Machine learning-based detection of aberrant deep learning segmentations of target and organs at risk for prostate radiotherapy using a secondary segmentation algorithm

Claessens, Michael; Vanreusel, Verdi; De Kerf, Geert; Mollaert, Isabelle; Lofman, Fredrik; Gooding, Mark J.; Brouwer, Charlotte; Dirix, Piet; Verellen, Dirk

Published in:
Physics in Medicine and Biology

DOI:
10.1088/1361-6560/ac6fad

PAPER • OPEN ACCESS

Machine learning-based detection of aberrant deep learning segmentations of target and organs at risk for prostate radiotherapy using a secondary segmentation algorithm

To cite this article: Michaël Claessens *et al* 2022 *Phys. Med. Biol.* **67** 115014

IOP Publishing Phys. Med. Biol. 67 (2022) 115014 https://doi.org/10.1088/1361-6560/ac6fad

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
18 August 2021

REVISED
26 April 2022

ACCEPTED FOR PUBLICATION
13 May 2022

PUBLISHED
27 May 2022

PAPER

Machine learning-based detection of aberrant deep learning segmentations of target and organs at risk for prostate radiotherapy using a secondary segmentation algorithm

Michaël Claessens[1,2], Verdi Vanreusel[1], Geert De Kerf[1], Isabelle Mollaert[1], Fredrik Löfman[3], Mark J Gooding[4], Charlotte Brouwer[5], Piet Dirix[1,2] and Dirk Verellen[1,2]

[1] Department of Radiation Oncology, Iridium Network, Wilrijk (Antwerp), Belgium
[2] Centre for Oncological Research (CORE), Integrated Personalized and Precision Oncology Network (IPPON), University of Antwerp, Belgium
[3] Department of Machine Learning, RaySearch Laboratories AB, Stockholm, Sweden
[4] Mirada Medical Ltd, Oxford, United Kingdom
[5] University of Groningen, University Medical Center Groningen, Department of Radiation Oncology, The Netherlands

E-mail: michael.claessens@uantwerpen.be

Keywords: quality assurance, auto-segmentation, artificial intelligence, machine learning, deep learning, time-saving

Abstract

Objective. The output of a deep learning (DL) auto-segmentation application should be reviewed, corrected if needed and approved before being used clinically. This verification procedure is labour-intensive, time-consuming and user-dependent, which potentially leads to significant errors with impact on the overall treatment quality. Additionally, when the time needed to correct auto-segmentations approaches the time to delineate target and organs at risk from scratch, the usability of the DL model can be questioned. Therefore, an automated quality assurance framework was developed with the aim to detect in advance aberrant auto-segmentations. *Approach.* Five organs (prostate, bladder, anorectum, femoral head left and right) were auto-delineated on CT acquisitions for 48 prostate patients by an in-house trained primary DL model. An experienced radiation oncologist assessed the correctness of the model output and categorised the auto-segmentations into two classes whether minor or major adaptations were needed. Subsequently, an independent, secondary DL model was implemented to delineate the same structures as the primary model. Quantitative comparison metrics were calculated using both models' segmentations and used as input features for a machine learning classification model to predict the output quality of the primary model. *Main results.* For every organ, the approach of independent validation by the secondary model was able to detect primary auto-segmentations that needed major adaptation with high sensitivity (recall $= 1$) based on the calculated quantitative metrics. The surface DSC and APL were found to be the most indicated parameters in comparison to standard quantitative metrics for the time needed to adapt auto-segmentations. *Significance.* This proposed method includes a proof of concept for the use of an independent DL segmentation model in combination with a ML classifier to improve time saving during QA of auto-segmentations. The integration of such system into current automatic segmentation pipelines can increase the efficiency of the radiotherapy contouring workflow.

Glossary

ACROP	Advisory Committee for Radiation Oncology Practice
AI	Artificial Intelligence
ART	Adaptive Radiation Therapy
APL	Added Path Length
RT	Radiation Therapy

CNN	Convolutional Neural Networks
CT	Computed Tomography
DL	Deep Learning
DSC	Dice Similarity Coefficient
ESTRO	European Society for Radiotherapy and Oncology
FOV	Field Of View
HD	Hausdorff Distance
ML	Machine Learning
MRI	Magnetic Resonance Imaging
OAR	Organ at risk
QA	Quality Assurance
RO	Radiation Oncologist
RTOG	Radiation Therapy Oncology Group
SMOTe	Synthetic Minority Over-Sampling Technique
SVM	Support Vector Machine
TPS	Treatment Planning System

1. Introduction

In the world of medicine, radiation therapy (RT) is established as a cornerstone for curative treatment used as a single or combined modality in more than 40% of the cancer patients (Borras *et al* 2016). The main goal of the RT treatment is to deliver a high conformal radiation dose to the target to achieve an optimal, curative effect and, at the same time, spare the nearby healthy organs at risk (OARs) preventing acute radiation toxicity and chronic complications. A prominent step during this treatment process is an accurate, spatial delineation of the target and OARs (2011, Kosmin *et al* 2019). The segmentation is usually performed by physicians or experienced planners, which is a laborious and time-consuming procedure accompanied with an intra- and inter-observer variability (Brouwer *et al* 2012, Vinod *et al* 2016, van der Veen *et al* 2019, Chen *et al* 2020a). With the introduction of deep learning (DL) in RT, convolutional neural network (CNN)-based segmentation models have shown to improve consistency and efficiency of this process (Lustberg *et al* 2018, Cardenas *et al* 2019, Unkelbach *et al* 2020). Such DL architecture typically categories every voxel in an image to a target or OAR based on features of the position and intensity of the voxel and surrounding voxels. Due to extensive research by different institutions, these models are now outperforming traditional auto-contouring methods and have reached the accuracy of expert delineations (Boldrini *et al* 2019). In addition, due to lower resolution of dose grids compared to planning image data, as well as uncertainties in patient positioning and anatomical changes during the course of treatment, plans containing contouring deviations in the range of several millimetres may still produce a very similar dose (Sharp *et al* 2014). As a result, the implementation of auto-segmented contours may not have any significant negative effect on treatment efficacy. However, auto-segmentation errors can still occur that range from minor (e.g. boundary miss) to major (e.g. missing slices) geometric deviations during clinical implementation. Prominent causes are the individual patient anatomy, which can differ from the model training (e.g. surgical removal of tissue) and imaging acquisition protocols, which can differ across RT centers and over time. In this context, verification of auto-segmentations remains necessary, although clinical experts could spend a substantial amount of time examining and modifying the segmentations slice by slice. Ultimately, this could decrease the potential benefit of automated segmentation (Chen *et al* 2020b).

To tackle this issue, several methods have already been developed to automate this verification step (Claessens and Oria 2022). First attempts calculated 2D and/or 3D volumetric features of historical patients to create statistical feature probability distributions per contour (Altman *et al* 2015, Hui *et al* 2018). Instead of only considering the geometric attributes of contours (centroid, volume, and shape), the spatial relationship of neighbouring structures, as well as the anatomical similarity of individual contours among patients, can also be used as training input. That way, several machine learning (ML) models were able to characterize the inter-structural centroid and volume variations and the intra-structural shape variations of each individual structure (McIntosh *et al* 2013, Chen *et al* 2015). The use of DL approaches for QA of auto-segmentation is still an emerging area. One study proposed the use of 2D computed tomography (CT) images with its corresponding 2D probability map and uncertainty map to predict the Dice Similarity Coefficient (DSC) scores between ground

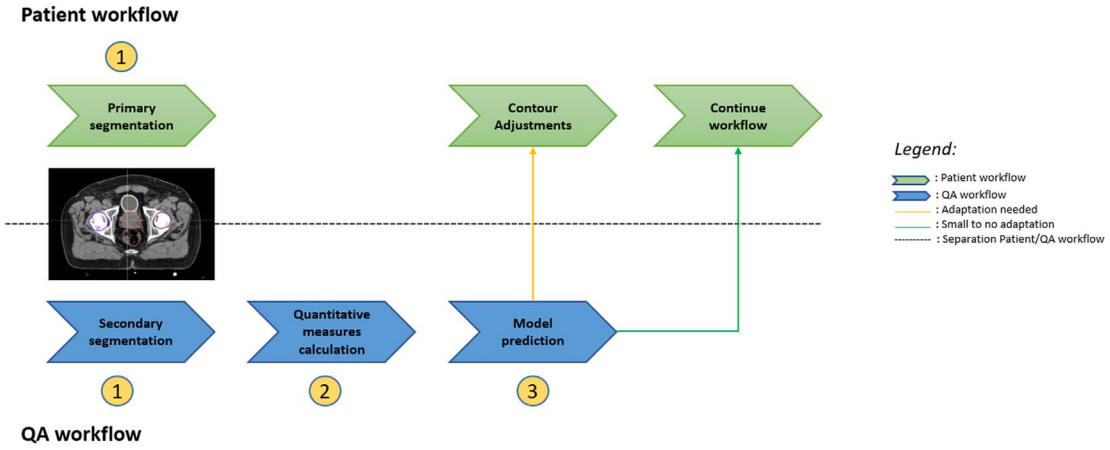

Figure 1. Automated QA workflow for DL auto-segmentations for prostate radiotherapy.

truth and auto-segmentation (Chen *et al* 2020b). Another similar approach could be to create spatial probability maps by applying Monte Carlo Dropout and Gaussian distribution (van Rooij *et al* 2021).

In this study, we propose an automatic QA platform to assess the accuracy of the auto-segmentations of an in-house trained primary model. An experiment was designed where prostate patients were auto-delineated both by the primary model and an independent, secondary DL segmentation algorithm. This approach is similar to introducing a secondary (independent) dose calculation algorithm to validate the clinical dose calculation algorithm. The latter requires high accuracy and tight QA procedures, whereas the former can be more generic and may be less accurate, yet able to detect clinically relevant deviations. Subsequently, nine different quantitative metrics (e.g. DSC,…) were calculated with the aim of defining a state of similarity or difference between both auto-segmentations. Finally, these metrics were used as input features for a ML classifier to predict the quality of the primary model segmentations. In this way, such predictive algorithm can serve as a guardian that detects in advance minor auto-segmentations, which can be directly presented to the clinical expert.

2. Material and methods

2.1. Automated QA workflow

The overall architecture of the automated QA workflow for auto-segmentations is illustrated in figure 1. The presented pipeline runs through three, consecutive steps: (1) obtaining auto-segmentations by the primary and secondary DL model; (2) calculating segmentation metrics between both auto-segmentations as input features for ML classifier; (3) ML-based classification to highlight unacceptable aberrations.

2.2. Patient cohort

A cohort of 48 prostate patients with or without lymph node invasion were randomly selected, of which 42 patients underwent conventional RT treatment (60 Gy/20 fractions) and 6 patients stereotactic RT treatment (35 Gy/5 fractions). The CT images used for the treatment planning were selected, which were acquired by two different systems: a Toshiba Aquilion CT-simulator (*Toshiba Healthcare, Tokyo, Japan*) and Phillips Big Bore CT (*Philips Healthcare, Best, the Netherlands*). These CT images were characterised by 1.074 mm in plane resolution and a slice thickness of 3 or 1 mm, respectively to their conventional (87.5%) or stereotactic treatment (12,5%). In this study, the primary and secondary model delineated automatically five different regions of interest (ROIs) on these CT images: prostate (with or without lymph node invasion), bladder, anorectum and both femoral heads. The selected 48 CT images were not included in the training and validation set of the primary model. The secondary model was trained and validated on data from outside the department.

2.3. The auto-segmentation models

2.3.1. Primary auto-segmentation model

In collaboration with RaySearch laboratories (*Stockholm, Sweden*), the Iridium Network developed an in-house trained auto-segmentation model for prostate patients. The training dataset consisted of CT images acquired by the same imaging systems as the patient cohort (see the Toshiba Aquilion CT-simulator and the Phillips Big Bore CT). These CT images were characterised by 1.074 mm in plane resolution and only 3 mm slice thickness. The CT segmentations of the clinical expert were delineated according to ESTRO-ACROP and RTOG guidelines,

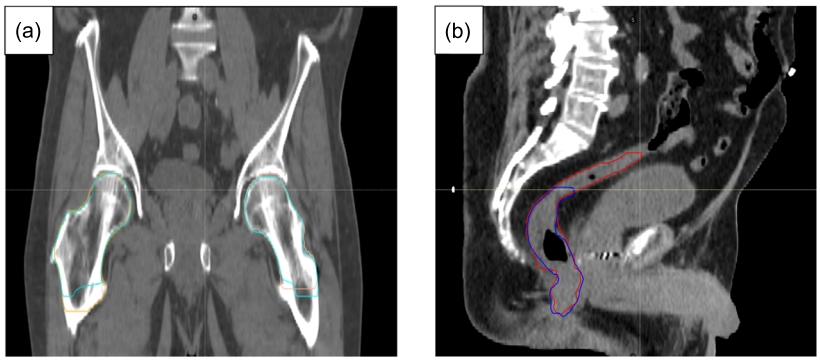

Figure 2. The pre-analysis of the secondary model before metric calculation. (a) For the femoral head structure, differences in caudal segmentation between the primary (orange) and secondary model (blue) can be observed (coronal view). (b) Differences in segmentation of the anorectum by primary (blue) and secondary model (red) in the cranial region (sagittal view).

which were clinically approved, peer-reviewed and used in delivered RT treatment plans. For every ROI, the training, validation and test set consisted of +/−200/25/25 segmented CT images respectively. These were randomly selected featuring significant variability in anatomy. Patients with hip prosthesis were excluded from the training and test set.

2.3.2. Secondary auto-segmentation model

One of the main criteria of this study is the independent character of the secondary model in comparison with the primary one. Independence in this context means that the DL model had not been trained on clinical data of the Iridium Network. Therefore, the auto-segmentation model DLCExpert, constructed by Mirada Medical feq *(Oxford, UK)*, was selected as a secondary algorithm, which is trained on 437 prostate cases contoured according to another institution's internal guidelines. These CT images were characterised by 0.98 mm in plane resolution and 3 mm slice thickness with no-contrast agents used.

2.4. Qualitative analysis

The accuracy of the primary model auto-segmentations was analysed on all 48 CT images by a RO with multiple years of experience in prostate cancer treatment. Initially, the segmentations were categorised into four classes primarily based on a time-saving criterion: class 1 (see starting from scratch), class 2 (see 5–10 min adaptations), class 3 (see small adaptations) and class 4 (see no adaptations needed). This type of quotation arose from the fact that the main objective to use an in-house trained DL model clinically is to reduce time needed to delineate rather than completely eliminate manual intervention. Regarding also the persisting intra-and inter-observer variability, achieving an accuracy comparable to this variability is generally considered as sufficiently accurate. When time-saving is the rationale, determining if and how much manual editing is required to a structure to meet the clinical guidelines is an important result. During the evaluation, a Turing test like setup was installed where the RO quoted the primary auto-segmentation without visualisation of the clinical approved ground truth segmentation (Gooding *et al* 2018). By examining the distribution over the four categories, a large imbalance was observed for every structure. Therefore, class 1 and 2 as well as class 3 and 4 were merged to end with a binary classification problem. Such approach can be justified in the sense of preventing possible overfitting during model training. In this scenario, a proper distinction must be found between segmentations that require minor or major adaptations, where the latter is directly presented to the RO for further revision.

2.5. Quantitative analysis

2.5.1. Pre-analysis

Before quantitative metrics were calculated between the two auto-segmentations, the overall performance of the secondary algorithm was verified on ten, randomly selected test patients in comparison to the primary model (figure 2). In case of the femoral heads, a difference in segmentation level in the caudal section was observed (figure 2(a)). Because the in-house clinical treatment protocol only defines a maximum dose constraint on the femoral head structure and the lower transversal slices usually are not incorporated into the treated Field Of View (FOV), both segmentations were cropped to the same caudal level before metric calculation. This ensured that the metrics reflected differences in the segmentations that matter, rather than those that might be attributed to difference in contouring for training the models. Furthermore, the secondary model segmented the rectum and anus separately and required an additionally merging step to acquire the anorectum structure. Additionally, the primary model delineates this structure up to the sigmoid transition (see ESTRO-ACROP guidelines)

Table 1. Quantitative metrics for comparison of online and offline segmentation.

Type of metric	Metrics used as input features
Overlap metrics	Overlap of the contour volume/surface: (1) Volumetric DSC (2) Mean surface DSC (tolerance: 1 mm, 2 mm, 3 mm)
Distance metrics	Distance between two contours volumes: (1) Hausdorff distance (95th-, 99th-,100th percentile) (2) Mean added path length[a]
Volume	Difference in absolute volume between two contours

[a] The Mean Added Path Length was calculated as metric, because the dataset consisted of CTs with different slice thickness, 3 mm (conventional) and 1 mm (SBRT).

(Salembier *et al* 2018), whereas the secondary model contours the anorectum beyond this level in cranial/anterior direction (figure 2(b)). For prostate and bladder, no consistent differences between both segmentations were observed.

2.5.2. Segmentation evaluation metrics

The quality assessment of contours is usually performed by calculating metrics that quantify the degree of overlap or distance between the manual and the automated segmentation. In this study, the following nine metrics were calculated between both primary and secondary auto-segmentations: the volumetric DSC, the surface DSC at 1 mm, 2 mm, and 3 mm tolerances, the 95th, 99th and 100th percentile Hausdorff Distance (HD), the mean Added Path Length (APL) and the difference in absolute volume. An overview of used metrics is given in table 1.

Considering two different volumes A and B, the volumetric DSC is twice the overlap between these volumes, divided by their sum. A DSC of 1 indicates perfect overlap, while 0 indicates no overlap. The surface DSC is calculated by the same formula as the volumetric DSC, but compares the two external surfaces (Nikolov *et al* 2021). To permit small differences between surfaces without consequences, a tolerance parameter can be used: if points in two surfaces are separated by a distance that is within the tolerance parameter, they are considered part of the intersection of A and B. The HD calculates the minimum distance from every point in surface A to every point in surface B, and vice versa, arranges all distances in ascending order, and returns the maximum distance (100th percentile) or another percentile if so specified (e.g. 95th percentile). The APL is the path length of a contour that had to be added to meet the institutional guidelines for contouring.

In this study, it should be emphasized that neither model was assumed to be superior to the other, and the metrics were calculated to serve as input features for a ML classifier with the aim to detect primary auto-segmentations that need further adaptation. All metrics calculations were made using custom Python scripts that leveraged common scientific libraries.

2.6. Binary classification

Three different ML classifiers architectures were trained for every organ separately: logistic regression, support vector machine (SVM) and random forest (RF). The nine, calculated quantitative metrics were used as input features to predict if the primary auto-segmentation needed minor or major adaptation.

The original dataset of every ROI was divided into a training and test set, where the latter contained a same ratio of classes. In case of the femoral heads, no distinction was made between both sides and were merged in one dataset (= 96 cases). Because imbalanced data was observed between both classes, Synthetic Minority Over-Sampling Technique (SMOTe) was implemented to oversample the minority class. During the training phase, stratified k-fold cross-validation ($k = 3$) was used, and various hyperparameter combinations were exhausted by grid search. In table 2, an overview of the used hyperparameters per model is shown. During every fold, 33% of the dataset was selected as validation set with the preservation of the relative class frequencies before data augmentation of the remaining training set (66%) with SMOTe. This approach avoids the risk of inadvertently testing on data derived from the training set. The mean and standard deviation of the balanced accuracy was recorded during each fold.

After cross-validation, the performance of the overall model on the independent test set was characterised by the classification report, which contain the evaluation metrics recall, precision and $f1$-score. The latter was only executed for the model architecture with highest mean balanced average the cross-validation. In addition, feature importance was also investigated to determine which quantitative metrics had the most clinical value to

Table 2. Overview of the tuned hyperparameters during training of the different ML models.

ML Model	Tuned hyperparameters
Logistic regression	(1) *C*-value: [0.001–1000] (2) Solver: [newton-cg, lbfgs, liblinear, sag, saga] (3) Class weight
Support vector machine	(1) Kernel: linear (*C*-value: [0.001–1000]) (2) Kernel: rbf (*C*-value: [0.001–1000], Gamma: [0.0001, 0.001, 0.01, 0.1, 0.2]) (3) Kernel: poly (*C*-value: [0.001–1000])
Random forest	(1) Bootstrap (2) Max depth: [2, 3, 4, 5] (3) Max features: [2, 3, 4, 5] (4) Min samples leaf: [2, 3, 4, 5] (5) Min samples split: [2, 4, 6] (6) Number of estimators: [10, 20, 40, 60, 80, 100, 200] (7) Oob score

detect deviating segmentations. Scripting of these models was performed in Python, using dedicated ML libraries.

3. Results

3.1. Femoral heads

The balanced accuracy of all ML models trained on femoral head data is shown in table 3(a). In terms of mean balanced accuracy, the RF model performed slightly better in comparison to the other models.

The performance on the independent test set of the overall RF model after cross-validation is shown in the classification report in table 3(b). One can observe that a perfect classification can be made with a recall and precision equal to 1 for both classes. In figure 3(A), the feature importance of the overall RF model is shown. The Surface Dice (all tolerances levels), volumetric Dice and mean APL were considered as the three, most prominent features to distinct femoral head auto-segmentations that need minor or major adaptations.

3.2. Bladder

The balanced accuracy of all ML models trained on bladder data is shown in table 3(c). In terms of mean balanced accuracy, the logistic regression model performed significantly better in comparison to the other models.

The performance on the independent test set of the overall logistic regression model after cross-validation is shown in the classification report in table 3(d). The segmentations that need major adaptation were all detected by the algorithm (see recall score of 1.00), whereas 17% of the segmentations quoted by the RO that needed minor adaptation were wrong assigned to the other class (see recall score of 0.87). In figure 3(B), the mean APL, Hausdorff distance (99th interval) and Surface Dice (tolerance 1 mm) were considered as the three, most prominent features to distinct auto-segmentations that need minor or major adaptations.

3.3. Prostate

The balanced accuracy of all ML models trained on prostate data is shown in table 3(e). In terms of mean balanced accuracy, the logistic regression model performed significantly better in comparison to the other models.

The performance on the independent test set of the overall logistic regression model after cross-validation is shown in the classification report in table 3(f). Despite the fact that all major deviated auto-segmentations were detected by the ML model (see recall score of 1.00), 50% of the minor cases were assigned to the wrong class (see recall of 0.50). In figure 3(C), the Surface Dice (tolerance 1 and 3 mm) and Hausdorff distance (95th interval) and Surface Dice (tolerance 1 mm) were considered as the three major features to distinct auto-segmentations that need minor or major adaptations.

Table 3. (a) ML models balanced accuracy for femoral heads auto-segmentations achieved across three different training folds. (b) Classification report of the overall RF model on the independent test set. (c) ML models balanced accuracy for bladder auto-segmentations achieved across three different training folds. (d) Classification report of the overall logistic regression model on the independent test set. (e) ML models balanced accuracy for prostate auto-segmentations achieved across three different training folds. (f) Classification report of the overall logistic regression model on the independent test set. (g) ML models balanced accuracy for anorectum auto-segmentations achieved across three different training folds. (h) Classification report of the overall logistic regression model on the independent test set.

Femoral heads					
(a)	ML Model architecture	Mean	Standard deviation		
	Logistic regression	0.9639	0.0302		
	Support vector machine	0.9503	0.0467		
	Random forest	0.9653	0.086		
(b)		Precision	Recall	*F*1-score	Number of cases
	Major adaptation	1.00	1.00	1.00	15
	No to minor adaptation	1.00	1.00	1.00	28
Bladder					
(c)	ML Model architecture	Mean	Standard deviation		
	Logistic Regression	0.8501	0.0390		
	Support Vector Machine	0.7591	0.0819		
	Random Forest	0.7835	0.0236		
(d)		Precision	Recall	*F*1-score	Number of cases
	Major adaptation	0.75	1.00	0.86	6
	No to minor adaptation	1.00	0.83	0.91	12
Prostate					
(e)	ML Model architecture	Mean	Standard deviation		
	Logistic Regression	0.5785	0.0258		
	Support Vector Machine	0.4253	0.0667		
	Random Forest	0.4237	0.0568		
(f)		Precision	Recall	*F*1-score	Number of cases
	Major adaptation	0.64	1.00	0.78	7
	No to minor adaptation	1.00	0.50	0.67	8
Anorectum					
(g)	ML Model architecture	Mean	Standard deviation		
	Logistic Regression	0.6240	0.0524		
	Support Vector Machine	0.7197	0.1560		
	Random Forest	0.6943	0.0406		
(h)		Precision	Recall	*F*1-score	Number of cases
	Major adaptation	0.89	1.00	0.94	8
	No to minor adaptation	1.00	0.92	0.96	13

3.4. Anorectum

The balanced accuracy of all ML models trained on anorectum data is shown in table 3(g). In terms of mean balanced accuracy, the SVM model performed better in comparison to the other models.

The performance on the independent test set of the svm model after cross-validation is shown in the classification report in table 3(h). The confusion matrix of the best performed SVM model on the test set is shown in figure 3.4. In this case, no feature importance can be shown, because the overall model is characterised by a non-linear rbf kernel. Although SVMs are often interpreted as transforming your features into a high-dimensional space and fitting a linear classifier in the new space, the transformation is implicit and cannot be easily retrieved. In this scenario, all major deviations were detected (see recall of 1) with 8% of the minor cases assigned to the opposite class (see recall of 0.92).

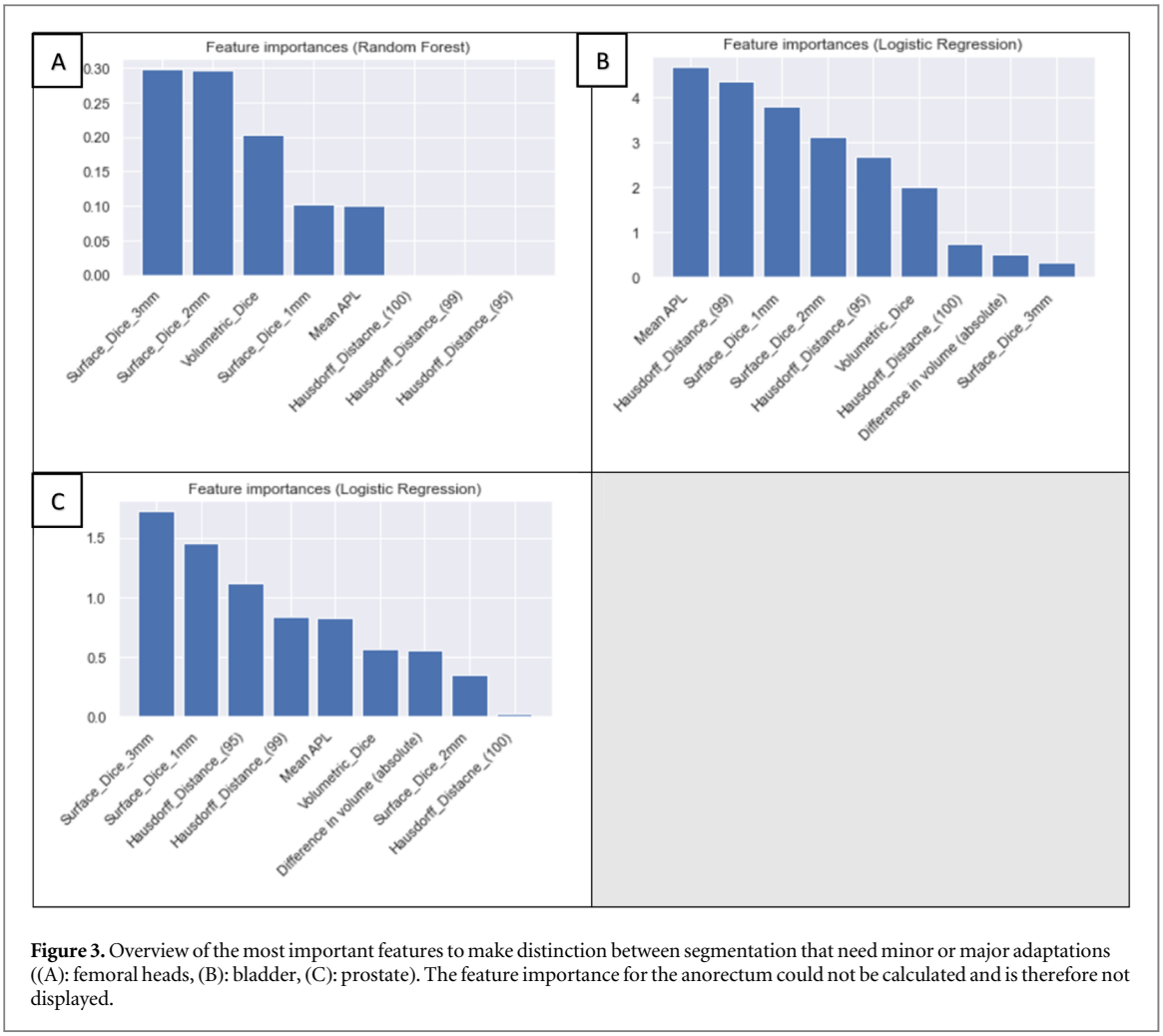

Figure 3. Overview of the most important features to make distinction between segmentation that need minor or major adaptations ((A): femoral heads, (B): bladder, (C): prostate). The feature importance for the anorectum could not be calculated and is therefore not displayed.

4. Discussion

In this study, a QA method is illustrated to automate the verification process of the quality of an in-house trained primary auto-segmentation model for prostate patients. Five ROIs of prostate patients: prostate (target), bladder, anorectum and femoral heads, were delineated simultaneously by the primary model and an independent, secondary model. Subsequently, quantitative metrics were calculated (table 1) and used as input for a ML binary classifier to distinguish between high- and lower-quality primary auto-segmentations, where the latter is prioritized to the RO in order to suggest further adjustments (figure 1). Such automation in the segmentation QA process is necessary to enrol an efficient adaptive radiation therapy (ART) workflow, where the burden of auto-segmentations increases and subsequently the physician's workload (Sonke *et al* 2019).

The idea to implement a secondary DL model in segmentation QA was inspired by the well-standardized usage of independent dose calculation in the treatment planning QA process. The treatment planning system (TPS) as well as the independent dose engine produce a full 3D dose volume which are compared by gamma-analysis. In cases where the passing rate is below the institutional tolerance (e.g. 95% agreement level), the TPS dose distribution is flagged to be verified and if necessary adapted (e.g. reduce the level of plan complexity). In comparison to this study, the level of agreement was expressed by standard segmentation quantitative metrics whereas the tolerance level was defined by a trained ML model whether an auto-segmentation needed minor or major adaptation. In addition, preference was given to interpretable ML systems instead of black boxes to understand the relevance of the different input features to assign differences and to facilitate the later usage of the QA approach in a clinical environment (El Naqa *et al* 2021).

Whereas the clinically primary DL model requires high accuracy to delineate target and OARs, the secondary DL model can have a more generic character and may be less accurate, yet able to detect clinically relevant deviations (see treatment planning process). However, in order to avoid feedback on deviations between both models in non-significant segmentation areas (see outside treatment fields), pre-analysis of both models for a small cohort should be performed (figure 2). That way, specific segmentation deviations in areas of low importance can be detected and cropped out before metric calculation.

The degree of contour similarity between both auto-segmentations was quantified by commonly-used geometry-based metrics such as the DSC or HD (Gooding *et al* 2018). Despite their general use and ease for interpretation, these metrics have a low correlation to the time needed to adjust contours (Vaassen *et al* 2020, Sherer *et al* 2021). Because auto-segmentation techniques are nowadays frequently introduced in clinical practice to reduce contouring time, it is desirable to estimate this time-saving. Ideally, a RO or experienced planner should be able to score the quality of the auto-segmented contour before starting the editing process by visual inspection to determine if they should edit the auto-contour or start from scratch (Vaassen *et al* 2020). Therefore, two recently developed evaluation metrics, namely surface DSC and APL were also included in the calculated metrics. These metrics have been shown to be better indicators for the clinical delineation time saved (Vaassen *et al* 2020, Kiser *et al* 2021). They may provide additional objectively quantifiable surrogates for assessing time-saving and clinical applicability and quality of automatically generated contours in the delineation process (Gooding *et al* 2018). Based on a combination of the calculated metrics, the ML models highlighted all primary auto-segmentations with high accuracy that needed major adaptation for every ROI of prostate patients. Having a system with high sensitivity for anomalies has the possibility to streamline the QA process by directly flagging the low-quality segmentations to the RO for correction. Despite the amount of false negatives for anorectum, bladder and prostate (resp. 8%, 17% and 50%), this only affects the burden of contours needed to be re-checked, which is in this case of lower importance than undetected major deviations. By investigating the feature importance in depth, both the surface DSC (see femoral heads and anorectum) and APL (see bladder) were considered to be most informative. By converting these findings into certain thresholds levels for surface DSC and APL, manual editing could be recommended.

A methodological limitation in this study is the limited number of cases during model training and testing (with exception for femoral heads), where future research should incorporate more examples. Instead of completely re-training the different models to enhance the prediction ability, an open-loop strategy can be obtained to progressively add new auto-segmentations to upgrade the database and preserve better the anatomical variety. By implementing such iterative learning process, subsequent models can be generated that capture the actual clinical practice. Another future point of investigation is the impact of the slice thickness (see 3 and 1 mm) on the performance of the primary DL model. A large imbalance was observed between both groups (87,5% with 3 mm and 12.5% with 1 mm), so that no significant results could be collected.

To conclude, this proposed method includes a proof of concept for the use of an independent DL segmentation model in combination with a ML classifier to improve time saving during QA of auto-segmentations. The integration of this detection system into current automatic segmentation pipelines can increase the efficiency of the RT workflow.

Acknowledgments

Michaël Claessens was supported by a grant of the Flemish League against Cancer, Belgium (ref: 000 019 356)

ORCID iDs

Mark J Gooding https://orcid.org/0000-0002-1177-5608

References

Altman M B *et al* 2015 A framework for automated contour quality assurance in radiation therapy including adaptive techniques *Phys. Med. Biol.* **60** 5199–209

Boldrini L, Bibault J E, Masciocchi C, Shen Y and Bittner M I 2019 Deep learning: a review for the radiation oncologist *Front. Oncol.* **9** 1–9

Borras J M *et al* 2016 How many new cancer patients in Europe will require radiotherapy by 2025? An ESTRO-HERO analysis *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **9** 5–11

Brouwer C L *et al* 2012 3D Variation in delineation of head and neck organs at risk *Radiat. Oncol.* **7** 7–15

Cardenas C E, Yang J, Anderson B M, Court L E and Brock K B 2019 Advances in auto-segmentation *Semin. Radiat. Oncol.* **29** 185–97

Chen H-C *et al* 2015 Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: a general strategy *Med. Phys.* **42** 1048–59

Chen M Y, Woodruff M A, Dasgupta P and Rukin N J 2020a Variability in accuracy of prostate cancer segmentation among radiologists, urologists, and scientists *Cancer Med.* **9** 7172–82

Chen Xinyuan *et al* 2020b CNN-based quality assurance for automatic segmentation of breast cancer in radiotherapy *Front. Oncol.* **10** 524–533

Claessens M and Oria C S 2022 Quality assurance for AI-based applications in radiation therapy *Semin. Radiat. Oncol.* (Special Issue)

Gooding M J *et al* 2018 Comparative evaluation of autocontouring in clinical practice: a practical method using the Turing test *Med. Phys.* **45** 5105–15

Coen R N RMulti-Institutional Target Delineation in Oncology Group *et al* 2011 Human-computer interaction in radiotherapy target volume delineation: a prospective, multi-institutional comparison of user input devices *J. Digit. Imaging* **24** 794–803

El Naqa I *et al* 2021 AI in Medical Physics: Guidelines for Publication *Med. Phys.* **48** 4711–4
Hui C B *et al* 2018 Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach *Med. Phys.* **45** 2089–96
Kiser K J, Barman A, Stieb S, Fuller C D and Giancardo L 2021 Novel autosegmentation spatial similarity metrics capture the time required to correct segmentations better than traditional metrics in a thoracic cavity segmentation workflow *J. Digit. Imaging* **34** 541–53
Kosmin M *et al* 2019 Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **135** 130–40
Lustberg T *et al* 2018 Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **126** 312–7
McIntosh C, Svistoun I and Purdie T G 2013 Groupwise conditional random forests for automatic shape classification and contour quality assessment in radiotherapy planning *IEEE Trans. Med. Imaging* **32** 1043–57
Nikolov S *et al* 2021 Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study *J. Med. Internet Res.* **23** e26151
Salembier C *et al* 2018 ESTRO ACROP consensus guideline on CT- and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **127** 49–61
Sharp G *et al* 2014 Vision 20/20: perspectives on automated image segmentation for radiotherapy *Med. Phys.* **41** 50902
Sherer M V *et al* 2021 Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: a critical review *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **160** 185–91
Sonke J-J, Aznar M and Rasch C 2019 Adaptive radiotherapy for anatomical changes *Semin. Radiat. Oncol.* **29** 245–57
Unkelbach J *et al* 2020 The role of computational methods for automating and improving clinical target volume definition *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **153** 15–25
Vaassen F *et al* 2020 Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy *Phys. Imaging Radiat. Oncol.* **13** 1–6
van Rooij W, Verbakel W F, Slotman B J and Dahele M 2021 Using spatial probability maps to highlight potential inaccuracies in deep learning-based contours: facilitating online adaptive radiation therapy *Adv. Radiat. Oncol.* **6** 100658
van der Veen J, Gulyban A and Nuyts S 2019 Interobserver variability in delineation of target volumes in head and neck cancer *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **137** 9–15
Vinod S K, Jameson M G, Min M and Holloway L C 2016 Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **121** 169–79