Strategies for tackling the class imbalance problem of oropharyngeal primary tumor segmentation on magnetic resonance imaging

Outeiral, R.R.; Bos, P.; Hulst, H.J. van der; Al-Mamgani, A.; Jasperse, B.; Simoes, R.; Heide, U.A. van der

Physics and Imaging in Radiation Oncology 23 (2022) 144–149

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Strategies for tackling the class imbalance problem of oropharyngeal primary tumor segmentation on magnetic resonance imaging

Roque Rodríguez Outeiral [a,*], Paula Bos [b,c], Hedda J. van der Hulst [b], Abrahim Al-Mamgani [a], Bas Jasperse [b], Rita Simões [a], Uulke A. van der Heide [a]

[a] *Department of Radiation Oncology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands*
[b] *Department of Radiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands*
[c] *Department of Head and Neck Oncology and Surgery, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands*

ARTICLE INFO

Keywords:
Oropharyngeal cancer
Convolutional neural network
Segmentation
Class imbalance, MRI
Two-stage approach

ABSTRACT

Background and purpose: Contouring oropharyngeal primary tumors in radiotherapy is currently done manually which is time-consuming. Autocontouring techniques based on deep learning methods are a desirable alternative, but these methods can render suboptimal results when the structure to segment is considerably smaller than the rest of the image. The purpose of this work was to investigate different strategies to tackle the class imbalance problem in this tumor site.
Materials and methods: A cohort of 230 oropharyngeal cancer patients treated between 2010 and 2018 was retrospectively collected. The following magnetic resonance imaging (MRI) sequences were available: T1-weighted, T2-weighted, 3D T1-weighted after gadolinium injection. Two strategies to tackle the class imbalance problem were studied: training with different loss functions (namely: Dice loss, Generalized Dice loss, Focal Tversky loss and Unified Focal loss) and implementing a two-stage approach (i.e. splitting the task in detection and segmentation). Segmentation performance was measured with Sørensen–Dice coefficient (Dice), 95th Hausdorff distance (HD) and Mean Surface Distance (MSD).
Results: The network trained with the Generalized Dice Loss yielded a median Dice of 0.54, median 95th HD of 10.6 mm and median MSD of 2.4 mm but no significant differences were observed among the different loss functions (p-value > 0.7). The two-stage approach resulted in a median Dice of 0.64, median HD of 8.7 mm and median MSD of 2.1 mm, significantly outperforming the end-to-end 3D U-Net (p-value < 0.05).
Conclusion: No significant differences were observed when training with different loss functions. The two-stage approach outperformed the end-to-end 3D U-Net.

1. Introduction

Radiotherapy is one of the common treatment options for head and neck cancer patients [1,2]. One key step of the radiotherapy workflow is tumor contouring. While contouring of organs at risk is increasingly being automated in clinical practice, tumor contouring is still done manually. This is time consuming and suffers from high interobserver variability [3].

Deep learning methods, particularly Convolutional Neural Networks (CNNs), are the current state of the art for automatic segmentation of medical images. Several review papers have been published on deep learning applied to radiotherapy and automatic segmentation is often discussed as one of the main applications [4–7]. For the particular case of head and neck cancer, various works have focused on the automatic segmentation of organs at risk with deep learning [8], some of them achieving clinically acceptable performance and being commercially available [9]. For the case of tumor contouring, the literature is more scarce and those algorithms are still not implemented in the clinic.

In our previous work [10], we segmented the oropharyngeal primary tumor on magnetic resonance imaging (MRI) and showed that combining multiple anatomical MRI sequences improved the segmentation performance compared to single-sequence. We also proposed a semi-automatic approach that improved the segmentation performance by splitting the segmentation task in manual detection and segmentation. To the best of our knowledge, there is only one other work where the authors segmented the oropharyngeal primary tumor on MRI [11].

[*] Corresponding author.
E-mail addresses: r.rodriguez.outeiral@nki.nl (R. Rodríguez Outeiral), u.vd.heide@nki.nl (U.A. van der Heide).

https://doi.org/10.1016/j.phro.2022.08.005
Received 15 November 2021; Received in revised form 29 July 2022; Accepted 10 August 2022
Available online 13 August 2022

The authors studied the impact of combining different anatomical (T1 weighted and T2 weighted) and quantitative images (ADC, K^{trans} and v_e) as input channels to a CNN and showed that combining anatomical sequences significantly improved the performance.

A known issue in the field of deep learning for medical image segmentation is class imbalance, meaning that the structure to be segmented is present in a smaller amount of voxels compared to the rest of the image. Class imbalance can result in suboptimal solutions because the network is exposed to proportionally less relevant information during the training process. Several works in the field of medical image segmentation have focused on this problem, either by modifying the input data to the network [12,13] or by defining different loss functions [14–16]. This problem is even more critical in the case of tumor segmentation, given that tumors tend to be smaller than other structures and they are heterogeneous in their location, shape and size. This is also the case for the oropharyngeal primary tumor.

Several loss functions have been designed with the aim of tackling class imbalance, such as the Generalized Dice loss [17], the Focal loss [14], the (focal) Tversky loss [15,18] and the Unified Focal loss [16]. Although the choice of the loss function can be critical for the training of a CNN, comprehensive loss function comparisons for specific tumor sites or anatomies are not commonly performed. Ma et al. [19] showed that the influence in performance of the loss function varies greatly depending on the segmentation task. To the best of our knowledge, this has not been studied yet in the particular case of oropharyngeal cancer segmentation.

Other works have implemented two-stage approaches (i.e. detection and segmentation) that resulted in more accurate segmentations than their one-stage counterparts [20–22]. By locating the tumor first, the context around the tumor is reduced. Consequently, two-stage approaches are a possible way of tackling class imbalance. The semi-automatic approach from our previous work [10] consisted of having human observers outlining a box around the tumor to provide a first approximation of the tumor location and consequently ease the segmentation task. However, the semi-automatic approach still needed manual intervention. The implementation of a two-stage approach will also allow us to fully automate the semi-automatic approach proposed in our previous work [10].

The aim of this study was to investigate two different strategies for tackling the class imbalance problem for oropharyngeal primary tumor segmentation: training with different loss functions and implementing a fully automatic two-stage approach.

2. Materials and methods

2.1. Data

A cohort of 230 patients treated at our institute between January 2010 and May 2018 was used for this project. The mean age of the patients was 61 years (standard deviation ± 7 years) and 66% of the patients were male. Further details on tumor stage and HPV status can be found in the Supplemental Material (Table S.1). All patients had histologically proven primary oropharyngeal squamous cell carcinoma and received a pre-treatment MRI for primary staging. The institutional review board approved the study (IRBd18047). Informed consent was waived by the institutional review board considering the retrospective design. The cohort was extended from our previous work [10]. A total of 59 new patients were included.

The scans were acquired on 1.5 T ($n = 108$) or 3.0 T ($n = 122$) MRI scanners (Philips Medical System, Best, The Netherlands). The imaging protocol included: 2D T1-weighted fast spin-echo, 2D T2-weighted fast spin-echo with fat suppression, 3D T1-weighted high-resolution isotropic volume excitation after gadolinium injection with fat suppression. Further details on the MRI protocols are given in the Supplemental Material (Table S.2). The primary tumors were manually contoured in 3D Slicer (version 4.8.0, https://www.slicer.org/) by one observer with 1 year of experience (P.B. or H.H.). Afterwards, they were reviewed and adjusted, if needed, by a radiologist with 7 years of experience (B.J.). All tumor volumes were delineated on the 3D sequence but the observers were allowed to consult the other sequences.

For the experimental set-up, the data set was split in three subsets: a training set ($n = 190$), a validation set ($n = 20$) and a test set ($n = 20$). The test set was not used for training or hyper-parameter tuning. We stratified the three subsets for tumor volume, subsite, and aspect ratio since these features are likely relevant for segmentation. Subsites were defined as tonsillar tissue, soft palate, base of tongue and posterior wall. The aspect ratio was defined as the ratio between the shortest and the longest axis of the tumor. All images were resampled to a voxel size of 0.8 mm × 0.8 mm × 0.8 mm.

2.2. Baseline model architecture

The 3D U-Net architecture [23,24] was used as the basis for our experiments. The Adam optimizer [25] and early stopping were used for training. Dropout and data augmentation were used for regularization. Further details on the training procedure can be found in Table S.4 and in the code which is publicly available in: https://github.com/RoqueRouteiral/oroph_segm_ts.

2.3. Training with different loss functions

We trained the 3D U-Net with four different loss functions: Dice loss [26], Generalized Dice loss [17], Focal Tversky loss [18] and Unified Focal loss [16]. For the particular case of the Unified Focal loss, Yeung et al. [16] showed that the choice of the γ hyperparameter can affect the performance. Consequently, we trained four networks with the Unified Focal loss for different values of its hyperparameter γ ($\gamma = [0.2, 0.4, 0.6, 0.8]$). We compared the segmentation performance of all the networks among each other.

2.4. Two-stage approach

In our previous work, we demonstrated that the segmentation of the oropharyngeal primary tumor was more accurate when the input image was manually cropped with a clipbox around the tumor before being fed to a segmentation network.

In this work, we fully automated this two-stage approach (Fig. 1). The first stage consisted of roughly detecting the tumor by automatically selecting a clipbox around it. In the second stage, this clipbox was used to crop the image which was then used as input to a segmentation network. The loss function chosen for both stages was the Generalized Dice loss function. The loss was backpropagated through each network separately.

For the detection stage, a 3D U-Net was trained using the bounding box of the tumor as ground truth. At inference time, the output of the detection was computed as the bounding box of the output.

For the segmentation stage, the same architecture as in our previous work was used [10]. This segmentation network was trained with only the information contained inside the clipboxes. In every training iteration, the clipboxes were randomly shifted by an amount of up to 25 mm in different directions to make the network robust to possible displacements in the detection. At inference time the input images were cropped by the clipboxes defined by the detection network. Similarly to our previous work, the clipboxes were dilated by 5 mm.

2.5. Statistics

To confirm that the three subsets were balanced in subsite, volume and aspect ratio, a Kruskal-Wallis test was used for continuous variables (volume and aspect ratio) and a chi-square test for independence for the categorical data (subsite).

Predicted segmentations and the segmentations from the human

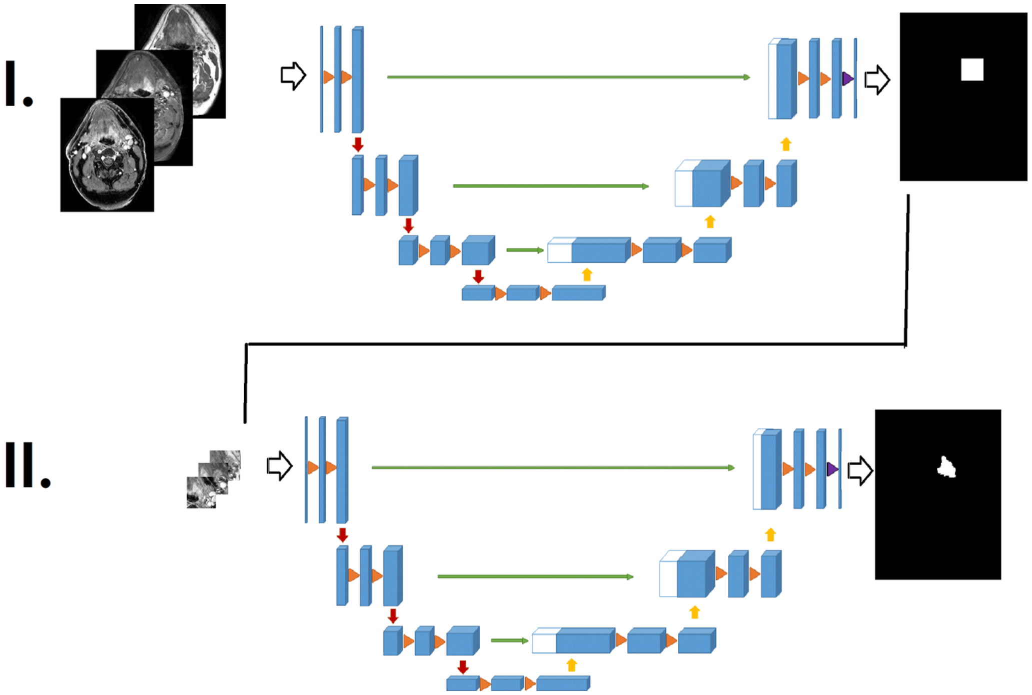

Fig. 1. Overview of the two-stage approach.

observers were compared for the patients on the unseen test set. Common segmentation metrics were used: Sørensen–Dice coefficient (Dice), 95th Hausdorff Distance (HD) and Mean surface distance (MSD). The metrics were implemented using the Python package from DeepMind (https://github.com/deepmind/surface-distance). For the two-stage approach, the detection was evaluated by measuring the absolute mean shift in all 6 directions between the tumor bounding box and the detected clipbox for the patients on the unseen test set. The average shift of the boxes for the observers from our previous work was used for comparison [10]. Differences among the loss function experiments were assessed by the Friedman test whereas the two-stage approach experiments were assessed by the Wilcoxon signed-ranked test. P-values below 0.05 were considered statistically significant.

All networks were retrained four times. Reported results are the mean of the results of the four versions of each network. We opted for this approach over *N*-fold cross-validation to account for the random initialization of the network while ensuring proper stratification in the three sets for all the folds.

3. Results

3.1. Summary of tumor characteristics

Table S.3 shows the tumor characteristics (location, volume and aspect ratio) of our cohort. No significant differences were found in the distributions of subsite, volume and aspect ratio among the training, validation and test sets.

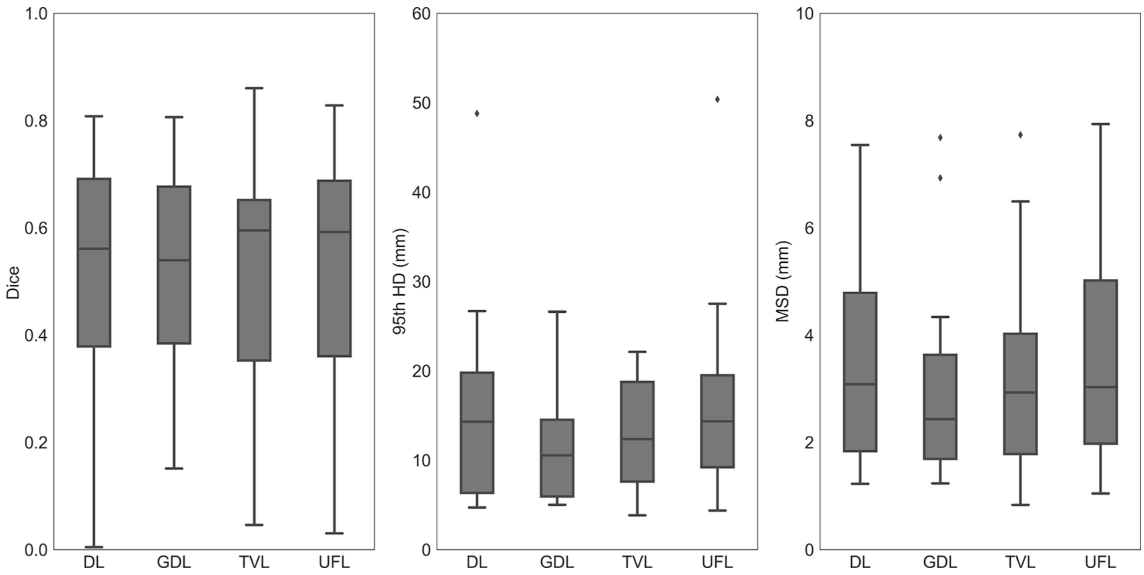

Fig. 2. Segmentation performance of the 3D U-Net trained with different loss functions: Dice Loss (DL), Generalized Dice Loss (GDL), Tversky Loss (TVL) and Unified Focal Loss (UFL).

3.2. Training with different loss functions

When comparing the performance of the networks trained with different loss functions no significant differences were found (p-value > 0.25 for the three metrics). Lower variance in the MSD and Dice can be observed for the network trained with the Generalized Dice loss (Fig. 2). The network achieved a median Dice of 0.54, median 95th HD of 10.6 mm and median MSD of 2.4 mm. Non-significant differences were observed when training the network with different γ values for the Unified Focal loss (Fig. S.1).

3.3. Two-stage approach

The mean shift for the detection network was of 8.9 mm (Table 1) and no significant differences were found when comparing to the detection of observer 2 from our previous work (p-value = 0.40). Significant differences were found when comparing the detection of this work to the detection of the observer 1 from out previous work (p-value < 0.001). When separating the mean shift per direction, we observed a mean shift of 10.0 mm in the cranial-caudal direction, 8.4 mm in the medial–lateral direction and 7.7 mm in the dorsal–ventral direction.

The segmentation results of the two-stage approach were significantly better for Dice (p-value = 0.03) and MSD (p-value = 0.02) than the results of the end-to-end 3D UNet (Table 1). The fully automated two-stage approach yielded a median Dice of 0.64, median HD of 8.7 mm and median MSD of 2.1 mm. One patient was missed in the detection of the two-stage approach for one of the folds, and thus removed from that fold for the analysis.

3.4. Qualitative results

Examples of segmentations obtained by the end-to-end 3D U-Net, the two-stage approach and ground truth segmentation are shown in Fig. 3. The end-to-end 3D U-Net approach oversegmented (Fig. 3a–c) the tumor, where the two-stage approach showed better segmentation comparison to the ground truth. Fig. 3b shows cases where the segmentation end-to-end 3D U-Net rendered additional false positive structures on the image.

4. Discussion

This work investigated two different strategies to tackle the class imbalance problem for the task of oropharyngeal primary tumor segmentation: training with the different loss functions and implementing a two-stage approach. Additionally, the proposed two-stage approach fully automated the semi-automatic approach described in our previous work [10].

When training the networks with different loss functions, no significant improvements were observed in the segmentation metrics. Hyperparameter tuning for the γ hyperparameter of the Unified Focal loss did not yield significantly better results either. This result is consistent with the work of Ma et al. [19], where they concluded that Dice-related losses are often optimal for medical image segmentation tasks. Additionally, it is also in line with the conclusions described by Isensee et al. and their proposed "no new Net" (nnU-Net) [27]. They showed that a tailored-to-task method configuration is more relevant than specific setup choices when designing a segmentation deep learning pipeline.

The two-stage approach achieved significantly better results compared to the conventional end-to-end approach. The high complexity of the task may make the end-to-end training of the network suboptimal, while focusing on two simpler tasks can render better results. In our previous work [10], a semi-automatic approach in which an observer selected a clipbox around the tumor was implemented. When comparing the current detection results to the semi-automatic approach of our previous work, we noted that one of the observers (Obs. 1) selected a tighter box (although all the tumors were included inside the clipboxes) compared to that of our two-stage approach which resulted in significantly different detection performance. However, we did not observe significant differences with the detection performance of the semi-automatic approach for the other observer (Obs. 2), showing that a fully automatic two-stage approach can be a feasible alternative to a semi-automatic approach. Also, the time spent on delineating in the clinical practice is aimed to be as low as possible. We reported in our previous work that the time spent on drawing the boxes was lower for observer 2 than for observer 1, making the delineations of observer 2 a more realistic representation of what is expected in the clinic. In the present work, the whole pipeline is automated, which can save time in the clinic. That said, further efforts in improving the detection are of interest to improve the segmentation performance of the two-stage approach.

The literature on automatic segmentation for the oropharyngeal tumor on MRI is scarce and its aims are heterogeneous. Besides our previous work [10], only Wahid et al. [11] have focused on the segmentation of this tumor site on MRI. Their work focused on studying the value of multiparametric MRI on the segmentation performance, both for qualitative and quantitative imaging. Other works focused on the automatic segmentation on multiparametric MRI of the head and neck cancer in general, rather than on the particular subset of oropharyngeal cancer: Bielak et al. [28] used diffusion weighted imaging while Schouten et al. [29] proposed a multiview CNN architecture. To the best of our knowledge, only our work is focused on tackling the class imbalance problem for head and neck cancer segmentation on MRI, and particularly for the oropharyngeal subsite.

In 2020, the first head and neck tumor segmentation challenge, known as HECKTOR challenge, was launched [30]. The main subsite of the challenge was the oropharyngeal tumor and the winner of the challenge achieved a mean Dice of 0.76, but the image modalities used were PET/CT. Additionally, Ren et al. [31] compared the use of PET/CT/MRI as different input image combinations for the automatic segmentation of head and neck GTV and observed that, when including PET, the segmentation performance improved. Considering all the above, it is possible that PET is a useful modality for the task of head and neck tumor segmentation. However, the differences in resolution between imaging modalities may be reflected in the detail of the manual ground truth delineations used for training and evaluation. Potentially, this can also explain the difference in performance of the MRI-based task. That said, we argue that the strategies to tackle class imbalance in this work can be useful in the development of autocontouring tools for the case of oropharyngeal cancer.

This study has limitations. Firstly, there is a high interobserver variability on this tumor subsite, especially in case of tonsillar fossa and base of tongue tumor which are rich in lymphatic tissue, so it is possible that the ground truth delineations used in this work are partially biased. However, one observer corrected the other's delineation, reducing this observer variation. Secondly, validation of our results is still needed

Table 1
Detection and segmentation performance of the two stage approach and comparison to results of the previous work [10].

	Detection	Segmentation		
	Avg. shift (mm) – [SD]	Dice	HD (mm)	MSD (mm)
This work				
3D end-to-end UNet	–	0.54	10.6	2.4
Two stage approach	8.7 [8.2]	0.64	8.7	2.1
Previous work				
Semi-automatic approach (Obs. 1)	3.0 [3.9]	0.74	4.6	1.2
Semi-automatic approach (Obs. 2)	8.9 [6.9]	0.67	7.2	1.7

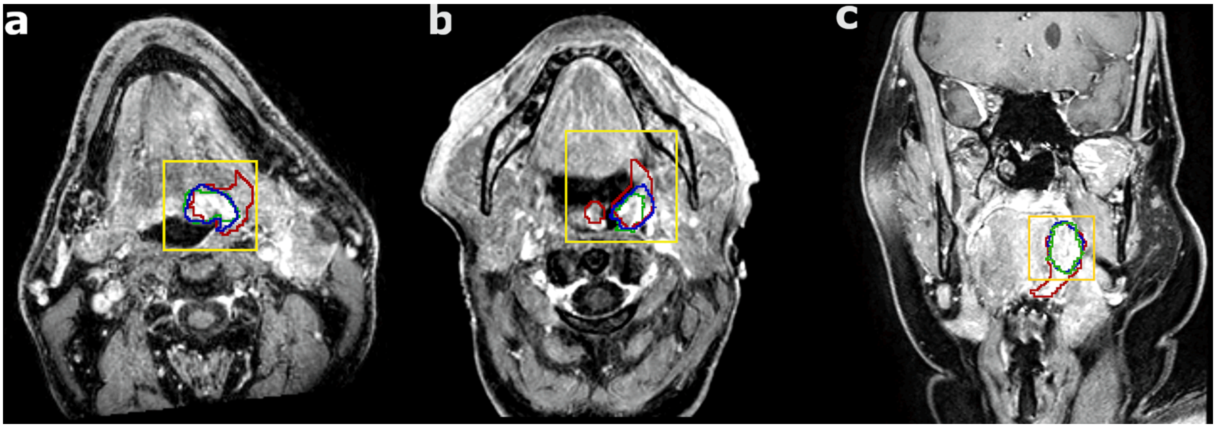

Fig. 3. Comparison of the oropharyngeal segmentations in three different patients (a, b, c) trained with the end-to-end 3D U-Net (red), with the two-stage approach (blue) and the manual delineation (green). The yellow boxes are drawn by detection network from the two-stage approach. All the images correspond to the 3D sequence.

with an independent cohort in a multi-center study. Thirdly, the performance could also be improved by making different decisions on the training setup, such as using larger batch sizes or non downsampled data, but other strategies to mitigate memory limitations would be needed. Finally, there is a certain variability in the scan protocols. However, variability in the training set can be desirable as it makes the network robust to protocol differences.

In conclusion, the loss functions designed to tackle class imbalance performed comparably among each other. The approach of splitting the problem into localization and segmentation outperformed the end-to-end network, proving an effective strategy to mitigate the class imbalance problem in oropharyngeal cancer segmentation.

Data statement

The data that has been used in this study is confidential. The institutional review board approved the study (IRBd18047). Informed consent was waived considering the retrospective design.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2022.08.005.

References

[1] Delaney G, Jacob S, Barton M. Estimation of an optimal external beam radiotherapy utilization rate for head and neck carcinoma. Cancer 2005;103:2218. https://doi.org/10.1002/cncr.21084.

[2] Barton MB, Jacob S, Shafiq J, Wong K, Thompson SR, Hanna TP, et al. Estimating the demand for radiotherapy from the evidence: a review of changes from 2003 to 2012. Radiother Oncol 2014;112:140–4. https://doi.org/10.1016/j.radonc.2014.03.024.

[3] Blinde S, Mohamed ASR, Al-Mamgani A, Newbold K, Karam I, Robbins JR, et al. Large interobserver variation in the international MR-LINAC oropharyngeal carcinoma delineation study. Int J Radiat Oncol Biol Phys 2017;99:639–40.

[4] Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. Med Phys 2019;46:1–36. https://doi.org/10.1002/mp.13264.

[5] Brouwer CL, Dinkla AM, Vandewinckele L, Crijns W, Claessens M, Verellen D, et al. Machine learning applications in radiation oncology: current use and needs to support clinical implementation. Phys Imaging Radiat Oncol 2020;16:144–8. https://doi.org/10.1016/j.phro.2020.11.002.

[6] Boldrini L, Bibault JE, Masciocchi C, Shen Y, Bittner MI. Deep learning: a review for the radiation oncologist. Front Oncol 2019;9:977. https://doi.org/10.3389/fonc.2019.00977.

[7] Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. Comput Biol Med 2018;98:126–46. https://doi.org/10.1016/j.compbiomed.2018.05.018.

[8] Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods. Med Phys 2020;47:929–50. https://doi.org/10.1002/mp.14320.

[9] Brunenberg EJL, Steinseifer IK, van den Bosch S, Kaanders JHAM, Brouwer CL, Gooding MJ, et al. External validation of deep learning-based contouring of head and neck organs at risk. Phys Imaging Radiat Oncol 2020;15:8–15. https://doi.org/10.1016/j.phro.2020.06.006.

[10] Rodríguez Outeiral R, Bos P, Al-Mamgani A, Jasperse B, Simões R, van der Heide UA. Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning. Phys Imaging Radiat Oncol 2021;19:39–44. https://doi.org/10.1016/j.phro.2021.06.005.

[11] Wahid KA, Ahmed S, He R, van Dijk LV, Teuwen J, McDonald BA, et al. Evaluation of deep learning-based multiparametric MRI oropharyngeal primary tumor auto-segmentation and investigation of input channel effects: results from a prospective imaging registry. Clin Transl Radiat Oncol 2022;32:6–14. https://doi.org/10.1016/j.ctro.2021.10.003.

[12] Small H, Ventura J. Handling Unbalanced Data in Deep Image Segmentation. [https://svds.com/learning-imbalanced-classes/]; 2017.

[13] Kochkarev A, Khvostikov A, Korshunov D, Krylov A, Boguslavskiy M. Data balancing method for training segmentation neural networks. CEUR Workshop Proc 2020:27441–9. https://doi.org/10.51130/graphicon-2020-2-4-19.

[14] Lin T-Y. Focal Loss for Dense Object Detection (RetinaNet). 13C-NMR Nat Prod 2017:30–3. https://doi.org/10.1109/ICCV.2017.324.

[15] Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. Lect Notes Comput Sci. 2017;10541 LNCS:379–87. doi: 10.1007/978-3-319-67389-9_44.

[16] Yeung M, Sala E, Schönlieb CB, Rundo L. Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. 95. Comput Med Imaging Graph 2022. https://doi.org/10.1016/j.compmedimag.2021.102026.

[17] Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge CM. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. Lect Notes Comput Sci 2017:10553. https://doi.org/10.1007/978-3-319-67558-9_28.

[18] Abraham N, Khan NM. A novel focal tversky loss function with improved attention u-net for lesion segmentation. Proc - Int Symp Biomed Imaging 2019:683–7. https://doi.org/10.1109/isbi.2019.8759329.

[19] Ma J, Chen J, Ng M, Huang R, Li Y, Li C, et al. Loss odyssey in medical image segmentation. Med Image Anal 2021;71:102035. https://doi.org/10.1016/j.media.2021.102035.

[20] Feng X, Qing K, Tustison NJ, Meyer CH, Chen Q. Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images. Med Phys 2019;46:2169–80. https://doi.org/10.1002/mp.13466.

[21] Balagopal A, Kazemifar S, Nguyen D, Lin MH, Hannan R, Owrangi A, et al. Fully automated organ segmentation in male pelvic CT images. Phys Med Biol 2018;63:245015. https://doi.org/10.1088/1361-6560/aaf11c.

[22] Wang Y, Zhao L, Wang M, Song Z. Organ at risk segmentation in head and neck CT images using a two-stage segmentation framework based on 3D U-Net. IEEE Access 2019;7:144591–602. https://doi.org/10.1109/access.2019.2944958.

[23] Ronneberger O, Fischer P, U-net BT. Convolutional networks for biomedical image segmentation. In: Lecture notes in computer science; 2015. p. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28.

[24] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation BT– MICCAI 2016. 424–32.

[25] Kingma DP, Ba JL. Adam: A method for stochastic optimization. ICLR 2015 - Conference Track Proceedings. 2015.

[26] Milletari F, Navab N, Ahmadi S-A. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Fourth international conference on 3D vision (3DV). IEEE; 2016. p. 565–71. https://doi.org/10.1109/3dv.2016.79.

[27] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11. https://doi.org/10.1038/s41592-020-01008-z.

[28] Bielak L, Wiedenmann N, Nicolay NH, Lottner T, Fischer J, Bunea H, et al. Automatic tumor segmentation with a convolutional neural network in multiparametric mri: Influence of distortion correction. Tomography 2019;5:292–9. https://doi.org/10.18383/j.tom.2019.00010.

[29] Schouten JPE, Noteboom S, Martens RM, Mes SW, Leemans CR, de Graaf P, et al. Automatic segmentation of head and neck primary tumors on MRI using a multi-view CNN. Cancer Imaging 2022:1–9. https://doi.org/10.1186/s40644-022-00445-7.

[30] Andrearczyk V, Oreiller V, Jreige M, Vallières M, Castelli J, Elhalawani H, et al. Overview of the HECKTOR Challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT. 2021;12603 LNCS:1–21. doi: 10.1007/978-3-030-67194-5_1.

[31] Ren J, Eriksen JG, Nijkamp J, Korreman SS. Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation. Acta Oncol 2021;60:1399–406. https://doi.org/10.1080/0284186x.2021.1949034.