Computer Methods and Programs in Biomedicine 221 (2022) 106902

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Computer-aided detection and segmentation of malignant melanoma lesions on whole-body ^{18}F-FDG PET/CT using an interpretable deep learning approach

Ine Dirks[a,b,*], Marleen Keyaerts[c], Bart Neyns[d], Jef Vandemeulebroucke[a,b,e]

[a] Vrije Universiteit Brussel (VUB), Department of Electronics and Informatics (ETRO), Brussels, Belgium
[b] imec, Leuven, Belgium
[c] Vrije Universiteit Brussel (VUB), Universitair Ziekenhuis Brussel (UZ Brussel), Department of Nuclear Medicine, Brussels, Belgium
[d] Vrije Universiteit Brussel (VUB), Universitair Ziekenhuis Brussel (UZ Brussel), Department of Medical Oncology, Brussels, Belgium
[e] Vrije Universiteit Brussel (VUB), Universitair Ziekenhuis Brussel (UZ Brussel), Department of Radiology, Brussels, Belgium

ARTICLE INFO

Article history:
Received 10 December 2021
Revised 27 April 2022
Accepted 21 May 2022

Keywords:
Whole-body
^{18}F-FDG
PET/CT
Segmentation
Detection

ABSTRACT

Background and objective: In oncology, 18-fluorodeoxyglucose (^{18}F-FDG) positron emission tomography (PET) / computed tomography (CT) is widely used to identify and analyse metabolically-active tumours. The combination of the high sensitivity and specificity from ^{18}F-FDG PET and the high resolution from CT makes accurate assessment of disease status and treatment response possible. Since cancer is a systemic disease, whole-body imaging is of high interest. Moreover, whole-body metabolic tumour burden is emerging as a promising new biomarker predicting outcome for innovative immunotherapy in different tumour types. However, this comes with certain challenges such as the large amount of data for manual reading, different appearance of lesions across the body and cumbersome reporting, hampering its use in clinical routine. Automation of the reading can facilitate the process, maximise the information retrieved from the images and support clinicians in making treatment decisions. **Methods:** This work proposes a fully automated system for lesion detection and segmentation on whole-body ^{18}F-FDG PET/CT. The novelty of the method stems from the fact that the same two-step approach used when manually reading the images was adopted, consisting of an intensity-based thresholding on PET followed by a classification that specifies which regions represent normal physiological uptake and which are malignant tissue. The dataset contained 69 patients treated for malignant melanoma. Baseline and follow-up scans together offered 267 images for training and testing. **Results:** On an unseen dataset of 53 PET/CT images, a median F1-score of 0.7500 was achieved with, on average, 1.566 false positive lesions per scan. Metabolically-active tumours were segmented with a median dice score of 0.8493 and absolute volume difference of 0.2986 ml. **Conclusions:** The proposed fully automated method for the segmentation and detection of metabolically-active lesions on whole-body ^{18}F-FDG PET/CT achieved competitive results. Moreover, it was compared to a direct segmentation approach which it outperformed for all metrics.

1. Introduction

In oncology, disease assessment and treatment monitoring is commonly performed using positron emission tomography (PET) / computed tomography (CT) [1–3]. For PET, 18-fluorodeoxyglucose (^{18}F-FDG) is a widely used radiotracer. This glucose analogue indicates all areas with tracer buildup, including brain, bladder, certain regions of the abdomen and metabolically-active tumours. This functional information is joined with the anatomical knowledge provided by CT to specify the exact location of the areas that light up on the PET image. The high sensitivity and specificity of ^{18}F-FDG PET combined with the high resolution of CT allows for an accurate interpretation of disease status.

Malignant melanoma is the most lethal form of skin cancer. In 2020, it was responsible for over 300 000 new cases and over 63 000 deaths [4]. However, developments in new therapies with immune checkpoint inhibitors and targeted therapies have shown promising results [5–7]. Analysis of baseline and follow-up scans are imperative for proper evaluation of the response to treatment. Moreover, recent studies [8–11] have demonstrated that the images

* Corresponding author.
E-mail address: idirks@etrovub.be (I. Dirks).

https://doi.org/10.1016/j.cmpb.2022.106902

contain information that can contribute to treatment selection. Parameters like total metabolic tumour volume (TMTV) and total lesion glycolysis (TLG) hold predictive power for treatment response. Yet, these cannot be used in clinical practice because the current tumour segmentation procedure is labour intensive and time consuming. Also, the required user input makes the procedure subjective and leads to intra- and inter-observer variations. Automating the process can tackle these issues and assist in exploiting all available information, which is the subject of current work.

1.1. Related work

Several authors have proposed tools that partially or fully automate tumour segmentation from PET/CT imaging. Hirata et al. [12] proposed a semi-automated method to derive a reference standardised uptake value (SUV) from the liver. This was done by automatically placing a spherical volume of interest (VOI) of 30 mm diameter in a manually-drawn sphere enclosing the right liver lobe. The position was selected based on the coefficient of variation associated with a voxel (CV_v) on the PET image, defined as the standard deviation divided by the mean of the intensities included in a sphere surrounding the voxel. Although they achieved good agreement with the manual method, user interaction was still required and more difficult cases, like patients with more than two liver metastases, were excluded.

Gsaxner et al. [13] created a system for urinary bladder segmentation on CT using the FDG PET to automatically extract ground truth masks for the CNN. The method reached a true positive rate of 83.1 %, a true negative rate of 99.9 %, a dice score of 81.9 % and a Hausdorff distance of 11.9 pixels.

Zhao et al. [14] developed a lung tumour segmentation system consisting of separate V-nets for PET and CT, a summation step and four cascaded convolutional blocks. The method achieved a mean dice score of 0.85, a ratio of absolute volume difference to ground truth volume of 0.33 and a classification error of 0.15.

Zhong et al. [15] trained separate 3D U-Nets on PET and on CT from which the probability maps were combined into a final mask by a graph cut co-segmentation. Dice scores of 0.76 and 0.87 were obtained for PET and CT respectively.

Moe et al. [16] applied a U-Net to PET and CT separately as well as to the combined images of head-and-neck cancer patients. The ground truth segmentations were obtained by merging delineations of gross tumour volume by oncologists and extracted pathological lymph nodes. The model with both PET and CT achieved the best results with a dice score of 0.75, positive predictive value of 0.78, sensitivity of 0.74 and a specificity of 0.99.

Sibille et al. [17] examined the use of CNNs on whole-body ^{18}F-FDG PET/CT images to localise and classify uptake patterns into suspicious and non-suspicious regions. These classifications, together with the anatomical locations and a reconstructed maximum intensity projection (MIP) image served as inputs to the CNN. To obtain candidate regions, the volumes of interest extracted from the PET image were segmented using a fixed thresholding method [18]. The system reached an AUC of 0.98 with a sensitivity of 87.1% and specificity of 99.0% for lung cancer and a sensitivity of 75.4% and a specificity of 95.8% for lymphoma.

In a next step, reported by Capobianco et al. [19], this classification network was used to develop a fully automated method for TMTV estimation. The volumes of all regions classified suspicious by the aforementioned CNN were summed to a $TMTV_{PARS}$. The ground truth volumes were derived semi-automatically by 2 experienced nuclear medicine physicians. The remaining volumes were added to obtain $TMTV_{REF}$. For both the ground truth volumes and predicted volumes, an analysis was performed for progression-free survival and overall survival over four years. A log-rank test between the Kaplan-Meier curves decided if there were statistically significant differences between the ground truth and prediction. The ranked TMTV estimated showed significant correlation. The suspicious segmentations reached a median dice score across all patient of 73%, a median recall of 62% and median precision of 96%.

Li et al. [20] created DenseX-Net for lymphoma segmentation on 2D slices of whole-body ^{18}F-FDG PET/CT. The network consisted of two main pathways, one for supervised and one for unsupervised learning. The former handled feature extraction and semantic segmentation while the latter aimed to learn semantic representations in an unsupervised way by minimising the divergence between the input and output. During the joint training, per batch, one pathway was trained while keeping the other one fixed after which these trained convolution kernels were used to initialise the encoder layers of the other pathway. This was repeated while alternating between the pathways. In an ablation study comparing the proposed DenseX-Net to ConvU-Net [21], ConvX-Net, ResUnet, ResX-net and DenseU-Net [22], the DenseX-Net obtained the best dice score (0.7263) and recall (0.8079). ConvU-Net achieved the highest precision, 0.7599 compared to 0.7003 for the DenseX-Net. Also, with respect to other segmentation networks [23–25], dice and recall were highest with the DenseX-Net.

Jemaa et al. [26] developed a tumour segmentation system for whole-body ^{18}F-FDG PET/CT. It consisted of a modified U-Net to obtain 2D segmentations that were subsequently refined by one of three 3D V-Nets, depending on their anatomical location. For the latter, different networks were trained on patches extracted from the head and neck region, the chest and abdomen or the pelvis. These locations were determined with respect to the liver and lungs, which were localised through the method described in the work of Bauer et al. [27]. For a set of follicular Non-Hodgkin's lymphoma patients, the system reached an average dice score of 0.886, a voxel level sensitivity of 92.6 % and the derived total metabolic tumour volume had a Spearman's correlation of 0.97 with the ground truth, while for the SUV_{max} this was 0.96.

Kumar et al. [28] proposed a CNN for lesion detection and segmentation on FDG PET/CT from non-small cell lung cancer patients. Tumours were located based on the diagnostic imaging report written by an experienced imaging specialist and delineated through 40% peak SUV connected thresholding and manual adjustments. Ground truth for lungs and mediastinum were derived through adaptive thresholding and connected thresholding respectively. The CNN comprised two encoders, one for PET and one for CT, a co-learning part for feature fusion and a reconstruction component to derive the final segmentation. Evaluation was performed on foreground areas, including lungs, mediastinum and tumours, and other regions denoting for example high intensity PET noise. For tumour detection, the method reached 64.56% precision, 79.97% sensitivity, 99.89% specificity, 99.85% accuracy derived from the overlap of ground truth and prediction. For segmentation, a dice score of 63.85% was obtained.

Li et al. [29] proposed a system to segment tumours on PET/CT in non-small cell lung cancer. CT probability maps derived by a CNN were combined with the PET image through a fuzzy variational model. The method obtained a mean dice of 0.86, sensitivity of 0.86, positive predictive value of 0.87, a volume error of 0.16 and a classification error of 0.30.

Blanc-Durand et al. [30] applied the nnU-Net [31] to segment lesions for diffuse large B-cell lymphoma on whole-body FDG-PET/CT. For the ground truth, a 50 cm^3 sphere was manually placed in the liver. A PET threshold was applied, set at 1.5 times the mean liver uptake plus 2 times the standard deviation. The high-intensity regions were delineated at 41% of SUV_{max}. Two experienced physicians removed regions of physiological uptake, added low-intensity lesions, classified the lymphoma lesions and saved the locations. In a cross-validation, the network achieved a

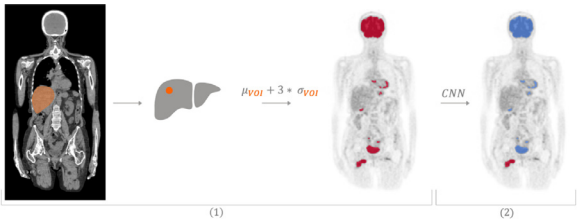

Fig. 1. Proposed two-step method. First, all regions of increased PET tracer uptake are segmented based on a threshold derived from an automatically identified VOI in the liver (1). Next, these areas are classified by a CNN as either physiological uptake (blue) or tumourous tissue (red) (2). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

mean dice of 73%, a mean Jaccard coefficient of 68% and, at voxel level, a sensitivity of 75%, specificity of 79%, positive predictive value of 83%, a negative predictive value of 99% and a difference in TMTV of 12 ml. For the independent test set, only the difference in TMTV was reported, which increased to 116 ml.

1.2. Goal and contributions of this study

The goal of this work is the development of an automated detection and segmentation procedure of metabolically-active lesions on whole-body ^{18}F-FDG PET/CT, with the aim to support medical experts in the disease assessment and treatment selection, specifically for malignant melanoma. Automation of the procedure renders it less time consuming and labour intensive and diminishes the inter- and intra-observer variability. This may render the detection of lesions more reliable, and could enable the use of additional measures such as TMTV and TLG for routine selection of the optimal therapy.

The main novelty of the approach lies in the design of a two-step procedure which closely resembles the manual reading approach performed by nuclear medicine specialists (but not feasible in clinical routine for whole-body images), and described in international guidelines, in order to facilitate the interaction between the system and the medical doctor. The proposed method exploits the highly sensitive nature of ^{18}F-FDG PET through an initial intensity-based candidate detection step. Next, a refinement is implemented in form of a false positive reduction step, incorporating also the CT as well as contextual information. The intermediate outputs allow to clarify how the system came to the final results, increasing the interpretability of the approach. Moreover, they enable the user to identify and correct errors when needed. The proposed procedure is compared to a direct segmentation method, currently state-of-the-art in the field, to evaluate the advantages and disadvantages of both.

2. Materials and methods

The proposed system, illustrated in Fig. 1, follows the general procedure used by nuclear medicine physicians. First, a spherical volume of interest is drawn in a uniform region of the liver. The intensities contained in the VOI are used to determine a threshold. All PET areas with an intensity exceeding this threshold are considered a potential lesion. This includes metabolically-active tumours, but also healthy tissue that consumes glucose, like the brain, bladder and parts of the abdomen. In the second step, these PET-positive areas are subjected to a classifier, in order to distinguish physiological tracer uptake from the final lesion segmentations.

2.1. Experimental data

The study included a retrospective analysis of 69 patients treated at Universitair Ziekenhuis Brussel (UZ Brussel, Brussels, Belgium) for malignant melanoma. Each patient received a baseline exam and between 0 and 9 follow-ups, resulting in a total of 267 PET/CT images. Most scans were acquired on a Philips Gemini TF (Koninklijke Philips N.V., Amsterdam, Netherlands) (108) or a Siemens Biograph mCT (Siemens Healthineers, Erlangen, Germany) (157). Two scans were taken on a GE Discovery 690 (GE Healthcare, Chicago, USA). The PET scans were converted to body-weight-corrected standardised uptake values (SUV_{bw}) according to

$$SUV_{bw} = \frac{C(t) * bw}{D}, \tag{1}$$

with $C(t)$ the activity concentration measured from the image acquired at time t [Bq/ml], bw the body weight of the patient [kg] and D the injected dose, corrected for the decay between tracer injection and scan acquisition [Bq]. On average, the injected dose was 276.4 ± 34.71 MBq [177.0 MBq - 634.0 MBq] and the body weight was 71.19 ± 15.72 kg [113 kg - 40 kg]. The PET images varied in size between 144 - 256 voxels left-right and anterior-posterior and 255 - 680 voxels inferior-superior with spacings between 2.73 mm and 4 mm. The CT images had a size of 512 voxels left-right and anterior-posterior and 204 - 2038 inferior-superior with spacings between 0.90 mm and 5 mm.

2.2. Ground truth annotations

The results of the proposed system were compared to the methods corresponding clinical practice in our institute. Though this is based on the PET Response Criteria in Solid Tumours (PERCIST), there are some differences. The ground truth tumour segmentations were derived semi-automatically by a physician using MIM Encore (MIM Software Inc., Cleveland, USA). A volume of interest of approximately 30 mm diameter was drawn in a homogeneous-looking part of the liver. On 48 scans this was more challenging due to the presence of liver metastases. In a few extreme cases, the VOI size was manually adapted to exclude any lesion tissue in the threshold calculation. The PET threshold was determined according to common clinical practice in the institute as

$$p = \mu_{VOI} + 3 * \sigma_{VOI}, \tag{2}$$

with p the threshold in SUV_{bw}, and μ_{VOI} and σ_{VOI} respectively the mean and standard deviation of the intensities in the liver VOI. Areas with an intensity above this threshold were manually classified as lesion or healthy tissue. Any volumes smaller than 1 ml were excluded. The VOI position, false positive removal and resulting lesion segmentations were checked by an experienced nuclear medicine expert. A lesion could be enlarged or reduced a few voxels to improve the delineation and, in a few cases, a low-intensity lesion was added. A continuous high-intensity region could contain both the lesion and healthy class.

2.3. Dataset division

The disease status of the patients was very diverse. The number of lesions per scan ranged from none to hundreds with total lesion volumes varying from 0 ml to more than 6500 ml. For training and validation of the classification step, stratified sampling ensured an equal distribution of these patients over the different data splits. The images were divided into five strata depending on the ratio of the number of lesions to the number of high-intensity areas. So, this ratio indicates which percentage of the total amount of patches contains a lesion. Moreover, the follow-ups of the same patient were included in the same set so that images of the same patient could not appear in different sets and lead to results are too optimistic. Also, a proportional distribution of the number of lesion candidates over the different sets was ensured. The training

Table 1
Division of the dataset.

Strata	# lesions / # patches	Train 60%	Validation 10%	Test 10%	Independent test 20%	Total
1	no lesions	90	10	14	29	143
2	1% - 10%	31	3	0	10	51
3	11% - 20%	16	6	7	6	29
4	21% - 30%	13	4	1	3	20
5	> 30%	15	2	2	5	24
Total # images		165	25	24	53	**267**
Total # patients		42	6	5	16	**69**
Total # lesion patches		594	111	90	198	**993**
Total # healthy patches		3761	599	645	1250	**6255**

set was used to train the weights of the CNN model. The validation set was used for assessing regularisation during training, evaluating the loss at the end of each epoch and determining the optimal epoch. Other hyperparameters, such as the optimal network topology, were tuned by looking at the results on the test set. The independent test set was only used three times: twice during the design of the CNN and once for generating the final results. Table 1 shows the division of the images over the different sets per strata and in total.

2.4. Threshold selection

In order to draw a VOI in the liver, this organ first has to be segmented. We previously proposed a method for automated liver and VOI localisation [32]. Here, a slightly modified approach is presented, optimised for robustness. In brief, the dense V-net developed by Gibson et al. [33] was altered to segment only the liver instead of eight different abdominal organs and the settings were tuned to work well with whole-body images. The network was retrained on the Liver Tumor Segmentation Challenge dataset [34] using an Adam optimiser with a learning rate of 0.001 in 3000 iterations. This previously proposed method [32] was extended by adding a morphological closing to remove any holes due to noise in the CT as well as a restriction to the right side of the body to manage leakage to the heart. Each possible position for the liver VOI inside this mask was judged based on its surrounding intensity variations. For both PET and CT, the standard deviation of the intensities enclosed in a 30 mm-diameter sphere was calculated and assigned to $\sigma_{\boldsymbol{x}}^{PET}$ and $\sigma_{\boldsymbol{x}}^{CT}$ respectively. Each voxel $\boldsymbol{x}$ was given a score

$$\sigma_{\boldsymbol{x}}^{total} = \sigma_{\boldsymbol{x}}^{PET} * \sigma_{\boldsymbol{x}}^{CT} \tag{3}$$

representing the homogeneity of the region.

To handle cases with liver lesions, voxels with a high intensity were excluded. Though there is no clear SUV threshold that can be determined, several studies [35–38] show that typically, healthy liver parenchyma will not exceed 5 SUV. To avoid the VOI being taken in a large, uniform lesion, voxels over 5 SUV_{bw} were excluded from the liver mask. Only positions where the VOI could be fully enclosed in the liver mask were considered. The voxel with the lowest $\sigma_{\boldsymbol{x}}^{total}$ was selected as centre of the VOI for the PET threshold calculation using (2).

2.5. False positive reduction

The PET and CT images were resampled to an isotropic voxel size of 4 mm and 2 mm respectively. To classify each area with a PET intensity higher than the threshold, a multimodal, multiscale CNN was trained. Centred on each ground truth lesion, patches of different sizes were extracted from both PET and CT and were labelled positive. After subtracting these lesions from the ground truth lesion candidate segmentation, similarly sized patches were extracted and labelled as negative. For each of the modalities, isotropic patches of 160 mm and of 80 mm were extracted to include contextual information. For PET, an additional set of zoomed in patches were extracted with a size of 40 mm in three dimensions. These extra patches were not included for CT as these smaller lesions are unlikely to cause anatomical changes perceivable on CT. Thus, in total, two sets of CT patches (160 mm and 80 mm) were collected and three sets of PET patches (160 mm, 80 mm and 40 mm). All patches were resampled to 20 x 20 x 20 voxels for PET and 40 x 40 x 40 voxels for CT. In the training set, there were about five times more patches with physiological uptake than with lesion tissue. This was balanced by five different translations on each lesion patch. Translations, rotations and flips were applied to increase the dataset to fifteen times its original size. This resulted in a balanced, augmented dataset of 168 300 patches of which 60 % were used for training.

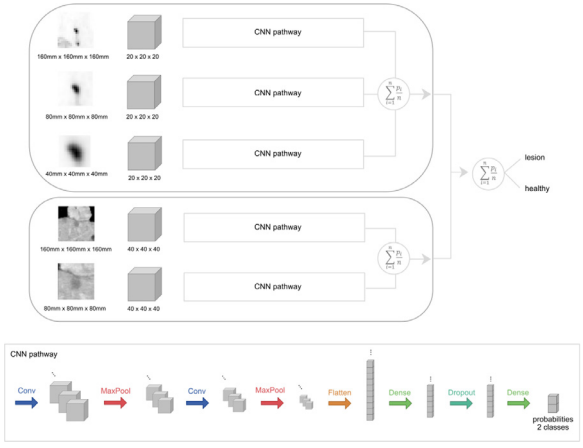

Fig. 2. Proposed multimodal, multiscale CNN.

The proposed network, shown in Fig. 2, consists of two main branches, one for PET and one for CT, so that the hyperparameters could be tuned to PET and CT separately as they contain very different information. Similarly, each branch handles the multiscale information through separate pathways allowing the parameters to be trained to the level of contextual information available in the patch. The pathways as well as the two main branches are fused through late averaging of the probability maps, thereby combining the information of the five different pathways, each associate to one patch type.

Each pathway consists of twice a convolutional step followed by a max pooling layer. After flattening the tensor, two dense layers, separated by a dropout layer, lead to a probability for the patch representing a lesion or healthy tissue. The optimised parameters included the amount of regularisation, the convolutional kernel size and stride, the number of filters per convolution, the

number of dense nodes in the final layer, the learning rate and the number of epochs.

2.6. Direct segmentation

The two-step procedure was compared to a direct segmentation approach. The open-source nnU-Net [31] was used as it won several recent medical image segmentations challenges and requires minimal model optimisation in terms of hyperparameters. The 3D network was trained and tested on the same dataset. Both PET and CT were resampled to an isotropic spacing of 4 mm and provided to the network in two channels. Patches of 128 x 128 x 128 voxels were extracted in a batch size of 2. An Adam optimiser was used instead of Stochastic Gradient Descend because this gave slightly better results. All other parameters were set as recommended in [31]. Similarly to the ground truth, any lesions with a volume smaller than 1 ml were excluded.

2.7. Evaluation

Since no ground truth liver segmentations were available and the method does not require a perfect segmentation, a visual check was performed to ensure a proper mask was created for every patient. The liver localisation network was trained on the publicly available Liver Tumor Segmentation Challenge dataset [34]. Therefore, the derivation of a suitable PET threshold was tested on the entire dataset of 267 cases. Evaluation of the automatically derived thresholds is not straightforward. The manually obtained thresholds can not strictly be considered ground truths as they are subject to intra- and interobserver variations. The automatically obtained values should be similar to those that were manually obtained, but are still expected to deviate within a normal range of observer variability. Both sets of thresholds were compared using Bland-Altman analysis to assess their relation and potential bias. To measure the agreement between the manual and proposed method, the intraclass correlation coefficient (ICC) was computed. For testing the full pipeline, the PET was thresholded with the automatically derived value to obtain the lesion candidates. Volumes smaller than 1 ml were excluded. Centred on each PET-positive connected component, the patches were extracted which were then given to the CNN to classify.

To assess the performance of the classification step, independently from the impact of the slightly different PET thresholds on the final tumour segmentations, we also evaluated the classification step by itself. For the latter, the patches for testing were extracted the same way as the training patches, using the manually derived thresholds.

In clinical routine, both the detection and segmentation of tumours are important. For detection, a lesion was counted as a true positive (TP) when the centroid distance between the ground truth and prediction was less than 10 mm. When the distance was larger or there was no prediction nearby, a false negative (FN) was added. Inversely, a prediction was considered a false positive (FP) if its centroid was more than 10 mm from the centroid of any ground truth lesion. In case there were multiple predictions within the required distance from a ground truth lesion, the closest one was counted as true positive. If there was no other lesion with which the predictions could correspond, the remaining predictions were false positives. Model performance was assessed with the F1-score, combining recall (R) and precision (P) in one metric according to

$$R = \frac{TP}{TP+FN}, \tag{4}$$

$$P = \frac{TP}{TP+FP}, \tag{5}$$

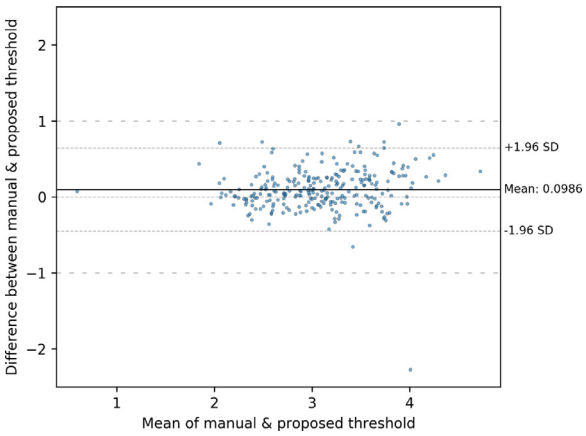

Fig. 3. Bland-Altman comparison of manually and automatically derived PET thresholds.

$$F1 = 2 * \frac{P*R}{P+R} = \frac{TP}{TP+0.5*(FP+FN)}. \tag{6}$$

For segmentation, the dice similarity coefficient (DSC) and absolute volume difference (AVD), defined as

$$DSC = \frac{2*TP}{2*TP+FP+FN}, \tag{7}$$

$$AVD = |V_{groundtruth} - V_{prediction}| \tag{8}$$

were evaluated. The true positives, false positives and false negatives were determined at voxel level. Cases without ground truth lesions were excluded from F1 and dice calculations.

3. Results

3.1. Threshold selection

The automatically and manually obtained thresholds are compared in a Bland-Altman analysis in Fig. 3. Except for one case, all thresholds were within 1 SUV_{bw} of the manually defined values. Including this outlier, the mean deviation of the thresholds was 0.0986 ± 0.279 SUV_{bw}. The mean absolute difference was 0.207 ± 0.212 SUV_{bw}. The average dice score between the segmentations obtained when thresholding the PET with the manually derived value and with the automatically derived threshold was 0.965. An ICC of 0.903 (95% CI: 0.880, 0.920, $p < 0.001$) indicated an excellent agreement [39] between both sets of thresholds.

The mean SUV_{bw} value in the liver VOI has an average deviation from the ground truth of 0.0127 ± 0.265 which is within the expected range of inter-observer variability [12,40].

For the outlier, thresholds of 2.87 SUV_{bw} and 5.14 SUV_{bw} were derived by the manual and proposed method respectively. The high value for the latter indicates that lesion tissue was included in the VOI. Indeed, as shown in Fig. 4, it was not possible to find a 30 mm-diameter liver VOI consisting of only healthy tissue. Verification revealed that the manually drawn VOI was given a smaller diameter to fit in between the lesion tissue.

3.2. Tumour detection and segmentation

The median detection and segmentation results are summarised in Table 2. A comparison is made with the results for the individual steps of the proposed method as well as the full pipeline and the direct segmentation approach.

The row corresponding to the full pipeline presents the results for the fully automated, two-step procedure.

Table 2
Median detection and segmentation results.

Evaluated on	Method	F1	# FP	DSC	AVD
Test set	Classification step	0.733 ± 0.204	2.00 ± 1.42	0.892 ± 0.352	4.00 ± 11.6
	Full pipeline	0.774 ± 0.191	1.50 ± 1.56	0.868 ± 0.334	5.41 ± 25.8
	nnU-Net	0.667 ± 0.199	2.00 ± 1.33	0.637 ± 0.293	6.92 ± 45.0
Independent test set	Classification step	0.857 ± 0.196	0 ± 3.58	0.900 ± 0.402	0 ± 187
	Full pipeline	0.750 ± 0.275	0 ± 4.73	0.849 ± 0.390	0.299 ± 187
	nnU-Net	0.634 ± 0.222	2.00 ± 2.91	0.500 ± 0.288	12.5 ± 86.1

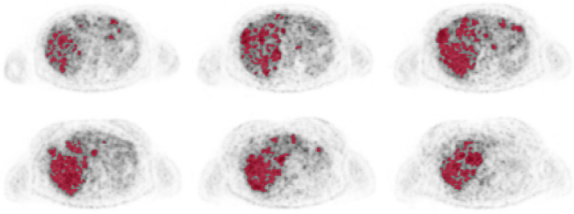

Fig. 4. Axial slices from the PET image considered an outlier in the threshold calculation with lesion overlay in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

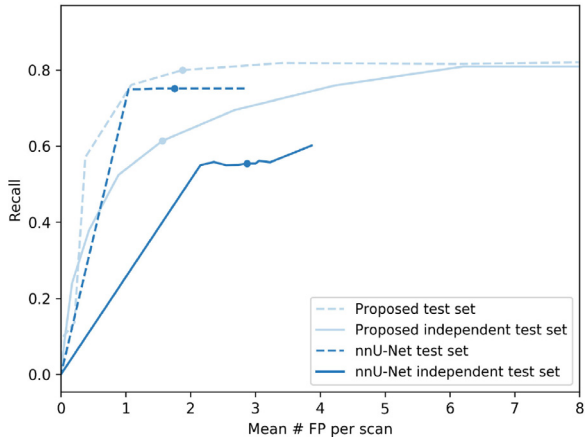

Fig. 5. Recall plotted against the mean number of false positives per scan for the proposed method and the nnU-Net evaluated on the test set and independent test set. The dots correspond to a lesion probability threshold of 50 %.

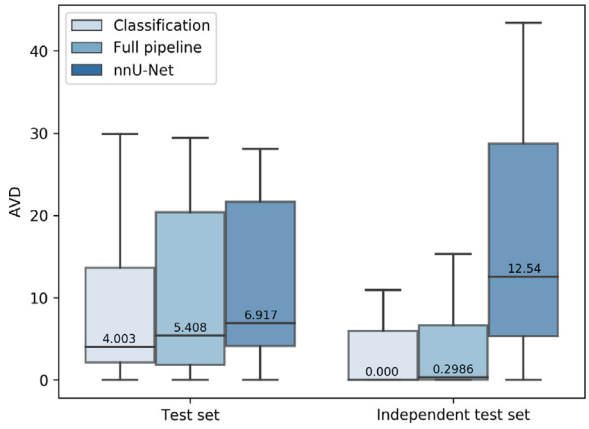

Fig. 6. Boxplots representing the absolute volume differences achieved by the classification step alone, the full pipeline and the nnU-Net for both the intermediate and independent test set. Outliers were omitted for visibility purposes.

The row corresponding to the classification step only allows to asses the impact of the difference in threshold, as it considers the results when applying the manually obtained thresholds and removing false positives with the trained CNN. In this case, the best results are obtained for most metrics, indicating excellent performance of the false positive reduction step.

The proposed full pipeline outperforms the direct segmentation approach in terms of detection and segmentation metrics. The difference between the methods for the independent test set was significant for AVD (Wilcoxon signed-rank test, $p = 0.00236$), but not for dice (Wilcoxon signed-rank test, $p = 0.0615$), partially due to a lower number of comparisons as cases with no ground truth lesions were excluded. The curves plotting recall against the average number of false positives per scan for both approaches are drawn in Fig. 5.

Overall, detection and segmentation metrics were slightly worse for the independent test set compared to the test set, and associated with larger variation. This may indicate some overfitting and/or differences in test and independent test set data distribution. The degradation in performance was largest for the direct segmentation approach. The distributions of the AVD results are summarised in Fig. 6. Considerably better results are obtained on the independent test set. Dice values were associated with large variations. This is expected to be due to the fact that this is a less appropriate metric for small structures [41].

Table 3 summarises the detection and segmentation results of an ablation study. A comparison is made between networks limited to PET images or to CT images, as well as using only 80 mm patches, only 160 mm patches, a combination of both and the proposed network. The proposed method leads to the best results for most metrics in case of the test set and was found to give a good overall compromise on the independent test set.

4. Discussion

A fully automated procedure for tumour detection and segmentation on whole-body ^{18}F-FDG PET/CT was proposed. In line with how manual reading is performed, a PET threshold is derived to delineate all regions with glucose retention. Next, the areas of physiological uptake are suppressed to obtain the final lesion segmentations. In addition to evaluating the performance of the system, it is of value to consider the output given to the user, illustrated in Fig. 7. Along with the final segmentations, our approach visualises the VOI selected in the liver, and the corresponding PET-positive regions classified as physiological uptake. These allow a better understanding of how the final tumour masks are obtained, giving the system a higher interpretability compared to a direct method. Moreover, supported by these additional visualisations, the system's performance can be reviewed and modified if necessary.

In terms of computational time, the proposed method is suited to be used in clinical practice. Once the entire model is trained, it takes on average 3 to 4 min to go from the DICOM PET/CT scans to the final lesion segmentations, which can be performed in the background. Adaptation to the liver VOI or excluded physiological uptake regions, leads to a new result in a matter of seconds.

Table 3
Median detection and segmentation results of the ablation study .

Evaluated on	Network	F1	# FP	DSC	AVD
Test set	PET only	0.690 ± 0.199	2.00 ± 1.71	0.663 ± 0.302	7.02 ± 43.1
	CT only	0.667 ± 0.230	3.00 ± 2.46	0.657 ± 0.295	10.5 ± 355
	PET/CT 80 mm patches	0.571 ± 0.198	3.00 ± 2.14	0.626 ± 0.308	19.9 ± 32.8
	PET/CT 160 mm patches	0.667 ± 0.176	2.00 ± 1.78	0.747 ± 0.267	3.76 ± 41.6
	PET/CT 80 mm & 160 mm patches	0.727 ± 0.190	1.50 ± 1.67	0.867 ± 0.333	6.04 ± 18.7
	Proposed	0.774 ± 0.191	1.50 ± 1.56	0.868 ± 0.334	5.41 ± 25.8
Independent test set	PET only	0.800 ± 0.283	1.00 ± 3.89	0.776 ± 0.382	2.44 ± 222
	CT only	0.667 ± 0.267	0 ± 5.90	0.668 ± 0.400	1.54 ± 268
	PET/CT 80 mm patches	0.720 ± 0.288	1.00 ± 5.66	0.862 ± 0.331	5.38 ± 209
	PET/CT 160 mm patches	0.776 ± 0.278	0 ± 4.06	0.862 ± 0.408	0 ± 207
	PET/CT 80 mm & 160 mm patches	0.760 ± 0.277	0 ± 5.09	0.865 ± 0.361	0.448 ± 189
	Proposed	0.750 ± 0.275	0 ± 4.73	0.849 ± 0.390	0.299 ± 187

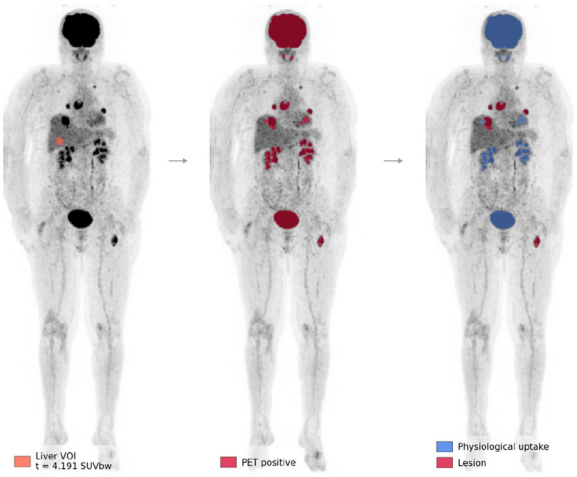

Fig. 7. Example showing the outputs of the proposed system on MIPs. The position of the liver VOI with the derived threshold (left), segmentations after thresholding (centre) and the classification into physiological uptake or lesion tissue (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In line with how the ground truths were created in our institute, the automated VOI selection process does not favour certain regions of the liver as would be the case when following PERCIST guidelines. From the conducted experiments, we observed that avoiding liver metastases and vessels when selecting a uniform region were more important than favouring a certain area of the liver. This is also supported by the work of Viner et al. [40].

The set of 267 automatically derived thresholds shows an excellent agreement with the manually acquired ones. A small bias towards lower values was observed due the fact that the automated system searches for the area in the liver with the lowest intensity variation, which is difficult to perform manually. The results include one outlier, due to the presence of extensive liver lesions. In such cases, a physician would reduce the VOI or look to the mediastinum to derive a blood pool value. Implementing such approaches could further improve the robustness of the method. Alternatively, the system could give an alert if the PET threshold or variation exceeds a predetermined value to draw the physician's attention to the fact that possibly lesion tissue was included in the VOI. A manual intervention can then be performed to adjust the VOI if necessary.

To remove the areas of natural glucose consumption from the lesion segmentations, a CNN was developed to classify each region of high PET intensity as healthy or malignant. The network consists of different pathways corresponding to different modalities and scales. The probability outputs were combined through late averaging to obtain a final decision. Earlier feature fusion was tested in various architectures, but these more complex models did not provide better performance, likely due to the lack of training data. To illustrate the need for false positive reduction, we can compare the result with those of simple thresholding. There are about ten times more false positives per scan (20.0 ± 6.61 for the test set, 16.0 ± 9.63 for the independent test set) with a median F1-score that is five to six times lower (0.133 ± 0.226 for the test set, 0.186 ± 0.219 for the independent test set) for the masks generated before the classification step.

A two-step approach was chosen over a direct lesion segmentation to maximise the supporting role the system can play for the clinical staff. This leads to some limitations over direct segmentation approaches. Firstly, no distinction can be made between healthy and malignant tissue depicted as one connected component after thresholding. Secondly, lesions with PET intensity below the threshold can not be detected. On the other hand, with a direct approach, a higher chance was observed for severe over- or undersegmentation per lesion and wrongful segmentation of low-intensity regions leading to additional false positives. Besides the network trained on images with isotropic 4 mm spacing, a higher resolution with 2 mm spacing was investigated as well. However, this gave considerably worse results which may be due to the following reason. The maximum patch size that could fit in the GPU memory was 128 x 128 x 128. While the higher level of detail due to the lower voxel spacing is expected to improve results, increasing the resolution led to a significant decrease in contextual information given to the network, which may have negatively impacted the performance. The results summarised in Table 2 and Figs. 5 and 6 show that the proposed full pipeline outperforms the direct segmentation performed by nnU-Net for all metrics. From this, the advantages of the proposed method seem to outweigh the drawbacks. Additionally, the proposed method offers segmentation of PET-positive regions that are physiological uptake as shown in Fig. 7. This further facilitates interaction with the user. On visual inspection of the segmentations, it takes less time to select one of these areas to include in the total tumour load than to manually delineate an extra lesion.

Fig. 8 provides a qualitative comparison of segmentation outputs for three different cases taken from the independent test set comparing the ground truth, the final outputs of the proposed method and the direct approach in maximum intensity projection (MIP) images. The top row contains the segmentations for a patient with a dice score close to the median value for both methods. For this case, the two-step approach achieves a dice score of 0.807 with an AVD of 6.66 ml while for the nnU-Net this is respectively 0.514 and 26.4 ml. There is one lesion present which is identified by both methods without any false positives. However, oversegmentation by the nnU-Net reduces the evaluation metrics.

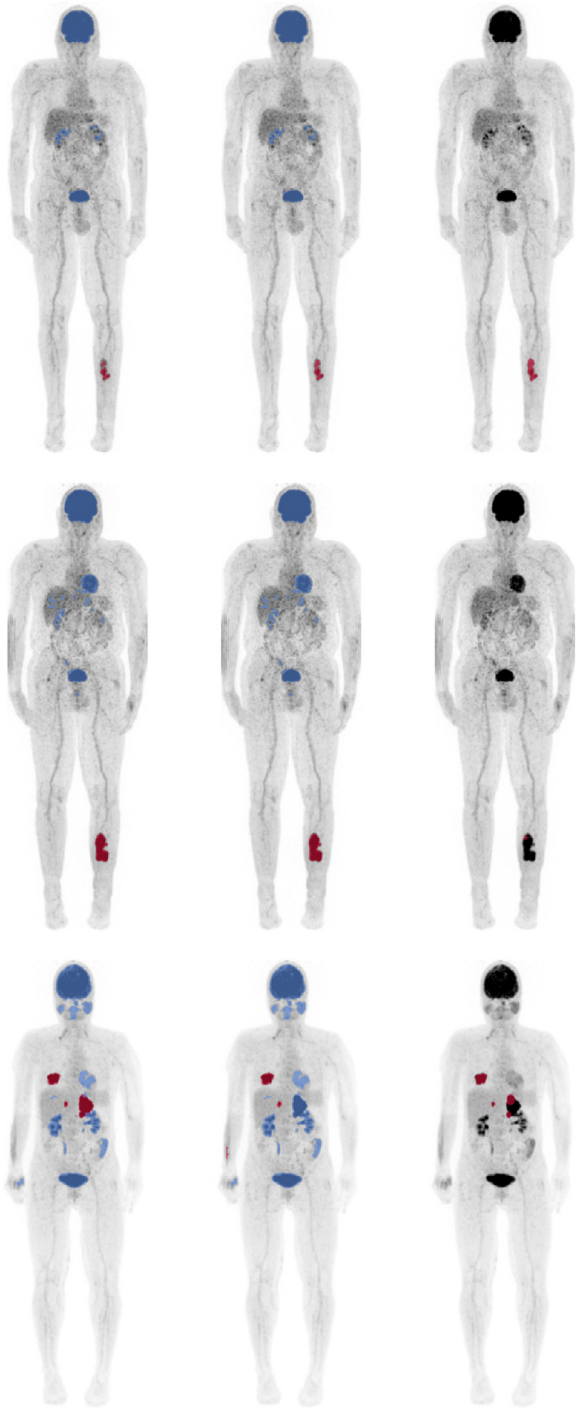

Fig. 8. Coronal MIP images showing the ground truth (left), final output of the proposed system (centre) and the outputs of the nnU-Net (right) for a case with a close-to-median dice score (top row), an example of an undersegmentation by the direct method (centre row) and a case with a tumour missed by the proposed method (bottom row). Areas of physiological uptake are indicated in blue, lesions in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The centre row contains the segmentations for a patient that illustrate the higher chance of severe undersegmentation within the same lesion for direct method. The proposed method reaches an AVD of 0.448 ml and a dice of 0.998 where the slight deviation is due to the use of a different PET threshold. The nnU-Net recognises the lesion, but severely undersegments it, resulting in an AVD of 131 ml and a dice score of only 0.0276.

The last row in Fig. 8 illustrates a disadvantage of the two-step approach. For this patient one lesion is missed, corresponding to only one wrongly classified patch. However, the patch belongs to a relatively large tumour, leading to an AVD of 134 ml and a dice score of 0.461. The direct segmentation method does pick up on this lesion. Though not the entire tumour is segmented, it achieves a better AVD of 109 ml and dice score of 0.606.

Table 2 shows a difference in performance between the smaller test set used for hyperparameter tuning and the unseen dataset. Though dice scores are more distributed, the median value is similar for the proposed method. The direct method shows a shift in dice score of 0.138 when going to the independent set. For the proposed method, the median AVD is better for the unseen data. This is partly due to the patients with no lesions. Out of 53 cases, there are 29 without any lesions. The proposed method correctly recognises 26 of them, while the direct method only identifies 4 images as lesion free. When excluding the cases without lesions, the median AVD for the independent set by the proposed method is 5.75 ml while this is 27.9 ml for the nnU-Net.

The results of the ablation experiment included in Table 3 indicate a clear benefit of each component of the proposed network. On the intermediate test set, the proposed network scored best for F1, number of false positives and dice. The difference with the best AVD value is only 1.65 ml. When evaluating on the independent test set, there is not one network that scores best on all metrics. However, the proposed method proofs to work well. Looking at the smaller test set, there is a clear increase in dice score when combining both modalities. The larger patches of 160 mm reach a higher dice score than the patches of 80 mm, though the value is highest when combining both scales. Including more context in the patch improves performance, but combining the information from different scales is even better. Adding the 40 mm PET patches only increases dice and F1 and decreases the AVD slightly. As expected, these patches are helpful for the cases with small lesions.

Though it is not possible to directly compare our results to those of methods reported in literature due to differences in pathologies and data, we can observe that the obtained performance is competitive, but does not outperform all previously reported works in terms of segmentation metrics. We consider current proposed method is advantageous for integration in clinical software as the approach and output are more suited for aiding the physician in the detection and segmentation.

Malignant melanoma is an aggressive disease that can metastasise anywhere in the body. Lesions can have any shape or volume and will have different appearances throughout the whole-body PET/CT scan. The study included patients at varying disease status, from lesion free to severely metastasised. That being said, the dataset of only 69 patients is considered the main limitation of this work. It is expected that increasing the dataset will further improve the results.

Properly identifying the PET threshold is important and can have a large impact on the final performance of the algorithm. Low-dose and low-contrast CT may negatively impact $\sigma_{\boldsymbol{x}}^{CT}$ of Eq. 3. Future research is needed to test the performance in such imaging settings.

5. Conclusion

A fully automated method was developed for the segmentation and detection of metabolically-active lesions on whole-body ^{18}F-FDG PET/CT. The proposed system consists of two steps, in line with clinical practice, in order to facilitate interpretation and interaction with clinicians and promote potential future integration in a clinical image analysis system. The system achieved competitive results in terms of detection and segmentation metrics, and outperformed a direct segmentation approach trained on the same

dataset. Future work should include the expansion of the dataset to improve the representation of possible disease states.

Declaration of Competing Interest

There are no competing interests to declare.

References

[1] J. Czernin, M. Allen-Auerbach, D. Nathanson, K. Herrmann, PET/CT In oncology: current status and perspectives, Curr. Radiol. Rep. 1 (3) (2013) 177–190, doi:10.1007/s40134-013-0016-x.

[2] J. Li, Y. Xiao, Application of FDG-PET/CT in radiation oncology, Front. Oncol. 3 (2013), doi:10.3389/fonc.2013.00080.

[3] M. Juweid, B. Cheson, Positron-Emission tomography and assessment of cancer therapy, N. Engl. J. Med. 354 (5) (2006) 496–507, doi:10.1056/nejmra050276.

[4] H. Sung, J. Ferlay, R. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 71 (3) (2021) 209–249, doi:10.3322/caac.21660.

[5] R. Mason, L. Au, A. Ingles Garces, J. Larkin, Current and emerging systemic therapies for cutaneous metastatic melanoma, Expert Opin. Pharmacother. 20 (9) (2019) 1135–1152, doi:10.1080/14656566.2019.1601700.

[6] Y. Jansen, E. Rozeman, R. Mason, S. Goldinger, M. Geukes Foppen, L. Hoejberg, H. Schmidt, J. van Thienen, J. Haanen, L. Tiainen, I. Svane, S. Mäkelä, T. Seremet, A. Arance, R. Dummer, L. Bastholt, M. Nyakas, O. Straume, A. Menzies, G. Long, V. Atkinson, C. Blank, B. Neyns, Discontinuation of anti-PD-1 antibody therapy in the absence of disease progression or treatment limiting toxicity: clinical outcomes in advanced melanoma, Ann. Oncol. (2019), doi:10.1093/annonc/mdz110.

[7] J. Berk-Krauss, J. Stein, J. Weber, D. Polsky, A. Geller, New systematic therapies and trends in cutaneous melanoma deaths among US whites, 1986–2016, Am. J. Public Health 110 (5) (2020) 731–733, doi:10.2105/AJPH.2020.305567.

[8] G. Awada, I. Özdemir, J. Schwarze, E. Daeninck, O. Gondry, Y. Jansen, T. Seremet, M. Keyaerts, H. Everaert, B. Neyns, Baseline total metabolic tumor volume assessed by 18FDG-PET/CT predicts outcome in advanced melanoma patients treated with pembrolizumab, Ann. Oncol. 29 (supplement 10) (2018), doi:10.1093/annonc/mdy493.019.

[9] G. Awada, J.K. Schwarze, O. Gondry, Y. Jansen, S. Ong, K.M. Gorman, S. Warren, M. Kockx, T. Seremet, M. Keyaerts, H. Everaert, B. Neyns, Baseline biomarkers correlated with outcome in advanced melanoma treated with pembrolizumab monotherapy, J. Clin. Oncol. 38 (15) (2020), doi:10.1200/JCO.2020.38.15_suppl.e22041.

[10] G. Awada, Y. Jansen, J. Schwarze, J. Tijtgat, L. Hellinckx, O. Gondry, S. Vermeulen, S. Warren, K. Schats, P. van Dam, M. Kockx, M. Keyaerts, H. Everaert, T. Seremet, A. Rogiers, B. Neyns, A comprehensive analysis of baseline clinical characteristics and biomarkers associated with outcome in advanced melanoma patients treated with pembrolizumab, Cancers (Basel) 13 (2) (2021) 1–18, doi:10.3390/cancers13020168.

[11] K. Vekens, H. Everaert, B. Neyns, B. Ilsen, L. Decoster, The value of 18F-FDG PET/CT in predicting the response to PD-1 blocking immunotherapy in advanced NSCLC patients with high-Level PD-L1 expression, Clin. Lung Cancer (2021) 1–9, doi:10.1016/j.cllc.2021.03.001.

[12] K. Hirata, K. Kobayashi, K. Wong, O. Manabe, A. Surmak, N. Tamaki, S. Huang, A semi-automated technique determining the liver standardized uptake value reference for tumor delineation in FDG PET-CT, PLoS ONE 9 (8) (2014), doi:10.1371/journal.pone.0105682.

[13] C. Gsaxner, P.M. Roth, J. Wallner, J. Egger, Exploit fully automatic low-level segmented PET data for training high-level deep learning algorithms for the corresponding CT data, PLoS ONE 14 (3) (2019), doi:10.1371/journal.pone.0212550.

[14] X. Zhao, L. Li, W. Lu, S. Tan, Tumor co-segmentation in PET / CT using multimodality fully convolutional neural network, Phys. Med. Biol. 64 (1) (2018), doi:10.1088/1361-6560/aaf44b.

[15] Z. Zhong, Y. Kim, L. Zhou, K. Plichta, B. Allen, J. Buatti, X. Wu, 3D Fully Convolutional Networks for Co-segmentation of Tumors on PET-CT Images, in: Physiology & behavior, 2018, pp. 228–231, doi:10.1109/ISBI.2018.8363561.

[16] Y. Moe, A. Groendahl, M. Mulstad, O. Tomic, U. Indahl, E. Dale, E. Malinen, C. Futsaether, Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers, in: Medical Imaging with Deep Learning, volume 1, 2019, pp. 1–4.

[17] L. Sibille, R. Seifert, N. Avramovic, T. Vehren, B. Spottiswoode, S. Zuehlsdorff, M. Schäafers, F-FDG PET / CT Uptake classification in lymphoma and lung cancer by using deep convolutional neural, Radiology (2019), doi:10.1148/radiol.2019191114.

[18] Y. Erdi, O. Mawlawi, S. Larson, M. Imbriaco, H. Yeung, R. Finn, J. Humm, Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding, Cancer 80 (12 SUPPL) (1997) 2505–2509, doi:10.1002/(SICI)1097-0142(19971215)80:12+<2505::AID-CNCR24>3.0.CO;2-F.

[19] N. Capobianco, M. Meignan, A. Cottereau, L. Vercellino, L. Sibille, B. Spottiswoode, S. Zuehlsdorff, O. Casasnovas, C. Thieblemont, I. Buvat, Deep-learning 18F-FDG uptake classification enables total metabolic tumor volume estimation in diffuse large B-Cell lymphoma, J. Nucl. Med. 62 (1) (2021) 30–36, doi:10.2967/jnumed.120.242412.

[20] H. Li, H. Jiang, S. Li, M. Wang, Z. Wang, G. Lu, J. Guo, Y. Wang, DenseX-Net: an end-to-End model for lymphoma segmentation in whole-Body PET/CT images, IEEE Access 8 (2020) 8004–8018, doi:10.1109/ACCESS.2019.2963254.

[21] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241, doi:10.1007/978-3-319-24574-4_28.

[22] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2017-July(2017) 1175–1183. 10.1109/CVPRW.2017.156

[23] Z. Xu, Z. Wu, J. Feng, CFUN: Combining Faster R-CNN and U-net Network for Efficient Whole Heart Segmentation, Technical Report, arxiv: 1812.04914, 2018.

[24] G. Zeng, X. Yang, J. Li, L. Yu, P. Heng, G. Zheng, 3D U-net with multi-level deep supervision: fully automatic segmentation of proximal Femur in 3D MR images, Machine Learning in Medical Imaging. MLMI 2017. Lecture Notes in Computer Science, volume 10541, 2017, doi:10.1007/978-3-319-67389-9_32.

[25] N. Souly, C. Spampinato, M. Shah, Semi supervised semantic segmentation using generative adversarial network, Proceedings of the IEEE International Conference on Computer Vision 2017-Octob (2017) 5689–5697. 10.1109/ICCV.2017.606

[26] S. Jemaa, J. Fredrickson, R. Carano, T. Nielsen, A. de Crespigny, T. Bengtsson, Tumor segmentation and feature extraction from whole-body FDG-PET/CT using cascaded 2D and 3D convolutional neural networks, J. Digit. Imaging (2020), doi:10.1007/s10278-020-00341-1.

[27] C. Bauer, S. Sun, W. Sun, J. Otis, A. Wallace, B.J. Smith, J.J. Sunderland, M.M. Graham, M. Sonka, J.M. Buatti, R.R. Beichel, Automated measurement of uptake in cerebellum, liver, and aortic arch in full-body FDG PET/CT scans, Med. Phys. 39 (6) (2012) 3112–3123, doi:10.1118/1.4711815.

[28] A. Kumar, M. Fulham, D. Feng, J. Kim, Co-Learning feature fusion maps from PET-CT images of lung cancer, IEEE Trans. Med. Imaging 39 (1) (2020) 204–217, doi:10.1109/TMI.2019.2923601.

[29] L. Li, X. Zhao, W. Lu, S. Tan, Deep learning for variational multimodality tumor segmentation in PET/CT, Neurocomputing 392 (2020) 277–295, doi:10.1016/j.neucom.2018.10.099.

[30] P. Blanc-Durand, S. Jégou, S. Kanoun, A. Berriolo-Riedinger, C. Bodet-Milin, F. Kraeber-Bodéré, T. Carlier, S. Le Gouill, R. Casasnovas, M. Meignan, E. Itti, Fully automatic segmentation of diffuse large B cell lymphoma lesions on 3D FDG-PET/CT for total metabolic tumour volume prediction using a convolutional neural network, Eur. J. Nucl. Med. Mol. Imaging (2020), doi:10.1007/s00259-020-05080-7.

[31] F. Isensee, P. Jaeger, S.A. Kohl, J. Petersen, K. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Method. 18 (2021) 203–211, doi:10.1038/s41592-020-01008-z.

[32] I. Dirks, M. Keyaerts, B. Neyns, J. Vandemeulebroucke, Automated threshold selection on whole-body 18F-FDG PET/CT for assessing tumor metabolic response, in: Proc. SPIE 11313, Medical Imaging 2020: Image Processing, 2020, p. 62, doi:10.1117/12.2549796.

[33] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. Pereira, M. Clarkson, D. Barratt, Automatic multi-organ segmentation on abdominal CT with dense V-networks, IEEE Trans. Med. Imaging 37 (2018) 8.

[34] P. Bilic, P.F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser, S. Kadoury, T. Konopczynski, M. Le, C. Li, X. Li, J. Lipkovà, J. Lowengrub, H. Meine, J.H. Moltz, C. Pal, M. Piraud, X. Qi, J. Qi, M. Rempfler, K. Roth, A. Schenk, A. Sekuboyina, E. Vorontsov, P. Zhou, C. Hülsemeyer, M. Beetz, F. Ettlinger, F. Gruen, G. Kaissis, F. Lohöfer, R. Braren, J. Holch, F. Hofmann, W. Sommer, V. Heinemann, C. Jacobs, G.E.H. Mamani, B. van Ginneken, G. Chartrand, A. Tang, M. Drozdzal, A. Ben-Cohen, E. Klang, M.M. Amitai, E. Konen, H. Greenspan, J. Moreau, A. Hostettler, L. Soler, R. Vivanti, A. Szeskin, N. Lev-Cohain, J. Sosna, L. Joskowicz, B.H. Menze, The Liver Tumor Segmentation Benchmark, (LiTS) (2019) https://competitions.codalab.org/competitions/17094.

[35] N. Paquet, A. Albert, J. Foidart, R. Hustinx, Within-patient variability of 18F-FDG: standardized uptake values in normal tissues, J. Nucl. Med. 45 (5) (2004) 784–788.

[36] A. Thie, Understanding the standardized uptake value, its methods, and implications for usage, J. Nucl. Med. 45 (9) (2004) 1431–1434.

[37] K. Perry, M. Tann, M. Miller, Which reference tissue is best for semiquantitative determination of FDG activity? J. Nucl. Med. 69 (supplement 1) (2008) 425.

[38] F. Hofheinz, R. Bütof, I. Apostolova, K. Zöphel, I.G. Steffen, H. Amthauer, J. Kotzerke, M. Baumann, J. van den Hoff, An investigation of the relation between tumor-to-liver ratio (TLR) and tumor-to-blood standard uptake ratio (SUR) in oncological FDG PET, EJNMMI Res. 6 (1) (2016), doi:10.1186/s13550-016-0174-y.

[39] J. Landis, G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1) (1977) 159–174.

[40] M. Viner, G. Mercier, F. Hao, A. Malladi, R. Subramaniam, Liver SULmean at FDG PET/CT: interreader agreement and impact of placement of volume of interest, Radiology 267 (2) (2013) 596–601, doi:10.1148/radiol.12121385.

[41] A. Reinke, M. Eisenmann, M. Tizabi, C. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M. Cardoso, V. Cheplygina, K. Farahani, B. Glocker, D. Heckmann-Nötzel, F. Isensee, P. Jannin, C. Kahn, J. Kleesiek, T. Kurc, M. Kozubek, B. Landman, G. Litjens, K. Maier-Hein, B. Menze, H. Müller, J. Petersen, M. Reyes, N. Rieke, B. Stieltjes, R. Summers, S. Tsaftaris, B. van Ginneken, A. Kopp-Schneider, P. Jäger, L. Maier-Hein, Common Limitations of Image Processing Metrics: A Picture Story, 2021, pp. 1–11. 2104.05642