applied sciences

MDPI

Article

Clinical Evaluation of Commercial Deep Learning and Model-Based Segmentation Algorithms for Male Pelvic Structures in Prostate Cancer Computed Tomography Scans

Nicola Maffei [1], Marco Saguatti [1,2,*], Ercole Mazzeo [3], Marco Vernaleone [3], Giulia Miranda [4], Maria Victoria Gutierrez [1], Domenico Finocchiaro [1], Giulia Stocchi [3], Dario Corbelli [4], Maria Pia Morigi [5], Bruno Meduri [3], Alessio Bruni [3,4,*] and Gabriele Guidi [1]

1 Medical Physics Unit, AOU di Modena, 41125 Modena, Italy
2 Post Graduate School in Medical Physics, UNITO, 10125 Torino, Italy
3 Radiotherapy Unit, Department of Oncology and Hematology, Azienda Ospedaliero-Universitaria di Modena, 41125 Modena, Italy
4 Department of Medical and Surgical Sciences for Mother, Child and Adult, UNIMORE, 41125 Modena, Italy
5 Department of Physics and Astronomy, UNIBO, 40127 Bologna, Italy
* Correspondence: saguatti.marco@aou.mo.it (M.S.); bruni.alessio@aou.mo.it (A.B.)

Abstract

The performances of two autosegmentation algorithms were evaluated on 28 anonymized pelvic CT scans as a pilot study for the clinical implementation of a semi-automatic workflow. Four organs at risk (OARs), namely the rectum, bladder, and femoral heads, were contoured manually by an expert radiation oncologist (RO)—considered as the ground truth (GT)—and by model-based segmentation (MBS) and deep learning (DL) algorithms. Autocontouring performances were evaluated using a qualitative scoring system, contouring time analysis, and five geometrical indices: the 95th percentile Hausdorff Distance (95HD), Dice Similarity Coefficient (DSC), Surface Dice Similarity Coefficient (SDSC), Added Path Length (APL), and Relative Added Path Length (RAPL). Considering total median value for the four OARs, both MBS and DL showed clinically acceptable results with differences between the two algorithms being not statistically significant for almost all indices. The DL autocontouring algorithm achieved high geometric accuracy, high scores from the ROs, and consistent performances with all validation indices for every OAR. The MBS algorithm achieved high geometric accuracy for the femoral heads and bladder. The DL algorithm required 30 s to contour all the OARs, and the MBS algorithm required 90 s, showing a time gain compared with the manual contours, which took 20 min for each case. The DL autocontouring algorithm obtained promising but preliminary results with every evaluation metric and for every analyzed OAR. The application of the MBS algorithm as the only contouring tool still presents challenges.

Keywords: autocontouring; deep learning segmentation; model-based segmentation; geometrical indexes; pelvis

Academic Editor: Arkady Voloshin

Received: 17 December 2025
Revised: 26 January 2026
Accepted: 27 January 2026
Published: 29 January 2026

1. Introduction

In the last decades, modern radiotherapy (RT) techniques (e.g., 3D conformal radiation therapy, volumetric modulated arc therapy (VMAT), and helical tomotherapy) have been widely used for prostate cancer treatment for their capability to deliver a highly conformal dose distribution to the tumor target area [1,2]. Generally, these techniques use images

from computed tomography (CT) scans to precisely plan the treatment area, then radiation beams are conformed to the tumor from multiple directions to deliver higher radiation doses while sparing the normal tissues and organs around it. For this reason, an accurate delineation of the organs at risk (OARs) in the planning images (e.g., CT and MRI scans) is required to guarantee a safe and effective treatment [3].

In clinical practice, OARs' and tumor volumes' segmentations are manually performed by expert radiation oncologists (ROs) following RTOG contouring guidelines [4].

However, due to poor soft tissue contrast in CT images and complex anatomical shapes, this procedure is highly time-consuming, it requires training, and it is subjected to intra- and inter-observer variability [5,6]. These uncertainties can lead to severe consequences (e.g., acute or late rectal and urinary toxicities) [7], while the long time required can limit the number of contoured images [8].

Automatic segmentation, i.e., the generation of contours of OARs and target volumes on a digital image by a computer algorithm, can help ROs to delineate the required structures with greater accuracy and consistency and with less time consumption [9,10]. Deep learning techniques have demonstrated impressive performances in computer vision and medical image analysis applications, and thus, thanks to remarkable improvements in computer hardware, they can also be a potential solution for autocontouring in clinical practice [11–13]. However, to use these algorithms, a clinical evaluation must be performed in advance.

The model-based (MBS) and deep learning (DL) segmentation algorithms for the automatic delineation of the OARs and target volumes have been developed by RaySearch Laboratories® and are provided with the RayStation® 11B treatment planning system (TPS). A preliminary study for the clinical evaluation of these two algorithms has been performed for prostate cancer OARs to assess if their use can be beneficial in clinical applications. The performances of the autocontouring algorithms were evaluated using a Python (version 3.9.16) script with five geometrical indices, a qualitative scoring system, and contouring time analysis.

2. Materials and Methods

The workflow for the evaluation of the two autocontouring algorithms is depicted in Figure 1. For every anonymized CT scan, OARs were contoured manually by an expert RO, considered as the ground truth (GT), and by the MBS and the DL autocontouring algorithms of the RayStation® TPS. Then, the RT image and the RT structures were exported from the TPS. A Python script was written to load these data, to convert each contour to a binary label map, and to compare the manual and automatic contours using five geometrical indices: the 95th percentile Hausdorff distance (95HD), Dice Similarity Coefficient (DSC), Surface Dice Similarity Coefficient (SDSC), Added Path Length (APL), and Relative Added Path Length (RAPL). Qualitative scores (ranging from 1 (completely unacceptable contour) to 4 (contour clinically adequate without the need for manual corrections)) were assigned by two expert ROs to each OAR. A time analysis was finally performed.

2.1. Pelvis Dataset

A database containing 28 anonymized CT scans of prostate cancer was used to clinically evaluate two autocontouring algorithms for pelvic structures. All images were acquired by a Toshiba Aquilion® Large Bore CT (Canon Medical®, Minato, Japan) system with a slice thickness of 3 mm, an image matrix of 512 × 512 pixels, and a pixel spacing of 0.98 mm × 0.98 mm.

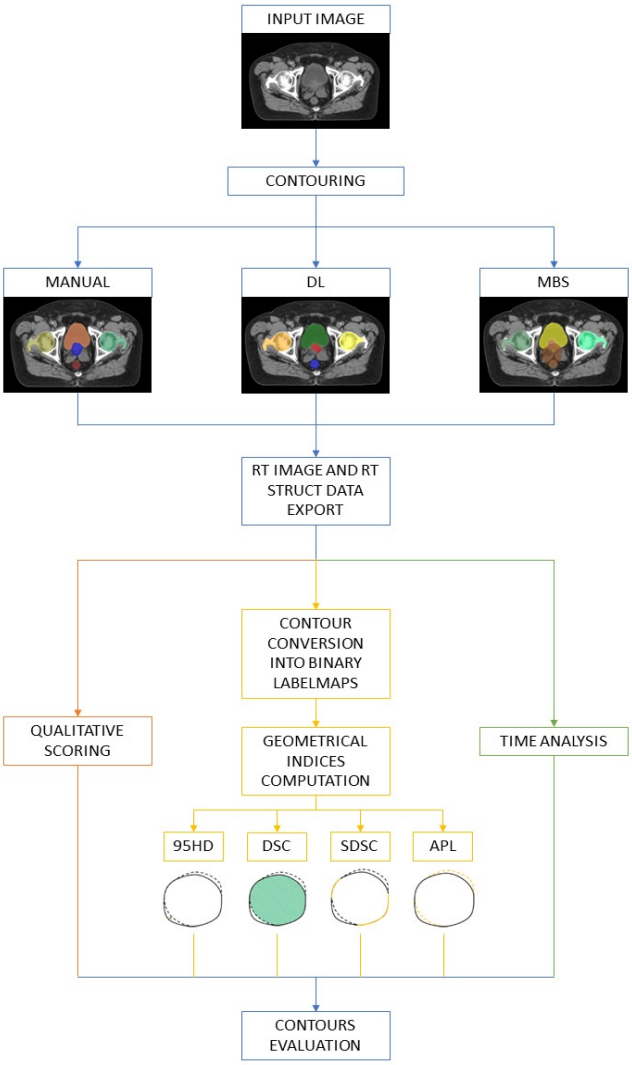

Figure 1. Workflow for the clinical evaluation of the two autocontouring algorithms.

For every CT image, four OARs were manually contoured, i.e., the rectum, bladder, and left and right femoral heads. The OARs were also contoured using both the MBS and the DL segmentation algorithms. A total of 336 contours were obtained and analyzed (Figure 2).

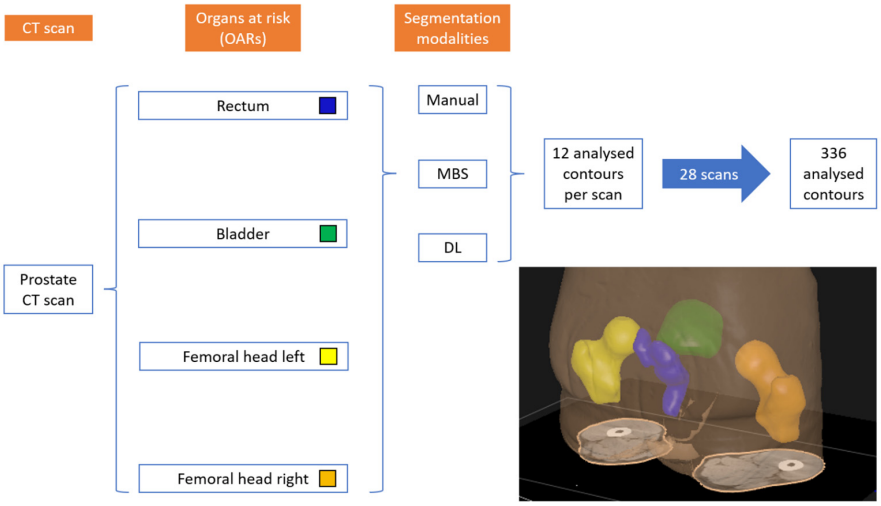

Figure 2. OARs contoured manually and by the MBS and DL segmentation algorithms.

2.2. RayStation® Autocontouring

The autocontouring task was performed using MBS and DL RayStation® proprietary algorithms.

2.2.1. MBS

The MBS algorithm is a semi-automatic tool for the delineation of organs in CT, cone beam CT (CBCT), and magnetic resonance image (MRI) sets. It combines greyscale gradients and shape statistics information to match 3D statistical shape models of the organs to new image data [14].

The model is automatically initialized on the basis of rigid and deformable image registration. Then, given the model and some greyscale information on the structures that must be contoured, image feature points are extracted and are matched to the model without violating the shape constraints of the model. This process is an iterative non-linear optimization where new feature points are extracted at each step [14]. This is a semi-automatic method because user interaction is needed to provide the greyscale information.

2.2.2. DL Segmentation

The DL model takes a 3D image as input and produces a labeled image as output for the delineation of anatomical structures in CT, CBCT, and MR image sets [14]. A voxel classifier neural network assigns each voxel its probability of belonging to a certain anatomical region [15]. The specific network architecture is a 3D version of U-net [16,17], i.e., a convolutional network consisting of a contracting path to capture context and a symmetric expansive path to enable precise localization.

DL networks take as input bounding boxes centered around one or more ROIs of the image and perform the segmentation using patch-based or non-patch-based models. For patch-based submodels the bounding box is split into multiple patches that match the dimensions of the submodel. Then, running the submodel for every patch, the entire bounding box is covered without resolution losses. For non-patch-based submodels, the bounding box is modified so that it matches the resolution of the submodel. Then, running the submodel only once, a segmentation of the entire bounding box region is achieved but with a loss of resolution [18].

2.3. Indices for Clinical Evaluation

To use autocontouring algorithms in clinical practice, their clinical utility and consistency must be assessed. However, a unique standard for the evaluation of autosegmentation algorithms does not exist. To choose the best metrics for the purpose, Michael V. Sherer et al. suggest considering the goals of autosegmentation, i.e., a reduction in contouring time, a decrease in intra- and inter-observer variability, and improvement of dose consistency and accuracy [19].

To clinically evaluate the DL and the MBS segmentation algorithms, the following geometric indices were used.

- The Dice Similarity Coefficient (DSC) ranges from 0 to 1 and measures the spatial overlap between two segmentation volumes, and it is defined as:

$$DSC = \frac{2\left|S_g^1 \cap S_t^1\right|}{\left|S_g^1\right| + \left|S_t^1\right|}. \quad (1)$$

Although it is one of the simplest metrics used to assess the validity of automatically segmented volumes, its correlation with physician-reported time savings is weak [20].

Moreover, it is not effective in distinguishing random from systematic errors [21] and discovering boundary errors [22].

- The Hausdorff Distance (HD) is a distance index measured in cm and is defined as:

$$H^{LM}(A,B) = \max\left(h^{L}(A,B), h^{M}(B,A)\right), \quad (2)$$

where:

$$h^{L}(A,B) = L^{th}_{a\in A}\min_{b\in B}\|a-b\|. \quad (3)$$

Surface distance-based metrics results are intuitive and quantitative. However, due to high sensitivity to small regions of poor segmentation, they are not always correlated with qualitative scoring and time saving in clinical practice [23].

- The Surface Dice Similarity Coefficient (SDSC) ranges from 0 to 1 and is an overlap-based metric. It measures the agreement between just the surfaces of two segmentation volumes above a clinically determined tolerance parameter τ [19]:

$$SDSC^{(\tau)}_{i,j} = \frac{\left|S_i \cap B_j^{(\tau)}\right| + \left|S_j \cap B_i^{(\tau)}\right|}{|S_i| + |S_j|}. \quad (4)$$

It is another geometric performance measure that showed a correlation with time saving in clinical practice [24]. A tolerance parameter τ = 3 mm was used.

- Added Path Length (APL) is defined as the absolute path length of a contour that has to be added to (or removed from) the automatic segmentation during editing to match the GT segmentation. It is measured in voxels and it is not normalized by volume [24].
- Relative Added Path Length (RAPL): because the APL varies between patients and different structures, Trimpl et al. [25] proposed to report the APL relative to the ground truth contour length.

A Python script was developed to compute these quantitative metrics and to extract data; the script is publicly available at GitHub (https://github.com/ (accessed on 25 January 2026)).

Moreover, a qualitative scoring system, similar to the ones proposed by Maffei et al. [26] and by Rauseo et al. [27], was used for the evaluation of the two autocontouring algorithms. Two expert ROs visually analyzed and evaluated every automatically generated contour with a score ranging from 1 to 4. A score of 1 was assigned to completely inaccurate contours considered clinically unacceptable. A score of 2 was assigned to significantly unacceptable contours that require numerous manual changes to make them clinically acceptable. A score of 3 was assigned to clinically acceptable contours that require minor editing to make them completely match the guidelines. A score of 4 was assigned to good quality contours that completely match the guidelines considered for the segmentation without needing any manual editing.

Finally, a time analysis between a full manual and the two automatic workflows, followed by manual correction of the contours, was carried out to assess if time savings can be achieved using the algorithms in clinical practice.

3. Results

Table 1 shows median values and 95% confidence intervals (95CIs) of the quantitative evaluation indices for the four OARs contoured by the MBS and DL algorithms, both compared with the GT. Due to a non-normal distribution of the data, a Mann–Whitney U test with significance level of α = 0.05 was carried out to assess if the distributions of the MBS and DL validation indices are statistically significantly different or not.

Table 1. Comparison of MBS and DL contours for the 4 pelvic structures on 28 CT scans using 95HD, DSC, SDSC, APL, and RAPL.

Evaluation Parameter	OAR	MBS (Median [95CI])	DL (Median [95CI])	*p*-Value
95HD (cm)	Rectum	3.12 [1.96–3.59]	1.06 [0.85–2.12]	<0.05
	Bladder	0.60 [0.39–1.17]	0.30 [0.30–0.32]	<0.05
	Fem. head (left)	0.90 [0.60–1.17]	1.20 [0.86–1.56]	0.22
	Fem. head (right)	0.43 [0.30–0.60]	1.36 [0.90–1.83]	<0.05
	Total	0.94 [0.60–1.24]	0.90 [0.66–1.20]	0.48
DSC	Rectum	0.61 [0.50–0.71]	0.84 [0.80–0.85]	<0.05
	Bladder	0.93 [0.89–0.95]	0.94 [0.94–0.95]	0.06
	Fem. head (left)	0.93 [0.90–0.94]	0.92 [0.90–0.92]	0.16
	Fem. head (right)	0.95 [0.93–0.96]	0.91 [0.90–0.92]	<0.05
	Total	0.91 [0.89–0.91]	0.91 [0.90–0.92]	0.99
SDSC	Rectum	0.51 [0.43–0.60]	0.85 [0.82–0.88]	<0.05
	Bladder	0.92 [0.80–0.94]	0.97 [0.96–0.98]	<0.05
	Fem. head (left)	0.90 [0.89–0.91]	0.91 [0.88–0.93]	0.88
	Fem. head (right)	0.85 [0.93–0.97]	0.90 [0.88–0.91]	<0.05
	Total	0.89 [0.83–0.91]	0.91 [0.89–0.92]	<0.05
APL (10^3 voxels)	Rectum	6.72 [5.27–7.42]	4.90 [3.90–5.53]	<0.05
	Bladder	7.15 [3.97–10.78]	7.24 [5.50–8.64]	0.92
	Fem. head (left)	4.12 [3.11–5.96]	5.71 [4.97–6.50]	0.12
	Fem. head (right)	3.87 [2.50–5.17]	6.09 [5.71–6.58]	<0.05
	Total	5.12 [4.43–5.98]	5.71 [5.47–6.14]	0.25
RAPL (10 voxels/cc)	Rectum	9.29 [7.39–10.48]	6.89 [5.72–8.03]	<0.05
	Bladder	3.08 [2.04–3.89]	2.95 [2.45–3.53]	0.79
	Fem. head (left)	2.34 [1.80–3.16]	3.22 [2.79–3.71]	0.12
	Fem. head (right)	2.11 [1.43–2.65]	3.36 [2.77–3.77]	<0.05
	Total	2.96 [2.45–3.45]	3.46 [3.11–3.77]	0.28

Figure 3 summarizes the results obtained with SDSC for all OARs and for both autocontouring algorithms (analogous boxplots for the other quantitative indices are reported in the Supplementary Materials).

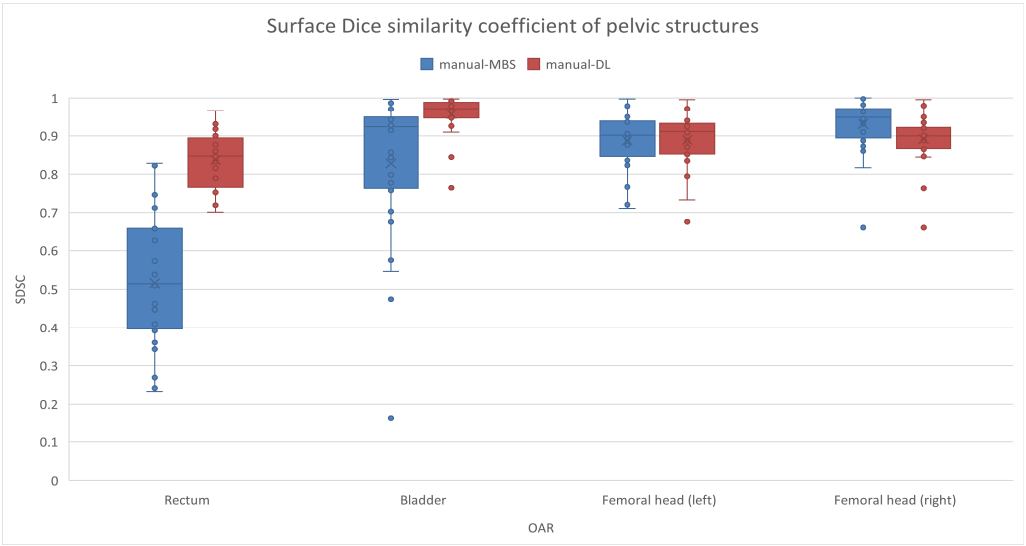

Figure 3. Boxplot of SDSC for the five OARs contoured using the MBS (blue boxes) and DL (red boxes) algorithms.

As shown in Table 1, considering the total median value for all four OARs, both MBS and DL showed clinical acceptable results, with differences between the two algorithms

being not statistically significant for almost all the indices (e.g., total median HD: 0.94 cm and 0.90 cm, and total median DSC: 0.91 and 0.91 for MBS and DL, respectively). SDSC better discriminated the results of the two autocontouring algorithms, showing a statistically significant difference with a p-value < 0.05 and a median value of 0.89 and 0.91 for MBS and DL, respectively.

In detail, considering every geometrical index individually, 95HD obtained the best results for the DL segmented bladder (Figure A1), which showed the lowest median value and a narrow 95CI (0.30 [0.30–0.36] cm), while the highest median value was the one of the MBS for the rectum (3.12 [1.96–3.59] cm). DSC obtained the best results for the MBS-segmented right femoral head (0.95 [0.93–0.96]), as shown in Figure A2. SDSC obtained the best results for the DL-segmented bladder (0.97 [0.96–0.98]), as shown in Figure 3. On the contrary, DSC and SDSC obtained the worst results for the MBS for the rectum (0.61 [0.50–0.71] and 0.51 [0.43–0.60], respectively). Both APL and RAPL obtained the best results for the MBS for the right femoral head (3.87 [2.50–5.17] $\times 10^3$ voxels and 2.11 [1.43–2.65] $\times$ 10 voxels/cc, respectively). With APL, the worst accuracy was obtained with the DL-segmented bladder (7.24 $\times 10^3$ voxels) and the worst precision with the MBS-segmented bladder ([3.97–10.78] $\times 10^3$ voxels), while with RAPL, the worst results were obtained with the MBS-segmented rectum (9.29 [7.39–10.48] $\times$ 10 voxels/cc).

A Mann–Whitney U test with significance level of $\alpha = 0.05$ was used to compare the two autocontouring algorithms. In almost every comparison, the p-value was lower than 0.05, highlighting a statistically significant difference in the performances of the two algorithms. For the left femoral head only, none of the geometrical indices showed a statistically significant difference, while for the bladder, DSC, APL, and RAPL did not show a statistically significant difference.

The qualitative scores assigned by the two expert ROs showed at least clinically acceptable contours (i.e., score $\geq$ 3) for more than 66% and 98% of the MBS- and DL-generated contours, respectively. Table 2 shows the distribution of scores for each OAR and each segmentation method. Higher scores were assigned to DL-generated contours with respect to the MBS ones for the rectum and bladder, in agreement with the results obtained with the quantitative indices (Figure 4 shows an example of a contoured bladder). Different results were obtained for the femoral heads, where DL contours obtained higher scores even if the quantitative indices highlighted the higher accuracy of the MBS contours.

Table 2. Score distribution for each OAR and each autocontouring method.

OAR	Segmentation Method	Score			
		1	2	3	4
Rectum	MBS	12	28	13	3
	DL	0	0	14	42
Bladder	MBS	5	24	26	1
	DL	0	3	15	38
Femoral head (left)	MBS	0	5	50	1
	DL	0	1	9	46
Femoral head (right)	MBS	0	2	51	3
	DL	0	1	8	47

The time comparison between fully manual procedures and automatic procedures followed by eventual manual editing showed that the MBS autocontouring algorithm required, on average, 90 s to perform the segmentation of the four OARs. The DL algorithm required, on average, 30 s to perform the same task. Moreover, contours that did not obtain an evaluation of 4 from the expert ROs required some manual editing to make

them match the GT contours. The time required to correct the automatically generated segmentations was compared with the time required to manually contour them entirely, which takes around 20 min for each patient. The time spent by the ROs to correct the automatic contours evaluated as 1 or 2 was higher than time required for exclusive manual contouring. However, for contours with an evaluation of 3, the time required to adjust them was smaller compared with fully manual contouring.

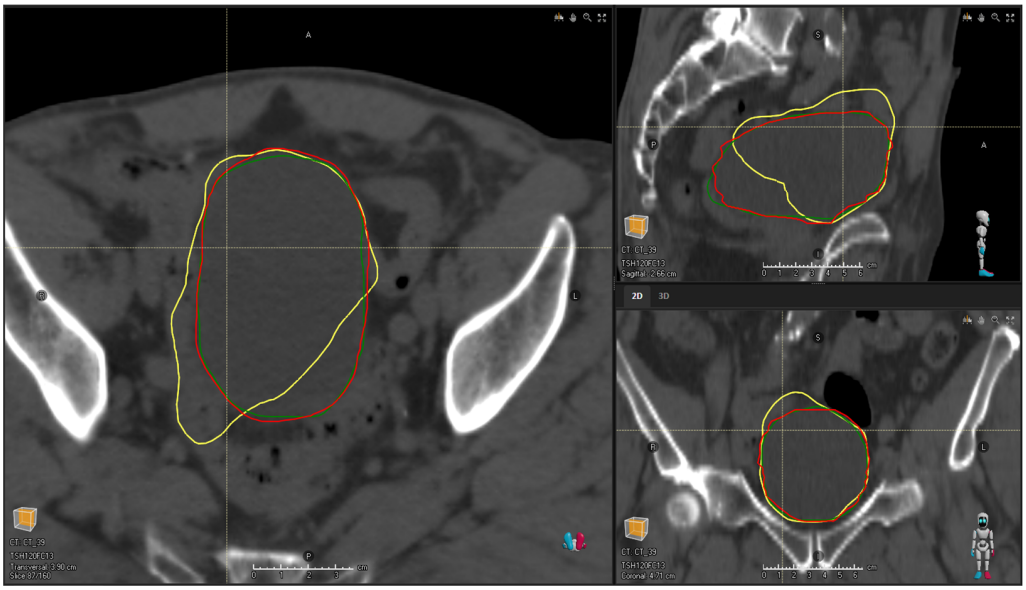

Figure 4. Example of a bladder contoured by the ROs (red), the DL algorithm (green), and the MBS algorithm (yellow). The MBS contour is a contour evaluated as 1 by the ROs, while the DL one is a contour evaluated as 3.

4. Discussion

In the last few years, an expansion of the use of machine learning and artificial intelligence (AI) algorithms in clinical practice has been observed, but guidelines about proper uses of these new tools are still missing or barely under development. This study was not only about evaluating the performances of two autocontouring algorithms but also a preliminary effort to define a set of useful metrices to consider for a more aware use of AI tools before their complete integration in clinical routine.

The autocontouring performances of the MBS and DL segmentation algorithms for pelvic structures were evaluated on 28 anonymized male pelvis CT scans, considering the manual contours as the GT. Some of the most used geometric indices, i.e., DSC and HD, as well as novel geometric performance metrics, i.e., SDSC, APL, and RAPL, have been used for the evaluation, along with non-geometrical indices. A qualitative score was assigned by two expert ROs to the automatically generated segments to assess the quality of the contours, based on clinical applicability. Then, a speed comparison between a completely manual procedure and an automatic procedure, followed by manual editing of the contours, was carried out to assess if time saving can be achieved by using the algorithms in clinical practice.

The DL autocontouring algorithm achieved high geometric accuracy (with both DSC and SDSC > 0.70 [28]) and high scores from the expert ROs (median values greater than 3 out of 4) for every OAR. Moreover, it showed very consistent performances, with small 95CIs, for almost all geometrical indices and all analyzed OARs. Thus, its application in clinical practice could bring significant time savings, with the need for small manual corrections only in a limited number of cases.

The MBS algorithm achieved high geometric accuracy for the femoral heads (also higher than the DL) and the bladder, while a low accuracy (DSC and SDSC < 0.70) was achieved for the rectum. The consistency of this algorithm was lower than the one achieved with DL for every geometrical index and every OAR. Moreover, the time required to run this algorithm was three times higher than the time required to run the DL one (less than 2 min in both cases). These results can be due to an input grey level that is more suitable for femoral head segmentation than for the other OARs. Running the algorithm with different input grey levels can bring better results for all OARs but with increased time required to obtain the contours.

The qualitative scores assigned by the ROs to MBS-generated contours were lower than the scores assigned to the DL ones for every OAR. The qualitative scores showed that the best results were obtained by the DL-segmented femoral heads and rectum, while the worst ones were achieved by the MBS-segmented rectum.

The discrepancy between quantitative indices and qualitative scores for femoral heads is due to a difference in the number of contoured CT slices with the three segmentation methods. The number of slices contoured by DL is greater than the number of slices contoured by the ROs and MBS.

The time comparison between fully manual procedures and automatic procedures followed by manual editing highlighted that time savings in clinical practice are strictly related to contour quality. Poorly rated contours could require major RO editing that take more time to be performed than a completely manual contouring procedure.

This study has some strengths, i.e., the usage of a wide range of metrics and the integration of quantitative metrics, qualitative scores, and time analysis, but it also has a few limitations. Only 28 CT scans were used for the validation; however, due to the number of contours for the evaluation of the two algorithms (112 for each algorithm), it was considered a good starting point to perform a pilot study. Only prostate tumor CT scans acquired with the same imaging device and scanning protocol were involved in the study; this choice was made to avoid an increase in variability during evaluation. For the same reason, CT was used as the only imaging modality. Only one RO expert in prostate cancer was involved in the manual contouring and editing to have unique GT contours. This limitation can be assessed in a future multidisciplinary study.

For future research, bigger datasets that include CT scans acquired with different imaging systems and scanning protocols, as well as multi-center comparisons, could be assessed to increase the substantial evidence, also including MR image sets as the imaging modality. Moreover, different pathologies other than pelvis tumors could be used to conduct further research.

5. Conclusions

The feasibility of the implementation of a semi-automatic workflow for the segmentation of pelvic OARs in our clinical practice was evaluated in this pilot study. Five geometrical indices, a qualitative scoring system, and time analysis were used to clinically evaluate the MBS and DL autocontouring algorithms provided by a commercial TPS for the pelvic anatomical region. Because there is not a unique index that can be used for the evaluation, their combination provides more complete information that better reflects the clinical complexity.

The DL algorithm showed high geometric accuracy and high consistency with every geometrical index for every OAR. These results were in line with the qualitative scores assigned by the expert ROs. The MBS algorithm showed high geometric accuracy for the femoral heads. Both qualitative and quantitative evaluations stated that major manual correction could be needed for rectum and bladder contours. According to the findings of

this preliminary study, the application of this algorithm as the only (unsupervised by ROs) contouring tool still presents challenges.

Even though the use of autocontouring tools could speed up the clinical procedure (up to 90% of time gained), human intervention is still needed for the evaluation and the judgement of the entire procedure. As stated by Jarret et al. [29]: "machines may excel at replicating, automating and standardizing human behavior on manual chores, meanwhile the conceptual clinical challenges relating to definition, evaluation, and judgement remain in the realm of human intelligence and insight".

Supplementary Materials: The Python script used for the extraction of the geometrical indices can be downloaded at https://github.com/MarcoSaguatti/Multimetric_analysis_of_automatic_contours (accessed on 25 January 2026).

Author Contributions: Conceptualization and Methodology: N.M., M.S. and E.M. Data curation: E.M., G.M. and D.C. Resources: B.M. Software: M.S. Investigation: E.M. and A.B. Formal analysis: D.F., M.V.G. and G.S. Writing—original draft preparation: N.M., M.S. and E.M. Writing—review and editing: A.B., G.G., M.V. and M.P.M. Supervision and visualization: A.B. and G.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethical Committee of Area Vasta Emilia Nord (AVEN) (protocol code 129/12 of 10 July 2012).

Informed Consent Statement: The study used 28 anonymized planning CT scans from patients enrolled in a clinical trial evaluating a hypofractionated radiotherapy regimen for prostate cancer. The present analysis did not assess the clinical outcomes of the original trial. All patients provided written informed consent.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

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

Abbreviations

The following abbreviations are used in this manuscript.

CT	Computed Tomography
OARs	Organs At Risk
RO	Radiation Oncologist
GT	Ground Truth
MBS	Model-Based Segmentation
DL	Deep Learning
95HD	95th Percentile Housdorff Distance
DSC	Dice Similarity Coefficient
SDSC	Surface Dice Similarity Coefficient
APL	Added Path Length
RAPL	Relative Added Path Length
RT	Radiotherapy
VMAT	Volumetric Modulated Arc Therapy
MRI	Magnetic Resonance Imaging
MR	Magnetic Resonance
TPS	Treatment Planning System
CBCT	Cone Beam Computed Tomography
95CIs	95% Confidence Intervals
AI	Artificial Intelligence

Appendix A. Boxplots

In this section, the boxplots of all geometrical indices other than SDSC are reported.

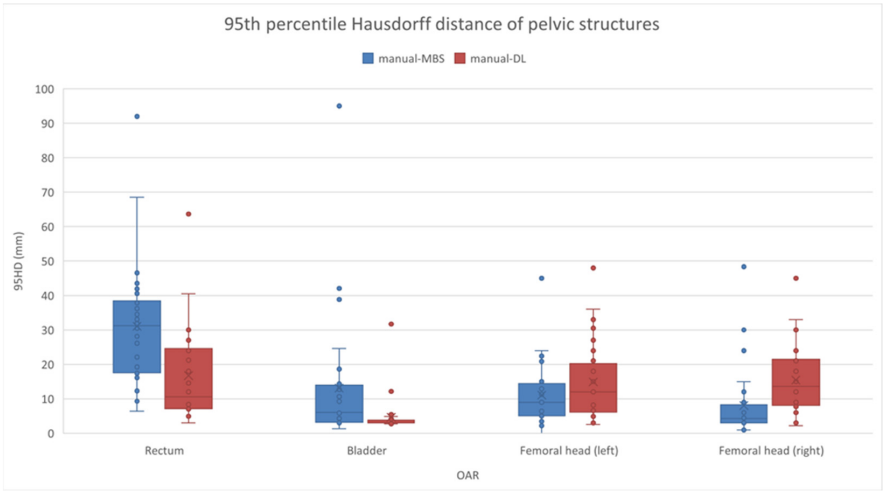

Figure A1. Boxplot of 95HD for the five OARs contoured using the MBS (blue boxes) and DL (red boxes) algorithms.

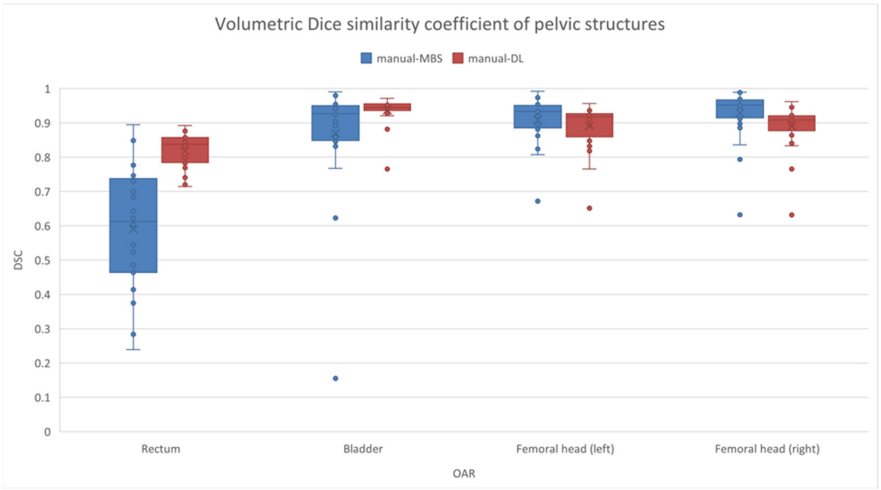

Figure A2. Boxplot of DSC for the five OARs contoured using the MBS (blue boxes) and DL (red boxes) algorithms.

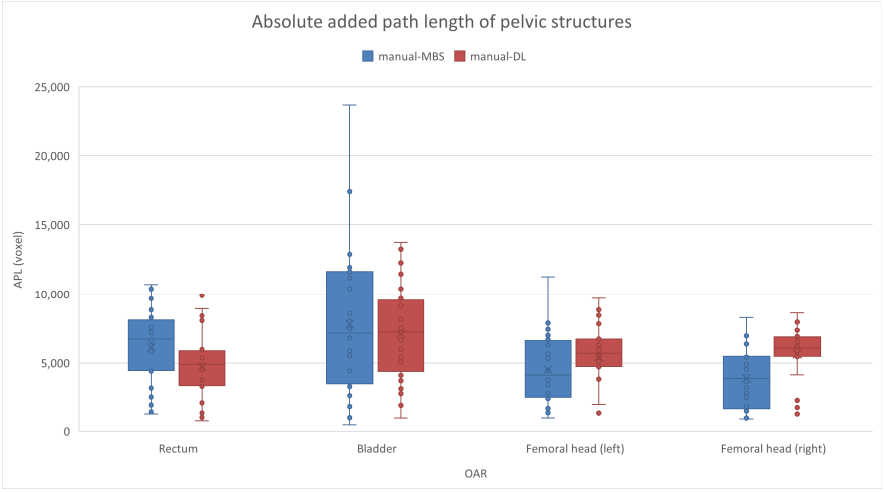

Figure A3. Boxplot of APL for the five OARs contoured using the MBS (blue boxes) and DL (red boxes) algorithms.

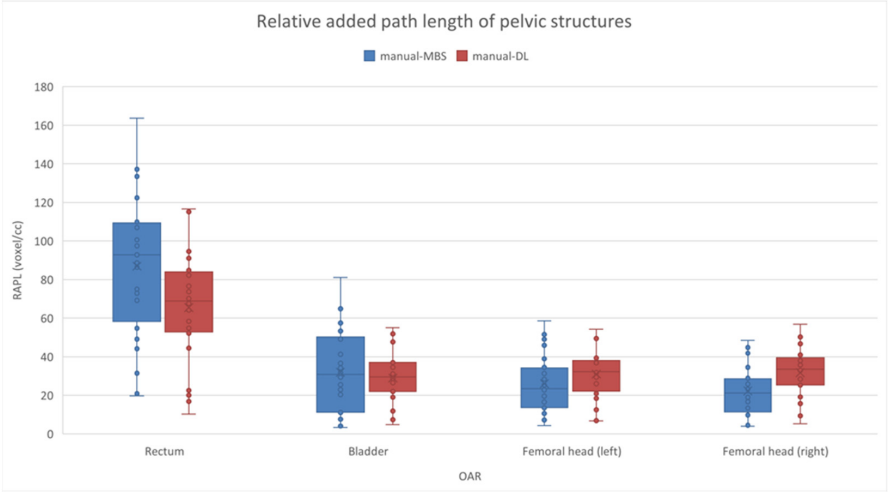

Figure A4. Boxplot of RAPL for the five OARs contoured using the MBS (blue boxes) and DL (red boxes) algorithms.

References

1. Tran, A.; Zhang, J.; Woods, K.; Yu, V.; Nguyen, D.; Gustafson, G.; Rosen, L.; Sheng, K. Treatment Planning Comparison of IMPT, VMAT and 4π Radiotherapy for Prostate Cases. *Radiat. Oncol.* **2017**, *12*, 10. [CrossRef]
2. Wang, W.; Wang, Q.; Jia, M.; Wang, Z.; Yang, C.; Zhang, D.; Wen, S.; Hou, D.; Liu, N.; Wang, P.; et al. Deep Learning-Augmented Head and Neck Organs at Risk Segmentation from CT Volumes. *Front. Phys.* **2021**, *9*, 743190. [CrossRef]
3. Doolan, P.J.; Charalambous, S.; Roussakis, Y.; Leczynski, A.; Peratikou, M.; Benjamin, M.; Ferentinos, K.; Strouthos, I.; Zamboglou, C.; Karagiannis, E. A Clinical Evaluation of the Performance of Five Commercial Artificial Intelligence Contouring Systems for Radiotherapy. *Front. Oncol.* **2023**, *13*, 1213068. [CrossRef]
4. Gay, H.A.; Barthold, H.J.; O'Meara, E.; Bosch, W.R.; El Naqa, I.; Al-Lozi, R.; Rosenthal, S.A.; Lawton, C.; Lee, W.R.; Sandler, H.; et al. Pelvic Normal Tissue Contouring Guidelines for Radiation Therapy: A Radiation Therapy Oncology Group Consensus Panel Atlas. *Int. J. Radiat. Oncol. Biol. Phys.* **2012**, *83*, E353–E362. [CrossRef]
5. Kalantar, R.; Lin, G.; Winfield, J.M.; Messiou, C.; Lalondrelle, S.; Blackledge, M.D.; Koh, D.M. Automatic Segmentation of Pelvic Cancers Using Deep Learning: State-of-the-Art Approaches and Challenges. *Diagnostics* **2021**, *11*, 1964. [CrossRef] [PubMed]
6. Acosta, O.; Dowling, J.; Drean, G.; Simon, A.; de Crevoisier, R.; Haigron, P. Multi-Atlas-Based Segmentation of Pelvic Structures from CT Scans for Planning in Prostate Cancer Radiotherapy. In *Abdomen and Thoracic Imaging: An Engineering & Clinical Perspective*; El-Baz, A.S., Saba, L., Suri, J., Eds.; Springer: New York, NY, USA, 2014. [CrossRef]
7. Zelefsky, M.J.; Fuks, Z.; Hunt, M.; Yamada, Y.; Marion, C.; Ling, C.C.; Amols, H.; Venkatraman, E.S.; Leibel, S.A. High-Dose Intensity Modulated Radiation Therapy for Prostate Cancer: Early Toxicity and Biochemical Outcome in 772 Patients. *Int. J. Radiat. Oncol. Biol. Phys.* **2002**, *53*, 1111–1116. [CrossRef] [PubMed]
8. Guo, D.; Jin, D.; Zhu, Z.; Ho, T.-Y.; Harrison, A.P.; Chao, C.-H.; Xiao, J.; Ku, L. Organ at Risk Segmentation for Head and Neck Cancer Using Stratified Learning and Neural Architecture Search. *arXiv* **2020**, arXiv:2004.08426. [CrossRef]
9. Greenham, S.; Dean, J.; Fu, C.K.K.; Goman, J.; Mulligan, J.; Tune, D.; Sampson, D.; Westhuyzen, J.; McKay, M. Evaluation of Atlas-Based Auto-Segmentation Software in Prostate Cancer Patients. *J. Med. Radiat. Sci.* **2014**, *61*, 151–158. [CrossRef]
10. Teguh, D.N.; Levendag, P.C.; Voet, P.W.J.; Al-Mamgani, A.; Han, X.; Wolf, T.K.; Hibbard, L.S.; Nowak, P.; Akhiat, H.; Dirkx, M.L.P.; 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*, 950–957. [CrossRef]
11. Ibragimov, B.; Xing, L. Segmentation of Organs-at-Risks in Head and Neck CT Images Using Convolutional Neural Networks. *Med. Phys.* **2017**, *44*, 547–557. [CrossRef]
12. Li, X.; Li, L.; Li, M.; Yan, P.; Feng, T.; Luo, H.; Zhao, Y.; Yin, S. Knowledge Distillation and Teacher–Student Learning in Medical Imaging: Comprehensive Overview, Pivotal Role, and Future Directions. *Med. Image Anal.* **2026**, *107*, 103819. [CrossRef]
13. Li, X.; Li, L.; Jiang, Y.; Wang, H.; Qiao, X.; Feng, T.; Luo, H.; Zhao, Y. Vision-Language Models in Medical Image Analysis: From Simple Fusion to General Large Models. *Inf. Fusion* **2025**, *118*, 102995. [CrossRef]
14. RaySearch Laboratories. *RayStation 11B User Manual*; RaySearch Laboratories: Stockholm, Sweden, 2021.
15. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Dian, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *J. Big Data* **2021**, *8*, 53. [CrossRef]

16. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597. [CrossRef]
17. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *arXiv* **2016**, arXiv:1606.06650. [CrossRef]
18. RaySearch Laboratories. *Machine Learning Reference Manual*; RaySearch Laboratories: Stockholm, Sweden, 2021.
19. Sherer, M.V.; Lin, D.; Elguindi, S.; Duke, S.; Tan, L.-T.; Cacicedo, J.; Dahele, M.; Gillespie, E.F. Metrics to Evaluate the Performance of Auto-Segmentation for Radiation Treatment Planning: A Critical Review. *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **2021**, *160*, 185–191. [CrossRef] [PubMed]
20. Cha, E.; Elguindi, S.; Onochie, I.; Gorovets, D.; Deasy, J.O.; Zelefsky, M.; Gillespie, E.F. Clinical Implementation of Deep Learning Contour Autosegmentation for Prostate Radiotherapy. *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **2021**, *159*, 1–7. [CrossRef] [PubMed]
21. Valentini, V.; Boldrini, L.; Damiani, A.; Muren, L.P. Recommendations on How to Establish Evidence from Auto-Segmentation Software in Radiotherapy. *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **2014**, *112*, 317–320. [CrossRef] [PubMed]
22. Taha, A.A.; Hanbury, A. Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool. *BMC Med. Imaging* **2015**, *15*, 29. [CrossRef]
23. Gooding, M.J.; Smith, A.J.; Tariq, M.; Aljabar, P.; Peressutti, D.; van der Stoep, J.; Reymen, B.; Emans, D.; Hattu, D.; van Loon, J.; et al. Comparative Evaluation of Autocontouring in Clinical Practice: A Practical Method Using the Turing Test. *Med. Phys.* **2018**, *45*, 5105–5115. [CrossRef]
24. Vaassen, F.; Hazelaar, C.; Vaniqui, A.; Gooding, M.; Van der Heyden, B.; Canters, R.; Van Elmpt, W. Evaluation of Measures for Assessing Time-Saving of Automatic Organ-at-Risk Segmentation in Radiotherapy. *Phys. Imaging Radiat. Oncol.* **2020**, *13*, 1–6. [CrossRef]
25. Trimpl, M.J.; Boukerroui, D.; Stride, E.P.J.; Vallis, K.A.; Gooding, M.J. Interactive Contouring through Contextual Deep Learning. *Med. Phys.* **2021**, *48*, 2951–2959. [CrossRef]
26. Maffei, N.; Fiorini, L.; Aluisio, G.; D'Angelo, E.; Ferrazza, P.; Vanoni, V.; Lohr, F.; Meduri, B.; Guidi, G. Hierarchical Clustering Applied to Automatic Atlas Based Segmentation of 25 Cardiac Sub-Structures. *Phys. Medica* **2020**, *69*, 70–80. [CrossRef] [PubMed]
27. Rauseo, E.; Omer, M.; Amir-Khalili, A.; Sojoudi, A.; Le, T.-T.; Cook, S.A.; Hausenloy, D.J.; Ang, B.; Toh, D.-F.; Bryant, J.; et al. A Systematic Quality Scoring Analysis to Assess Automated Cardiovascular Magnetic Resonance Segmentation Algorithms. *Front. Cardiovasc. Med.* **2021**, *8*, 816985. [CrossRef]
28. Zijdenbos, A.P.; Dawant, B.M.; Margolin, R.A.; Palmer, A.C. Morphometric Analysis of White Matter Lesions in MR Images: Method and Validation. *IEEE Trans. Med. Imaging* **1994**, *13*, 716–724. [CrossRef] [PubMed]
29. Jarrett, D.; Stride, E.; Vallis, K.; Gooding, M.J. Applications and Limitations of Machine Learning in Radiation Oncology. *Br. J. Radiol.* **2019**, *92*, 20190001. [CrossRef] [PubMed]