Autopet Challenge 2023: nnUNet-based whole-body 3D PET-CT Tumour Segmentation

Anissa Alloula[1], Daniel R McGowan[2,3][0000−0002−6880−5687], and Bartłomiej W. Papież[1]

[1] Big Data Institute, University of Oxford, England
[2] Department of Oncology, University of Oxford, England
[3] Department of Medical Physics and Clinical Engineering, Oxford University Hospitals NHS FT, Oxford, England. anissa.alloula@kellogg.ox.ac.uk

Abstract. Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) combined with Computed Tomography (CT) scans are critical in oncology to the identification of solid tumours and the monitoring of their progression. However, precise and consistent lesion segmentation remains challenging, as manual segmentation is time-consuming and subject to intra- and inter-observer variability. Despite their promise, automated segmentation methods often struggle with false positive segmentation of regions of healthy metabolic activity, particularly when presented with such a complex range of tumours across the whole body. In this paper, we explore the application of the nnUNet to tumour segmentation of whole-body PET-CT scans and conduct different experiments on optimal training and post-processing strategies. Our best model obtains a Dice score of 69% and a false negative and false positive volume of 6.27 and 5.78 mL respectively, on our internal test set. This model is submitted as part of the autoPET 2023 challenge. Our code is available at: https://github.com/anissa218/autopet_nnunet

1 Introduction

Positron Emission Tomography / Computed Tomography (PET/CT) imaging is a vital element of the diagnostic process and clinical management for a wide range of solid tumours [1, 2]. PET imaging, usually conducted with Fluorodeoxyglucose (FDG) radiotracer, a glucose analogue, provides metabolic information of areas with high glucose consumption, and these can be indicative of the presence of tumours [2, 3]. On the other hand, CT images provide anatomic information, which can help a clinician determine whether abnormal glucose uptake corresponds to a normal healthy metabolic organ or a malignant tumour [1]. The accurate detection and quantification of tumours is crucial for diagnosis, tumour staging and characterisation, and monitoring of disease evolution. Manual tumour segmentation is a costly and long process, which may be subject to intra- and inter- clinician variability [4]. For these reasons, automated segmentation of these tumours is of particular importance.

Previous research has shown that it is possible to accurately segment tumours with a variety of deep learning algorithms such as convolutional neural networks

like the U-Net or transformer-based architectures [5–8]. However, whole-body PET/CT tumour segmentation is a particularly challenging problem given the inherent multi-modality of the data as well as the extent of anatomical coverage and morphological variability of the tumours that can be present across the body[9]. Notably, last year's challenge results showed accurate segmentation of tumours in whole-body PET/CT scans, with a Dice score of 0.79 for the winning model [9]. However, performance of the algorithms on images acquired in a different hospital to the ones they were trained on was substantially inferior, particularly in terms of Dice accuracy [9]. This year's challenge, Autopet-ii, aims to extend this work by focusing on generalisation of the model to other acquisition protocols and sites. Generalisation beyond a single scanner or acquisition site is challenging because of domain shift, for instance due to different image resolutions, varying levels of noise, and spatial variations [10, 11]. This is also hindered by the lack of publicly available segmented PET/CT datasets on which robust models can be trained. To this end, the Autopet-ii organisers provide an extensive dataset of whole-body PET/CT scans of patients with and without tumours [12].

In this paper, we extend nnUNet, a semantic segmentation method which automatically adapts to a given dataset, and which has shown state-of-the art results across a wide range of medical image segmentation tasks [13]. We also investigate different post-processing methods to the improve final prediction. Indeed, nnU-Net based methods consistently outperformed other architectures in last year's challenge [9].

2 Methods

2.1 Data

The challenge provided a set of 1016 whole-body PET/CT scans publicly available on The Cancer Imaging Archive (TCIA) as well as the mask as segmented by two radiologists [12]. These were acquired from University Hospital Tübingen and University Hospital of the LMU with Siemens Biograph mCT, mCT Flow and Biograph 64, GE Discovery 690 PET/CT scanners. 900 patients were involved; approximately half had no cancer, and the other half presented with histologically-proven malignant melanoma, lymphoma, or lung cancer.

In addition, a hidden test-set of 205 images were used to evaluate the models, which was drawn in part from the same source distribution (1/4) and in part (3/4) from a different distribution. 5 of these images were used for preliminary testing, and the remainder will be used for final evaluation at the end of the challenge (not available to the authors at the time of submission).

One of the submitted models was also trained on a dataset including both the TCIA PET/CT scans and an additional 200 PET/CT head and neck scans from the Head and Neck Tumour Segmentation and Outcome Prediction in PET/CT 2022 challenge (HECKTOR) [14]. The aim was increase the diversity of the training data in order to improve model generalisability.

2.2 Pre-processing

DICOM files were resampled (CT to PET imaging resolution, ranging from 200-677x400x400) and normalised. This involves Z-score intensity normalisation for PET images and global dataset percentile clipping and Z-score normalisation for CT. Subsequent pre-processing was done according to standard nnUNet procedure [13]. The PET and CT images were concatenated into 2-channel 3D images with a median size of 2x236x400x400, and 90% were kept for training and cross-validation, 10% were kept as a held-out internal test set.

2.3 Architecture and training

A standard 3D full resolution U-Net was trained and cross-validated on a Tesla P100-SXM2-16GB GPU for 5 folds with the specifications shown in Table 1. PET-CT channels were concatenated and cropped to patches as input. For each batch of 2 patches, oversampling was implemented so that at least one of the two contained a positive label in the ground truth segmentation. This was done to ensure the model was trained with enough examples of lesions.

Table 1. Final training strategy

Training strategy	**Final Implementation**
Number of epochs	1500
Iterations per epoch	250
Batch size	2
Patch size	128x128x128
Foreground batch oversampling	>50%
Learning rate	0.0001
Learning scheduler	Poly learning rate schedule
Loss function	Soft dice loss and cross entropy loss

Standard nnUNet data augmentation was used during training. This includes random rotation, scaling, Gaussian noise and blur, brightness, contrast augmentation, gamma correction, simulation of low resolution, and mirroring.

2.4 Post-processing

After training, different methods of post-processing were evaluated, based on the removal of predicted tumour areas of low size. This was done through connected component analysis of positive voxels [15].

2.5 Evaluation

Inference was conducted on the held-out set of 101 images and all of the models were evaluated based on three metrics. At a later stage, the preliminary test set

(released by the organisers) of 5 images was also used for evaluation. For the images where tumours were present, Dice overlap score of the segmented lesions, as well as volume of false positive connected components that do not overlap with true positives, were measured. In all images, false negative volume was also evaluated. This corresponds to the volume of positive connected components in the ground truth label that did not overlap with any positive segmented mask.

3 Results

Models were trained with a variety of nnUNet parameters in order to determine the optimal configuration, and results are presented in Table 2. These results suggest that increasing training duration improved the model's performance on both the cross-validation and testing set. Removing mirroring as a data augmentation step also showed similar benefits, perhaps as there are important specifics to the organs on the left and right side of the PET/CT scans. Moreover, increasing the input patch size and maximum number of feature maps used in the model's architecture did not translate to better performance in the test set, perhaps due to over-fitting of the model. Finally, reducing the number of positive images presented to the network at each batch did not improve performance.

Table 2. Dice, false negative, and false positive scores on cross-validation (CV) images across all folds, as well as on 101 held-out internal test images. Removal of segmented regions of less than 10 connected voxels was also performed on the test predictions and metrics are shown in parentheses. Inference was not performed for the final g model due to time constraints.

		Cross-validation data			Internal testing data		
Model	Changes relative to baseline	Dice	FN	FP	Dice	FN	FP
baseline		0.716	8.345	11.916	0.664	7.651	11.883
h	1500 epochs	0.734	**7.430**	10.258	0.680	7.376	8.043
a	1500 epochs, no mirroring	0.732	8.265	8.577	**0.685**	**6.270**	5.778
e	1500 epochs, 0.01 lr	0.670	9.926	22.255	0.611	14.878	13.408
c	1500 epochs, 192 patch size, 512 features, 0.01 lr	**0.743**	13.960	**6.746**	0.673	13.807	3.788
d	192 patch size, 512 features	0.716	8.345	11.916	0.682	8.923	**3.239**
f	33% oversampling	0.714	7.93	13.2	0.660	8.190	12.412
g	10% oversampling	0.722	**7.227**	23.79			

The examination of the model's prediction showed that a high proportion of the segmentations were of small size (see Figure 1). As can be seen in Table

3, removing these tumours caused a minor reduction in false positive volumes compared to no removal (min size of 0). A threshold of 10 appeared as optimal, maintaining a high Dice score while slightly lowering false positive scores.

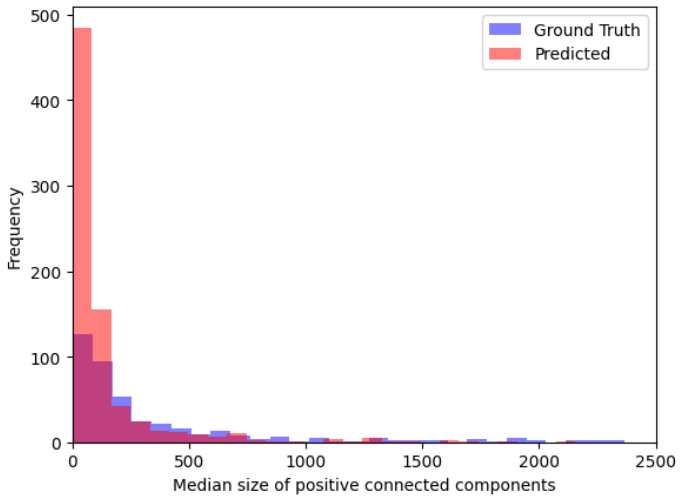

Fig. 1. Distribution of median positive connected component size across all training images. The purple bars show the sizes of the components in the ground truth labels while the red bars show the median sizes predicted by model "a".

Table 3. Effect of the removal of segmented lesions below a certain threshold size of connected components on the three challenge metrics in the test set.

Min size threshold	Dice	FP	FN
0	0.685	5.778	6.270
5	0.685	5.763	6.510
10	0.686	5.744	6.723
20	0.686	5.679	7.909
40	0.653	5.576	9.195
80	0.621	5.312	11.623

4 Discussion

As model "a" obtained the best results on the internal test set, it was submitted as our final model to the Autopet-2023 challenge.

However, in the future, many aspects remain to be investigated. Firstly, the model outputs high volumes of false positive and false negative components, despite having a relatively high Dice overlap score. Ensembling the predictions of multiple models of different architectures and trained on different modalities of input data (for instance just PET, just CT, or both) may help maximise robustness and minimise false predictions. Indeed, in last year's challenge and other similar medical image segmentation tasks, ensembles of multiple diverse models often outperform any individual model [9, 16, 17].

This includes developing an end-to-end model which predicts fewer very small false positive connected components and therefore does not require post-processing. Modifying the loss function, for instance with generalised Dice overlap, which weighs the contribution of each label by the inverse of its volume, may be a way to target this by penalising false small predictions [18].

Moreover, in order to build a more generalisable model, it would be of interest to quantify the uncertainty associated with the model's predictions. This would give an indication of how trustworthy each prediction is, which would be particularly useful when doing inference on out-of-distribution images [19–21]. Incorporating uncertainty during model training could also be beneficial by maximising the learning of examples which the model is uncertain about [20].

This work highlights the strength and adaptability of the nnUNet, which, with very little parameter tuning, accurately segmented lesions. This work and the Autopet-ii challenge represent crucial steps towards the development of reliable and robust PET/CT segmentation algorithms, with significant potential for valuable clinical application.

5 Acknowledgements

This work was supported by the EPSRC grant number EP/S024093/1 and the Centre for Doctoral Training in Sustainable Approaches to Biomedical Science: Responsible and Reproducible Research (SABS: R3) Doctoral Training Centre, University of Oxford. The authors acknowledge the AUTOPET challenge for the free publicly available PET/CT images used in this study. The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. Michael D. Farwell, Daniel A. Pryma, and David A. Mankoff. PET/CT imaging in cancer: Current applications and future directions. *Cancer*, 120(22):3433–3445, 2014.
2. Sandip Basu, Thomas C. Kwee, Suleman Surti, Esma A. Akin, Don Yoo, and Abass Alavi. Fundamentals of PET and PET/CT imaging. *Annals of the New York Academy of Sciences*, 1228(1):1–18, 2011.
3. Floriane Legot, Florent Tixier, Minea Hadzic, Thomas Pinto-Leite, Christelle Gallais, Rémy Perdrisot, Xavier Dufour, and Catherine Cheze-Le-Rest. Use of baseline 18F-FDG PET scan to identify initial sub-volumes with local failure after concomitant radio-chemotherapy in head and neck cancer. *Oncotarget*, 9(31):21811–21819, 2018.
4. Sara Sheikhbahaei Mohammad S. Sadaghiani, Steven P. Rowe. Applications of artificial intelligence in oncologic 18F-FDG PET/CT imaging: a systematic review. *Ann Transl Med*, 9(9):823, 2021.
5. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.
6. Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.
7. Eren M., Veziroglu MS, Faraz Farhadi BS, Navid Hasani BS, Moozhan Nikpanah MD, Mark Roschewski MD, Ronald M. Summers MD PhD, and Babak Saboury MD MPH. Role of artificial intelligence in PET/CT imaging forlymphoma diagnosis. *Seminars in Nuclear Medicine*, 53(3):426–448, 2023.
8. Hossein Arabi, Azadeh AkhavanAllaf, Amirhossein Sanaat, Isaac Shiri, and Habib Zaidi. The promise of artificial intelligence and deep learning in pet and spect imaging. *Physica Medica*, 83:122–137, 2021.
9. Matthias Fabritius et al. Sergios Gatidis, Marcel Früh. The autopet challenge: Towards fully automated lesion segmentation in oncologic pet/ct imaging. 2023.
10. Yixi Xu, Ivan Klyuzhin, Sara Harsini, Anthony Ortiz, Shun Zhang, François Bénard, Rahul Dodhia, Carlos F. Uribe, Arman Rahmim, and Juan Lavista Ferres. Automatic segmentation of prostate cancer metastases in PSMA PET/CT images using deep neural networks with weighted batch-wise dice loss. *Computers in Biology and Medicine*, 158:106882, 2023.
11. Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J. Wood, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 39(7):2531–2540, 2020.
12. Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Christina Pfannenberg Konstantin Nikolaou, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Sci Data*, 9(601), 2022.
13. F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

14. Vincent Andrearczyk, Valentin Oreiller, Mathieu Hatt, and Adrien Depeursinge. Head and neck tumor segmentation and outcome prediction. third challenge, HECKTOR 2022 held in conjunction with MICCAI 2022, singapore, september 22, 2022, proceedings. In *Lecture notes in computer science*, 2023.
15. William Silversmith. cc3d: Connected components on multilabel 3D and 2D images., 2021.
16. Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Andreea-Iuliana Miron. Diversity-promoting ensemble for medical image segmentation, 2022.
17. Vajira Thambawita, Steven A. Hicks, Pål Halvorsen, and Michael A. Riegler. Divergentnets: Medical image segmentation by network ensemble, 2021.
18. Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R.S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248, Cham, 2017. Springer International Publishing.
19. Ke Zou, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. Tbrats: Trusted brain tumor segmentation, 2022.
20. Ke Zou, Xuedong Yuan, Xiaojing Shen, Yidi Chen, Meng Wang, Rick Siow Mong Goh, Yong Liu, and Huazhu Fu. Evidencecap: Towards trustworthy medical image segmentation via evidential identity cap, 2023.
21. L. Huang, S. Ruan, P. Decazes, and T. Denœux. Evidential segmentation of 3d pet/ct images. In *International Conference on Belief Functions*, pages 159–167. Springer, 2021.