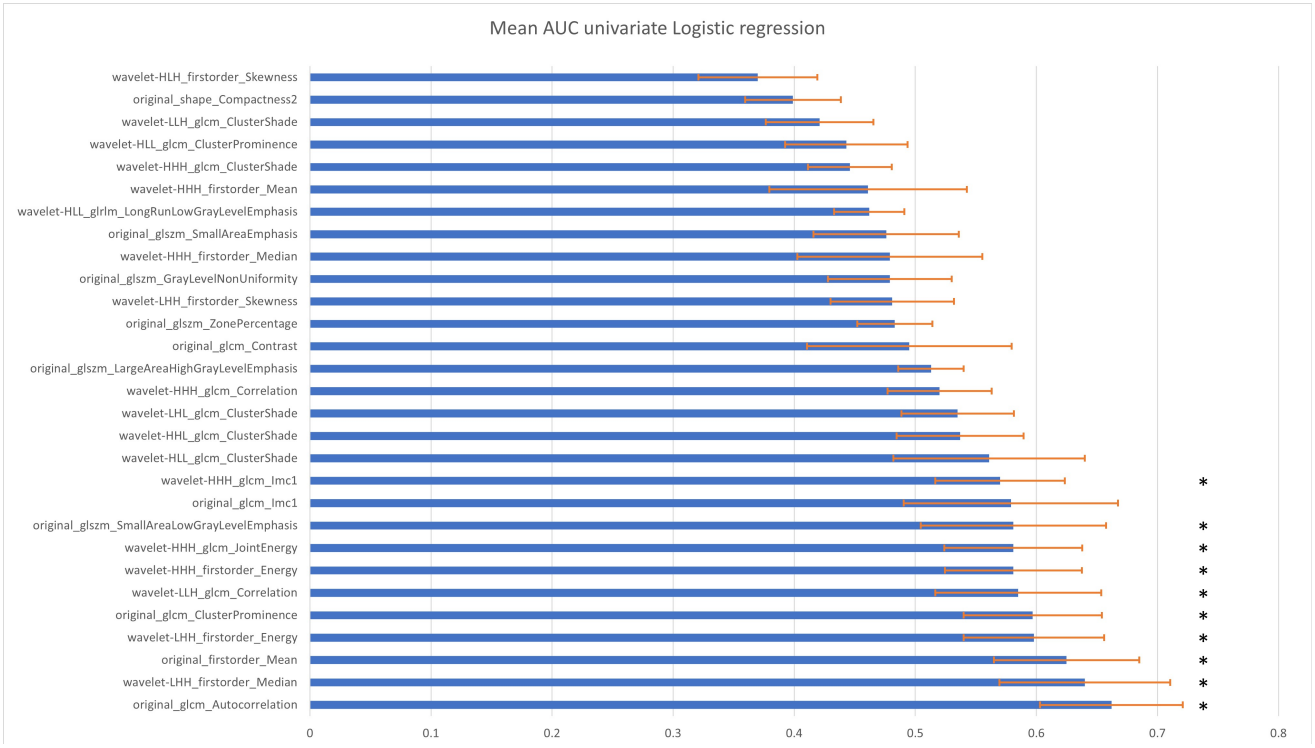

Figure 4. AUC for univariate logistic regression. The blue bar represents the mean AUC obtained in the 10-fold cross-validation, whereas the orange bar depicts $\frac{1}{2}$ standard deviations. Features marked with * satisfy the condition for being retained after univariate analysis.

Statistical Analysis

A statistical analysis was conducted in order to investigate how the features distribute in D1 between patients with adenocarcinoma and patients characterised by other histological subtypes. The Mann–Whitney U statistical test was used for unpaired comparisons between the features of subjects with adenocarcinoma and subjects with other histological types of cancer. A correction for multiple testing, using the Benjamini–Hochberg method was conducted for all the resulting p-values. Corrected p-values lower than 0.05 were considered significant. The statistical analysis revealed three significant features, namely the `original_glcm_Autocorrelation` ($p = 0.001$), the `wavelet-LHH_firstorder_Median` ($p = 0.003$) and the `original_firstorder_Mean` ($p = 0.016$). Figure 5 shows the box plots for the features selected in the features reduction step. In particular, the distributions of the statistically significant features are different among patients with adenocarcinoma

with respect to subjects with other histological types of cancer. For visualisation purposes, normalised values have been clipped in the range $[-4, 4]$, so that outliers do not alter the range of visibility of the boxplots. This difference between the distributions of the statistically significant features can also be seen in Table 2, where summary statistics for the distribution of the features, z-score normalised, is reported. For each feature, Table 2 reports the mean, the median, the standard deviation and the interquartile range for each of the two conditions, and the corrected p-value. The statistical analysis was carried out with Python 3.7, using the `scikit-learn 0.22.2`, `numpy 1.21.5` and `scipy 1.7.3` libraries. Visualisation was performed with `matplotlib 3.5.1` and `seaborn 0.10.0` libraries.

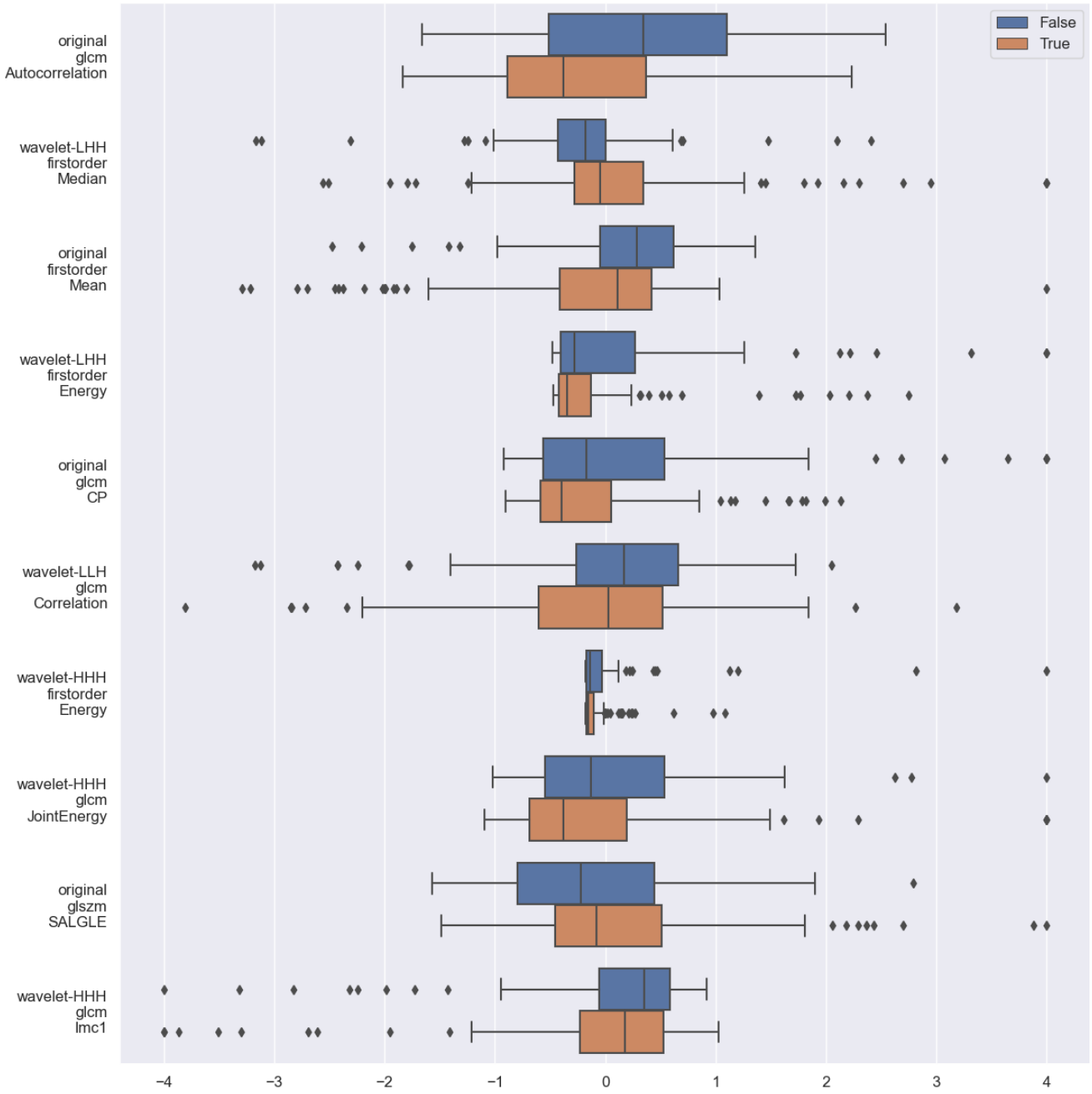

Figure 5. Box plots of the selected features. Feature distributions are shown on the data from D1 images. Normalized values have been clipped in the range $[-4, 4]$, to avoid that some outliers will after the nature of the boxplots. Features `original_glcm_Autocorrelation`, `wavelet-LHH_firstorder_Median`, and `original_firstorder_Mean`, have p-value < 0.05 in Mann–Whitney U test. The p-values are, respectively, 0.001, 0.004, and 0.021.

Table 2. Statistics for features distribution between adenocarcinoma and other cases. Std is the standard deviation, IQR is the interquartile range. Mann–Whitney U tests was used to calculate the *p*-values. The multiple comparisons problem was handled with Benjamini–Hochberg (BH) correction. * indicates statistically significant features (*p*-value < 0.05).

Feature Name	Mean		Std		Median		IQR		*p*-Value
	LUAD	Other	LUAD	Other	LUAD	Other	LUAD	Other	BH
O_glcm_Autocorrelation	−0.2449	0.3470	0.8658	1.0672	−0.3783	0.3439	1.2597	1.6090	0.0011 *
O_glcm_CP	−0.1698	0.2334	0.6434	1.3142	−0.3974	−0.1727	0.6376	1.0980	0.1406
O_glszm_SALGLE	0.1244	−0.1560	1.0337	0.9384	−0.0800	−0.2190	0.9632	1.2388	0.1993
O_firstorder_Mean	−0.1230	0.1965	1.1544	0.6577	0.1154	0.2908	0.8318	0.6643	0.0207 *
W-HHH_glcm_JointEnergy	−0.0526	0.0776	1.0925	0.8696	−0.3825	−0.1291	0.8819	1.0823	0.1591
W-HHH_glcm_Imc1	−0.0565	0.0869	1.0679	0.9050	0.1817	0.3549	0.7541	0.6453	0.1993
W-HHH_firstorder_Energy	−0.0917	0.1253	0.1835	1.5198	−0.1577	−0.1340	0.0651	0.1342	0.1591
W-LHH_firstorder_Energy	−0.1404	0.1940	0.5804	1.3641	−0.3493	−0.2788	0.2868	0.6737	0.1406
W-LHH_firstorder_Median	0.1818	−0.2496	1.1321	0.7358	−0.0498	−0.1823	0.6231	0.4321	0.0036 *
W-LLH_glcm_Correlation	−0.0680	0.0908	1.0421	0.9493	0.0318	0.1667	1.1192	0.9215	0.1993

3.3. Classification

The main task of this work was to classify the histological subtype of lung nodules using radiomic features. In order to accomplish this task, we trained several classifiers considering a variable number of features. Specifically, features were added to the input pattern of each classifier according to their mean univariate AUC calculated in the features selection phase (as described in the previous section whose results are reported in Figure 4).

The employed classification models include Logistic Regression (LogReg), Support Vector Machines (SVM) with a linear kernel, AdaBoost (AdaBo), Random Forest (RF), Multi-layer Perceptron (MLP), Gradient Boost (GB); these models were selected since they are widely used in radiomic studies for medical classification purposes [41,42].

We also evaluated an ensemble of five models (En5) composed by the models mentioned above, except for GB. The ensemble model makes decisions following a voting strategy; the classification output by LogReg, SVM, and MLP was weighted 2; RF and AdaBo prediction was weighted as 1. Simpler models tend to be more robust and so they have received a larger weight. GB was not included since its adding to the ensemble did not lead to performance improvements. Moreover, it already suffered from low accuracies in many cases when more features were considered. The hyperparameters of all the models are reported in Table 3.

To assess the performance of the classifiers and the reliability of the radiomic features considered, we performed four experiments:

Experiment 1: Internal cross-validation on each dataset (D1, D2, D3) separately; this experiment aimed to investigate if the radiomic signatures found by our approach could be discriminative for this classification task;

Experiment 2: To train models on D1 and validate them on D2 and D3; this experiment aimed to investigate if the classifiers trained on a single dataset could perform on different datasets;

Experiment 3: To train models on D1 and D2 (as a single dataset) and validate them on D3; this experiment aimed to investigate if increasing the training sample size could have improved the classification performance on the external validation;

Experiment 4: Internal cross-validation merging D1, D2 and D3; this experiment aimed to investigate if considering all the datasets together could have led to a better performance than considering datasets separately, even at the cost of lower generalisation capabilities of the models.

In all the experiments, cross-validation was performed with 10-folds and stratified samples, i.e., the distribution of classes remains the same across folds. Figure 6 shows the workflows implemented for the four classification experiments in terms of input data and output results.

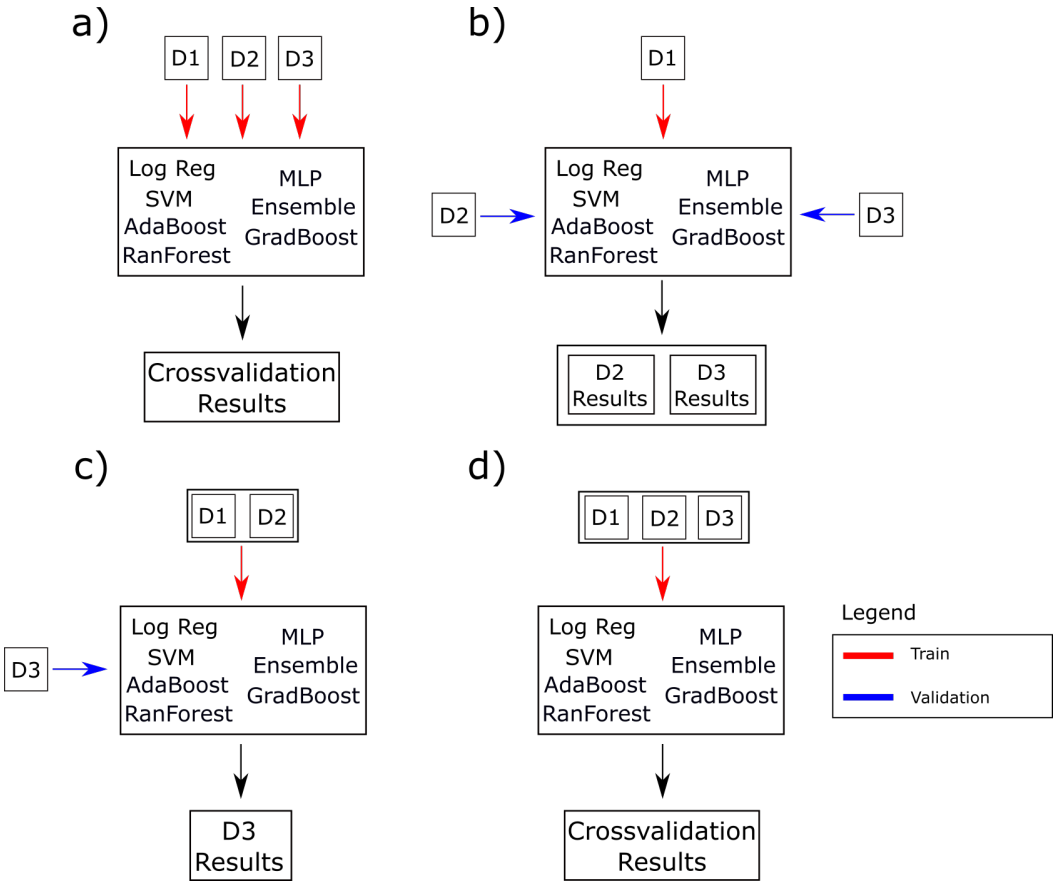

Figure 6. Experimental Setup. (**a**) Workflow for experiment 1. It consists of separate internal cross-validation for each cohort. (**b**) Workflow for experiment 2. It consists of using D1 for training and D2 and D3 as external validation cohorts. (**c**) Workflow for experiment 3. It consists of using D1 and D2 for training and D3 as external validation cohort. (**d**) Workflow for experiment 4. It consists of performing internal cross-validation considering D1, D2, and D3 as a single large cohort.

Table 3. Training hyperparameters.

Model	Hyperparameters	Value
LogReg	`max_iter`	100
	`penalty`	12
	`C`	1.0
SVM	`kernel`	`rbf`
	`C`	1.0
	`penalty`	12
AdaBo	`n_estimators`	50
	`learning_rate`	1.0
RF	`n_estimators`	100
	`criterion`	`gini`
	`max_depth`	None

Table 3. *Cont.*

Model	Hyperparameters	Value
MLP	neurons_per_layer	100
	hidden_layers	1
	b1	0.9
	b2	0.999
	solver	adam
	learning_rate_init	0.001
	max_iter	200
	early_stop	None
	penalty	12
	alpha	0.0001
	activation	relu
GB	learning_rate	0.1
	n_estimators	100
	criterion	friedman_mse

4. Results and Discussion

The results of each experiment are reported in Figure 7. Each matrix in the figure shows the mean AUC per model per number of features of the input pattern.

Experiment 1 aimed to assess if the radiomic signature obtained from Dataset1, thanks to the implemented approach for features reduction, could be discriminative for classifying LUAD and other histological types also on the other datasets. To do this, we trained the classification models on the selected features from D1, D2 and D3, respectively. The performances of these models are reported in Figure 7(I-a, I-b and I-c), respectively.

The classifiers reached the highest robustness in classifying samples from D1, regardless of the number of features and the classifier itself. Specifically, the mean (±standard deviation) AUC for all the considered models and features was 0.64 (±0.04), 0.54 (±0.06), and 0.56 (±0.09), considering data from D1, D2 and D3, respectively. Considering input data from D1, the best model was LogReg, which allowed us to obtain an AUC of 0.70 when trained with the input pattern size of six features. Internal cross-validation with input data from D2, instead, showed an increased variability among models' performances, regardless of the number of features, as can be noticed in the heatmap reported in Figure 7(I-b). In this case, the best model was the SVM trained on only one feature, which allowed us to achieve an AUC of 0.65. Considering the models trained on input data from D3, LogReg and MLP allowed us to achieve the best results in the univariate version; also in this case, with a mean AUC of 0.75.

Experiment 2 aimed to assess if models trained on the histotype signature from D1 could generalise to make decisions also on D2 and D3. The results obtained show that GB, trained on two features, and AdaBo, trained on four features, allowed us to achieve an AUC of 0.62 considering data from D2 (Figure 7(II-a)). However, better performances were obtained, validating models on D3, with an AUC of 0.70 for SVM trained with one, five, and six features (Figure 7(II-b)).

In **Experiment 3**, instead, we investigated if increasing the training sample would allow us to obtain a higher performance on D3 as a validation set with respect to **Experiment 2**. As shown in Figure 7III, the MLP trained with an input pattern of seven features allowed us to obtain an AUC of 0.69 on data from D3 considering a joint train set composed of D1 and D2. Comparing this experiment to the previous one, a small improvement in terms of mean performances were obtained; in fact, this training configuration allowed us to achieve a mean (±standard deviation) AUC of 0.58 (±0.07), instead of 0.59 (±0.09), on the external validation set D3.

Concerning **Experiment 4**, it has been designed to evaluate how the models would have performed if D1, D2 and D3 would be merged. The internal cross-validation revealed that merging D1, D2 and D3 led to more robust performance of the classifiers, except for the

comparison with the case of internal cross-validation performed on D1 (in **Experiment 1**). LogReg, RF, and Ensemble-5 models allowed us to achieve an AUC of 0.65 in their best case, i.e., trained considering two, six and seven features, respectively (as shown in Figure 7 IV). The overall mean (±standard deviation) AUC obtained in this experiment is 0.61 (±0.03).

Since the feature set has been chosen considering data in D1, we observed in some cases, i.e., Experiments 1 and 2, that a single feature allowed us to achieve the best performances in some external validation settings. It should be noted that, in the case of input patterns composed of a single feature, MLPs tend to behave very similarly to LogReg, considering that MLP is a set of neurons each with the capability of LogReg; thus, this configuration could not extract more information in the presence of only one feature.

The best ROC curves in external validation settings for cases II-a, II-b and III are reported in Figure 8.

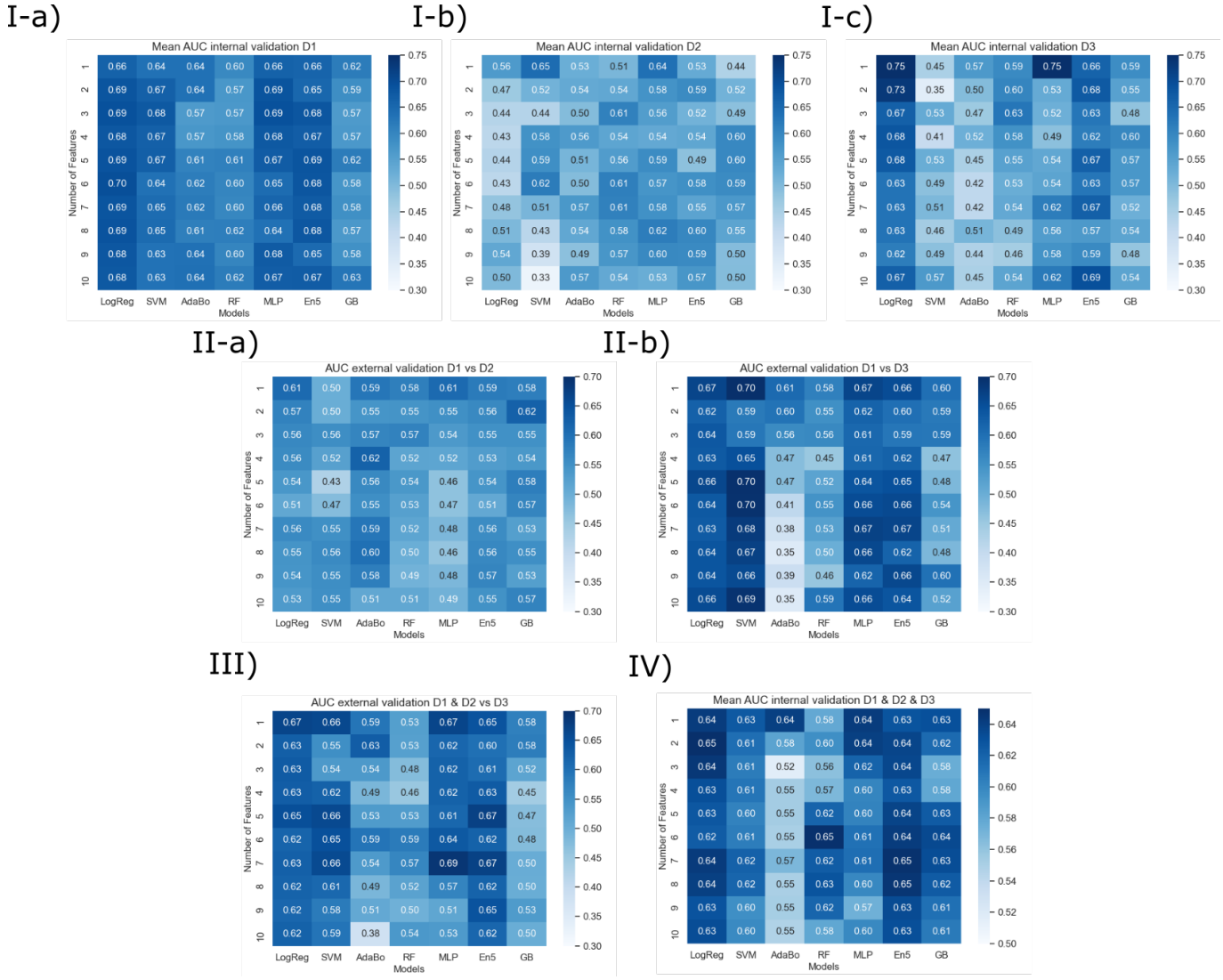

Figure 7. Heatmaps with AUC for all the experimental setups. (**I-a**,**I-b**,**I-c**) Heatmaps with mean AUC of internal cross-validation from experiment 1 for D1, D2, and D3, respectively. (**II-a**,**II-b**) AUC of external validation, from experiment 2, on D2 and D3, respectively. (**III**) AUC of external validation, from experiment 3, on D3. (**IV**) Mean internal cross-validation AUC from experiment 4 on D1, D2, and D3 considered as a single cohort.

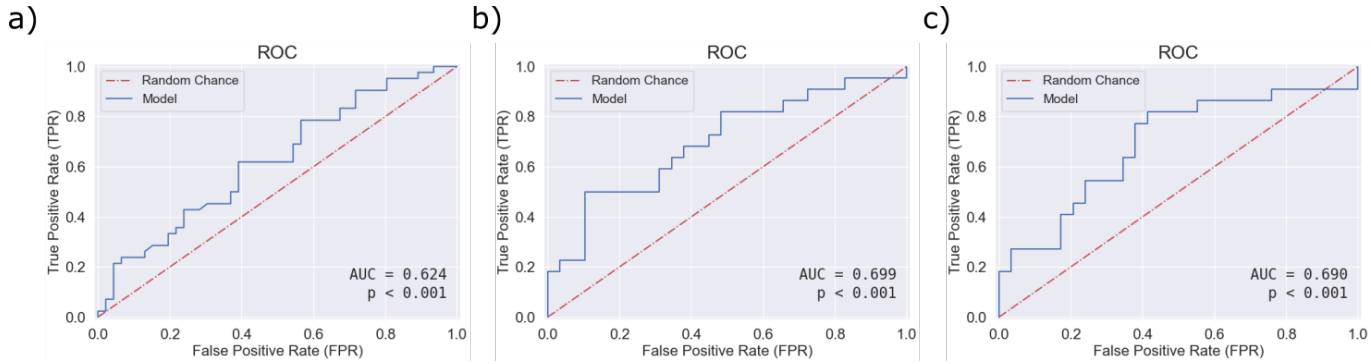

Figure 8. ROC curves for the considered external validations. (**a**) External validation of the SVM classifier (5 features) trained on D1 and validated on D2. (**b**) External validation of the AdaBoost classifier (7 features) trained on D1 and validated on D3. (**c**) External validation of the AdaBoost Classifier (10 features) trained on the merged dataset D1, D2 and validated on D3.

5. Conclusions

Radiomics is a research field in which several algorithms are exploited to extract quantitative high-dimensional characteristics from images. Several authors have already used this methodology to address classification problems that involve radiological images. The use of algorithms for the processing and reduction of features may allow the design and implementation of more robust classification models that are able to improve the capabilities of automatic intelligent systems to support decisions in the biomedical field. However, the reproducibility of radiomic studies on external validation data is considered a crucial point.

In this study, we extracted a reduced set of radiomic features to classify the histological subtype in lung nodules to discriminate between adenocarcinoma cases from other histological classes. In this work, we used three datasets, namely D1 and D2, which were provided by Grossman et al., and D3, which was built starting from a local cohort of lung CTs opportunely processed to extract radiomic features.

Starting from D1, we designed and implemented a pipeline based on univariate analysis to select the most discriminative features to discriminate between lung adenocarcinoma and other histological types.

Then, four experiments were performed to investigate whether and how the number of features, the input dataset, and the model affect the classification performance and the reproducibility of radiomic studies.

Our analyses revealed that the best configuration for reproducibility on our local dataset D3 is the one using D1 for training the model. However, also combining D1 and D2 as a single training set led to good performances on data from D3.

In any case, internal cross-validations evaluated in Experiment 1 allowed us to reach higher performances, highlighting the importance of considering external validation datasets when performing radiomics studies, so that the real effectiveness of the discovered feature set can be measured. Eventually, the best result for the D3 dataset comes from internal cross-validation, but then this result would be less generalisable on unseen data.

Such a conclusion is also endorsed by other authors, such as Wu et al., who also performed external validations in a similar case study, but with a reduced cohort for external validations, obtaining similar results (AUC of 0.72 with Naïve Bayes' classifier) [31].

Future works may include the comparison between different kinds of enhancement modalities from CT scans before performing the radiomics feature extraction and the correlation of radiomic signatures with genomics data. The collection of new data may also help in the direction of designing a robust radiomics pipeline for histological and genomic classification.

Author Contributions: Conceptualization, B.P., M.G. and V.B.; methodology, A.B., N.A., D.B. and B.P.; formal analysis, A.B., N.A. and B.P.; data curation, N.A., E.G., F.C., M.G. and B.P.; writing—original draft preparation, A.B., N.A., D.B. and B.P.; writing—review and editing, all authors; visualization,

B.P. and N.A.; supervision, A.B., M.G. and V.B.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki. Ethical review and approval were waived for this study since this was a retrospective observational study with anonymised data.

Informed Consent Statement: Patient consent was waived due to the fact that this was a retrospective observational study with anonymised data, already acquired for medical diagnostic purposes.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AdaBo	AdaBoost
LogReg	Logistic Regression
MLP	Multi Layer Perceptron
RF	Random Forest
SVM	Support Vector Machine
AUC	Area Under Curve
CT	Computed Tomography
C2	`Compacteness2`
CP	`ClusterProminence`
CS	`ClusterShade`
fMed	`Firstorder Median`
fS	`Firstorder Skewness`
GLNU	`GrayLevelNonUniformity`
LAHGLE	`LargeAreaHighGrayLevelEmphasis`
LRLGLE	`LongRunLowGrayLevelEmphasis`
SAE	`SmallAreaEmphasis`
SALGLE	`SmallAreaLowGrayLevelEmphasis`

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Altini, N.; Cascarano, G.D.; Brunetti, A.; Marino, F.; Rocchetti, M.T.; Matino, S.; Venere, U.; Rossini, M.; Pesce, F.; Gesualdo, L.; et al. Semantic segmentation framework for glomeruli detection and classification in kidney histological sections. *Electronics* **2020**, *9*, 503. [CrossRef]
3. Bevilacqua, V.; Brunetti, A.; Trotta, G.F.; Dimauro, G.; Elez, K.; Alberotanza, V.; Scardapane, A. A novel approach for Hepatocellular Carcinoma detection and classification based on triphasic CT Protocol. In Proceedings of the 2017 IEEE Congress on Evolutionary Computation (CEC), Donostia, Spain, 5–8 June 2017; pp. 1856–1863. [CrossRef]
4. Robinson, P.N. Deep phenotyping for precision medicine. *Hum. Mutat.* **2012**, *33*, 777–780. [CrossRef]
5. Tan, W.L.; Jain, A.; Takano, A.; Newell, E.W.; Iyer, N.G.; Lim, W.T.; Tan, E.H.; Zhai, W.; Hillmer, A.M.; Tam, W.L.; et al. Novel therapeutic targets on the horizon for lung cancer. *Lancet. Oncol.* **2016**, *17*, e347–e362. [CrossRef]
6. Bevilacqua, V.; Brunetti, A.; Guerriero, A.; Trotta, G.F.; Telegrafo, M.; Moschetta, M. A performance comparison between shallow and deeper neural networks supervised classification of tomosynthesis breast lesions images. *Cogn. Syst. Res.* **2019**, *53*, 3–19. [CrossRef]
7. Aerts, H.J. The potential of radiomic-based phenotyping in precision medicine: a review. *JAMA Oncol.* **2016**, *2*, 1636–1642. [CrossRef]
8. Bevilacqua, V. Three-dimensional virtual colonoscopy for automatic polyps detection by artificial neural network approach: New tests on an enlarged cohort of polyps. *Neurocomputing* **2013**, *116*, 62–75. [CrossRef]
9. European Society of Radiology (ESR) communications@myesr.org. Medical imaging in personalised medicine: A white paper of the research committee of the European Society of Radiology (ESR). *Insights Into Imaging* **2015**, *6*, 141–155. [CrossRef]
10. Lee, G.; Lee, H.Y.; Park, H.; Schiebler, M.L.; van Beek, E.J.R.; Ohno, Y.; Seo, J.B.; Leung, A. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art. *Eur. J. Radiol.* **2017**, *86*, 297–307. [CrossRef]

11. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; Van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **2012**, *48*, 441–446. [CrossRef]
12. Rahmim, A.; Schmidtlein, C.R.; Jackson, A.; Sheikhbahaei, S.; Marcus, C.; Ashrafinia, S.; Soltani, M.; Subramaniam, R.M. A novel metric for quantification of homogeneous and heterogeneous tumors in PET for enhanced clinical outcome prediction. *Phys. Med. Biol.* **2015**, *61*, 227. [CrossRef] [PubMed]
13. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **2020**, *295*, 328–338. [CrossRef] [PubMed]
14. Kumar, V.; Gu, Y.; Basu, S.; Berglund, A.; Eschrich, S.A.; Schabath, M.B.; Forster, K.; Aerts, H.J.; Dekker, A.; Fenstermacher, D.; et al. Radiomics: The process and the challenges. *Magn. Reson. Imaging* **2012**, *30*, 1234–1248. [CrossRef] [PubMed]
15. Rizzo, S.; Botta, F.; Raimondi, S.; Origgi, D.; Fanciullo, C.; Morganti, A.G.; Bellomi, M. Radiomics: The facts and the challenges of image analysis. *Eur. Radiol. Exp.* **2018**, *2*, 36. [CrossRef]
16. Loconsole, C.; Cascarano, G.D.; Brunetti, A.; Trotta, G.F.; Losavio, G.; Bevilacqua, V.; Di Sciascio, E. A model-free technique based on computer vision and sEMG for classification in Parkinson's disease by using computer-assisted handwriting analysis. *Pattern Recognit. Lett.* **2019**, *121*, 28–36. [CrossRef]
17. Buongiorno, D.; Barsotti, M.; Barone, F.; Bevilacqua, V.; Frisoli, A. A linear approach to optimize an EMG-driven neuromusculoskeletal model for movement intention detection in myo-control: A case study on shoulder and elbow joints. *Front. Neurorobot.* **2018**, 74. [CrossRef]
18. Le, N.Q.K.; Kha, Q.H.; Nguyen, V.H.; Chen, Y.C.; Cheng, S.J.; Chen, C.Y. Machine learning-based radiomics signatures for EGFR and KRAS mutations prediction in non-small-cell lung cancer. *Int. J. Mol. Sci.* **2021**, *22*, 9254. [CrossRef]
19. Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **2014**, *5*, 4006. [CrossRef]
20. Bevilacqua, V.; Altini, N.; Prencipe, B.; Brunetti, A.; Villani, L.; Sacco, A.; Morelli, C.; Ciaccia, M.; Scardapane, A. Lung Segmentation and Characterization in COVID-19 Patients for Assessing Pulmonary Thromboembolism: An Approach Based on Deep Learning and Radiomics. *Electronics* **2021**, *10*, 2475. [CrossRef]
21. Chaddad, A.; Kucharczyk, M.J.; Daniel, P.; Sabri, S.; Jean-claude, B.J.; Niazi, T.; Abdulkarim, B. Radiomics in Glioblastoma: Current Status and Challenges Facing Clinical Implementation. *Front. Oncol.* **2019**, *9*, 374. [CrossRef]
22. Wakabayashi, T.; Ouhmich, F.; Gonzalez, C.; Emanuele, C.; Saviano, A.; Agnus, V.; Savadjiev, P.; Baumert, T.F.; Pessaux, P.; Marescaux, J.; et al. Radiomics in hepatocellular carcinoma: A quantitative review. *Hepatol. Int.* **2019**, *13*, 546–559. [CrossRef] [PubMed]
23. Valdora, F.; Houssami, N.; Rossi, F.; Calabrese, M.; Stefano, A.; Pet, C.T. Rapid review: Radiomics and breast cancer. *Breast Cancer Res. Treat.* **2018**, *169*, 217–229. [CrossRef] [PubMed]
24. Park, J.E.; Park, S.Y.; Kim, H.J.; Kim, H.S. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. *Korean J. Radiol.* **2019**, *20*, 1124–1137. [CrossRef]
25. Parmar, C.; Leijenaar, R.T.; Grossmann, P.; Velazquez, E.R.; Bussink, J.; Rietveld, D.; Rietbergen, M.M.; Haibe-Kains, B.; Lambin, P.; Aerts, H.J. Radiomic feature clusters and Prognostic Signatures specific for Lung and Head & neck cancer. *Sci. Rep.* **2015**, *5*, 11044. [CrossRef]
26. Zhang, Y.; Oikonomou, A.; Wong, A.; Haider, M.A.; Khalvati, F. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. *Sci. Rep.* **2017**, *7*, 46349. [CrossRef] [PubMed]
27. Van Griethuysen, J.J.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef]
28. Ferreira-Junior, J.R.; Koenigkam-Santos, M.; Magalhães Tenório, A.P.; Faleiros, M.C.; Garcia Cipriano, F.E.; Fabro, A.T.; Näppi, J.; Yoshida, H.; de Azevedo-Marques, P.M. CT-based radiomics for prediction of histologic subtype and metastatic disease in primary malignant lung neoplasms. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 163–172. [CrossRef]
29. Linning, E.; Lu, L.; Li, L.; Yang, H.; Schwartz, L.H.; Zhao, B. Radiomics for Classifying Histological Subtypes of Lung Cancer Based on Multiphasic Contrast-Enhanced Computed Tomography. *J. Comput. Assist. Tomogr.* **2019**, *43*, 300–306. [CrossRef]
30. Zhai, T.T.; Wesseling, F.; Langendijk, J.A.; Shi, Z.; Kalendralis, P.; van Dijk, V.L.; Hoebers, F.; Steenbakkers, R.J.H.M.; Dekker, A.; Wee, L.; et al. External validation of nodal failure prediction models including radiomics in head and neck cancer. *Oral Oncol.* **2021**, *112*, 105083. [CrossRef]
31. Wu, W.; Parmar, C.; Grossmann, P.; Quackenbush, J.; Lambin, P.; Bussink, J.; Mak, R.; Aerts, H.J. Exploratory study to identify radiomics classifiers for lung cancer histology. *Front. Oncol.* **2016**, *6*, 71. [CrossRef]
32. Grossmann, P.; Stringfield, O.; El-Hachem, N.; Bui, M.M.; Velazquez, E.R.; Parmar, C.; Leijenaar, R.T.; Haibe-Kains, B.; Lambin, P.; Gillies, R.J.; et al. Defining the biological basis of radiomic phenotypes in lung cancer. *eLife* **2017**, *6*, e23421. [CrossRef] [PubMed]
33. Yushkevich, P.A.; Piven, J.; Cody Hazlett, H.; Gimpel Smith, R.; Ho, S.; Gee, J.C.; Gerig, G. User-Guided {3D} Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability. *Neuroimage* **2006**, *31*, 1116–1128. [CrossRef] [PubMed]

34. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **2016**, *278*, 563–577. [CrossRef] [PubMed]
35. Haralick, R.M.; Dinstein, I.; Shanmugam, K. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
36. Galloway, M.M. Texture analysis using gray level run lengths. *Comput. Graph. Image Process.* **1975**, *4*, 172–179. [CrossRef]
37. Chu, A.; Sehgal, C.M.; Greenleaf, J.F. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognit. Lett.* **1990**, *11*, 415–419. [CrossRef]
38. Thibault, G.; Fertil, B.; Navarro, C.; Pereira, S.; Cau, P.; Levy, N.; Sequeira, J.; Mari, J.l. Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification. In Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 140–145.
39. Sun, C.; Wee, W.G. Neighboring gray level dependence matrix for texture classification. *Comput. Vision, Graph. Image Process.* **1983**, *23*, 341–352. [CrossRef]
40. Amadasun, M.; King, R. Texural Features Corresponding to Texural Properties. *IEEE Trans. Syst. Man Cybern.* **1989**, *19*, 1264–1274. [CrossRef]
41. Wang, P.; Pei, X.; Yin, X.P.; Ren, J.L.; Wang, Y.; Ma, L.Y.; Du, X.G.; Gao, B.L. Radiomics models based on enhanced computed tomography to distinguish clear cell from non-clear cell renal cell carcinomas. *Sci. Rep.* **2021**, *11*, 13729. [CrossRef]
42. Altini, N.; Brunetti, A.; Mazzoleni, S.; Moncelli, F.; Zagaria, I.; Prencipe, B.; Lorusso, E.; Buonamico, E.; Carpagnano, G.E.; Bavaro, D.F.; et al. Predictive Machine Learning Models and Survival Analysis for COVID-19 Prognosis Based on Hematochemical Parameters. *Sensors* **2021**, *21*, 8503. [CrossRef]