




## Article

# Phenotyping the Histopathological Subtypes of Non-Small-Cell Lung Carcinoma: How Beneficial Is Radiomics?

Giovanni Pasini <sup>1,2</sup>, Alessandro Stefano <sup>2,\*</sup>, Giorgio Russo <sup>2</sup>, Albert Comelli <sup>2,3</sup>, Franco Marinozzi <sup>1</sup>  
and Fabiano Bini <sup>1</sup>

<sup>1</sup> Department of Mechanical and Aerospace Engineering, Sapienza University of Rome, Eudossiana 18, 00184 Rome, Italy

<sup>2</sup> Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Contrada, Pietrapollastra-Pisciotta, 90015 Cefalù, Italy

<sup>3</sup> Ri.MED Foundation, Via Bandiera 11, 90133 Palermo, Italy

\* Correspondence: alessandro.stefano@ibfm.cnr.it

**Abstract:** The aim of this study was to investigate the usefulness of radiomics in the absence of well-defined standard guidelines. Specifically, we extracted radiomics features from multicenter computed tomography (CT) images to differentiate between the four histopathological subtypes of non-small-cell lung carcinoma (NSCLC). In addition, the results that varied with the radiomics model were compared. We investigated the presence of the batch effects and the impact of feature harmonization on the models' performance. Moreover, the question on how the training dataset composition influenced the selected feature subsets and, consequently, the model's performance was also investigated. Therefore, through combining data from the two publicly available datasets, this study involves a total of 152 squamous cell carcinoma (SCC), 106 large cell carcinoma (LCC), 150 adenocarcinoma (ADC), and 58 no other specified (NOS). Through the matRadiomics tool, which is an example of Image Biomarker Standardization Initiative (IBSI) compliant software, 1781 radiomics features were extracted from each of the malignant lesions that were identified in CT images. After batch analysis and feature harmonization, which were based on the ComBat tool and were integrated in matRadiomics, the datasets (the harmonized and the non-harmonized) were given as an input to a machine learning modeling pipeline. The following steps were articulated: (i) training-set/test-set splitting (80/20); (ii) a Kruskal–Wallis analysis and LASSO linear regression for the feature selection; (iii) model training; (iv) a model validation and hyperparameter optimization; and (v) model testing. Model optimization consisted of a 5-fold cross-validated Bayesian optimization, repeated ten times (inner loop). The whole pipeline was repeated 10 times (outer loop) with six different machine learning classification algorithms. Moreover, the stability of the feature selection was evaluated. Results showed that the batch effects were present even if the voxels were resampled to an isotropic form and whether feature harmonization correctly removed them, even though the models' performances decreased. Moreover, the results showed that a low accuracy (61.41%) was reached when differentiating between the four subtypes, even though a high average area under curve (AUC) was reached (0.831). Further, a NOS subtype was classified as almost completely correct (true positive rate ~90%). The accuracy increased (77.25%) when only the SCC and ADC subtypes were considered, as well as when a high AUC (0.821) was obtained—although harmonization decreased the accuracy to 58%. Moreover, the features that contributed the most to models' performance were those extracted from wavelet decomposed and Laplacian of Gaussian (LoG) filtered images and they belonged to the texture feature class. In conclusion, we showed that our multicenter data were affected by batch effects, that they could significantly alter the models' performance, and that feature harmonization correctly removed them. Although wavelet features seemed to be the most informative features, an absolute subset could not be identified since it changed depending on the training/testing splitting. Moreover, performance was influenced by the chosen dataset and by the machine learning methods, which could reach a high accuracy in binary classification tasks, but could underperform in multiclass problems. It is, therefore, essential that the scientific community propose a more systematic radiomics approach, focusing on multicenter studies, with clear and solid guidelines to facilitate the translation of radiomics to clinical practice.



**Citation:** Pasini, G.; Stefano, A.; Russo, G.; Comelli, A.; Marinozzi, F.; Bini, F. Phenotyping the Histopathological Subtypes of Non-Small-Cell Lung Carcinoma: How Beneficial Is Radiomics? *Diagnostics* **2023**, *13*, 1167. <https://doi.org/10.3390/diagnostics13061167>

Academic Editor: Alessio Imperiale

Received: 8 February 2023

Revised: 16 March 2023

Accepted: 16 March 2023

Published: 18 March 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** radiomics; CT; non-small-cell lung carcinoma; NSCLC; phenotyping; multicenter; harmonization; machine learning

## 1. Introduction

Lung cancer is the second most common type of cancer, both in men and in women. Moreover, despite its incidence rate being less than prostate cancer and breast cancer, its mortality rate is higher. Indeed, lung cancer is the leading cause of cancer deaths in 2022 [1,2]. However, the percentage of people still living 3 years after diagnosis is increasing, especially thanks to advances in surgical techniques, early detection, and targeted therapies. However, this is largely confined to non-small-cell lung cancer (NSCLC) [1].

NSCLC is the type of lung cancer that occurs most often (85%) between the two main forms of lung cancer—which are NSCLC and small-cell lung cancer (SCLC). Moreover, the World Health Organization (WHO) has further classified NSCLC into three main groups: adenocarcinoma (ADC, ~40%), squamous cell carcinoma (SCC, 25–30%), and large cell carcinoma (LCC, 5–10%). Furthermore, difficulties in the NSCLC diagnosis have led to the creation of a fourth class, named “not otherwise specified” (NOS), which includes NSCLCs that do not have the characteristics of the three main subtypes [3,4].

Typically, NSCLC is diagnosed in cases of advanced-stage disease, which is when patients that show symptoms—such as coughing, hemoptysis chest pain, and dyspnea—undergo medical imaging that is followed by biopsy. Furthermore, the efficacy of treatment strongly depends on the cancer stage, and it is mostly based on surgery, chemotherapy, immunotherapy, and radiotherapy [3,5].

Recently, precision medicine has emerged as an innovative approach to refine the treatment of NSCLC, according to histological and molecular subtypes, thus improving disease outcome [6]. However, the early detection, differential diagnoses of NSCLC subtypes, and the cancer staging remain the main challenges in terms of allowing treatment personalization. To address this issue, radiomics could be a useful tool to support the clinical decision process [7].

Radiomics is an emerging research field that involves the use of artificial intelligence to analyze medical images, either through using classical machine learning pipelines or through advanced deep learning methods. Its aim is to extract quantitative metrics that can be used to build predictive models that are able to respond to a specific clinical question. Therefore, radiomics can be used to identify the most predictive biomarkers in a disease, to perform differential diagnoses between cancer subtypes and lung diseases, and to predict overall survival and responses to therapy [8–10]. Radiomics is used to obtain information that is not even visible to the eyes of expert clinicians and radiologists, such as the shape, the distribution of gray levels, and the texture of a lesion, thus increasing the amount of data available. Its main advantage over biopsy is that it is non-invasive, less time consuming, and can be integrated in automated pipelines [11].

However, despite the Imaging Biomarker Standardization Initiative (IBSI) [12] being provided, the guidelines on how radiomics features should be computed, the lack of standardization in the parameters used during the feature extraction process (as well as in the algorithms that are used to perform feature selection), and machine learning are all still major challenges and represent the limits of radiomics itself [13].

Moreover, when medical images come from different centers, dissimilarities in the protocols and in the technical specifications of the imaging manufacturers can lead to the generation of batch effects. This can have a negative impact on radiomics analysis. To address this issue, batch analysis and feature harmonization are needed [14,15].

In oncology, several studies demonstrated the predictive and prognostic power of radiomics, such as those focused on the detection and localization of prostate cancer, the Gleason Score (GS), and the recurrence prediction [16,17], while others were focused on the differential diagnoses between lung cancer types and the prediction of lung nodules’

malignancy [18,19]. Radiomics was also applied to breast cancer detection and several studies demonstrated that its integration with mammography and magnetic resonance imaging (MRI) improved diagnostic accuracy [20]. Other applications are those that involve radiomics for the evaluation of both gastrointestinal tumors [21] and brain tumors [22]. In neuroscience, radiomics is applied for the early detection of neurodegenerative diseases, such as Alzheimer's [23] and Parkinson's [24] disease. In addition, many deep learning methods have been proposed [25], such as those that were developed for automated multiple sclerosis detection [26]. It has also been applied to predict the expanded disability status scale (EDSS) [27] of patients who were affected by multiple sclerosis. Meanwhile, certain studies also focused on the prediction of epilepsy in patients affected by frontal glioma [28].

Regarding NSCLC, radiomics studies have focused on predicting mediastinal lymph node metastasis [29]. More specifically, they were mostly focused on only differentiating between the two NSCLC subtypes, such as SCC and ADC [30–34], rather than focusing on multi-subtype classification [35,36], which is crucial since treatment strategies strongly depend on the NSCLC subtype. Moreover, the differentiation between lung adenocarcinoma and lung squamous cell carcinoma subtypes has also been investigated in the field of genomics, and it was through a novel explainable AI(XAI)-based deep learning framework [37] that promising results were achieved.

### 1.1. Related Works

In this section, we divided the existing studies related to NSCLC classification problems into two groups: those whose aim was a multiclass classification problem (i.e., four classes: NOS, ADC, LCC, and SCC) and those whose aim was a binary classification problem (i.e., two classes: SCC and ADC). We summarize their main characteristics in Tables 1 and 2. Meanwhile, the comparison between machine learning methods and results will be illustrated in Section 4.1. Moreover, we limit the comparison to studies that only used classical machine learning models, as we did not use deep learning in our study. Table 1 shows that two studies that evaluated binary classification were multicenter studies [33,34], but none of them investigated the presence of batch effects, and neither performed feature harmonization. Moreover, certain details regarding image pre-processing were lacking in all the studies—except in Haga et al. [31]—which was particularly related to the software that was used for the feature extraction and for the extracted features. For example, discretization was needed for the extraction of texture features [12], but discretization parameters were lacking in [30,32–34]. The only study that correctly reported the discretization parameters (bin count 225, bin size 25 HU) was [31]. Moreover, in certain studies [32–34], whose filtering and/or wavelet decomposition were applied, the parameters regarding filters and the method for wavelet decomposition were lacking. Indeed, only two studies [30,31] reported the method that was used for wavelet decomposition (namely, *coiflet*). Furthermore, the isotropic voxel dimension was correctly reported in all the studies in which images were resampled, but only in Song et al. [34] was the interpolator (namely, bilinear interpolation) specified. Neither of the studies [35,36] in Table 2 specified all of the parameters that were used for the pre-processing. Indeed, in Khodabakhshi et al. [36], the isotropic voxel (1 mm<sup>3</sup>), discretization parameters (64 bins), and sigma (0.5–5; 0.5 step) that were used for the LoG filter were specified, but the methods used for the wavelet decomposition and the spatial resampling were missing. Meanwhile, in Liu et al. [35], the wavelet decomposition method (namely, *coiflet*) was specified, but no information about discretization or feature extractor were given. Moreover, both studies were not multicenter studies, because only one dataset was used in [36], and two datasets, but coming from the same center (i.e., the MAASTRO clinic), were used in [35]. Finally, none of the studies that were based on PyRadiomics as the extractor specified if the remaining parameters were set to PyRadiomics' default values.

**Table 1.** The binary classification studies (SCC vs. ADC). M: multicenter, B: batch analysis, and H: harmonization. The number in the round brackets in the dataset column refers to the number of patients that were involved after a selection from the total available amount.

Work	Dataset	Modality	Feature Extraction	M/B/H
[30] (2016)	1 public dataset [38] (192) and lung 2 (152)	CT	Pre-processing: not all specified Extractor: Matlab 2012 Extracted Features: shape, 1st-order statistics, texture features (440)	No/no/no
[31] (2017)	Private dataset (40)	CT	Pre-processing: all specified Extractor: Mvalliers package Extracted Features: shape, 1st-order statistics, texture features (476)	No/no/no
[32] (2021)	Private dataset (1419)	PET + CT	Pre-processing: not all specified Extractor: PyRadiomics Extracted Features: 1st-order statistics, texture Features (688)	No/no/no
[33] (2021)	Private dataset (302) and 2 public datasets [38] (203), [39] (140)	CT	Pre-processing: not all specified Extractor: PyRadiomics Extracted Features: shape, 1st-order statistics, texture features (788)	Yes/no/no
[34] (2023)	8 public datasets [38–45] (868).	CT	Pre- processing: not all specified, interpolator specified Extractor: PyRadiomics Extracted Features: shape, 1st-order statistics, texture features (1409).	Yes/no/no
ours	2 public datasets: [38,39] (302)	CT	Pre-processing: all specified Extractor: PyRadiomics Extracted Features: shape, 1st-order statistics, texture features (1433)	Yes/Yes/Yes

**Table 2.** Multiclass classification studies. M: multicenter, B: batch analysis, and H: harmonization. The number in the round brackets in the dataset column refers to the number of patients that were involved after a selection from the total available amount.

Work	Dataset	Modality	Feature Extraction	M/B/H
[35] (2019)	2 public datasets [38] (278), [45] (71)	CT	Pre-processing: not all specified. Extractor: not specified Extracted Features: shape, 1st-order statistics, texture features (440)	No/no/no
[36] (2021)	1 public dataset: [38] (354)	CT	Pre-processing: not all specified, Extractor: PyRadiomics Extracted Features: shape, 1st-order statistics, texture features (1433)	No/no/no
ours	2 public datasets: [38,39] (466)	CT	Pre-processing: all specified Extractor: PyRadiomics Extracted Features: shape, 1st-order statistics, texture features (1433)	Yes/Yes/Yes

### 1.2. Research Motivation and Contribution

Even if IBSI [12] provides guidelines on how to compute radiomics features, they are still strongly influenced by the software that is used and the pre-processing parameters. This aspect, together with a non-exhaustive report of the extraction parameters, makes radiomics studies less robust and more difficult to reproduce. Moreover, few studies have focused on the multiclass classification of NSCLC subtypes, while many of them only focused on differentiating between the two subtypes. Furthermore, only a few studies (e.g., [34]),

investigated the potentiality of models that were based on multicenter data, which we strongly believe is the next frontier for the creation of more robust and generalizable models. To the best of our knowledge, this is the first study which investigated the presence of batch effects and the impact of feature harmonization on multicenter radiomics CT-based models for the phenotyping of four different NSCLC subtypes.

Following this last line of research, this study

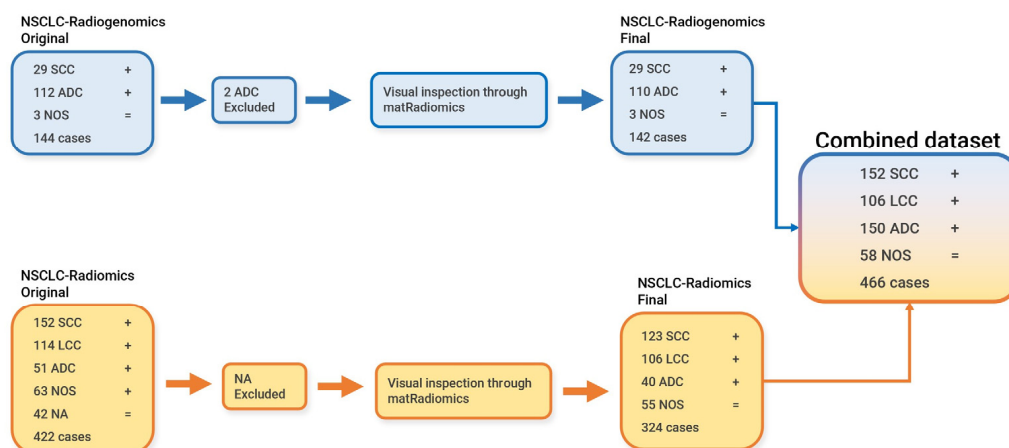
- i. Investigates the possibility of building a multiclass multicenter radiomics-based machine learning model, which is capable of differentiating between the NSCLC subtypes with the aim of improving treatment personalization;
- ii. Evaluates the presence of batch effects in a multicenter study with the aim to assess the impact of feature harmonization on machine learning models;
- iii. Evaluates the stability of the feature selection procedure by iterating the machine learning modeling pipeline 10 times with the aim of reducing the result variability;
- iv. Shows that selected feature subsets are influenced by training/testing set splitting;
- v. Provides a scientific and critical analysis of the advantages and disadvantages of radiomics in the absence of well-defined standard guidelines with the aim of promoting the need for reproducible and repeatable radiomics studies.

The article is organized as follows: the used datasets, image segmentation, image pre-processing, feature extraction, batch analysis, feature harmonization, and the machine learning modeling pipeline are all described in the Section 2. The feature stability, feature harmonization, and model performance results are described in the Section 3. The Supplementary Figures and Tables are provided in the Supplementary Materials. The Sections 4 and 6 provide explanations for the obtained results in comparison with previous studies, and with the current issues of radiomics. A recap of the relevant feature extraction parameters is given in Appendix A, while the acronyms are reported in Appendix B.

## 2. Materials and Methods

### 2.1. Data Preparation and Image Segmentation

The medical images used in this study were obtained by combining the data from two publicly available datasets, namely the NSCLC-Radiomics [38] dataset and the NSCLC-Radiogenomics dataset [39]. The data preparation workflow is illustrated in Figure 1.



**Figure 1.** Data preparation workflow.

### 2.2. NSCLC-Radiomics Dataset

The NSCLC-Radiomics dataset contains CT images from 422 non-small-cell lung cancer patients and their associated segmentations, both in the digital imaging and communication in medicine (DICOM) format. Meanwhile, the histopathological information is provided in a .xls file. The segmentations were manually performed by experts [38]. The original dataset contains 152 patients who were diagnosed with SCC, 114 patients



that were diagnosed with LCC, 51 patients who were diagnosed with ADC, 63 patients specified as NOS, and 42 patients who were specified as NA (because the diagnosis was not available). The NA cases were excluded from the original dataset. Using the matRadiomics software [18], a visual inspection of all the cases and their associated segmentation was carried out. Due to the quality of images and segmentations, the total number of patients was further reduced to 324, composed of the 123 SCC cases, 106 LCC cases, 40 ADC cases, and the 55 NOS cases.

### 2.3. NSCLC-Radiogenomics

The NSCLC-Radiogenomics dataset contains computed tomography images from 144 non-small-cell lung cancer patients and their associated segmentations, both in the DICOM format. Meanwhile, the histopathological information is provided in a .csv file. The segmentations were performed through a semiautomatic algorithm and were manually refined by an expert [39]. The original dataset contains 29 patients who were diagnosed with SCC, 112 patients diagnosed with ADC, and 3 patients specified as NOS. Two patients that belonged to the ADC class were excluded from the original dataset after visual inspection through matRadiomics. Therefore, we only considered a total of 142 patients.

#### 2.3.1. CT Images' Pixel Spacing, Slice Thickness, Matrix Dimension, and Manufacturers

Since the images were acquired in different institutions—i.e., the MAASTRO clinic [46] (NSCLC-Radiomics dataset), the Stanford University Medical Center, and the Palo Alto Veterans Affairs Healthcare System (VA) (NSCLC-Radiogenomics dataset)—different scanning devices were used. Therefore, we inspected the DICOM attributes related to each CT scan through the matRadiomics software to investigate if differences in the pixel spacing, slice thickness, and matrix dimensions were present [47]. The data are shown in Table 3.

**Table 3.** The pixel spacing (mm), slice thickness (mm), and matrix dimension of each CT scan.

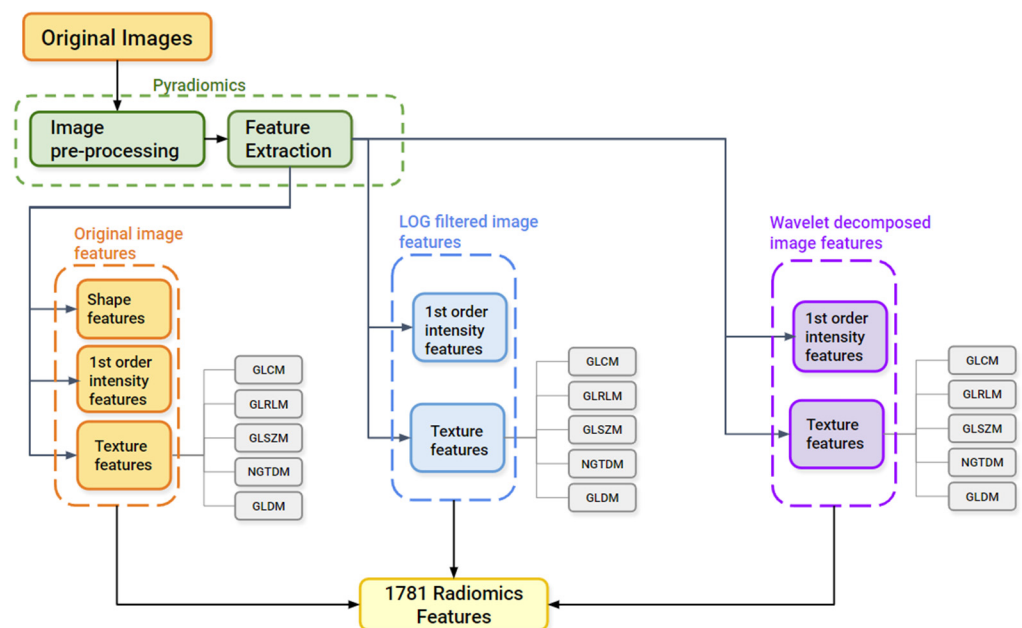
Datasets	Pixel Spacing [x, y] (n°)	Slice Thickness [z] (n°)	Matrix Dimension [x, y] (n°)
NSCLC-Radiomics	[0.97656250, 9765625] (247)	3 (247)	[512 × 512] (247)
NSCLC-Radiomics	[0.9770, 0.9770] (77)	3 (77)	[512 × 512] (77)
NSCLC-Radiogenomics	In a range between [0.589844, 0.976562] (142)	In a range between [0.625, 3] (142)	[512 × 512] (142)

As shown in Table 3, the differences in pixel spacing and slice thickness were individuated. Moreover, the 247 CT scans with [0.9765625, 0.9765625] pixel spacing were acquired using Siemens devices, while the 77 CT scans with [0.9770, 0.9770] pixel spacing were acquired using CMS devices, as reported in the manufacturer DICOM attribute (tag 0008, 0070); furthermore, they all belonged to the NSCLC-Radiomics dataset. Since in the NSCLC-Radiogenomics dataset voxel spacing and slice thickness values varied greatly for each patient, the division in the groups was not carried out and the values were reported within a range. In this case, the images of the 142 patients were acquired using Siemens, GE Medical Systems, and Philips devices. Therefore, four different device manufacturers were individuated: Siemens, CMS, GE Medical Systems, and Philips.

#### 2.3.2. Image Pre-Processing and Feature Extraction

Using matRadiomics, 1781 radiomics features were extracted from each of the patients included in this study. All the quantitative metrics can be grouped into three major classes: (i) shape/morphological features, (ii) first order statistical features, and (iii) texture features, which are the gray level co-occurrence matrix (GLCM), gray level run length

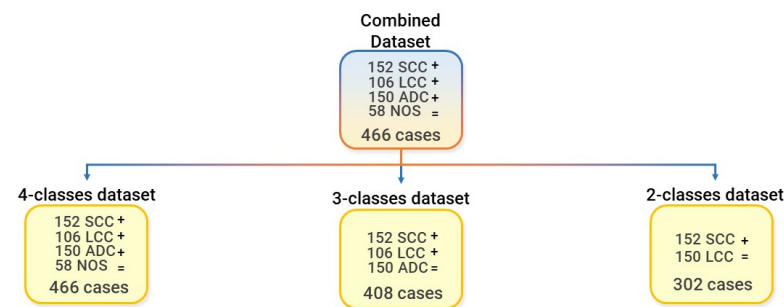
matrix (GLRLM), gray level size zone matrix (GLSZM), neighboring gray tone difference matrix (NGTDM), and the gray level dependence matrix (GLDM)). Since matRadiomics uses the PyRadiomics [19] package to perform feature extraction, the image pre-processing parameters were set according to a previous study [36]. The parameters used for image pre-processing are reported as follows: the voxels were resampled to an isotropic voxel ( $1 \times 1 \times 1 \text{ mm}^3$ ), using the *sitkLinear* interpolator, while the gray levels were discretized into 64 bins using the *bin count* option. Prior to feature extraction, the LoG was used to filter the original images with different *sigma* values (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5), and the 8 wavelet Haar transform was used to decompose (HHH (high-high-high), HHL (high-high-low), HLH (high-low-high), LHH (low-high-high), LLL (low-low-low), LHL (low-high-low), HLL (high-low-low), and LLH (low-low-high)) the original images. All the other parameters were left to PyRadiomics' default (<https://pyradiomics.readthedocs.io/en/latest/customization.html#feature-extractor-level>, accessed on 1 January 2023). A recap of all the feature extraction parameters is provided in Table A1 in Appendix A. The application of the LoG filter and the 8-wavelet decomposition made it possible to obtain the first order statistical features, as well as the texture features that were computed both on the LoG-filtered images and on the decomposed images, not only on the original images. The feature extraction workflow is schematically illustrated in Figure 2.



**Figure 2.** Feature extraction workflow.

### 2.3.3. Dataset Preparation for Batch Analysis, Feature Harmonization, and for the Machine Learning Modeling Pipeline

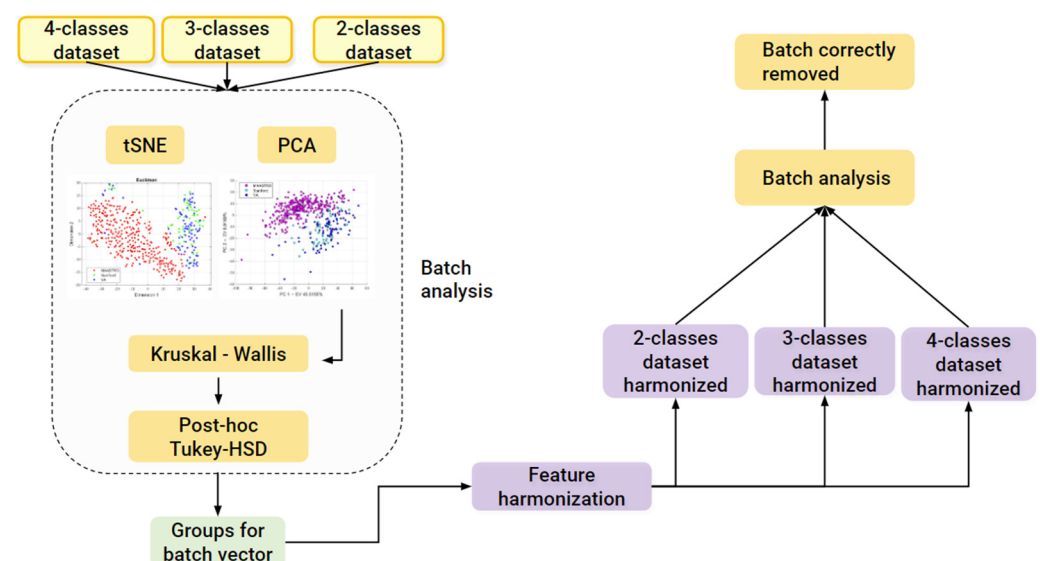
After feature extraction, the combined dataset, containing the 1781 features that were extracted for each of the 466 patients, was further subdivided into three datasets. The first one, namely the “4-classes dataset (4-c)”, contains all the classes (SCC, LCC, ADC, and NOS) of the combined dataset, and essentially coincides with it. The second one, namely “3-classes dataset (3-c)”, contains only three classes (SCC, LCC, and ADC), while the third one, namely “2-classes dataset (2-c)”, contains only the most numerous classes (SCC and LCC). The final dataset subdivision is shown in Figure 3.



**Figure 3.** Final dataset subdivision into the 4-classes dataset, 3-classes dataset, and the 2-classes dataset.

### 2.3.4. Batch Analysis and Feature Harmonization

Since CT images were obtained in three different centers (MAASTRO, Stanford, and VA) while using four different scanners, we investigated the presence of the batch effects. This analysis was preliminarily conducted on all the datasets (the 4-classes dataset, 3-classes dataset, and the 2-classes dataset) and results were used to know which classes to use in order to build the batch vector, which was further used to perform feature harmonization across all the datasets. Therefore, principal component analysis (PCA) was performed to project the data in a space of reduced dimensions. We further performed the visual inspection of the 2D plot, consisting of only the first and second principal components (PC1, PC2) in order to investigate the presence of clusters. Moreover, we used the t-distributed stochastic neighbor embedding (tSNE) to check for the batch effects using four different distance methods: Euclidean, cityblock, Minkowski, and Chebychev. Finally, the Kruskal–Wallis test was performed to assess if a statistically significant difference was present between the clusters. To assess which groups were significantly different, the Kruskal–Wallis test was followed by the post-hoc Tukey–HSD test. Therefore, to remove the batch effects, we performed feature harmonization through the ComBat package (<https://github.com/Jfortin1/ComBatHarmonization>, accessed on 1 January 2023), which was integrated in matRadiomics and aim of which was to standardize the mean and variance across the batches. Finally, we verified if the batches were correctly removed and whether they performed feature harmonization on all of the datasets. Therefore, three more datasets were generated, the 4-classes harmonized dataset (4-c-h), the 3-classes harmonized dataset (3-c-h), and the 2-classes harmonized dataset (2-c-h). The Figure 4 pipeline shows the batch analysis and feature harmonization workflows.



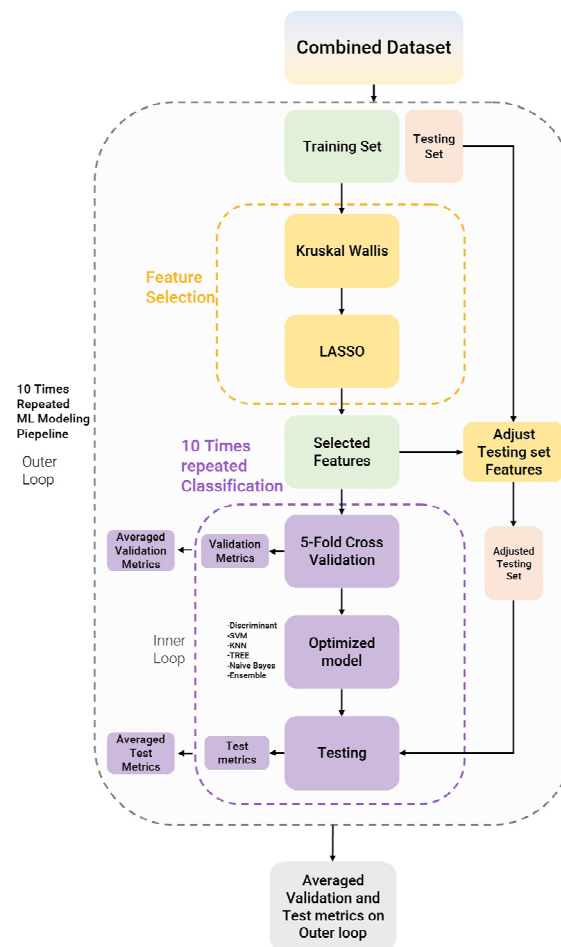
**Figure 4.** Batch analysis and harmonization pipeline.



### 2.3.5. Machine Learning Modeling Pipeline

All the datasets, both harmonized and non-harmonized, were given as the input to the machine learning modeling pipeline. The adopted pipeline consists in repeating the following scheme 10 times (i.e., the outer loop technique): (i) training-set/testing-set stratified splitting (80/20 ratio), (ii) a Kruskal–Wallis analysis followed by the well-known least absolute shrinkage and selection operator (LASSO) [48–50] for feature selection, (iii) model training, (iv) model validation and hyperparameter optimization, and (v) model testing.

The machine learning modeling pipeline is illustrated in Figure 5.



**Figure 5.** Modeling pipeline. The feature selection process is in yellow and the model building process is in purple.

### 2.4. Feature Selection and Feature Stability

Since all the datasets, both non-harmonized and harmonized, are high-dimensional, with a number of features (1781) that is much greater than the number of cases (466, 408 and 302), feature reduction and selection are needed. This procedure was carried out only on the training sets. Moreover, given the results on the training sets, the testing sets were adjusted leaving the selected features and discarding the rest. A Kruskal–Wallis analysis was only used to retain the features with a  $p$ -value that was less than a specified threshold. Therefore, three increasing thresholds were set ( $p$ -value  $< 0.005$ ,  $p$ -value  $< 0.01$ , and  $p$ -value  $< 0.05$ ) and the following scheme was adopted to switch between the thresholds: (i) If no feature meets the first threshold ( $p < 0.005$ ), then the  $p$ -value increases to the second threshold; (ii) if no feature meets the second threshold, then the  $p$ -value increases to the third threshold; and (iii) if none of the features meets the third threshold, then all the features are given as an input to the LASSO.

LASSO, a well-known feature selection algorithm in radiomics, was used to select the most important features. In addition, the optimal value of lambda was obtained through a 10-fold cross validation procedure. Finally, we obtained ten subsets of the selected features, and for each selected feature we computed its frequency. A 100% frequency means that the selected feature appears in each subset of the selected features.

### 2.5. Classification

A 5-fold stratified cross-validation was used for model validation, during which hyperparameters tuning was also performed. Bayesian optimization was selected for quicker hyperparameter tuning. The whole validation procedure was repeated ten times (inner loop) and the performance results were averaged, both on the inner loop and on the outer loop. Then, the optimized models were tested on the testing set and test performance metrics were averaged on the outer loop. Finally, six machine learning models, discriminant analysis (DA) [51], tree [52], K-nearest neighbors (KNN) [53], support vector machines (SVM) [54], Naïve Bayes (NB) [55], and ensemble [52] were trained, validated, and tested. For each model, the accuracy, AUC, sensitivity, specificity, precision, and the f-score were obtained.

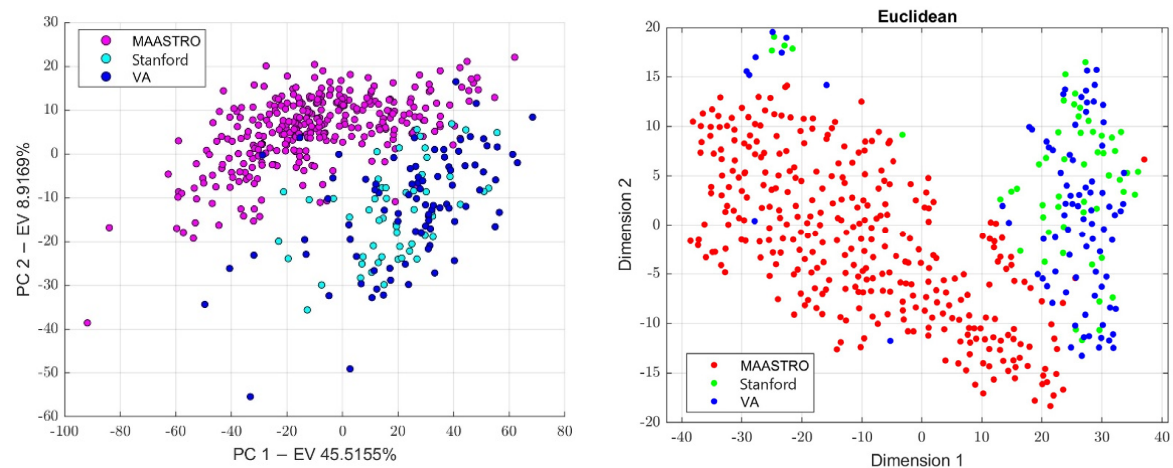
### Software Used for The Radiomics Analysis

matRadiomics [18], an IBSI compliant freeware, was used to perform a visual inspection of the images and segmentation, as well as performing the image pre-processing, feature extraction, and feature harmonization. MATLAB R2022b Update 1 was used to perform batch analysis, feature selection, and classification.

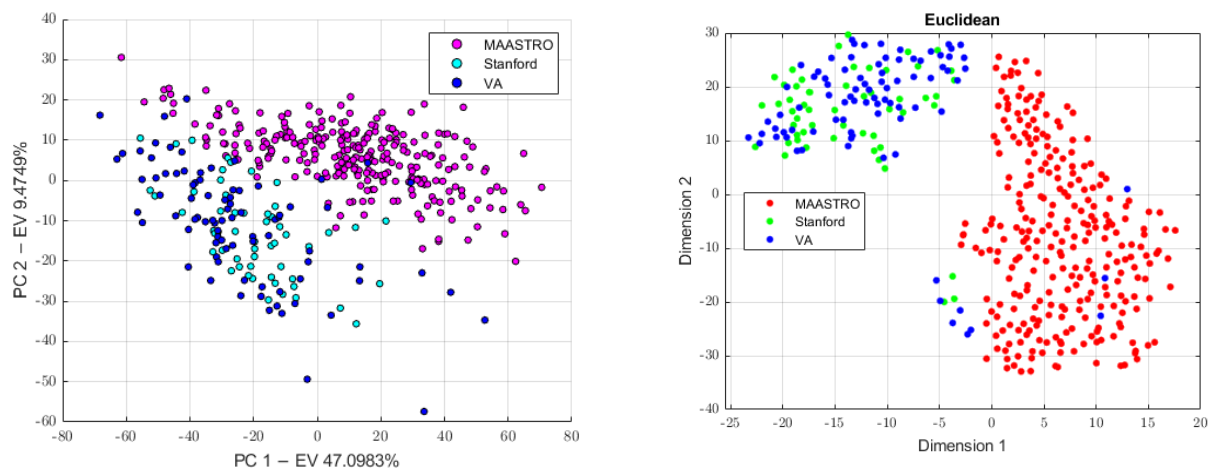
## 3. Results

### 3.1. Batch Analysis and Feature Harmonization

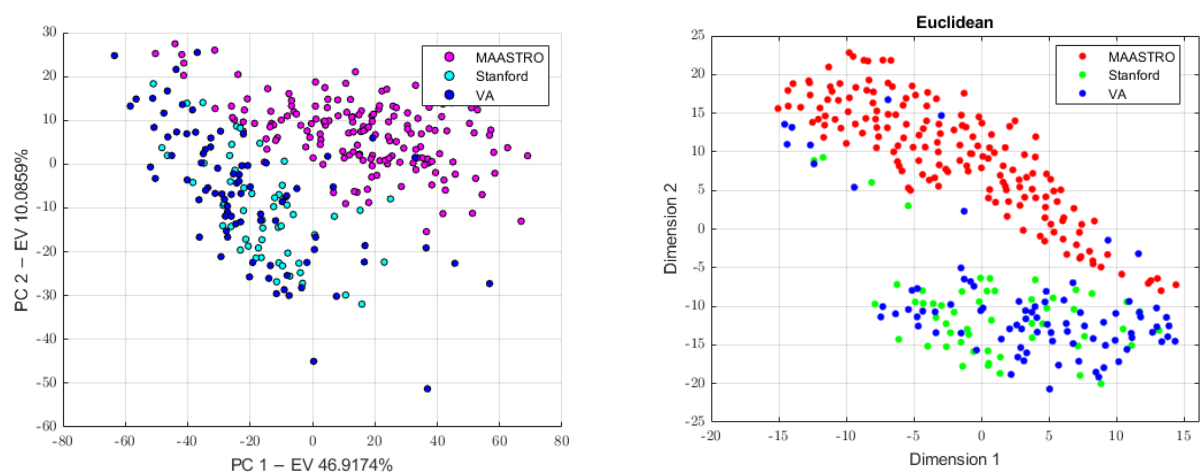
First, we conducted a preliminary batch analysis on all the datasets (4-classes dataset, 3-classes dataset, and 2-classes dataset) performing PCA and tSNE, whose plots show the presence of clusters, as illustrated in Figures 6–8. The Stanford and VA groups were well separated from the MAASTRO group, with only few outliers, while a visual separation between the Stanford and VA groups could not be observed. Therefore, for each dataset, we performed a Kruskal–Wallis test on both of the principal components' data to assess the statistical significance ( $p$ -value threshold = 0.05, null hypothesis, data in each group comes from the same distribution). Since the Kruskal–Wallis test suggested the statistical significance ( $p < 0.05$ )—both on the PC1 and PC2 data for all datasets, as shown in the Supplementary File (see Figures S1–S3)—we performed a post-hoc Tukey–HSD test to identify which groups were significantly different. For all the datasets, the post-hoc test confirmed that both the mean ranks of the Stanford and VA groups were significantly different from the mean rank of the MAASTRO group, both for the PC1 and PC2 data, while no significant difference was present between the mean ranks of the Stanford and VA groups, both on the PC1 (4-classes:  $p = 0.574$ , 3-classes:  $p = 0.6495$ , 2-classes:  $p = 0.5048$ ) and PC2 (4-classes:  $p = 0.783$ , 3-classes:  $p = 0.5010$ , 2-classes:  $p = 0.2289$ ) data. Based on the batch analysis, we constructed the batch vector that was to be used for the feature harmonization. Since the only significant difference was found between the Stanford and MAASTRO groups, and also between the VA and MAASTRO groups we built our batch vector with only two classes, grouping Stanford and VA in a single class (1: MAASTRO and 2: Stanford + VA). Therefore, we performed the feature harmonization and verified if the procedure removed the batch effects (see Figure 9). The Kruskal–Wallis test suggested no significant difference both in the PC1 (4-classes:  $p = 0.137$ , 3-classes:  $p = 0.1193$ , 2-classes:  $p = 0.1665$ ) and PC2 (4-classes:  $p = 0.964$ , 3-classes:  $p = 0.784$ , 2-classes:  $p = 0.3964$ ) data, as is shown in the Supplementary File (see Figures S4–S6).



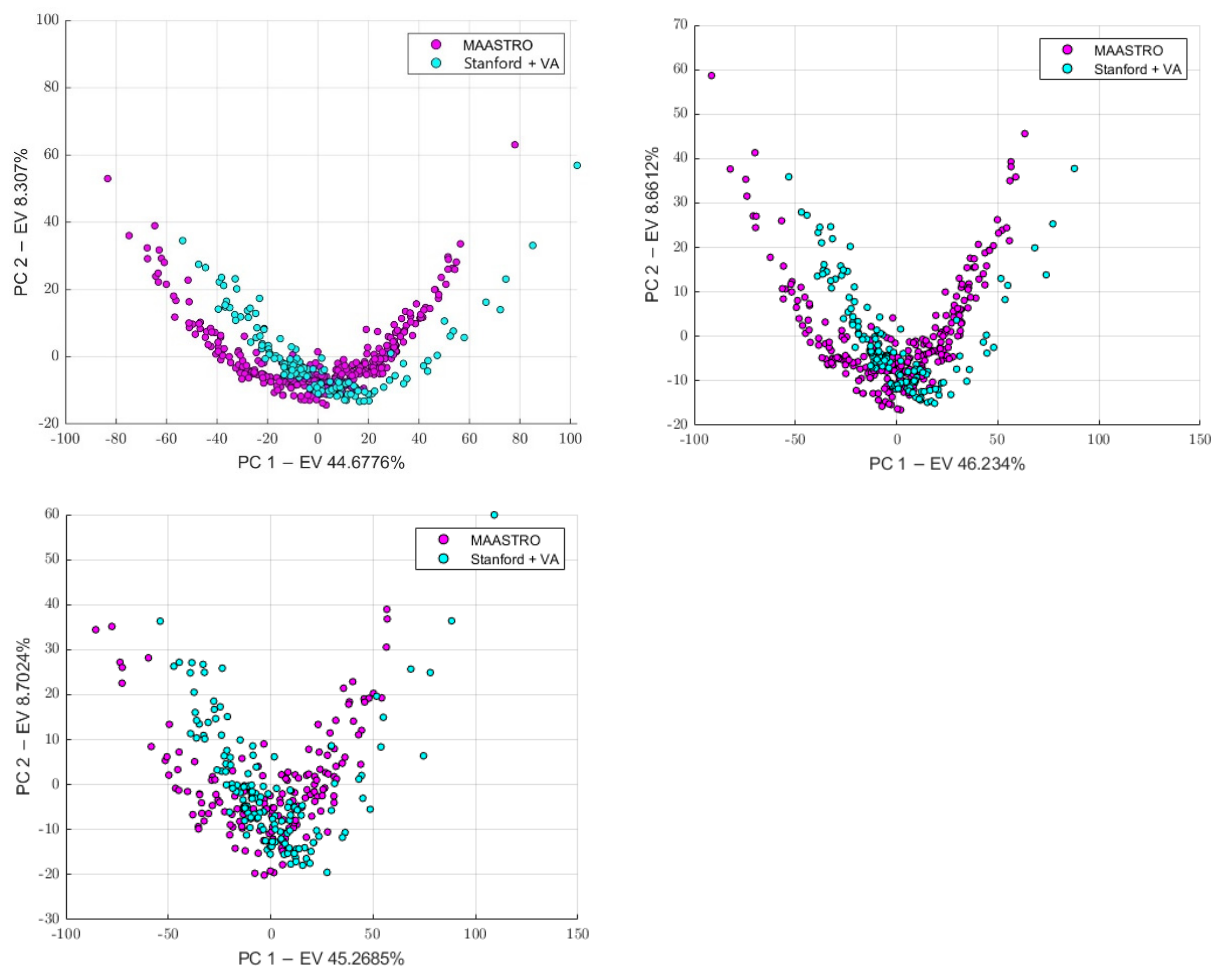
**Figure 6.** The PCA scatter plot on the (left) and the tSNE scatter plot on the (right) for the 4-classes dataset.



**Figure 7.** The PCA scatter plot on the (left) and the tSNE scatter plot on the (right) for the 3-classes dataset.



**Figure 8.** The PCA scatter plot on the (left) and the tSNE scatter plot on the (right) for the 2-classes dataset.



**Figure 9.** The PCA scatter plot after harmonization. The 4-classes are on the **top-left**, the 3-classes are on the **top-right**, the 2-classes are on the **bottom-left**.

### 3.2. Feature Selection and Stability

The feature selection process produced feature subsets of the different sizes at each outer loop iteration. Therefore, we reported, for each dataset, the minimum and the maximum size encountered, as shown in Table 4. Moreover, for each dataset, we reported the 20 features with the highest frequency, as is shown in the Supplementary File (Figures S7–S9). For all the datasets, the selected features had a  $p$ -value  $< 0.005$ . Moreover, the features that had a frequency greater than or equal to 80% belonged in a majority in relation to the texture class. Therefore, 11 features and 9 features in the 4-classes were in the non-harmonized and 4-classes harmonized datasets, respectively. In addition, 4 features were in both of the 3-classes' non-harmonized and harmonized datasets, and 2 features were in both of the 2-classes harmonized and non-harmonized datasets, which were found to have a frequency greater than or equal to 80%. The details are shown in Table 5.

**Table 4.** Minimum and maximum size subsets for each dataset.

Size	4-c-nh	4-c-h	3-c-nh	3-c-h	2-c-nh	2-c-h
Minimum	16	14	9	2	8	1
Maximum	28	22	23	14	22	9

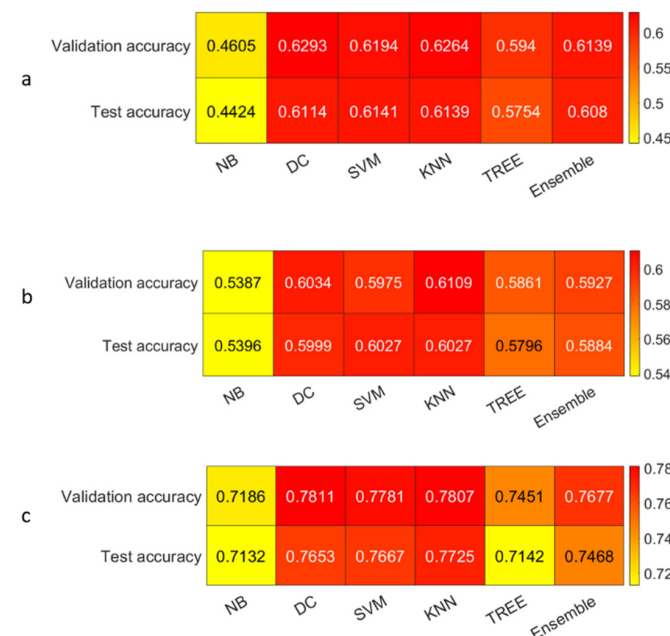
**Table 5.** Selected features with a frequency greater than or equal to 80%.

		4-c-nh	4-c-h	3-c-nh	3-c-h	2-c-nh	2-c-h
Class	Amount	11	9	4	2	2	1
	Shape	3	1	0	0	0	0
	First order	1	1	1	0	0	0
	Texture	<b>7</b>	<b>7</b>	<b>3</b>	<b>2</b>	<b>2</b>	<b>1</b>
Image Type	Original	3	1	0	0	0	0
	LoG	<b>4</b>	1	0	0	0	<b>1</b>
	Wavelet	<b>4</b>	<b>7</b>	<b>4</b>	<b>2</b>	<b>2</b>	<b>0</b>

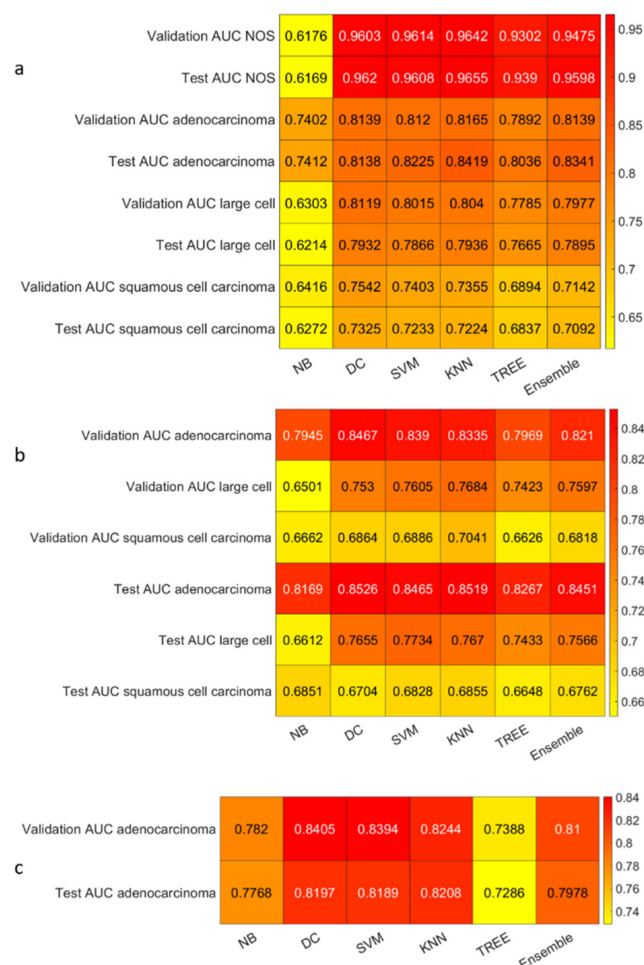
The highest values of interest are highlighted in bold.

### 3.3. Classification

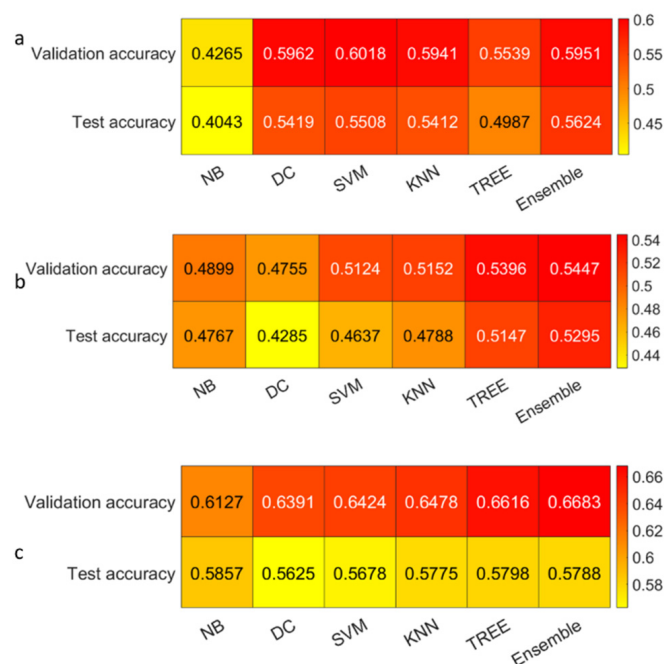
We reported the accuracy and the AUC that were averaged on the outer loop of the machine learning modeling pipeline for all the six different classifiers (DA, KNN, SVM, Naïve Bayes, Tree, and Ensemble), as well as for both of the non-harmonized and harmonized datasets in Figures 10–13. Moreover, we reported in the Supplementary File (see Tables S1–S6), the accuracy, AUC, sensitivity, specificity, precision, and f-score, together with the 95% confidence interval; in addition, the outer loop of the machine learning modeling pipeline for the classifiers that obtained the highest test accuracy was also averaged. Following this, SVM obtained the highest test accuracy ( $0.6141 \pm 0.0317$ ) in the 4-classes dataset; ensemble obtained the highest test accuracy ( $0.5624 \pm 0.0555$ ) in the 4-classes harmonized dataset; KNN obtained the highest test accuracy ( $0.6027 \pm 0.0347$ ) in the 3-classes dataset; ensemble obtained the highest test accuracy in the 3-classes harmonized dataset ( $0.5295 \pm 0.0555$ ); KNN obtained the highest test accuracy in the 2-classes dataset ( $0.7725 \pm 0.0437$ ); and Naïve Bayes obtained the highest accuracy ( $0.5857 \pm 0.0418$ ) in the 2-classes non-harmonized dataset. Furthermore, in Figures 14 and 15, we reported examples of the ROCs (receiver operating characteristics), which were computed for both the harmonized and non-harmonized datasets, and which were also only for the classifiers that obtained the highest test accuracy.

**Figure 10.** Accuracy heatmaps: (a) The 4-classes datasets, (b) 3-classes datasets, and (c) 2-classes datasets.

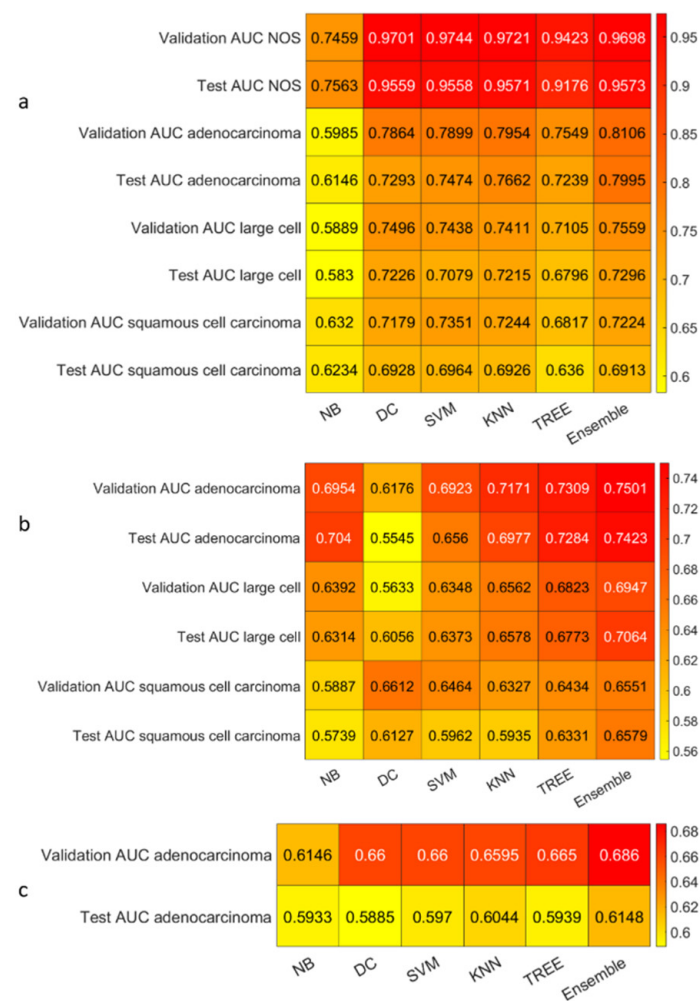




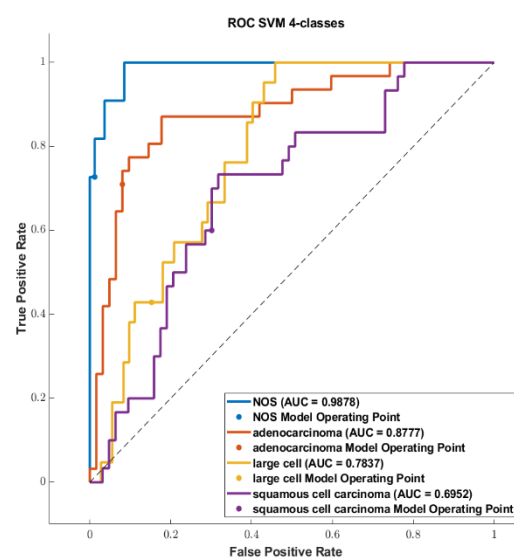
**Figure 11.** AUC heatmaps: (a) The 4-classes datasets, (b) 3-classes datasets, and (c) 2-classes datasets.



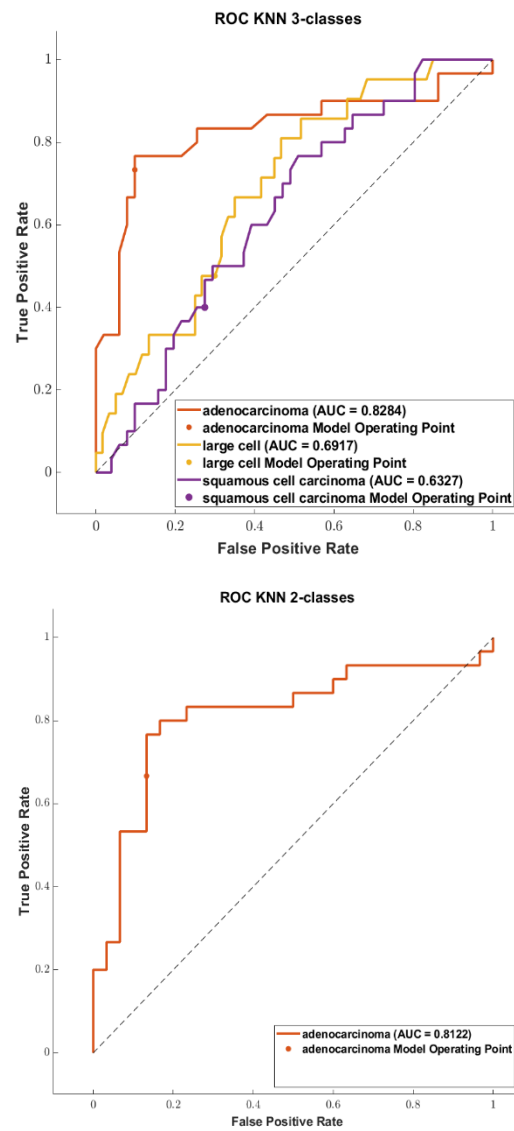
**Figure 12.** Accuracy heatmaps for the harmonized datasets: (a) The 4-classes datasets, (b) 3-classes datasets, and (c) 2-classes datasets.



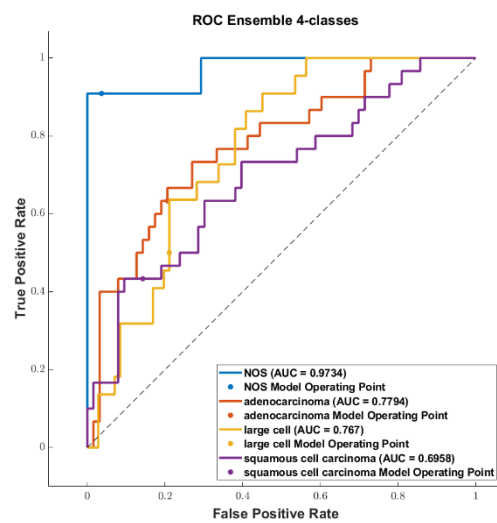
**Figure 13.** AUC heatmaps for the harmonized datasets: (a) The 4-classes datasets, (b) 3-classes datasets, and (c) 2-classes datasets.



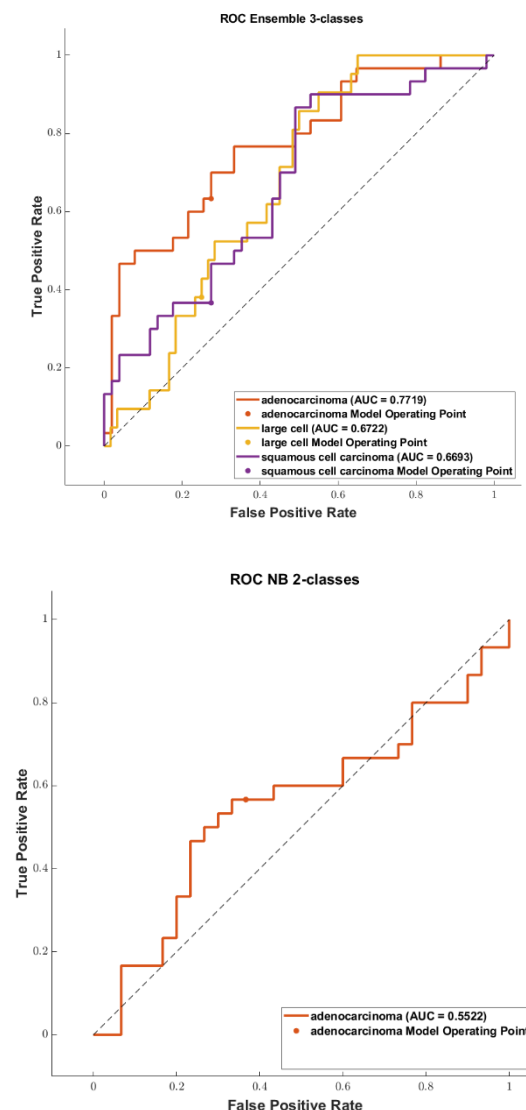
**Figure 14.** Cont.



**Figure 14.** ROC curves: (top) The 4-classes datasets, (middle) 3-classes datasets, and (bottom) 2-classes datasets.



**Figure 15.** Cont.



**Figure 15.** ROC curves for the harmonized datasets: **(top)** The 4-classes datasets, **(middle)** 3-classes datasets, and **(bottom)** 2-classes datasets.

#### 4. Discussion

To date, most studies focusing on the differentiation of NSCLC subtypes have used private datasets or publicly available non-multicenter datasets. Meanwhile, only a few studies have focused on combining the datasets from different centers [33,34], although they did not investigate the presence of batch effects and the impact of feature harmonization on the models' performances. Moreover, many studies have focused on the differentiation of only two NSCLC subtypes (SCC and ADC), without building a multiclass classifier [30–34]. Furthermore, due to the lack of large datasets, certain studies evaluated the performance of the classifiers only on the validation cohort, thereby resulting in a high risk of overfitting and low generalization.

Therefore, considering aims (i) and (ii), which were illustrated in Section 1.2, we investigated the possibility of building a multiclass classifier that would be able to differentiate between the four NSCLC subtypes (NOS, SCC, LCC, and ADC), based on two publicly available multicenter datasets. We evaluated the performance of the models both on the validation and on the testing cohorts, the impact of feature harmonization on models' performance, and the performance variability that arose from adopting our modeling pipeline.

Our hypothesis that multicenter data could be affected by batch effects was confirmed by the experimental results, and, although ComBat feature harmonization is effective in removing it, the models' performance of the harmonized datasets were always lower than those of the non-harmonized datasets. Moreover, a batch effect was present even if the voxel spacing was resampled to be isotropic, meaning that several causes could lead to the presence of batch effects in the multicenter studies, such as the kernel reconstruction and in kVp, among others. Furthermore, the different image segmentation methods adopted in the two datasets—manual for the NSCLC-Radiomics dataset, and semiautomatic with manual refinement for the NSCLC-Radiogenomics dataset—could have contributed to the data clusterization that was shown in the PCA scatter plot. Therefore, the combination of the two datasets may artificially increase the performance of the non-harmonized datasets, because the models learnt to distinguish between the two different segmentation methods. In any case, our experimental results show that the models' accuracy was low. It was not possible to build a reliable and accurate model that was able to distinguish between the four different NSCLC subtypes with our proposed modeling pipeline (highest accuracy = 61.41%), even though an AUC equal to 0.831 was reached (and which was averaged on the 4 classes). Therefore, we decided to simplify the classification problem, reducing the dataset to three, and using only two classes. The worst results were obtained in the 3-classes' datasets, both harmonized and non-harmonized (four classifiers out of six always showed a lower accuracy than those that were built using the 4-classes dataset), while the highest performance was obtained in the 2-classes datasets. In fact, the 2-classes non-harmonized dataset obtained the highest performance (highest accuracy = 77.25%) in comparison with the other datasets. However, similar to that which is mentioned above, this could be due to an artificial increase in performance, which is a result that is removed after feature harmonization (highest accuracy = 0.5857). Moreover, it is evident that the NOS class can be better separated (AUC ~ 0.96%) from the other classes (ADC, LCC, and SCC), and that this could be since ADC, LCC, and SCC have more physical properties in common. Meanwhile, the NOS class groups together all the types that do not have characteristics that are associated with the 3 main classes. The 3-classes' datasets obtained the lowest accuracy because the more separable NOS class was removed.

Furthermore, it is not possible to assess which was the best classifier in absolute terms since this changed depending on the dataset being considered (i.e., 4-classes, 3-classes, 2-classes datasets) in both non-harmonized and harmonized data, as shown in the Section 3.

The third and fourth aims (see Section 1.2) of the proposed study was to evaluate the stability of the feature selection procedure. It is evident that the feature stability decreased as the number of classes decreased, and that the more stable features (i.e., a frequency  $\geq 80\%$ ) were those that belonged to the wavelet and LoG image types and texture class, as is shown in several studies [56,57]. Furthermore, the selected features changed by varying the number of classes and, consequently, the amount of data, even when considering them between the harmonized and non-harmonized datasets. Therefore, it is difficult to establish in absolute terms which features contributed the most to the final models. In general, those with a frequency greater than or equal to 80% were more likely to have contributed the most.

The last aim of this study was to promote the need for reproducible and repeatable radiomics studies. Therefore, in Section 4.1, we discuss our results and the adopted modeling pipeline by comparing them with related works, while also considering the issues regarding the reporting of feature extraction parameters that were already discussed in Sections 1.1 and 1.2.

#### 4.1. Comparison of the Highest Accuracies and AUCs with Related Works

Tables 6 and 7 summarize the existing works related to NSCLC classification problems in two groups: those whose aim is a multiclass classification (i.e., the four classes: NOS, ADC, LCC, and SCC) and those whose aim is a binary classification (i.e., the two classes: SCC and ADC). Table 6 shows that all the modeling methods (feature selection +



classification) reached a test accuracy of  $\geq 0.74$ , except for [31], which reached an accuracy that was equal to 0.656. The highest test accuracy and AUC (0.794, 0.863) were reached in study [32], in which a combination of  $\ell_{2,1}$  norm regularization and linear discriminant analysis was used, while our method (Kruskal–Wallis + LASSO + KNN) obtained the second highest test accuracy and AUC (0.7725, 0.821). The differences between our work and [32] could be found in the machine learning methods adopted, in the dataset that was used (private datasets in [32] and multicenter datasets in our study), and in the images used. This is because they used PET/CT fused images, and due to the validation and testing schemes that were adopted. Indeed, we repeated both the training/testing splitting and the cross-validation procedure to obtain the average metrics that consider training and testing data variability. Moreover, due to having used multicenter data, a more robust performance was guaranteed. Indeed, a robust performance was obtained in the biggest multicenter study [33], in which the third highest accuracy and AUC (0.766, 0.815) were reached. Additionally, in that case, different classifiers were used and the synthetic minority over-sampling technique (SMOTE) was used to balance the training classes, thus leading to data expansion.

**Table 6.** A binary classification (SCC and ADC) comparison between the works, in which machine learning methods, validation and testing schemes, as well as the best averaged results are reported. Acronyms: Gini index (GINI), information gain (IG), gain ratio (GR), minimum description length (MDL), DKM (author names), Laplacian score (LS), spectral feature selection (SPEC),  $\ell_{2,1}$ -norm regularization ( $\ell_{2,1}$ NR), efficient and robust feature selection (RFS), multi-cluster feature selection (MCFS), chi-square score (CSS), Fisher score based on statistics (FS), t-score (TS), redundancy maximum relevance feature selection (mRMR), sequential forward selection (SFS), and least absolute shrinkage and selection operator (LASSO), random forest (RF), Naïve Bayes (NB), Gaussian Naïve Bayes (GNB), K-nearest neighbors (KNN), AdaBoost (AdaB), extreme gradient boosting (XGBoost), bagging (BAG), decision tree (DT), gradient boosting decision tree (GDBT), logistic regression (LR), multilayer perceptron (MLP), linear discriminant analysis (LDA), and support vector machines (SVM).

Work	ML Methods	Training and Testing Sets	Validation and Testing Schemes	Results
[30] (2016)	Selection: correlation + GINI, IG, GR, MDL, DKM, ReliefF + variants. Classification: RF, NB, and KNN	Training: 192 (public dataset) Testing: 152 (Lung 2)	External testing	ReliefFdistance + NB: Test AUC 0.72
[31] (2017)	Selection: univariate analysis $p < 0.05$ + interobserver variation analysis $p < 0.1$ + cross correlation analysis $r < 0.7$ . Classification: NB	Training: 28 (private dataset) Testing: 12 (private dataset)	Training/testing sets splitting: 30 times repeated (70/30)	Test accuracy: 0.656 Test AUC: 0.725
[32] (2021)	Selection: LS, ReliefF, SPEC, $\ell_{2,1}$ NR, RFS, MCFS, CSS, FS, TS, and GINI. Classification: AdaB, BAG, DT, NB, KNN, LR, MLP, LDA, and SVM	Training: 1136 (private dataset) Testing: 283 (private dataset)	Training/testing sets splitting: (80/20). Validation: 10-fold cross validation	$\ell_{2,1}$ NR + LDA: test accuracy 0.794 $\ell_{2,1}$ NR + LDA: test AUC 0.863
[33] (2021)	Selection: mRMR, SFS, and LASSO Classification: LR, SVM, and RF	Since the ratio of training/testing split is not specified (merged dataset), the number of samples used for training and testing is unknown	Training/testing sets splitting: ratio not specified Validation: 5-fold cross validation	Test accuracy 0.74 Test AUC 0.78

Table 6. Cont.

Work	ML Methods	Training and Testing Sets	Validation and Testing Schemes	Results
[34] (2023)	Selection: wrapper $\ell_{2,1}$ norm minimization + 10-fold cross validated LR Classification: LR, SVM, RF, MLP, KNN, GNB, GBDT, AdaB, BAG, and XGBoost	Training: from 560 (5 merged datasets) to 940 after SMOTE. Testing: 140 (5 merged datasets) + external testing (168-3 datasets)	Training/Testing sets splitting: (80/20) Validation: 10-fold cross validation External testing	Bagging-AdaBoost-LR (ensemble): test accuracy 0.766 Bagging-AdaBoost-SVM (Ensemble): test AUC 0.815
ours	Selection: Kruskal Wallis + LASSO Classification: NB, DA, KNN, SVM, TREE, and ensemble	Training: 242 (2 datasets merged) Testing: 60 (2 datasets merged)	Training/Testing sets splitting: 10 times repeated (80/20) Validation: 10 times repeated 5-fold cross validation	KNN: test accuracy 0.7725 KNN: test AUC 0.821

**Table 7.** Multiclass classification comparison between works, in which machine learning methods, validation and testing schemes, and the best averaged results are reported. Acronyms: least absolute shrinkage and selection operator (LASSO), Naïve Bayes (NB), K-nearest neighbors (KNN), discriminant analysis (DA), random forest (RF), and support vector machines (SVM).

Work	ML Methods	Training and Testing Sets	Validation and Testing Schemes	Results
[35] (2019)	Selection: wrapper $\ell_{2,1}$ norm minimization + SVM Classification: SVM	Training: from 279 (public dataset) to 760 after SMOTE Testing: 70 (public dataset)	Training/Testing sets splitting: (80/20) Validation: 10-fold cross validation	NO SMOTE test accuracy 0.67 SMOTE Test accuracy 0.86
[36] (2021)	Selection: Wrapper algorithm, multivariate adaptive regression splines Classification: Multinomial logistic regression	Training: 354 (public dataset)	Validation: 1000 bootstrapping	Validation accuracy: 0.865 Validation AUC: 0.747
ours	Selection: Kruskal Wallis + LASSO Classification: NB, DA, KNN, SVM, TREE, and ensemble	Training: 373 (2 datasets merged) Testing: 93 (2 datasets merged)	Training/Testing sets splitting: 10 times repeated (80/20) Validation: 10 times repeated 5-fold cross validation	DC: validation accuracy 0.6293 SVM: test accuracy 0.6141 DC: validation AUC 0.826 (averaged on 4 classes) KNN: test AUC 0.831 (averaged on 4 classes)

Other than the above, Table 7 shows that the best accuracy was reached in [36] (accuracy: 0.865, AUC: 0.747), but only on the validation set since a testing set was not available. Instead, our accuracy is closer to the accuracy reported in [35] when SMOTE was not used (overall testing accuracy  $\sim 0.67$ ). Our proposed model obtained the lowest accuracy but the highest AUC (0.831) averaged on the 4 classes, and it was also the only multicenter study among them (see Section 1.1). Furthermore, in [35], the training set/testing set split (80/20%) was performed only once, and this could have led to more optimistic training and testing sets, especially when dealing with a not very large dataset. In addition, the SMOTE was used not only to balance the dataset but to also increase the number of samples. Nevertheless, in a real life clinical scenario, data are not balanced, and it is also crucial to test the model on an imbalanced dataset. Moreover, oversampling techniques, such as SMOTE, have certain drawbacks, with overfitting being the most common one [58]. However, the modeling pipeline proposed in our study was different, and the 10-repeated approach (inner/outer loops) was allowed to cope with the overoptimistic performance,

which might be due to a good validation/test split, and provided information on how dataset composition influences selected features subsets, especially when small datasets are used (a common situation in the medical field). Moreover, results could be influenced by pre-processing parameters that strongly influence the feature extraction and model performance [59,60] and, as is shown in Section 1.1, they were not fully reported in the studies of [35,36]. Even if we used the same parameters of study [36] for the isotropic resampling and discretization, differences could be present in the resampling interpolator and in the wavelet method since they were not reported in the study itself.

As can be deduced, reproducibility is one of the major issues in radiomics studies. This is due to the different software, different ML methods, different choice of pre-processing parameters, and the different datasets being used. Moreover, if pre-processing parameters are not fully reported a complete comparison between radiomics studies cannot be carried out.

However, it seems that the accuracies were closer in the binary classification problems among all the studies, rather than in the multiclass classification problem. This could be due to the increased difficulty in differentiating between the four classes.

## 5. Limitation

The first limitation of this study was the dataset dimension. Indeed, even if we adopted all the strategies to reduce overfitting, the larger multicenter datasets are needed, especially for multiclass classification. Moreover, our 4-classes dataset was not balanced, and we did not adopt a strategy, such as SMOTE, to balance it. Furthermore, we showed the presence of the batch effects, but we did not extensively investigate their main cause. This could be due to different kernel reconstruction algorithms and kVp values. We also limited our study to only machine learning techniques while deep learning is becoming much more popular and is also showing promising results in classifying NSCLC subtypes. Indeed, in study [32], it was shown how a VGG-16 convolutional neural network outperformed (accuracy: 0.841, AUC: 0.903) classical machine learning methods in differentiating between the two NSCLC subtypes. Furthermore, in study [61], a self-supervised learning approach reached an AUC that was equal to 0.8641.

## 6. Conclusions

In our study, we investigated the possibility of building a multiclass radiomics model based on multicenter data that was capable of differentiating between the four different NSCLC subtypes. Unfortunately, when adopting our modeling pipeline a low accuracy was obtained (0.6141), thus meaning that an accurate model—which is capable of differentiating between NOS, SCC, ADC, and LCC—could not be built when adopting our modelling pipeline, even if a high average AUC was obtained (0.831) when KNN is used as the classifier. A high accuracy (0.7725) and high AUC (0.821) were achieved when only two classes were considered and when the KNN method was used as the final classifier. The results also showed how it was difficult to identify which was the best classifier capable of differentiating between the NSCLC subtypes since it changed depending on the dataset used (i.e., the 4-classes, 3-classes, and 2-classes) for both harmonized and non-harmonized data. Moreover, to the best of our knowledge, this is the first study that has investigated the presence of batch effects and the impact of feature harmonization on multicenter radiomics CT-based models for the phenotyping of the four different NSCLC subtypes. Therefore, we showed that feature harmonization through ComBat was useful in terms of removing batch effects; however, the performance was found to always decrease, which is an aspect that should be further investigated in future studies. We also showed how isotropic resampling is not enough to homogenize images that come from different centers. Furthermore, as shown in Section 1.1, the comparison between radiomics studies is even more difficult since not all the parameters are fully reported in every study. The results are strongly influenced by the imaging pre-processing techniques, feature selection, and machine learning algorithms used during the radiomics analysis. Indeed, we showed that a modeling pipeline that works well for a task (such as binary classification) may

not work well for another one (such as multiclass classification), thereby limiting the potentiality of radiomics. Moreover, it is important that all the parameters used in the pre-processing techniques are well reported because they have a strong impact on the extracted features. Additionally, dataset composition influences the performance results and selected features. Indeed, through our repeated modeling pipeline we showed that is difficult to find an absolute subset of selected features since it depends on the training/testing splitting. Therefore, we strongly believe that larger multicenter datasets are needed to build reliable models, as well as to reduce overfitting and to improve generalization. As for future directions, a deep investigation of why feature harmonization reduced the models' performance should be investigated, together with the investigation of the main cause that generated the batch effects. Moreover, SMOTE should be introduced and tried in our modeling pipeline for data balancing, with the expectation of an increase in accuracy results. Finally, a deep learning approach using transfer learning and/or the newest self-supervised approaches should be pursued and compared to classical machine learning methods.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/diagnostics13061167/s1>.

**Author Contributions:** Conceptualization, A.S.; Methodology, A.S.; Validation, A.S.; Formal analysis, G.P.; Investigation, G.P.; Resources, A.C.; Data curation, G.P.; Writing—original draft, G.P. and A.S.; Writing—review & editing, G.P. and A.S.; Visualization, G.P.; Supervision, A.S., G.R., A.C., F.M. and F.B.; Project administration, A.S.; Funding acquisition, G.R. 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 data used in this study was downloaded from a public database, so this study did not require the approval of the ethics committee.

**Informed Consent Statement:** Not applicable as the data used in this study was downloaded from a public database.

**Data Availability Statement:** Not applicable.

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

## Appendix A

Here, we provide a recap of the feature extraction parameters.

**Table A1.** Recap of the feature extraction parameters.

Extractor	Pre-Processing	Extracted Features
PyRadiomics v.3.0.1	Voxel resampling: $1 \times 1 \times 1 \text{ mm}^3$	GLCM
	Resampling interpolator: Sitk Linear.	GLSZM
	LoG filter: sigma (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5)	GLRLM
	Wavelet decomposition method: Haar	NGTDM
	Other parameters: PyRadiomics' default.	GLDM

## Appendix B

Here, we provide a table with all the acronyms and their meaning.

**Table A2.** Recap of all the acronyms and their meaning.

Acronym	Stands for
CT	Computed tomography
NSCLC	Non-small-cell lung cancer

**Table A2.** *Cont.*

Acronym	Stands for
SCC	Squamous cell carcinoma
LCC	Large cell carcinoma
ADC	Adenocarcinoma
NOS	No other specified
IBSI	Image Biomarker Standardization Initiative
AUC	Area under curve
LoG	Laplacian of Gaussian
SCLC	Small-cell lung cancer
WHO	World Health Organization
GS	Gleason score
MRI	Magnetic resonance imaging
EDSS	Expanded disability status scale
DICOM	Digital imaging and communication in medicine
GLCM	Gray level co-occurrence matrix
GLRLM	Gray level run length matrix
GLSZM	Gray level size zone matrix
NGTDM	Neighboring gray tone difference matrix
GLDM	Gray level dependence matrix
HHH	High-high-high
HHL	High-high-low
HLH	High-low-high
LHH	Low-high-high
LLL	Low-low-low
LHL	Low-high-low
HLL	High-low-low
LLH	Low-low-high
4-c	4-classes dataset
3-c	3-classes dataset
2-c	2-classes dataset
PCA	Principal component analysis
PC1	First principal component
PC2	Second principal component
tSNE	t-distributed stochastic neighbor embedding
4-c-h	4-classes harmonized dataset
3-c-h	3-classes harmonized dataset
2-c-h	2-classes harmonized dataset
LASSO	Least absolute shrinkage and selection operator
DA	Discriminant analysis
KNN	K-nearest neighbors
SVM	Support vector machines



Table A2. Cont.

Acronym	Stands for
NB	Naïve Bayes
SMOTE	Synthetic minority over-sampling technique
GINI	Gini index
IG	Information gain
GR	Gain ratio
LS	Laplacian score
MDL	Minimum description length
SPEC	Spectral feature selection
$\ell_2,1$ NR	$\ell_2,1$ -norm regularization
RFS	robust feature selection
MCFS	Multi-cluster feature selection
CSS	Chi-square score
FS	Fisher score based on statistics
TS	T-score
mRMR	Redundancy maximum relevance feature selection
SFS	Sequential forward selection
RF	Random forest
GNB	Gaussian Naïve Bayes
AdaB	AdaBoost
XGBoost	Extreme gradient boosting
BAG	Bagging
DT	Decision tree
GDBT	Gradient boosting decision tree
LR	Logistic regression
MLP	Multilayer perceptron
LDA	Linear discriminant analysis

## References

1. Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer statistics. *CA Cancer J. Clin.* **2022**, *72*, 7–33. [\[CrossRef\]](#)
2. Dalmartello, M.; La Vecchia, C.; Bertuccio, P.; Boffetta, P.; Levi, F.; Negri, E.; Malvezzi, M. European cancer mortality predictions for the year 2022 with focus on ovarian cancer. *Ann. Oncol.* **2022**, *33*, 330–339. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Duma, N.; Santana-Davila, R.; Molina, J.R. Non-Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment. *Mayo Clin. Proc.* **2019**, *94*, 1623–1640. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Travis, W.D.; Brambilla, E.; Nicholson, A.G.; Yatabe, Y.; Austin, J.H.M.; Beasley, M.B.; Chirieac, L.R.; Dacic, S.; Duhig, E.; Flieder, D.B.; et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *J. Thorac. Oncol.* **2015**, *10*, 1243–1260. [\[CrossRef\]](#)
5. Xing, P.; Zhu, Y.; Wang, L.; Hui, Z.; Liu, S.; Ren, J.; Zhang, Y.; Song, Y.; Liu, C.; Huang, Y.; et al. What are the clinical symptoms and physical signs for non-small cell lung cancer before diagnosis is made? A nation-wide multicenter 10-year retrospective study in China. *Cancer Med.* **2019**, *8*, 4055–4069. [\[CrossRef\]](#)
6. Thomas, A.; Liu, S.V.; Subramaniam, D.S.; Giaccone, G. Refining the treatment of NSCLC according to histological and molecular subtypes. *Nat. Rev. Clin. Oncol.* **2015**, *12*, 511–526. [\[CrossRef\]](#)
7. Cuocolo, R.; Cipullo, M.B.; Stanzione, A.; Ugga, L.; Romeo, V.; Radice, L.; Brunetti, A.; Imbriaco, M. Machine learning applications in prostate cancer magnetic resonance imaging. *Eur. Radiol. Exp.* **2019**, *3*, 35. [\[CrossRef\]](#)
8. Mayerhoefer, M.E.; Materka, A.; Langs, G.; Häggström, I.; Szczypiński, P.; Gibbs, P.; Cook, G. Introduction to Radiomics. *J. Nucl. Med.* **2020**, *61*, 488–495. [\[CrossRef\]](#)

9. Comelli, A.; Stefano, A.; Coronello, C.; Russo, G.; Vernuccio, F.; Cannella, R.; Salvaggio, G.; Lagalla, R.; Barone, S. Radiomics: A New Biomedical Workflow to Create a Predictive Model. In *Proceedings of the Medical Image Understanding and Analysis*; Papież, B.W., Namburete, A.I.L., Yaqub, M., Noble, J.A., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 280–293.
10. Bharati, S.; Podder, P.; Mondal, M.R.H. Hybrid deep learning for detecting lung diseases from X-ray images. *Inform. Med. Unlocked* **2020**, *20*, 100391. [[CrossRef](#)] [[PubMed](#)]
11. Afshar, P.; Mohammadi, A.; Plataniotis, K.N.; Oikonomou, A.; Benali, H. From Handcrafted to Deep-Learning-Based Cancer Radiomics: Challenges and Opportunities. *IEEE Signal Process. Mag.* **2019**, *36*, 132–160. [[CrossRef](#)]
12. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; 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](#)]
13. Van Timmeren, J.E.; Cester, D.; Tanadini-Lang, S.; Alkadhi, H.; Baessler, B. Radiomics in medical imaging—“how-to” guide and critical reflection. *Insights Imaging* **2020**, *11*, 91. [[CrossRef](#)]
14. Johnson, W.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118–127. [[CrossRef](#)] [[PubMed](#)]
15. Fortin, J.-P.; Cullen, N.; Sheline, Y.I.; Taylor, W.D.; Aselcioglu, I.; Cook, P.A.; Adams, P.; Cooper, C.; Fava, M.; McGrath, P.J.; et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* **2018**, *167*, 104–120. [[CrossRef](#)]
16. Alongi, P.; Stefano, A.; Comelli, A.; Laudicella, R.; Scalisi, S.; Arnone, G.; Barone, S.; Spada, M.; Purpura, P.; Bartolotta, T.V.; et al. Radiomics analysis of 18F-Choline PET/CT in the prediction of disease outcome in high-risk prostate cancer: An explorative study on machine learning feature classification in 94 patients. *Eur. Radiol.* **2021**, *31*, 4595–4605. [[CrossRef](#)] [[PubMed](#)]
17. Cutaia, G.; La Tona, G.; Comelli, A.; Vernuccio, F.; Agnello, F.; Gagliardo, C.; Salvaggio, L.; Quartuccio, N.; Sturiale, L.; Stefano, A.; et al. Radiomics and Prostate MRI: Current Role and Future Applications. *J. Imaging* **2021**, *7*, 34. [[CrossRef](#)] [[PubMed](#)]
18. Pasini, G.; Bini, F.; Russo, G.; Comelli, A.; Marinozzi, F.; Stefano, A. matRadiomics: A Novel and Complete Radiomics Framework, from Image Visualization to Predictive Model. *J. Imaging* **2022**, *8*, 221. [[CrossRef](#)]
19. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [[CrossRef](#)]
20. Tagliafico, A.S.; Piana, M.; Schenone, D.; Lai, R.; Massone, A.M.; Houssami, N. Overview of radiomics in breast cancer diagnosis and prognostication. *Breast* **2020**, *49*, 74–80. [[CrossRef](#)]
21. Cannella, R.; La Grutta, L.; Midiri, M.; Bartolotta, T.V. New advances in radiomics of gastrointestinal stromal tumors. *World J. Gastroenterol.* **2020**, *26*, 4729–4738. [[CrossRef](#)]
22. Russo, G.; Stefano, A.; Comelli, A.; Savoca, G.; Richiusa, S.; Sabini, M.; Cosentino, S.; Alongi, P.; Ippolito, M. Radiomics features of 11[C]-MET PET/CT in primary brain tumors: Preliminary results on grading discrimination using a machine learning model. *Phys. Med.* **2021**, *62*, S44–S45. [[CrossRef](#)]
23. Alongi, P.; Laudicella, R.; Panasiti, F.; Stefano, A.; Comelli, A.; Giaccone, P.; Arnone, A.; Minutoli, F.; Quartuccio, N.; Cupidi, C.; et al. Radiomics Analysis of Brain [<sup>18</sup>F]FDG PET/CT to Predict Alzheimer’s Disease in Patients with Amyloid PET Positivity: A Preliminary Report on the Application of SPM Cortical Segmentation, Pyradiomics and Machine-Learning Analysis. *Diagnostics* **2022**, *12*, 933. [[CrossRef](#)]
24. Shu, Z.; Cui, S.; Wu, X.; Xu, Y.; Huang, P.; Pang, P.; Zhang, M. Predicting the progression of Parkinson’s disease using conventional MRI and machine learning: An application of radiomic biomarkers in whole-brain white matter. *Magn. Reson. Med.* **2021**, *85*, 1611–1624. [[CrossRef](#)] [[PubMed](#)]
25. Shoeibi, A.; Khodatars, M.; Jafari, M.; Ghassemi, N.; Moridian, P.; Alizadehsani, R.; Ling, S.H.; Khosravi, A.; Alinejad-Rokny, H.; Lam, H.; et al. Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: A review. *Inf. Fusion* **2023**, *93*, 85–117. [[CrossRef](#)]
26. Shoeibi, A.; Khodatars, M.; Jafari, M.; Moridian, P.; Rezaei, M.; Alizadehsani, R.; Khozeimeh, F.; Gorriz, J.M.; Heras, J.; Panahiazar, M.; et al. Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. *Comput. Biol. Med.* **2021**, *136*, 104697. [[CrossRef](#)] [[PubMed](#)]
27. Nepi, V.; Pasini, G.; Bini, F.; Marinozzi, F.; Russo, G.; Stefano, A. MRI-Based Radiomics Analysis for Identification of Features Correlated with the Expanded Disability Status Scale of Multiple Sclerosis Patients. In *Proceedings of the Image Analysis and Processing*; Mazzeo, P.L., Frontoni, E., Sclaroff, S., Distant, C., Eds.; ICIAP 2022 Workshops; Springer International Publishing: Cham, Switzerland, 2022; pp. 362–373. [[CrossRef](#)]
28. Gao, A.; Yang, H.; Wang, Y.; Zhao, G.; Wang, C.; Wang, H.; Zhang, X.; Zhang, Y.; Cheng, J.; Yang, G.; et al. Radiomics for the Prediction of Epilepsy in Patients With Frontal Glioma. *Front. Oncol.* **2021**, *11*, 725926. [[CrossRef](#)] [[PubMed](#)]
29. Huang, Y.; Jiang, X.; Xu, H.; Zhang, D.; Liu, L.-N.; Xia, Y.-X.; Xu, D.-K.; Wu, H.-J.; Cheng, G.; Shi, Y.-H. Preoperative prediction of mediastinal lymph node metastasis in non-small cell lung cancer based on 18F-FDG PET/CT radiomics. *Clin. Radiol.* **2023**, *78*, 8–17. [[CrossRef](#)]
30. Wu, W.; Parmar, C.; Grossmann, P.; Quackenbush, J.; Lambin, P.; Bussink, J.; Mak, R.; Aerts, H.J.W.L. Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology. *Front. Oncol.* **2016**, *6*, 71. [[CrossRef](#)]

31. Haga, A.; Takahashi, W.; Aoki, S.; Nawa, K.; Yamashita, H.; Abe, O.; Nakagawa, K. Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: Interobserver delineation variability analysis. *Radiol. Phys. Technol.* **2018**, *11*, 27–35. [\[CrossRef\]](#)
32. Han, Y.; Ma, Y.; Wu, Z.; Zhang, F.; Zheng, D.; Liu, X.; Tao, L.; Liang, Z.; Yang, Z.; Li, X.; et al. Histologic subtype classification of non-small cell lung cancer using PET/CT images. *Eur. J. Nucl. Med.* **2020**, *48*, 350–360. [\[CrossRef\]](#)
33. Yang, F.; Chen, W.; Wei, H.; Zhang, X.; Yuan, S.; Qiao, X.; Chen, Y.-W. Machine Learning for Histologic Subtype Classification of Non-Small Cell Lung Cancer: A Retrospective Multicenter Radiomics Study. *Front. Oncol.* **2021**, *10*, 608598. [\[CrossRef\]](#)
34. Song, F.; Song, X.; Feng, Y.; Fan, G.; Sun, Y.; Zhang, P.; Li, J.; Liu, F.; Zhang, G. Radiomics feature analysis and model research for predicting histopathological subtypes of non-small cell lung cancer on CT images: A multi-dataset study. *Med. Phys.* **2023**, *early view*. [\[CrossRef\]](#)
35. Liu, J.; Cui, J.; Liu, F.; Yuan, Y.; Guo, F.; Zhang, G. Multi-subtype classification model for non-small cell lung cancer based on radiomics: SLS model. *Med. Phys.* **2019**, *46*, 3091–3100. [\[CrossRef\]](#)
36. Khodabakhshi, Z.; Mostafaei, S.; Arabi, H.; Oveisi, M.; Shiri, I.; Zaidi, H. Non-small cell lung carcinoma histopathological subtype phenotyping using high-dimensional multinomial multiclass CT radiomics signature. *Comput. Biol. Med.* **2021**, *136*, 104752. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Dwivedi, K.; Rajpal, A.; Rajpal, S.; Agarwal, M.; Kumar, V.; Kumar, N. An explainable AI-driven biomarker discovery framework for Non-Small Cell Lung Cancer classification. *Comput. Biol. Med.* **2023**, *153*, 106544. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Aerts, H.J.W.L.; Wee, L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; et al. Data From NSCLC-Radiomics 2019. The Cancer Imaging Archive. Available online: <https://doi.org/10.7937/k9/tcia.2015.pf0m9rei> (accessed on 1 January 2023).
39. Bakr, S.; Gevaert, O.; Echegaray, S.; Ayers, K.; Zhou, M.; Shafiq, M.; Zheng, H.; Zhang, W.; Leung, A.; Kadoch, M.; et al. Data for NSCLC Radiogenomics Collection 2017. The Cancer Imaging Archive. Available online: <http://doi.org/10.7937/K9/TCIA.2017.7hs46erv> (accessed on 1 January 2023).
40. Grove, O.; Berglund, A.E.; Schabath, M.B.; Aerts, H.J.W.L.; Dekker, A.; Wang, H.; Velazquez, E.R.; Lambin, P.; Gu, Y.; Balagurunathan, Y.; et al. Quantitative Computed Tomographic Descriptors Associate Tumor Shape Complexity and Intratumor Heterogeneity with Prognosis in Lung Adenocarcinoma. *PLoS ONE* **2015**, *10*, e0118261. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Li, P.; Wang, S.; Li, T.; Lu, J.; HuangFu, Y.; Wang, D. A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis. 2020. Available online: <https://doi.org/10.7937/TCIA.2020.NNC2-0461> (accessed on 1 January 2023).
42. Wee, L.; Aerts, H.J.; Kalendralis, P.; Dekker, A. Data from NSCLC-Radiomics-Interobserver1 2019. Available online: <https://doi.org/10.7937/tcia.2019.cwvlpd26> (accessed on 1 January 2023).
43. Kirk, S.; Lee, Y.; Kumar, P.; Filippini, J.; Albertina, B.; Watson, M.; Rieger-Christ, K.; Lemmerman, J. Radiology Data from The Cancer Genome Atlas Lung Squamous Cell Carcinoma [TCGA-LUSC] Collection. 2016. Available online: <https://doi.org/10.7937/k9/tcia.2016.tygkfmq> (accessed on 1 January 2023).
44. Albertina, B.; Watson, M.; Holback, C.; Jarosz, R.; Kirk, S.; Lee, Y.; Rieger-Christ, K.; Lemmerman, J. Radiology Data from The Cancer Genome Atlas Lung Adenocarcinoma [TCGA-LUAD]. 2016. Available online: <https://doi.org/10.7937/k9/tcia.2016.jgnihep5> (accessed on 1 January 2023).
45. 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. Data From NSCLC-Radiomics-Genomics. 2015. Available online: <https://doi.org/10.7937/k9/tcia.2015.l4fret6z> (accessed on 1 January 2023).
46. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Orhac, F.; Frouin, F.; Nioche, C.; Ayache, N.; Buvat, I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. *Radiology* **2019**, *291*, 53–59. [\[CrossRef\]](#)
48. Cho, H.-H.; Park, H. Classification of low-grade and high-grade glioma using multi-modal image radiomics features. In Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Republic of Korea, 11–15 July 2017; pp. 3081–3084. [\[CrossRef\]](#)
49. Zhou, Y.; Ma, X.-L.; Zhang, T.; Wang, J.; Zhang, T.; Tian, R. Use of radiomics based on 18F-FDG PET/CT and machine learning methods to aid clinical decision-making in the classification of solitary pulmonary lesions: An innovative approach. *Eur. J. Nucl. Med.* **2021**, *48*, 2904–2913. [\[CrossRef\]](#)
50. Bertolini, M.; Trojani, V.; Botti, A.; Cucurachi, N.; Galaverni, M.; Cozzi, S.; Borghetti, P.; La Mattina, S.; Pastorello, E.; Avanzo, M.; et al. Novel Harmonization Method for Multi-Centric Radiomic Studies in Non-Small Cell Lung Cancer. *Curr. Oncol.* **2022**, *29*, 5179–5194. [\[CrossRef\]](#)
51. Comelli, A.; Stefano, A.; Bignardi, S.; Russo, G.; Sabini, M.G.; Ippolito, M.; Barone, S.; Yezzi, A. Active contour algorithm with discriminant analysis for delineating tumors in positron emission tomography. *Artif. Intell. Med.* **2019**, *94*, 67–78. [\[CrossRef\]](#)
52. Artzi, M.; Bressler, I.; Ben Bashat, D. Differentiation between glioblastoma, brain metastasis and subtypes using radiomics analysis. *J. Magn. Reson. Imaging* **2019**, *50*, 519–528. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Comelli, A.; Stefano, A.; Russo, G.; Bignardi, S.; Sabini, M.G.; Petrucci, G.; Ippolito, M.; Yezzi, A. K-nearest neighbor driving active contours to delineate biological tumor volumes. *Eng. Appl. Artif. Intell.* **2019**, *81*, 133–144. [\[CrossRef\]](#)

54. Licari, L.; Salamone, G.; Campanella, S.; Carfi, F.; Fontana, T.; Falco, N.; Tutino, R.; De Marco, P.; Comelli, A.; Cerniglia, D.; et al. Use of the KSVM-based system for the definition, validation and identification of the incisional hernia recurrence risk factors. *Il Giornale di Chirurgia. G. Di Chir. J. Surg.* **2019**, *40*, 32–38.
55. Comelli, A.; Stefano, A.; Bignardi, S.; Coronello, C.; Russo, G.; Sabini, M.G.; Ippolito, M.; Yezzi, A. Tissue Classification to Support Local Active Delineation of Brain Tumors. In *Proceedings of the Medical Image Understanding and Analysis*; Zheng, Y., Williams, B.M., Chen, K., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 3–14.
56. Demircioğlu, A. The effect of preprocessing filters on predictive performance in radiomics. *Eur. Radiol. Exp.* **2022**, *6*, 40. [[CrossRef](#)]
57. Stefano, A.; Leal, A.; Richiusa, S.; Trang, P.; Comelli, A.; Benfante, V.; Cosentino, S.; Sabini, M.G.; Tuttolomondo, A.; Altieri, R.; et al. Robustness of PET Radiomics Features: Impact of Co-Registration with MRI. *Appl. Sci.* **2021**, *11*, 10170. [[CrossRef](#)]
58. Abd Elrahman, S.M.; Abraham, A. A review of class imbalance problem. *J. Netw. Innov. Comput.* **2013**, *1*, 332–340.
59. Schwier, M.; van Griethuysen, J.; Vangel, M.G.; Pieper, S.; Peled, S.; Tempny, C.; Aerts, H.J.W.L.; Kikinis, R.; Fennessy, F.M.; Fedorov, A. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Sci. Rep.* **2019**, *9*, 9441. [[CrossRef](#)]
60. LaRue, R.T.H.M.; Van Timmeren, J.E.; De Jong, E.E.C.; Feliciani, G.; Leijenaar, R.T.H.; Schreurs, W.M.J.; Sosef, M.N.; Raat, F.H.P.J.; Van Der Zande, F.H.R.; Das, M.; et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: A comprehensive phantom study. *Acta Oncol.* **2017**, *56*, 1544–1553. [[CrossRef](#)]
61. Tavolara, T.E.; Gurcan, M.N.; Niazi, M.K.K. Contrastive Multiple Instance Learning: An Unsupervised Framework for Learning Slide-Level Representations of Whole Slide Histopathology Images without Labels. *Cancers* **2022**, *14*, 5778. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.