EJNMMI Physics

ORIGINAL RESEARCH Open Access

Interobserver ground-truth variability limits performance of automated glioblastoma segmentation on [^{18}F]FET PET

Selene De Sutter[1*], Ine Dirks[1,5], Laurens Raes[2], Wietse Geens[3], Hendrik Everaert[2], Sophie Bourgeois[2], Johnny Duerinck[3] and Jef Vandemeulebroucke[1,4,5]

*Correspondence:
selene.de.sutter@vub.be

[1] Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Pleinlaan 9, Elsene, 1050 Brussels, Belgium
[2] Department of Nuclear Medicine, Vrije Universiteit Brussel (VUB), Universitair Ziekenhuis Brussel (UZ Brussel), Brussels, Belgium
[3] Department of Neurosurgery, Vrije Universiteit Brussel (VUB), Universitair Ziekenhuis Brussel (UZ Brussel), Brussels, Belgium
[4] Department of Radiology, Vrije Universiteit Brussel (VUB), Universitair Ziekenhuis Brussel (UZ Brussel), Brussels, Belgium
[5] Imec, Leuven, Belgium

Abstract

Background: Positron emission tomography (PET) with a [^{18}F]fluoroethyl)-L-tyrosine ([^{18}F]FET) tracer is of growing importance in the management of glioblastoma for the estimation of tumor extent and extraction of diagnostic and prognostic parameters. Robust and accurate glioblastoma segmentation methods are essential to maximize the benefits of this imaging modality. Given the importance of setting the foreground threshold during manual tumor delineation, this study investigates the added value of incorporating such prior knowledge to guide the automated segmentation and improve performance. Two segmentation networks were trained based on the nnU-Net guidelines: one with the [^{18}F]FET PET image as sole input, and one with an additional input channel for the threshold map. For the latter, we investigate the benefit of manually obtained thresholds and explore automated prediction and generation of such maps. A fully automated pipeline was constructed by selecting the best performing threshold prediction approach and cascading this with the tumor segmentation model.

Results: The proposed two-channel network shows increased performance with guidance of threshold maps originating from the same reader whose ground-truth tumor label the prediction is compared to (DSC = 0.901). When threshold maps were generated by a different reader, performance reverted to levels comparable to the one-channel network and inter-reader variability. The proposed full pipeline achieves results on par with current state of the art (DSC = 0.807).

Conclusions: Incorporating a threshold map can significantly improve tumor segmentation performance when it aligns well with the ground-truth label. However, the current inability to reliably reproduce these maps—both manually and automatically—or the ground-truth tumor labels, restricts the achievable accuracy for automated glioblastoma segmentation on [^{18}F]FET PET, highlighting the need for more consistent definitions of such ground-truth delineations.

Keywords: Brain, Glioblastoma, Positron emission tomography, Deep learning, Segmentation

Background

Positron emission tomography (PET) with amino acid tracers is increasingly recognized for their importance in the clinical management of glioblastoma, and has been included in the recommendations of the Response Assessment in Neuro-Oncology (RANO) group due to its additional value compared to magnetic resonance imaging (MRI) [1, 2]. One of the most commonly used amino acid tracers is O-(2-[^{18}F]fluoroethyl)-L-tyrosine ([^{18}F]FET), which, contrary to glucose tracers, shows relatively low uptake in normal brain tissue, allowing for good tumor-to-background contrast for glioblastoma imaging [3, 4].

[^{18}F]FET PET imaging displays significant potential across various disease stages for glioblastoma management. Reportedly, [^{18}F]FET PET allows a better estimation of the tumor extension compared to the contrast-enhanced boundaries found on MRI, a finding supported by multiple biopsy-controlled studies [4–6] and a volumetric study [7], which can enable enhanced tumor coverage in treatment planning [1]. In addition to mean and maximum tumor-to-background ratio (TBR_{mean} and TBR_{max}) [8], the metabolic tumor volume (MTV) was shown to be a strong prognostic factor for progression-free and overall survival, independent of the extent of resection [9]. Kinetic features derived from dynamic imaging have been shown to allow the differentiation between low- and high-grade glioma [10]. Both MTV and kinetic features have shown potential to serve as biomarkers predictive of the isocitrate dehydrogenase (IDH) mutation status [11, 12]. During response assessment, [^{18}F]FET PET can help differentiate between tumor progression and radiation-induced changes [13]. Moreover, studies have demonstrated that MTV changes can serve as early predictors for therapy response [14, 15].

Extraction of such parameters, however, requires a volume of interest (VOI) of the lesion and therefore introduces the need for accurate and reproducible segmentation of the tumor. The currently recommended workflow for such manual delineation [16] contains a thresholding step, which is based on the mean background activity. The mean background activity is extracted from a manually annotated background VOI in the hemisphere contralateral to the lesion [17]. To obtain the threshold, this value is multiplied with 1.6, a factor determined by a biopsy-guided study to optimally separate tumorous and healthy tissue on [^{18}F]FET PET [4]. Subsequently, manual correction and removal of non-tumorous tissue is applied on the thresholded image to achieve tumor segmentation. Expectedly, this process is prone to variations in acquiring the background VOI and subjectivity in the manual corrections, which in their turn can lead to significant differences in lesion boundaries. Unterrainer et al. [17] reported median inter-reader coefficients of variation of the background activity of 3.83%, 4.02% and 2.14%, assessed by circle-, sphere-, and crescent-shaped background VOIs, respectively, while a mean inter-reader difference in MTV of 4.1 mL and an overlap of Dice Similarity Score (DSC) equal to 0.68 was reported by Rahimpour et al. [18]. Additionally, the process poses a time-consuming task. Therefore, more robust and automated methods for glioblastoma segmentation are essential to fully exploit the benefits of [^{18}F]FET PET imaging for the management of this pathology.

Thus far, only a limited amount of research has been conducted on fully automated glioma segmentation on [^{18}F]FET PET. An initial feasibility study considering a small dataset ($n = 37$) revealed the potential of convolutional neural networks (CNN) for this

task, using a 3D U-Net and yielding a DSC of 0.79 and 0.82 without and with post-processing, respectively [19]. Rahimpour et al. [18] exploited a larger dataset ($n = 84$) and a full-resolution nnU-Net architecture [20] for a multi-label model which allowed the prediction of the brain mask and crescent shape background VOI in addition to the tumor contour, achieving a DSC of 0.746 and allowing the automated extraction of parameters such as TBR_{mean} and TBR_{max}. A large dataset ($n = 699$) was collected by Gutsche et al. [21] and used for the training of an nnU-Net. The study achieved a DSC of 0.75 and 0.81 on validation and test set, respectively.

Aforementioned studies demonstrated that automated segmentation using CNNs is feasible and can yield promising results. Considering the impact of determining the foreground threshold during manual tumor delineation, we hypothesize that inclusion of such knowledge in the segmentation process may improve its performance. Threshold awareness was implicitly included by Rahimpour et al. [18] through the inclusion of the background VOI as a segmentation label. For the segmentation of melanoma lesions on whole-body FDG-PET, Dirks et al. [22] showed that explicitly guiding the lesion segmentation, by adding the foreground threshold map as an additional input channel, significantly enhanced performance.

This study therefore explores the added value of guiding the segmentation by expanding the input with an additional channel that includes the threshold map, hypothesizing that this guides the network to focus on the most critical regions. Moreover, we explore multiple approaches for predicting the threshold and generating such threshold map automatically. Finally, we integrate our findings in a fully automated pipeline that incorporates both threshold prediction and tumor segmentation.

Materials and methods

Data and pre-processing

[18F]FET PET data

An in-house dataset was retrospectively collected, including patients with histopathologically confirmed glioblastoma who received a [^{18}F]FET PET scan between March 2004 and February 2023. Lesions smaller than 0.5 mL, defined as non-measurable disease in PET RANO 1.0 [23], were excluded. Demographics and imaging characteristics of the 174 glioblastoma subjects included in this study are summarized in Table 1.

Scans were acquired on a Siemens Biograph mCT 20 ($n = 45$), Biograph mCT 128 ($n = 84$) and ECAT 923 ($n = 23$) (Siemens Healthineers, Erlangen, Germany), and a Philips Gemini TF ($n = 22$) (Koninklijke Philips N.V., Amsterdam, Netherlands) system. Details on the images, their reconstruction and corrections can be found in Supplementary materials.

Manual ground-truth annotations

Ground-truth labels were created by four different experts according to current guidelines [16] using MIM Software version 7.3.3 (MIM Software Inc., Cleveland, USA), including two nuclear medicine physicians (HE, 35 years of experience; SB, 11 years of experience), one nuclear medical physicist (LR, 5 years of experience), and one neurosurgeon (WG, 4 years of experience). As illustrated in Fig. 1, ground-truth annotations for all patients were equally divided between these experts (annotation set A), and for

Table 1 Subject characteristics. Static PET images are acquired from 20 to 40 min post-injection, while dynamic images are summed 20–40 min post injection

		Overall, (n = 174)	Training set, (n = 125)	Test set, (n = 49)
Age	Mean ± STD [years]	62 ± 12	62 ± 13	61 ± 12
Sex	Male	122	94	28
	Female	52	31	21
Timepoint	Newly diagnosed	85	59	26
	Recurrent	89	66	23
Scanner	Siemens Biograph mCT 20	45	32	13
	Siemens Biograph mCT 128	84	63	21
	Siemens ECAT 923	23	15	8
	Philips Gemini TF	22	15	7
Image type	Static	91	63	28
	Dynamic	83	62	21
Injected dose	Mean ± STD [MBq]	197.10 ± 24.77	197.37 ± 25.66	196.42 ± 22.61
MTV	Mean ± STD (min–max) [mL]	32.10 ± 32.26 (0.68–203.80)	31.60 ± 32.56 (0.68–203.80)	33.37 ± 31.80 (1.18–145.69)

STD standard deviation, *MTV* metabolic tumor volume

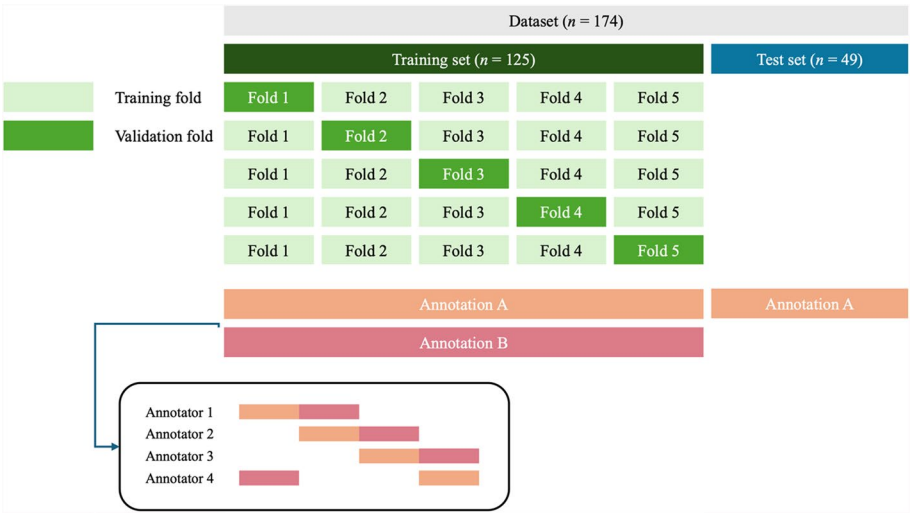

Fig. 1 Overview of data partitioning and annotation strategy, including fivefold cross-validation for training, and a fully independent test set for validation of the final model

each patient in the training set, segmentation was performed by a second expert to allow for analyzing inter-reader variability (annotation set B).

Firstly, a three-dimensional crescent background VOI with predetermined shape (with a cross-sectional diameter of 20 mm and a volume of 33 mL) was drawn in the hemisphere contralateral to the lesion covering both white and gray matter. From this, a threshold value was derived by multiplying the mean intensity of the background VOI with 1.6. Subsequently, the resampled PET image was thresholded according to this value, resulting in a binary threshold map. Lastly, non-tumorous tissue was excluded from the latter by manually defining a VOI around the tumorous region, arriving to the tumor segmentation label.

Models

Tumor segmentation with manual threshold guidance

The architecture of our models consisted of the full-resolution U-Net implementation as defined in the nnU-Net framework [20], similarly to [21]. The architecture was trained in two configurations. A first model (1 C-U-Net) used one input channel for the [^{18}F]FET PET image and predicted the tumor segmentation label at the output, while a second model (2 C-U-Net) relied on an additional input channel for the threshold map.

The training of 1 C-U-Net was equivalent to the network trained from Gutsche et al. [21], which adhered to the same guidelines [20], only with the exception of the patch size, allowing for comparison of our proposed networks' performances with the current state of the art. Moreover, their publicly available, pretrained network (JuST_BrainPET) was also evaluated on our data. Additionally, we implemented the approach described by Rahimpour et al. [18], namely a U-Net architecture with a multi-label output containing labels for the tumor, the crescent-shaped background and the whole brain (U-Net$_{\text{multi}}$).

Threshold estimation

To fully automate the use of 2 C-U-Net, the need for automated threshold prediction that allows the estimation of such threshold maps, was raised. In correspondence with the manual workflow, three methods of automated threshold prediction were explored, summarized in Fig. 2. Further details on training and parameters of these networks can be found in Sect. Training and implementation and the Supplementary Materials.

The first method (U-Net$_{\text{BKG}}$) was a segmentation network with the same U-Net architecture as for the tumor segmentation, which segmented the background VOI from the image. The model was optimized with respect to the loss computed compared to the manually annotated background VOIs. The threshold was subsequently determined by multiplying the mean background intensity with 1.6 and used to build the threshold map.

The second method (DenseNet$_{\text{TH}}$) was a regression network with a DenseNet121 [24] architecture, optimized using the threshold values extracted from the manually annotated background VOIs. The model estimated the threshold value directly from the image, which was then used to construct the threshold map.

The last method (U-Net$_{\text{TM}}$) consisted of another segmentation network with U-Net architecture, but in this case, the threshold map was segmented, and directly used as input for the tumor segmentation network. Training involved optimization of the loss computed compared to the manually attained threshold map segmentations. Corresponding threshold values were derived by taking the 5 th percentile of the intensity values contained within the threshold map.

Fully automated pipeline

The fully automated pipeline was constructed follows (see Fig. 2): firstly, the PET image was fed to the threshold prediction model, which automatically generated a threshold map. The PET image was then combined with the threshold map and

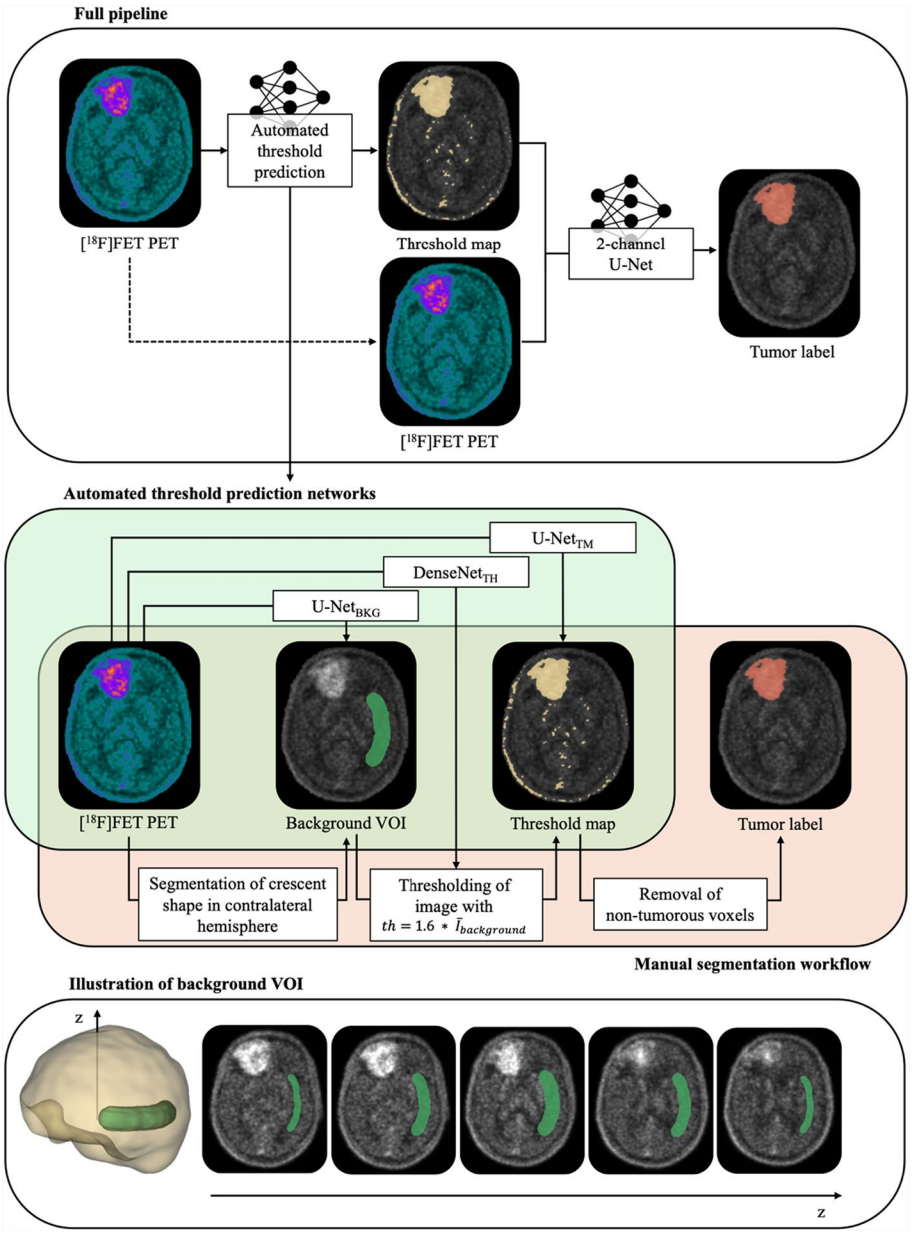

Fig. 2 Overview of the proposed approach. The full pipeline consists of initial prediction of a threshold map from the [^{18}F]FET PET image using an automated threshold estimation network. An overview of the investigated threshold prediction networks is shown (green) in correspondence the manual segmentation workflow (orange): from the PET image, U-Net$_{BKG}$ predicts the background VOI, DenseNet$_{TH}$ predicts the threshold value, and U-Net$_{TM}$ predicts the threshold map. The image and threshold map are subsequently fed as input channels to the segmentation network, a two-channel U-Net, for the prediction of the tumor label. A multi-slice representation of the background VOI is shown below. VOI = Volume Of Interest

inputted into the segmentation model (2 C-U-Net), resulting in the tumor segmentation. The final pipeline was chosen using the best performing threshold prediction model, selected as the model with the highest performance for the eventual tumor

segmentation when its predicted threshold map was used in combination with the 2 C-U-Net.

Training and implementation

Approximately 30% of the data ($n = 49$) was set apart for testing of the final model, while the remainder ($n = 125$) was used for fivefold cross-validation, as illustrated in Fig. 1. Training and hyperparameter tuning were conducted using the latter, where the training set was evenly divided into 5 folds, with each fold subsequently used for validation. All sets were stratified based on tumor volume. Models were trained on the training set with annotation labels A as ground truths and using cross-validation with 5 folds for 1000 epochs. Predictions for the test set were acquired through averaging of the predictions of the networks of each fold.

All segmentation networks, including those for tumor, background and threshold map segmentation, were trained according to the guidelines described in [20]. The patch size was 128 ×128x128 mm^3, while the batch size was 2. The loss function consisted of a combination of Dice and cross-entropy loss. An SGD optimizer with Nesterov momentum ($\mu = 0.99$) with an initial learning rate of 1e-2 and a decay according to $(1 - epoch/epoch_{max})^{0.99}$. Images were normalized using Z-scoring. No dropout was used, and data augmentation included random flipping along the sagittal plane, random rotations along all three axis and random zooming.

The regression network for threshold prediction was trained using an Adam optimizer and an initial learning rate of 1e-4 and a weight decay of 1e-5. The loss function consisted of L1 Loss. Images were resized to 128 ×128 ×128 mm^3 and intensities were clipped at 0 and 5, and rescaled to [0,1]. Data augmentation was similar to the segmentation networks.

All models were trained on an Nvidia A100 Ampere GPU with 40 GB RAM together with a 16-core AMD EPYC 7282 CPU. MONAI version 1.3.0 [25] was used for model implementations.

We refer to the full list of hyperparameters in Supplementary Materials. Moreover, implementation for training and trained models developed in this work are made publicly available.[1]

Evaluation

Segmentation performance

In accordance with the recommendations in [26], tumor segmentation predictions of each model were evaluated using the DSC, Normalized Surface Dice (NSD), and Absolute Volume Error (AVE), calculated in comparison with the manual ground-truth segmentations (annotation set A). NSD requires the tuning of tolerance parameter τ [27]. To evaluate boundary alignment between the predictions and ground truths, τ corresponded to the voxel spacing (1 mm). Additionally, to compare network performances with respect to inter-reader variability, the inter-reader NSD (IR-NSD) was defined by setting the τ value equal to the smallest value for which the median inter-rater NSD

[1] https://github.com/sdesutter/GBM-Segmentation-PET.

equaled 1. Given the skewed nature of our data, we reported the median and inter-quartile range (IQR) for each metric.

Furthermore, an inter-reader analysis was conducted on the subjects within the training set. For each pair of reader annotations, two thresholds were extracted from both background labels, MTVs were calculated from both tumor labels and DSC, NSD, and AVE were calculated between both manually delineated tumor volumes.

The 2C-U-Net was evaluated using the threshold maps from the annotations it was trained on (threshold maps A), as well as from the second set of annotations (threshold maps B). Predictions from both were compared to the tumor labels of both readers (A and B) to assess the impact of reader variability on the threshold guidance. Moreover, for each threshold prediction model, the threshold map was derived and fed to the tumor segmentation model (2 C-U-Net), allowing evaluation of the influence of the threshold prediction on the segmentation performance. Finally, the full pipeline was applied on the test set and compared to the state-of-the-art implementations.

A Wilcoxon signed-ranks test with $\alpha = 0.05$ was conducted to assess statistical significance when comparing models. In case of multiple comparisons with one model, Bonferroni correction was applied by adjusting the significance level considering the number of comparisons.

Detection performance

To assess the network's performance in tumor detection, the full pipeline was evaluated accordingly for both validation and test set. True positive (TP) was defined as cases where a prediction's bounding box overlapped with the bounding box surrounding the sphere drawn by the annotator to select tumorous voxels. Any degree of overlap was considered sufficient to classify a detection as a TP, as long as the volume of the predicted lesion exceeded 0.5 mL. This lenient criterion was chosen to account for the variability in lesion boundaries and to focus on the overall detection of lesions rather than the lesion's subregions, for which we found this analysis to be the most representative. False negative (FN) was defined as cases where the ground-truth bounding box had no overlap with predicted bounding boxes, while false positive (FP) was defined as a predicted bounding box outside this region (with a lesion volume exceeding 0.5 mL).

Due to the absence of negative scans without uptake above the 1.6-threshold required to compute true negative (TN) values, an approximation was derived by isolating the healthy hemisphere. Cases where no prediction exceeding the threshold volume of 0.5 mL was present in this region were considered TN. Subjects with bilateral lesions or lesions located in the midline were excluded from this approximation and calculation of any related metrics.

Results

The performance in tumor segmentation of the different models and inter-reader metrics on the validation set, are reported in Table 2. These results include the tumor segmentations predicted from the 2C-U-Net using the manually obtained threshold maps for combinations of both set of labels, as well as the threshold maps generated by the various threshold prediction models, as described in Sect. Models. Tuning the IR-NSD based on inter-reader variability yielded a τ value of 17 mm.

Table 2 Performance metrics for tumor segmentation on validation set

	Threshold map*	GT †	DSC ↑	NSD ↑ (τ = 1 mm)	IR-NSD ↑ (τ = 17 mm)	AVE [mL] ↓
1 C-U-Net	–	A	0.772 ± 0.248	0.612 ± 0.382	0.999 ± 0.050	4.89 ± 7.97
2 C-U-Net	Manual A	A	0.901 ± 0.167	0.804 ± 0.243	0.998 ± 0.042	2.05 ± 4.42
	Manual A	B	0.768 ± 0.285	0.630 ± 0.414	0.999 ± 0.025	4.02 ± 7.10
	Manual B	A	0.765 ± 0.284	0.601 ± 0.463	0.999 ± 0.042	4.83 ± 9.37
	Manual B	B	0.884 ± 0.189	0.781 ± 0.272	0.999 ± 0.024	2.00 ± 5.75
	$U\text{-}Net_{BKG}$	A	0.745 ± 0.307	0.619 ± 0.428	0.998 ± 0.063	4.95 ± 9.39
	$DenseNet_{TH}$	A	0.643 ± 0.373	0.400 ± 0.521	0.988 ± 0.100	9.03 ± 16.98
	$U\text{-}Net_{TM}$	A	0.799 ± 0.217	0.672 ± 0.306	0.999 ± 0.038	4.23 ± 7.61
Inter-reader	–	–	0.787 ± 0.211	0.652 ± 0.439	1.000 ± 0.017	4.27 ± 9.37

Reported numbers are median ± IQR

GT ground truth, *DSC* dice similarity coefficient, *(IR-)NSD* (inter-reader) normalized surface dice, *AVE* absolute volume error

*Origin of threshold map used for inference

†Ground truth label to which prediction is compared

Tumor segmentation with manual threshold guidance

The 2 C-U-Net with manually attained threshold maps significantly outperforms its one-channel counterpart for DSC, NSD and AVE ($p < 0.006$) when using labels from the same readers, increasing the DSC from 0.772 up to 0.901. Additionally, the metrics are significantly higher than those of the inter-reader analysis ($p < 0.006$, except for AVE using labels B), where a DSC of 0.787 is found. However, when combining labels from different readers, using one for the threshold map at the input and another for the compared ground-truth tumor label, DSC of 0.768 and 0.765 are found, and results remain on par with the one-channel method ($p > 0.006$). Metrics are also on par with inter-reader variability when using threshold maps A ($p > 0.006$), and lower when using threshold maps B ($p < 0.006$), with the exception of AVE. In terms of IR-NSD, no significant differences are found ($p > 0.006$).

Threshold estimation

The automatically obtained thresholds and MTV's of the predicted segmentations are compared with the manual thresholds (i.e., thresholds derived from the manually annotated background VOI) and ground-truth MTV's in a Bland–Altman plot, and set against the inter-reader differences, see Figs. 3 and 4. The performance of the underlying methods with respect to the segmentation of background VOI, threshold prediction and threshold map segmentation is provided in Supplementary Materials.

Thresholds extracted using $U\text{-}Net_{BKG}$ align reasonably well with the ground-truth thresholds, illustrated by the small bias and narrow limits of agreement in Fig. 3a, and are in line with the inter-reader threshold variability (see Fig. 3d). However, diminished tumor segmentation performance (DSC equals 0.745, see Table 2), also appreciable in the wider limits of agreement concerning MTV prediction in Fig. 4a, are found. Closer analysis revealed the model failed to segment a region entirely in some instances (5 out of 125), preventing the extraction of a threshold and further guidance of the segmentation, resulting in a threshold being assigned as 0 instead. All five cases

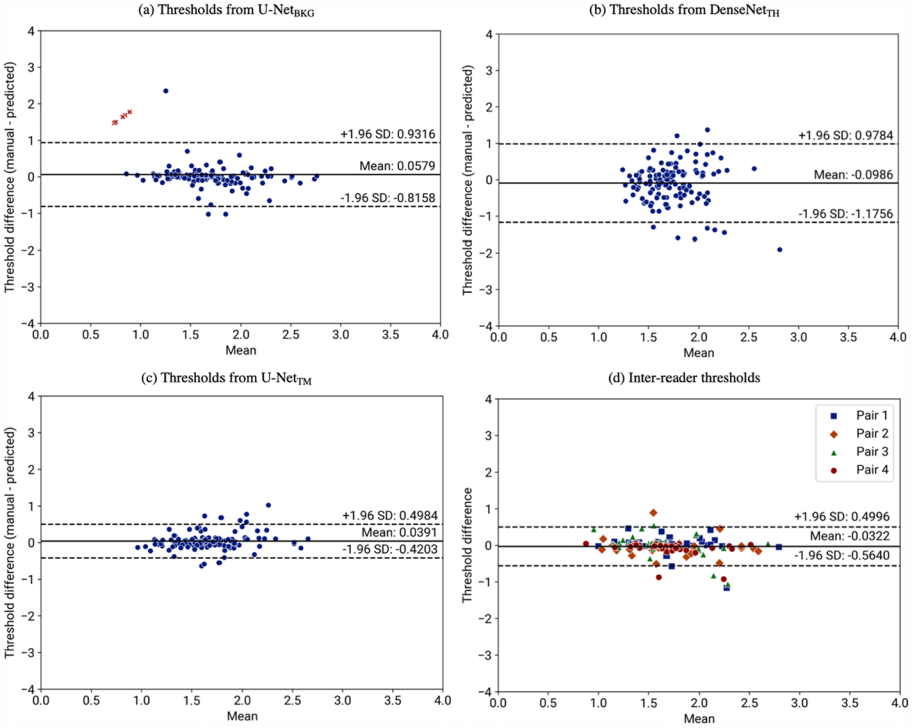

Fig. 3 Bland–Altman plots illustrating the differences between ground-truth thresholds and thresholds predicted using the various approaches for automated threshold prediction. Each point corresponds to a pair of predicted and ground-truth threshold values. Inter-reader differences are shown in (d) for different pairs of readers (1–4), where each point corresponds to a pair of threshold values, both determined by a different reader. The plots display the mean difference (bias) and 95% limits of agreement. Red crosses in (a) indicate cases where the network failed to segment a background VOI, resulting in a threshold set to 0. SD = Standard Deviation

originated from the same scanner (Siemens Biograph mCT 128), with no apparent image artifacts. One lesion showed low contrast and two were located near the midline. However, we hypothesize that the primary reason for failure lies in the absence of a consistent anatomical structure contained within the background volume, which may limit the model's ability to learn a stable segmentation pattern.

In Fig. 3b, a larger bias with wider limits of agreement is found for DenseNet$_{TH}$. With a DSC of 0.643, tumor segmentations using this method are found to have the lowest performing metrics (see Table 2) and worst agreement in MTV (see Fig. 4b). This method is the only one that demonstrates significantly lower performance compared to the others in terms of IR-NSD ($p < 0.006$).

U-Net$_{TM}$ shows high agreement in threshold value prediction (see Fig. 3c), with results that are in line with inter-reader variability. In terms of tumor segmentation, U-Net$_{TM}$ achieves a DSC of 0.799 and performs best among the automated threshold prediction methods ($p < 0.006$), see Table 2 and Fig. 4c. This method performs significantly worse compared to when using the manually obtained threshold map with same-reader labels ($p < 0.006$) and similar with respect to that method evaluated with different-reader labels ($p > 0.006$). The method performs similar to the inter-reader metrics ($p > 0.006$).

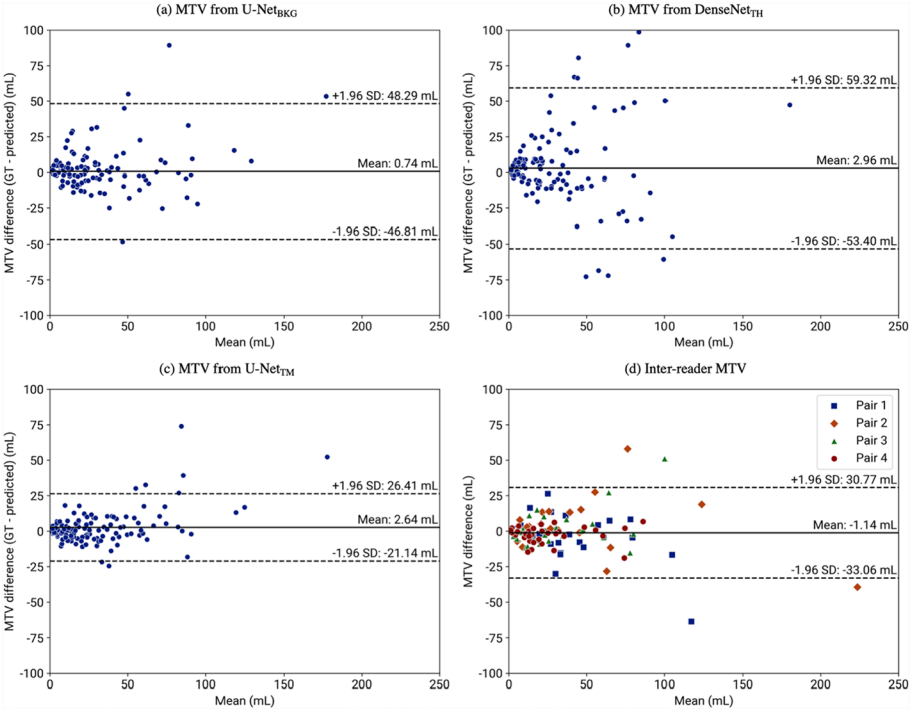

Fig. 4 Bland–Altman plots illustrating the differences between ground-truth MTV and MTV predicted using 2 C-U-Net with threshold maps generated by the various automated threshold prediction approaches. Each point corresponds to a pair of predicted and ground-truth volumes. Inter-reader differences are shown in (d) for the different pairs of readers (1–4), where each point corresponds to a pair of tumor volumes, both determined by a different reader. The plots display the mean difference (bias) and 95% limits of agreement. GT = Ground Truth; MTV = Metabolic Tumor Volume; SD = Standard Deviation

Table 3 Performance metrics for tumor segmentation on test set

	DSC ↑	NSD ↑ ($\tau = 1$ mm)	IR-NSD ↑ ($\tau = 17$ mm)	AVE [mL] ↓
Full pipeline (proposed)	0.807 ± 0.227	0.699 ± 0.368	1.000 ± 0.017	4.33 ± 7.52
1 C-U-Net (retrained)	0.805 ± 0.167	0.716 ± 0.320	0.999 ± 0.019	4.44 ± 5.83
U-Net$_{multi}$ (retrained) [18]	0.817 ± 0.169	0.699 ± 0.321	1.000 ± 0.014	4.11 ± 7.13
JuST_BrainPET (pretrained) [21]	0.760 ± 0.302	0.632 ± 0.625	0.999 ± 0.061	7.28 ± 17.67

Reported numbers are median ± IQR

DSC dice similarity coefficient, *(IR-)NSD* (inter-reader) normalized surface dice, *AVE* absolute volume error

Fully automated pipeline

Segmentation performance

Following the results above, the full pipeline was constructed by combining 2 C-U-Net with automatically obtained threshold maps from U-Net$_{TM}$. Subsequently, it was applied on the test set and compared to the state-of-the-art implementations, for which metrics are reported in Table 3. The full pipeline achieves a DSC of 0.807 and performance is found statistically on par with the state-of-the-art methods ($p > 0.017$), with the exception of the AVE metric when comparing the proposed network to U-Net$_{multi}$. Significantly higher performance was found in comparison to JuST_BrainPET ($p < 0.017$).

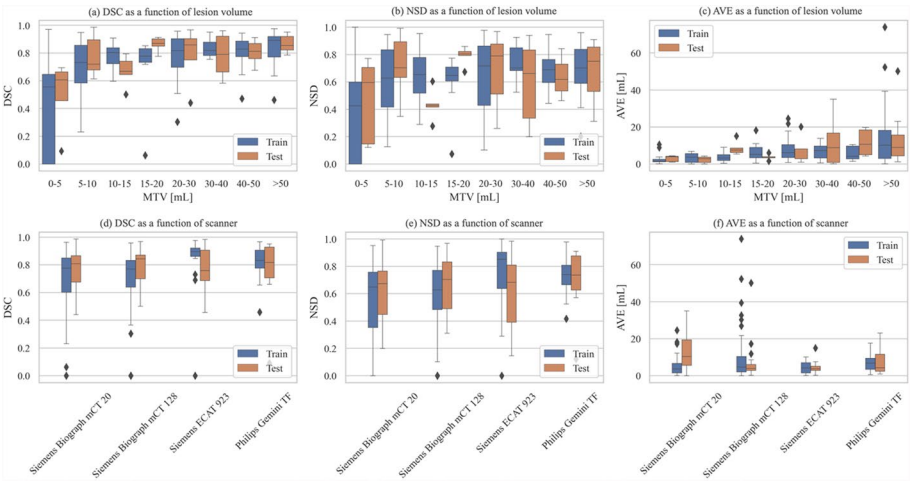

Fig. 5 Performance of the full pipeline as a function of lesion volume (**a**–**c**) and scanner (**d**–**f**). DSC = Dice Similarity Coefficient; MTV = Metabolic Tumor Volume; NSD = Normalized Surface Dice; AVE = Absolute Volume Error

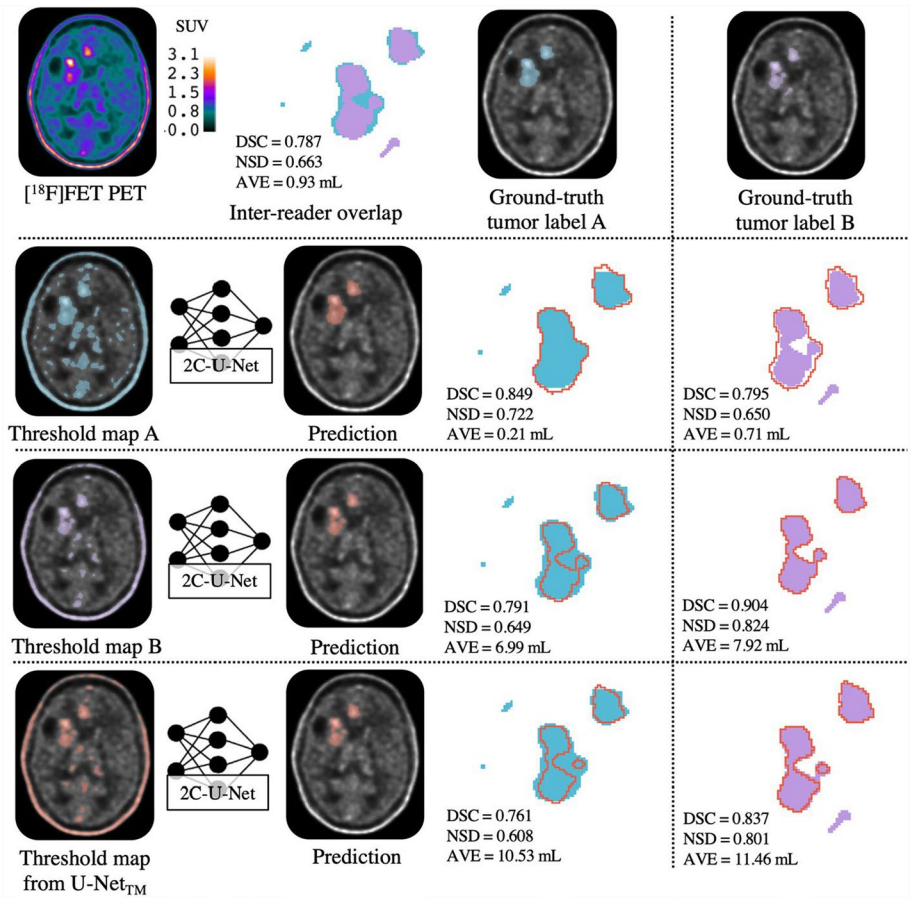

Fig. 6 Example segmentations of representative subject. Threshold map from reader A, from reader B and automatically generated from U-Net$_{TM}$ are shown in the first column. Tumor label predictions from 2 C-U-Net using these threshold maps are shown in the second column and compared to ground-truth labels of both readers. Overlap between labels of both readers are visualized with corresponding metrics. AVE = Absolute Volume Error; DSC = Dice Similarity Coefficient; NSD = Normalized Surface Dice

Additionally, performance of the full pipeline in function of lesion volume and different scanners is reported in Fig. 5.

Per illustration, a representative subject (with inter-reader variability DSC equal to the median), along with the corresponding threshold maps and segmentations, is shown in Fig. 6.

Detection performance

In the validation set, which comprised 132 lesions from 125 patients, the full pipeline correctly identified 124 lesions. The model produced 10 FP, including segmentation of vascular structures in two cases and sinusoidal uptake in one case, while the remaining corresponded to uptake regions not identified by the reader. Additionally, 8 FN were observed, with four cases involving missed lesions in patients with multiple lesions. Notably, 7 out of 8 FN cases were considered small lesions (volumes between 1 mL and 3.8 mL). This resulted in a sensitivity of 94%, a precision of 93%, and an F1-score of 93%. Using the healthy-hemisphere approximation for TN, the specificity and accuracy were determined to be 89% and 92%, respectively.

The test set contained 52 lesions from 49 patients, of which all were correctly identified. Among five FP, two involved vascular structure segmentation, while the others corresponded to uptake regions not annotated by the reader. For this set, the model achieved a sensitivity of 100%, a precision of 91%, and an F1-score of 95%. Specificity and accuracy were found to be 96% and 98%, respectively.

Discussion

In this work, we have explored the added value of threshold guidance for automated tumor segmentation from [^{18}F]FET PET imaging.

Results in Table 2 show that the addition of a threshold map at the input can strongly boost the performance of the tumor segmentation network, as long as the threshold map originates from the same reader whose ground-truth tumor label the prediction is compared to. When threshold maps were generated by a different reader, performance reverted to levels comparable to the one-channel network and inter-reader variability. While inter-reader DSC was found higher than reported in literature (0.787 compared to 0.68), AVE was similar (4.27 mL compared to 4.1 mL) [18].

Among threshold prediction methods, U-Net$_{TM}$ demonstrated superior performance, achieving better tumor segmentation metrics, and the closest alignment in threshold value and MTV to expert annotations. Therefore, the full pipeline was constructed by combining 2 C-U-Net with said threshold map prediction, achieving an eventual median DSC, NSD and AVE of 0.807, 0.699 and 4.33 mL, respectively. Performance of the full pipeline is compared to the state of the art in Table 3, demonstrating that our proposed network performs on par to previously proposed approaches. The increased performance of our model compared to JuST_BrainPET is likely attributable to differences in training data, as both models were trained on data from their respective institutions. Our data may better represent the test set distribution, thereby influencing performance.

These results indicate that threshold guidance may be of added value for automated glioblastoma segmentation, when the threshold map sufficiently aligns with the ground-truth label of the lesion, a finding that can be appreciated in Fig. 6. The figure

demonstrates that 2 C-U-Net effectively learns to select the tumor region within a provided threshold map, leading to high performance when comparing predictions to same-reader ground-truth tumor labels. In contrast, when evaluated using a threshold map from a different origin, the alignment of the prediction with this threshold map results in weak alignment with the ground-truth tumor. While this might be interpreted as reduced model performance, it in fact highlights the underlying disagreement between threshold maps from different readers. Indeed, the reproducibility of such maps—whether generated manually or through automated methods trained on manual annotations—remains limited. This inter-reader variability emerges as a key limitation for automated glioblastoma segmentation from [1^8F]FET PET, as it affects both the consistency of training and the reliability of evaluation, ultimately imposing a ceiling on achievable performance. Notably, on the test set, our proposed method and the state-of-the-art techniques achieve statistically similar performance, touching upon said upper limit. This is confirmed by observing the IR-NSD, which was configured to account for the uncertainty due to inter-reader variability, and thus measures how well predictions fall within the expected margins of variability between ground-truth segmentations. All methods achieve near-perfect IR-NSD scores, suggesting that the high variability in ground truths constrains the ability to detect meaningful performance differences.

From evaluating the proposed full pipeline in function of lesion volume and scanner type, lower and more variable performance was found for lower volumes, as appreciable in Fig. 5, indicating the model can have difficulties in correctly identifying small lesions. Across the training set, a slight bias towards images acquired on a Siemens ECAT 923 system was found, although this bias did not translate throughout the test set.

The pipeline demonstrates robust performance in tumor detection, achieving high sensitivity (94% and 100% for the training and test sets, respectively), precision (93% and 91%), and F1-scores (93% and 95%). The observed false negatives were primarily associated with small tumor sizes, again suggesting a potential limitation in detecting smaller lesions. A limited number of false positives were attributed to the segmentation of vascular structures, while the majority resulted from uptake regions not annotated by the reader. This again highlights the potential influence of inter-reader variability in ground truth segmentation, which may contribute to these discrepancies. Furthermore, the model achieved high specificity (89% and 96%) and accuracy (92% and 98%); however, these values should be interpreted with caution, as true negatives were estimated using an approximative method.

Our findings underscore the need for more consistent definitions for manual tumor boundary delineation on [^{18}F]FET PET imaging. While consensus-based annotations from multiple readers could partly reduce this bottleneck, more detailed guidelines or semi-automated procedures aiding in the placement of the background VOI and removal of false positive voxels should be investigated. Better reproducibility achieved as such would likely enable automated threshold estimation and segmentation methods to show higher agreement with the ground truth. It remains to be assessed whether in such cases threshold-guided approaches show a benefit over direct segmentation. However, threshold prediction models may hold value beyond segmentation-oriented applications, for example, in estimating diagnostic and prognostic parameters such as TBR_{mean} and TBR_{max}.

A limitation of our work lies in the uni-institutional nature of our dataset. Although [^{18}F]FET PET imaging is of growing importance for glioblastoma management, the available data remains scarce and large, openly available benchmark dataset initiatives with data from multiple centers that allow comprehensive training and comparable evaluation of algorithms, are lacking. Another limitation is the lack of negative scans in our dataset, i.e., those with lesions below the clinically established threshold or without increased tracer uptake, which restricts the evaluation of the network in terms of its specificity and false positive rate.

Conclusion

In this work, we investigated the added value of threshold-map guidance for automatic [^{18}F]FET PET-based glioblastoma segmentation. Our results demonstrate that incorporating a threshold map can significantly enhance segmentation performance when well aligned with the ground-truth label. Direct segmentation of the threshold map provided the most robust method for its automatic generation and integrating both in a full pipeline generated results comparable to the current state of the art (DSC = 0.807). However, the lack of reproducibility of threshold maps, whether generated manually or automatically, and ground truth tumor segmentations on [^{18}F]FET PET remains a major limitation, limiting the overall segmentation accuracy.

Abbreviations

AVE	Absolute volume error
CNN	Convolutional neural networks
DSC	Dice similarity score
DH	Isocitrate dehydrogenase
FN	False negative
FP	False positive
IQR	Inter-quartile range
MRI	Magnetic resonance imaging
MTV	Metabolic tumor volume
NSD	Normalized surface dice
PET	Positron emission tomography
RANO	Response assessment in neuro-oncology
SUV	Standardized uptake value
TBRmax	Maximum tumor-to-background ratio
TBRmean	Mean tumor-to-background ratio
TN	True negative
TP	True positive
VOI	Volume of interest
[18 F]FET	O-(2-[18F]fluoroethyl)-L-tyrosine

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40658-025-00767-y.

Additional file 1.

Acknowledgements
Not applicable.

Author contributions
All authors contributed to the study conception and design. Data collection, preparation and annotation were performed by LR, WG, HE and SB. Data analysis and model implementations were performed by SDS. The first draft of the manuscript was written by SDS and all authors commented on previous ersions of the manuscript. All authors read and approved the final manuscript.

Funding
This study was funded by the EU-funded HosmartAI project (grant number 101016834). Resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government.

Data availability
The datasets generated and/or analysed during the current study are not publicly available due to ethical/privacy reasons.

Declarations

Ethics approval and consent to participate
This single-center retrospective study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Universitair Ziekenhuis Brussel (Commissie Medische Ethiek; protocol code EC-2021–137; date of approval 28–07-2021). This study is a retrospective analysis of data obtained during a prospective study (Axig (NCT01562197), GliAvAx (NCT03291314), and GlitIpNi (NCT03233152)), during which all patients signed informed consent for the use of their data.

Consent for publication
The authors affirm that human research participants provided informed consent for publication of the images in Figs. 2 and 6.

Competing interests
The authors declare that they have no competing interests.

Received: 9 December 2024 Accepted: 9 May 2025
Published online: 06 June 2025

References

1. Galldiks N, Niyazi M, Grosu AL, et al. Contribution of PET imaging to radiotherapy planning and monitoring in glioma patients—a report of the PET/RANO group. Neuro Oncol. 2021;23(6):881–93. https://doi.org/10.1093/neuonc/noab013.
2. Albert NL, Weller M, Suchorska B, et al. Response assessment in neuro-oncology working group and European association for neuro-oncology recommendations for the clinical use of PET imaging in gliomas. Neuro Oncol. 2016;18(9):1199–208. https://doi.org/10.1093/neuonc/now058.
3. Pauleit D, Stoffels G, Bachofner A, et al. Comparison of 18F-FET and 18F-FDG PET in brain tumors. Nucl Med Biol. 2009;36(7):779–87. https://doi.org/10.1016/j.nucmedbio.2009.05.005.
4. Pauleit D, Floeth F, Hamacher K, et al. O-(2-[18F] fluoroethyl)-L-tyrosine PET combined with MRI improves the diagnostic assessment of cerebral gliomas. Brain. 2005;128(3):678–87. https://doi.org/10.1093/brain/awh399.
5. Pöpperl G, Götz C, Rachinger W, Gildehaus F-J, Tonn J-C, Tatsch K. Value of O-(2-[18F] fluoroethyl)-L-tyrosine PET for the diagnosis of recurrent glioma. Eur J Nucl Med Mol Imaging. 2004;31:1464–70. https://doi.org/10.1007/s00259-004-1590-1.
6. Floeth FW, Pauleit D, Wittsack HJ, et al. Multimodal metabolic imaging of cerebral gliomas: positron emission tomography with [18F]fluoroethyl-L-tyrosine and magnetic resonance spectroscopy. J Neurosurg. 2005;102(2):318–27. https://doi.org/10.3171/jns.2005.102.2.0318.
7. Lohmann P, Stavrinou P, Lipke K, et al. FET PET reveals considerable spatial differences in tumour burden compared to conventional MRI in newly diagnosed glioblastoma. Eur J Nucl Med Mol Imaging. 2019;46:591–602. https://doi.org/10.1007/s00259-018-4188-8.
8. Celli M, Caroli P, Amadori E, et al. Diagnostic and prognostic potential of 18F-FET PET in the differential diagnosis of glioma recurrence and treatment-induced changes after chemoradiation therapy. Front Oncol. 2021;11: 721821. https://doi.org/10.3389/fonc.2021.721821.
9. Suchorska B, Jansen NL, Linn J, et al. Biological tumor volume in 18FET PET before radiochemotherapy correlates with survival in GBM. Neurology. 2015;84(7):710–9. https://doi.org/10.1212/WNL.0000000000001262.
10. Suchorska B, Jansen NL, Kraus T, Giese A, Bartenstein P, Tonn J. Correlation of dynamic 18FET PET with IDH 1 mutation for prediction of outcome in anaplastic astrocytoma WHO° III independently from tumor vascularisation. Am Soc Clin Oncol. 2015. https://doi.org/10.1200/jco.2015.33.15_suppl.2037.
11. Unterrainer M, Winkelmann I, Suchorska B, et al. Biological tumour volumes of gliomas in early and standard 20–40 min 18F-FET PET images differ according to IDH mutation status. Eur J Nucl Med Mol Imaging. 2018;45:1242–9. https://doi.org/10.1007/s00259-018-3969-4.
12. Kaiser L, Quach S, Zounek A, et al. Enhancing predictability of IDH mutation status in glioma patients at initial diagnosis: a comparative analysis of radiomics from MRI, [18F] FET PET, and TSPO PET. Eur J Nucl Med Mol Imaging. 2024;51:2371–81. https://doi.org/10.1007/s00259-024-06654-5.
13. Lohmann P, Elahmadawy MA, Gutsche R, et al. FET PET radiomics for differentiating pseudoprogression from early tumor progression in glioma patients post-chemoradiation. Cancers. 2020;12(12):3835. https://doi.org/10.3390/cancers12123835.
14. Ceccon G, Lohmann P, Werner JM, et al. Early treatment response assessment using 18F-FET PET compared with contrast-enhanced MRI in glioma patients after adjuvant temozolomide chemotherapy. J Nucl Med. 2021;62(7):918–25. https://doi.org/10.2967/jnumed.120.254243.

15. Wollring MM, Werner JM, Bauer EK, et al. Prediction of response to lomustine-based chemotherapy in glioma patients at recurrence using MRI and FET PET. Neuro Oncol. 2023;25(5):984–94. https://doi.org/10.1093/neuonc/noac229.
16. Law I, Albert NL, Arbizu J, et al. Joint EANM/EANO/RANO practice guidelines/SNMMI procedure standards for imaging of gliomas using PET with radiolabelled amino acids and [18F] FDG: version 10. Eur J Nucl Med Mol Imaging. 2019;46:540–57. https://doi.org/10.1007/s00259-018-4207-9.
17. Unterrainer M, Vettermann F, Brendel M, et al. Towards standardization of 18F-FET PET imaging: do we need a consistent method of background activity assessment? EJNMMI Res. 2017;7:1–8. https://doi.org/10.1186/s13550-017-0295-y.
18. Rahimpour M, Boellaard R, Jentjens S, Deckers W, Goffin K, Koole M. A multi-label CNN model for the automatic detection and segmentation of gliomas using [18F] FET PET imaging. Eur J Nucl Med Mol Imaging. 2023;50(8):2441–52. https://doi.org/10.1007/s00259-023-06193-5.
19. Blanc-Durand P, Van Der Gucht A, Schaefer N, Itti E, Prior JO. Automatic lesion detection and segmentation of 18F-FET PET in gliomas: a full 3D U-net convolutional neural network study. PLoS ONE. 2018;13(4):0195798. https://doi.org/10.1371/journal.pone.0195798.
20. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–11. https://doi.org/10.1038/s41592-020-01008-z.
21. Gutsche R, Lowis C, Ziemons K, et al. Automated brain tumor detection and segmentation for treatment response assessment using amino acid PET. J Nucl Med. 2023;64(10):1594–602. https://doi.org/10.2967/jnumed.123.265725.
22. Dirks I, Keyaerts M, Neyns B, Vandemeulebroucke J. Computer-aided detection and segmentation of malignant melanoma lesions on whole-body 18F-FDG PET/CT using an interpretable deep learning approach. Comput Methods Programs Biomed. 2022;221: 106902. https://doi.org/10.1016/j.cmpb.2022.106902.
23. Albert NL, Galldiks N, Ellingson BM, et al. PET-based response assessment criteria for diffuse gliomas (PET RANO 1.0): a report of the RANO group. Lancet Oncol. 2024;25(1):e29–41.
24. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017;4700–4708. https://doi.org/10.48550/arXiv.1608.06993.
25. Cardoso MJ, Li W, Brown R, et al. MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701. 2022. https://doi.org/10.48550/arXiv.2211.02701.
26. Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024;21(2):195–212. https://doi.org/10.1038/s41592-023-02151-z.
27. Nikolov S, Blackwell S, Zverovitch A, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. J Med Internet Res. 2021;23(7): e26151. https://doi.org/10.2196/26151.

Publisher's Note