# Head and Neck Cancer Primary Tumor Auto Segmentation using Model Ensembling of Deep Learning in PET/CT Images

Mohamed A. Naser[1[0000-0003-1020-4966]], Kareem A. Wahid[1[0000-0002-0503-0175]], Lisanne V. van Dijk[1[0000-0002-9515-5616]], Renjie He[1[0000-0001-9166-6286]], Moamen Abobakr Abdelaal[1[0000-0003-4476-2122]], Cem Dede[1[0000-0002-0543-9325]], Abdallah S.R. Mohamed[1[0000-0003-2064-7613]], and Clifton D. Fuller[1[0000-0002-5264-3994]]

[1] Department of Radiation Oncology, The University of Texas MD Anderson Cancer, Houston, Texas 77030, USA
manaser@mdanderson.org

**Abstract.**

Auto-segmentation of primary tumors in oropharyngeal cancer using PET/CT images is an unmet need that has the potential to improve radiation oncology workflows. In this study, we develop a series of deep learning models based on a 3D Residual Unet (ResUnet) architecture that can segment oropharyngeal tumors with high performance as demonstrated through internal and external validation of large-scale datasets (training size = 224 patients, testing size = 101 patients) as part of the 2021 HECKTOR Challenge. Specifically, we leverage ResUNet models with either 256 or 512 bottleneck layer channels that demonstrate internal validation (10-fold cross-validation) mean Dice similarity coefficient (DSC) up to 0.771 and median 95% Hausdorff distance (95% HD) as low as 2.919 mm. We employ label fusion ensemble approaches, including Simultaneous Truth and Performance Level Estimation (STAPLE) and a voxel-level threshold approach based on majority voting (AVERAGE), to generate consensus segmentations on the test data by combining the segmentations produced through different trained cross-validation models. We demonstrate that our best performing ensembling approach (256 channels AVERAGE) achieves a mean DSC of 0.770 and median 95% HD of 3.143 mm through independent external validation on the test set. Our DSC and 95% HD test results are within 0.01 and 0.06 mm of the top ranked model in the competition, respectively. Concordance of internal and external validation results suggests our models are robust and can generalize well to unseen PET/CT data. We advocate that ResUNet models coupled to label fusion ensembling approaches are promising candidates for PET/CT oropharyngeal primary tumors auto-segmentation. Future investigations should target the ideal combination of channel combinations and label fusion strategies to maximize segmentation performance.

**Keywords:** PET, CT, Tumor Segmentation, Head and Neck Cancer, Oropharyngeal Cancer, Deep Learning, Auto-contouring.

# 1 Introduction

Oropharyngeal cancer (OPC) is a type of head and neck squamous cell carcinoma that affects a large number of individuals across the world [1]. Radiation therapy is an effective component of OPC treatment but is highly dependent on accurate segmentation of gross tumor volumes [2], i.e., visible gross disease that is informed by clinical examination and radiographic findings. Importantly, precise tumor delineation is crucial to ensure adequate radiation therapy dose to target volumes while minimizing dose to surrounding healthy tissues. The combination of computed tomography (CT) with positron emission tomography (PET) allows for sufficient anatomic detail in determining tumor location coupled to underlying physiologic information [3]. However, tumor segmentation in OPC has long been seen as an inefficient and potentially inconsistent process as multiple studies have demonstrated high human inter- and intra-observer segmentation variability [4, 5]. Therefore, developing automated tools, such as those based on deep learning [6–9], to reduce the variability in OPC PET/CT tumor segmentation while retaining reasonable performance is imperative for improving the radiation therapy workflow.

The annual Medical Image Computing and Computer Assisted Intervention Society (MICCAI) Head and Neck Tumor Segmentation Challenge (HECKTOR) has provided an avenue to systematically evaluate different OPC primary tumor auto-segmentation methodologies through the release of high-quality, multi-institutional training and testing PET/CT data. We previously participated in the 2020 HECKTOR challenge and achieved reasonable results using deep learning approaches [10]. Subsequently, we improve upon our previous approach through various architectural modifications, ensembling of independent models' predictions, and additional provided training/testing data, that ultimately leads to improved segmentation performance. This work presents the results of our OPC primary tumor auto-segmentation model based on a ResUnet deep learning model applied to the 2021 HECKTOR Challenge PET/CT training and testing data.

# 2 Methods

We developed deep learning models (2.3) for auto-segmentation of primary tumors of OPC patients using co-registered $^{18}$F-FDG PET and CT imaging data (2.1). The ground truth manual segmentations of the tumors and the normalized imaging data (2.2) were used to train the models (2.4). The performance of the trained models for auto-segmentation were validated using a 10-fold cross-validation approach (2.5).

## 2.1 Imaging Data

The data set used in this study, which was released through AIcrowd [11] for the HECKTOR Challenge at MICCAI 2021 [12–14], consists of co-registered $^{18}$F-FDG PET and CT scans for 325 OPC patients (224 patients used for training and 101 patients used for testing, previously partitioned by the HECKTOR Challenge organizers). All imaging data in the training set (224 patients) was paired with ground

truth manual segmentations of the OPC primary tumors derived from clinical experts (HECKTOR Challenge organizers). All training and testing data were provided in Neuroimaging Informatics Technology Initiative (NIfTI) format.

### 2.2 Image Processing

All images (i.e., PET, CT, and tumor segmentation masks) were cropped to fixed bounding box volumes, provided with the imaging data (2.1) by the HECKTOR Challenge organizers [11], of size 144x144x144 $mm^3$ in the x, y and z dimensions. To mitigate the variable resolution of the PET and CT images, the cropped images were resampled to a fixed image resolution of 1 mm in the x, y, and z dimensions. We used spline interpolation of order 3 for resampling the PET/CT images and nearest-neighbor interpolation for resampling the segmentation masks. We based our cropping and resampling work on the code provided by the HECKTOR Challenge organizers (https://github.com/voreille/hecktor). The CT intensities were truncated in the range of [-200, 200] Hounsfield Units (HU) to increase soft tissue contrast and then were normalized to a [-1, 1] scale. The intensities of PET images were normalized with z-score normalization ([intensity-mean]/standard deviation). We used the Medical Open Network for AI (MONAI) [15] software transformation packages to rescale and normalize the intensities of the PET/CT images. Image processing steps used in this manuscript are displayed in **Figure 1**.

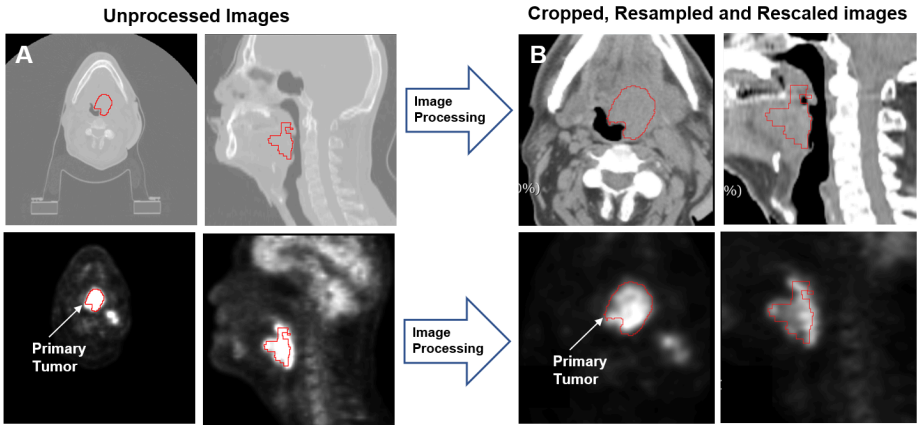


**Fig. 1.** An illustration of the workflow used for image processing. (A) Overlays of the provided ground truth tumor segmentation masks (red outline) and the original CT (top) and PET (bottom) images. (B) Overlays of the provided ground truth tumor segmentation masks (red outline) and the processed CT (top) and PET (bottom) images.

### 2.3 Model Architecture

A deep learning convolutional neural network model based on the ResUnet architecture included in the MONAI software package was used for the analysis. As shown in **Figure 2**, the network consisted of 4 convolution blocks in the encoding and decoding branches and a bottleneck convolution block between the two branches. All convolution layers used a kernel size of 3 except one convolution layer in the

bottleneck, which used a kernel size of 1. The number of output channels for each convolution layer is given above each layer, as shown in **Figure 2**. Each convolution block in the encoding branch was composed of a two-strided convolution layer and a residual connection that contained a two-strided convolution layer and a one-strided convolution layer. In the bottleneck, the residual connection contained two one-strided convolution layers. In the decoding branch, each block contained a two strided convolution transpose layer, a one strided convolution layer and a residual connection. Batch normalization and parametric ReLU (PReLU) activation functions were used throughout the architecture. The PET/CT images acted as two channel inputs to the model, while a two-channel output provideed the tumor segmentation mask (i.e., 0 = background, 1 = tumor). The architecture shown in **Figure 2** corresponds to a ResUnet with a maximum of 512 channels in the bottleneck layer (512 Model) where the number of channels in the convolution layers was (32, 64, 128, 256, and 512). We also implemented a model using a maximum of 256 channels in the bottleneck layer (256 Model), which has the same structure as the 512 Model, but the number of channels in the convolution layers was (16, 32, 64, 128, and 256).

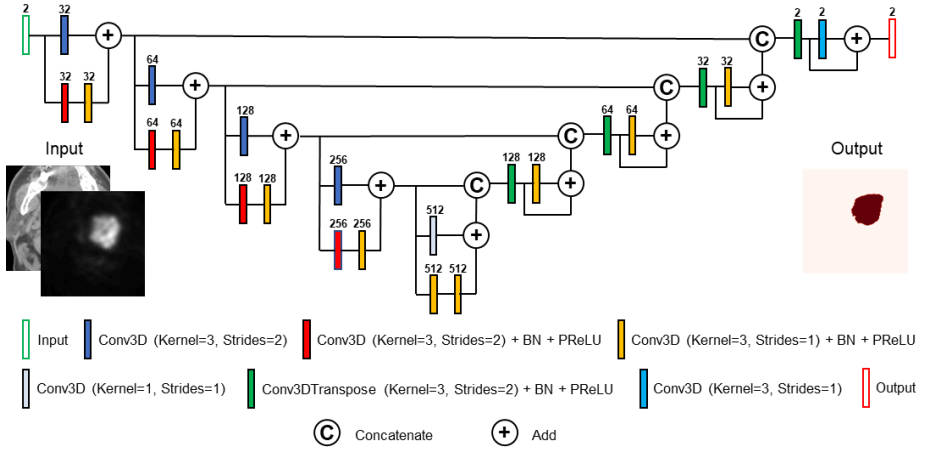


**Fig. 2.** Schematic of the ResUnet architecture used for the segmentation model. The number of channels (32, 64, 128, 256, and 512) is given above each block. The batch normalization and the parametric ReLU layers are annotated by (BN) and PReLU, respectively. The channels given in the figure are for the 512 model, while for the 256 model the channels are (16, 32, 64, 128, and 256).

## 2.4 Model Implementation

We used a 10-fold cross-validation approach where the 224 patients from the training data were randomly divided into ten non-overlapping sets. Each set (22 patients) was used for model validation while the remaining 202 patients in the remaining sets were used for training, i.e., each set was used once for testing and nine times for training. The processed PET, CT, and tumor masks (2.2) were randomly cropped to four random fixed-sized regions (patches) of size (96, 96, 96) per patch per patient. The random spatial cropping considered the patch center of mass as foreground (i.e., a tumor - positive) or background (i.e., non-tumor - negative) with a 50% probability

for both the positive and negative cases as shown in **Figure 3A**. We used a batch size of 2 patients' images and, therefore, a total of 8 patches of images. The shape of the input tensor provided to the network (2.3) for a batch size of 2, patches per image of 4, a two-channel input (PET/CT), and patch size of (96, 96, 96) is (8, 2, 96, 96, 96). The tumor mask was used as the ground truth target to train the segmentation model. The shape of the target tensor provided was (8, 1, 96, 96, 96). To minimize overfitting, in addition to the random spatial cropping to patch the images and masks, we implemented additional data augmentation to both image and mask patches which includes random horizonal flips of 50%, and random affine transformations with an axial rotation range of 12 degrees and scale range of 10%. We used Adam as the optimizer and Dice loss as the loss function. The model was trained for 700 iterations with a learning rate of $2\times10^{-4}$ for the first 550 iterations and $1\times10^{-4}$ for the remaining 150 iterations. The image processing (2.2), data augmentation, network architecture, and loss function were used from the software packages provided by the MONAI framework [15]; code for these packages can be found at "https://github.com/Project-MONAI/".

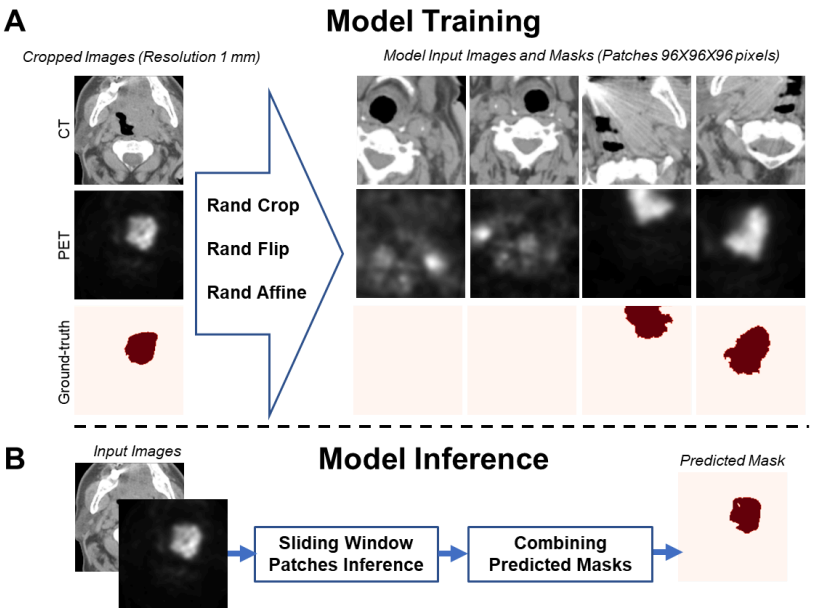


**Fig. 3.** An illustration of the workflow of the training and inference phases of the segmentation model. (A) Data transformation and augmentation is used to produce input data to the model. An example of four patches of CT and PET images and the corresponding ground truth tumor segmentation with at 50% representation of the tumor in these patches used for training the segmentation model (four patches of images per patient - 96 x 96 x 96 voxels each). (B) Segmentation model prediction using sliding window inferences (96X96X96 voxels each) and combining the predicted masks from all patches to provide the final mask.

### 2.5 **Model Validation**

For each validation fold (i.e., 22 patients), we trained the 256 and 512 ResUnet models (2.3) on the remaining 202 patients. Therefore, we obtained 10 different models for the 256 and 512 networks each from 10-fold cross-validation. We applied an argmax function to the two-channel output of each model to generate the predicted tumor segmentation mask (i.e., 0 = background, 1 = tumor). We evaluated the performance of each separate model on the corresponding validation set using metrics of spatial overlap (Sørensen–Dice similarity coefficient [DSC] [16], recall, and precision) and surface distance (surface DSC [17], 95% Hausdorff distance [95% HD] [18]) between generated and ground truth segmentations. The surface distance metrics were calculated using the surface-distances Python package by DeepMind [17]. A tolerance of 3.0 mm was chosen for calculation of surface DSC based on previous investigations [19, 20] as a reasonable estimate of human inter-observer error.

For the test set (101 patients), we implemented two different model ensembling approaches post-hoc (after training) to estimate the predicted tumor masks. In the first approach, we use used the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm [21] as a method to fuse labels generated by applying the 10 models produced during the 10-fold cross-validation on the test data set, i.e., generate the consensus predicted masks from the different generated predicted masks (STAPLE approach). The STAPLE algorithm was derived from publicly available Python code (https://github.com/fepegar/staple). In the second approach, we implemented a simple threshold of agreement based on all cross-validation fold models at the voxel level (AVERAGE approach). The total number of cross-validation models used in thresholding could be modulated as a parameter for this approach. For our purposes, we selected a threshold of 5 cross-validation folds as a proxy for majority voting, i.e. at least 5 cross-validation models must consider a voxel to be a tumor (label = 1) for that final voxel label to be considered as a tumor (label =1). Majority voting in this context was chosen since it is common in other model ensembling approaches [22].

## 3 Results

The performances of the segmentation models are illustrated in **Fig. 4**, which shows Boxplots of the DSC, recall, precision, surface DSC, and 95% HD distributions obtained using the 10-fold cross-validation approach described in (2.5). The mean ± standard deviation values of the DSC, recall, precision, surface DSC, and 95% HD achieved by the 256 and 512 Models are 0.771 ± 0.039 and 0.768 ± 0.041, 0.807 ± 0.042 and 0.793 ± 0.038, 0.788 ± 0.038 and 0.797 ± 0.038, 0.892 ± 0.042 and 0.890 ± 0.044, and 6.976 ± 2.405 and 6.807 ± 2.357 respectively. The mean and median values of these metrics are summarized in **Table 1**. Notably, one case did not return a segmentation prediction (CHUS028) for either the 256 or 512 models, which led to the spurious prediction of surface distance metrics. Therefore, this case has been excluded in the analysis of surface DSC and 95% HD.

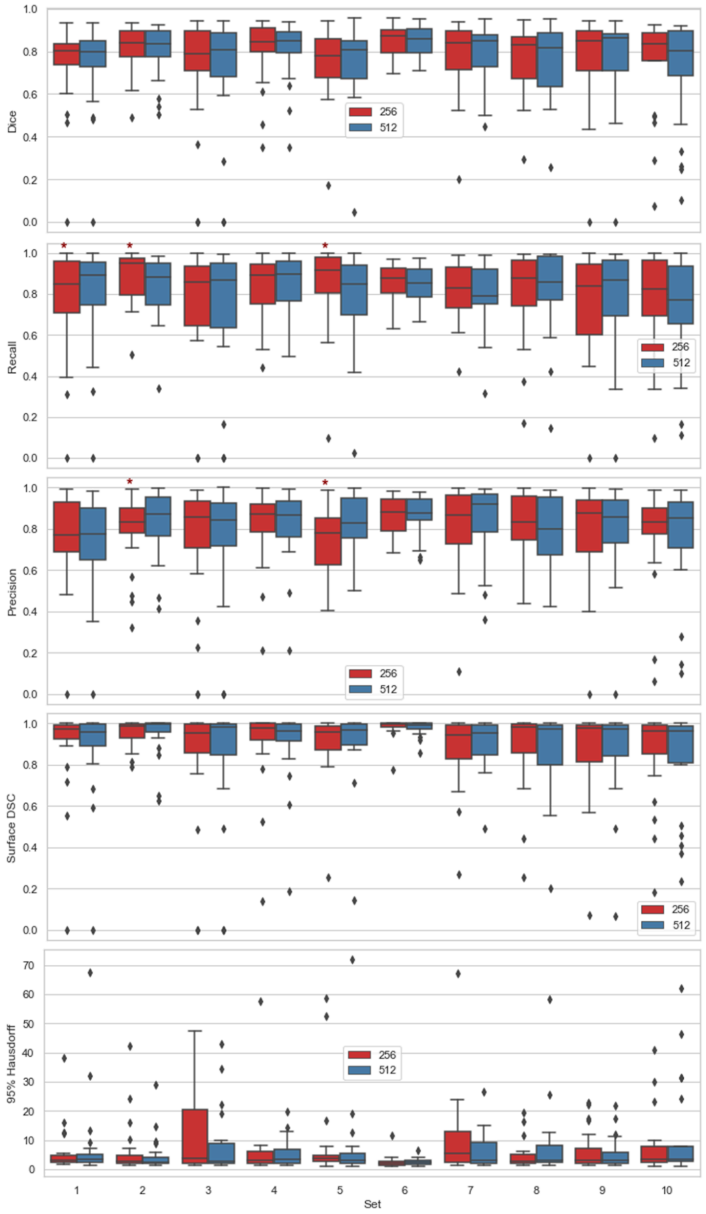


**Fig. 4**. Boxplots of the DSC, recall, precision, surface DSC, and 95% HD distributions for the 10-fold cross-validation data sets (Set 1 to Set 10 – 22 patients each*) used for the 256 and 512 ResUnet models. The lines inside the boxes refer to the median values. The stars refer to significant differences in the results by the two models (p-value < 0.05) using two-sided Wilcoxon signed-rank test. [1] One patient in Set 1 (CHUS028) did not return a segmentation prediction for either model and was thus excluded from the analysis of surface distance metrics (surface DSC, 95% HD).

**Table 1.** 256 and 512 ResUnet model performance metrics. [1] One case (CHUS028) in a cross-validation fold contained no segmentation prediction for either model and led to erroneous surface distance calculations; therefore, this case was excluded from the presented surface distance metric results.

| Model | DSC | Recall | Precision | Surface DSC[1] | 95% HD[1] (mm) |
|---|---|---|---|---|---|
| 256 (mean) | 0.771 ± 0.039 | 0.807 ± 0.042 | 0.788 ± 0.038 | 0.892 ± 0.042 | 6.976 ± 2.405 |
| 256 (median) | 0.829 ± 0.024 | 0.873 ± 0.039 | 0.841 ± 0.037 | 0.970 ± 0.016 | 3.192 ± 0.816 |
| 512 (mean) | 0.768 ± 0.041 | 0.793 ± 0.038 | 0.797 ± 0.038 | 0.890 ± 0.044 | 6.807 ± 2.357 |
| 512 (median) | 0.828 ± 0.024 | 0.854 ± 0.040 | 0.849 ± 0.038 | 0.972 ± 0.013 | 2.919 ± 0.391 |

To visually illustrate the internal validation performance of the segmentation model, samples of overlays of CT and PET images with the outlines of tumor masks using ground truth and model segmentations from the validation data sets are shown in **Fig. 5**. The figure shows representative segmentation results for DSC values of 0.54, 0.77, and 0.96 which are below, comparable, and above the segmentation model's mean DSC of 0.77, respectively.

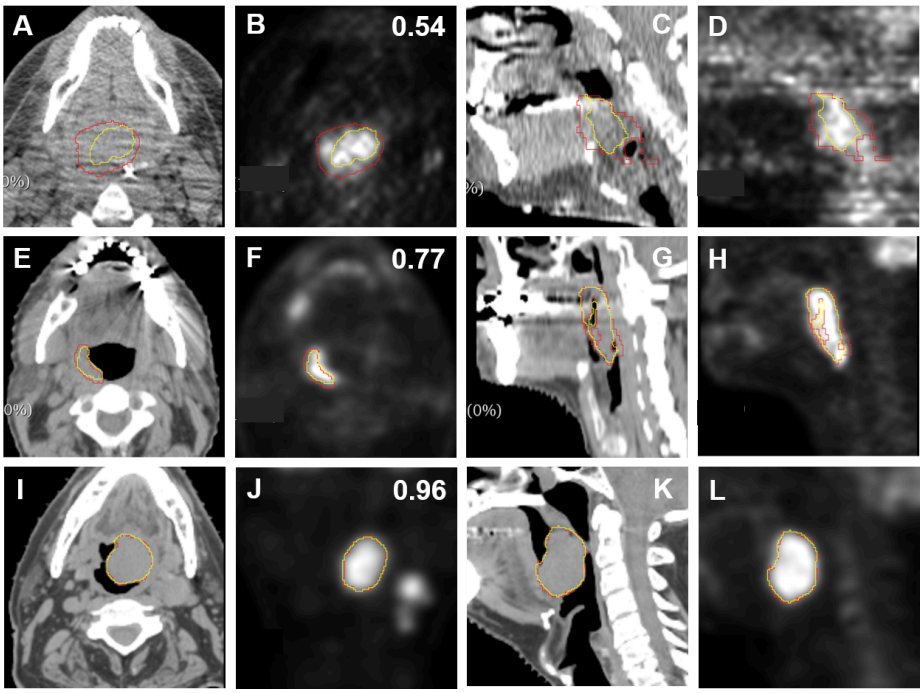


**Fig. 5**. Illustrative examples overlaying the ground truth tumor segmentations (red) and predicted tumor segmentations (yellow) on the CT images (first and third columns) and PET images (second and forth columns) with different 3D volumetric DSC values (below, equivalent, and above the mean estimated DSC value of 0.77) given at the right top corners of the PET images in the second column.

Finally, our models' external validation (test set) performances based on ensembling of cross-validation folds previously described are shown in **Table 2**. Mean DSC and median 95% HD for our best model (256 AVERAGE) was 0.770 and 3.143 mm, respectively (standard deviation or confidence intervals not provided by the HECKTOR 2021 submission portal). Our best model was ranked 8$^{th}$ place in the competition. Compared to the top ranked submission on the HECKTOR 2021 submission portal (pengy), our DSC and 95% HD results are within 0.01 and 0.06 mm, respectively.

**Table 2.** Test set results for ensemble models. Metrics are reported from the HECKTOR 2021 submission portal.

| Model | Mean DSC | Median 95% HD (mm) |
|---|---|---|
| 256 STAPLE | 0.763 | 3.433 |
| 512 STAPLE | 0.759 | 3.291 |
| 256 AVERAGE | 0.770 | 3.143 |

| | | |
|---|---|---|
| 512 AVERAGE | 0.761 | 3.155 |

## 4 Discussion

In this study, we have trained and evaluated OPC primary tumor segmentation models based on the 3D ResUnet deep learning architecture applied to large-scale and multi-institutional PET/CT data from the 2021 HECKTOR Challenge. Moreover, we investigate a variety of architectural modifications (512 vs. 256 channels in bottleneck layer) and ensembling techniques (STAPLE vs. AVERAGE) for test set predictions. Our approaches yield high and consistent segmentation performance in internal validation (cross-validation) and external validation (independent test set), thereby providing further empirical evidence for the feasibility of deep learning-based primary tumor segmentation for fully-automated OPC radiotherapy workflows.

Through internal validation procedures on the training set (10-fold cross-validation), we attain mean DSC, recall, and precision values of 0.771, 0.807, and 0.788 for the 256 model and 0.768, 0.793, and 0.797, for the 512 model, respectively. While the 512 model offers a greater number of channels that could provide greater contextual information, maximum DSC performance is achieved with the 256 model. This may indicate the 256 model led to less over-fitting on the training data evaluation procedure. Interestingly, there was a tradeoff between recall and precision for the two models tested, with the 256 model offering higher recall at the cost of precision compared to the 512 model. Regardless, both these internal validation results improve upon our 3D models implemented in the 2020 HECKTOR Challenge, which only achieved a mean DSC of 0.69 [10]. These improved results potentially highlight the utility and importance of residual connections in a Unet architecture for this task. Moreover, image processing approaches that improve target class balance (i.e., tumor vs non-tumor) in the provided images significantly improve model sensitivity. Finally, we have further investigated the performance of our models using surface distance metrics, as these metrics have been suggested to be more closely linked to clinically meaningful endpoints [23, 24]. We observe minimal differences between the two models for the surface DSC, with both models showing strong performance. However, the 512 model has a slightly lower 95% HD, which may be favorable when more precise tumor boundary definitions are desired.

When models were evaluated on the test data (external validation), we demonstrate high performance consistent with the internal validation results. Generally, the AVERAGE method outperformed the STAPLE method in terms of both DSC and HD. Typically, the AVERAGE method led to more conservative estimates than the STAPLE method, which could indicate ground truth segmentations in the test set tended to be more conservative when considering tumor boundaries. Interestingly, while a tradeoff between DSC and HD exists based on the channel number for the STAPLE method (256 = better DSC, 512 = better HD), this tradeoff is not present with the AVERAGE method, as the 256 model has better DSC and HD compared to the 512 model. Compared to our entry for the 2020 HECKTOR Challenge [10], our mean DSC test results were improved by a sizable degree from the original DSC of

0.637 (0.133 increase for our best model). Moreover, we also improve upon the performance of the winning submission in the 2020 HECKTOR Challenge [25], which achieved a DSC of 0.759 (0.011 increase for our best model). Our positive results may in part be due to the inclusion of ensembling coupled to our improved network modifications (as indicated in the internal validation). The utility of ensembling for PET/CT OPC tumor segmentation has been previously noted since the winning entry in the 2020 challenge used an ensembling approach based on leave-one-center-out cross-validation models to yield the best performing DSC results [25]. Therefore, our results further incentivize the ensembling of model predictions for OPC tumor segmentation data. While our results were not ranked particularly highly within the 2021 HECKTOR leaderboard ($8^{th}$ place), it is worth noting our best model, and most of the models within the top 10-15 entries, scored highly similarly for both DSC and 95% HD. This may indicate a theoretical upper limit on this segmentation task, regardless of model implementations.

In recent years, there has been increasing evidence suggesting the utility of applying deep learning for fully-automated OPC tumor auto-segmentation in various imaging modalities [20, 26–28]. PET/CT has recently shown excellent performance when used as inputs to deep learning models, partly due to the large and highly curated datasets provided by the HECKTOR Challenge [12]. While direct comparison of performance metrics between segmentation studies is often ill-advised, the HECKTOR Challenge offers a systematic method for directly compare segmentation methods with each other. Moreover, since it has been suggested that the mean interobserver DSC for head and neck tumors in human experts is approximately 0.69 [29], our results indicate the potential for further testing to develop auto-segmentation workflows. However, it should be noted that before any definitive statements could be said about the clinical value of an auto-segmentation tool, the dosimetric impact and clinical acceptability of auto-segmeneted structures should be thoroughly evaluated through further studies [24].

One limitation of our study is the reliance of our loss function purely on DSC as an optimization metric. We have chosen the DSC loss since it has provided excellent results in previous investigations and due to its widespread acceptance. However, other loss functions such, as cross entropy [10] and focal loss [25], can be combined with the DSC loss for model optimization which may require further investigation. Moreover, additional measures of spatial similarity, such as surface DSC and 95% HD, are relevant in auto-segmentation for radiotherapy applications [24], and therefore may be attractive candidates for use in model loss optimization [30]. The importance of additional measures of spatial similarity seems to have been noted by the HECKTOR Challenge organizers, as the 95% HD has now become a metric used in the leaderboard to rank contestant performance. An additional limitation of our study is we have only tested a few label fusion approaches as ensembling techniques for our models. For example, we have selected STAPLE as a label fusion method because of its general ubiquity and widely available implementations. However, STAPLE has been criticized in the past [31]; therefore, additional label fusion approaches may be necessary to test in this framework [32]. Moreover, for the AVERAGE ensembling method, the specific threshold in the number of

cross-validation models used to determine final label fusion can be seen as an additional parameter to tune. While we have chosen a 5-model threshold as a proxy for majority voting, alternative thresholding strategies can lead to more conservative or liberal estimates of tumor segmentation.

## 5 Conclusion

This study presented the development and validation of deep learning models using a 3D ResUnet architecture to segment OPC primary tumors in an end-to-end automated workflow based on PET/CT images. Using a combination of pre-processing steps, architectural design decisions, and model ensembling approaches, we achieve promising internal and external validation segmentation performance, with external validation mean DSC and median 95% HD of 0.770 and 3.143 mm, respectively, for our best model. Our method notably improves upon our previous iteration of our model submitted in the 2020 HECKTOR Challenge. Future studies should seek to further optimize these methods for improved OPC tumor segmentation performance in forthcoming iterations of the HECKTOR Challenge.

**Acknowledgements.** M.A.N. is supported by a National Institutes of Health (NIH) Grant (R01 DE028290-01). K.A.W. is supported by a training fellowship from The University of Texas Health Science Center at Houston Center for Clinical and Translational Sciences TL1 Program (TL1TR003169), the American Legion Auxiliary Fellowship in Cancer Research, and a NIDCR F31 fellowship (1 F31 DE031502-01). C.D.F. received funding from the National Institute for Dental and Craniofacial Research Award (1R01DE025248-01/R56DE025248) and Academic-Industrial Partnership Award (R01 DE028290), the National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant (NSF 1557679), the NIH Big Data to Knowledge (BD2K) Program of the National Cancer Institute (NCI) Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (1R01CA214825), the NCI Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program (1R01CA218148), the NIH/NCI Cancer Center Support Grant (CCSG) Pilot Research Program Award from the UT MD Anderson CCSG Radiation Oncology and Cancer Imaging Program (P30CA016672), the NIH/NCI Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award (P50 CA097007) and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) Research Education Program (R25EB025787). He has received direct industry grant support, speaking honoraria and travel funding from Elekta AB.

## References

1. Hay, A., Nixon, I.J.: Recent advances in the understanding and management of oropharyngeal cancer. F1000Research. 7, (2018).
2. Njeh, C.F.: Tumor delineation: The weakest link in the search for accuracy in

radiotherapy. J. Med. physics/Association Med. Phys. India. 33, 136 (2008).
3. Foster, B., Bagci, U., Mansoor, A., Xu, Z., Mollura, D.J.: A review on segmentation of positron emission tomography images. Comput. Biol. Med. 50, 76–96 (2014).
4. Segedin, B., Petric, P.: Uncertainties in target volume delineation in radiotherapy–are they relevant and what can we do about them? Radiol. Oncol. 50, 254–262 (2016).
5. Rasch, C., Steenbakkers, R., van Herk, M.: Target definition in prostate, head, and neck. In: Seminars in radiation oncology. pp. 136–145. Elsevier (2005).
6. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Med. Image Anal. 63, 101693 (2020).
7. Zhou, T., Ruan, S., Canu, S.: A review: Deep learning for medical image segmentation using multi-modality fusion. Array. 3, 100004 (2019).
8. Bakator, M., Radosav, D.: Deep learning and medical diagnosis: A review of literature. Multimodal Technol. Interact. 2, 47 (2018).
9. Naser, M.A., Deen, M.J.: Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images. Comput. Biol. Med. 121, 103758 (2020). https://doi.org/10.1016/j.compbiomed.2020.103758.
10. Naser, M.A., van Dijk, L. V, He, R., Wahid, K.A., Fuller, C.D.: Tumor segmentation in patients with head and neck cancers using deep learning based-on multi-modality PET/CT images. In: 3D Head and Neck Tumor Segmentation in PET/CT Challenge. pp. 85–98. Springer (2020).
11. AIcrowd MICCAI 2020: HECKTOR Challenges.
12. Andrearczyk, V., Valentin, O., Mario, J., Vallières, M., Castelli, J., Elhalawani, H., Boughdad, S., Prior, J.O., Depeursinge, A.: Overview of the HECKTOR challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT.
13. Vincent Andrearczyk, Valentin Oreiller, Sarah Boughdad, Catherine Chez Le Rest, Hesham Elhalawani, Mario Jreige, John O. Prior, Martin Vallières, Dimitris Visvikis, Mathieu Hatt, A.D.: Overview of the HECKTOR challenge at MICCAI 2021: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT images. LNCS challenges. (2021).
14. Valentin Oreiller et al.: Head and Neck Tumor Segmentation in PET/CT: The HECKTOR Challenge. Med. Image Anal. (2021).
15. The MONAI Consortium: Project MONAI, (2020). https://doi.org/http://doi.org/10.5281/zenodo.4323059.
16. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology. 26, 297–302 (1945).
17. Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B.: Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. J. Med. Internet Res. 23, e26151 (2021).
18. Taha, A.A., Hanbury, A.: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging. 15, 1–28 (2015).
19. Blinde, S., Mohamed, A.S.R., Al-Mamgani, A., Newbold, K., Karam, I., Robbins, J.R., Thomson, D., Fuller, C.D., Raaijmakers, C.P., Terhaard, C.: Large interobserver variation in the international MR-LINAC oropharyngeal carcinoma delineation study. Int. J. Radiat. Oncol. Biol. Phys. 99, E639–E640 (2017).
20. Wahid, K.A., Ahmed, S., He, R., van Dijk, L. V, Teuwen, J., McDonald, B.A., Salama, V., Mohamed, A.S.R., Salzillo, T., Dede, C., Taku, N., Lai, S.Y., Fuller, C.D., Naser, M.A.: 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. 32, 6–14 (2022). https://doi.org/https://doi.org/10.1016/j.ctro.2021.10.003.

21. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging. 23, 903–921 (2004).
22. Zhou, Z.-H.: Ensemble learning. In: Machine Learning. pp. 181–210. Springer (2021).
23. Kiser, K.J., Barman, A., Stieb, S., Fuller, C.D., Giancardo, L.: 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. 1–13 (2021).
24. Sherer, M. V, Lin, D., Elguindi, S., Duke, S., Tan, L.-T., Cacicedo, J., Dahele, M., Gillespie, E.F.: Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiother. Oncol. (2021).
25. Iantsen, A., Visvikis, D., Hatt, M.: Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined PET and CT images. In: 3D Head and Neck Tumor Segmentation in PET/CT Challenge. pp. 37–43. Springer (2020).
26. Outeiral, R.R., Bos, P., Al-Mamgani, A., Jasperse, B., Simões, R., van der Heide, U.A.: Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning. Phys. imaging Radiat. Oncol. 19, 39–44 (2021).
27. Andrearczyk, V., Oreiller, V., Vallières, M., Castelli, J., Elhalawani, H., Jreige, M., Boughdad, S., Prior, J.O., Depeursinge, A.: Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans. In: Medical Imaging with Deep Learning. pp. 33–43. PMLR (2020).
28. Moe, Y.M., Groendahl, A.R., Mulstad, M., Tomic, O., Indahl, U., Dale, E., Malinen, E., Futsaether, C.M.: Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers. arXiv Prepr. arXiv1908.00841. (2019).
29. Gudi, S., Ghosh-Laskar, S., Agarwal, J.P., Chaudhari, S., Rangarajan, V., Nojin Paul, S., Upreti, R., Murthy, V., Budrukkar, A., Gupta, T.: 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, 184–192 (2017). https://doi.org/10.1016/j.jmir.2016.11.003.
30. Karimi, D., Salcudean, S.E.: Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. IEEE Trans. Med. Imaging. 39, 499–513 (2019).
31. Van Leemput, K., Sabuncu, M.R.: A cautionary analysis of staple using direct inference of segmentation truth. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 398–406. Springer (2014).
32. Robitaille, N., Duchesne, S.: Label fusion strategy selection. Int. J. Biomed. Imaging. 2012, (2012).