Contents lists available at ScienceDirect

Physica Medica

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

Radiomics classifier to quantify automatic segmentation quality of cardiac sub-structures for radiotherapy treatment planning

Nicola Maffei [a], Luigi Manco [a], Giovanni Aluisio [b], Elisa D'Angelo [b], Patrizia Ferrazza [c], Valentina Vanoni [c], Bruno Meduri [b], Frank Lohr [b], Gabriele Guidi [a,*]

[a] A.O. U. di Modena, Medical Physics Unit, Modena, Italy
[b] A.O. U. di Modena, Radiotherapy Unit, Dept. of Oncology, Modena, Italy
[c] Ospedale S. Chiara, Radiotherapy Unit, Trento, Italy

ARTICLE INFO

Keywords:
Artificial intelligence
Automatic segmentation
Radiomics
Cardiac structures

ABSTRACT

Purpose: A radiomics features classifier was implemented to evaluate segmentation quality of heart structures. A robust feature set sensitive to incorrect contouring would provide an ideal quantitative index to drive auto-contouring optimization.
Methods: Twenty-five cardiac sub-structures were contoured as regions of interest in 36 CTs. Radiomic features were extracted from manually-contoured (MC) and Hierarchical-Clustering automatic-contouring (AC) structures. A robust feature-set was identified from correctly contoured CT datasets. Features variation was analyzed over a MC/AC dataset. A supervised-learning approach was used to train an Artificial-Intelligence (AI) classifier; incorrect contouring cases were generated from the gold-standard MC datasets with translations, expansions and contractions. ROC curves and confusion matrices were used to evaluate the AI-classifier performance.
Results: Twenty radiomics features, were found to be robust across structures, showing a good/excellent intra-class correlation coefficient (ICC) index comparing MC/AC. A significant correlation was obtained with quantitative indexes (Dice-Index, Hausdorff-distance). The trained AI-classifier detected correct contours (CC) and not correct contours (NCC) with an accuracy of 82.6% and AUC of 0.91. True positive rate (TPR) was 85.1% and 81.3% for CC and NCC. Detection of NCC at this point of the development still depended strongly on degree of contouring imperfection.
Conclusions: A set of radiomics features, robust on "gold-standard" contour and sensitive to incorrect contouring was identified and implemented in an AI-workflow to quantify segmentation accuracy. This workflow permits an automatic assessment of segmentation quality and may accelerate expansion of an existing autocontouring atlas database as well as improve dosimetric analyses of large treatment plan databases.

1. Introduction

Image segmentation is a fundamental task in the RT workflow because the contoured treatment targets and organs at risk (OARs) are used to both optimize and evaluate treatment plans. Segmentation quality can have an impact on patient treatment and related analyses (i.e. radiomics analysis). In a standard clinical workflow, segmentation is still mainly carried out manually by an expert human contourer (Manual Contouring, MC) [1].

This approach is a repetitive and time-consuming process [1]. Moreover, reports also showed that there is significant interobserver-variation, which may be the consequence of a general lack of a clear ground truth for a given situation but may also be a consequence of variations in individual training level [1–13].

Auto-contouring (AC) has the potential to accelerate the treatment planning workflow and to facilitate on-line Adaptive Radiation Therapy (ART) strategies. According to Cardenas et al., auto-segmentation algorithms can be grouped as Atlas-Based Segmentation (ABS), Model-Based Segmentation (MBS) and Machine Learning-based segmentation. Although all above mentioned methods have been evaluated with prom31ising results, accurate commissioning, periodic QA and patient-specific manual verification must be performed to evaluate image segmentation uncertainty appropriately and avoid possible errors in the segmentation process [1]. According to Jungo et al. medical image

[*] Corresponding author.
E-mail address: guidi.gabriele@aou.mo.it (G. Guidi).

https://doi.org/10.1016/j.ejmp.2021.05.009
Received 15 December 2020; Received in revised form 30 April 2021; Accepted 3 May 2021
Available online 13 May 2021

segmentation uncertainty can be evaluated at three levels: the voxel-wise uncertainty, the uncertainty at the level of a segmented instance, the subject-level uncertainty [14]. Automatic quality assurance of auto-segmentations has been investigated in the literature, evaluating ROI specific characteristics such as centroid, volume, shape and use statistical approaches to determine variations in contoured ROIs [15–17]. Court et al. used the results of a primary segmentation algorithm and compare these to a secondary, independent, verification algorithm [18]. Unsupervised segmentation quality assessment is of high interest in medical imaging fields. Robinson et al. evaluated the segmentation quality metrics by the reverse classification accuracy (RCA) using image registration and manually contoured atlas images [19]. The input image is registered to the atlas images generating later a set of surrogate reference segmentation by reversely transforming the manual segmentations using a quality metric between the candidate segmentation and the set of surrogate reference segmentations. Zhou et al. implemented two CNNs to use features related to segmentation to improve the robustness of the quality regression network. A reconstruction network and a quality regression network were developed to reconstruct the image masked by the provided segmentation and to predict the segmentation quality based on the reconstruction difference image and the provided segmentation, respectively [20]. According to them, deep learning methods, however, may fail due to many factors, such as domain shift, adversarial noise and low image quality and robustness problems if the input images have a different distribution from the training datasets for the regress network [20]. Jin et al. investigated the accuracy of automatic segmentation of a multiple U-net based algorithms and related radiomics features in US images of ovarian cancer patients [21].

It is important to remember that the quality of the data (both the images and segmentations) could be closely correlated with segmentation algorithm results. Moreover, the use of different contouring guidelines between institutions could have significant impact on the performance of an algorithm when tested on a new dataset [1]. All auto-segmentations methods should therefore be used as a decision support tool and should be carefully reviewed and approved by the local clinical staff before use in a treatment plan [1].

Recent studies have shown the potential of radiomics to significantly improve the ability to stratify patients according to treatment response or treatment side effects beyond conventional prognostic factors, leading to more accurate personalized cancer care [22,23]. Radiomics focus on the extraction of quantitative imaging features to be used for the development of decision support systems, e.g. to accurately estimate patient risk and improve personalized treatment selection and monitoring [24]. The hypothesis behind radiomics is that mineable data can be extracted from medical images to provide additional information (e.g. gene protein, tumor phenotype) useful for patient care [25].

As outlined by Owens et al., the generic workflow of radiomics studies includes 4 steps [26]: Image acquisition; ROI segmentation (drawn manually or generated with an automatic tool), feature extraction and feature analysis. Features can be classified into the following 3 categories [9,23]: First order features describe the histogram of voxel intensity values contained within the volume of interest (VOI), shape features describe the 3D shape and size of the VOI and texture features reveal heterogeneity differences.

In a radiomics study, a high number of features (typically more than a hundred) are extracted characterizing a given ROI in different ways [26]. In a second step, the features are tested as prognosticators. Moreover, to be clinically applicable, features have to be selected carefully regarding feature robustness and sensitivity towards the delineation process [23].

This, in turn, also opens another potential use of radiomic features and distributions as a possible application might be the radiomics-based generation of regions/volumes of interest (ROI/VOI) with certain characteristics to improve anatomic auto-contouring. To automatize the assessment of contouring quality, radiomic features may be used to assess and quantify contouring accuracy. The major sources of uncertainty in the contouring process are, indeed, image quality, different experience level of physicians and inter/intra observer variability. Therefore, integrating radiomic features into the AC process might be an effective approach to quantify and reduce the uncertainty of ROIs. Suitable features must be robust between a gold standard and a well contoured test dataset and sensitive towards contouring errors.

To date, a small number of studies have been performed to assess the usability of radiomic features in the quantification of contouring precision [23,26]. Substructures of the heart, such as proposed by Duane et al. [27] to standardize reporting, are a particularly relevant AC target [28] and AC of small structures has not been fully accurate, particularly on non-contrast imaging for treatment planning purposes.

In a previous study the Dice similarity coefficient (DSC), the Average Hausdorff Distance (AHD) and the volume comparisons were used to evaluate the performance of heart substructure AC [29]. While these parameters, when used simultaneously, sufficiently describe the accuracy of an autocontouring result in comparison to a gold standard, convolving their information into a potential driver of AC is not straightforward.

This manuscript focuses first on identifying a set of radiomic features that correlates robustly with the conventional similarity metrics when comparing manually and automatically generated heart substructures; in a second step, this set of radiomic features was used to drive an Artificial Intelligence Classifier (AIC) to quantify segmentation quality, a necessary step to automatize atlas database expansion and potentially further improve AC results being used as an additional driver of contour optimization.

2. Material and methods

2.1. Data set cohort

Thirty-six anonymized female Computed Tomography (CT) scans (without intravenous (i.v.) contrast), originally treated with 3D-CRT for breast tumors (right or left), were retrospectively analyzed. The planning CT was acquired in supine position, with the arms above the head, using a standard positioning system. All images were acquired by an Aquilion® Large Bore CT (Canon Medical®), with a slice thickness of 3 mm, an image matrix of 512x512 pixels and pixel size of 0.098 cm × 0.098 cm.

2.2. Cardiac sub-structures and manual contouring to define ground truth

All cardiac sub-structures were contoured by an expert radiation oncologist based on the atlas recently proposed by Duane et al. [27] and this contouring was considered the ground truth. The 25 contoured sub-structures with related acronyms are summarized in Table 1.

2.3. Atlas based autocontouring

For the automatic generation of the contour dataset underlying the current analysis the RayStation® (RaySearch Laboratories®) ABS tool was used with a Hierarchical Clustering Workflow previously established [29] as the auto-segmentation solution for cardiac sub-structures. In its standard implementation, based on the ANAtomically CONstrained Deformation Algorithm (ANACONDA), the hybrid DIR algorithm combines image information with anatomical information as provided by contoured image sets [30]. The segmentation algorithm combines rigid image registration (RIR) and deformable image registration (DIR) during the atlas-based contouring initialization [30,31]. Hierarchical Clustering, supported by IronPython® scripts, was added to the standard workflow as described previously. The “heart” structure is used as an external ROI to guide the segmentation of other cardiac substructures in a top-down approach. The developed ABS method was previously evaluated using both a qualitative and a quantitative

Table 1
List of cardiac sub-structures with related acronyms.

Cardiac sub-structure	ROI Acronym
Heart	Heart
Left Atrium	LA
Right Atrium	RA
Left Ventricle	LV
Right Ventricle	RV
Anterior Left Ventricle	AntLV
Apical Left Ventricle	ApLV
Lateral Left Ventricle	LatLV
Inferior Left Ventricle	InfLV
Septal Left Ventricle	SepLV
Ascending Aorta	Aorta
Pulmonary Artery	PA
Left Main Coronary Artery	LMCA
Proximal Right Coronary Artery	ProxRCA
Mid Right Coronary Artery	MidRCA
Distal Right Coronary Artery	DistRCA
Posterior Descending Right Coronary Artery	DescRCA
Proximal Left Anterior Descending Coronary Artery	ProxLADCA
Mid Left Anterior Descending Coronary Artery	MidLADCA
Distal Left Anterior Descending Coronary Artery	DistLADCA
Proximal Circumflex Coronary Artery	ProxCCA
Distal Circumflex Coronary Artery	DistCCA
Coronary Sinus	CS
Inferior Vena Cava	IVC
Superior Vena Cava	SVC

approach to evaluate automatic contours compared with the results of MC. The previous study showed that subjective physician scoring was good or acceptable for 70% of automatically contoured ROIs (AC-Good/Acceptable as opposed to AC-Nonacceptable) [29]. Inter-observer evaluation showed that contours obtained by the Hierarchical Clustering method are statistically comparable with those obtained by a second, independent, expert contourer [29]. This high-quality subgroup of the AC dataset (25 AC-Good/Acceptable datasets) was therefore, together with the MC dataset, used as the ground-truth dataset to identify a robust feature set common to all the 25 ROIs.

2.4. Radiomic feature extraction

Fig. 1 showed the robust radiomic features extraction workflow. The 25 cardiac sub-structured were manually and automatically segmented in each of the 36 CTs. All the 900 ROIs obtained from the manual segmentation were considered Ground truth. Considering the automatic segmentation, only 630 ROIs (70% of the total) were considered good/acceptable by the evaluating radiation oncologist and so considered ground truth. The remaining 30% of not correct automatically generated contours were used in the training and test dataset, respectively, of the classifier being real “negative” cases, essential to be identified in clinical routine.

Using the SlicerRadiomics® tool, an extension for 3DSlicer® (v4.10 [32]) that encapsulates the PyRadiomics® library [33–37], a total of 96 3D-Radiomic features were extracted from each CT dataset both for manual and automatic contours of each of the 25 cardiac sub-structures. A detailed explanation of features used in this study and their respective nomenclature can be found, for example, in the documents published by the Image Biomarker Standardization Initiative (IBSI) [23,33–37].

2.5. Statistical analysis

After feature extraction, the reproducibility of each feature between the MC and AC-Good/Acceptable ROIs was evaluated by the intra-class correlation coefficient (ICC). The ICC is a statistical measure, ranging between 0 (null reproducibility) and 1 (perfect reproducibility) [26,38]. McGraw and Wong [39] defined 10 forms of ICC. For the analysis reported here the interrater reliability was calculated reflecting the variation between 2 or more raters who measure the same group of subjects [38]. The two-way mixed effects, absolute agreement, single rater/measurement model was used.

Among the features that exhibited good and excellent ICC values, the Kruskal-Wallis test was performed in order to determine whether a specific feature class was significantly more reproducible than another.

To compare the feature range between values extracted from manual and automatic contours, a Z-score normalization was applied calculating the normalized feature according to Owens et al. [26].

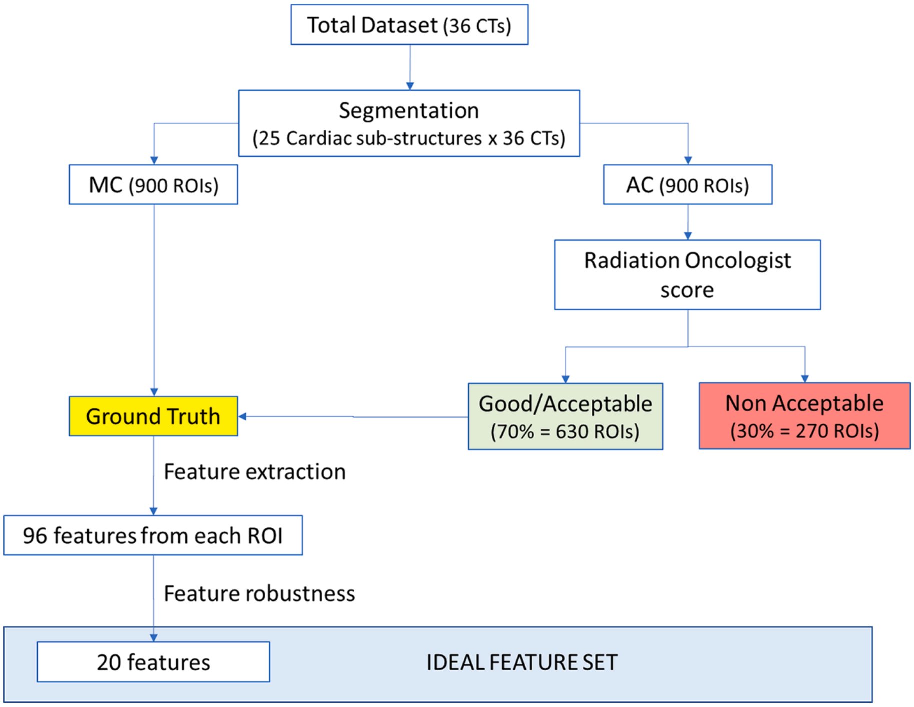

Total Dataset (36 CTs)
Segmentation
(25 Cardiac sub-structures x 36 CTs)
MC (900 ROIs)
AC (900 ROIs)
Radiation Oncologist score
Good/Acceptable (70% = 630 ROIs)
Non Acceptable (30% = 270 ROIs)
Ground Truth
Feature extraction
96 features from each ROI
Feature robustness
20 features
IDEAL FEATURE SET

Fig. 1. Robust radiomic features extraction workflow.

The Mann-Whitney test was performed to compare feature values between manual and automatic segmentation.

Finally, a correlation between ICC (calculated over the robust feature set) and commonly used quantitative indexes (i.e. DSC and AHD) comparing the MC-dataset with the AC-dataset (this time comprising all quality levels of contouring, thus AC-Good/Acceptable and AC-Nonacceptable) was evaluated by using the r-test in order to validate our hypothesis that the addition of a new radiomic index might simplify the assessment of contouring quality/be used as a driver for autocontouring.

2.6. AI-Classifier implementation

The total dataset of 36 patients was divided into the training set (26 patients corresponding to the 72% of the total) and test set (10 patients corresponding to the 28% of the total). The training set was composed of 650 correct manually contoured (CC-MC) ROIs, 455 correct automatically contoured ROIs (CC-AC Good/Acceptable segmented by the Hierarchical Clustering approach), and 2795 non-correct-contours (NCC) composed by 2600 not correct contoured ROIs created by synthetic degradation of MC (NCC-S) plus 195 clinical unacceptable automatically contoured ROIs (NCC-AC). The four not correct contoured ROIs created by synthetic degradation of MC were obtained by isotropic expansion and contraction of 1 mm (for ROI with volume $\leq$ 1 cc) and 3 mm (for ROI with volume $>$ 1 cc) of the ROI with a center of mass (COM) shifted of 1 mm and 3 mm from the MC.

A total of 78,000 features considering all the 25 ROIs for all the 26 training cases were included in the AI Classifier. The labels used as input to train the model were summarized in the Supplementary Material (Table 3) and shown in Fig. 2 considering the example of the Aorta.

The test set is composed of 250 CC-MC ROIs, 175 CC-AC ROIs, and 1075 NCC composed by 1000 NCC-S ROIs plus 75 NCC-AC ROIs. For the test set, the total number of features was thus 30000.

2.7. AI-Classifier architecture

The AIC was implemented using the Classification Learner app in the Machine Learning and Deep Learning group of Matlab® (The MathWorks, Inc. [40]). A supervised machine learning approach was selected to train the model to identify incorrect contours of cardiac sub-structures in order to support physician decision making and automatize ATLAS database expansion. The different steps of the general workflow are shown in Fig. 3 and can be summarized as follows: the 25 cardiac sub-structures of a new, blank, CT were automatically segmented using a Hierarchical Clustering ABS, the robust radiomic features were extracted from each ROI and the trained AI classifier evaluated the quality of the obtained contours. If the contours were classified as clinically acceptable, the CT with related structures were uploaded into the ABS database in order to expand the atlas dataset improving so the possible anatomical matching. If the contours were classified as clinically unacceptable, the CT with related structures were evaluated by an expert radiation oncologist. At this point, if a brief manual editing can fix the automatical contour errors, the CT with related structures were uploaded into the ABS database; otherwise, they were discarded.

We implemented and trained the decision trees, starting at the top node, using the robust radiomic feature set as predictor; at each decision, the AI Classifier assesses the values of the predictors to decide which branch to follow. When the branches reach a leaf node, the data is classified either as Correct Contour (CC) or Not Correct Contour (NCC).

The principal component analysis (PCA) was used to reduce the dimensionality of the predictor space. Reducing the dimensionality can create classification models that help prevent overfitting. PCA linearly transforms predictors in order to remove redundant dimensions and generates a new set of variables (i.e. principal components). Each principal component is a linear combination of the original variables; all the principal components are orthogonal to each other, so there is no redundant information [40]. The model was trained allowing PCA to keep only the components that explain 95% of the variance; a higher value could generate overfitting, while a lower value could remove useful dimensions.

Between the different implemented and tested model types, the Fine Tree model was selected for our purposes. It modelled feature data of different contoured ROIs using a decision tree with many leaves that makes fine distinctions between classes with a maximum number of 100 splits.

3. Results

3.1. Feature robustness

The results of the analysis have identified a group of reliable radiomics features for cardiac sub-structures contoured similarly using MC and AC segmentation methods. The first requisite of a feature set that can be used to discriminate correctly from incorrectly contoured datasets is that it is robust across perfectly contoured "ground truth" datasets. Only for this subset of features that are robust across perfectly

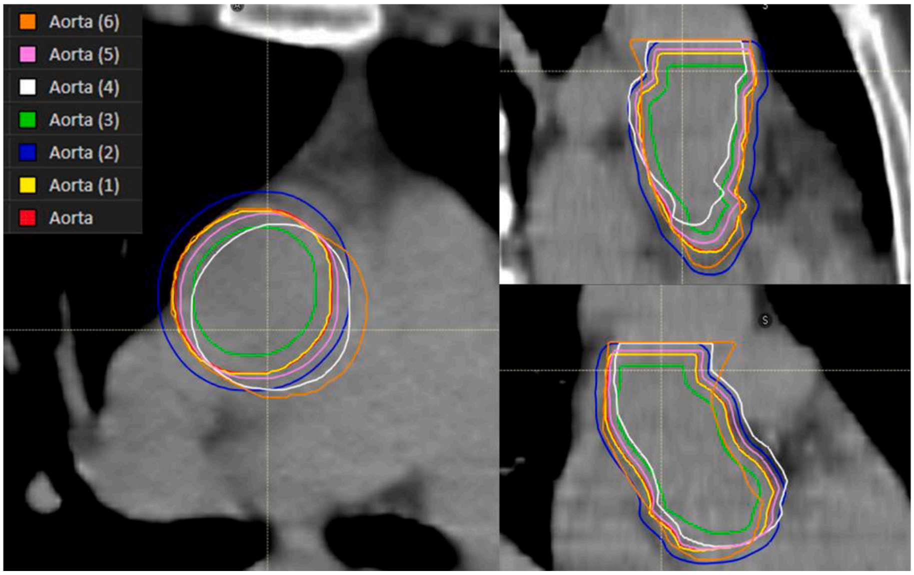


Fig. 2. Example of a training case of the AIC considering the Aorta contours. "Aorta" (red) is the ROI contoured manually (CC-MC), "Aorta (1)" (yellow) is the ROI auto-contoured correctly by the Hierarchical Clustering approach (CC-AC), "Aorta (2)" (blue) is the ROI with isotropic expansion of 3 mm obtained from MC (NCC-S), "Aorta (3)" (green) is the ROI with isotropic contraction of 3 mm (NCC-S), "Aorta (4)" (white) is the ROI with 3 mm a shift of the COM of MC (NCC-S), "Aorta (5)" (pink) is the ROI with 1 mm shift of the COM (NCC-S), "Aorta (6)" (orange) is the ROI not correctly contoured by Hierarchical Clustering autocontouring approach (NCC-AC). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

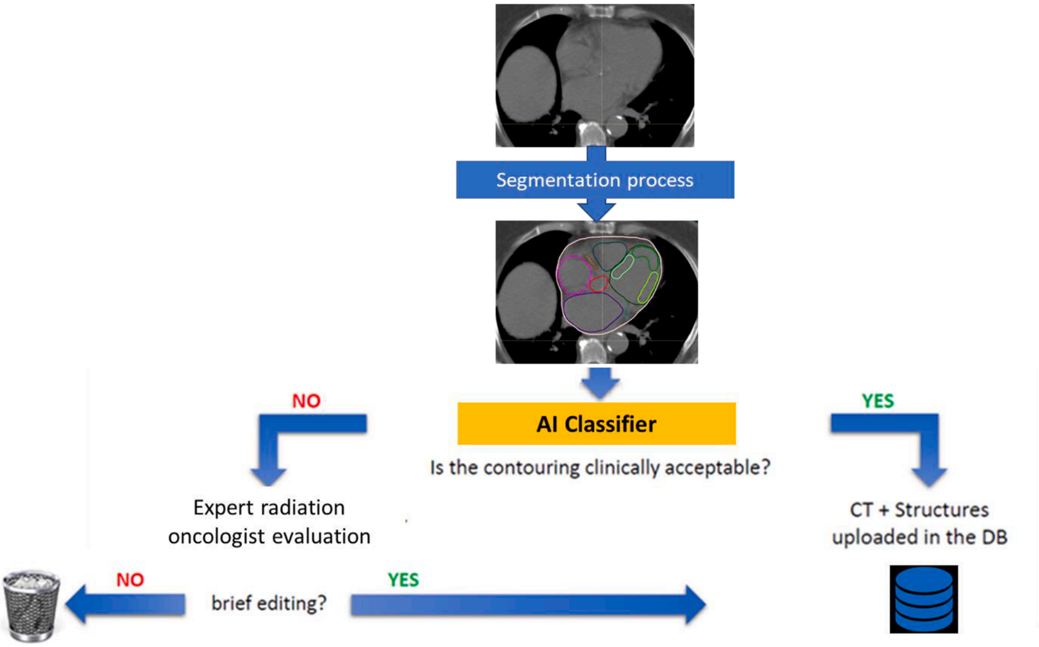

Segmentation process
NO
AI Classifier
YES
Is the contouring clinically acceptable?
Expert radiation oncologist evaluation
CT + Structures uploaded in the DB
NO
brief editing?
YES

Fig. 3. General workflow to create/expand an existing atlas database for autocontouring.

contoured datasets does it make sense to then evaluate the sensitivity to incorrect contouring.

Based on the approach chosen by Owens et al. [26], all the 96 extracted features were merged into 4 groups according to their ICC values: poor reproducibility ($ICC < 0.4$), fair reproducibility ($0.4 \leq ICC < 0.60$), good reproducibility ($0.60 \leq ICC < 0.75$) and excellent reproducibility ($ICC \geq 0.75$). Results are reported in Supplementary Materials (Fig. 7).

Each of the 25 cardiac sub-structures displayed a structure-specific robust feature set. In order to identify a common set for all the 25 contoured ROIs, an ICC threshold has been used. Features with an $ICC \geq 0.60$ for more than half of the cardiac sub-structures were considered robust (supplementary materials Fig. 8). The list of the 20/96 (21%) features identified as robust are shown in Supplementary Materials (Table 4).

Among the features that exhibited good and excellent ICC values, the Kruskal-Wallis test was performed in order to determine whether a specific feature class (i.e. First Order, Glcm, Gldm, Glrlm, Glszm) was significantly more reproducible than another feature category. No features classes were found to be more reproducible than another (sign. 0.466).

3.2. Feature range analysis

To assess the feature range with good or excellent ICC, a Z-score normalization was applied. According to Owens et al. [26], a simple comparison can be performed using the normalization, allowing to plot features for different scales. The minimum and maximum normalized

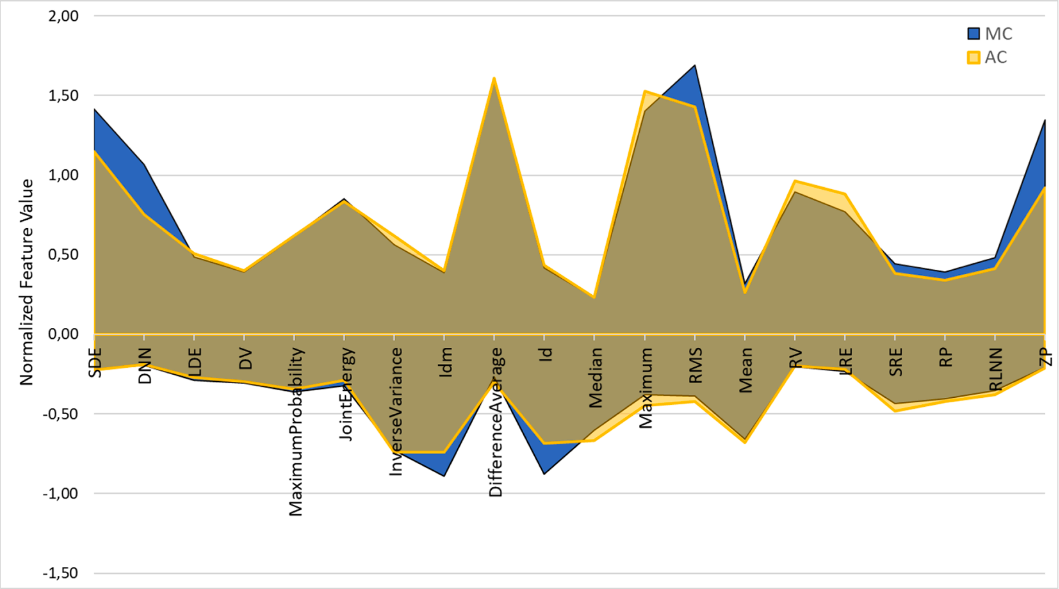


Fig. 4. Normalized feature range. Comparison of normalized feature range between MC (blue) and AC (orange) segmentation using z-score normalization. The minimum and maximum values are plotted for each feature with good or excellent ICC over the entire structure sets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

features values for MC (blue) and AC-Good/Acceptable (orange) (over the entire structure sets) were plotted in Fig. 4.

The Mann-Whitney test was performed to compare feature values between MC and AC-Good/Acceptable (Supplementary Materials Table 5). No statistical difference was observed for each of the 20 robust features on all the 25 ROIs, so AC and MC are comparable and both clinically acceptable.

3.3. Correlation between robust features and quantitative indexes

To evaluate a correlation between the feature robustness and the accuracy of the segmentation, the 20 robust features detected were correlated with quantitative indexes previously used (i.e. DSC and AHD) to compare agreement between MC and AC (this time, both subcohorts, AC-Good/Acceptable and AC-Nonacceptable) by using the r-test (Table 2).

For most features, a significant correlation between the ICC of robust features and the standard quantitative indexes used to measure the automatic segmentation performance was observed (Sign. < 5%).

3.4. AI based classification

After the training process the AI Classifier (Fig. 5) achieved an accuracy of 76.8%. The addition of PCA improved the classifier performance, reaching an accuracy of 82.6%.

In the training set, there are “good” and “bad” contours, the good of which were the gold standard created either manually or automatically, the bad cases (NCC) used to train the model were those obtained from expansion, contraction and shifts of the gold standard plus contours created automatically but considered clinically unacceptable. Receiver operating characteristic (ROC) curve analyses were used to evaluate the true positive rate (TPR) versus false positive rates (FPR) for correct and not correct contours. The area under the curve (AUC) was 0.91. The red point of the ROC curve represents the coordinates of TPR and FPR of the model (i.e. correct cases).

A confusion matrix was obtained for the two output classes: CC and NCC. The TPR for CC was 85.1% (362/425 ROIs). The TPR for NCC was 81.3% (874/1075 ROIs); 852/1000 ROIs NCC-S and 22/75 ROIs NCC-AC were correctly classified.

The false negative rates (FNR) were 14.9% (63/425 ROIs), 18.7% (201/1075 ROIs) for Correct and NCC, respectively. The Positive Predicted Values (PPV) and False Discovery Rates (FDR) were 69.5% and 30.5% for CC, 91.6% and 8.4% for NCC.

Table 2
r test between robust feature ICC and quantitative indexes (i.e. DSC and AHD). Significance level: 5%.

Features	DSC		AHD	
	r_0	Sign.	r_0	Sign.
DNN	0.90	<0.05%	−0.56	0.50%
DV	0.93	<0.05%	−0.82	<0.05%
DifferenceAverage	0.77	<0.05%	−0.43	3.50%
Id	0.84	<0.05%	−0.55	0.70%
Idm	0.83	<0.05%	−0.48	1.90%
InverseVariance	0.92	<0.05%	−0.70	<0.05%
JointEnergy	0.90	<0.05%	−0.60	0.30%
LDE	0.87	<0.05%	−0.60	0.30%
LRE	0.89	<0.05%	−0.57	0.40%
Maximum	0.52	0.70%	−0.57	0.30%
MaximumProbability	0.93	<0.05%	−0.70	<0.05%
Mean	0.41	4.80%	−0.24	26.20%
Median	0.80	<0.05%	−0.62	0.10%
RMS	0.39	8.10%	−0.44	4.30%
RLNN	0.78	<0.05%	−0.45	3.30%
RP	0.88	<0.05%	−0.53	0.90%
RV	0.94	<0.05%	−0.73	<0.05%
SRE	0.80	<0.05%	−0.54	0.80%
SDE	0.79	<0.05%	−0.44	3.30%
ZP	0.83	<0.05%	−0.62	0.10%

Specific TPR analysis considering all the 25 cardiac sub-structures was shown in Fig. 6. For each ROI, the TPR for “Correct” and “not Correct” contours were assessed.

Considering results summarized in Fig. 7 of the Supplementary materials, ICC results of some ROIs (e.g. DistLADCA) are poorer compared with other rightmost ROIs (e.g. SepLV). As expected, the classification of bad/good contours is generally better for the rightmost structures in Fig. 7. As an example, the DistLADCA has a TPR of 50% and 85% for correct and not correct contours; the SepLV has a TPR of 100% and 92% for correct and not correct contours.

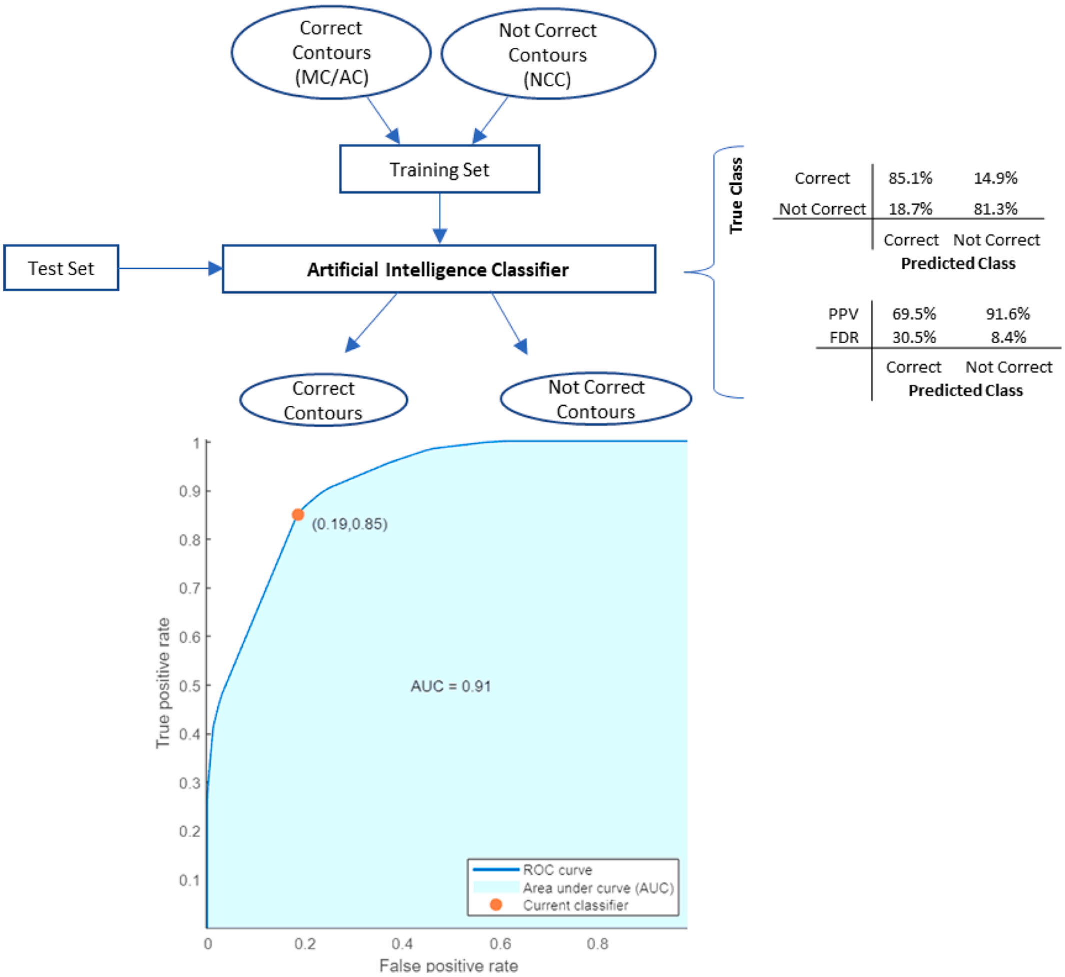
4. Discussion

ROI contouring is a fundamental task of the RT workflow; tumor control and OAR toxicity are potentially correlated with the accuracy of delineation. Inaccuracies in this process can have an impact on patient care as these ROIs are used to optimize and evaluate radiotherapy treatment plans. Therefore, the quality of RT treatment and also subsequent analyses (i.e. radiomics analysis) depends on segmentation quality [27-29]. High quality auto-contouring is the single most important element to further accelerate the treatment planning workflow and to facilitate on-line adaptive radiotherapy (ART) strategies. Moreover, a recent review study focused on machine learning and deep learning in imaging highlighted a great interest on the possibility to extract useful features directly from raw images despite segmentation related challenges [41]. Identifying contouring methods that improve feature reliability, helps to reduce feature uncertainties caused by inconsistent contouring.

In a previous study a Hierarchical Clustering Atlas based algorithm was developed and tested. 70% of automatically contoured ROIs was judged good or acceptable by physician scoring [29] which is already an acceptable yield for retrospective DVH analyses but leaves considerable room for improvement and still requires extensive editing in a lot of patients when used clinically.

A recent study was conducted by Owens et al. to evaluate the uncertainty of radiomics features from CT scans of non-small cell lung cancer for both manual and semi-automatic segmentation due to intra-observer, inter-observer, and inter-software reliability. The authors observed that to minimize the uncertainty in radiomics studies one contouring approach should be used; moreover, auto-contouring is a fundamental step because it reduces human uncertainty. Finally, radiomics features extracted from semi-automatic contours showed improved reproducibility and reliability than those obtained from manual segmentation [26].

In an attempt to automatize the assessment of contouring quality, radiomic features may, reversing the process of radiomic analysis, be used to assess and quantify contouring accuracy. Suitable features must be robust between a gold standard and a well contoured test dataset and sensitive towards contouring errors. A radiomic based index was investigated and correlated with commonly used quantitative indexes (i.e. DSC and AHD). In order to find a robust feature set in a group of gold standard contours (MC and AC) of cardiac sub-structures, ICC and Z-normalization score were applied to a set of 36 CTs. The feature set “ideal” for contour evaluation consists of the 20 features (of a total of 96 studied features) that were identified as robust across the contoured CT-datasets with “ground truth” contour quality. ICC values were similar when extracted from manual or automatic gold standard contours. Focusing on each ROI, it is possible to observe that some of them has a greater number of robust features (Supplementary materials Fig. 7). For example, Heart, LA, LV, RA, SepLV had 61 features (64%), 52 features (54%), 46 features (48%), 49 features (51%), 65 features (68%), respectively, with good or excellent reproducibility. A possible correlation between the number of robust features and the structure dimension could be further investigated. Indeed, among the 25 ROIs analyzed in this analysis, these structures had a mean volume of 183 cc [12 ÷ 620



True Class	Correct	Not Correct
Correct	85.1%	14.9%
Not Correct	18.7%	81.3%

Predicted Class

	Correct	Not Correct
PPV	69.5%	91.6%
FDR	30.5%	8.4%

Predicted Class

Fig. 5. AI Classifier training and test datasets with related results. At the bottom: ROC curves related to the classification of test datasets, AUC = 0.91; the red point of the ROC curve represents the coordinates of TPR and FPR of the model (i.e. correct cases). At the right Confusion matrix for Correct and Not Correct contours between true and predicted classes. TPR: True Positive rates, FNR: False Negative Rates, PPV: Positive Predicted Values, FDR: False Discovery Rates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

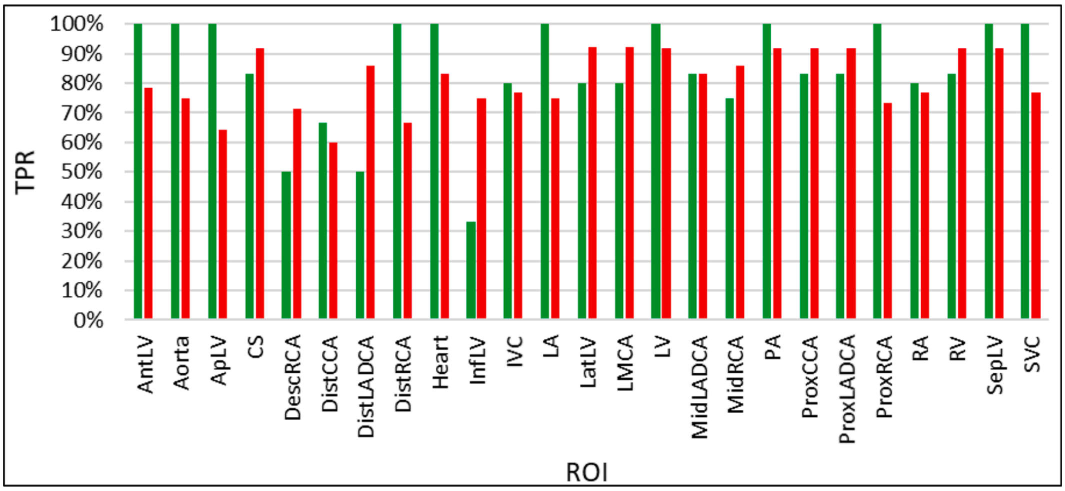


Fig. 6. Specific TPR analysis for all the 25 cardiac sub-structures. Green bar represents the TPR for correct contoured class, red bar represents the TPR for not correct contoured class. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cc]. Furthermore, as suggested by Parmar et al., it becomes important to determine whether the features extracted from automatic segmentations capture the same image properties as with manual delineations [23]. Therefore, after a Z-score normalization of every feature value, the normalized range between manual and automatic segmentation groups were analyzed. It is possible to observe that the features extracted from the AC overlapped in range spread when compared to those of the MC. Moreover, the Mann-Whitney test was performed to compare feature values between manual and automatic segmentation. No feature class was found to be statistically more reproducible than another.

To automatically evaluate segmentation quality of heart structures, all manual contours and the "well autocontoured" ROIs of 26/36 patients were selected as the "well contoured" dataset for training the classifier while a set of "badly countoured" ROIs was obtained from 4 different degradation of all manual contours plus clinical unacceptable automatically contoured ROIs. The remaining 10 patients of the original 36-patient group were chosen as the test set.

The classifier was trained and implemented using a decision tree algorithm. The decision tree was chosen as it follows a non-parametric method; meaning, it is distribution-free and does not depend on probability distribution assumptions. It can work on high-dimensional data such as radiomic features with good accuracy. Unlike other classification algorithm, in decision trees, nonlinear relationships between parameters do not influence the trees performance. Another advantage of the decision tree is that possible missing values in the data do not affect the classification building process; it is useful in situations such as small cardiac sub-structures segmentation with possible blank points in the data analysis results (see Fig. 7 in Supplementary materials). Cases of missing values and outliers have less significance on the decision tree's performance than on the performance of other methods. Moreover, the training time of decision tree is faster compared to other methods.

In order to evaluate the sensitivity of a radiomics driven approach towards the identification of NCC the classifier was trained as restrictive regarding the correct recognition of NCC cases as this permit to expand an existing atlas database for an atlas-based autocontouring system only with high quality contours that need very little editing before being approved as an atlas. With an AUC of 0.91 and an accuracy of 82.6%, the AI based model can detect contours with radiomics features that do not correspond with those of the gold standard dataset. Such a low rate of false negative datasets (incorrectly contoured but falsely considered correct) minimizes the chances of including incorrect ROIs in the atlas database and also reduces the workload for final manual verification and potential residual manual editing as all grossly insufficiently contoured datasets have likely already been excluded by the classifier. This approach enables a semiautomatic expansion of the atlas database and might be used to prescreen autocontouring results that are applied to large databases for patient-individual DVH-analysis, a paradigm that has recently become important to further refine our knowledge of normal tissue tolerance to therapeutic radiation.

Detection of NCC was, at this stage of the development, still much better for synthetically degraded structures than for "bad" autocontours. There are two main reasons for this observation: On one hand, the training dataset was heavily dominated by synthetically degraded structures with comparatively large deviations from the ground truth. On the other hand, "bad" automatic contours had mostly minor deviations from the ground truth (the threshold to label them "bad" was relatively low) and therefore likely only NCCs with very significant deviations from the ground truth were classified as NCC. Our analysis concentrated on proof of principle and therefore mainly on the well controlled synthetically degraded contours. Further refinements of the classifier will likely significantly improve performance on NCC created by autosegmentation.

Another support to the hypothesis that a robust feature-set could be sensitive to contour quality evaluation is given by results of specific TPR analysis for all the 25 cardiac sub-structures. Classification of bad/good contours is generally improved for ROIs that showed a greater ICC.

The image sets used to create the atlas database were composed of CTs without I.V. contrast. While I.V. contrast would improve autocontouring results for all structures in clinical practice dramatically and likely also further facilitate a radiomics driven analysis of contouring accuracy, non-contrast CTs are the imaging standard for breast cancer radiotherapy treatment planning and therefore retrospective treatment plan databases will mostly consist of non-contrast images, which motivated our choice for this disease paradigm.

To further improve the overall performance of the proposed autocontouring workflow, an a-priori integration of contour delimiters using biomechanical information (to improve interface detection, and limit contours to what is biomechanically possible) and the improvement of the detection capability of the AI Classifier, training with MC structure sets that are non-uniformly expanded/contracted (again integrating information based on biomechanically imposed limitations) could be investigated to create NCC-datasets that more closely resemble real anatomic variations.

5. Conclusions

A set of radiomics features that are robust on "gold-standard" contour datasets AND sensitive to incorrect contouring was identified and implemented in an AI workflow to quantify segmentation accuracy. This workflow will after further refinement permit an automatic assessment of segmentation quality and may therefore accelerate the expansion of an existing autocontouring atlas database as well as improve the quality of retrospective autocontouring-based dosimetric analyses of large treatment plan databases.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2021.05.009.

References

[1] Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. Semin Radiat Oncol 2019;29(3):185–97.

[2] Walker GV, Awan M, Tao R, Koay EJ, Boehling NS, Grant JD, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. Radiother Oncol 2014;112(3):321–5.

[3] Sims R, Isambert A, Grégoire V, Bidault F, Fresco L, Sage J, et al. A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. Radiother Oncol 2009;93(3):474–8.

[4] Dean JA, Welsh LC, McQuaid D, Wong KH, Aleksic A, Dunne E, et al. Assessment of fully-automated atlas-based segmentation of novel oral mucosal surface organ-at-risk. Radiother Oncol 2016;119(1):166–71.

[5] Teguh DN, Levendag PC, Voet PWJ, Al-Mamgani A, Han X, Wolf TK, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. Int J Radiat Oncol Biol Phys 2011;81(4):950–7.

[6] Jarrett D, Stride E, Vallis K, Gooding MJ. Applications and limitations of machine learning in radiation oncology. Br J Radiol 2019;92(1100):20190001. https://doi.org/10.1259/bjr.20190001.

[7] van Rooij W, Dahele M, Ribeiro Brandao H, Delaney AR, Slotman BJ, Verbakel WF. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. Int J Radiation Oncol Biol Phys 2019;104(3):677–84.

[8] Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother Oncol 2018;126(2):312–7.

[9] Giraud P, et al. Radiomics and machine learning for radiotherapy in head and neck cancers. Front Oncol 2019;9:174.

[10] Nikolov N et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv:1809.04430v1, 2018.

[11] Boon I, Au Yong T, Boon C. Assessing the role of artificial intelligence (AI) in clinical oncology: utility of machine learning in radiotherapy target volume delineation. Medicines 2018;5(4):131. https://doi.org/10.3390/medicines5040131.

[12] van der Veen J, Willems S, Deschuymer S, Robben D, Crijns W, Maes F, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. Radiother Oncol 2019;138:68–74.

[13] Kosmin M, Ledsam J, Romera-Paredes B, Mendes R, Moinuddin S, de Souza D, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. Radiother Oncol 2019;135:130–40.

[14] A Jungo, M Reyes Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation.. arXiv:1907.03338v2 [eess.IV] 2019.

[15] Chen HC, et al. Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: a general strategy. Med Phys 2015;42:1048–59.

[16] Rhee DJ, Cardenas CE, Elhalawani H, McCarroll R, Zhang L, Yang J, et al. Automatic detection of contouring errors using convolutional neural networks. Med Phys 2019;46(11):5086–97.

[17] Hui CB, Nourzadeh H, Watkins WT, Trifiletti DM, Alonso CE, Dutta SW, et al. Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach. Med Phys 2018;45(5):2089–96.

[18] Court LE, Kisling K, McCarroll R, Zhang L, Yang J, Simonds H, et al. Radiation planning assistant – a streamlined, fully automated radiotherapy treatment planning system. J Vis Exp 2018;(134). https://doi.org/10.3791/57411.

[19] Robinson R, et al. In: Automatic quality control of cardiac mri segmentation in large-scale population imaging. Springer; 2017. p. 720–7.

[20] Zhou L et al. Robust Image Segmentation Quality Assessment without Ground Truth. arXiv:1903.08773v1 [cs.CV] 20 Mar 2019.

[21] Jin J et al. Multiple U-Net-Based Automatic segmentations and Radiomics Feature Stability on Ultrasound Images for Patients With Ovarian Cancer. Front. Oncol. 2021.

[22] R J. Gillies et al. Radiomics: Images Are More than Pictures, They Are Data. Radiology. February 2016; 278(2): 563–577.

[23] S SF Yip et al. Applications and limitations of radiomics. Phys Med Biol. 2016 July 7; 61(13).

[24] Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS ONE 2014;9(7):e102107.

[25] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48(4):441–6.

[26] Owens CA, Peterson CB, Tang C, Koay EJ, Yu W, Mackin DS, et al. Lung tumor segmentation methods: impact on the uncertainty of radiomics features for non-small cell lung cancer. PLoS ONE 2018;13(10):e0205003.

[27] Duane F, Aznar MC, Bartlett F, Cutter DJ, Darby SC, Jagsi R, et al. A cardiac contouring atlas for radiotherapy. Radiother Oncol 2017;122(3):416–22.

[28] Darby SC, Ewertz M, McGale P, Bennet AM, Blom-Goldman U, Brønnum D, et al. Risk of ischemic heart disease in women after radiotherapy for breast cancer. N Engl J Med 2013;368(11):987–98.

[29] Maffei N, Fiorini L, Aluisio G, D'Angelo E, Ferrazza P, Vanoni V, et al. Hierarchical clustering applied to automatic atlas based segmentation of 25 cardiac sub-structures. Physica Med 2020;69:70–80.

[30] Weistrand O, Svensson S. The ANACONDA algorithm for deformable image registration in radiotherapy. Med Phys 2015;42(1):40–53.

[31] Delpon G, Escande A, Ruef T, Darréon J, Fontaine J, Noblet C, et al. Comparison of automated atlas-Based segmentation software for Postoperative Prostate cancer radiotherapy. Radiotherapy Front Oncol 2016;6. https://doi.org/10.3389/fonc.2016.00178.

[32] Kikinis R, Pieper SD, Vosburgh KG. In: Intraoperative Imaging and Image-Guided Therapy. New York, NY: Springer New York; 2014. p. 277–89. https://doi.org/10.1007/978-1-4614-7657-3_19.

[33] van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–7.

[34] https://www.slicer.org/wiki/Documentation/4.10/Extensions/Radiomicshttps://www.slicer.org/wiki/Documentation/4.10/Extensions/Radiomics.

[35] https://www.radiomics.io/pyradiomics.html www.radiomics.io/pyradiomics.html.

[36] Zwanenburg A et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020; volume 295, issue 2 / 328–338.

[37] https://pyradiomics.readthedocs.io/en/latest/features.html#.

[38] Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropract Med 2016;15(2):155–63.

[39] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1(1):30–46.

[40] https://www.mathworks.com/.

[41] Manco L, Maffei N, Strolin S, Vichi S, Bottazzi L, Strigari L. Basic of machine learning and deep learning in imaging for medical physicists. Phys Med 2021;83:194–205.