



A convolutional neural network with self-attention for fully automated metabolic tumor volume delineation of head and neck cancer in [¹⁸F] FDG PET/CT

Pavel Nikulin¹ · Sebastian Zschaecck^{2,3} · Jens Maus¹ · Paulina Cegla⁴ · Elia Lombardo⁵ · Christian Furth⁷ · Joanna Kaźmierska^{10,11} · Julian M. M. Rogasch^{3,7} · Adrien Holzgreve⁸ · Nathalie L. Albert⁸ · Konstantinos Ferentinos⁹ · Iosif Strouthos⁹ · Marina Hajjianni^{2,3} · Sebastian N. Marschner⁵ · Claus Belka^{5,6} · Guillaume Landry⁵ · Witold Cholewinski^{4,10} · Jörg Kotzerke¹² · Frank Hofheinz¹ · Jörg van den Hoff^{1,12}

Received: 14 November 2022 / Accepted: 14 March 2023 / Published online: 20 April 2023
© The Author(s) 2023, corrected publication 2023

Abstract

Purpose PET-derived metabolic tumor volume (MTV) and total lesion glycolysis of the primary tumor are known to be prognostic of clinical outcome in head and neck cancer (HNC). Including evaluation of lymph node metastases can further increase the prognostic value of PET but accurate manual delineation and classification of all lesions is time-consuming and prone to interobserver variability. Our goal, therefore, was development and evaluation of an automated tool for MTV delineation/classification of primary tumor and lymph node metastases in PET/CT investigations of HNC patients.

Methods Automated lesion delineation was performed with a residual 3D U-Net convolutional neural network (CNN) incorporating a multi-head self-attention block. 698 [¹⁸F]FDG PET/CT scans from 3 different sites and 5 public databases were used for network training and testing. An external dataset of 181 [¹⁸F]FDG PET/CT scans from 2 additional sites was employed to assess the generalizability of the network. In these data, primary tumor and lymph node (LN) metastases were interactively delineated and labeled by two experienced physicians. Performance of the trained network models was assessed by 5-fold cross-validation in the main dataset and by pooling results from the 5 developed models in the external dataset. The Dice similarity coefficient (DSC) for individual delineation tasks and the primary tumor/metastasis classification accuracy were used as evaluation metrics. Additionally, a survival analysis using univariate Cox regression was performed comparing achieved group separation for manual and automated delineation, respectively.

Results In the cross-validation experiment, delineation of all malignant lesions with the trained U-Net models achieves DSC of 0.885, 0.805, and 0.870 for primary tumor, LN metastases, and the union of both, respectively. In external testing, the DSC reaches 0.850, 0.724, and 0.823 for primary tumor, LN metastases, and the union of both, respectively. The voxel classification accuracy was 98.0% and 97.9% in cross-validation and external data, respectively. Univariate Cox analysis in the cross-validation and the external testing reveals that manually and automatically derived total MTVs are both highly prognostic with respect to overall survival, yielding essentially identical hazard ratios (HR) ($HR_{\text{man}} = 1.9$; $p < 0.001$ vs. $HR_{\text{cnn}} = 1.8$; $p < 0.001$ in cross-validation and $HR_{\text{man}} = 1.8$; $p = 0.011$ vs. $HR_{\text{cnn}} = 1.9$; $p = 0.004$ in external testing).

Conclusion To the best of our knowledge, this work presents the first CNN model for successful MTV delineation and lesion classification in HNC. In the vast majority of patients, the network performs satisfactory delineation and classification of primary tumor and lymph node metastases and only rarely requires more than minimal manual correction. It is thus able to massively facilitate study data evaluation in large patient groups and also does have clear potential for supervised clinical application.

P. Nikulin and S. Zschaecck contributed equally to this article.

This article is part of the Topical Collection on Advanced Image Analyses (Radiomics and Artificial Intelligence).

✉ Pavel Nikulin
p.nikulin@hzdr.de

Extended author information available on the last page of the article

Keywords FDG PET · Metabolic tumor volume · MTV · Head and neck cancer · HNC · Convolutional neural network

Introduction

Primary treatment approaches for localized head and neck cancer (HNC) include either definitive radiochemotherapy or surgery. The latter is often followed by adjuvant radiotherapy or radiochemotherapy. Treatment related side effects are considerable and differ between both primary treatment approaches, as shown by the randomized ORATOR trial in oropharyngeal carcinomas [1]. For primary radiochemotherapy, radiosensitivity differs considerably between individual patients and local tumor recurrences remain an important clinical issue. Biomarkers for an improved personalized treatment include quantitative PET parameters, notably metabolic tumor volume (MTV), total lesion glycolysis, and SUV_{max} of the primary tumor which have been shown to be prognostic of clinical outcome in patients with HNC [2–6]. Evaluation of lymph node (LN) metastases in addition to the primary tumor has potential to further increase the prognostic value of PET [7]. Such analysis requires, however, accurate delineation and classification of all lesions which is very time-consuming when performed manually. Additionally, the tumor volumes can be prone to interobserver variability which hampers reproducibility of the results. The problem of accelerating tumor delineation in PET has previously been addressed by several groups using semi-automated methods such as fixed or adaptive thresholding, fuzzy locally adaptive Bayesian segmentation, region growing method, etc. [8]. Such approaches provide satisfactory results at sufficiently high target to background contrast but become increasingly more inaccurate with decreasing contrast which makes manual intervention regularly necessary. Consequently, notable time demands are imposed on the user, especially for LN delineation.

The recent emergence of deep learning-based methods for medical image analysis [9–13] allowed for significant progress in the tasks of therapy response [14–19] and clinical outcome [20–28] prediction, image registration [29–34], exam and object classification [35–46], object detection [47–54], and, finally, object delineation [55–68]. More specifically, the approaches to HNC cancer lesion delineation mostly rely on similar U-Net-like architectures but differ regarding the choice of target volume definition, considered patient population, and employed imaging modalities. Some researchers have exclusively considered the morphological modalities, i.e. CT [69–74] and MRI [75–79] or a combination of both [80], with Dice similarity coefficients (DSCs) reaching 0.74 for primary tumor and 0.66 for LN metastases in CT and 0.65 for primary

tumor and 0.58 for LN metastases in MRI. For the special case of MRI in nasopharyngeal cancer a much higher DSC of up to 0.90 has been reported [78]. Furthermore, many studies report that combining CT or MRI with PET improves the network's performance considerably [70–74, 81] compared to usage of only a single modality. The majority of state-of-the-art designs utilizes PET/CT which has been shown to be slightly superior to PET/MR [80] and also is much more widely available. Examples include primary tumor delineation in oropharyngeal cancer [72, 82, 83] (DSC = 0.61), primary tumor + LN metastases delineation in squamous cell carcinoma of the oral cavity, oropharynx, hypopharynx and larynx [73, 74, 80] (DSC = 0.75), and primary + LN metastases delineation in a non-specified HNC [70, 71, 81] (DSC = 0.82). The number of proposed solutions to the problem at hand increased drastically with creation of the HECKTOR challenge aiming on primary gross tumor volume (GTV) delineation in oropharyngeal cancer using a substantial PET/CT database [82]. The most recent challenge included contributions from 20 teams scoring DSCs of [0.63–0.78] [84]. Interestingly, despite a large variety of proposed solutions, the winning contribution relied on the well known 3D U-Net architecture with only minimal modifications [85].

It is important to emphasize that the above-mentioned studies aimed at GTV rather than MTV delineation and that these two volumes are not identical in general. MTV is mainly used in a diagnostic context and for therapy response assessment. Therefore, sensitivity and specificity should be well balanced to avoid overdiagnosis. In contrast, GTV is mainly utilized in radiotherapy planning where a higher sensitivity might be preferred at the expense of a lower specificity to reduce the risks of underdosage of malignant lesions. As a consequence, MTV might possess higher prognostic power compared to GTV [86]. So far, only a single study investigated the CNN-based fully automated MTV (PET-based GTV) delineation in HNC [87]. The authors considered multiple CNN architectures and loss functions in a population of 470 patients achieving DSC = 0.87, demonstrating generally comparable performance with different configurations. However, to the best of our knowledge the possibility of automated differentiation between primary tumors and LN metastases has not been thoroughly investigated so far. Therefore, our goal was development of an automated tool for MTV delineation and classification of primary tumor and lymph node metastases in HNC in PET/CT. Additionally, our aim was to compare the manually and CNN derived PET parameters

regarding outcome prediction of patients in an independent external cohort of patients.

Methods

Patients and data acquisition

1133 patients available from a retrospective cohort of an ongoing clinical multicenter investigation [4] were considered for inclusion in the present study. Exclusion occurred as follows:

- no CT data sets: $N = 165$
- severe metal artifacts in the CT data sets: $N = 45$
- no sizable [^{18}F]FDG uptake, lesion identification/delineation in PET thus impossible: $N = 44$

Ultimately, 879 patients could thus be included in the current study.

The data were split into a main dataset used for training, validation, and testing and a dataset for external testing using only data from sites which were not included in network model generation. The main dataset consisted of 698 [^{18}F]FDG PET/CT scans of head and neck squamous cell carcinoma (HNSCC) patients (535 men and 163 women, mean age 61 years, range 25–87) from three clinical sites (Berlin, Germany ($N = 175$), Dresden, Germany ($N = 24$), Poznan, Poland ($N = 22$)) and 5 public databases (Data from Head-Neck-PET-CT [88] ($N = 269$), Data from Head-Neck-Radiomics-HN1 [89] ($N = 34$), Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy [90] ($N = 32$), Radiology Data from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma [TCGA-HNSC] collection [91] ($N = 11$), Data From QIN-HEADNECK [92] ($N = 131$)). The dataset for external testing included $N = 15$ patients from Limassol, Cyprus and $N = 166$ patients from Munich, Germany (138 men and 43 women, mean age 63 years, range 28–89).

The main dataset included 643 primary tumors and 1078 LN metastases with mean (median) volumes of 13.21 (8.16) ml and 5.45 (2.74) ml, respectively. The external dataset contained 175 primary tumors and 397 LN metastases with mean (median) volumes of 15.62 (8.05) ml and 4.30 (0.70) ml, respectively. The most frequent localizations of primary tumor in the two datasets (main/external, respectively) were oropharynx (63%/41%), larynx (17%/17%), oral cavity (6%/19%), hypopharynx (6%/14%), and nasopharynx (6%/6%). The majority of the patients were staged UICC IV (69%/67%). Details on the respectively utilized PET/CT systems, data acquisition, and image reconstruction can be found in Supplementary Materials as well as in [4] and citations therein.

Ground truth definition

The interactive lesion delineations performed in the context of the above-mentioned multicenter investigation served as ground truth for network training and evaluation. For this delineation, the metabolically active areas of, both, primary tumor and lymph node metastases were identified in the PET data by a semi-automatic algorithm based on adaptive thresholding considering the local background [93, 94] using the ROVER software (version 3.0.41; ABX GmbH, Radeberg, Germany). Each proposed region of interest (ROI) delineation was individually verified by one of two experienced observers and manually corrected (also using the ROVER software) where this was deemed necessary. For the primary tumor manual correction was required in 41 out of 879 patients. Manual correction was necessary more frequently in lymph nodes (716 out of 1475 lesions). The majority of corrections concerned lesions with diffuse low tracer accumulation. Furthermore, in a few patients where tumor and lymph nodes were in close vicinity, the ROVER algorithm was not able to generate separate ROIs and erroneously fused the neighboring lesions in a single ROI.

Network architecture, data preprocessing, and training procedure

Automated lesion delineation was performed with a residual 3D U-Net CNN modified by inclusion of a Multi-Head Self-Attention (MHSA) block [95] at the bottom of the U-Net in order to improve global context awareness during lesion classification. More details on the CNN design are provided in Supplementary Materials. The proposed architecture was implemented using the Apache MXNet (version 1.9.0) package for the R language and environment for statistical computing (version 4.2.0) [96].

The network was trained using pairs of PET and CT volumes as input. The data were pre-processed as follows. First, all image volumes were resampled to a common voxel size of $2.5 \times 2.5 \times 2.5$ mm and centrally cropped to a matrix size of 128×128 in the transaxial plane (corresponding to 32×32 cm field coverage). A further variable axial crop was performed preserving the head and neck region of the respective PET/CT image volume. In the next step, image patches of size $128 \times 128 \times 32$ were extracted with a partial overlap of at least 75% in axial direction. After windowing the CT intensity values to a range of $[-150, 150]$ HU, PET and CT volumes were individually normalized to the range $[0, 1]$. The ground truth delineations were encoded into one-hot format for the three classes — background, primary tumor, and LN — using a voxel grid matching the one holding the PET/CT data as described above. The whole process results in a total of 9535 data samples in the main dataset.

A 5-fold cross-validation scheme with 5 equally sized folds was employed in order to assess the network performance in all data in the main dataset. For each of the 5 training runs, 64% of the data were assigned for training, 16% for validation, and 20% for testing, respectively. Network training was performed for 200 epochs with the *Adadelta* optimizer (batch size = 16) using Dice + Cross-Entropy as loss function. The training process was monitored by calculating the soft DSC in the validation data. More details on the loss function and evaluation metric as well as the training logs are provided in Supplementary Materials. Training was stopped if no improvement in the evaluation metric was recorded for 30 epochs. The 5 models achieving the highest scores in the respective validation data were selected for further evaluation.

Network evaluation

Each of the five resulting CNN models was used to predict primary tumor and LN metastases probability maps in its respective test subset of the main dataset. Since the network does not directly predict the probability for the whole image but only for image patches, the predictions for the entire image volume was derived from the predictions in the separate patches by calculating the output probability in the overlap areas as a weighted sum of probabilities. The weights were chosen to be 3D Gaussian with the respective full-widths at half-maximum equal to the patch side lengths. Such weighting is based on our empirical finding that the CNN predictions are more reliable in the center of the patch. Each voxel was assigned to a class (background: 0, primary tumor: 1, LN metastases: 2) according to the highest probability in the derived maps. Accordingly, the union of all voxels with class 1 (class 2) defines the primary tumor (LN metastases) ROI. ROIs with volumes < 0.1 ml (both, manually and automatically delineated) were excluded from further analysis. Finally, the predictions obtained in the disjunct test data of the different folds were pooled, i.e., the complete available data set was considered for further analysis of network performance rather than analyzing in turn each of the folds separately. The complete data analysis for the present investigation was performed on the above-mentioned image volumes resampled to cubic (2.5 mm)³ voxels that were processed by the CNN. It should be noted that for possible applications of the network beyond MTV determination (notably in the context of radiation treatment planning), it would be necessary to transform the CNN outputs back to the original voxel grid prior to the voxel class assignment.

The evaluation was additionally performed in an external test dataset to assess the capability of the network to generalize to data from so far unseen sources. Separate runs were performed with all 5 CNN models and class membership

was determined using probability maps obtained by averaging of the individual model outputs.

Spatial concordance

The spatial concordance between manual and automatic delineations was quantified using the DSC for primary tumor, LN metastases, and the union of all lesions representing the total tumor burden (TTB), respectively. We calculated, both, cohort DSC (determined for the union of all delineations across all patients) as well as individual DSC (determined for each patient) together with mean and median values of the corresponding distributions. Furthermore, the mean absolute difference between manual and automated TTB delineations as well as the corresponding correlation coefficient were computed. In these calculations, 1% of the data exhibiting the highest absolute TTB differences were rejected to reduce the influence of outliers.

Classification capabilities

The network's capability to distinguish between primary tumor and LN metastases was assessed by considering the subset of voxels included in both manual and CNN delineations. Voxels of primary tumor and LN metastases (as defined in manual delineation) which were correctly classified by the CNN were counted as true primary tumor (T_{PT}) and true LN (T_{LN}), respectively. Primary tumor voxels classified as LN metastases and LN metastases voxels classified as primary tumor were counted as false LN (F_{LN}) and false primary (F_{PT}), respectively. Classification performance was quantified by the true positive rate of "primary tumor" labeled voxels $TPR_{PT} = T_{PT}/(T_{PT} + F_{LN})$, the corresponding true positive rate $TPR_{LN} = T_{LN}/(T_{LN} + F_{PT})$ of "lymph node metastasis" labeled voxels, and the classification accuracy $ACC = (T_{PT} + T_{LN})/(T_{PT} + F_{PT} + T_{LN} + F_{LN})$.

The analysis was performed with the R language and environment for statistical computing (version 4.2.0) [96].

Structure-wise analysis

The ability of the network to identify and classify individual lesions was assessed via structure-wise analysis as proposed in [74]. Shortly, for each lesion in the ground truth and CNN delineation, a coverage fraction by the complementary delineation was calculated. A ground truth lesion was considered as identified (true positive with respect to manual delineation, TP_{man}) if it was at least 50% covered by the CNN delineation and as missed by the CNN (false negative, FN) if coverage was below 50%. CNN delineated structures with coverage over 50% were considered true positive with respect to CNN delineation (TP_{cnn}) and the

remaining CNN delineations were not corresponding to a ground truth lesion and were therefore considered false positives (FP). Based on this classification, true positive rate $TPR_{str} = TP_{man}/(TP_{man} + FN)$ and positive predictive value $PPV_{str} = TP_{cnn}/(TP_{cnn} + FP)$ were calculated.

Survival analysis

We also investigated the impact of the differences between manual and automated delineation on a survival analysis of the patient data. The full survival analysis of these data is the objective of the above mentioned still ongoing clinical study. In the present investigation, we therefore only have exemplarily considered the prognostic value of TTB for overall survival (OS). All patients satisfying the following criteria were included into this analysis: primary chemoradiotherapy, [^{18}F]FDG PET/CT prior to therapy, minimum follow up time of 6 months, no distant metastases, and no surgery ($N = 585$ patients from 10 institutions in main dataset; $N = 142$ patients from 2 institutions in external test dataset). The median TTB of the manual delineation in the cross-validation data was used as cutoff value for differentiating high and low risk groups in, both, manual as well as CNN delineated lesions (TTB_{man} and TTB_{cnn} , respectively). The same cutoff was also used in analysis of the external data. For all delineations, an univariate Cox regression and a Kaplan-Meier analysis was performed. Results with $p < 0.05$ were considered significant.

Results

Spatial concordance

A summary of the delineation performance is given in Table 1. In the cross-validation experiment, delineation of all malignant lesions with the proposed network achieves a cohort DSC of 0.870 when not discriminating between primary tumor and lymph nodes. Treating primary tumor and lymph node metastases as distinct classes yields cohort DSCs of 0.885 and 0.805, respectively. In the external test data, cohort DSC reaches 0.850, 0.724, and 0.823 for primary tumor, LN metastases and their union, respectively. The frequency distributions of the individual respective DSCs obtained in different patients is given in Fig. 1. The delineation failed (DSC = 0) in 5 cases (0.7%) when not discriminating between primary tumor and lymph nodes. When treating primary tumor and LN metastases as distinct, delineation failed in 48 (6.9%) and 65 (9.3%) cases, respectively. In the external dataset, the delineation failed in 9 (5.0%) and 27 (14.9%) cases when treating primary tumor and LN metastases as distinct, respectively, and it never failed for the union.

Table 1 Delineation performance with respect to the target volume in cross-validation and external testing

Target volume	DSC				N failed (DSC = 0)
	Cohort	Mean	Median	50% CI	
Cross-validation ($N = 698$)					
Primary tumor	0.885	0.815	0.924	[0.856, 0.948]	48
LN metastases	0.805	0.750	0.871	[0.688, 0.948]	65
Primary + metastases	0.870	0.840	0.894	[0.814, 0.929]	5
External testing ($N = 181$)					
Primary tumor	0.850	0.805	0.896	[0.827, 0.926]	9
LN metastases	0.724	0.622	0.756	[0.307, 0.930]	27
Primary + metastases	0.823	0.808	0.866	[0.803, 0.909]	0

Figure 2 demonstrates the degree of correlation between the manually and automatically derived TTB in, both, cross-validation and external testing experiments ($R^2 = 0.95$ and $R^2 = 0.85$, respectively, excluding the outliers). The mean absolute TTB difference was 2.62 ml and 4.29 ml in cross-validation and external testing data, respectively.

Classification capabilities

In cross-validation data, the overall classification accuracy was 98.0%. 640 of the scans (91.7%) did not exhibit any classification errors. In external data, the classification accuracy was 97.9% with 147 of the scans (81.2%) free of any classification errors. The corresponding true positive rates for primary and LN metastases classification and the full contingency tables are given in Table 2.

Structure-wise analysis

In cross-validation data, in 44/252 cases non-pathological uptake was delineated and marked as primary tumor/LN metastases (false positives). 65 primaries and 229 LNs were not recognized by the network (false negatives). In external data (primary/LN), 10/53 false positive delineations were produced and 11/204 lesions were missed by the CNN. The complete statistics as well as the network's PPV_{str} and TPR_{str} is provided in Table 3.

Examples

Figure 3 shows exemplary delineations of four patients (A-D) from the cross-validation test subset. Patient A demonstrates that the trained CNN is able to accurately determine the contours of both primary tumor and LN metastasis and to correctly classify them as such. Patient B

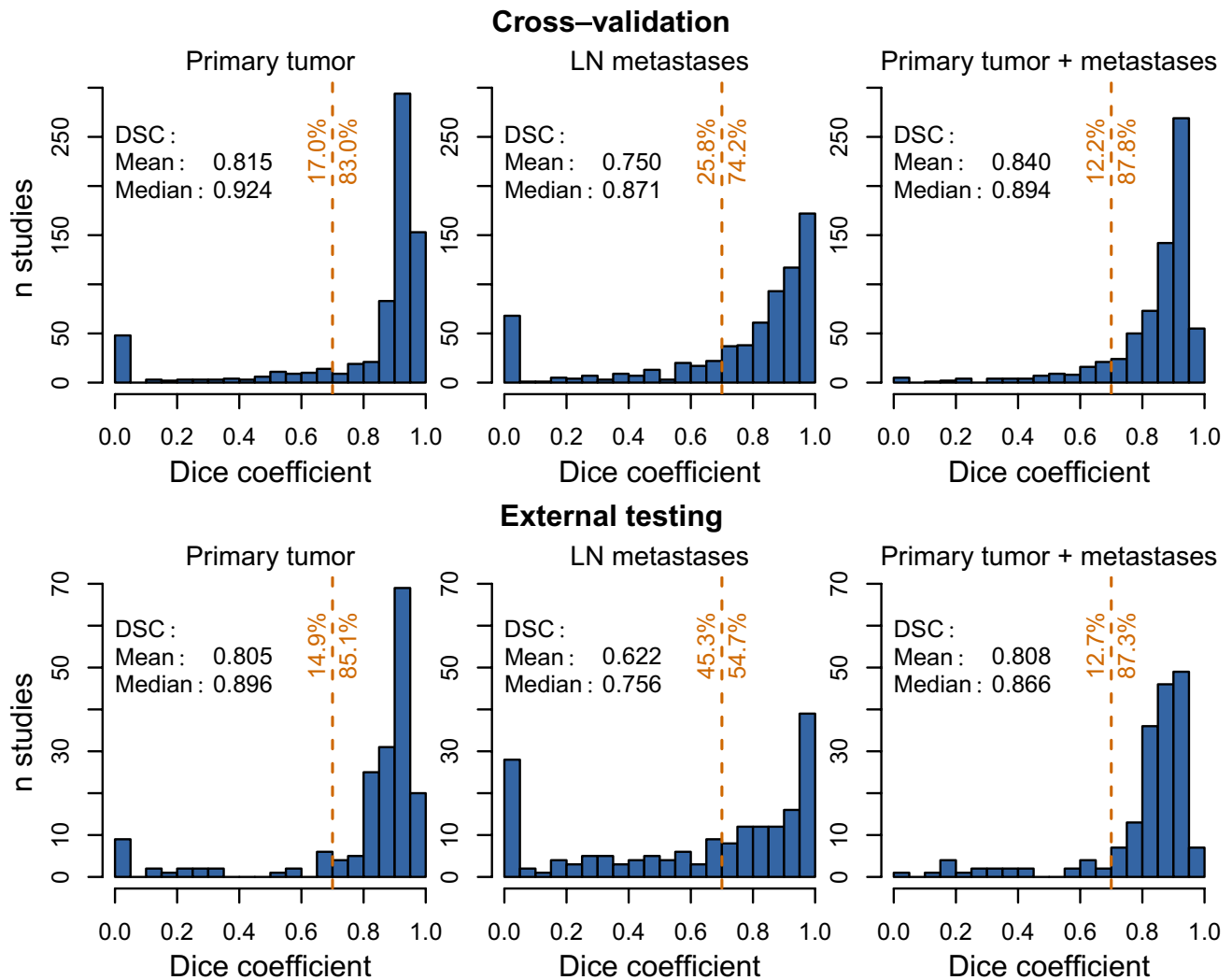


Fig. 1 Frequency distribution of the observed Dice coefficients (CNN vs. manual delineation/labeling) for primary tumor (left), LN metastases (middle), and the union of both (right) in cross-validation (top, $N = 698$ patients) and external testing data (bottom, $N = 181$ patients). The dashed vertical line indicates the location of a

DSC = 0.7 threshold, a value which might be considered acceptable for practical use. The numbers to the left and to the right of the line specify the percentage of cases yielding a DSC below and above that threshold, respectively

demonstrates the ability of the CNN to delineate relatively large lesions. Note that the human observer and the CNN consistently excluded the necrotic core of the tumor from MTV delineation. Patient C demonstrates that the CNN is able to delineate multiple structures exhibiting different relative contrast ($SUV_{max} = 22.0$ vs $SUV_{max} = 7.4$ for primary tumor and LN metastases, respectively) simultaneously. Patient D illustrates the CNN's capability to distinguish small lesions with low uptake ($SUV_{max} = 4.0$) from comparable physiological focal uptake in other regions of the same image.

Figure 4 shows example cases where CNN delineation failed for some lesions. In example A, the LN metastasis

was misclassified as primary tumor. In example B, the CNN produced a spurious LN metastasis ROI. In examples C and D, the CNN missed the primary tumor and LN metastasis, respectively.

Survival analysis

Univariate Cox regression in the cross-validation data consistently revealed TTB_{man} as well as TTB_{cnn} as highly prognostic factors for OS with practically identical hazard ratios (HR) ($HR = 1.9$; $p < 0.001$ and $HR = 1.8$; $p < 0.001$, respectively). In 5.3% of the cases binarization led to a different classification, where classification was incorrect in

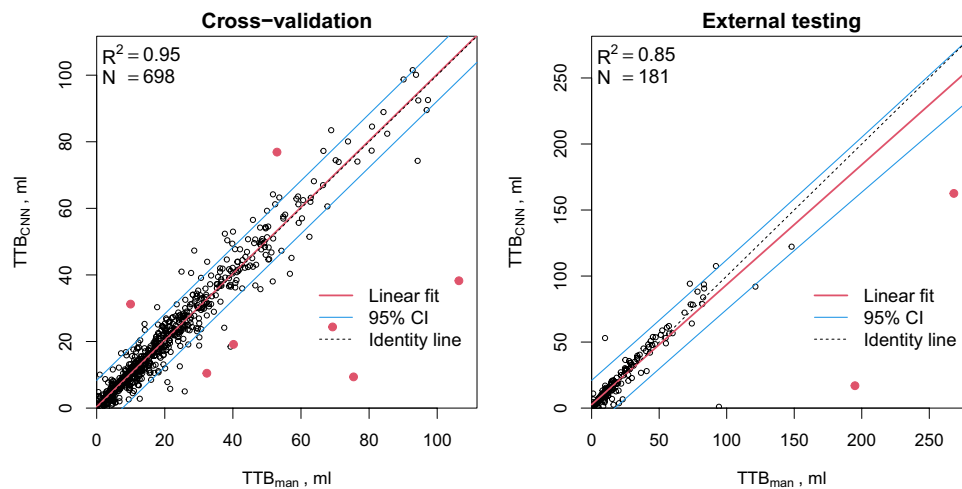


Fig. 2 Correlation between manually and automatically derived total tumor burden (TTB: sum of primary tumor and LN metastases) in the cross-validation (left) and external testing (right) data. Note the difference in scale between the plots. Solid red points indicate outliers, defined as data points where the deviation of CNN from manual delineation exceeds the 99% percentile (i.e., the top 1%). These outli-

ers were excluded from regression analysis. The red line represents the least squares fit of a straight line to the remaining data. The blue lines delineate the corresponding 95% prediction (tolerance) interval of expected scatter of individual data points around the regression line

Table 2 Classification performance in cross-validation and external testing. The contingency tables are normalized so that the sum over the respective table’s elements equals 1 (100%)

Dataset	Contingency table				TPR _{PT}	TPR _{LN}	ACC
	T _{PT}	F _{PT}	T _{LN}	F _{LN}			
Cross-validation (801384 voxels)	60.8%	1.5%	37.2%	0.5%	99.2%	96.1%	98.0%
External testing (229448 voxels)	67.0%	1.2%	30.9%	0.9%	98.7%	96.3%	97.9%

Table 3 Lesion detection performance with respect to the target volume in cross-validation and external testing

Target volume	TP _{man}	TP _{cnn}	FN	FP	TPR _{str}	PPV _{str}
Cross-validation (N = 698)						
Primary tumor	578	592	65	44	89.9%	93.1%
LN metastases	849	893	229	252	78.8%	78.0%
Primary + metastases	1427	1485	294	296	82.9%	83.4%
External testing (N = 181)						
Primary tumor	164	160	11	10	93.7%	94.1%
LN metastases	193	198	204	53	48.6%	78.9%
Primary + metastases	357	358	215	63	62.4%	85.0%

2.6% for TTB_{man} and in 2.7% for TTB_{cnn}. Both, TTB_{man} and TTB_{cnn} prognostic factors also reached significance in external data and exhibited virtually the same hazard ratios as in the cross-validation dataset (HR = 1.8; *p* = 0.011 and HR = 1.9; *p* = 0.004, respectively). The fraction of cases resulting in different classification was higher in the external dataset than in the cross-validation one reaching 7.7%, where classification was incorrect in 5.6% for TTB_{man} and in 2.1% for TTB_{cnn}. The corresponding Kaplan-Meier curves are shown in Fig. 5.

Discussion

In this investigation we have demonstrated that fully automated simultaneous delineation and classification of metabolically active lesions in HNC, discriminating between primary tumor and lymph node metastases, is feasible with a suitable CNN architecture trained on combined PET/CT patient data. The achieved cohort DSC of 0.870 indicates state of the art performance of our network and is in line with the mean DSC of 0.87 reported in [87] for

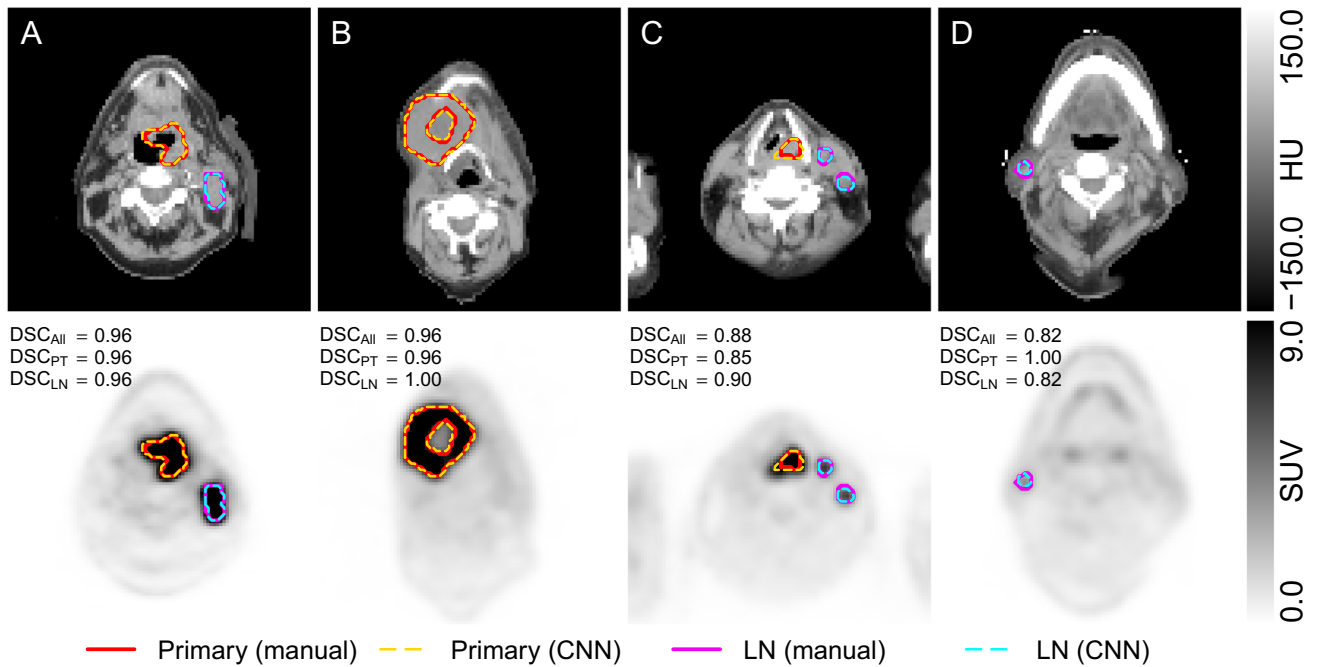


Fig. 3 Manual and CNN-based delineations of primary tumor and lymph node metastases in 4 selected patients. Relevant transaxial PET/CT slices are shown (top: CT, bottom: PET). The dice coefficients (in the presented plane) for primary tumor (DSC_{PT}), LN metastases (DSC_{LN}), and their union (DSC_{All}) are indicated. Patient A:

oropharyngeal cancer with LN metastasis; patient B: oropharyngeal cancer with necrotic core; patient C: hypopharyngeal cancer with 2 LN metastases; patient D: cancer of oral cavity (not visible in this slice) exhibiting a low uptake LN metastasis

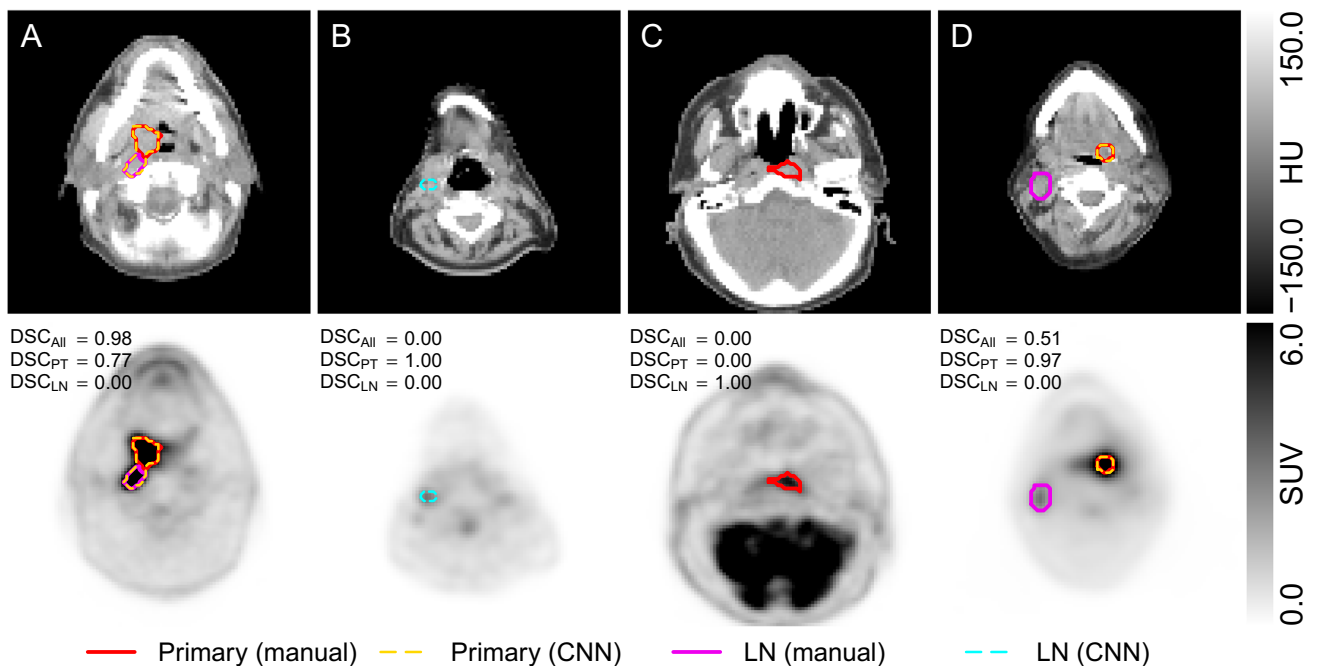
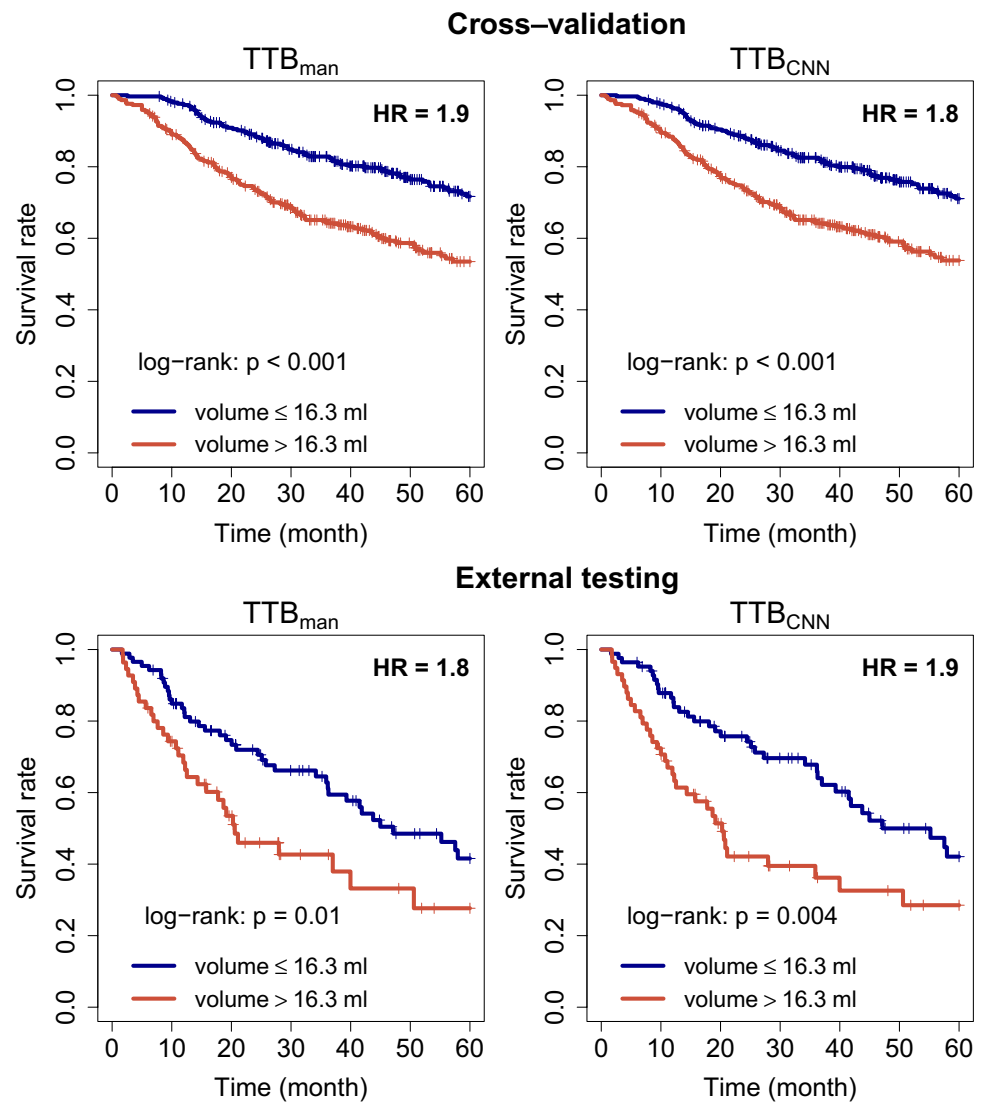


Fig. 4 Examples of CNN delineation errors in 4 selected patients. Relevant transaxial PET/CT slices are shown (top: CT, bottom: PET). The dice coefficients (in the presented plane) for primary tumor (DSC_{PT}), LN metastases (DSC_{LN}) and their union (DSC_{All}) are indicated. Patient A with laryngeal cancer and multiple LN metastases (only one in plane): LN metastasis incorrectly classified as primary;

patient B with oropharyngeal cancer and LN metastasis (both out of plane): CNN produced spurious LN metastasis ROI; patient C with nasopharyngeal cancer and LN metastases (out of plane): primary missed by CNN; patient D with oropharyngeal cancer and low and diffuse uptake LN metastasis: LN metastasis missed by CNN

Fig. 5 Kaplan–Meier curves with respect to OS in cross-validation (top) and external testing (bottom) data



PET-based GTV delineation. However, differences in the used dataset and data preprocessing scheme do not allow for direct comparison between that study and our present investigation. Similarly, only indirect comparison in lesion detection performance is possible to [74] reporting mean patient-wise (TPR/PPV) of (0.86/0.33) for primary tumor + LN metastases detection as compared to cohort (TPR/PPV) of (0.83/0.83) in the present study.

As far as primary tumor delineation alone is concerned, one might validly ask whether there are practical advantages of a CNN-based delineation for tumor entities with usually high tumor to background uptake ratios such as HNC. In fact, much simpler approaches have previously been shown to work adequately in such circumstances. E.g., Ha et al. used interactively defined ellipsoidal ROIs and fixed SUV thresholds within those ROIs to successfully determine

MTV and TLG in head and neck soft tissue sarcoma [86] and demonstrated the prognostic value of the such derived parameters. Regarding the data underlying the present investigation we, too, have previously resorted to a semi-automatic adaptive threshold based method developed in our group [97] for delineation in the context of the previously published clinical study [4] and obtained adequate results in about 95% of the delineations.

However, when extending the task to delineation of lymph node metastases one faces the problem that many lymph nodes exhibit only modest [^{18}F]FDG uptake with typical SUVs of about 3–4 (or lower). This is comparable to the level of the physiological [^{18}F]FDG uptake of various structures in the nasopharyngeal region, e.g., tonsils, minor salivary glands, brown adipose tissue, vocal cords or, in some cases, also muscles. In this situation, threshold

based methods tend to fail and manual intervention or fully manual delineation becomes frequently necessary which is very tedious and time-consuming.

It is exactly this context of delineation at low target to background ratios where the CNN-based approach proves to be distinctly superior. This has recently also been demonstrated by Han and coworkers [98] in a different tumor entity (thymic epithelial tumor) which seems to pose a challenge comparable to the one encountered in the present study regarding the lymph node metastases.

However, the distinguishing advantage of the presently proposed approach is its ability to not only provide decent delineation for, both, high and low contrast structures but also to perform fully automated identification of primary tumor and LN metastases thus providing additional classification information for inclusion into further analysis.

The level of concordance between CNN and the human observer provided ground truth in the present study is superior to typically encountered human interobserver concordance which has, e.g., been reported in [99] as DSC = 0.69 for GTV delineation in PET/CT. Although interobserver concordance might be expected to be somewhat higher for MTV delineation [100], experience tells that it generally will not exceed the degree of concordance between CNN and human observer reported in the present study. This can be rephrased as stating that our trained network overall is capable to perform mostly comparably to an experienced human reader (specifically, the reader(s) having provided the ground truth delineation used in training the network). While it is not able to replace said human observer (due to the remaining sporadic incidences of failure), it is well suited to be utilized as an efficient delineation assistance tool in clinical and research contexts. As has also been reported elsewhere, utilizing such tools will considerably reduce (without manual intervention: eliminate) interobserver variability while also providing obvious speed benefits compared to fully manual or semi-automatic delineation [101]. We believe that the presently proposed CNN especially has potential to facilitate large-scale clinical study evaluations and in the next step could allow to utilize translation of findings from such studies to the clinical routine without imposing intolerable time demands on the clinician.

For example, in the present study the native CNN-based TTB determination yielded outcome predictions for HNC patients (regarding overall survival) of a quality fully competitive to and concordant with prediction based on manually derived TTBs (Fig. 5). This observation can be traced back to the very decent correlation between manual and CNN-based TTB volumes ($R^2 = 0.95$) and the low number of definite outliers as demonstrated in Fig. 2. Further added value is provided by the network's classification capabilities allowing to derive metabolic volumes separately for primary tumor and lymph node metastases which provides

the prerequisite to further tailor the decision support process for specific cancer types [102, 103].

A fundamental concern regarding adoption of deep learning approaches in diagnostic imaging and data evaluation is a possible inability of the trained network to generalize from training data to new, so far unseen data [104]. It is theoretically conceivable that the trained CNN actually has incorporated ("learned") very specific inherent characteristics of the training data and requires their presence in any new input in order to perform successful delineation. A similar problem appears in the context of radiomics where it is addressed via data harmonization procedures [105, 106]. Consequently, it has been suggested to use data harmonization to tackle the generalization problem in CNN-based delineation as well [84]. However, there is currently not enough evidence supporting the usefulness of this strategy. In the present study we have approached the problem from the opposite direction: rather than aiming at harmonization of the training data as well as any "new" data to which the trained network should be applied, we intentionally included "heterogeneous" training data from as many varied sources and institutions as we found doable. Altogether, we were able to collect 698 scans from eight independent data sources. Our working hypothesis was that exposing the network to images with different image characteristics would force the CNN to learn common properties of the images and promote generalization. This hypothesis was then tested in an additional independent external dataset. Even though the overall DSC decreased from 0.870 in the cross-validation results to 0.823 in these external data, it remained high enough to prove good generalizability of the developed network. Survival analysis confirms this conclusion revealing that automatically derived TTB remains prognostic of overall survival in external data, too, with virtually identical hazard ratio to those observed in both manual delineation and cross-validation data. The main driver for reduction in overall DSC was reduced performance of LN metastases delineation ($DSC_{LN} = 0.724$ vs $DSC_{LN} = 0.805$ in external and cross-validation data, respectively). This behavior can be potentially explained by the differences in the distribution of volumes of LN metastases across the datasets (median volume = 2.74 ml vs 0.70 ml, in main and external datasets, respectively) as smaller lesions are generally more difficult to detect and unambiguously delineate.

The representative examples shown in Fig. 3 demonstrate that the trained CNN is principally able to provide fully satisfactory MTV delineation across a wide range of image characteristics regarding tumor and LN metastases location and target/background contrasts. As Fig. 3 B demonstrates, the network also is able to correctly delineate diverse tumor shapes and to exclude necrotic tumor areas. Figure 3 D demonstrates the arguably most important capability of a CNN-based delineation approach, namely

the ability to differentiate elevated “physiological” uptake from malignant lesions.

However, despite satisfactory performance in the majority of cases (obviating any need for manual correction), delineation and classification errors of different severity including occasional manifest failure were also observed. Figure 4 demonstrates some of the most common patterns.

Figure 4 A shows an instance of misclassification of LN metastasis and primary tumor which is reflected in correspondingly reduced DSCs. Such partial misclassification does not affect the TTB parameter at all and can be rather easily corrected manually if deemed necessary. The latter is also true for a related type of error, namely misclassification of physiological focal uptake (e.g., inflammation) as malignant lesion as shown in Fig. 4 B. Actually, this concerns a considerable fraction (22.0%) of the generated ROIs for LN metastases. This can intuitively be understood as a consequence of the fact that inflamed and metastatic LN cannot be discriminated unambiguously based on the image data alone. Discrimination between them by the clinician usually is based on additional clinical information that is not accessible to the CNN. This type of error, too, is relatively easy to correct manually by deleting the spurious ROIs but it occurs in a notable fraction of all cases (25.6%) and will therefore contribute noticeably to the remaining time demands required for user intervention when performing CNN-assisted delineation.

Figure 4 C and D demonstrate instances of failure to identify the primary tumor (C) or LN metastases (D). For such cases, manual intervention/delineation would obviously be required. Such failures occurred in 9.3% (22.2%) of the patients for the primary tumor (LN metastases). However, the mean volume of the missed lesions was 8.16 and 2.78 ml for primary tumor and LN metastases, respectively, which is almost a factor of two lower than the respective mean volumes of 13.21 and 5.45 ml of all lesions, suggesting that mainly small lesions were not detected. In the affected patients, the missed lesions contributed on average 33.6% (median: 16.9%) of the ground truth TTB, indicating that their impact on the derived TTB values was limited in most of the cases, however large errors also occurred.

Due to the black box nature of neural networks, it is inherently difficult to identify specific image characteristics that tend to cause misdelineation/classification. What we have noticed is that sizable errors predominantly occurred for tumors of unusual composition (e.g., Fig. 4 A, where a seemingly singular lesion is in fact the primary tumor and the LN metastasis in direct vicinity of each other) or localization (Fig. 4 C, nasopharyngeal cancer contributed only 6% of the cases in the main dataset). Furthermore, differentiation between malignant and benign [^{18}F]FDG-positive LNs is complicated even for experienced human observers,

particularly in the cases of low (Fig. 4 B) or diffuse (Fig. 4 D) uptake and small lesion size.

Consequently, in such circumstances the ground truth manual delineation and classification itself is not completely unambiguous which inherently limits the obtainable degree of concordance between different observers (either human or neural network). This is our tentative explanation for the observation that overall concordance between network and our ground truth delineation was better for the primary tumors than for lymph nodes. In this context, it has also to be noted, that ground truth definition was based on a single manual delineation per lesion which constitutes an obvious limitation of the present investigation.

In fact, it is quite likely that the increase of training data afforded by multiple independent delineations for each patient would further improve the performance of the resulting network. However, recruiting further experienced observers for the very time-intensive task of performing hundreds of delineations was not feasible within the limits of the present investigation.

Another potential limitation is the omission of data augmentation frequently employed to prevent overfitting which turned out not to be doable due to limited capabilities of the utilized deep learning framework MXNet when dealing with 3D image volumes (affine and warp transforms not available). We compensated for this deficiency by heavy sampling of all 3D data sets with 75% overlap between patches which effectively functions as image shift augmentation. The comparatively large number of tomographic data sets available for network training within this study should further reduce the benefits of additional data augmentation.

Conclusion

To the best of our knowledge, this study presents the first CNN for simultaneous MTV delineation and lesion classification for [^{18}F]FDG PET/CT in HNC patients. Our network allows fast delineation and classification of primary tumor and lymph node metastases in HNC while rarely requiring more than minimal manual corrections. It thus is a capable tool able to massively accelerate and facilitate study data evaluation in large patient groups which also does have clear potential for supervised clinical application.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00259-023-06197-1>.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was partly supported by the Berliner Krebsgesellschaft (ZSF201720) and by the German Federal Ministry of Education and Research (BMBF contract 03ZIK42/OncoRay). The funders had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication.

Availability of data and material Sources of data from public databases are listed in the “Methods” section. All other data are available from the institutes contributing the respective data to the present study upon reasonable request. The developed `mxnet` CNN models are available from the authors upon reasonable request.

Declarations

Ethics approval The studies were approved by the Institutional Review Boards of the participating centers. Retrospective analysis was conducted in accordance with the guidelines of the International Conference on Harmonization/Good Clinical Practice and the principles of the Declaration of Helsinki.

Consent to participate All patients provided written informed consent.

Consent for publication Not applicable

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References


- Nichols AC, Theurer J, Prisman E, Read N, Berthelet E, Tran E, et al. Radiotherapy versus transoral robotic surgery and neck dissection for oropharyngeal squamous cell carcinoma (ORATOR): an open-label, phase 2, randomised trial. *Lancet Oncol*. 2019;20(10):1349–59.
- Wang L, Bai J, Duan P. Prognostic value of 18F-FDG PET/CT functional parameters in patients with head and neck cancer. *Nucl Med Commun*. 2019;40(4):361–9. <https://doi.org/10.1097/mnm.0000000000000974>.
- Zschaek S, Li Y, Lin Q, Beck M, Amthauer H, Bauersachs L, et al. Prognostic value of baseline [18F]-fluorodeoxyglucose positron emission tomography parameters MTV, TLG and asphericity in an international multicenter cohort of nasopharyngeal carcinoma patients. *PLoS ONE*. 2020;15(7): e0236841. <https://doi.org/10.1371/journal.pone.0236841>.
- Zschaek S, Weingärtner J, Lombardo E, Marschner S, Hajjianni M, Beck M, et al. 18F-Fluorodeoxyglucose positron emission tomography of head and neck cancer: location and HPV specific parameters for potential treatment individualization. *Front Oncol*. 2022;12. <https://doi.org/10.3389/fonc.2022.870319>.
- Marschner S, Lombardo E, Minibek L, Holzgreve A, Kaiser L, Albert N, et al. Risk stratification using 18F-FDG PET/CT and artificial neural networks in head and neck cancer patients undergoing radiotherapy. *Diagnostics*. 2021;11(9):1581. <https://doi.org/10.3390/diagnostics11091581>.
- Wang Y, Lombardo E, Avanzo M, Zschaek S, Weingärtner J, Holzgreve A, et al. Deep learning based time-to-event analysis with PET, CT and joint PET/CT for head and neck cancer prognosis. *Comput Methods Programs Biomed*. 2022;222: 106948. <https://doi.org/10.1016/j.cmpb.2022.106948>.
- Castelli J, Depeursinge A, Devillers A, Campillo-Gimenez B, Dicente Y, Prior JO, et al. PET-based prognostic survival model after radiotherapy for head and neck cancer. *Eur J Nucl Med Mol Imaging*. 2018;46(3):638–49. <https://doi.org/10.1007/s00259-018-4134-9>.
- Foster B, Bagci U, Mansoor A, Xu Z, Mollura DJ. A review on segmentation of positron emission tomography images. *Comput Biol Med*. 2014;50(1):76–96. <https://linkinghub.elsevier.com/retrieve/pii/S0010482514001000>.
- Sadaghiani MS, Rowe SP, Sheikhabaei S. Applications of artificial intelligence in oncologic 18F-FDG PET/CT imaging: a systematic review. *Ann Transl Med*. 2021, 9(9), pp. 823–823. <https://doi.org/10.21037/atm-20-6162>
- Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digit Med*. 2021;4(1). <https://doi.org/10.1038/s41746-021-00438-z>.
- Bera K, Braman N, Gupta A, Velcheti V, Madabhushi A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat Rev Clin Oncol*. 2021;19(2):132–46. <https://doi.org/10.1038/s41571-021-00560-7>.
- Chen X, Wang X, Zhang K, Fung KM, Thai TC, Moore K, et al. Recent advances and clinical applications of deep learning in medical image analysis. *Med Image Anal*. 2022;79: 102444. <https://doi.org/10.1016/j.media.2022.102444>.
- Li S, Liu J, Wang Z, Cao Z, Yang Y, Wang B, et al. Application of PET/CT-based deep learning radiomics in head and neck cancer prognosis: a systematic review. *Radiology Science*. 2022;1(1). <https://doi.org/10.15212/radsci-2022-0006>.
- Ha R, Chin C, Karcich J, Liu MZ, Chang P, Mutasa S, et al. Prior to initiation of chemotherapy, can we predict breast tumor response? Deep learning convolutional neural networks approach using a breast MRI tumor dataset. *J Digit Imaging*. 2018;32(5):693–701. <https://doi.org/10.1007/s10278-018-0144-1>.
- Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin Cancer Res*. 2019;25(11):3266–75. <https://doi.org/10.1158/1078-0432.ccr-18-2495>.
- Braman N, Adoui ME, Vulchi M, Turk P, Etesami M, Fu P, et al. Deep learning-based prediction of response to HER2-targeted neoadjuvant chemotherapy from pre-treatment dynamic breast MRI: a multi-institutional validation study. 2020.
- Jiang Y, Jin C, Yu H, Wu J, Chen C, Yuan Q, et al. Development and validation of a deep learning CT signature to predict survival and chemotherapy benefit in gastric cancer. *Ann Surg*. 2020;274(6):e1153–61. <https://doi.org/10.1097/sla.00000000000003778>.
- Yang J, Chen J, Kuang K, Lin T, He J, Ni B. MIA-Prognosis: a deep learning framework to predict therapy response. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020*. Springer International Publishing; 2020. p. 211–20.
- Mu W, Jiang L, Zhang J, Shi Y, Gray JE, Tunali I, et al. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. *Nat Commun*. 2020;11(1). <https://doi.org/10.1038/s41467-020-19116-x>.
- Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, et al. A deep learning-based radiomics model for prediction of survival in Glioblastoma Multiforme. *Sci Rep*. 2017;7(1). <https://doi.org/10.1038/s41598-017-10649-8>.
- Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication:

- a retrospective multi-cohort radiomics study. *PLoS Med.* 2018;15(11): e1002711. <https://doi.org/10.1371/journal.pmed.1002711>.
22. Kickingereder P, Isensee F, Tursunova I, Petersen J, Neuberger U, Bonekamp D, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* 2019;20(5):728–40. [https://doi.org/10.1016/s1470-2045\(19\)30098-1](https://doi.org/10.1016/s1470-2045(19)30098-1).
 23. Wang S, Liu Z, Rong Y, Zhou B, Bai Y, Wei W, et al. Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. *Radiother Oncol.* 2019;132:171–7. <https://doi.org/10.1016/j.radonc.2018.10.019>.
 24. Peng H, Dong D, Fang MJ, Li L, Tang LL, Chen L, et al. Prognostic value of deep learning PET/CT-Based radiomics: potential role for future individual induction chemotherapy in advanced nasopharyngeal carcinoma. *Clin Cancer Res.* 2019;25(14):4271–9. <https://doi.org/10.1158/1078-0432.ccr-18-3065>.
 25. Drukker K, Edwards A, Papaioannou J, Giger M. Deep learning predicts breast cancer recurrence in analysis of consecutive MRIs acquired during the course of neoadjuvant chemotherapy. In: *Medical Imaging 2020: Computer-Aided Diagnosis*, Hahn HK, Mazurowski MA, editors, vol. 11314. International Society for Optics and Photonics. SPIE. p. 1131410. <https://doi.org/10.1117/12.2549044>.
 26. Zhou T, Fu H, Zhang Y, Zhang C, Lu X, Shen J, et al. M²-Net: multi-modal multi-channel network for overall survival time prediction of brain tumor patients. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*. Springer International Publishing; 2020. p. 221–31.
 27. Starke S, Leger S, Zwanenburg A, Leger K, Lohaus F, Linge A, et al. 2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma. *Sci Rep.* 2020;10(1). <https://doi.org/10.1038/s41598-020-70542-9>.
 28. Zhang Y, Lobo-Mueller EM, Karanicolas P, Gallinger S, Haider MA, Khalvati F. Improving prognostic performance in resectable pancreatic ductal adenocarcinoma using radiomics and deep learning features fusion in CT images. *Sci Rep.* 2021;11(1). <https://doi.org/10.1038/s41598-021-80998-y>.
 29. Sokooti H, de Vos B, Berendsen F, Lelieveldt BPF, Išgum I, Staring M. Nonrigid image registration using multi-scale 3D convolutional neural networks. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Springer International Publishing. 2017;232–239.
 30. Balakrishnan G, Zhao A, Sabuncu MR, Dalca AV, Guttag J. An unsupervised learning model for deformable medical image registration. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. p. 9252–60. <https://doi.org/10.1109/cvpr.2018.00964>.
 31. de Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, Išgum I. A deep learning framework for unsupervised affine and deformable image registration. *Med Image Anal.* 2019;52:128–43. <https://doi.org/10.1016/j.media.2018.11.010>.
 32. Wang J, Zhang M. DeepFLASH: an efficient network for learning-based medical image registration. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. p. 4443–51. <https://doi.org/10.1109/cvpr42600.2020.00450>.
 33. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. Deep learning in medical image registration: a review. *Phys Med Biol.* 2020;65(20):20TR01. <https://doi.org/10.1088/1361-6560/ab843e>.
 34. Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. *Mach Vis Appl.* 2020;31(1-2). <https://doi.org/10.1007/s00138-020-01060-x>.
 35. Kumar A, Kim J, Lyndon D, Fulham M, Feng D. An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE J Biomed Health Inform.* 2017;21(1):31–40. <https://doi.org/10.1109/jbhi.2016.2635663>.
 36. Wang SH, Phillips P, Sui Y, Liu B, Yang M, Cheng H. Classification of Alzheimer's Disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. *J Med Syst.* 2018;42(5). <https://doi.org/10.1007/s10916-018-0932-7>.
 37. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing.* 2018;321:321–31. <https://doi.org/10.1016/j.neucom.2018.09.013>.
 38. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, et al. Attention gated networks: learning to leverage salient regions in medical images. *Med Image Anal.* 2019;53:197–207. <https://doi.org/10.1016/j.media.2019.01.012>.
 39. Xie Y, Zhang J, Xia Y. Semi-supervised adversarial model for benign-malignant lung nodule classification on chest CT. *Med Image Anal.* 2019;57:237–48. <https://doi.org/10.1016/j.media.2019.07.004>.
 40. Sharma H, Jain JS, Bansal P, Gupta S. Feature extraction and classification of Chest X-Ray images using CNN to detect pneumonia. In: *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE. p. 227–31. <https://doi.org/10.1109/confluence47617.2020.9057809>.
 41. Abbas A, Abdelsamea MM, Gaber MM. Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *Appl Intell.* 2020;51(2):854–64. <https://doi.org/10.1007/s10489-020-01829-7>.
 42. Li X, Jia M, Islam MT, Yu L, Xing L. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Trans Med Imaging.* 2020;39(12):4023–33. <https://doi.org/10.1109/tmi.2020.3008871>.
 43. Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Thorax disease classification with attention guided convolutional neural network. *Pattern Recogn Lett.* 2020;131:38–45. <https://doi.org/10.1016/j.patrec.2019.11.040>.
 44. Shorfuzzaman M, Hossain MS. MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. *Pattern Recogn.* 2021;113: 107700. <https://doi.org/10.1016/j.patcog.2020.107700>.
 45. Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, et al. Big self-supervised models advance medical image classification. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE. p. 3458–68. <https://doi.org/10.1109/iccv48922.2021.00346>.
 46. Chen L, Dohopolski M, Zhou Z, Wang K, Wang R, Sher D, et al. Attention guided lymph node malignancy prediction in head and neck cancer. *Int J Radiat Oncol Biol Phys.* 2021;110(4):1171–9. <https://doi.org/10.1016/j.ijrobp.2021.02.004>.
 47. Zheng Y, Liu D, Georgescu B, Nguyen H, Comanicu D. 3D deep learning for efficient and robust landmark detection in volumetric data. In: *Lecture Notes in Computer Science*. Springer International Publishing; 2015. p. 565–72.
 48. Ding J, Li A, Hu Z, Wang L. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Springer International Publishing; 2017. p. 559–67.

49. Yan K, Tang Y, Peng Y, Sandfort V, Bagheri M, Lu Z, et al. MULAN: multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In: *Lecture Notes in Computer Science*. Springer International Publishing; 2019. p. 194–202.
50. Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation. *Med Image Anal*. 2020;59: 101557. <https://doi.org/10.1016/j.media.2019.101557>.
51. Wang D, Zhang Y, Zhang K, Wang L. FocalMix: Semi-supervised learning for 3D medical image detection. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. p. 3950–9. <https://doi.org/10.1109/cvpr42600.2020.00401>.
52. Liu Y, Zhang F, Zhang Q, Wang S, Wang Y, Yu Y. Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. p. 3811–21. <https://doi.org/10.1109/cvpr42600.2020.00387>.
53. Mei J, Cheng MM, Xu G, Wan LR, Zhang H. SANet: A slice-aware network for pulmonary nodule detection. *IEEE Trans Pattern Anal Mach Intell*. 2021;1–1. <https://doi.org/10.1109/tpami.2021.3065086>.
54. Baur C, Denner S, Wiestler B, Navab N, Albarqouni S. Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Med Image Anal*. 2021;69: 101952. <https://doi.org/10.1016/j.media.2020.101952>.
55. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. 2015.
56. Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE. p. 565–71. <https://doi.org/10.1109/3dv.2016.79>.
57. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: a nested U-net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer International Publishing; 2018. p. 3–11.
58. Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, et al. Automatic multi-organ segmentation on abdominal CT with dense V-networks. *IEEE Trans Med Imaging*. 2018;37(8):1822–34. <https://doi.org/10.1109/tmi.2018.2806309>.
59. Xue Y, Xu T, Zhang H, Long LR, Huang X. SegAN: adversarial network with multi-scale L1 loss for medical image segmentation. *Neuroinformatics*. 2018;16(3–4):383–92. <https://doi.org/10.1007/s12021-018-9377-x>.
60. Yu L, Wang S, Li X, Fu CW, Heng PA. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: *Lecture Notes in Computer Science*. Springer International Publishing; 2019. p. 605–13.
61. Bai W, Chen C, Tarroni G, Duan J, Guitton F, Petersen SE, et al. Self-supervised learning for cardiac MR image segmentation by anatomical position prediction. In: *Lecture Notes in Computer Science*. Springer International Publishing; 2019. p. 541–9.
62. Baumgartner CF, Tezcan KC, Chaitanya K, Hötter AM, Muehlethaler UJ, Schawkat K, et al. PHiSeg: capturing uncertainty in medical image segmentation. In: *Lecture Notes in Computer Science*. Springer International Publishing; 2019. p. 119–27.
63. Wang X, Han S, Chen Y, Gao D, Vasconcelos N. Volumetric attention for 3D medical image segmentation and detection. In: *Lecture Notes in Computer Science*. Springer International Publishing; 2019. p. 175–84.
64. Fan DP, Zhou T, Ji GP, Zhou Y, Chen G, Fu H, et al. Inf-Net: Automatic COVID-19 lung infection segmentation from CT images. *IEEE Trans Med Imaging*. 2020;39(8):2626–37. <https://doi.org/10.1109/tmi.2020.2996645>.
65. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. 2021.
66. Zuo Q, Chen S, Wang Z. R2AU-Net: attention recurrent residual convolutional neural network for multimodal medical image segmentation. *Secur Commun Netw*. 2021;2021:1–10. <https://doi.org/10.1155/2021/6625688>.
67. Samarasinghe G, Jameson M, Vinod S, Field M, Dowling J, Sowmya A, et al. Deep learning for segmentation in radiation therapy planning: a review. *J Med Imaging Radiat Oncol*. 2021;65(5):578–95.
68. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. UNETR: Transformers for 3D medical image segmentation. In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE. p. 1748–58. <https://doi.org/10.1109/wacv51458.2022.00181>.
69. Li S, Xiao J, He L, Peng X, Yuan X. The tumor target segmentation of Nasopharyngeal Cancer in CT images based on deep learning methods. *Technol Cancer Res Treat*. 2019;18:153303381988456. <http://journals.sagepub.com/doi/10.1177/1533033819884561>.
70. Guo Z, Guo N, Gong K, Zhong S, Li Q. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. *Phys Med Biol*. 2019;64(20):205015. <https://iopscience.iop.org/article/10.1088/1361-6560/ab440d>.
71. Moe YM, Groendahl AR, Mulstad M, Tomic O, Indahl U, Dale E, et al. Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers. 2019.
72. Andrearczyk V, Oreiller V, Vallières M, Castelli J, Elhalawani H, Jreige M, et al. Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans. In: *Proceedings of Machine Learning Research*. 2020. p. 33–43. <https://tinyurl.com/yc5sq5jy>.
73. Groendahl AR, Knudtsen IS, Huynh BN, Mulstad M, Moe YM, Knuth F, et al. A comparison of methods for fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers. *Phys Med Biol*. 2021;66(6): 065012. <https://doi.org/10.1088/1361-6560/abe553>.
74. Moe YM, Groendahl AR, Tomic O, Dale E, Malinen E, Futsaether CM. Deep learning-based auto-delineation of gross tumour volumes and involved nodes in PET/CT images of head and neck cancer patients. *Eur J Nucl Med Mol Imaging*. 2021;48(9):2782–92. <https://doi.org/10.1007/s00259-020-05125-x>.
75. Zhao B, Soraghan J, Caterina GD, Grose D. Segmentation of Head and Neck Tumours Using Modified U-net. In: *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE; p. 1–4. <https://doi.org/10.23919/eusipco.2019.8902637>.
76. Bielak L, Wiedenmann N, Berlin A, Nicolay NH, Gunashekar DD, Hägele L, et al. Convolutional neural networks for head and neck tumor segmentation on 7-channel multiparametric MRI: a leave-one-out analysis. *Radiat Oncol*. 2020;15(1). <https://doi.org/10.1186/s13014-020-01618-z>.
77. Outeiral RR, 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>.
78. Badrigilan S, Nabavi S, Abin AA, Rostampour N, Abedi I, Shirvani A, et al. Deep learning approaches for automated classification and segmentation of head and neck cancers and brain tumors in magnetic resonance images: a meta-analysis study. *Int J Comput Assist Radiol Surg*. 2021;16(4):529–42. <https://doi.org/10.1007/s11548-021-02326-z>.

79. Schouten JP, 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;22(1). <https://doi.org/10.1186/s40644-022-00445-7>.
80. 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(11):1399–406. <https://doi.org/10.1080/0284186x.2021.1949034>.
81. Guo Z, Guo N, Gong K, Li Q. Automatic multi-modality segmentation of gross tumor volume for head and neck cancer radiotherapy using 3D U-Net. In: Mori K, Hahn HK, editors. *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950. International Society for Optics and Photonics. SPIE; p. 1095009. <https://doi.org/10.1117/12.2513229>.
82. Andrearczyk V, Fontaine P, Oreiller V, Castelli J, Jreige M, Prior JO, et al. Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer. In: Rekić I, Adeli E, Park SH, Schnabel J, editors. *Predictive Intelligence in Medicine*. Springer International Publishing, Cham; p. 147–156. ISBN 978-3-030-87602-9.
83. Sobirov I, Nazarov O, Alasmawi H, Yaqub M. Automatic segmentation of head and neck tumor: how powerful transformers are? 2022.
84. Andrearczyk V, Oreiller V, Boughdad S, Rest CCL, Elhalawani H, Jreige M, et al. Overview of the HECKTOR Challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT Images. In: *Lecture Notes in Computer Science*. Springer International Publishing. 2022. p. 1–37.
85. Xie J, Peng Y. The Head and Neck Tumor Segmentation Based on 3D U-Net. In: *Lecture Notes in Computer Science*. Springer International Publishing. 2022. p. 92–98.
86. Ha SC, Oh JS, Roh JL, Moon H, Kim JS, Cho KJ, et al. Pretreatment tumor SUV max predicts disease-specific and overall survival in patients with head and neck soft tissue sarcoma. *Eur J Nucl Med Mol Imaging*. 2017;44:33–40.
87. Shiri I, Arabi H, Sanaat A, Jenabi E, Becker M, Zaidi H. Fully automated gross tumor volume delineation from pet in head and neck cancer using deep learning algorithms. *Clin Nucl Med*. 2021;46(11):872–83. <https://doi.org/10.1097/rlu.00000000000003789>.
88. Vallières M, Kay-Rivest E, Perrin L, Liem X, Furstoss C, Khaouam N, et al. Data from Head-Neck-PET-CT. 2017.
89. Wee L, Dekker A. Data from Head-Neck-Radiomics-HN1. 2019.
90. Grossberg AJ, Mohamed ASR, Elhalawani H, Bennett WC, Smith KE, Nolan TS, et al. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Sci Data*. 2018;5(1). <https://doi.org/10.1038/sdata.2018.173>.
91. Zuley ML, Jarosz R, Kirk S, Lee Y, Colen R, Garcia K, et al. Radiology data from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma [TCGA-HNSC] collection. 2016.
92. Beichel RR, Ulrich EJ, Bauer C, Wahle A, Brown B, Chang T, et al. Data from QIN-HEADNECK. 2015.
93. Hofheinz F, Pötzsch C, Oehme L, Beuthien-Baumann B, Steinbach J, Kotzerke J, et al. Automatic volume delineation in oncological PET. Evaluation of a dedicated software tool and comparison with manual delineation in clinical data sets. *Nuklearmedizin*. 2012;51:9–16.
94. Hofheinz F, Langner J, Petr J, Beuthien-Baumann B, Steinbach J, Kotzerke J, et al. An automatic method for accurate volume delineation of heterogeneous tumors in PET. *Med Phys*. 2013;40(8): 082503.
95. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc.; p. 5999–6009. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
96. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for statistical computing, Vienna, Austria. 2022.
97. Hofheinz F, Langner J, Petr J, Beuthien-Baumann B, Steinbach J, Kotzerke J, et al. An automatic method for accurate volume delineation of heterogeneous tumors in PET. *Med Phys*. 2013;40(8): 082503.
98. Han S, Oh JS, Kim Yi, Seo SY, Lee GD, Park MJ, et al. Fully automatic quantitative measurement of 18F-FDG PET/CT in Thymic Epithelial tumors using a convolutional neural network. *Clin Nucl Med*. 2022;47(7):590–598.
99. Gudi S, Ghosh-Laskar S, Agarwal JP, Chaudhari S, Rangarajan V, Paul SN, et al. 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*. 2017;48(2):184–92. <https://doi.org/10.1016/j.jmir.2016.11.003>.
100. Miyabe J, Hanamoto A, Tatsumi M, Hamasaki T, Takenaka Y, Nakahara S, et al. Metabolic tumor volume of primary tumor predicts survival better than T classification in the larynx preservation approach. *Cancer Sci*. 2017;108(10):2030–8. <https://doi.org/10.1111/cas.13345>.
101. Lin L, Dou Q, Jin YM, Zhou GQ, Tang YQ, Chen WL, et al. Deep learning for automated contouring of primary tumor volumes by MRI for Nasopharyngeal Carcinoma. *Radiology*. 2019;291(3):677–86. <https://doi.org/10.1148/radiol.2019182012>.
102. Cho JK, Hyun SH, Choi N, Kim MJ, Padera TP, Choi JY, et al. Significance of lymph node metastasis in cancer dissemination of head and neck cancer. *Translational Oncology*. 2015;8(2):119–25. <https://doi.org/10.1016/j.tranon.2015.03.001>.
103. Huang Y, Feng M, He Q, Yin J, Xu P, Jiang Q, et al. Prognostic value of pretreatment 18F-FDG PET-CT for nasopharyngeal carcinoma patients. *Medicine*. 2017;96(17): e6721. <https://doi.org/10.1097/md.00000000000006721>.
104. Ahmed KB, Goldgof GM, Paul R, Goldgof DB, Hall LO. Discovery of a generalization gap of convolutional neural networks on COVID-19 X-Rays Classification. *IEEE Access*. 2021;9:72970–9. <https://doi.org/10.1109/access.2021.3079716>.
105. Aide N, Lasnon C, Veit-Haibach P, Sera T, Sattler B, Boellaard R. EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. *Eur J Nucl Med Mol Imaging*. 2017;44(S1):17–31. <https://doi.org/10.1007/s00259-017-3740-2>.
106. Mali SA, Ibrahim A, Woodruff HC, Andrearczyk V, Müller H, Primakov S, et al. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. *J Perinat Med*. 2021;11(9):842. <https://doi.org/10.3390/jpm11090842>.

Authors and Affiliations

Pavel Nikulin¹  · Sebastian Zschaeck^{2,3} · Jens Maus¹ · Paulina Cegla⁴ · Elia Lombardo⁵ · Christian Furth⁷ · Joanna Kaźmierska^{10,11} · Julian M. M. Rogasch^{3,7} · Adrien Holzgreve⁸ · Nathalie L. Albert⁸ · Konstantinos Ferentinos⁹ · Iosif Strouthos⁹ · Marina Hajjianni^{2,3} · Sebastian N. Marschner⁵ · Claus Belka^{5,6} · Guillaume Landry⁵ · Witold Cholewinski^{4,10} · Jörg Kotzerke¹² · Frank Hofheinz¹ · Jörg van den Hoff^{1,12}

¹ Helmholtz-Zentrum Dresden-Rossendorf, PET Center, Institute of Radiopharmaceutical Cancer Research, Bautzner Landstrasse 400, 01328 Dresden, Germany

² Department of Radiation Oncology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

³ Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany

⁴ Department of Nuclear Medicine, Greater Poland Cancer Centre, Poznan, Poland

⁵ Department of Radiation Oncology, University Hospital, Ludwig-Maximilians-University (LMU) Munich, Munich, Germany

⁶ German Cancer Consortium (DKTK), Partner Site Munich, Munich, Germany

⁷ Department of Nuclear Medicine, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

⁸ Department of Nuclear Medicine, University Hospital, Ludwig-Maximilians-University (LMU) Munich, Munich, Germany

⁹ Department of Radiation Oncology, German Oncology Center, European University Cyprus, Limassol, Cyprus

¹⁰ Electroradiology Department, University of Medical Sciences, Poznan, Poland

¹¹ Radiotherapy Department II, Greater Poland Cancer Centre, Poznan, Poland

¹² Department of Nuclear Medicine, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany