MDPI

Article

Application of nnU-Net for Automatic Segmentation of Lung Lesions on CT Images and Its Implication for Radiomic Models

Matteo Ferrante [1,†,‡], Lisa Rinaldi [2,†,‡], Francesca Botta [1,*], Xiaobin Hu [3], Andreas Dolp [3], Marta Minotti [4,‡], Francesca De Piano [4,‡], Gianluigi Funicelli [4,‡], Stefania Volpe [5,6], Federica Bellerba [7], Paolo De Marco [1], Sara Raimondi [7], Stefania Rizzo [8,9], Kuangyu Shi [10,11], Marta Cremonesi [2], Barbara A. Jereczek-Fossa [5,6], Lorenzo Spaggiari [6,12], Filippo De Marinis [13], Roberto Orecchia [4,14] and Daniela Origgi [1]

1   Medical Physics Unit, IEO European Institute of Oncology IRCCS, via Ripamonti 435, 20141 Milan, Italy
2   Radiation Research Unit, IEO European Institute of Oncology IRCCS, via Ripamonti 435, 20141 Milan, Italy
3   Department of Informatics, Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany
4   Division of Radiology, IEO, European Institute of Oncology IRCCS, via Ripamonti 435, 20141 Milan, Italy
5   Division of Radiation Oncology, IEO European Institute of Oncology IRCCS, via Ripamonti 435,
    20141 Milan, Italy
6   Department of Oncology and Hemato-Oncology, University of Milan, via Festa del Perdono 7,
    20122 Milan, Italy
7   Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, via Ripamonti 435,
    20141 Milan, Italy
8   Clinica di Radiologia EOC, Istituto Imaging della Svizzera Italiana (IIMSI), via Tesserete 46,
    6900 Lugano, Switzerland
9   Faculty of Biomedical Sciences, Università della Svizzera Italiana (USI), via G. Buffi 13,
    6900 Lugano, Switzerland
10 Chair for Computer-Aided Medical Procedures, Department of Informatics, Technical University of Munich,
    Arcisstraße 21, 80333 Munich, Germany
11 Department of Nuclear Medicine, Bern University Hospital, University of Bern, Freiburgstrasse 18,
    3010 Bern, Switzerland
12 Division of Thoracic Surgery, IEO, European Institute of Oncology IRCCS, via Ripamonti 435,
    20141 Milan, Italy
13 Division of Thoracic Oncology, IEO, European Institute of Oncology IRCCS, via Ripamonti 435,
    20141 Milan, Italy
14 Scientific Direction, IEO, European Institute of Oncology IRCCS, via Ripamonti 435, 20141 Milan, Italy
*   Correspondence: francesca.botta@ieo.it
†   These authors contributed equally to the work.
‡   Affiliation at the time of the study.

Abstract: Radiomics investigates the predictive role of quantitative parameters calculated from radiological images. In oncology, tumour segmentation constitutes a crucial step of the radiomic workflow. Manual segmentation is time-consuming and prone to inter-observer variability. In this study, a state-of-the-art deep-learning network for automatic segmentation (nnU-Net) was applied to computed tomography images of lung tumour patients, and its impact on the performance of survival radiomic models was assessed. In total, 899 patients were included, from two proprietary and one public datasets. Different network architectures (2D, 3D) were trained and tested on different combinations of the datasets. Automatic segmentations were compared to reference manual segmentations performed by physicians using the DICE similarity coefficient. Subsequently, the accuracy of radiomic models for survival classification based on either manual or automatic segmentations were compared, considering both hand-crafted and deep-learning features. The best agreement between automatic and manual contours (DICE = 0.78 ± 0.12) was achieved averaging 2D and 3D predictions and applying customised post-processing. The accuracy of the survival classifier (ranging between 0.65 and 0.78) was not statistically different when using manual versus automatic contours, both with hand-crafted and deep features. These results support the promising role nnU-Net can play in automatic segmentation, accelerating the radiomic workflow without impairing the models' accuracy. Further investigations on different clinical endpoints and populations are encouraged to confirm and generalise these findings.

1. Introduction

In Europe, lung cancer is the second most common malignancy in men and the third most common in women with higher incidence rates in developed over undeveloped countries. Non-small cell lung cancer (NSCLC) accounts for 80–90% of lung malignancies, and includes adenocarcinoma, squamous cell carcinoma and large cell carcinoma [1,2].

The usual diagnostic pathway for lung cancer is based on computed tomography (CT) scan of the head, thorax and superior abdomen (to include the adrenals), and PET-CT. The histological confirmation of the presence of malignancy is usually achieved by CT-guided biopsy, when the lesion is in the outer part of the lung, or by fine needle aspiration under endoscopic and ultrasound guidance, when the lesion is in the inner part of the lungs. The latter is also used in cases of positive lymph nodes at PET-CT, to confirm the presence of nodal involvement [3,4].

After diagnosis, the treatment of these malignancies can involve different clinical pathways.

Advanced methods can be applied to the radiological images, especially CT, to derive useful information for the physicians in order to improve diagnostic accuracy and choose the best treatment for each patient [5,6]. Among these radiomics, the calculation of a wide amount of quantitative parameters from medical images has shown promising results for the prediction of therapy outcomes, survival probabilities and other clinical endpoints, including tumour type, stage, mutation status and presence of metastasis [7–12]. In addition, artificial-intelligence (AI) -based methodologies have been increasingly used, either within the radiomic workflow or alone [13–19]. The final aim of these approaches is to mine high-level information from radiological images, which are not visible to the human eye but might be relevant for clinical purpose. Such methodologies are rapidly emerging as complementary to, and in some cases replacing, the previous approaches used for lung image analysis, including the use of semantic features [20] and of computer-aided diagnosis tools [21]. In particular, AI-based techniques appear promising for overcoming the main issues preventing the clinical implementation of the previous and current methodologies, mainly represented by lack of standardisation, generalisability [6,22] and the time resources that would be required for their use in the daily routine.

The current traditional radiomic approach is based on the calculation of descriptors, named radiomic features, from the numerical content of the medical images. These descriptors quantify different properties of the area of the image under investigation, among them shape, signal intensity and texture, and can be analysed for possible association with clinical endpoints using different methodologies, including statistical approaches or machine-learning AI techniques [23]. In the most advanced AI applications, such descriptors can be learnt by deep-learning algorithms rather than being calculated with hand-crafted tools [24–27].

One of the main bottlenecks of radiomic studies is the segmentation of the volume of interest (VOI), typically the lesion in oncological applications. This is a crucial step, since all subsequent operations of the radiomic workflow (feature extraction and model development) are concerned with the VOI only. Segmentation is most often performed manually by one or more expert physicians, and it is a very demanding task, especially for being extremely time-consuming. For this reason, semi-automatic techniques were introduced in the field of medical segmentation in the past decades [28–30]. In these approaches, even though most of the segmentation is performed by an algorithm, human intervention is still required to guide the procedure. As a result, in addition to needing time from a dedicated resource, the semi-automatic segmentation techniques might suffer from an important limitation of the fully manual contouring: the inter- and intra-reader variability [30–34]. AI can play a role in this context as well, with the introduction of fully

automatic techniques [35,36]. Once conveniently trained and validated, such algorithms can drastically reduce the time spent by radiologists in monitoring lesion segmentation, and at the same time would allow the number—and therefore the reliability—of the datasets under investigation to be increased.

Automatic segmentation has in recent years also been investigated for the contouring of lung tumours. In particular, deep-learning algorithms were extensively adopted in the segmentation field [18,37–39] because of their flexibility and generalisability. Most studies focused on different variants of the basic architecture U-Net, a famous convolutional neural network (CNN) that down-samples and up-samples image features sharing information between the two processes at each level [35]. Among these, nnU-Net is currently considered the state-of-the-art framework; it includes ad-hoc automated pre-processing and outperforms most of the other variants in a wide variety of experiments [40]. In addition, nnU-Net is available online, offering the possibility to test it, train it on wider datasets and under different configurations, ultimately promoting generalisability.

The present study fits into this scenario and has two purposes.

The first aim was to train different configurations of nnU-Net on a wide, real-world dataset of CT images acquired from lung cancer patients at different stages in order to identify the configuration with the highest accuracy in comparison to the manual segmentation performed by physicians, taken as reference standard.

The second aim was to assess to what extent the use of automatic segmentation, as a replacement for manual contouring, affects the performance of a radiomic predictive model. To this purpose, radiomic models for the prediction of five-years survival were built on a population of early-stage NSCLC patients, comparing the performances obtained when extracting hand-crafted or deep radiomic features from manual versus automatic segmentation.

2. Materials and Methods

2.1. Patient Population and CT Acquisition

In this study, three different datasets of patients diagnosed with NSCLC were analysed, two retrospectively collected at the European Institute of Oncology (Milano, Italy)—dataset A and dataset B—and one publicly available, dataset C.

Patients included in datasets A and B were staged before pre-treatment by a contrast-enhanced CT scan and a PET-CT, according to the current guidelines. More specifically, dataset A included 270 patients staged up to pT3pN1M0, undergoing surgery soon after acquisition of a diagnostic contrast-enhanced CT at the European Institute of Oncology without pre-operative chemotherapy. The clinical characteristics of this population have been described elsewhere [10]. The information on the acquisition parameters of the CT image can be found in the Supplementary Materials—Table S2.

Dataset B included 217 patients extracted from a database of 261 patients affected by advanced NSCLC, undergoing chemotherapy after the acquisition of a diagnostic contrast-enhanced CT image. Clinical characteristics of this population are reported in Supplementary Materials—Table S1, while CT acquisition information is listed in Supplementary Materials—Table S2. Only patients undergoing CT imaging at the European Institute of Oncology were selected.

The Institutional Review Board of the European Institute of Oncology approved this retrospective study (UID-2078), waiving the need for informed consent.

Dataset C included 412 patients extracted from the *Lung1* public dataset [41] of the NSCLC-Radiomics collection on The Cancer Imaging Archive platform (TCIA) [42]. The full dataset is composed of 422 patients; of these, 10 were excluded due to metadata instability or corrupted labels (i.e., clashes between the images and segmentations in eight cases, respiratory motion artefacts in two cases). Clinical information can be retrieved from [43] and the TCIA website (https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics, accessed on 1 July 2021).

2.2. Segmentation

2.2.1. Manual Segmentation

One lesion per patient was segmented on the axial CT images by manually delineating the border of the lesion slice by slice.

Datasets A and B were contoured by three radiologists with different degrees of experience, after a common agreement on the procedure. The common criteria among radiologists included the adoption of both the lung and the mediastinal visualisation windows (width of 1500 HU and level of −600 HU and width of 350 HU and level of 40 HU, respectively) according to the lesion location and to the contrast with the surrounding tissues; moreover, vessels were excluded and opacity along the lesion edges was included. Segmentation was performed on the AWServer platform (v. 3.2 Ext. 2.0 tool, GE Healthcare, Chicago, IL, USA) and saved in RT structure format.

Original segmentations from dataset C segmentations were downloaded from the online archive and imported on 3D Slicer version 4.10.2 (NIH, Bethesda, MD, USA) [44]. Each segmentation was revised by a radiation oncologist, who edited the volume of interest as needed. Specifically, all nodal areas were excluded from the gross tumour volume to overcome possible inconsistencies between features extracted from the primary tumour and regional lymph nodes. Blood vessels were excluded, as well.

The final masks were then saved in nearly raw raster data (nrrd) format.

2.2.2. Automatic Segmentation: Training and Testing

The nnU-Net, previously developed for automatic segmentation tasks in medical imaging [40], was adopted for the present study. Based on a U-Net architecture, it represents a state-of-the-art framework, improving on previous ones by introducing empirical rules for image pre-processing which allow for the achievement of high performance in the segmentation task. The net is trained by providing a set of images (training set) as input along with the corresponding manual segmentation performed by the physicians, considered as the reference standard to be learnt. During the training phase, the parameters of the network are learnt in order to produce automatic segmentations as similar as possible to the manual reference segmentation. Then, in inference, it is tested on previously unseen images (test set) for which the network provides the automatic segmentation as output without receiving the manual segmentation in input.

The model training and testing was repeated in twelve different configurations to investigate the variation of the model performance with different source data and network architecture. The configurations were chosen by using different combinations of the three datasets (A, B and C) as training or test sets, and varying the spatial resolution of the images and of the manual segmentation masks. The split between training and test sets was performed randomly, without overlaps between them to avoid overfitting. This procedure aimed to increase the intrinsic variability of the tested datasets, and to identify the model yielding the best performance in terms of computing power, training timing and segmentation performance. Images and masks were used either at full resolution (512 × 512) or reduced to half spacing (256 × 256) using the SimpleITK Python library for voxel resampling [45]. Resampling was used to assess whether the computation time and memory footprint could be lightened without losing performances. As soon as more computational power became available during the study, we opted for an approach without a-priori resampling, referring for this step to the specific architecture chosen (fullres or lowres) which internally already performs these operations.

In Table 1, all the configurations investigated in this study are listed along with the description of the training modality (described in detail in the following), the initial image resolution, and the number of cases used for training and testing.

Table 1. Investigated combinations of the three datasets for the training and testing of nnU-Net.

#	Configuration ^ Training Dataset	Testing Dataset	Training Modality	Initial Image Resolution	# pts Training	# pts Testing
1	A	A *	ensemble(2D, 3D fullres)	256 × 256	220	50
2	A + B	A + B *	ensemble(2D, 3D fullres)	256 × 256	296	128
3	A + B	A + B *	ensemble(cascade, 3D fullres)	512 × 512	340	147
4	C	C *	ensemble(2D, 3D fullres)	512 × 512	328	84
5	C	A	ensemble(2D, 3D fullres)	512 × 512	328	79
6	C	B	ensemble(2D, 3D fullres)	512 × 512	328	66
7	C	A + B	ensemble(2D, 3D fullres)	512 × 512	328	147
8	A + B + C	A	ensemble(2D, 3D fullres)	512 × 512	668	80
9	A + B + C	B	ensemble(2D, 3D fullres)	512 × 512	668	67
10	A + B + C	C	ensemble(2D, 3D fullres)	512 × 512	668	84
11	A + B + C	A + B + C *	ensemble(2D, 3D fullres)	512 × 512	668	231
12	B + C	A	ensemble(2D, 3D fullres)	512 × 512	629	270

^ The two "configuration" columns, one for the training and the other for the test sets, report the name of the dataset/datasets from which the images were extracted. * For this configuration, the training and test sets were created starting from the same dataset/datasets, but the patients used for the two sets were not the same, meaning that there was no overlap between the two groups. # stands for "number of".

Regarding training modality, at least two different models (architectures) were trained for each configuration, and the performance of the ensemble model was evaluated. The ensemble configuration combines the outputs of multiple architectures by averaging, for each voxel, the probabilities of belonging to the lesion instead of the background predicted by each architecture. Except for configuration #3, an ensemble (2D, 3D fullres) approach was used, which combines a 2D and a full resolution (fullres) 3D U-Net architecture. In the case of 2D architecture, the training is performed by taking each slice separately as input and using a 2D convolutional kernel. In the 3D architecture, the CT images are analysed considering also the adjacent slices with 3D kernels in order to catch the volumetric (inter-slices) information and thus provide the learning process with the maximum context. In configuration #3, on the other hand, a cascade approach was applied. The cascade architecture consists first of all of a 3D U-Net, trained using down-sampled inputs; the outputs of the segmentation are then up-resampled to the original size and used as additional inputs for another 3D U-Net, which is trained at full resolution. This approach was tested in one configuration only, due to the intensive requirements in terms of computational resources.

In all cases, the network was trained using the nnU-Net "all" flag, except in the case of the "cascade" model. For the latter, the training was repeated five times by operating a five-fold cross validation. This made it possible to test numerous configurations in acceptable times by measuring the variations of the performance.

More information about the nnU-Net parameters and available architecture modalities can be found in the Supplementary Materials (Additional information on the segmentation network (nnU-Net)).

A customised post-processing algorithm—different from the one proposed in the original nnU-Net paper—was developed for the present study and applied to the output of the network. In its original version, the nnU-Net package applies a post-processing, returning only the largest connected component with a non-maxima suppression algorithm. We calculated, instead, the connected components in three dimensions using the Python connected components 3D library (https://pypi.org/project/connected-components-3d/#description, accessed on 7 September 2021) to individuate and separate all the segmented regions, in this way taking into account cases in which the network identifies multiple lesions or false positives. In particular, we separated each connected volume estimated as a lesion and computed its volume. Estimates of the ensemble and 3D approaches were proposed with decreasing confidence in order of volume size, so that it was possible to choose the most suitable segmentation, even when multiple lesions were identified, and discard false positives.

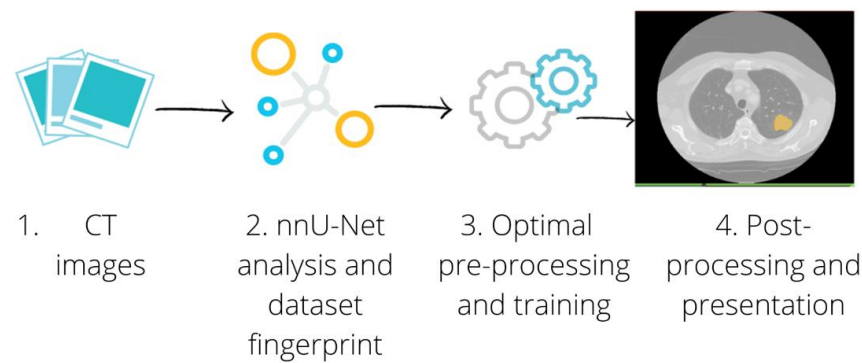A pipeline of the segmentation procedure is shown in Figure 1.

Figure 1. Segmentation Pipeline. The CT images of patients with a lung lesion were collected, along with the manual contours used as the ground truth (**1**). For each configuration we separated the cases into training and test set. The training set was passed through the nnU-Net framework (**2**), which automatically adapted the pre-processing to perform the optimal training (**3**). We finally applied the post-processing to the best performing model to present each connected component provided by the network separately (**4**).

The segmentation network was trained for a minimum of 400 epochs, defined as a 250 batch size as in the original nnU-Net paper on a server with a Nvidia RTX 2080 Ti GPU Card with 11 GB of dedicated RAM memory. Each training modality required between 10 and 20 h to be completed.

2.2.3. Automatic Segmentation Performance

The performance of the automatic segmentation pipeline on each testing dataset was assessed by computing the DICE similarity coefficient, a measure of the degree of volume overlapping between the manual (ground truth) and the automatic segmentation produced by the network [46]. It ranges between 0, when the two segmentation masks do not share common voxels, and 1, when they are perfectly overlapped.

In addition, the percentage of lesions correctly identified by the network was assessed by calculating the number of lesions with a DICE greater than 0 out of the total number of cases in each test set. Similarly, the percentage of lesions with good and excellent correspondence with the ground truth was assessed by calculating the number of lesions with a DICE greater than 0.5, and greater than 0.8 from the total cases in each test set (the values being arbitrarily chosen with reference to the mean DICE value obtained in the different configurations).

2.3. Survival Prediction

Radiomic models for classifying patients according to survival were performed on dataset A, extracting radiomic features from either manual or automatic segmentation, with hand-crafted and deep-learning methodologies, as detailed in the next paragraphs. To this purpose, the automatic contours generated by training the nnU-Net on datasets B and C and by testing on dataset A were adopted (configuration #12 in Table 1). The reason for choosing this configuration was related to the availability of survival follow-up data for patients in dataset A.

Two additional steps were included in the segmentation pipeline: first, among all the identified connected components, the one with the highest DICE, meaning the highest overlap with the ground truth, was selected. Then, all the cases for which the DICE coefficient was lower than 0.3 were excluded from the analysis [47]. These two steps, applied taking advantage of the availability—in this development phase—of the ground truth manual segmentation for all patients, were intended to simulate the role of the physician in a real clinical application of the automatic algorithm, where manual segmentation is not available. In this latter case, indeed, the physician is expected to revise the segmentations proposed by the automatic algorithm, rejecting the cases for which the segmented object is not the

lesion (step 1) and the cases for which the lesion is correctly identified by the automatic algorithm, but contoured in a bad way (the concept of bad contour being translated, in terms of DICE score, as: DICE score < 0.3, step 2 [47]). For this reason, hereafter we will refer to this refinement operation as a "radiologist simulation" procedure.

2.3.1. Hand-Crafted Radiomic Features Extraction

Hand-crafted features were extracted from both the manual and the automatic contours using the open-source tool Pyradiomics (v. 2.2.0) [48]. The features were extracted from the "original" images, meaning that no filter was applied to the CT image before extraction. Moreover, we analysed only features extracted in 2D from each axial slice and averaged among all the slices of the mask, as recommended when the voxels are not isotropic (https://arxiv.org/pdf/1612.07003.pdf).

As part of the pre-processing techniques, pixel size in the axial plane was resampled ("sitkBSpline" interpolation, the default in Pyradiomics), and the grey-level intensities were discretised to a fixed bin width of 25 HU (Hounsfield Units).

The categories of features analysed in this study were: shape, first order, grey level co-occurrence matrix (glcm), grey level run length matrix (glrlm), grey level size zone matrix (glszm), neighbouring grey tone difference matrix (ngtdm) and grey level dependence matrix (gldm). A total of 165 "original" features were available within Pyradiomics. Among these, 12 features were discarded, either because they were intrinsically dependent on the number of voxels inside the VOI or because they actually coincided with other features for the chosen configuration/parameters. More details on the reasons for exclusion, and the full list of the 153 included features can be found in the Supplementary Materials in [49].

For each feature, the interclass correlation coefficient (ICC) [50] was calculated to compare the values extracted from manual versus automatic contours. The ICC ranges between 0 and 1, indicating, respectively, no or perfect agreement between measurements. The *irr* package (v. 0.84.1) in R was used to calculate the ICC. The analysis was also stratified for different DICE coefficient intervals: [0.30, 0.50], [0.50, 0.70], [0.70, 0.80), [0.80, 0.90) and [0.90, 1.00], in order to investigate the dependence of features agreement in relation to the contours' agreement.

2.3.2. Deep-Learning Feature Extraction

A pipeline involving a neural network was developed to select and extract features autonomously in order to compare or integrate the traditional hand-crafted radiomic approach with a deep one.

Due to the limited size of dataset A, a transfer-learning approach was adopted, relying on models pre-trained for more general purposes.

More precisely, a pre-processing pipeline was developed using the Medical Open Network for AI library (MONAI) [51]. First, the image was resampled at a fixed size of 1×1 mm in the axial plane and 1.5 mm in the perpendicular plane (spline interpolation), and then it was cut to a fixed size of $224 \times 224 \times 152$, centred at the centre of the segmented volume, in order to be compatible with the chosen network architecture. Subsequently, the ACSConv library [52], based on PyTorch [53,54], was used to transform a pre-trained 2D model into a 3D model. Basically, the 2D convolutional layers were replaced with their three-dimensional versions by processing the 2D weights to derive the corresponding 3D version, resulting in a three-dimensional pre-trained model, particularly suitable in medical images for being focused on the processing of axial, coronal and sagittal views.

The ResNet152 model [55] was finally applied to extract 2048 deep features from the latest AdaptiveAvgPool3d layer of the PyTorch implementation.

2.3.3. Survival Model Implementation

Survival characteristics of the dataset A population, updated in October 2021, were collected. In particular, for each patient, the survival expressed in the months after the date of the CT examination was recorded. Based on this information, patients were dichotomised

according to 5-year survival [56], and the possibility for predicting the survival class (below/above 5 years) based on radiomic features was investigated using machine-learning (ML) techniques. The radiomic features used as input variables were either the hand-crafted features, deep features, or hybrid features (the latter being a combination of hand-crafted and deep features obtained by concatenating these two groups of features), extracted from either manual or automatic contours. This analysis made it possible to investigate to what extent the use of manual or automatic segmentation can affect the performance of this kind of radiomic-based dichotomous survival model.

A preliminary analysis was performed in order to compare different ML algorithms, and different combinations of hyperparameters for each algorithm, with the purpose of identifying the algorithm and hyperparameters with the best model performance (cross-validation). This preliminary phase involved only features extracted from manual segmentation, since it represents the gold standard segmentation in this study; it considered hand-crafted, deep and hybrid features separately. The investigated algorithms were: Random Forest (RF), Support Vector Machine (SVM) and Multilayer Perceptron models (MLP); all of them incorporate a feature selection step to identify the features mostly associated with the outcome. For each model, 50 hyperparameter combinations were considered. To this purpose, the Weight & Biases (W&B) [57] Python library was used.

Subsequently, once the best ML model and the best combination of hyperparameters were identified according to an accuracy metric, the selected ML model was trained for each of the six cases under analysis: using hand-crafted features, deep features or hybrid features, extracted from either manual or automatic contours, as the input.

Finally, a ten-fold cross validation was performed on the entire dataset in order to obtain a statistic measure of the accuracy. The accuracy results of the ten-fold cross validation for the six cases were compared using a t-test to check if the distributions were statistically different.

In this way, the impact on model performance of using manual versus automatic segmentation and hand-crafted radiomic features versus deep features versus hybrid features was assessed.

The pipeline for the survival model implementation is schematised in Figure 2.
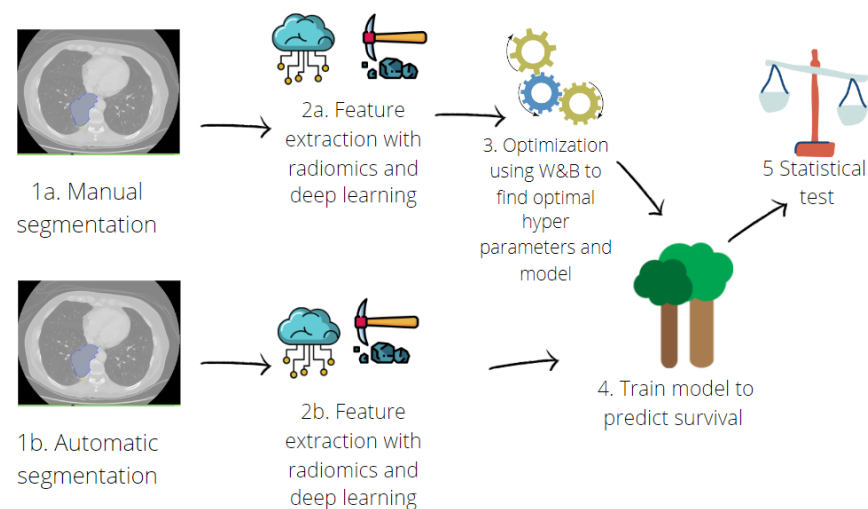

Figure 2. Pipeline for predictive model implementation. Hand-crafted and deep features were extracted (**2a**) from manual contours (**1a**). Then, an optimisation step was performed (**3**) with a random search across the hyperparameters of three different machine learning algorithms: Random Forest, Support Vector Machine and Multilayer Perceptron. Once the best model was found, deep-learning and radiomic features were also extracted from contours obtained with automatic segmentation (**1b**,**2b**). The predictive model was trained (**4**) on features extracted from both manual and automatic segmentation, and a ten-fold cross validation was performed. Finally, performances were compared using a *t*-test (**5**).

The entire code, including the segmentation tool, was developed with Python v 3.8 using popular deep-learning and machine learning libraries, including PyTorch [53,54], MONAI [51] and Scikit-Learn [58].

3. Results

3.1. Manual and Automatic Segmentation

According to the manual segmentation, mean lesion volume was 38.5 cm^3 (range 0.2–511.9 cm^3) for dataset A, 57.2 cm^3 (range 0.1–708.3 cm^3) for dataset B, and 68.2 cm^3 (range 0.5–648.9 cm^3) for dataset C.

The performance of the automatic segmentation pipeline without applying any post-processing is shown in Table 2 for each of the configurations investigated, listed in Table 1. The mean DICE coefficient ($\pm$ standard deviation) and the percentage of lesions correctly identified using the automatic tool (DICE > 0) are reported, along with the percentage of lesions with a good (>0.50) and excellent (>0.80) DICE value. The mean DICE value is calculated considering only the outputs of the network with a DICE > 0. The network correctly identified (DICE > 0) from 82% up to 94% of lesions, depending on the configuration, with an average DICE among all configurations equal to 0.70 (min = 0.65, max = 0.77). For the majority of lesions, the automatic contours achieved a DICE value larger than 0.50 in all configurations (range: 68–82%); a DICE over 0.80 was obtained for less than 50% of the lesions in almost all configurations.

Table 2. Results of segmentation, before post-processing and including only the outputs with a DICE > 0. DICE is intended as mean $\pm$ standard deviation. The last three columns correspond to the number of lesions with a DICE > 0, DICE > 0.5 and DICE > 0.80, respectively, across the total number of test cases.

#	Configuration ^		DICE	% correctly Identified (DICE > 0) Lesions	% Lesions with DICE > 0.50	% Lesions with DICE > 0.80
	Training Dataset	Testing Dataset				
1	A	A *	0.65 $\pm$ 0.29	94%	74%	38%
2	A + B	A + B *	0.74 $\pm$ 0.28	93%	82%	51%
3	A + B	A + B *	0.66 $\pm$ 0.32	93%	71%	41%
4	C	C *	0.68 $\pm$ 0.33	86%	71%	32%
5	C	A	0.69 $\pm$ 0.33	83%	69%	35%
6	C	B	0.71 $\pm$ 0.32	85%	73%	40%
7	C	A + B	0.70 $\pm$ 0.33	84%	71%	37%
8	A + B + C	A	0.71 $\pm$ 0.32	88%	70%	48%
9	A + B + C	B	0.77 $\pm$ 0.31	87%	79%	48%
10	A + B + C	C	0.67 $\pm$ 0.32	83%	68%	31%
11	A + B + C	A + B + C *	0.71 $\pm$ 0.32	86%	72%	42%
12	B + C	A	0.72 $\pm$ 0.29	91%	78%	45%

^ The two "configuration" columns, one for the training and the other for the test sets, report the name of the dataset/datasets from which the images were extracted. * For this configuration, the training and test sets were created starting from the same dataset/datasets, but the patients used for the two sets were not the same, meaning that there was no overlap between the two groups.

The inclusion of dataset C in the pipeline reduced the segmentation performance in terms of percentage of correctly identified lesions. However, the performance in terms of average DICE was similar to the configurations without dataset C. On the other hand, adding a completely independent cohort—as dataset C—made it possible to increase the intrinsic variability of the training set, and thus the generalisability of the network.

Two examples of the output of the segmentation pipeline are reported in Figure 3, with automatic segmentation displayed in red, superimposed onto the manual segmentation (ground truth) displayed in yellow. Figure 3a shows a case with a low DICE (0.43), meaning a not-optimal overlap between the automatic segmentation and the manual one. Figure 3b is instead an example of an excellent DICE (0.88).
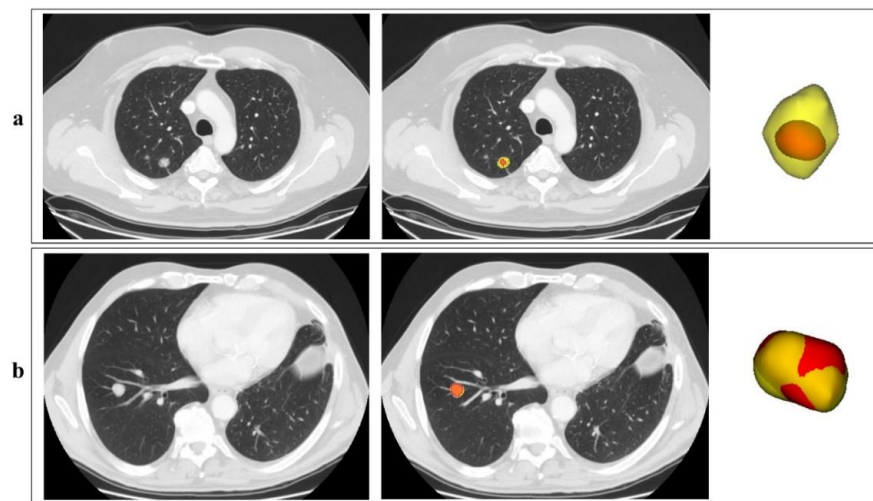
Figure 3. Examples of the automatic segmentation outputs. In the picture, two outputs of the segmentation network, corresponding to two different patients (**a**,**b**), are reported in red, superimposed onto the manual segmentation in yellow. For each patient, one axial CT slice including the lesion is reported, without (**left**) and with (**right**) indication of the lesion contours; the 3D visual representation of the contours is also reported. Although the two lesions are quite similar in shape and volume, the lesion for patient in Figure 3a is characterised by a low-density edge which is not captured by the automatic algorithm.

In the case of configuration #12 (training on datasets B and C, testing on dataset A, used for the subsequent survival analysis), when the post-processing and the "radiologist simulation" procedure were added, an improvement of the DICE coefficient was observed: 90% (242/270) of the lesions had a DICE coefficient greater than 0.3, with a mean DICE equal to 0.78 ± 0.12 (compared to 0.72 ± 0.29 without any post-processing). The distribution of the DICE coefficient over the dataset A in this configuration is shown in Figure 4.
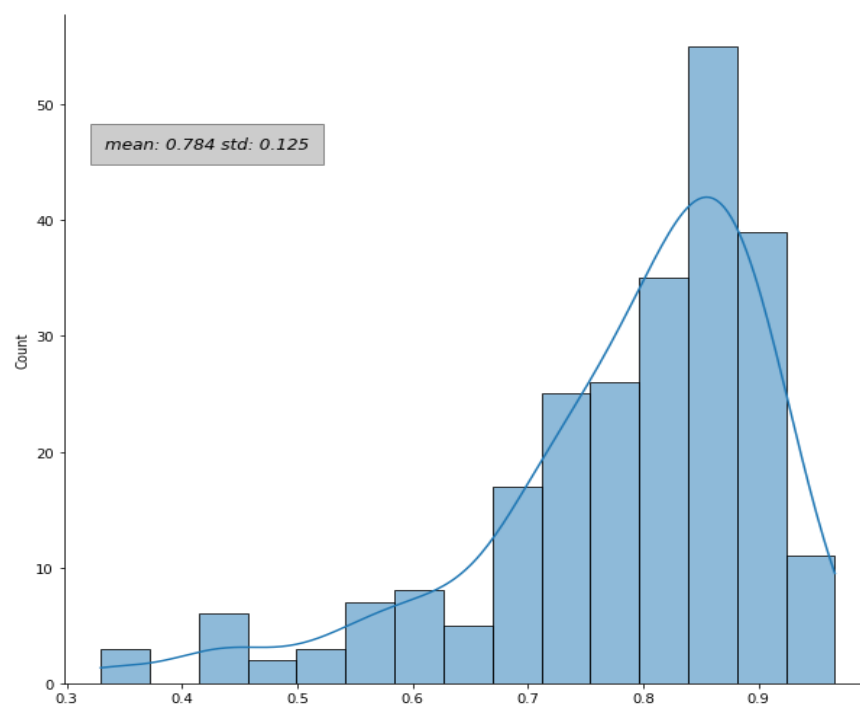
Figure 4. Distribution of the DICE coefficient. The DICE coefficients were evaluated for the patients in dataset A by comparing the automatic and the manual contours in configuration #12. The plot refers to the results after the application of the post-processing and the "radiologist simulation" steps.

To further check the segmentation performance, the volume of the predicted segmentation ($V_{automatic}$) was compared to the ground truth (V_{manual}) by computing the ratio $V_{automatic}/V_{manual}$; a mean value of 1.04 was obtained, indicating that on average the predictions were just 4% apart in terms of volume, with respect to the manual reference.

When comparing the radiomic features extracted from manual versus automatic contours, we observed that 11% (17/153) and 22% (34/153) of the hand-crafted features had an excellent (ICC $\geq$ 0.90) and a good (0.75 $\leq$ ICC < 0.90) agreement between manual and automatic contours. The majority of the features (47%), instead, had a moderate agreement (0.50 $\leq$ ICC < 0.75), and the remaining 19% (29/153) had an ICC value lower than 0.50 (poor agreement).

When analysing the agreement separately in each DICE coefficient range ([0.30, 0.50), [0.50, 0.70), [0.70, 0.80), [0.80, 0.90), [0.90, 1.00]), an improvement of the overall ICC was observed for the group of patients with higher DICE coefficient values. In general, for each category of features analysed, a trend was observed with the ICC increasing as a function of the DICE coefficient. A graphical overview of this behaviour is reported in the Supplementary Materials—Figure S1. Nonetheless, some features might exhibit poor or good agreement irrespective of the DICE value.

3.2. Survival Model

As previously reported, when the post-processing and the "radiologist simulation" procedure were applied to configuration #12, 242 patients from dataset A had a valid segmentation; they were all included in the survival prediction analysis. A total of 132 patients had a survival longer than 5 years, and the remaining 110 fell in the class with survival below 5 years.

Using the W&B library to keep track of the dependence of the hyperparameters on the three different ML models tested, the best performances were achieved with RF. During the optimisation phase, the *n_estimates* (number of trees in the RF model) and a low value of *ccp_alpha* (regularisation parameter for the cost-complexity pruning algorithm) were positively correlated with accuracy in the training set. As a result, a RF classifier with 1000 trees and a ccp_alpha of 0.01 were selected for the implementation of the classification model, which was trained six times in the six different cases according to the type of features (handcrafted/deep/hybrid) and to the type of contour used for feature calculation.

In Table 3, the accuracy of the model achieved using hand-crafted features, deep features, and a combination of them (hybrid) is reported, separately for features calculated on manual or automatic contours.

Table 3. Accuracy results of the survival prediction using the RF model, comparing manual and automatic contours. The model performances are given for the three types of features extracted: hand-crafted features (extracted using conventional radiomics), deep-learning features (extracted using a deep-learning algorithm), and hybrid features (obtained by concatenating the first two types).

	Manual	**Automatic**
Hand-crafted features	0.73	0.78
Deep features	0.65	0.78
Hybrid features	0.70	0.78

The *p*-values of the t-test comparing the accuracies obtained from ten-fold cross validation on the entire dataset are reported in Figure 5. No statistical difference was observed (all $p > 0.05$) between the model accuracy obtained with hand-crafted, deep or hybrid features, for both manual and automatic contours, nor when comparing manual versus automatic contours for a given group of features (hand-crafted, deep or hybrid).
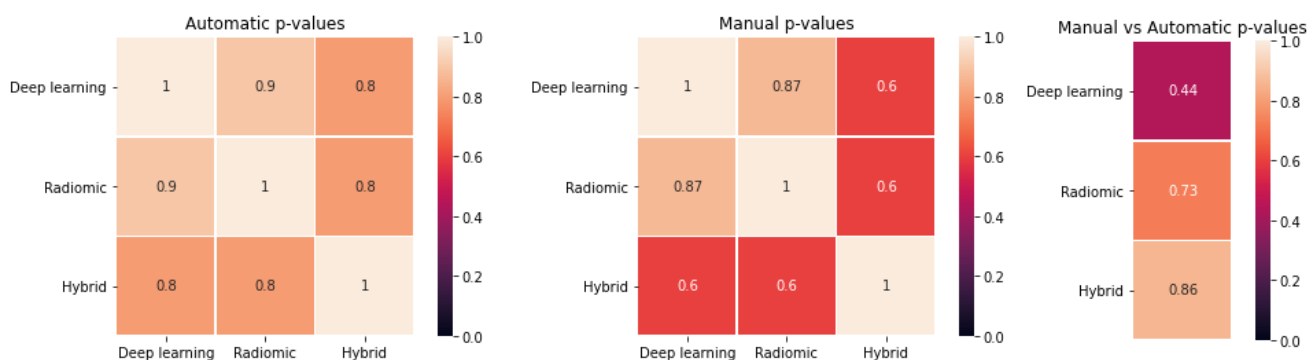
Figure 5. *T*-test results of the comparison between the different approaches. (**Left**): *p*-values for deep-learning versus hand-crafted versus hybrid features for automatic segmentation. (**Centre**): *p*-values for deep-learning versus hand-crafted versus hybrid features for manual segmentation. (**Right**): *p*-values for manual versus automatic in case of deep-learning, hand-crafted and hybrid features.

4. Discussion

In recent years, many studies and methodologies have emerged with the aim of extracting quantitative information from medical images, potentially relevant to predict clinical outcomes. In the near future, some of these might prove their usefulness in simplifying and guiding choices towards personalised treatments. However, to be efficiently integrated into the clinical practice, besides being robust and generalisable they need to be easily implementable on a routine basis.

Segmentation is in most cases an unavoidable and crucial step of these procedures. Automatic algorithms capable of producing reliable segmentations in a short time could drastically reduce the time required for clinicians for this type of activity, allowing access to much larger segmented datasets during the model development phase. In this regard, it might not be crucial to train automatic algorithms segmenting the region of interest with equal accuracy as the manual gold standard. Indeed, the identification of algorithms segmenting the region of interest with enough accuracy to produce comparable predictive models as those obtained with manual segmentation might equally serve the purpose.

In this study, a pipeline based on a state-of-the-art framework, nnU-Net, was developed for automatic segmentation of NSCLC tumours using CT images, testing different network architectures on different dataset configurations. The ability to perform a pre-processing aimed at and suited to the characteristics of a CT thoracic image dataset was exploited, and a post-processing algorithm was optimised for lung lesion segmentation. The automatic contours were compared with manual ones, drawn by experienced physicians. A survival predictive model was then built, based on radiomic features calculated from either manual or automatic segmentations, and the model performances were compared. In addition, we investigated the difference in model performance for the clinical outcome under investigation, when using hand-crafted radiomic features, deep-learning features, or a hybrid approach.

Concerning the automatic segmentation procedure, good results were obtained both for the volume of overlap between the ground truth and the output of the network (evaluated with the DICE coefficient) and for the number of correctly detected lesions. The average DICE among all the dataset configurations was equal to 0.70 and the percentage of correctly identified lesions was higher than 82% in all the investigated cases (Table 2).

These results are largely in agreement with others reported in the literature. Gan et al. [59] created a combination of a 3D and 2D network to segment lung lesions using a total of 260 patients from a private dataset, achieving a mean DICE coefficient of 0.72 ± 0.10. Yang et al. [52] developed an ACS (axial–coronal–sagittal) network and tested it on the LIDC–IDRI dataset of lung nodules [60]. The main idea behind this architecture was to use 2D kernels on the three views separately and then combine them to obtain a 3D output. The best DICE was obtained on the ACS pretrained model architecture, and it was

equal to 0.76, outperforming both the 3D (best DICE equal to 0.75) and 2D (DICE equal to 0.69) pretrained networks. Better results were obtained by Haarburger et al. [47]. In this study, the authors applied a probabilistic segmentation algorithm based on a 2D U-Net, named PHiSeg network [61], to three datasets of different tumours (lung, kidney and liver lesions) and obtained a DICE metric of 0.85 (IQR between 0.77 and 0.89) for the LIDC–IDRI lung dataset.

As previously anticipated, a perfect agreement between different segmentations (DICE equal to 1) might not be fundamental for the purpose of predictive model creation. An automatic segmentation not entirely matching the manual gold standard contour but leading to predictive models with comparable accuracy to those built on manual contours reaches the scope of overcoming the manual contouring limitations, without impairing the final goal for the clinics. We tested this for a specific clinical endpoint, namely the classification of patients according to survival.

To this purpose, one of the configurations analysed in the segmentation pipeline was selected, the one using datasets B and C for training and dataset A for testing.

As a first step, in this specific configuration, the customised post-processing was applied to the masks obtained from the segmentation network, improving the DICE coefficient in the subset of selected lesions from 0.72 ± 0.29 to 0.78 ± 0.12.

The predictive models obtained using hand-crafted, deep or hybrid features calculated from manual or automatic segmentations had comparable accuracy (Table 3). Although the models built on automatic contours provided slightly higher accuracy than those obtained on manual contours (0.78 versus 0.65–0.73), and deep features appeared to perform worse than hand-crafted features in the manual setting (0.65 versus 0.73), a non-significant difference was found (*t*-test *p*-values > 0.05) between the performances obtained during a ten-fold cross validation with the two types of segmentation and the three categories of extracted features (Figure 5).

In the case of hand-crafted features, the comparable models' accuracy was achieved although the differences between the features extracted from manual versus automatic contours were in some cases not negligible, especially for contours exhibiting low DICE coefficient values (Figure S1 in Supplementary Materials).

This was a relevant result proving that a perfect agreement between manual and automatic contours, and the corresponding radiomic features, might not be necessary for the purpose of radiomic-based predictive model creation. This might be related to the fact that while the disagreement between manual and automatic contours occurs mostly on the lesion edge, the features relevant for the clinical outcome of interest might be those capturing the properties of the internal voxels, affected to a lower extent by the outer voxels. Nonetheless, it should be noted that, despite the features' agreement generally increasing when the contours' agreement does, some exceptions might occur, with some features exhibiting good or bad agreement irrespective of the DICE value. In this regard, it might be interesting to repeat the comparison between the performances of manual versus automatic predictive models separating the patients according to the value of DICE coefficient. However, it must be pointed out that this result was obtained for the specific clinical endpoint considered in this study (dichotomous classification of survival) and it was tested on a single, multicentre sample. Generalisability of this finding to other clinical outcomes (e.g., mutational status), or other populations, will be properly investigated in dedicated datasets; the same populations will be used as additional, independent test sets in order to further assess the segmentation performance. In this study, multiple physicians were involved in the contouring of the lung lesions for the three datasets used in the segmentation pipeline. Even if common criteria were established among them, some variability may have been introduced, reducing the segmentation performance of the network. On the other hand, the presence of multiple operators performing reference, manual segmentation in the training set could also represent an advantage for algorithm generalisability. Further studies will be mandatory for the evaluation of the inter-reader variability in manual segmentation, and its impact on the training of automatic tools. Among other limitations, a

relatively small number of images was used for the development of the survival model. Larger datasets might allow us to obtain superior performance. However, in this study we did not focus on the validation of a survival model. Rather, our primary goal was the comparison of the model performance in different scenarios (segmentation modality and/or feature type) when the same group of patients was involved.

5. Conclusions

An automatic tool for lung tumour segmentation in CT images was adopted based on the nnU-Net framework and properly adapted with customised post-processing. After testing different network architectures on multiple datasets, the best model achieved an average DICE coefficient of 0.78 ± 0.12 after the application of ad hoc post-processing techniques, correctly finding and segmenting 90% of the tested lesions.

The radiomic features extracted from thus-obtained automatic contours resulted in survival predictive models having comparable accuracy to the ones obtained extracting features from the reference manual contours (accuracy statistically not distinguishable). In addition, hand-crafted and deep radiomic features provided comparable results in terms of predictive accuracy with both segmentation modalities.

These findings support the idea that segmentation tools based on deep learning can be effectively included in the image analysis workflow, dramatically reducing the physician's workload without impairing the accuracy in comparison to the use of manual segmentation. If confirmed on other populations and for other clinical endpoints, this could simplify access to large datasets and accelerate the identification of reliable tools and their translation to the clinical practice.

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

Abbreviations

CT = computed tomography; NSCLC = Non-small Cell Lung Cancer; PET-CT = positron emission tomography—computed tomography; VOI = volume of interest; HU = Hounsfield unit; ML = machine learning; RF = Random Forest; SVM = Support Vector Machine; MLP = Multilayer Perceptron models; Std = standard deviation.

References

1. Planchard, D.; Popat, S.; Kerr, K.; Novello, S.; Smit, E.F.; Faivre-Finn, C.; Mok, T.S.; Reck, M.; Van Schil, P.E.; Hellmann, M.D.; et al. Metastatic Non-Small Cell Lung Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Ann. Oncol.* **2018**, *29*, iv192–iv237. [CrossRef] [PubMed]
2. 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 A Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
3. Fanti, S.; Farsad, M.; Battista, G.; Monetti, F.; Montini, G.C.; Chiti, A.; Savelli, G.; Petrella, F.; Bini, A.; Nanni, C.; et al. Somatostatin Receptor Scintigraphy for Bronchial Carcinoid Follow-Up. *Clin. Nucl. Med.* **2003**, *28*, 548–552. [CrossRef]
4. Guarize, J.; Casiraghi, M.; Donghi, S.; Diotti, C.; Vanoni, N.; Romano, R.; Casadio, C.; Brambilla, D.; Maisonneuve, P.; Petrella, F.; et al. Endobronchial Ultrasound Transbronchial Needle Aspiration in Thoracic Diseases: Much More than Mediastinal Staging. *Can. Respir. J.* **2018**, *2018*, 4269798. [CrossRef] [PubMed]
5. Forghani, R.; Savadjiev, P.; Chatterjee, A.; Muthukrishnan, N.; Reinhold, C.; Forghani, B. Radiomics and Artificial Intelligence for Biomarker and Prediction Model Development in Oncology. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 995–1008. [CrossRef]
6. Ibrahim, A.; Primakov, S.; Beuque, M.; Woodruff, H.C.; Halilaj, I.; Wu, G.; Refaee, T.; Granzier, R.; Widaatalla, Y.; Hustinx, R.; et al. Radiomics for Precision Medicine: Current Challenges, Future Prospects, and the Proposal of a New Framework. *Methods* **2021**, *188*, 20–29. [CrossRef]
7. Huang, Y.; Liu, Z.; He, L.; Chen, X.; Pan, D.; Ma, Z.; Liang, C.; Tian, J.; Liang, C. Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non—Small Cell Lung Cancer. *Radiology* **2016**, *281*, 947–957. [CrossRef]
8. 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]
9. De Jong, E.E.; Van Elmpt, W.; Rizzo, S.; Colarieti, A.; Spitaleri, G.; Leijenaar, R.T.H.; Jochems, A.; Hendriks, L.E.L.; Troost, E.G.C.; Reymen, B.; et al. Applicability of a Prognostic CT-based Radiomic Signature Model Trained on Stage I-III Non-Small Cell Lung Cancer in Stage IV Non-Small Cell Lung Cancer. *Lung Cancer* **2018**, *124*, 6–11. [CrossRef]
10. Botta, F.; Raimondi, S.; Rinaldi, L.; Bellerba, F.; Corso, F.; Bagnardi, V.; Origgi, D.; Minelli, R.; Pitoni, G.; Petrella, F.; et al. Association of a CT-Based Clinical and Radiomics Score of Non-Small Cell Lung Cancer (NSCLC) with Lymph Node Status and Overall Survival. *Cancers* **2020**, *12*, 1432. [CrossRef]
11. Ninatti, G.; Kirienko, M.; Neri, E.; Sollini, M.; Chiti, A. Imaging-Based Prediction of Molecular Therapy Targets in NSCLC by Radiogenomics and AI Approaches: A Systematic Review. *Diagnostics* **2020**, *10*, 359. [CrossRef]
12. Cucchiara, F.; Petrini, I.; Romei, C.; Crucitta, S.; Lucchesi, M.; Valleggi, S.; Scavone, C.; Capuano, A.; De Liperi, A.; Chella, A.; et al. Combining Liquid Biopsy and Radiomics for Personalized Treatment of Lung Cancer Patients. State of the Art and New Perspectives. *Pharmacol. Res.* **2021**, *169*, 105643. [CrossRef]
13. Hosny, A.; Parmar, C.; Coroller, T.P.; Grossmann, P.; Zeleznik, R.; Kumar, A.; Bussink, J.; Gillies, R.J.; Mak, R.H.; Aerts, H.J.W.L. Deep Learning for Lung Cancer Prognostication: A Retrospective Multi-Cohort Radiomics Study. *PLoS Med.* **2018**, *15*, e1002711. [CrossRef]
14. Coudray, N.; Ocampo, P.S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A.L.; Razavian, N.; Tsirigos, A. Classification and Mutation Prediction from Non–Small Cell Lung Cancer Histopathology Images Using Deep Learning. *Nat. Med.* **2018**, *24*, 1559–1567. [CrossRef]
15. Xu, Y.; Hosny, A.; Zeleznik, R.; Parmar, C.; Coroller, T.; Franco, I.; Mak, R.H.; Aerts, H.J.W.L. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clin. Cancer Res.* **2019**, *25*, 3266–3275. [CrossRef]
16. Lakshmanaprabu, S.K.; Mohanty, S.N.; Shankar, K.; Arunkumar, N.; Ramirez, G. Optimal Deep Learning Model for Classification of Lung Cancer on CT Images. *Future Gener. Comput. Syst.* **2019**, *92*, 374–382. [CrossRef]
17. Avanzo, M.; Stancanello, J.; Pirrone, G.; Sartor, G. Radiomics and Deep Learning in Lung Cancer. *Strahlenther. Onkol.* **2020**, *196*, 879–887. [CrossRef]
18. Binczyk, F.; Prazuch, W.; Bozek, P.; Polanska, J. Radiomics and Artificial Intelligence in Lung Cancer Screening. *Transl. Lung Cancer Res.* **2021**, *10*, 1186–1199. [CrossRef]
19. Jiao, Z.; Li, H.; Xiao, Y.; Dorsey, J.; Simone, C.B.; Feigenberg, S.; Kao, G.; Fan, Y. Integration of Deep Learning Radiomics and Counts of Circulating Tumor Cells Improves Prediction of Outcomes of Early Stage NSCLC Patients Treated With Stereotactic Body Radiation Therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **2022**, *112*, 1045–1054. [CrossRef]

20. Yip, S.S.F.; Liu, Y.; Parmar, C.; Li, Q.; Liu, S.; Qu, F.; Ye, Z.; Gillies, R.J.; Aerts, H.J.W.L. Associations between Radiologist-Defined Semantic and Automatically Computed Radiomic Features in Non-Small Cell Lung Cancer. *Sci. Rep.* **2017**, *7*, 3519. [CrossRef]

21. El-Baz, A.; Beache, G.M.; Gimel'farb, G.; Suzuki, K.; Okada, K.; Elnakib, A.; Soliman, A.; Abdollahi, B. Computer-Aided Diagnosis Systems for Lung Cancer: Challenges and Methodologies. *Int. J. Biomed. Imaging* **2013**, *2013*, e942353. [CrossRef] [PubMed]

22. Zwanenburg, A. Radiomics in Nuclear Medicine: Robustness, Reproducibility, Standardization, and How to Avoid Data Analysis Traps and Replication Crisis. *Eur. J. Nucl. Med. Mol. Imaging* **2019**, *46*, 2638–2655. [CrossRef] [PubMed]

23. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **2016**, *278*, 563–577. [CrossRef] [PubMed]

24. Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L.H.; Aerts, H.J.W.L. Artificial Intelligence in Radiology. *Nat. Rev. Cancer* **2018**, *18*, 500–510. [CrossRef] [PubMed]

25. 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]

26. Castiglioni, I.; Rundo, L.; Codari, M.; Di Leo, G.; Salvatore, C.; Interlenghi, M.; Gallivanone, F.; Cozzi, A.; D'Amico, N.C.; Sardanelli, F. AI Applications to Medical Images: From Machine Learning to Deep Learning. *Phys. Med.* **2021**, *83*, 9–24. [CrossRef]

27. Papadimitroulas, P.; Brocki, L.; Christopher Chung, N.; Marchadour, W.; Vermet, F.; Gaubert, L.; Eleftheriadis, V.; Plachouris, D.; Visvikis, D.; Kagadis, G.C.; et al. Artificial Intelligence: Deep Learning in Oncological Radiomics and Challenges of Interpretability and Data Harmonization. *Phys. Med.* **2021**, *83*, 108–121. [CrossRef]

28. Pham, D.L.; Xu, C.; Prince, J.L. Current Methods in Medical Image Segmentation. *Annu. Rev. Biomed. Eng.* **2000**, *2*, 315–337. [CrossRef]

29. Dehmeshki, J.; Amin, H.; Valdivieso, M.; Ye, X. Segmentation of Pulmonary Nodules in Thoracic CT Scans: A Region Growing Approach. *IEEE Trans. Med. Imaging* **2008**, *27*, 467–480. [CrossRef]

30. Owens, C.A.; Peterson, C.B.; Tang, C.; Koay, E.J.; Yu, W.; Mackin, D.S.; Li, J.; Salehpour, M.R.; Fuentes, D.T.; Court, L.E.; et al. Lung Tumor Segmentation Methods: Impact on the Uncertainty of Radiomics Features for Non-Small Cell Lung Cancer. *PLoS ONE* **2018**, *13*, e0205003. [CrossRef]

31. Parmar, C.; Rios Velazquez, E.; Leijenaar, R.; Jermoumi, M.; Carvalho, S.; Mak, R.H.; Mitra, S.; Shankar, B.U.; Kikinis, R.; Haibe-Kains, B.; et al. Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation. *PLoS ONE* **2014**, *9*, e102107. [CrossRef]

32. Pavic, M.; Bogowicz, M.; Würms, X.; Glatz, S.; Finazzi, T.; Riesterer, O.; Roesch, J.; Rudofsky, L.; Friess, M.; Veit-Haibach, P.; et al. Influence of Inter-Observer Delineation Variability on Radiomics Stability in Different Tumor Sites. *Acta Oncol.* **2018**, *57*, 1070–1074. [CrossRef]

33. Joskowicz, L.; Cohen, D.; Caplan, N.; Sosna, J. Inter-Observer Variability of Manual Contour Delineation of Structures in CT. *Eur. Radiol.* **2019**, *29*, 1391–1399. [CrossRef]

34. Bianconi, F.; Fravolini, M.L.; Palumbo, I.; Pascoletti, G.; Nuvoli, S.; Rondini, M.; Spanu, A.; Palumbo, B. Impact of Lesion Delineation and Intensity Quantisation on the Stability of Texture Features from Lung Nodules on CT: A Reproducible Study. *Diagnostics* **2021**, *11*, 1224. [CrossRef]

35. Siddique, N.; Paheding, S.; Elkin, C.P.; Devabhaktuni, V. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access* **2021**, *9*, 82031–82057. [CrossRef]

36. Yu, X.; Jin, F.; Luo, H.; Lei, Q.; Wu, Y. Gross Tumor Volume Segmentation for Stage III NSCLC Radiotherapy Using 3D ResSE-Unet. *Technol. Cancer Res. Treat.* **2022**, *21*, 153303382210908. [CrossRef]

37. Kido, S.; Hirano, Y.; Mabu, S. Deep Learning for Pulmonary Image Analysis: Classification, Detection, and Segmentation. In *Deep Learning in Medical Image Analysis: Challenges and Applications*; Advances in Experimental Medicine and Biology; Lee, G., Fujita, H., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 47–58, ISBN 978-3-030-33130-6.

38. Liu, X.; Li, K.-W.; Yang, R.; Geng, L.-S. Review of Deep Learning Based Automatic Segmentation for Lung Cancer Radiotherapy. *Front. Oncol.* **2021**, *11*, 717039. [CrossRef]

39. Bianconi, F.; Fravolini, M.L.; Pizzoli, S.; Palumbo, I.; Minestrini, M.; Rondini, M.; Nuvoli, S.; Spanu, A.; Palumbo, B. Comparative Evaluation of Conventional and Deep Learning Methods for Semi-Automated Segmentation of Pulmonary Nodules on CT. *Quant. Imaging Med. Surg.* **2021**, *11*, 3286–3305. [CrossRef]

40. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. NnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]

41. Aerts, H.J.W.L.; Wee, L.; Rios Velazquez, E.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; et al. Data From NSCLC-Radiomics [Data Set]. *Cancer Imaging Arch.* **2019**. [CrossRef]

42. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J. Digit. Imaging* **2013**, *26*, 1045–1057. [CrossRef] [PubMed]

43. 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]

44. Fedorov, A.; Beichel, R.; Kalpathy-Cramer, J.; Fillion-Robin, J.-C.; Pujol, S.; Bauer, C.; Jennings, D.; Fennessy, F.; Sonka, M.; Buatti, J.; et al. 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **2013**, *28*, 1323–1341. [CrossRef] [PubMed]

45. Yaniv, Z.; Lowekamp, B.C.; Johnson, H.J.; Beare, R. SimpleITK Image-Analysis Notebooks: A Collaborative Environment for Education and Reproducible Research. *J. Digit. Imaging* **2018**, *31*, 290–303. [CrossRef] [PubMed]

46. Rizwan, I.; Haque, I.; Neubert, J. Deep Learning Approaches to Biomedical Image Segmentation. *Inform. Med. Unlocked* **2020**, *18*, 100297. [CrossRef]

47. Haarburger, C.; Müller-Franzes, G.; Weninger, L.; Kuhl, C.; Truhn, D.; Merhof, D. Radiomics Feature Reproducibility under Inter-Rater Variability in Segmentations of CT Images. *Sci. Rep.* **2020**, *10*, 12688. [CrossRef]

48. 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]

49. Rinaldi, L.; Pezzotta, F.; Santaniello, T.; De Marco, P.; Bianchini, L.; Origgi, D.; Cremonesi, M.; Milani, P.; Mariani, M.; Botta, F. HeLLePhant: A Phantom Mimicking Non-Small Cell Lung Cancer for Texture Analysis in CT Images. *Phys. Med.* **2022**, *97*, 13–24. [CrossRef]

50. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef]

51. MONAI Consortium MONAI: Medical Open Network for AI. 2020. Available online: https://monai.io/ (accessed on 24 September 2021).

52. Yang, J.; Huang, X.; He, Y.; Xu, J.; Yang, C.; Xu, G.; Ni, B. Reinventing 2D Convolutions for 3D Images. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3009–3018. [CrossRef]

53. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in PyTorch. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.

54. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, 8024–8035.

55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27 June 2016; IEEE: Manhattan, NY, USA, 2016; pp. 770–778.

56. Goldstraw, P.; Chansky, K.; Crowley, J.; Rami-Porta, R.; Asamura, H.; Eberhardt, W.E.E.; Nicholson, A.G.; Groome, P.; Mitchell, A.; Bolejack, V.; et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *J. Thorac. Oncol.* **2016**, *11*, 39–51. [CrossRef]

57. Biewald, L. Experiment Tracking with Weights and Biases. Available online: https://wandb.ai/site/ (accessed on 24 August 2022).

58. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

59. Gan, W.; Wang, H.; Gu, H.; Duan, Y.; Shao, Y.; Chen, H.; Feng, A.; Huang, Y.; Fu, X.; Ying, Y.; et al. Automatic Segmentation of Lung Tumors on CT Images Based on a 2D & 3D Hybrid Convolutional Neural Network. *BJR* **2021**, *94*, 20210038. [CrossRef]

60. Armato, S.G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans: The LIDC/IDRI Thoracic CT Database of Lung Nodules. *Med. Phys.* **2011**, *38*, 915–931. [CrossRef]

61. Baumgartner, C.F.; Tezcan, K.C.; Chaitanya, K.; Hötker, A.M.; Muehlematter, U.J.; Schawkat, K.; Becker, A.S.; Donati, O.; Konukoglu, E. PHiSeg: Capturing Uncertainty in Medical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, Shenzhen, China, 13–17 October 2019; Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 119–127.