Deep learning aided oropharyngeal cancer segmentation with adaptive thresholding for predicted tumor probability in FDG PET and CT images

Alessia De Biase¹,², Nanna M Sijtsema¹, Lisanne V van Dijk¹,³, Johannes A Langendijk¹ and Peter M A van Ooijen¹,²

¹ Department of Radiation Oncology, University Medical Center Groningen, Groningen, 9700RB, The Netherlands
² Data Science Center in Health (DASH), University Medical Center Groningen, Groningen, 9700RB, The Netherlands
³ Department of Radiation Oncology, University of Texas MD Anderson Cancer Center, Houston, TX-77030, Texas, United States of America

E-mail: a.de.biase@umcg.nl

Keywords: adaptive segmentation, deep learning, head and neck cancer

Abstract

Objective. Tumor segmentation is a fundamental step for radiotherapy treatment planning. To define an accurate segmentation of the primary tumor (GTVp) of oropharyngeal cancer patients (OPC) each image volume is explored slice-by-slice from different orientations on different image modalities. However, the manual fixed boundary of segmentation neglects the spatial uncertainty known to occur in tumor delineation. This study proposes a novel deep learning-based method that generates probability maps which capture the model uncertainty in the segmentation task.

Approach. We included 138 OPC patients treated with (chemo)radiation in our institute. Sequences of 3 consecutive 2D slices of concatenated FDG-PET/CT images and GTVp contours were used as input. Our framework exploits inter and intra-slice context using attention mechanisms and bi-directional long short term memory (Bi-LSTM). Each slice resulted in three predictions that were averaged. A 3-fold cross validation was performed on sequences extracted from the axial, sagittal, and coronal plane. 3D volumes were reconstructed and single- and multi-view ensembling were performed to obtain final results. The output is a tumor probability map determined by averaging multiple predictions.

Main Results. Model performance was assessed on 25 patients at different probability thresholds. Predictions were the closest to the GTVp at a threshold of 0.9 (mean surface DSC of 0.81, median HD₉₅ of 3.906 mm). Significance. The promising results of the proposed method show that it is possible to offer the probability maps to radiation oncologists to guide them in a slice-by-slice adaptive GTVp segmentation.

1. Introduction

Image-guided deep learning (DL) tools have shown enormous potential in the field of radiation oncology. Some of the main goals are to speed up the radiotherapy workflow, to automate error prone tasks, and to reduce inter-observer variability. In recent years, analysis of the variability of target volume definitions across different and same observers, documented as inter- and intra-observer variability, have been published (Weiss and Hess 2003, Sadeghi et al 2021). Several studies have investigated the extent of inter-observer variability in target volume delineation of head and neck cancer (HNC), showing large discrepancies in the contoured structures (median DSC and HD₉₅ for the clinical target volume of the primary tumor (CTVp) ranged between 0.51 and 0.79 and 6.7 mm and 16.1 mm respectively) (van der Veen et al 2019). With the word tumor we refer to the macroscopic disease that can be visually identified on imaging, so called gross tumor volume (GTV). The clinical target volume (CTV) used for radiotherapy treatment planning is generated adding a margin to the GTV in order to include the microscopic tumor extensions. Tumor contouring variation between clinical specialists directly
influences the manual contours (Sadeghi et al 2021), which are used as ground truth during the training process of DL auto-segmentation models, resulting in a lower accuracy and general acceptance of the auto-segmented contours. In summary, because of the absence of a proper reference representing the absolute gold standard, the one-contour output provided by such models would not represent the most ideal contour for all clinicians.

Different strategies to improve tumor delineation and reduce inter-observer variability have been proposed. Some examples are consensus guidelines, reliable ground truth, discuss the contours in a multidisciplinary meeting or the use of proper and complementary imaging modalities (Gudi et al 2017, Sadeghi et al 2021). The use of FDG-PET/CT for GTV delineation, for example, showed to significantly reduce inter-observer variability for head and neck cancers compared to using CT only (Gudi et al 2017). Despite the above mentioned solutions may improve tumor delineation, they are not always the most efficient. Adding more image modalities, in fact, is expected to improve the contouring accuracy (Rodríguez Outeiral et al 2021), but makes the contouring process more time consuming and more complex (Anderson et al 2014). Thus, automatic segmentation of organs at risk (OARs) and clinical target volumes (CTVs) is of great interest, especially in cases where various imaging modalities are needed for radiotherapy treatment planning.

For the head and neck region, 3D deep learning segmentation models on Positron Emission Tomography (PET) and Computed Tomography (CT) images were designed and tested during the first and second edition of the HECKTOR challenge (Andrearczyk et al 2021a, 2022a). The obtained segmentation performance, in terms of Dice Similarity Coefficient (DSC) and 95th Hausdorff Distance (HD95), demonstrated the potential of deep learning methods in the tumor segmentation task. However, automatic tumor segmentation has demonstrated to achieve lower performance compared to automatic OARs delineation, since tumor volumes highly depend on the clinical scenario and clinical judgement of the treating physician (Wong et al 2020). Tumor segmentation, in fact, is more challenging, since tumors have variable presentation, i.e. they are different in location, size, and shape (see figure 1) (Diao et al 2022).

For the head and neck region, 3D deep learning segmentation models on Positron Emission Tomography (PET) and Computed Tomography (CT) images were designed and tested during the first and second edition of the HECKTOR challenge (Andrearczyk et al 2021a, 2022a). The obtained segmentation performance, in terms of Dice Similarity Coefficient (DSC) and 95th Hausdorff Distance (HD95), demonstrated the potential of deep learning methods in the tumor segmentation task. However, automatic tumor segmentation has demonstrated to achieve lower performance compared to automatic OARs delineation, since tumor volumes highly depend on the clinical scenario and clinical judgement of the treating physician (Wong et al 2020). Tumor segmentation, in fact, is more challenging, since tumors have variable presentation, i.e. they are different in location, size, and shape (see figure 1) (Diao et al 2022).

For the introduction in the clinic of DL auto-segmentation models and to gain the trust of the radiation oncologists in those methods, it would be beneficial to provide them with information about how reliable segmentations are and how confident the network is in predicting results (Wang et al 2019, Diao et al 2022). Uncertainty estimation of segmentation results, in fact, gained great importance in recent years. In medical image segmentation, the most common techniques to estimate uncertainty are non-Bayesian methods, which make use of multiple results obtained as output in the prediction phase. The different predictions can then be

Figure 1. Examples of manually performed organs at risk (OARs) segmentation (b), (f) and gross tumor volume of the primary tumor (GTVp) segmentation (d), (h) of two oropharyngeal cancer patients (first and second row). The delineation of the OARs (e.g. left (L) and right (R) parotid and submandibular glands in the figure) and the GTVp are two fundamental steps of radiotherapy treatment planning. OARs segmentation is usually performed on the CT scan only (a), (e). Multiple imaging modalities such as PET/CT (c), (g) and MRI are used to identify the GTVp in head and neck cancer patients. In the figure, the contoured OARs show little difference between patients (b), (f). On the contrary, the GTVp in (d) has different location, shape, and size of the GTVp in (h).
combined (so-called ensembling) or directly used for uncertainty measures calculation (predictive variance, predictive entropy). In segmentation tasks, uncertainty maps are the most commonly used uncertainty visualization tools. They are represented by pixel- or voxel-wise values obtained by uncertainty measures. The output of a segmentation model, in these cases, is twofold: a segmentation output and an uncertainty map (see figure 2(b)). The motivation for such an output in clinic is to speed up the review process drawing attention only to the segmented regions with a large uncertainty (Nair et al. 2020). However, the additional provided information is purely informative and does not provide a suggestion for alternative contouring options to the end user. We propose in this study to use an alternative method: to show a certainty/probability map, obtained as combination of different predictions, which could guide the end user in a more conscious decision for a final segmentation (see figure 2(a)).

The contribution of this study is threefold. Firstly, we proposed a deep learning segmentation method that is optimized towards clinical application, hence it mimics radiation oncologists in their task of slice-by-slice tumor contouring on PET-CT images. The deep learning framework takes sequences of three consecutive slices from the CT and PET images as input. The corresponding sequences extracted from the gross tumor volume of the primary tumor (GTVp) manually delineated by experts were used as ground truth. Bi-directional long short term memory (Bi-LSTM) and spatial and channel attention mechanisms were used to capture context information from the nearby slices and to enhance inter-slice dependencies. Secondly, we modelled probability maps exploiting the probabilistic outcome of the sigmoid function used to produce the output of the deep learning segmentation algorithm. Multiple predictions of a same image were averaged to obtain a wider range of outcome tumor probabilities. Those multiple predictions were obtained using different strategies. Each slice in a volume can be either first, second, or third in a sequence, resulting in three different predictions of the same slice in our method. The network was trained in parallel using sequences from the axial, coronal, and sagittal view, and 3-fold cross-validation was performed on each cross-sectional view, resulting in nine different predictions for each volume. Finally, we proposed the concept of probability maps as a tool to assist radiation oncologists in decision making, giving them a deep learning aided tool for tumor segmentation in radiotherapy treatment planning, where the optimal threshold for clinical acceptance could be customized and where the variability in agreement among models trained in the training set is available and displayed.

2. Materials and methods

2.1. Data collection

Data used in this study were manually collected from the Picture Archiving and Communication System (PACS) of the University Medical Center of Groningen (UMCG) in the Netherlands retrospectively. The data used for this research comprise confidential patient health information, which is protected and may not be released unless approved by the Committee of Ethics of the UMCG. We considered oropharyngeal cancer patients (OPC) who were treated, in our institute, with (chemo) radiation therapy between 2014 and 2017. 166 patients satisfied the eligibility criteria. For each case, planning CT and FDG-PET 3D images, and GTV primary tumor delineations, that were used in the radiotherapy treatment planning process, were collected. All imaging was
acquired in treatment position using fixation with a mask. The tumor delineations were reviewed by a specialized head and neck nuclear medicine physician and by a head and neck radiologist in cases where the MRI scan was also available. The tumor contours were approved after a multidisciplinary second reading session attended by all head and neck radiation oncologists.

Firstly, rigid image registration was performed. Radiation oncologists’ manual annotations of primary tumors were extracted from the DICOM RTSTRUCT files. Then, the imaging data were downloaded in DICOM format and transformed into NIfTI file format (Arnhold et al 1995) to facilitate data handling. PET images were transformed from raw to Standardized Uptake Values (SUV) (Ulaner 2019). For 9 patients, the PET image could not be registered with the planning CT, because either the PET image was missing or the images co-registration was not accurate due to differences in patient posture between the PET and CT images. Only patients having one primary tumor and who received radiation therapy as the first cancer treatment were included in this study. The final group satisfying the inclusion criteria consisted of 138 OPC patients (age: 61.95 ± 9.02 years). Participants provided written informed consent for their data to be used for research purposes. In Table 1, the dataset description including gender, T and N stages, and HPV status is reported.

### 2.2. Region of interest selection and pre-processing

Medical imaging data can be considerably large in dimensions. Since this study focused specifically on oropharyngeal cancer, a region of interest centered in the oral cavity was automatically extracted for each patient. The method used for the bounding box selection is inspired by Andrearczyk et al (2020). It consists of a two step approach: first, brain segmentation is performed identifying the brain as the largest connected component in the image after applying a fixed threshold for the SUV values of three in the PET image; second, a bounding box of fixed size (144 × 144 × 144) is determined with respect to the brain position, as in Andrearczyk et al (2020). We observed that using a fixed threshold of three to separate the brain from the rest of the head did not work for some of the patients in our dataset. The main problem occurred in cases where tumor activity on the PET image was included as it was part of the brain. To avoid manual assessment of the quality of the bounding boxes, we identified some metrics to guide us in manually correcting the threshold for the SUV values. We considered the brain volume (expected to be around 1260 cm³ in men and 1130 cm³ in women) and the maximum, the mean and the standard deviation of the Hounsfield units (HU) in the corresponding CT image of the segmented area of the brain. High values of maximum HU suggest that the segmentation is too large because bones have a much larger HU compared to soft tissues. Once the correct bounding boxes had been identified, CT and PET images were reconstructed into 144 × 144 × 144 pixels resulting in volumes of 144 × 144 × 144 mm³.

To facilitate tumor segmentation, a level of 40 HU and a window of 350 HU were chosen for the CT images. The PET images pixel values below 0 were set to 0. Finally, for both PET and CT images z-normalization was performed per patient.

### Table 1. Dataset description (n = 138).

| Variables   | Training (n = 113) | Test (n = 25) |
|-------------|--------------------|--------------|
| Gender      | Male 69 (61)       | 18 (72)      |
|             | Female 44 (39)     | 7 (28)       |
| T Stage     | T1 16 (14)         | 5 (20)       |
|             | T2 30 (27)         | 4 (16)       |
|             | T3 13 (11)         | 3 (12)       |
|             | T4 54 (48)         | 13 (52)      |
| N Stage     | N0 18 (16)         | 5 (20)       |
|             | N1 13 (11)         | 0 (0)        |
|             | N2a 6 (5)          | 0 (0)        |
|             | N2b 37 (33)        | 8 (32)       |
|             | N2c 36 (32)        | 10 (40)      |
|             | N3 3 (3)           | 2 (8)        |
| HPV status  | Negative 47 (42)   | 13 (52)      |
|             | Positive 47 (42)   | 9 (36)       |
|             | Unknown 19 (16)    | 3 (12)       |
2.3. Proposed method

To mimic clinicians’ job in the slice-by-slice segmentation task, we proposed a method which exploits the sequential nature of slices in a 3D volume. Our proposed method can be summarized in two steps:

(i) A deep learning network was trained on PET and CT images to identify tumor areas based on image features (see section 2.3.1). The input type of our network is a 3D volume of three consecutive slices extracted by the original bounding box. The backbone of the deep learning architecture is inspired by the work of Pan et al (2021) which addresses the problem of tumor segmentation in whole breast ultrasound images.

(ii) Probability maps were reconstructed for each patient. Firstly, single slice probability maps were obtained from the sequences generated by the deep learning network. Secondly, single-view (see section 2.3.2) and multi-view (see section 2.3.3) average based ensembling were performed.

2.3.1. Prediction model on sequences

The proposed deep learning segmentation model trains on sequences, as in Pan et al (2021). As training data, we used sequences of three consecutive 2D slices of PET \([pt_i, pt_{i+1}, pt_{i+2}]\) and CT images \([ct_i, ct_{i+1}, ct_{i+2}]\), concatenated in the channel domain (as shown in figure 3), together with three corresponding consecutive 2D slices of GTV primary tumor segmentation \([gTV_i, gTV_{i+1}, gTV_{i+2}]\), in the form of binary masks. The main reason to use only three slices was computational costs.

Tumor segmentation, as many medical imaging problems, suffers of class imbalance. In such a model, the problem is intensified, since negative slices represent the majority of the volume. To train and validate the network, sequence selection was needed. A sequence \([i, i + 1, i + 2] \) was always selected if its first slice \(i \) contained tumor pixels for more than 5% of the total area. The training/validation set consisted for 90% of those sequences. For the remaining 10% of the sequences in the training/validation, 5% were selected from sequences where the first slice was a negative slice and the other 5% from sequences containing tumor pixels for less than 5% of the total area. During the testing phase, all sequences from a volume were selected, containing both
positive and negative slices. In order to recognize negative examples as such, we also needed to include them in the training set.

As in Pan et al. (2021), the framework is divided into two main steps: first, an encoder decoder path, comprising spatial and channel attention blocks, extracts the main image features from each sequence; second, a bi-directional LSTM with spatial and channel attention aims to capture context information from the nearby slices. The network is described in all its parts in figure 3. Additional implementation details on the bi-directional LSTM and on the spatial and channel attention blocks can be found in Pan et al. (2021).

The loss function was obtained as the sum of the soft Dice loss calculated in three different steps during training (see 1, 2, and 3 in the yellow circles in figure 3). The soft Dice loss was calculated between the output predicted probability and the GTV primary tumor used as the ground truth. Dice Score Coefficient (DSC) was calculated on the validation set to monitor the training process. To perform this task, a threshold of 0.5 was applied to the pixel probability values of the predictions to transform segmentation results into binary masks. We saved only three different checkpoints: one corresponding to the last trained model after 150 epochs, the second one corresponding to the model with the highest mean DSC on the validation set, and the final one corresponding to the model which obtained the second highest value of mean DSC after the first 100 epochs, indicating the point where fluctuations in the loss functions decreased.

2.3.2. Single slice probability prediction
As shown in figure 3, the output of the proposed network is a sequence \([map_1, map_{i+1}, map_{i+2}]\) obtained as result of a sigmoid function, meaning that each pixel of the image corresponds to a probability indicating the chance of it belonging to tumor or not. During the testing phase, from each volume, a total of 142 sequences of three consecutive slices were extracted and predictions were calculated. For each slice in a volume, a minimum of one (for slice 1 and 144) and a maximum of three (from slice 3 to 142) different predictions were obtained. This is because slice number \(k\), with \(k\) from 3 to 142, will be first, second or third in a sequence.

\[
\begin{align*}
\text{seq}_1 &= [map_1, map_2, map_3] \\
\text{seq}_{k-1} &= [map_{k-2}, map_{k-1}, map_k] \\
\text{seq}_k &= [map_{k-1}, map_k, map_{k+1}] \\
\text{seq}_{k+1} &= [map_k, map_{k+1}, map_{k+2}] \\
\text{seq}_{142} &= [map_{142}, map_{143}, map_{144}].
\end{align*}
\]

Lastly, each final predicted slice of a volume \(\text{out}_k\) was calculated as the average of the total number of predictions for that slice:

\[
\begin{align*}
\text{out}_k = \left\{ \begin{array}{ll}
\text{seq}_k[1], & k = 1 \\
\frac{1}{2} (\text{seq}_k[1] + \text{seq}_{k-1}[2]), & k = 2 \\
\frac{1}{3} \sum_{i=1}^{3} \text{seq}_{k+i-1}[i], & 3 \leq k \leq 142 \\
\frac{1}{2} (\text{seq}_{k-1}[2] + \text{seq}_{k-2}[3]), & k = 143 \\
\text{seq}_{k-1}[3], & k = 144
\end{array} \right.
\]

Finally, the 3D volume was obtained by the sequential \(\text{out}_k\). The idea behind this last step is showing to the end user the tumor probability that the network assigned to each pixel.

2.3.3. Single- and multi-view average based ensembling
As last step, we used ensembling techniques to widen the range of probability of the output. Ensemble learning combines the predictions from multiple deep learning models to reduce the variance of predictions and reduce generalization error (Logan et al. 2021). It was shown that by averaging predictions over multiple models (deep ensemble) it is possible to capture model uncertainty (Lakshminarayanan et al. 2016). To exploit the inter-patient variability in tumour locations and imaging characteristics, we performed single-view average based ensembling. We averaged predictions obtained on the hold-out test set from the three models trained and validated during 3-fold cross validation. Finally, multi-view average based ensembling was performed fusing results from the three different orthogonal cross-sections. By training the network on sequences obtained slicing the volumes in different directions, we aimed to generate different predictions of a same spatial context. The operations of ensembling generated probability maps where high probability areas corresponded to higher agreement between trained models, and low probability areas to lower agreement.
2.4. Evaluation metrics and techniques

To evaluate the proposed method, we reconstructed each single slice from the predicted sequences of probabilities. For each patient we obtained 144 slices of probabilities for the axial, 144 for the sagittal and 144 for the coronal view. We obtained three 3D volumes from sequential 2D slices and one averaging multi-view prediction. We resampled each output to the original CT resolution and we calculated the metrics in 3D at different probability thresholds.

To evaluate the quality of the segmentation model we used the volumetric Dice similarity coefficient (volumetric DSC), the surface Dice similarity coefficient (surface DSC), and the 95th percentile Hausdorff distance (HD95), as defined in Nikolov et al (2018). Dice similarity coefficient is largely used in image segmentation tasks (Gudi et al 2017), because it quantifies the overlap between the prediction result and the ground truth image, which in our case is the GTV contour of the primary tumor. However, it was shown that it is not well suited when assessing the clinical applicability of the contours. Firstly, it does not take into account the distance from the surface when quantifying the mismatch between structures (Nikolov et al 2018); secondly, it is highly affected by the structure size; finally, it does not always correlate with the time needed for manual adjustments (Vaassen et al 2020). Surface DSC is a segmentation performance metric introduced by Nikolov et al (2018) which calculates the overlap between the surfaces of two structures at a specific tolerance. The tolerance represents the range within which a mismatch between the predicted and target structure can still be considered clinically acceptable. Using surface DSC, instead of volumetric DSC, represents a benefit in tasks such as tumor segmentation where the manual delineation of the ground truth suffers from inter-observer variability. The HD95 is the 95% percentile of the ordered distance measures between two contours and it is calculated in millimeter. It is widely used in clinic and it provides a more intuitive indication of segmentation quality.

3. Experimental results

3.1. Implementation details

The deep learning framework was implemented and trained using Python (v3.7.4), PyTorch (v1.6.0), MONAI (v0.8.0), CUDA (v10.1), and cuDNN (v7.6.4) on GPUs of the Peregrine Cluster of the University of Groningen. The training and validation set consisted of 113 patients, while the test set consisted of 25 patients. The same network was trained in parallel three times using, each time, different input data: sequences from the axial (Ma), sagittal (Ms) and coronal (Mc) view of the 3D volumes. Batch size of 1 was used during training, consisting of one sequence of three consecutive slices from the same patient, extracted from the same plane. To avoid overfitting, 3-fold cross validation was performed for Ms, Mc and Ms, resulting in a total of 9 final models. A fixed maximum number of epochs was set to 150. Adam optimizer with a fixed learning rate equal to 0.0002 was used for training. The publicly available package surface-distance4 was used to evaluate our method. For the surface DSC, a tolerance of 3.00 mm was chosen, as in Naser et al (2022).

3.2. Deep learning network performance analysis

Training and validation losses were monitored during training the network. During training, the mean DSC was calculated over the 3 validation sets, made of sequences selected using the criteria explained in section 2.3.1. To obtain final predictions, we selected the model which achieved the highest value of mean DSC (calculated after binarizing predictions using a threshold of 0.5) after the first 100 epochs, because the fluctuations in both loss functions decreased. The highest value of mean DSC was achieved between the 103th and the 149th epoch when trained on the selected sequences from the axial (0.67 ± 0.02 DSC), coronal (0.56 ± 0.04 DSC) and sagittal (0.64 ± 0.02 DSC) planes of the validation set.

3.3. Probability maps reconstruction results

As a final step, 142 consecutive sequences were extracted from each image volume of each patient in the validation set and predictions were obtained. Single slices and 3D volumes (multi-view results) were then reconstructed as explained in sections 2.3.2 and 2.3.3. The performance of the method on each patient was assessed by comparing the evaluation metrics of the reconstructed volumes for the Ms, Mc, Ma and multi-view results (Multi view). We used different probability thresholds, when transforming images into binary masks, and we recalculated the evaluation metrics.

Using a probability threshold of th = 0.9, when binarizing the output, resulted in an overall higher mean surface DSC and lower median HD95. In table 2 the validation and test results obtained by single- and multi-view average based ensembling, at a probability threshold of th = 0.9, are shown. When testing, in each direction,

---

4 https://github.com/deepmind/surface-distance
slices were obtained by ensembling predictions of the three optimized models trained using 3-fold cross validation. The higher performance in the test set than in the validation set confirmed the robustness of the ensemble technique. For the single-view results, the mean surface DSC of the $M_a$ achieved the lowest score. $M_a$ and $M_s$ showed comparable results in terms of mean surface DSC in the test set, despite their difference in performance in the validation set. The results from the Multi view outperformed single-view models in terms of both surface DSC and HD95.

In figure 4 the distributions of the volumetric DSC, surface DSC and HD95 obtained from the test Multi view results are reported at different probability thresholds using boxplots. An ascending pattern can be observed for the surface DSC, meaning that when the threshold of the predicted probability increases, the segmentation prediction surface overlaps better with the ground truth delineation of the tumor. The descendant pattern of the HD95 across thresholds confirms that the higher the threshold, the closest the generated contour is to the GTV primary tumor manually delineated. The volumetric DSC has an ascendant-descendant behavior across probability thresholds. The highest median result of 0.74 is achieved when $th = 0.7$.

In figure 5 boxplots showing the surface DSC distributions across probability thresholds for different T and N stages can be found. Cancer staging describes the severity of cancer: T stage is mainly related to the primary tumor size, N stage to pathological lymph nodes. An overall increasing pattern in mean surface DSC can be observed for T and N stages. However, predictions of T1, T2 and N3 stages show lower performance results. Predictions of T3 and T4 stages result always in higher surface DSC compared to predictions of lower T stages across probability thresholds. The same pattern can be observed for N0, N2b, and N2c stages compared to N3. Predictions of N3 stage result in a constant mean surface DSC until $th = 0.8$. In figures A.1 and A.2, in the
appendix, boxplots showing the volumetric DSC distributions across probability thresholds for different T and N stages are reported. Discrepancies in performance between T and N stages are consistent with the surface DSC results.

3.4. Qualitative results

In figure 6, we show how the probability map predictions of our network looks on the CT image of a patient. On the left side, the total range of probabilities is shown on few slices extracted from the axial plane. On the right side, four cases of different probability threshold settings are displayed. The range of probability shown on the image for each case corresponds to the one above the selected threshold up to 100%. Increasing the probability percentage shrinks the predicted area around the tumor.

After a qualitative assessment of the results, we acknowledged that the low performance of the models, in terms of surface DSC, corresponded to cases where big metastatic lymph nodes were present and primary tumor volume was smaller.

4. Discussion

The experimental results show the potential of the proposed deep learning-based method with adaptive thresholding for predicted tumor probability for oropharyngeal cancer in PET-CT images. Our method is a threshold-based approach intended for clinical use. The added value of the method is providing probability maps which could be used in form of clinical decision support as a starting point in the contouring process by the radiation oncologist. Providing the radiation oncologist with contours of the GTV primary tumor for different probabilities, makes it possible to choose the most appropriate contour. This contour could then be optimized based on clinical information and additional imaging. It is expected that by offering the radiation oncologist a GTV contour to start with, the contouring process can become faster and the inter-observer variability could be reduced (Vaassen et al 2020).

A recently introduced metric was preferred to measure the performance of our method. As shown already in the literature by Vaassen et al (2020), compared to volumetric DSC, surface DSC provides more clinically relevant and better quantitative measures for automatically-generated contour quality. In addition, for the type of output we propose, we believe that metrics such as surface DSC and HD$_{95}$ represent a better fit than volumetric DSC, which strictly measures the exact overlap of two volumes. During the second edition of the HECKTOR challenge (Andrearczyk et al 2022b), Xie and Peng (Xie and Peng 2022) ranked first in the task of automatic tumor segmentation using 3D PET-CT images and 3D U-Net with squeeze-and-excitation normalization. They achieved a value of mean volumetric DSC of 0.7785 and a median HD$_{95}$ of 3.088 mm on a test set of 101 cases from two different centers. In terms of median HD$_{95}$, our test result of 3.906 mm is comparable. However, in terms of volumetric DSC, our method showed comparable performance for T3 (0.82
mean volumetric DSC and T4 (0.74 mean volumetric DSC) tumor stages at a threshold of 0.7, but lower performance for T1 and T2 stages (see A.1). Not far from the winner result, Naser et al (2022) ranked 8th achieving a mean volumetric DSC of 0.770 and a median HD95 of 3.143 mm. In their work they used surface DSC at a tolerance of 3 mm as evaluation metric of the validation results achieving 0.892 mean surface DSC and 6.976 mm median HD95 versus our validation results of 0.72 mean surface DSC and 5.053 mm median HD95, and our test results of 0.81 mean surface DSC and 3.906 mm median HD95. The promising and comparable results show the potential of our method in guiding radiation oncologists in clinical practice.

Despite several studies showing promising results in recognizing head and neck tumors on multi-modal image data using deep learning (Oreiller et al 2022), the limitations involved are still an obstacle for their clinical implementation. In this paper, we suggest a solution that tackles the problem of high variability in ground truth images used for tumor segmentation, and that may enhance the utility of such a tool in clinical practice. The proposed method, in fact, displays on the different CT-slices a tumor probability derived by the tumor segmentation learning process and supports the end user towards a quicker and more attentive decision making. In tasks where the gold standard is not represented by one and only one ground truth, like the one of tumor segmentation, a large number of different annotations would be required for training a DL network. In several deep learning studies, one adopted solution, is to collect consistent tumor delineations obtained after consultations of different radiation oncologists for the whole dataset (Andrearczyk et al 2022b, Naser et al 2022). The dataset used, in these cases, is consistent, and the deep learning framework learns according to what has been decided to be ‘the truth’. However, relabeling a dataset is not practical and it is time consuming. New data will always be collected and used to update existing models and similar problems will again be encountered. In our paper, data has been collected between 2014 and 2017. In this time period, different radiation oncologists have been working in our center and additional image modalities became available for a more precise final tumor contour (e.g. MRI). Despite the delineation guidelines being the same, the variation among tumor delineations is large. Supervised learning methods reproduce human biases and errors, thus the problem of
inter-observer variability in target volume delineation is a perpetuate error. This can be one of the reasons why the standard deviation in our surface DSC results is higher than in Naser et al (2022) (0.35 versus 0.04).

Recently, providing the end user with information related to the confidence of the network in making certain choices gained greater importance in clinical applications. Wang et al (2019) introduced the concept of pixel-wise uncertainty estimation which is defined as the uncertainty of the segmentation result of a single pixel and represents the model’s confidence in single pixel classification. For uncertainty estimation the variance or the entropy of predictions for a given image are calculated. In Diao et al (2022) and Rosvoll Groendahl et al (2021) entropy-based loss functions are used to create output uncertainty maps. In Diao et al (2022), the concept of pixel-level uncertainty is explored in the tumor segmentation task. They designed a method which outputs the segmentation output, the pixel uncertainty maps and the case level uncertainty used for out-of-distribution detection. In Rosvoll Groendahl et al (2021) tumor segmentation is performed on PET-CT volumes and the model uncertainty derived by segmenting on each single modality is modelled using evidence theory. In our work we do not generate uncertainty maps, but probability maps where high probability areas correspond to higher agreement among trained models, and low probability areas to lower agreement. To widen the output probability range we averaged over multiple predictions of a same image. We obtained predictions by the bidirectional LSTM model, 3-folds cross validation, and multi-view outputs. The bi-directional LSTM network is a powerful tool for modeling the sequential dependencies of slices in a volume. By averaging predictions over multiple models trained using 3-folds cross validation, model uncertainty is captured. This concept is addressed by the term deep ensemble and it was firstly introduced in Lakshminarayanan et al (2016) for predictive uncertainty quantification. Compared to other uncertainty quantification methods, deep ensemble does not require changes in the network architectures and it is easy to implement. Finally, we performed multi-view average based ensembling, fusing results obtained by models trained on sequences extracted by single-views. In our method, we capture the variation between patients, images and contouring in the training set and we display the variation in contours between the different probability threshold values.

To evaluate the method presented in this paper, we calculated metrics at different probability thresholds to establish whether the outcome predictions were matching the manually delineated GTV primary tumor or not. The contours generated by choosing a specific probability threshold value are not post-processed after ensembling, hence areas of lower predicted probabilities can still be present as noise in the output (as shown in figure 6). We decided not to apply post-processing techniques (as keeping the largest connected component) to the final results because the generated contours are meant to be adapted for clinical practice and additional information can still be relevant to clinicians. Increasing the probability threshold, predictions showed to become more similar to what the oncologist defined as tumor delineation. When setting a probability threshold of 0.9, only three out of 25 tumors in the test set have a corresponding surface DSC of 0. In two cases the tumor dimension was small and really difficult to identify on PET, as also mentioned in the patient file. In one case the FDG PET uptake was absent in the primary tumor location. However, all three cases showed probabilities with lower predicted values in the tumor region, which indicates that some of the trained models assigned tumor probabilities to those regions but the agreement among models was low. In our study we included all patients satisfying the inclusion criteria, also the ones that were considered difficult to contour based on CT and PET images by radiation oncologists. The choice was made because the aim of our work was to provide a threshold-based approach to aid radiation oncologists in tumor segmentation.

In our work, we build on the fact that tumors are identified on a pixel level on multi-modal images. The combination of PET and CT images gives complementary information that makes the network assign a certain tumor probability to a pixel. The channel attention blocks of the first part of the framework (in figure 3) are intended to guarantee cross-modal learning. On one hand, different imaging modalities have a positive impact on contouring consistency, taking advantage of all the imaging information available for segmentation (Bird et al 2015). On the other hand, contour variations can result from differences in tumor visualization in various imaging modalities and not accurate multi-modal image co-registrations (Sadeghi et al 2021). PET imaging, in fact, does not allow a high delineation accuracy but is fundamental in identifying head and neck cancer. The implemented deep learning framework heavily relies on FDG uptake in PET images. Since the brain area is characterized by a large amount of pixels with high SUV values, these areas can also appear in the outcome prediction, but with low probabilities. This explains the large difference in surface DSC between probability thresholds below and above 0.3 in the axial and coronal view in A.3. When trained on sequences from the axial plane, the network appears to rely on PET intensity and on tumor location. In the presence of primary tumors with larger volume that expand to metastatic lymph nodes (higher N stages), the mean DSC values were lower.
5. Conclusions

In conclusion, this study proposes a novel deep learning aided tool for oropharyngeal tumor segmentation on FDG-PET/CT images using probability maps which is biased towards clinical application. The framework mimics experts in their task of slice-by-slice tumor contouring. The main objective is to obtain a probability map where the range of pixel values (probabilities) is wider and captures the model uncertainty in the segmentation task, which could represent useful information to assist radiation oncologists in tumor contouring for radiotherapy treatment planning. Context information from nearby slices is enhanced by training on sequences, thanks to the bi-directional long short term memory mechanism. Single-view average based ensembling is used to estimate the uncertainty related to inter-patient variability. Multi-view average based ensembling is performed to fuse information learnt in each single view to obtain one final prediction. The available wide range of probabilities is of interest to minimise the risk of geographical misses and to make available the variability derived from the tumor segmentation task. Experimental results show the potential of the proposed method in identifying the primary tumor in correspondence of the highest agreement among trained models and that the performance is comparable to published DL methods for contouring of head and neck tumors.

Acknowledgments

This research was supported by the Hanarth Fonds. We would like to acknowledge the contribution in data collection of the PhD students working on Head and Neck cancer research in the Department of Radiation Oncology, at the University Medical Hospital of Groningen, Netherlands. We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

Data availability statement

The data cannot be made publicly available upon publication due to legal restrictions preventing unrestricted public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Declaration of competing interest

The authors declare no conflict of interest.
Appendix

Figure A.1. Volumetric DSC of patients in test set grouped by T stages at different probability thresholds.
Figure A.2. Volumetric DSC of patients in test set grouped by N stages at different probability thresholds.
References

Anderson C M, Sun W, Buatti J M, Maley J E, Policeni B, Mott S L, Bayouth J E and Author C 2014 Interobserver and intermodality variability in GTV delineation on simulation CT, FDG-PET, and MR images of head and neck cancer Jacobs J. Radiat. Oncol. 11:006

Andrearczyk V, Oreiller V and Depeursinge A 2020 Oropharynx detection in PET-CT for tumor segmentation Irish Machine Vision and Image Processing Conference (IMVIP) (Sligo, Ireland)

Andrearczyk V, Oreiller V and Depeursinge A 2021a Head and Neck Tumor Segmentation (Cham: Springer International Publishing) (Lecture Notes in Computer Science LNCS, volume 12603) 1st ed

Andrearczyk V, Oreiller V, Hatt M and Depeursinge A 2022b Head and Neck Tumor Segmentation and Outcome Prediction (Cham: Springer International Publishing) (Lecture Notes in Computer Science 13209) 2nd ed

Andrearczyk V et al 2022a Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images doi:https://doi.org/10.48550/arXiv.2201.04138. URLhttps://arxiv.org/abs/2201.04138v1

Arnholt J C M D, Hanson D P and Robb R A 1995 NIFTY (NeuroImaging Functional Toolkit): an interactive program for functional MRI data processing and visualization Physiology and Function from Multidimensional Images vol 2433 (SPIE) pp 276–87

Bird D et al 2015 Multimodality imaging with CT, MR and FDG-PET for radiotherapy target volume delineation in oropharyngeal squamous cell carcinoma BMJ Cancer 15 844

Diao Z, Jiang H and Shi T 2022 A unified uncertainty network for tumor segmentation using uncertainty cross entropy loss and prototype similarity Knowl. Syst. 246 108739

Gudi S, Ghosh-Laskar S, Agarwal J P, Chaudhari S, Rangarajan V, Nojin Paul S, Upreti R, Murthy V, Budrukarka A and Gupta T 2017 Interobserver variability in the delineation of gross tumour volume and specified organs-at-risk during IMRT for head and neck cancers and the impact of FDG-PET/CT on such variability at the primary site J. Med. Imaging Radiat. Sci. 48 184–92

Lakshminarayanan B, Pritzl A and Blundell C 2016 Simple and scalable predictive uncertainty estimation using deep ensembles Adv. Neural Inf. Process. Syst. 2017 6403–14

Logan R, Williams B G, Ferreira da Silva M, Indani A, Scholcinov G, Nagy C, Maley J E et al 2021 Deep convolutional neural networks with ensemble learning and generative adversarial networks for Alzheimer’s disease MRI data classification Front. Aging Neurosci. 13 720236

Nair T, Precup D, Arnold D L and Arbel T 2020 Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation Med. Image Anal. 59 101557

Naser M A, Wahid K A, van Dijk L V, He R, Abdelal M A, Dede C, Mohamed A S and Fuller C D 2022 Head and neck cancer primary tumor auto segmentation using model ensembling of deep learning in PET/CT images 13209 121 Head and neck tumor segmentation and outcome prediction : second challenge, HECKTOR 2021, held in conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Head and Neck Tumor Segmentation Challenge (2nd, 2021)

Nikolov S et al 2022 Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study Journal of medical Internet Research 23 7 e26151

Figure A.3. Surface DSC of single- and multi-view results for patients in test set at different probability thresholds.
Orellier V et al 2022 Head and neck tumor segmentation in PET/CT: the HECKTOR challenge Med. Image Anal. 77 102336
Pan P, Chen H, Li Y, Cai N, Cheng L and Wang S 2021 Tumor segmentation in automated whole breast ultrasound using bidirectional LSTM neural network and attention mechanism Ultrasomics 110 106271
Rodriguez Outeiral R, Bos P, Al-Mamgani A, Jaspeure B, Simões R and van der Heide U A 2021 Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning Phys. Imaging Radiat. Oncol. 19 39–44
Rosvoll Groendahl A et al 2021 EFNNet: evidence fusion network for tumor segmentation from PET-CT volumes Phys. Med. Biol. 66 205005
Sadeghi S, Siavashpour Z, Sadr A V, Farzin M, Sharp R and Gholami S 2021 A rapid review of influential factors and appraised solutions on organ delineation uncertainties reduction in radiotherapy Biomed. Phys. Eng. Express 7 052001
Ulaner G A 2019 FDG PET/CT performance and reporting Fundamentals Oncol. PET/CT (Philadelphia: Elsevier) 2 5–8
Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Caners R and van Elmpt W 2020 Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy Phys. Imaging Radiat. Oncol. 13 1–6
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. Therapeutic Radiol. Oncol. 137 9–15
Wang G, Li W, Aertsen M, Deprest J, Ourselin S and Vercauteren T 2019 Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks Neurocomputing 358 34–45
Weiss E and Hess C F 2003 The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy theoretical aspects and practical experiences Strahlenther Onkol 179 21–30
Wong J, Fong A, McVicar N, Smith S, Giambattista J, Wells D, Kolbeck C, Giambattista J, Gondara L and Alexander A 2020 Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning Radiother. Oncol. 144 152–8
Xie J and Peng Y 2022 The head and neck tumor segmentation based on 3D U-net Head and Neck Tumor Segmentation and Outcome Prediction (Cham: Springer) 13209, 92–98 Head and neck tumor segmentation and outcome prediction: second challenge, HECKTOR 2021, held in conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Head and Neck Tumor Segmentation Challenge (2nd : 2021)