



# Automatic Segmentation of Head and Neck Cancer from PET-MRI Data Using Deep Learning

Joonas Liedes<sup>1</sup> · Henri Hellström<sup>1</sup> · Oona Rainio<sup>1</sup> · Sarita Murtojärvi<sup>3</sup> · Simona Malaspina<sup>1,2</sup> · Jussi Hirvonen<sup>1,4</sup> · Riku Klén<sup>1</sup> · Jukka Kempainen<sup>1,2</sup>

Received: 10 May 2023 / Accepted: 10 August 2023  
© The Author(s) 2023

## Abstract

**Purpose** Head and neck squamous cell carcinoma (HNSCC) is one of the most common cancer types globally. Due to the complex anatomy of the region, diagnosis and treatment is challenging. Early diagnosis and treatment are important, because advanced and recurrent HNSCC have a poor prognosis. Robust and precise tools are needed to help diagnose HNSCC reliably in its early stages. The aim of this study was to assess the applicability of a convolutional neural network in detecting and auto-delineating HNSCC from PET-MRI data.

**Methods** 2D *U*-net models were trained and tested on PET, MRI, PET-MRI and augmented PET-MRI data from 44 patients diagnosed with HNSCC. The scans were taken 12 weeks after chemoradiation therapy with a curative intention. A proportion of the patients had follow-up scans which were included in this study as well, giving a total of 62 PET-MRI scans. The scans yielded a total of 178 PET-MRI slices with cancer. A corresponding number of negative slices were chosen randomly yielding a total of 356 slices. The data was divided into training, validation and test sets ( $n = 247$ ,  $n = 43$  and  $n = 66$  respectively). Dice score was used to evaluate the segmentation accuracy. In addition, the classification capabilities of the models were assessed.

**Results** When true positive segmentations were considered, the mean Dice scores for the test set were 0.79, 0.84 and 0.87 for PET, PET-MRI and augmented PET-MRI, respectively. Classification accuracies were 0.62, 0.71 and 0.65 for PET, PET-MRI and augmented PET-MRI, respectively. The MRI based model did not yield segmentation results. A statistically significant difference was found between the PET-MRI and PET models ( $p = 0.008$ ).

**Conclusion** Automatic segmentation of HNSCC from the PET-MRI data with 2D *U*-nets was shown to give sufficiently accurate segmentations.

**Keywords** Head and neck neoplasms · Deep learning · Convolutional neural networks · Positron-emission tomography · Magnetic resonance imaging

✉ Joonas Liedes  
joonas.liedes@utu.fi

Henri Hellström  
henri.m.hellstrom@outlook.com

Oona Rainio  
ormrai@utu.fi

Sarita Murtojärvi  
sarita.murtojarvi@tyks.fi

Simona Malaspina  
simona.malaspina@tyks.fi

Jussi Hirvonen  
jussi.hirvonen@utu.fi

Riku Klén  
riku.klen@utu.fi

Jukka Kempainen  
Jukka.Kempainen@tyks.fi

<sup>1</sup> Turku PET Centre, University of Turku and Turku University Hospital, 20520 Turku, Finland

<sup>2</sup> Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, 20521 Turku, Finland

<sup>3</sup> Department of Otorhinolaryngology, Turku University Hospital, 20521 Turku, Finland

<sup>4</sup> Department of Radiology, University of Turku and Turku University Hospital, Turku, Finland

## Abbreviations

CNN	Convolutional neural network
HNC	Head and neck cancer
HNSCC	Head and neck squamous cell carcinomas
HPV	Human papilloma virus
IMRT	Intensity-modulated radiation therapy
MTV	Resulting metabolic tumour volumes
CRT	Chemoradiation therapy
FDG	<sup>18</sup> F-fluorodeoxyglucose
SUV	Standardized uptake value

## 1 Introduction

Head and neck squamous cell carcinomas (HNSCC) are globally the sixth most common type of cancer. HNSCC can arise from multiple sites within the otorhinolaryngeal space such as the nasal cavity, sinuses, oral cavity, oropharynx, larynx and the salivary glands. The incidence of HNSCC has been steadily increasing. In Finland, the incidence of HNSCC was approximately 900 in 2018 [1, 2].

The complex nature of the anatomical site gives rise to several problems when diagnosing and treating HNSCC. Upon diagnosis the tumour is often already at an advanced stage and the treatment options are limited. Any recurrences after treatment are similarly difficult to diagnose due to the complexity of the anatomy in combination with inflammation and possible scarring from treatment. Recurrence of the disease has a poor prognosis [3]. Various follow-up imaging protocols after chemoradiation therapy (CRT) have been proposed to mitigate this problem. Due to their limited ability to distinguish scar tissue or inflammation from malignant tissue, conventional methods of imaging, such as computed tomography (CT) and magnetic resonance imaging (MRI), are often supplemented by fusion imaging such as positron emission tomography (PET) combined with CT. The combination of anatomical data from the conventional modalities and the metabolic information from PET greatly enhances the sensitivity of detecting recurrences in their earlier stages. PET-CT is a well-established and validated method for detecting recurrences [4]. Lately PET-MRI has also been proposed as a superior alternative to PET-CT, due to the better soft-tissue contrast in MRIs [5]. Fusion imaging is a valid alternative to elective neck dissection and it plays an important role in the event of an unknown primary [6, 7]. Fusion imaging has also been shown to give additional value when planning radiation therapy [8].

In clinical practice, the physician analyses and combines the anatomical information of the conventional imaging (CT/MRI) with the metabolic information of the PET images using a specific software. The PET data is analysed visually and semi-quantitatively assessing the standardised uptake value (SUV) of specific regions or lesions. The SUV

indicates the tissues propensity to intake the radiotracer used for the imaging, most commonly <sup>18</sup>F-fluorodeoxyglucose (FDG). An elevated SUV is typically seen in malignant processes, but it is not specific to them. FDG-uptake is also regularly seen in the case of inflammation. There is no clear cut-off point to differentiate malignant processes from benign processes. The analysis of the imaging data is mainly done manually by the physician. The process is cumbersome and time consuming. In addition, manual analysis of the data might cause inter-observer variation and low reproducibility; both of which would have negative implications at the individual level and when striving for standardised treatment protocols [9].

With the rapid increase in computing power over the last decade, especially in the case of graphic processing units, enormous developments have been made in the field of machine learning. Machine learning is a subset of artificial intelligence which finds connections and makes predictions given a specific set of data. The machine learning algorithm is fed a set of data (training data) which has been labelled appropriately beforehand in order for the algorithm to be able to produce predictions. The algorithm is then tested with a similar set of data that it has not seen before (test data) and asked to supply predictions. As opposed to traditional algorithms, machine learning needs little or no human oversight and no specific instructions on how to produce the predictions.

A specific type of machine learning architecture called neural networks have been shown to work exceptionally well in the field of pattern recognition and with machine vision in general. Neural networks are designed to mimic the human brain as regards how the processing units are connected and act as if they were actual neurons. For the last decade a subset of machine learning called deep learning has been perhaps the most prevalent due to its ability to tackle vast amounts of data and thus more complex problems. This has been the case especially for machine vision [10]. Deep learning typically utilises very large neural networks accompanied by various pooling and convolution layers; these are designed to preserve the data quality whilst it is being analysed as well as keeping the data size manageable. These networks are often referred to as convolutional neural networks (CNNs). Deep learning methods have been applied successfully to medical imaging in recent years [11]. The deep learning algorithms have matched or even surpassed experienced clinicians in medical image recognition tasks [12–14]. This approach has been successfully adopted in the segmentation of head and neck cancer (HNC) from PET-CT images [15–18] and PET-MRI images [15] as well.

The aim of our study is to accurately segment HNSCC from PET-MRI data using deep learning. Furthermore, the use of MRI data should provide interesting insights regarding the choice of the source of the anatomical information

when conducting segmentation via deep learning. This is especially relevant as the majority of studies regarding automatic segmentation of fusion imaging have been conducted with PET-CT data.

## 2 Materials and Methods

### 2.1 Imaging and Annotation

HNSCC patients referred to Turku University Hospital after curative chemoradiotherapy for restaging PET-MRI from February of 2014 to May of 2017 were retrospectively included in this study. We obtained permission from the hospital district board. Written patient consent was waived due to the retrospective nature of the study. Patients were treated with intensity-modulated radiation therapy (IMRT) with concurrent chemotherapy including cisplatin or cetuximab. The inclusion criteria were histologically confirmed squamous cell carcinoma of the head and neck area and treatment with chemoradiotherapy. Patients meeting these criteria that underwent an FDG PET-MRI 12 weeks after their treatment were consecutively chosen for the study, yielding a total of 52 patients. A workflow of the study is depicted in Fig. 1.

Patients with only distant metastases outside of the head and neck area were excluded, leaving a total of 44 patients. A portion of the patients had had follow-up scans, altogether these yielded 62 PET-MRIs, which were included in the analysis. Locoregional recurrences were found in 18 instances and 44 PET-MRI scans were considered negative. Recurrence was confirmed with follow-up imaging or with histopathological sampling. Histological confirmation was collected from 8 patients considered to have a recurrent disease.

PET-MRI scans were performed with a sequential Ingenuity 3 T TF PET-MRI system (Philips Healthcare) using a SENSE neurovascular coil. The transaxial sequences used for the MRI scans were: T2 TSE, T1 TSE, T1 SPIR. T1 sequences focused on the area of the primary tumour. T1SPIR sequences were scanned with a contrast agent. In addition, the T2 sequences provided exact anatomical information from both the tumour and lymph node areas. For this study, T1 SPIR sequences were utilised in the 56 cases available. T1 TSE was used in the 5 instances where T1 SPIR was not available. Similarly, T2 TSE was used in 1 case where the forementioned sequences were not available.

Attenuation correction sequences based on a Dixon MRI were acquired from the forehead level to the groin level. The attenuation correction procedure was performed using a 3-segment model with a differentiation between air, lung and soft tissue.

PET imaging was performed immediately after the MRI. The transaxial field of view was 576 mm. Reconstruction of

the PET images was done using the default reconstruction algorithm “Blob-OS-TF”, a 3D ordered subset iterative TOF reconstruction technique with 3 iterations and 33 subsets. Using  $144 \times 144$  matrices, the final voxel size was  $4 \times 4 \times 4$  mm<sup>3</sup>. All reconstructions included the necessary corrections for image quantification: attenuation, randoms, scatter, dead-time, decay and detector normalisation.

The images were then co-registered and resliced into common dimensions, and the PET images were cropped to the dimensions of the corresponding MRI sequence. The recurrent tumour was manually delineated under the supervision of an experienced nuclear medicine specialist. Delineation was done utilising both the metabolic information from the PET image and the anatomical information from the MRI image to avoid false positives associated with annotating solely on PET images. In addition, any malignant lymph nodes or local metastases, if present, were similarly annotated. PET-MRI reports written by nuclear medicine specialists and radiologists were also utilised in the annotation process along with clinical information from the patients’ records. Based on this information image masks were created. The resulting metabolic tumour volumes (MTV) were considered as the ground truth for training and evaluating our CNN model. The re-slicing and annotation was conducted using Carimas software [19].

The 18 imaging instances considered to have recurrences resulted in 178 individual transaxial PET-MRI image slices that presented PET-positive malignant tissue based on the manual delineation. These image slices were included in the study. Reciprocally 178 image slices of the head and neck area were randomly chosen from the patients with negative PET-MRI.

### 2.2 Data Pre-processing

The images were randomly divided into training and test sub-sets patient wise. The training set consisted of 290 images (81%) and the test set consisted of 66 images (19%). The cohorts were stratified using the mask value as class labels to ensure a similar distribution of cancer positive and negative images in both sets. The training sub-set was further divided into a validation set consisting of 43 (15%) images.

All image slices and the corresponding masks were resized from  $512 \times 512$  pixels to  $128 \times 128$  pixels. Resizing was carried out to accommodate larger batch sizes and reduce the computing power needed for training. PET and MRI pixel values were normalised by linearly scaling between 0 and 1. Normalisation was carried out due to the differing pixel value ranges between the two modalities. Normalisation was done utilising the global minima and maxima for PET and MRI images, which were acquired from the training sub-set. If the test set had higher pixel values than those of the training set,

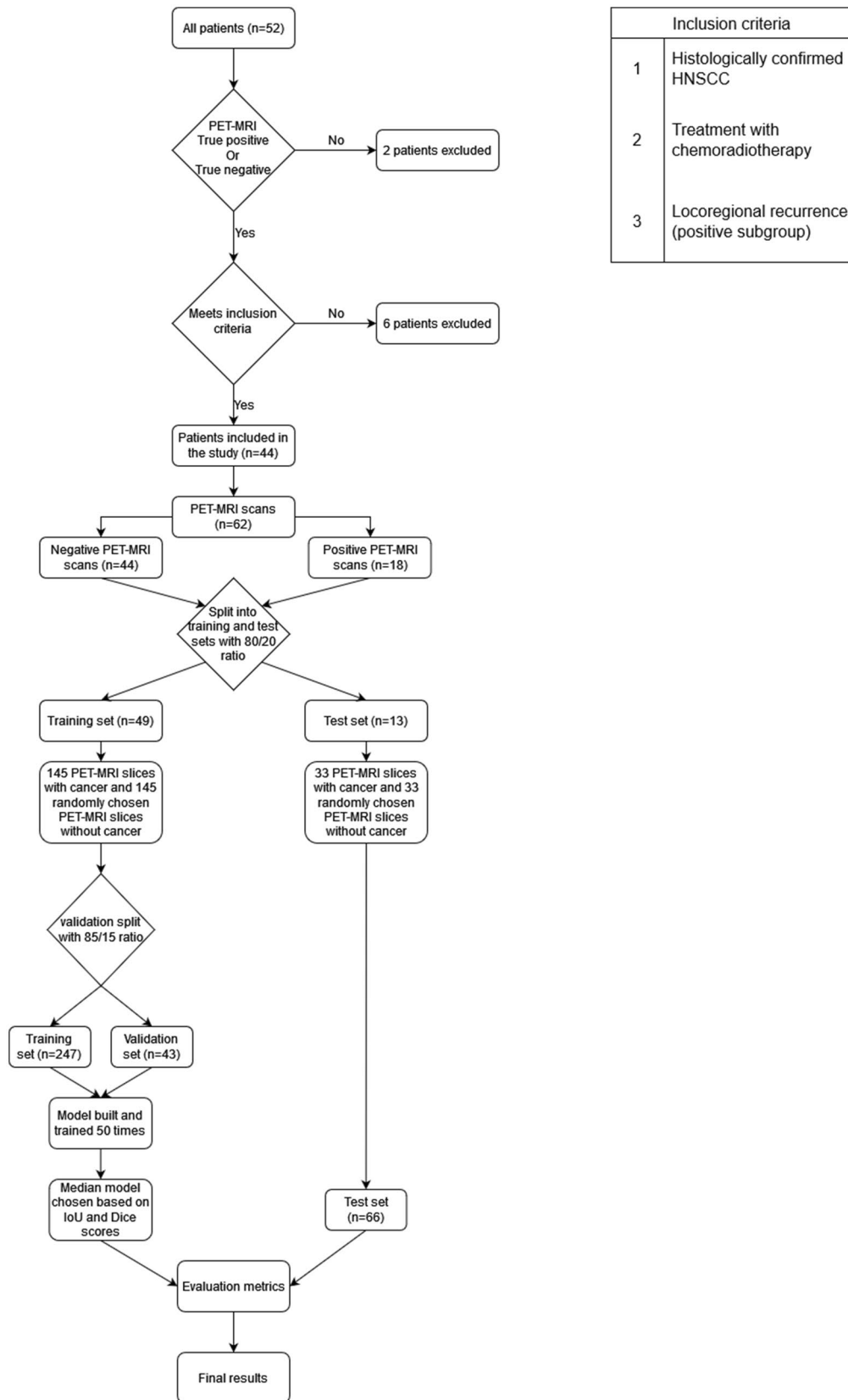


Fig. 1 Workflow of the study

the normalisation values were set to 1 and 0 for those pixel values that exceeded or fell below the given interval respectively.

An augmentation pipeline was constructed to evaluate the model with additional data. This approach has been shown to be a viable method of increasing the size of PET-MRI data sets [15]. Augmentation operations included flipping the images from left to right and upside down. In addition, the images were randomly translated from -10 to 10% and rotated between -15 and 15 degrees. Augmentation was conducted over the training sub-set and the corresponding masks to yield 5 augmented images from a single original image. This resulted in 1450 training images and masks.

### 2.3 Model Architectures and Training

A U-net model as described in [11] was constructed using Tensorflows Keras version 2.5.0 framework [20] in the Python version 3.7.10 [21]. The model was trained to segment the primary tumour and possible metastasis in the PET-MRI slices. In addition, models trained with PET slices alone and augmented PET-MRI slices were constructed. Binary cross-entropy was used as the loss function. Jaccard similarity coefficient [22] was used as the accuracy metric in training and the Adam optimizer with a learning rate of 0.001 was used to optimise the gradient descent. Early stopping was utilised, and epochs were set to 200 with a patience value of 50 epochs. The networks Jaccard similarity coefficient for the validation data was compared after each epoch and the highest value was maintained in the memory. After training, the best model configuration was restored based on this value and was chosen for further performance evaluation with the test set.

### 2.4 Model Performance Evaluation

The overall segmentation accuracy was evaluated by calculating the Jaccard similarity coefficient and Dice score for the test set.

For performance evaluation the model was constructed 50 times independently and the median model of this group was chosen based on the Dice scores. The median model was chosen so that its generalisation properties for a larger cohort would be optimal. After choosing the median model, an optimal threshold for yielding binary masks was chosen with a brute force method.

Let  $A$  be the predicted binary mask and  $B$  the ground truth binary mask. The Jaccard similarity coefficient is calculated as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where  $|A \cap B|$  depicts the intersection of  $A$  and  $B$ . Similarly,  $|A \cup B|$  represents the union.

The Dice score is calculated as follows:

$$D(A, B) = \frac{2 * |A \cap B|}{|A| + |B|}$$

where  $|A \cap B|$  depicts the intersection of the  $A$  and  $B$ , and  $|A|$  and  $|B|$  are the number of pixels with value 1 in  $A$  and  $B$  respectively.

The classification capabilities of the model were also evaluated. For this purpose, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) segmentations were determined. Since the images were cropped to a common size of  $128 \times 128$  from  $512 \times 512$ , one pixel in these images represents a corresponding height and width of 1.7 mm. The real-world fidelity of PET-MRI interpretation by radiologists and nuclear medicine specialists is around 5 mm. A cut off value of 9 segmented pixels per image was chosen to reflect this.

Sensitivity for the classification was calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity for the classification was calculated as follows:

$$Specificity = \frac{TN}{TN + FP}$$

Accuracy for the classification was calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

## 3 Results

### 3.1 Segmentation Performance

After training, our models were able to segment malignant tissue from the test images accurately. Dice scores and Jaccard similarity coefficients for the median model trained with PET-MRI images without augmentation were 0.81 and 0.68 respectively [Table 1]. The model trained solely on PET data achieved a median Dice score of 0.68 and a Jaccard similarity coefficient of 0.52. The model trained with

**Table 1** Segmentation performances of the median models for the whole test-set

Model	Median dice score	Median jaccard similarity coefficient
PET	0.68	0.52
PET-MRI	0.81	0.68
Augmented PET-MRI	0.71	0.56

augmented PET-MRI data achieved a Dice score of 0.71 and a Jaccard similarity coefficient of 0.56 for the entire test-set.

Segmentation performance was then evaluated as a mean of each individual pair of predictions and ground truth masks. For this evaluation, we included only the images where actual segmentation was performed [Table 2]. Similar evaluations where true negatives, false positives and false negatives were considered can be seen in the supplementary tables [Supplementary Table 2]. The performance of the models was also evaluated on the training set

[Supplementary Table 3]. An example of an accurate prediction of a cancer of the nasopharynx is shown in Fig. 2. Similarly, an example of an insufficient segmentation result can be seen in Fig. 3.

The models were then compared and tested for statistical differences using the Wilcoxon signed-rank test [Table 2]. A statistically significant difference between the PET based and PET-MRI based model was observed. Similar differences were not detected between PET-MRI and augmented PET-MRI or PET and augmented PET-MRI.

**Table 2** Segmentation performance of the median models for each individual image slice

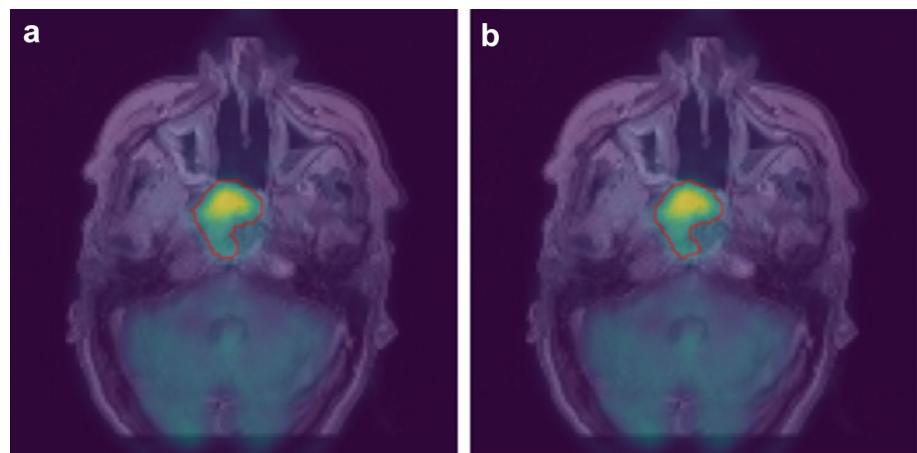
Model	Mean dice score $\pm$ SD	P-value	Mean jaccard similarity coefficient $\pm$ SD	P-value
PET	0.79 $\pm$ 0.16	0.008*	0.68 $\pm$ 0.20	0.008*
PET-MRI	0.84 $\pm$ 0.14	0.247§	0.75 $\pm$ 0.18	0.273 <sup>a</sup>
Augmented PET-MRI	0.87 $\pm$ 0.09	0.156†	0.78 $\pm$ 0.13	0.145 <sup>b</sup>

Only true positive segmentations were considered in these calculations. \*Compared to PET-MRI

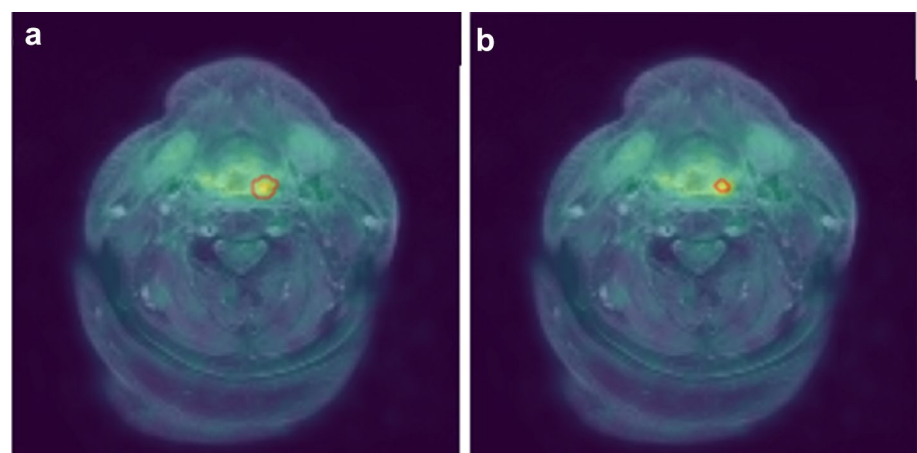
<sup>a</sup>Compared to augmented PET-MRI

<sup>b</sup>Compared to PET

**Fig. 2** Trans axial FDG PET-MRI images with example of a good segmentation result. The Dice score for this image pair was 0.95. Figure A represents the ground truth and Figure B represents the segmentation made by the model



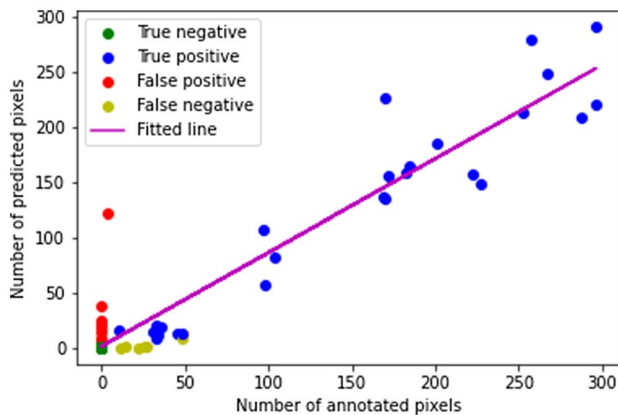
**Fig. 3** Trans axial FDG PET-MRI images with example of an insufficient segmentation result. The Dice score for this image pair was 0.43. Figure A represents the ground truth and Figure B represents the segmentation made by the model



Linear regression was done to evaluate the correlation between the number of segmented pixels in the ground truth labels and the respective predicted labels produced by the model [Fig. 4]. The coefficient of determination ( $R^2$ ) for the PET-MRI model trained with the original data set was 0.90.

### 3.2 Classification performance

With the classification fidelity set to the cut-off value of 9 pixels the PET-MRI based model yielded an accuracy of 0.71; with the specificity and sensitivity being 0.68 and 0.77, respectively. The solely PET based model was more prone to predicting false positives, however, its sensitivity was on par with the PET-MRI model [Table 3]. The model trained with augmented PET-MRI data achieved a sensitivity, specificity and accuracy of 0.53, 0.77 and 0.65, respectively. Similar results for the training set can be seen in Supplementary Table 5.



**Fig. 4** Y-axis depicts the number of pixels masked by the model given the test-set. X-axis depicts the number of pixels annotated manually. The labels represent predictions for individual image slices in the test-set and the colour coding refers to the classification of the image given a threshold of 9 pixels as seen in Table 3. This graph demonstrates how the number of annotated pixels positively correlates with the segmentation accuracy. This graph was produced based on the segmentation made by the median model trained with fused images without augmentation

## 4 Discussion

Our proposed method demonstrates that it is possible to build an accurate segmentation model despite having a limited amount of training data. In addition, this study suggests that the use of MRI provides additional value to segmentation tasks when fused with PET. Furthermore, we have shown that accurate segmentation at the level of individual image slices is feasible.

To the best of our knowledge only one study has been conducted on automatic HNC segmentation from PET-MRI data using deep learning [15]. In their study, the authors achieved a mean Dice score of 0.72 using a residual 3D-Unet. The key difference, compared to our approach, was a larger sample size that consisted of PET-MRI images with only cancer whereas our study included images with and without cancer. In addition, the authors used a residual 3D-Unet with a dual loss function combining Dice loss with focal loss. However, the code to reproduce this study was not available, therefore direct comparisons with our proposed model were not possible. In addition, a small number of studies utilising deep learning to auto delineate HNC from PET-CT data have been done [15–18]. The 2D CNN utilising PET-CT data described by Huang et al. [16] achieved a mean Dice score of 0.74 for the delineation of the primary tumour. The 3D-DenseNet as described by Guo et al. [18] achieved a median Dice score of 0.73 in automatic segmentation of the PET-CT images. The 2D-Unet described by Moe et al. [17] obtained a mean Dice score of 0.71 for automatic tumour delineation. These studies are summarised in Table 4.

Our median model obtained a Dice score of 0.81 over the entire test-set. A mean Dice score of 0.84 was achieved for individual image segmentations, when considering the image pairs classified as true positives. Our proposed method was able to achieve slightly higher segmentation results compared to those described previously. The use of MRI data along with PET data could be a contributing factor to this observed difference. It should be noted that the methods with which the reported segmentation scores are calculated vary and therefore are not strictly comparable. Similarly, the results obtained by the model trained with PET data only resemble those previously described in the literature. The information from the PET images seems

**Table 3** Classification performance of the models in the test-set

Data	True positives	True negatives	False positives	False negatives	Sensitivity	Specificity	Accuracy
PET	18	23	19	6	0.75	0.55	0.62
PET-MRI	20	27	13	6	0.77	0.68	0.71
Augmented PET-MRI	17	26	8	15	0.53	0.77	0.65

**Table 4** Comparison of segmentation models as described in the literature and our proposed method

Study	Modality used in training	Training group (Number of patients)	Validation group (Number of patients)	Test group (Number of patients)	Mean dice score
Ren et al	PET-MRI	92	31	30	0.72
Huang et al	PET-CT	21	1	0	0.74
Guo et al	PET-CT	140	35	75	0.73
Moe et al	PET-CT	142	15	40	0.71
Our proposed method	PET-MRI	42	7	13	0.84

essential for the performance as the mean Dice score for this PET based model was 0.79.

A purely MRI based model was also trained but did not yield segmentation results. This was most likely due to the limited dataset causing discrepancies among the anatomical locations of the malignant tissue between the training-set and the test-set. Similarly, the model trained with augmented PET-MRI data yielded slightly lower segmentation and classification results [Tables 1 and 3] compared to the PET-MRI model trained without augmentation. We suspect this causes the model to overfit to the augmented data thus accentuating the discrepancy between the training and test-sets. This might also explain why no significant difference was observed between the PET-based and the augmented PET-MRI based models.

Some limitations to our study should be taken into consideration. Firstly, the number of patients available for the training was limited and augmentation was used to increase the sample size; this might have caused the training population to be too homogenous for the model to generalise to wider patient populations. A homogenous training and testing material might also cause overestimation of the model's true accuracy in a real-world setting. Secondly, the negative images were chosen randomly to yield an even distribution of positives and negatives in the training and test sets to increase learning performance. In a real world setting the ratio between negative and positive image slices is often skewed towards the negatives. Furthermore, our proposed model is prone to segmenting false positives on PET images with high benign metabolic activity, which is evident when observing the classification capabilities [Table 3]. In a clinical setting the tendency to favour false positives rather than false negatives is preferable, yet efforts should be made to reduce the number of false positives since this directly affects the clinician's workload when deep learning systems are implemented in practice.

To produce a production ready deep learning model suitable for clinical practice, further investigation is needed to achieve better segmentation results and classification accuracy. Further refinement of the pre-processing protocols is required to ensure high quality training data while also increasing the sample size. For this study, a 2D-Unet

was chosen due to computational efficiency, however, more robust models capable of 3D-segmentation should be considered in the future. This is because it is our belief that these types of models will ultimately bring the most value in clinical practice. Furthermore, tactics to artificially produce quality training data, such as transfer learning and generative adversarial networks should be utilised [23, 24].

## 5 Conclusion

Our study demonstrates that deep learning with 2D U-nets can produce relatively accurate cancer segmentation results from HNSCC PET-MRI data, even with a limited sample size. Therefore, it is highly suitable for further development as a diagnostic tool.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s40846-023-00818-8>.

**Funding** Open Access funding provided by University of Turku (UTU) including Turku University Central Hospital. The first author was financially supported by Cancer Foundation Finland and State funding for university-level health research, Turku University Hospital (project number 11065). The second author was financially supported by State funding for university-level health research, Turku University Hospital (project number 11232). The third author was financially supported by Jenny and Antti Wihuri Foundation. The eighth author was financially supported by Cancer Foundation Finland and State funding for university-level health research, Turku University Hospital (project number 11065).

**Data Availability** Data not available due to ethical restrictions.

**Code Availability** Code available at: [https://github.com/joonasliedes/hnsc\\_ai](https://github.com/joonasliedes/hnsc_ai)

## Declarations

**Conflict of interest** The authors declare that there is no conflict of interest regarding the publication of this paper.

**Ethical Approval** Institutional Review Board approval was obtained from the Hospital District of Southwest Finland (project number 11065). Written informed consent was waived due to the retrospective nature of the study.

**Consent to Participate** Not applicable.

**Consent to Publish** Not applicable.

**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

1. Marur, S., & Forastiere, A. A. (2016). Head and neck squamous cell carcinoma: update on epidemiology, diagnosis, and treatment. *Mayo Clinic Proceedings*, *91*(3), 386–396.
2. <https://syoparekisteri.fi/tilastot/tautilastot/>.
3. Specenier, P. M., & Vermorken, J. B. (2008). Recurrent head and neck cancer: Current treatment and future prospects. *Expert Review of Anticancer Therapy*, *8*(3), 375–391.
4. Kao, J., Vu, H. L., Genden, E. M., Mocherla, B., Park, E. E., Packer, S., et al. (2009). The diagnostic and prognostic utility of positron emission tomography/computed tomography-based follow-up after radiotherapy for head and neck cancer. *Cancer*, *115*(19), 4586–4594.
5. Loeffelbein, D. J., Souvatzoglou, M., Wankerl, V., Martinez-Möller, A., Dinges, J., Schwaiger, M., et al. (2012). PET-MRI fusion in head-and-neck oncology: Current status and implications for hybrid PET/MRI. *Journal of Oral and Maxillofacial Surgery*, *70*(2), 473–483.
6. Mehanna, H., Wong, W. L., McConkey, C. C., Rahman, J. K., Robinson, M., Hartley, A. G. J., et al. (2016). PET-CT surveillance versus neck dissection in advanced head and neck cancer. *New England Journal of Medicine*, *374*(15), 1444–1454.
7. Miller, F. R., Hussey, D., Beeram, M., Eng, T., McGuff, H. S., & Otto, R. A. (2005). Positron emission tomography in the management of unknown primary head and neck carcinoma. *Archives of Otolaryngology–Head & Neck Surgery*, *131*(7), 626–629.
8. Koshy, M., Paulino, A. C., Howell, R., Schuster, D., Halkar, R., & Davis, L. W. (2005). F-18 FDG PET-CT fusion in radiotherapy treatment planning for head and neck cancer. *Head and Neck*, *27*(6), 494–502.
9. Riegel, A. C., Berson, A. M., Destian, S., Ng, T., Tena, L. B., Mitnick, R. J., et al. (2006). Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion. *International Journal of Radiation Oncology Biology Physics*, *65*(3), 726–732.
10. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.
11. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer International Publishing.
12. Hwang, E. J., Park, S., Jin, K. N., Kim, J. I., Choi, S. Y., Lee, J. H., et al. (2019). Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Network Open*, *2*(3), e191095–e191095.
13. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.
14. Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, *172*(5), 1122–1131.e9.
15. Ren, J., Eriksen, J. G., Nijkamp, J., & Korreman, S. S. (2021). Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation. *Acta Oncologica*, *60*(11), 1399–1406.
16. Huang, B., Chen, Z., Wu, P. M., Ye, Y., Feng, S. T., Wong, C. Y. O., et al. (2018). Fully automated delineation of gross tumor volume for head and neck cancer on PET-CT using deep learning: A dual-center study. *Contrast Media & Molecular Imaging*, *2018*, 8923028.
17. Moe, Y. M., Groendahl, A. R., Tomic, O., Dale, E., Malinen, E., & Futsaether, C. M. (2021). Deep learning-based auto-delineation of gross tumour volumes and involved nodes in PET/CT images of head and neck cancer patients. *European Journal of Nuclear Medicine and Molecular Imaging*, *48*(9), 2782–2792.
18. Guo, Z., Guo, N., Gong, K., Zhong, S., & Li, Q. (2019). Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. *Physics in Medicine & Biology*, *64*(20), 205015–205015.
19. Rainio, O., Han, C., Teuho, J., Nesterov, S. V., Oikonen, V., Piirola, S., et al. (2023). Carimas: An extensive medical imaging data processing tool for research. *Journal of Digital Imaging*, *36*(4), 1885–1893.
20. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. :19.
21. Python [Internet]. Python.org. [cited 2021 Nov 3]. Available from: <https://www.python.org/>
22. Iakubovskii P. Segmentation Models [Internet]. GitHub; 2019. Available from: [https://github.com/qubvel/segmentation\\_models](https://github.com/qubvel/segmentation_models)
23. Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 9.
24. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, *35*(1), 53–65.