European Journal of Nuclear Medicine and Molecular Imaging (2024) 51:1937–1954
https://doi.org/10.1007/s00259-024-06616-x

ORIGINAL ARTICLE

TMTV-Net: fully automated total metabolic tumor volume segmentation in lymphoma PET/CT images — a multi-center generalizability analysis

Fereshteh Yousefirizi[1] · Ivan S. Klyuzhin[1] · Joo Hyun O[2] · Sara Harsini[3] · Xin Tie[4] · Isaac Shiri[5] · Muheon Shin[4] · Changhee Lee[4] · Steve Y. Cho[4] · Tyler J. Bradshaw[4] · Habib Zaidi[5,6,7,8] · François Bénard[3,9] · Laurie H. Sehn[3,10] · Kerry J. Savage[3,10] · Christian Steidl[3,11] · Carlos F. Uribe[1,3,9] · Arman Rahmim[1,9,12,13]

Received: 4 October 2023 / Accepted: 15 January 2024 / Published online: 8 February 2024

Abstract

Purpose Total metabolic tumor volume (TMTV) segmentation has significant value enabling quantitative imaging biomarkers for lymphoma management. In this work, we tackle the challenging task of automated tumor delineation in lymphoma from PET/CT scans using a cascaded approach.

Methods Our study included 1418 2-[^{18}F]FDG PET/CT scans from four different centers. The dataset was divided into 900 scans for development/validation/testing phases and 518 for multi-center external testing. The former consisted of 450 lymphoma, lung cancer, and melanoma scans, along with 450 negative scans, while the latter consisted of lymphoma patients from different centers with diffuse large B cell, primary mediastinal large B cell, and classic Hodgkin lymphoma cases. Our approach involves resampling PET/CT images into different voxel sizes in the first step, followed by training multi-resolution 3D U-Nets on each resampled dataset using a fivefold cross-validation scheme. The models trained on different data splits were ensemble. After applying soft voting to the predicted masks, in the second step, we input the probability-averaged predictions, along with the input imaging data, into another 3D U-Net. Models were trained with semi-supervised loss. We additionally considered the effectiveness of using test time augmentation (TTA) to improve the segmentation performance after training. In addition to quantitative analysis including Dice score (DSC) and TMTV comparisons, the qualitative evaluation was also conducted by nuclear medicine physicians.

Results Our cascaded soft-voting guided approach resulted in performance with an average DSC of 0.68 ± 0.12 for the internal test data from developmental dataset, and an average DSC of 0.66 ± 0.18 on the multi-site external data ($n = 518$), significantly outperforming ($p < 0.001$) state-of-the-art (SOTA) approaches including nnU-Net and SWIN UNETR. While TTA yielded enhanced performance gains for some of the comparator methods, its impact on our cascaded approach was found to be negligible (DSC: 0.66 ± 0.16). Our approach reliably quantified TMTV, with a correlation of 0.89 with the ground truth ($p < 0.001$). Furthermore, in terms of visual assessment, concordance between quantitative evaluations and clinician feedback was observed in the majority of cases. The average relative error (ARE) and the absolute error (AE) in TMTV prediction on external multi-centric dataset were $\mathrm{ARE} = 0.43 \pm 0.54$ and $\mathrm{AE} = 157.32 \pm 378.12$ (mL) for all the external test data ($n = 518$), and $\mathrm{ARE} = 0.30 \pm 0.22$ and $\mathrm{AE} = 82.05 \pm 99.78$ (mL) when the 10% outliers ($n = 53$) were excluded.

Conclusion TMTV-Net demonstrates strong performance and generalizability in TMTV segmentation across multi-site external datasets, encompassing various lymphoma subtypes. A negligible reduction of 2% in overall performance during testing on external data highlights robust model generalizability across different centers and cancer types, likely attributable to its training with resampled inputs. Our model is publicly available, allowing easy multi-site evaluation and generalizability analysis on datasets from different institutions.

Keywords Total metabolic tumor volume, Generalizability, Artificial intelligence · Deep learning · 3D U-Net · PET/CT · Lymphoma

Extended author information available on the last page of the article

Introduction

The predictive potential of total metabolic tumor volume (TMTV), quantified through whole-body 2-[^{18}F]-fluorodeoxyglucose (FDG) positron emission tomography (PET)/computed tomography (CT) scans, has been extensively validated in the context of lymphoma [1–11]. As such, accurate lymphoma segmentation is important for clinical diagnosis and treatment planning in Hodgkin and non-Hodgkin lymphoma. The Lugano system [12] categorizes lymphoma into four stages based on lymph node involvement, but it may not fully represent disease burden. Deauville criteria in 2-[^{18}F]FDG PET/CT is also used for managing lymphoma patients. However, variability in the scores may arise due to differences in quantification reconstruction methods [13, 14]. Moreover, despite some quantitative aspects within these staging systems (Lugano and Deauville), its significant discretization might not precisely mirror the complete disease burden, unlike the continuous nature of TMTV. Although TMTV is a superior proxy for disease burden, providing a more precise measure of disease stage over time [15–20], yet it is commonly not quantified/reported at all [21, 22] since tumor segmentation that is needed for TMTV quantification is time-consuming and labor-intensive [23].

Segmentation of lymphoma lesions in 2-[^{18}F]FDG PET/CT scans presents a significant challenge, owing to the diverse distribution of lesions and the necessity for precise removal of physiological uptake and radiopharmaceutical clearance in organs like the brain, myocardium, liver, brown fat, kidneys, ureters, and bladder [21, 24, 25]. Most existing approaches for tumor segmentation in the clinical workflows are mainly based on maximum standardized uptake value (SUV) thresholding [26]. The tediousness of using currently available semi-automatic software and inherent variability requiring manual input from readers are significant obstacles to the widespread implementation of automated lymphoma segmentation in clinical practice.

While TMTV segmentation has been studied to some extent, there is still a vast potential for adapting artificial intelligence (AI)-based approaches in this area. Some methods have been proposed for lymphoma segmentation, including SUV-threshold-based [27], region-growing-based [28], and Convolutional Neural Network (CNN)-based methods on PET-only and PET/CT images [29, 30]. Although CNN-based segmentation methods have been used extensively, the U-Net and nnU-Net architecture are the most popular models that have been proposed for lymphoma segmentation in recent studies [31–33]. AI-based segmentation methods often exhibit poor precision when it comes to segmenting small lymphoma lesions observed in patients with limited-stage disease and/or small lesions [21, 25]. To improve the segmentation performance, previously some cascaded AI-based approaches such as 2D/3D and 3D/3D models were also applied utilizing a sequence of multiple networks that are interconnected in a sequential manner. Each network in the cascade processes the output of the previous network and refines it further [33, 34].

Despite the acceptable performance of CNN-based approaches, there is still a challenge in quantifying prediction uncertainty [35] including the uncertainty as a the result of the difference between training and testing datasets (domain shift) and/or model uncertainty that emerges from the limited training dataset and model misspecification. We previously suggested a CNN for segmenting whole-body PET/CT images across different cancer types using collective deep learning, presenting the potential to enable rapid assessment of whole-body tumor burden in PET/CT images [33]. We showed that the segmentation model, trained on diverse whole-body PET/CT datasets including primary mediastinal large B cell lymphoma (PMBCL), diffuse large B cell lymphoma (DLBCL), and non-small-cell lung cancer (NSCLC), outperformed models trained solely on DLBCL data that could be due to the varied size and location of DLBCL lesions. Based on this idea, in this work, to address these limitations, a fully automated method for segmenting TMTV was developed, utilizing whole-body PET/CT scans of lymphoma, lung cancer, and melanoma (from autoPET challenge [36]) and were tested extensively on multi-center whole-body PET/CT scans of lymphoma patients with different lymphoma subtypes and stages.

In this study, we propose a two-step cascaded segmentation approach to facilitate TMTV quantification and effectively handle the variability in lesion sizes and locations in lymphoma cases. Our primary objective is to improve the segmentation model generalizability, and our architectural choices are primarily designed to optimize it. The first step facilitates a comprehensive assessment of global connectivity, while the second step refines the process for a more intricate and finely detailed segmentation. More specifically, the multi-resolution approach, three steps, and ensembling have been deliberately made to effectively counter the challenges posed by the frequently encountered issue of dataset shift. These limitations necessitate a multi-site generalizability of deep learning model evaluations. To address these limitations, it becomes crucial to establish multi-site generalizability in deep learning model evaluations. To facilitate multi-site evaluation of our model, we containerized it and deployed it on our in-house developed user-friendly, cloud-based platform, Ascinta, enabling researchers to test our model on their datasets and perform generalizability analysis. We are also publicly sharing the codes and trained model on GitHub.

To address domain shift, test time augmentation (TTA) was proposed wherein the model performance on test examples is enhanced through various data augmentations and by minimizing the average entropy of the model [37]. Recent research has illuminated the potential of the test-time augmentation (TTA) to further enhance prediction robustness in critical areas such as image classification [38] and nodule detection [39]. In this study, we applied TTA as an additional step to our cascaded approach and state-of-the-art approaches to see if TTA is capable of improving the segmentation model performance on the external multi-site test data. For the clinical integration of TMTV-Net, figure of merit should be defined properly to quantify task performance as the components of claim [40]. Consequently, we also considered the bias and noise (i.e., variability) of our suggested technique, TMTV-Net in TMTV measurements along with DSC as recommended in [40, 41].

Material and methods

Patient population for model development and testing

Table 1 provides a comprehensive overview of the data employed in this study including the data we used for model development and testing and the external multi-site testing. We began training by utilizing the autoPET challenge 2022 dataset [36, 43, 44], which comprises patients diagnosed with histologically proven malignant melanoma, lymphoma, or lung cancer who underwent 2-[^{18}F]FDG PET/CT examinations at two major medical centers: University Hospital in Tubingen, Germany. To delineate the dataset, two expert radiologists, with 5 and 10 years of experience, respectively, segmented the lesions manually on axial slices. In total, the dataset included 900 cases, with half of the patients serving as negative controls.

To prevent data leakage, the dataset was divided into two separate sets (including development (training/validation) and test set with data splits of (70/15)/15%) at the patient level. Additionally, the development dataset was divided into five cross-validation sets, stratified by overall lesion volume, i.e., TMTV, to minimize the model variance trained on different splits. Stratification was performed to minimize the model variance trained on different data splits. The same splits were applied to both our proposed ensemble model and state-of-the-art models used for comparison.

Patient population for multi-site testing

For multi-site testing, we tested the model on data from different centers and lymphoma types. The study was conducted

Table 1 Details of multi-center PET dataset information from different lymphoma types

Data split	Center	Cancer type	# of cases	Average voxel spacing (mm^3)	Average injected radioactivity (MBq)	Scanner models
Development and test	autoPET challenge UHTG [36, 42–44]	Lung cancer ($n = 168$) Lymphoma ($n = 145$) (no specified subtype) Melanoma ($n = 188$)	450	2.04×2.04×3	314.7 ± 22.1	Siemens Biograph mCT, mCT Flow
Development and test	autoPET challenge UHTG [36, 42–44]	Negative cases	450	2.04×2.04×3	314.7 ± 22.1	Siemens Biograph mCT, mCT Flow
External testing	BCC	PMBCL	103	4.06×4.06×2.635	347.5 ± 52.6	GE (Discovery D600 and D690)
External testing	BCC	DLBCL (stage I to II)	86	3.65×3.65×3.27	335.9 ± 50.8	GE (Discovery D600 and D690)
External testing	BCC	Hodgkin lymphoma	30	3.65×3.65×3.27	363.91 ± 58.2	GE (Discovery D600 and D690)
External testing	SMSK	DLBCL (stages I to IV)	218	3.79×3.79×4.42	246 ± 47.5	GE (Discovery 710) ($n = 42$) Siemens (Biograph40 TruePoint) ($n = 176$)
External testing	UW	Classic Hodgkin ($n = 39$) Nodular lymphocyte-predominant ($n = 1$) DLBCL ($n = 41$)	81	3.63×3.63×3.09	472.2 ± 140.3	GE Discovery 710, GE Discovery MI, GE Discovery IQ

PMBCL primary mediastinal large B cell lymphoma, *DLBCL* diffuse large B cell lymphoma, *BCC* BC Cancer Canada, *SMSK* Saint Mary Hospital in South Korea, *UW* University of Wisconsin USA, *UHTG* University Hospital in Tubingen Germany in accordance with the Declaration of Helsinki (as revised in 2013). PET/CT images included patients with DLBCL from three different centers, BC Cancer Canada (BCC) (ethics number: H19-01866), Saint Mary Hospital in South Korea (SMSK) (ethics number: KC11EISI0293), and University of Wisconsin (UW) (ethics number: UW2016-0418). Patients with primary mediastinal large B cell lymphoma (PMBCL) were from BCC (ethics number: H19-01611). Patients with classic Hodgkin lymphoma were from BCC (ethic number: H19-001611) and UW (ethics number: UW2016-0418).

Manual segmentation was performed semi-automatically by the nuclear medicine physicians using MIM software. Datasets include the PET/CT DICOM series and their corresponding radiotherapy structure (RT-STRUCT) files, which should be parsed into arrays of voxel intensities and a binary mask that corresponds to a volume of interest (VOI) using a previously developed in-house tool [45]. Notably, the ground truth labels in UW dataset have 12 categories comprising non-equivocal (bone marrow lesion, osseous lesion, liver lesion, extra-nodal lesion, splenic lesion, and lymph-nodal lesion) and equivocal lesions (bone marrow lesion, osseous lesion, liver lesion, extra-nodal lesion, splenic lesion, and lymph-nodal lesion). Labels were obtained from a 3-reader adjudication process, in which a 2nd reader reviewed and edited the primary reader's annotations, and a 3rd reader adjudicated disagreements. We considered the segmentation capability of our model on both non-equivocal and equivocal lesions. In the dataset from centers BCC and SMSK, all extra-nodal lesions (such as bones and lung lesions) were included in the manual segmentation. Although, in PMBCL cases, the extra-nodal disease is rare but still the nuclear medicine expert considered them. We have no specific information about the type and extra-nodal disease of the auto-PET lymphoma data that was used for training in this study.

The 2-[^{18}F] FDG PET scans exhibited a diverse spectrum of normal uptake patterns (as illustrated in Fig. S1). Furthermore, among patients with lesions, a significant variability was observed, with some presenting bulky, disseminated patterns, while others displayed low uptake patterns. These findings underscore the complexity and diversity of 2-[^{18}F] FDG uptake patterns both in normal tissues and in lesion presentations.

Image preprocessing

First, we resampled the CT images to match the size of PET scans. Subsequently, we created five distinct series of inputs by combining the PET and CT images. PET SUV range was transformed from [0, 30] SUV to [0, 1] to capture a broader range of intensities. CT range was converted from [− 150, 300] Hounsfield units (HU) to [0, 1] to capture important patterns. CT Soft1 used [− 100, 100] HU, focusing on soft-tissue intensities. CT Soft2 used [− 1000, − 200] HU, focusing on lung tissue intensities. SUV hot used [2, 10] SUV range, aiding in mid-range intensity focus for lesions with low uptake. We resampled the PET/CT images into the voxel sizes of [2 mm]^3, [4 mm]^3, [6 mm]^3, and [8 mm]^3 and a random resampling (range = [2, 10] using linear interpolation for images and the nearest neighbor for ground truth images).

Data augmentation

We utilized elastic deformations for data augmentation to help the model to learn the varied size and shapes of the lesions. Also, the following augmentation techniques were applied to increase the complexity of the training data: random affine transformation includes random rotation (< 25°), random axis flip for all three dimensions, elastic deformations, and contrast transform; PET-only augmentation includes the Gamma transforms with γ sampled from the uniform distribution (0.8 and 1.2) and random Gaussian blur and brightness transform. We also used MixUp [46], a powerful and versatile data augmentation strategy that involves creating augmented samples by linearly interpolating between pairs of inputs and their corresponding labels and it is mainly used in semi-supervised learning.

Two-stage segmentation approach

Our approach includes two steps and a soft voting intermediate step (Fig. 1). To tackle the variation in size, location, and appearance of lymphoma, our biphasic methodology leverages a spacious receptive field in Step I by applying a series of 3D U-Nets to resampled PET/CT scans with different resolutions, facilitating coarse-grained analysis of global patterns and extensive dependencies. In Step II, the cascaded 3D U-Net takes as input the predictions from the first step and combines them with randomly resampled PET/CT scans, resulting in a more intricate and finely detailed segmentation. We incorporate both PET and CT images in our segmentation approach, as PET images are prone to blurring the contours of objects due to their low resolution and partial volume effect. We used deep supervision model to learn features at multiple levels of details [47].

Step I: multi-resolution 3D U-Nets. Step I enables a broader analysis of global patterns and extensive dependencies, ensuring a coarse-grained analysis achieved through a series of 3D models. Random resampling during training also enhances the model ability to learn from inputs with various resolutions. We assumed that large connectivity could be captured while small image details lost in the case of resampling at a larger voxel size (e.g., 6 and 8) [48]. In

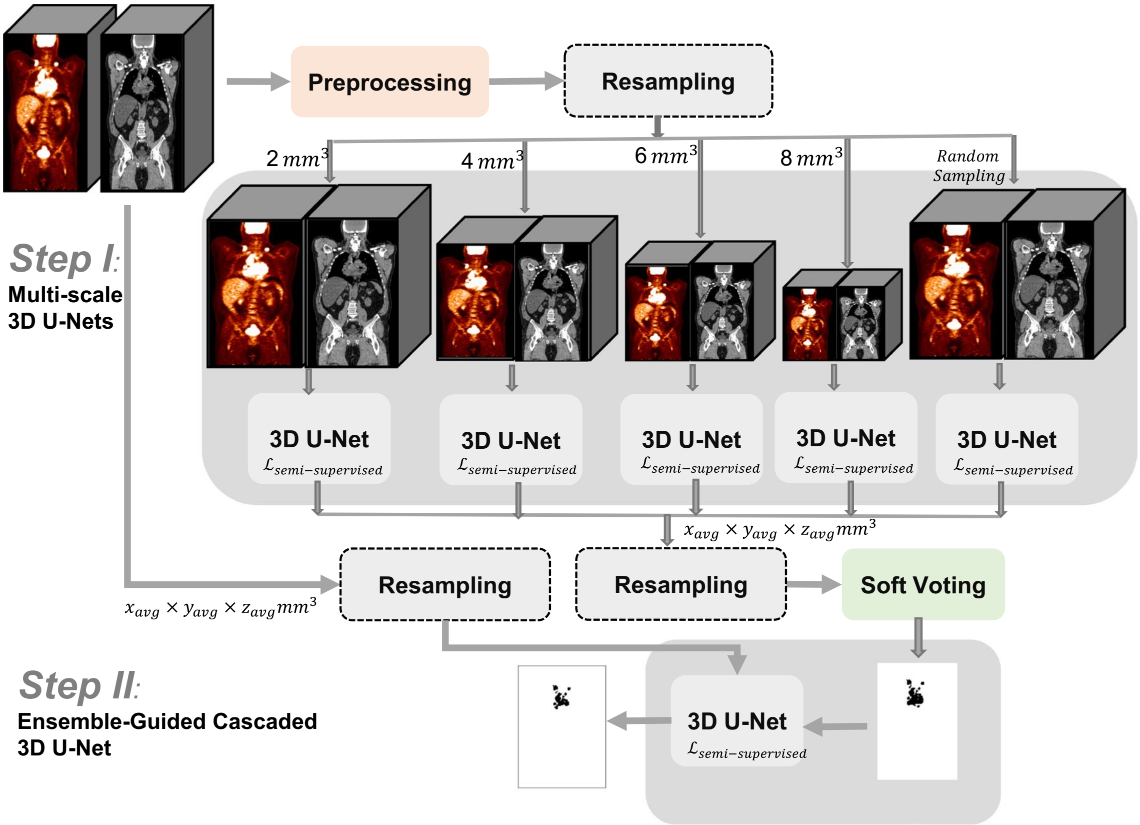
Preprocessing
Resampling
$2\,mm^3$
$4\,mm^3$
$6\,mm^3$
$8\,mm^3$
Random Sampling
Step I:
Multi-scale 3D U-Nets
3D U-Net $\mathcal{L}_{semi-supervised}$
3D U-Net $\mathcal{L}_{semi-supervised}$
3D U-Net $\mathcal{L}_{semi-supervised}$
3D U-Net $\mathcal{L}_{semi-supervised}$
3D U-Net $\mathcal{L}_{semi-supervised}$
$x_{avg} \times y_{avg} \times z_{avg} mm^3$
$x_{avg} \times y_{avg} \times z_{avg} mm^3$
Resampling
Resampling
Soft Voting
Step II:
Ensemble-Guided Cascaded 3D U-Net
3D U-Net $\mathcal{L}_{semi-supervised}$

Fig. 1 Overview of our two-stage cascaded approach for TMTV segmentation, TMTV-Net. In Step I, five resampled datasets were processed using the multi-scale 3D U-Net with a fivefold cross-validation (CV) strategy. The multi-resolution predicted masks of the five U-Net models were ensemble to create the segmentation mask first for CV and then averaged on predictions from different resolution using soft-voting to generate the intermediate mask. In Step II, a cascaded soft-voting-guided network approach was employed to further enhance segmentation performance

the first step, each of the five resampled datasets was processed with a 3D U-Net using a fivefold cross-validation (CV). Ensembles of five CV models trained on different data splits were generated on each resolution. Consistency in data splits was maintained across all resolution models.

Soft voting of multi-resolution predictions. Predicted mask of each 3D U-Nets was first resampled to the weighted average voxel size $[2.65 \times 2.65 \times 3.23\text{mm}]^3$ based on the training dataset. By soft voting (probability averaging) of the multi-resolution predictions in this step, we aimed to ensure that the majority of outputs have a greater influence on the final prediction of Step I. In fact, this averaged prediction is used to generate an initial segmentation, which is then used to guide the training of a subsequent network in Step II for further improvement.

Step II: soft-voting-guided cascaded segmentation. Resampling to the weighted average voxel size was first applied to the input PET/CT images. The soft-voted combination of the multi-resolution models was concatenated with the resampled PET/CT input images and fed into another 3D U-Net. This approach enabled the network to learn robust features and reduce false positives in the segmentation results. In this step, the cascaded approach comes into play, utilizing the predictions of the first step as well as incorporating information from randomly resampled PET/CT scans that lead to a finely detailed segmentation.

Segmentation model

We used a modified version of 3D U-Nets that incorporates a deep supervision architecture [47]. Deeply supervised 3D U-Net uses additional supervision by making predictions at intermediate decoder layers as well. These intermediate predictions are then combined during training to compute the final loss, which aids in alleviating the vanishing gradient problem and facilitates faster convergence during training. We used deep supervision module that applies deep supervision to intermediate layers and combines their outputs to compute the final loss. We used a semi-supervised loss function (Eq. (1)), composed of cross-entropy (CE) and Dice loss as supervised losses, and an unsupervised boundary-based loss term, namely, Mumford-Shah (MS) [49, 50]:

$$\mathcal{L}_{semi-supervised}(y,g;\theta) = \alpha\mathcal{L}_{MS}(y;\theta) + \beta\mathcal{L}_{Dice}(y,g;\theta) + \lambda\mathcal{L}_{CE}(y,g;\theta) \quad (1)$$

wherein y is the output of the network, g is the ground truth, and θ is the network parameter [51]. In our previous studies [50, 52], we conducted extensive semi-supervised training by systematically adjusting α, the weight of the Mumford-Shah term, and assessing its impact on the model performance. While we observed a significant role for the Mumford-Shah term in enhancing overall robustness particularly with limited datasets, in the current study, our model development does not involve scenarios with scarce labeled data, and consequently, the weight assigned to this term is relatively low. The tuned parameters were $\alpha = 10^{-5}$,$\beta = 1$, and $\lambda = 2$. The MS loss function helps the network utilize unlabeled images (Eq. (2)) and in the case of labeled images, it only needs the input image. In this study we used labeled images and in the case of unlabeled images, β and λ become 0:

$$\mathcal{L}_{MS} = \sum_{k=1}^{C} \sum_{j\epsilon\Omega} f_{jk}\|y_j - v_k\|^2 + \eta \sum_{k=1}^{C} \sum_{j\epsilon\Omega} \left|\nabla f_{jk}\right| \quad (2)$$

where f_{jk} is the softmax output of CNN, while $\sum_{k=1}^{c} f_{jk} = 1$ and $\left|\nabla f_{jk}\right|$ is the total variant of f_{jk} using the approximation $\nabla f_{jk} = f_{(j+1)k} - f_{jk}$. The average voxel intensity is shown by v_k here as well. The average voxel intensity is shown by v_k and $C = 2$ indicates the number of classes. All models were individually trained using the AdamW optimizer, with a learning rate of 10^{-3}, weight decay of 10^{-6}, and a decayed cosine warm restart scheduler with $T = 400$ epochs and a decay rate of 0.9 for each period. To ensure stable training, gradient clipping was also employed. Our proposed models were trained and validated within the computational environment of a Microsoft Azure virtual machine with Ubuntu 20.04.6. This machine consisted of 6 CPU cores (112 GiB RAM) and a single NVIDIA Tesla V100 GPUs (16 GiB RAM); we used Python 3.9 and Pytorch 1.11.

Test time augmentation

TTA involves generating multiple transformed copies of a test input and integrating the predictions obtained from these augmented images. By adopting TTA and considering the predictions from augmented images in addition to the "clean" images from the testing dataset, a more accurate final prediction can be achieved. This process usually involves averaging the predictions of each image, and it may also incorporate learnable weights to form a weighted average, ultimately contributing to superior segmentation outcomes and performance evaluation for various applications. The schematic of our TTA approach is shown in Fig. 2. TTA encompasses four steps: augmentation, prediction, inverse transform, and aggregation. During augmentation, the test image undergoes various transformations, including random rotation, elastic transform, random vertical flip, and Gaussian blur. Subsequently, predictions are generated for the original and augmented images, followed by an inverse transform to revert the transformations on the resulting predictions. Finally, these predictions are merged together to produce the final result. To optimize effectiveness, we fine-tuned the weights of each augmentation transform, seeking the combination that yielded the highest-performance enhancement. If Aug is the set of transformations, let x be a given input PET/CT dataset and Aug is a candidate subset of transformation (Tr_i) that are usually used for augmentation ($Aug = \left\{Tr_1, Tr_2, \ldots, Tr_n\right\}$):

$$y_{TTA} = \frac{1}{n} \sum_{i=1}^{n} \left(w_i \times Tr_i^{-1} \times \Theta(Tr_i(x))\right) \quad (3)$$

where Θ is the trained model, y_{tta} is the ensemble of the outputs after applying the inverse transforms (Tr_i^{-1}), and w_i are the weight for the inverse transforms to be applied to the model prediction.

We devised a strategy for determining the optimal weight set of the inverse transform in Eq. (3) through a random search. To begin, 1000 weight sets were randomly selected and subjected to thorough evaluation using CV. The resulting CV scores were then used to rank the weight sets, W,

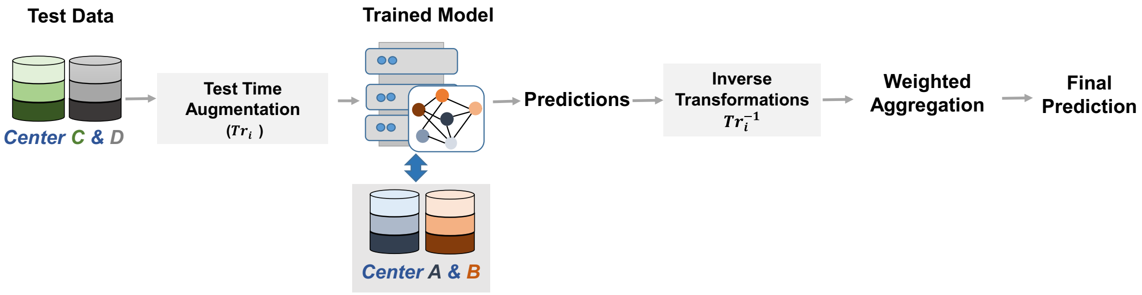

Fig. 2 Our test time augmentation approach, the model trained on data from center A and B, test data includes data from centers C and D

and to prevent any potential overfitting to the TTA training set, we carefully selected the top W hyperparameter sets for further consideration. For predicting on the TTA test set, we constructed an ensemble model that incorporates each y_{TTA}. This model randomly selects a weight set from the previously identified top W sets, thereby ensuring robustness and generalizability in our weight predictions. To fine-tune the selection process and identify the most suitable value for weights, we conducted an additional round of cross-validation, allowing us to optimize the performance of finding the optimal weights for TTA.

Multi-center external testing

We aimed to ensure that our models can be easily and reproducibly tested at different sites and evaluated clinically. To ensure that the cascaded model can be run as expected on different machines and/or sites without loss of performance, the model (including the component models) was containerized using Docker. All testing was performed with the containerized version of the model, with all dependencies and their versions specified. To ensure that the data is always supplied to the model in the correct format, a specialized secure DICOM-based API was designed to pass the data in and out of the container. The use of the API also allows running the containerized model on our in-house developed cloud-platform, Ascinta, which can be used for multi-site code-free model testing and clinical evaluation by radiologists. The platform generates lesion segmentation contour in the RTSTRUCT format and graphical PDF report with maximum intensity projection views along with the visualization of the segmented lesions, TMTV and lesion dissemination metric, D_{max}, and other metrics. The codes and trained model are also available here: https://github.com/qurit-frizi/TMTV-Net.

Evaluation analyses

To evaluate TMTV-Net on classical Hodgkin lymphoma cases ($n = 30$), in addition to DSC, we conducted a qualitative analysis. For comparison to SOTA approaches, we considered frameworks based on nnU-Net as deployed by Blanc-Durand et al. [31], deep evidential network by Huang et al. [53], and Swin UNETR by Hatamizadeh et al. [54] trained and tested on same data. We performed an ablation analysis on the cascaded segmentation approach to examine the key components of our proposed method. Specifically, we applied (i) a baseline approach using single-resolution 3D U-Net and (ii) only the first step (without cascaded refinement).

Results

The results of different experiments conducted in this work to evaluate TMTV-Net are presented as follows. Firstly, we present the findings from the ablation study, demonstrating the impact of integrating Step I, employing soft voting, and incorporating Step II for TMTV segmentation and quantification. This is followed by an exploration of external evaluation on multi-centric datasets and a meticulous assessment of segmentation performance in clinical contexts. Furthermore, we offer a comprehensive comparative analysis against state-of-the-art (SOTA) techniques, accompanied by an investigation into the influence exerted by the utilization of TTA.

Table 2 shows the performance of the models when evaluated on the held-out test split on autoPET dataset through segmentation analysis, along with the results of the ablation study. It should be mentioned that the negative cases with no segmentation were discarded from the average Dice score (DSC) evaluation. As the results shown in Table 2, the soft-voted multi-scale and cascaded refinement improved the segmentation performance in terms of DSC compared to single-scale 3D U-Net.

External validation and clinical evaluation of segmentation quality

We evaluated our segmentation model on the external datasets from multiple sites that are presented in Table 1 including classic Hodgkin, DLBCL (limited stage), and PMBCL cases from BCC and DLBCL cases from SMSK; the results are presented in Table 3. Some of the segmentation results on multi-site external testing datasets are shown in Fig. 3.

Qualitative analysis

Three physicians provided qualitative ratings on 10 cases. Physicians 1 and 3 suggested a rating scale of “bad/poor,” “average,” and “good,” while physician 2 suggested a ranking based on “incorrect and incomplete” (segmentation of

Table 2 Segmentation performance of the single-scale, soft-voted multi-scale and cascaded approaches on the test set

Approach	DSC
Single-scale 3D U-Net	0.59 ± 0.14
Soft-voted multi-scale	0.63 ± 0.18
Cascaded refinement	$\mathbf{0.68 \pm 0.12}$

The negative cases with no segmentation were discarded from the average Dice score (DSC) evaluation

The use of “bold” emphasis indicates statistical significance, denoted by a p-value<0.001

Table 3 Comparison to state-of-the-art (SOTA) approaches with our model on multi-site testing data (overall performance of our cascaded approach is Dice score (DSC) = 0.66 ± 0.18)

Study	Model	DSC on DLBCL cases from center SMSK	DSC on PMBCL cases from BCC	DSC on DLBCL cases from BCC	DSC on Hodgkin from BCC
Blanc-Durand et al. [31]	nnU-Net	0.61 ± 0.2	0.47 ± 0.25	0.58 ± 0.32	0.59 ± 0.18
Huang et al. [53]	Deep evidential network	0.57 ± 0.23	0.53 ± 0.33	0.48 ± 0.37	0.56 ± 0.17
Hatamizadeh et al. [54]	Swin UNETR	0.53 ± 0.25	0.51 ± 0.25	0.59 ± 0.30	0.45 ± 0.18
Our study	TMTV-Net	**0.70 ± 0.13**	**0.62 ± 0.15**	**0.67 ± 0.27**	**0.63 ± 0.17**

The use of "bold" emphasis indicates statistical significance, denoted by a p-value<0.001

false positives), "incomplete" (missing lesions that affect staging), "almost complete" (missing minor lesions that do not affect staging), and "complete." In most cases, quantitative evaluations and clinician feedback were consistent. However, there are some deviations between the qualitative considerations of physicians. Some of the selected results of the visual inspection along with their corresponding Dice scores are shown in Fig. 4.

Comparison to state-of-the-art approaches

The results before and after applying TTA are shown, respectively, in Tables 3 and 4. We performed a Wilcoxon signed-rank test to assess the significance of observed differences between our proposed model and state-of-the-art approaches. The test yielded p-value < 0.001, indicating robust statistical difference.

In Fig. 5(a) and (b), we present a comparison between our proposed cascaded 3D U-Net segmentation approach and SOTA methods using a DLBCL case from the BCC center. We provide maximum intensity projection (MIP) coronal and sagittal views (Fig. 5(a) and (b)) along with their respective ground truth (GT) and segmentation results for improved visual representation. Figure 5(d) and (e) show the effect of using TTA on the segmentation approaches to the same sample PET/CT scan of the multi-site external testing dataset. Additionally, the corresponding DSCs are reported in Fig. 5(c) and (f) to quantify the performance of the segmentation results.

External testing in UW center

We shared our model to be tested on the DLBCL and Hodgkin cases collected from UW hospitals and annotated by UW nuclear medicine physicians. The quantitative results of our model on this dataset are presented in Table 5. In addition, Fig. 6 shows four representative cases, highlighting the model capability to segment distributed (Fig. 6(a) and (c)) and small (Fig. 6(b) and (d)) lesions. The results on data from UW center (Table 5) demonstrated that TMTV-Net is capable of segmenting both non-equivocal and equivocal lesions with almost the same performance.

TMTV prediction evaluation

In the following series of evaluations, we considered the errors in TMTV quantification regardless of the accuracy of localization (detection).

Ablation study

The scatter plot of predicted and the ground truth (GT) TMTVs (left) and the Bland–Altman plot (right) are shown in Fig. 7(a) and (b). To assess the correlation between TMTV values, Pearson was used, which showed $R^2 = 0.89$ ($p < 0.0001$). Bland–Altman plot compared the predicted and GT TMTV using our suggested approach that shows the agreement between two measurements. Correlation and Bland–Altman analysis of using only the Step I are shown in Fig. 7(c) and (d). Correlation is reduced from 0.89 to 0.83 ($p < 0.001$), and the agreement based on Bland–Altman analysis occurred in a wider range between the measurements compared to the cascaded two-step approach.

Quantitative analysis of TMTV on multi-site external testing

The relative volume error of TMTV calculation based on different centers and different lymphoma types are shown in Fig. 8(a) and (b). The absolute error distribution and the relative error distribution as a function of total tumor volume ranges are shown in Fig. 8(c) and (d).

The predicted versus ground truth TMTV and the distribution of the estimated and real volumes are shown in Fig. S2. Performances for different centers are also shown in Fig. S3. Breakdowns of values are also shown in Table S1. The mean absolute error increased from 8.9 to 504.8 mL in the 5 volume bins, while mean relative errors decreased from 59 to 33%, with an overall uncertainly of 42%. Overally our external testing results showed that TMTV-Net worked well on a diverse dataset which is not expected from the existing segmentation networks for tumoral leasion of PET/CT scans in lymphoma cases.

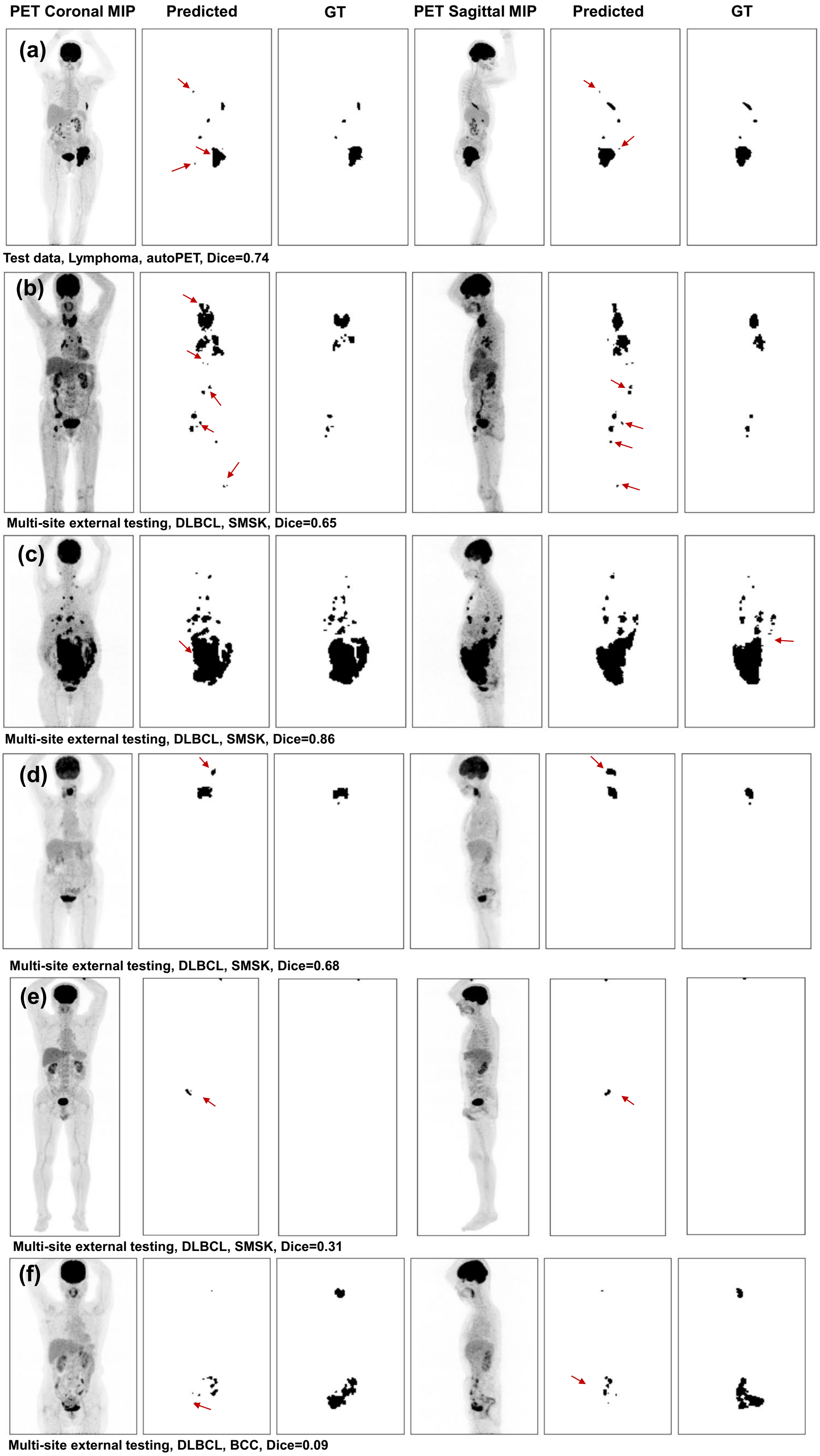
PET Coronal MIP
Predicted
GT
PET Sagittal MIP
Predicted
GT
(a)
Test data, Lymphoma, autoPET, Dice=0.74
(b)
Multi-site external testing, DLBCL, SMSK, Dice=0.65
(c)
Multi-site external testing, DLBCL, SMSK, Dice=0.86
(d)
Multi-site external testing, DLBCL, SMSK, Dice=0.68
(e)
Multi-site external testing, DLBCL, SMSK, Dice=0.31
(f)
Multi-site external testing, DLBCL, BCC, Dice=0.09

Fig. 3 Sample results on the unseen test set (**a**) and on the multi-site external testing (**b–d**). We also included two of our poor results (**e** and **f**). The red arrows show the different segmented regions with respect to the ground truths provided by nuclear medicine physicians

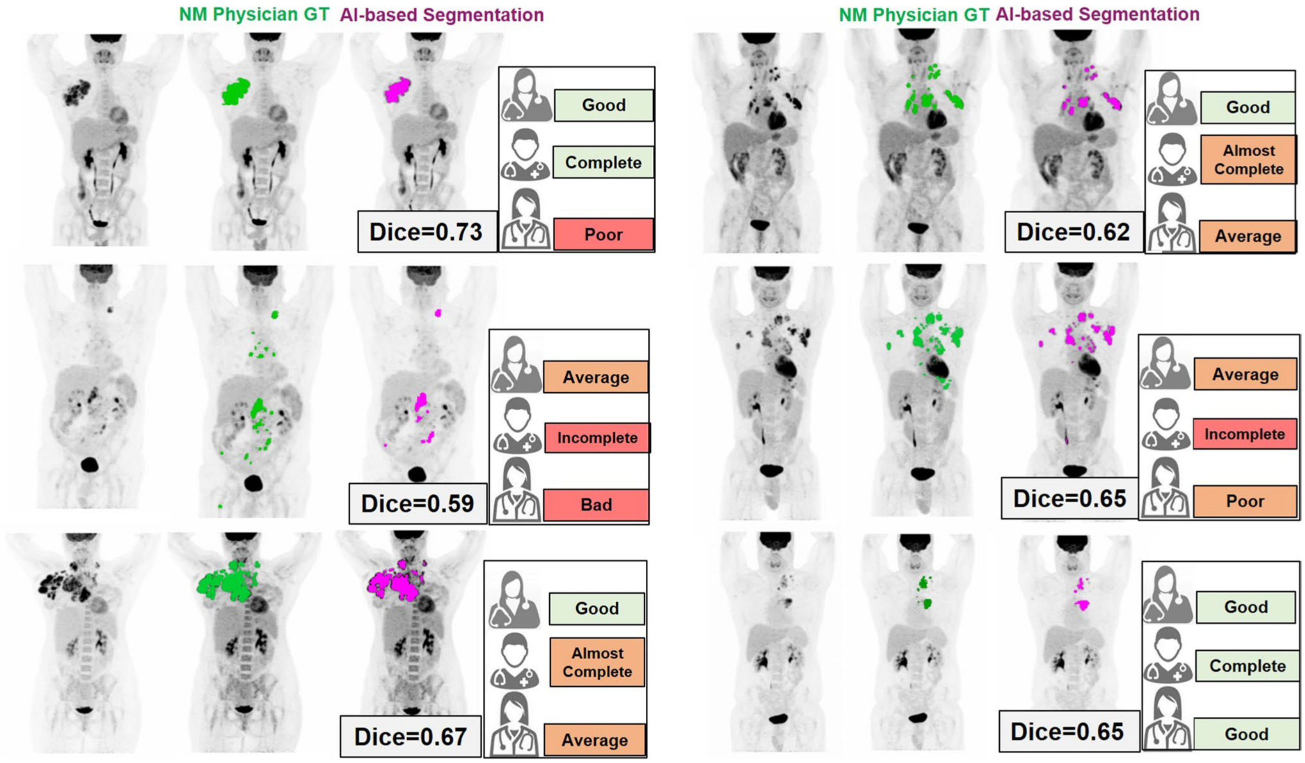

Fig. 4 Qualitative analysis (visual inspection) by three nuclear medicine physicians on the external Hodgkin cases from BCC. Maximum intensity projections are shown. The corresponding ground truth delineated by nuclear medicine physician (NM physician G. T.) and the TMTV-Net results are overlaid

Table 4 Comparison of applying test time augmentation (TTA) to state-of-the-art (SOTA) approaches with our model on multi-site testing data (overall performance of our cascaded approach on external multi-site testing Dice score (DSC = 0.66 ± 0.16))

Study	Model	DSC on DLBCL cases from center SMSK	DSC on PMBCL cases from BCC	DSC on DLBCL cases from BCC	DSC on Hodgkin from BCC
Blanc-Durand et al. [31]	nnU-Net	0.61 ± 0.2	0.49 ± 0.23	0.62 ± 0.30	0.60 ± 0.22
Huang et al. [53]	Deep evidential network	0.57 ± 0.22	0.54 ± 0.28	0.49 ± 0.37	0.56 ± 0.17
Hatamizadeh et al. [54]	Swin UNETR	0.53 ± 0.23	0.54 ± 0.23	0.62 ± 0.30	0.47 ± 0.16
Our study	TMTV-Net	**0.71 ± 0.13**	**0.62 ± 0.15**	**0.67 ± 0.27**	**0.63 ± 0.09**

The use of "bold" emphasis indicates statistical significance, denoted by a *p*-value<0.001

Discussion

Segmentation of lymphoma lesions poses a challenge for AI-based methods due to the wide range of lesion sizes and sites and the need to accurately exclude physiological uptake and radiopharmaceutical clearance in various organs, resulting in lower performance compared to segmentation of primary tumors in 2-[^{18}F]FDG PET/CT scans of various cancers [21, 25]. This is primarily due to the high variability in the distribution of lesions, as well as the need to account for physiological uptake (such as in the brain, myocardium, liver, and brown fat) and radiopharmaceutical clearance (such as in the kidneys, ureters, and bladder), which must be accurately trimmed to ensure precise segmentation (Fig. S1). To overcome this limitation, we suggested TMTV-Net, a comprehensive and fully automated approach for accurately segmenting tumoral lesions in lung cancer and melanoma, as well as lymphoma lesions. As PET/CT data from diverse cancers gain prominence, a promising opportunity emerges to train a single neural network capable of accurately quantifying tumor burden from various malignancies. By encompassing diverse types of cancer, including lung cancer, and melanoma, we aimed to improve the versatility and generalizability of our approach.

For model training, harmonization techniques were unnecessary for the mono-centric training data. However, for evaluating our segmentation model on an unseen external

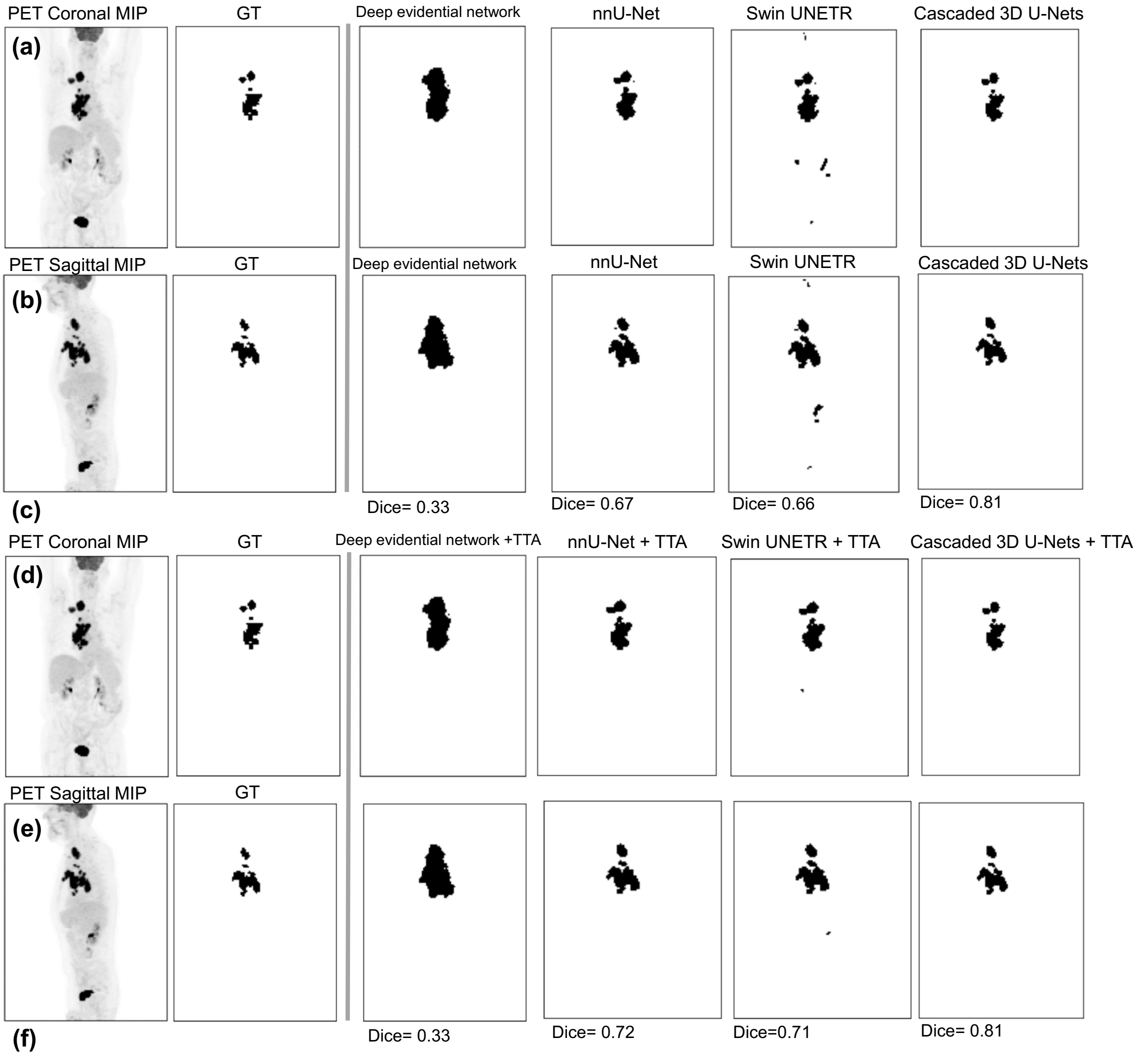

Fig. 5 Comparison of our suggested 3D segmentation approach to SOTA approaches on a DLBCL case from the BCC center. MIP views (**a** and **b**) and the corresponding GTs and the segmentations results are shown for better visualization. Comparison the effect of TTA on the segmentation performance of our suggested 3D segmentation approach and the SOTA approaches on a DLBCL case from the BCC center (**d** and **e**). The corresponding Dice scores are shown (**c** and **f**). *GT*: ground truth, *MIP*: maximum intensity projections

testing set, we used data from three distinct centers and acquired through six different scanners. This external testing dataset comprises instances of three lymphoma types at various stages as imaged using different scanners. Our study specifically aimed to assess how this diversity affects the model performance. In essence, one of the primary goals of this study is to assess the generalizability of our model under these diverse conditions.

Table 5 Performance evaluation of TMTV-Net in UW center on lymphoma data (Hodgkin and DLBCL)

Ground truth category	DSC	Precision	Recall
Non-equivocal lesions	0.694 ± 0.18	0.76 ± 0.15	0.71 ± 0.25
Equivocal and non-equivocal lesions	0.687 ± 0.19	0.76 ± 0.14	0.69 ± 0.25

To validate the effectiveness of our method, rigorous testing has been conducted on multi-center whole-body PET/CT scans of lymphoma patients. By encompassing data from diverse centers and lymphoma subtypes in our evaluation, we have ensured the robustness and generalizability of our approach (dataset description in Table 1 and sample results in Fig. 3). The overall performance of TMTV-Net on the multi-site external datasets is DSC: 0.66 ± 0.16 with TTA (0.66 ± 0.18 without TTA) which has a performance drop of 2% compared to the test set on autoPET dataset (0.68 ± 0.12). The resulted generalizability of our approach compared to the usual amount of expected performance drop for external testing in other

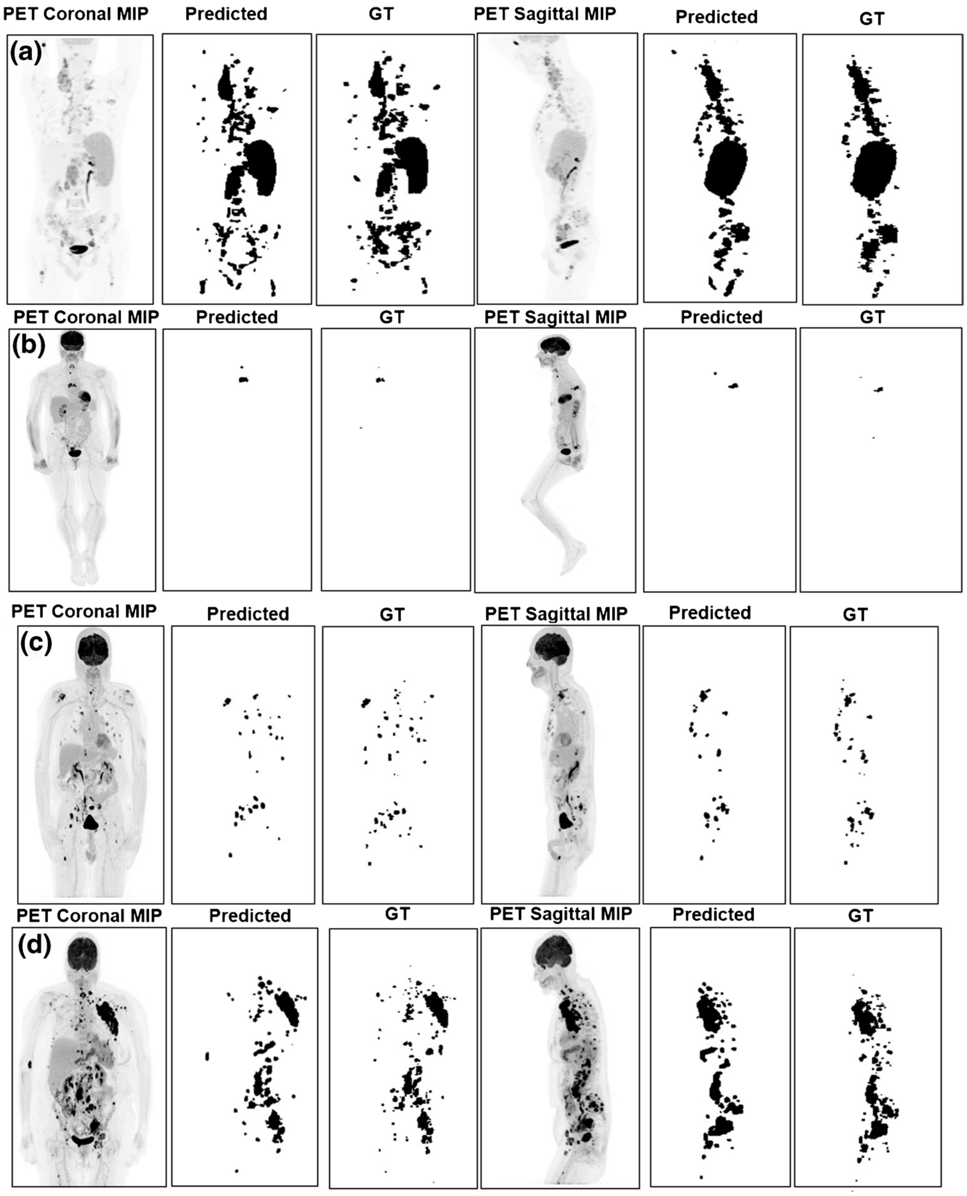

Fig. 6 The example results of the segmented TMTV on data from UW center. **a** Hodgkin case, DSC = 0.83, TMTV relative error = 0.18. **b** DLBCL case, DSC = 0.66, TMTV relative error = 0.10. **c** Hodgkins case, DSC = 0.76. **d** DLBCL, DSC = 0.67

studies which is around 10 percent underscores the robustness and reliability of our model to generalize effectively. The higher performance of the model on data from SMSK data is also notable. The results obtained from the UW center data (Table 5 and Fig. 6) underscore the capability of TMTV-Net in segmenting both non-equivocal and equivocal lesions, demonstrating nearly equivalent performance for both categories. Besides, in the external testing dataset, experts from different centers and cohorts additionally segmented extra-nodal disease. Our results on the external dataset revealed the ability of TMTV-Net to perform accurate segmentations.

Deep learning-based methods can handle variations in image appearance if the training dataset covers these variations adequately, but the resource-intensive demands of including such cases might be prohibitive [55]. Training the model with resampled inputs makes it more robust to resolution variations and reduces overfitting to a particular resolution, thereby improving its generalizability to new data. Our cascaded approach can be easily incorporated into various models and extended to address different segmentation tasks that may be affected by domain shifts or limited labeling resources in a plug-and-play fashion.

To validate the necessity of using both steps in our cascaded segmentation approach, we conducted an ablation analysis to investigate its key components. We compared two setups: a baseline approach using a single-scale 3D U-Net and the first step alone (without cascaded refinement). Our findings, as illustrated in Fig. 7(c) and (d), demonstrate that employing only the first step resulted in reduced correlation

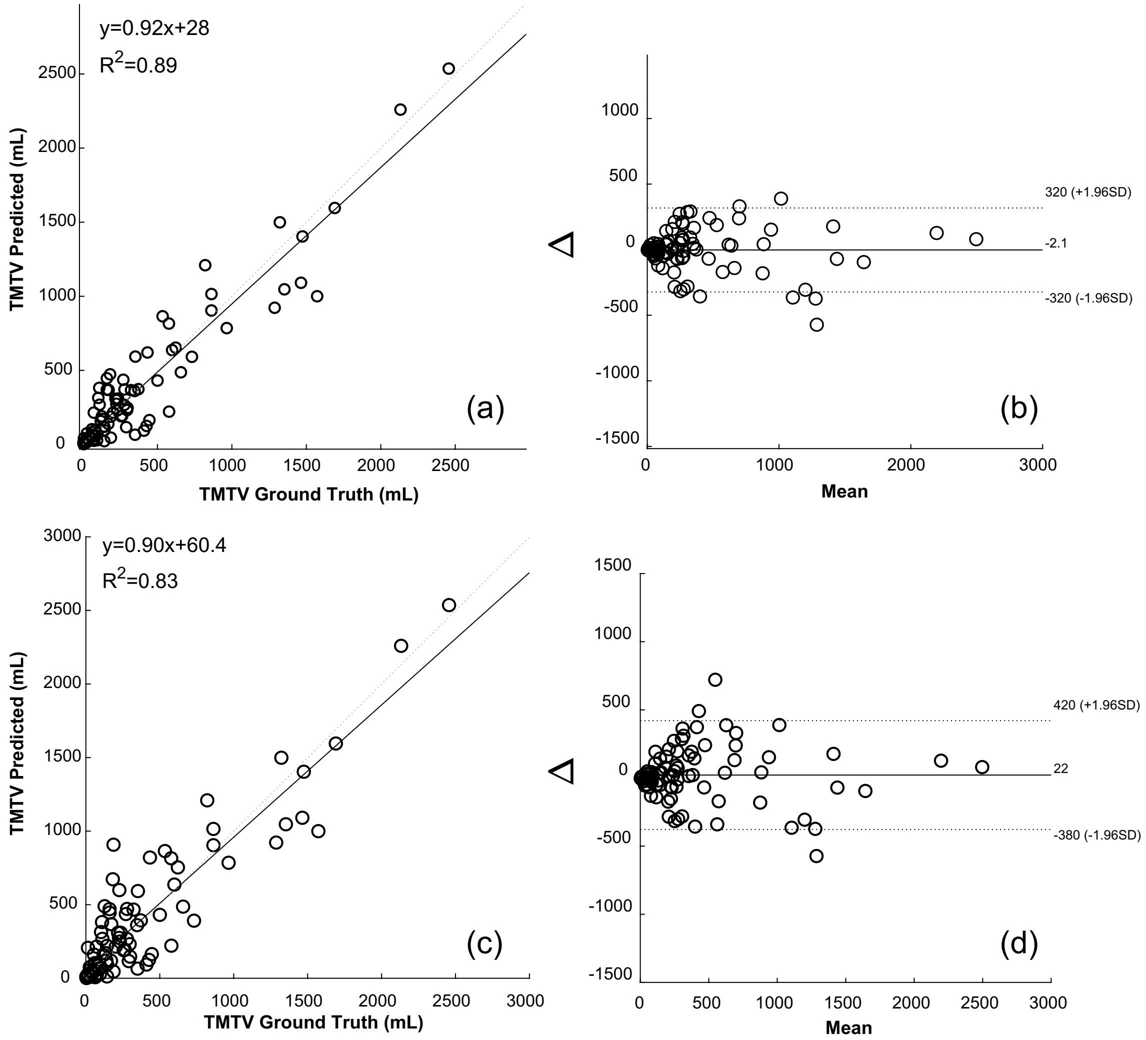

Fig. 7 Regression analysis (correlation) ($p < 0.0001$) (**a**) and Bland–Altman (**b**) plots of the correlations between ground truth and predicted TMTVs in the "unseen" test data, including negative cases with TMTV = 0 mL. Results from Step I, excluding the soft voting-guided Step II, revealed a lower correlation compared to the cascade approach in predicting TMTVs in the "unseen" test data with negative cases as TMTV = 0 mL. Specifically, a lower correlation of 0.83 ($p < 0.0001$) was obtained (**c**) and the Bland–Altman plot (**d**) showed agreement across only a broader range of measurements compared to using cascaded two-step approach

(from 0.89 to 0.83) and a wider range of agreement based on Bland–Altman analysis compared to the cascaded approach. It should be noted that the correlation values in the data from all the external test dataset are higher than 0.89 (Fig. S2). Additionally, the results presented in Table 2 indicate that the soft-voted multi-scale and cascaded refinement led to improved segmentation performance in terms of DSC compared to the single-scale 3D U-Net on test data.

Comparisons to SOTA methods trained, validated, and tested on the same datasets showed that our suggested cascaded 3D U-Nets had better performance compared to deep evidential network, nn-U-Net and SWIN UNETR. We also considered the effect of TTA on the segmentation performance of tour approach and the SOTA method. Basically our primary assumption was to expect the same effect of TTA of the performance of our model and SOTA techniques but our results showed that TTA worked better for nn-U-Net and SWIN UNETR (Table 4); for example, in the sample PET/CT scan in Fig. 5(d) and (e), TTA helped to decrease false positive rate by removing the regions that were erroneously included in the segmented volume but it could not improve our model performance significantly.

External testing was performed on a real-world dataset with ground truth manually segmented through a decentralized process, reflecting the challenges of intra- and inter-observer variabilities in ground truth generation. Our evaluation on this decentralized dataset demonstrated the effective performance of TMTV-Net. We evaluated our segmentation approach on 518 scans from different centers and lymphoma subtypes; for further generalizability evaluation we need more dataset from different

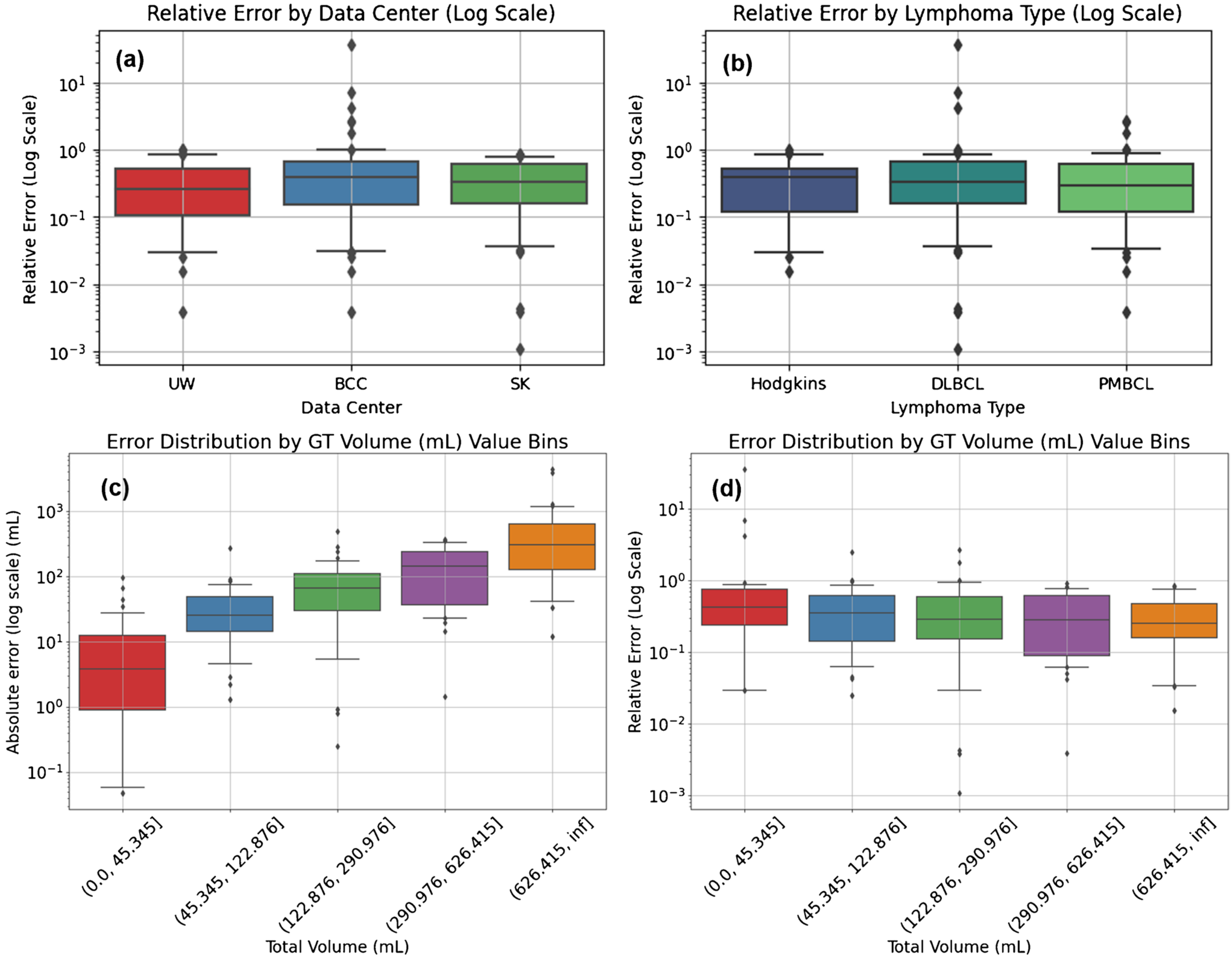

Fig. 8 Relative TMTV error (**a**) by center and (**b**) by lymphoma type. Also shown are plots of absolute error (**c**) and relative error (**d**) for 5 different tumor volume bins. Relative error is defined as absolute error between total tumor volumes calculated via TMTV-Net vs. ground truth, normalized by ground truth. UW, SK, and BCC are three different centers in this study. *PMBCL*: primary mediastinal B cell lymphoma, *DLBCL*: diffuse large B cell lymphoma

centers. Utilizing optimal approaches in model development involves the inclusion of multiple radiologists for annotating training datasets and assessing segmentation. Additionally, it is advisable to explore algorithms that exhibit data efficiency and the capability to incorporate unlabeled data for semi-supervised learning, considering the limited availability of experts for data labeling and annotation [56, 57]. Our multi-site external testing was performed on data for which ground truth segmentations were prepared by nuclear medicine physicians utilizing a previously validated [58] approach, PET-Edge, to define the ground truth for segmentation. An alternative and more accurate approach involves obtaining multiple manual delineations by expert clinicians and generating a statistical consensus using majority voting methods such as STAPLE [59]. This provides a more reliable ground truth for training our proposed model. However, such an approach requires considerable time and expertise, making it less practical, especially in large-scale studies or resource-limited settings [60, 61]. Moreover, imprecise ground truths can lead to reduced precision at the edges of predicted masks. In this study, our objective was to create a model with sufficient generalizability to maintain consistent segmentation performance across diverse centers, varying lymphoma lesion sizes, and inconsistent ground truth data. To address this issue, we additionally integrated the Mumford-Shah loss in our model as an unsupervised term, aiming to alleviate inconsistencies introduced by diverse edge refinement techniques employed in manual segmentation tools across different expert users.

Detection rate and recall or sensitivity could be used for evaluating the suggested approach to detect a lesion even by one voxel; the average detection rate for our model on the entire external test datasets is 0.68 ± 0.19. The average relative error (ARE) and the absolute error (AE) in

TMTV prediction on external multi-centric dataset were ARE = 0.43 ± 0.54 and AE = 157.32 ± 378.12 (mL) for all the external test data (n = 518) and ARE = 0.30 ± 0.22 and AE = 82.052 ± 99.778 (mL) when the 10% outliers (n = 53) were excluded that were mostly from the cases with high TMTV values. The model developed in this work and the necessary pre-processing steps are publicly available for multi-site testing and clinical evaluation.

TMTV-Net represents a pivotal advancement in the field of tumor segmentation using 2-[^{18}F] FDG PET/CT scans, offering a unique set of capabilities that contribute to superior generalizability when compared to existing approaches. Its integration of deep supervision within the framework of 3D U-Nets, coupled with the implementation of multi-resolution techniques, plays a pivotal role in mitigating the challenges posed by dataset shift. This strategic combination not only enhances the precision of tumor segmentation but also ensures the model reliability across diverse datasets. Future work includes the development of a user interface for active learning, which will allow physicians to be more involved in our segmentation framework. In addition, we are investigating the addition of a convolutional layer to automatically learn the best possible combination of multiple-resolution models.

Conclusion

We have introduced and validated TMTV-Net, a fully automated segmentation network for lymphoma PET/CT images using a cascaded framework. The proposed cascaded segmentation model trained on the PET/CT images of lung cancer, lymphoma, and melanoma patients achieved good performance for TMTV segmentation on data from multi-site external dataset with different lymphomas including DLBCL, PMBCL, and Hodgkin. The small 2% drop in the overall performance on the external testing data demonstrates the effectiveness of our suggested approach to be generalized well on unseen data from different centers. Training the model with resampled inputs makes it more robust to resolution variations and reduces overfitting to a particular resolution. Our cascaded approach can be easily incorporated into various models and extended to address different segmentation tasks that may be affected by domain shifts or limited labeling resources in a plug-and-play fashion. The model developed in this work and the necessary pre-processing steps are made available for multi-site testing and clinical evaluation through a cloud-based platform, which is user-friendly and requires no coding. Future work also includes implementing a user interface for an active learning approach to add the physician-in-the-loop option to our segmentation framework.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00259-024-06616-x.

Funding This research was in part supported by the Canadian Institutes of Health Research (CIHR) Project Grant PJT-173231, Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2019–06467, GE Healthcare, and through computational resources and services provided by Microsoft for Health.

Data Availability The data that was used for model development (autoPET) is publicly available. To protect study participant privacy, we cannot share the testing data from BC Cancer, University of Wisconsin and Seoul St. Mary's Hospital.

Declarations

Ethics approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The retrospective study was conducted with the following ethics numbers: H19-01611 for the PMBCL study and H19-01866 for the DLBCL study and H19-001611 for Hodgkin study at BC Cancer, as well as KC11EISI0293 for the DLBCL study at Seoul St. Mary's Hospital and UW2016-0418 for DLBCL and Hodgkin's cases at University of Wisconsin.

Conflict of interest Ivan Klyuzhin, Carlos Uribe, and Arman Rahmim are co-founders of Ascinta Technologies Inc.

References

1. Cottereau A-S, Lanic H, Mareschal S, Meignan M, Vera P, Tilly H, et al. Molecular profile and FDG-PET/CT total metabolic tumor volume improve risk classification at diagnosis for patients with diffuse large B-cell lymphoma. Clin Cancer Res. 2016;22:3801–9.
2. Kostakoglu L, Martelli M, Sehn LH, Belada D. Baseline PET-derived metabolic tumor volume metrics predict progression-free and overall survival in DLBCL after first-line treatment: results from the phase 3 …. Blood [Internet]. 2017; Available from: https://www.sciencedirect.com/science/article/pii/S000649711981340X.
3. Vercellino L, Cottereau A-S, Casasnovas O, Tilly H, Feugier P, Chartier L, et al. High total metabolic tumor volume at baseline predicts survival independent of response to therapy. Blood. 2020;135:1396–405.
4. Ceriani L, Martelli M, Zinzani PL, Ferreri AJM, Botto B, Stelitano C, et al. Utility of baseline 18FDG-PET/CT functional parameters in defining prognosis of primary mediastinal (thymic) large B-cell lymphoma. Blood. 2015;126:950–6.
5. Ceriani L, Milan L, Martelli M, Ferreri AJM, Cascione L, Zinzani PL, et al. Metabolic heterogeneity on baseline ^{18}FDG-PET/CT scan is a predictor of outcome in primary mediastinal B-cell lymphoma. Blood. 2018;132:179–86.
6. Cottereau A-S, Versari A, Loft A, Casasnovas O, Bellei M, Ricci R, et al. Prognostic value of baseline metabolic tumor volume in early-stage Hodgkin lymphoma in the standard arm of the H10 trial. Blood. 2018;131:1456–63.
7. Mikhaeel NG, Smith D, Dunn JT, Phillips M, Møller H, Fields PA, et al. Combination of baseline metabolic tumour volume and early response on PET/CT improves progression-free survival prediction in DLBCL. Eur J Nucl Med Mol Imaging. 2016;43:1209–19.
8. Song M-K, Yang D-H, Lee G-W, Lim S-N, Shin S, Pak KJ, et al. High total metabolic tumor volume in PET/CT predicts worse

prognosis in diffuse large B cell lymphoma patients with bone marrow involvement in rituximab era. Leuk Res. 2016;42:1–6.

9. Sasanelli M, Meignan M, Haioun C, Berriolo-Riedinger A, Casasnovas R-O, Biggi A, et al. Pretherapy metabolic tumour volume is an independent predictor of outcome in patients with diffuse large B-cell lymphoma. Eur J Nucl Med Mol Imaging. 2014;41:2017–22.
10. Toledano MN, Desbordes P, Banjar A, Gardin I, Vera P, Ruminy P, et al. Combination of baseline FDG PET/CT total metabolic tumour volume and gene expression profile have a robust predictive value in patients with diffuse large B-cell lymphoma. Eur J Nucl Med Mol Imaging. 2018;45:680–8.
11. Chang C-C, Cho S-F, Chuang Y-W, Lin C-Y, Chang S-M, Hsu W-L, et al. Prognostic significance of total metabolic tumor volume on ^{18}F-fluorodeoxyglucose positron emission tomography/computed tomography in patients with diffuse large B-cell lymphoma receiving rituximab-containing chemotherapy. Oncotarget. 2017;8:99587–600.
12. Cheson BD, Fisher RI, Barrington SF, Cavalli F, Schwartz LH, Zucca E, et al. Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. J Clin Oncol. 2014;32:3059–68.
13. Ly J, Minarik D, Edenbrandt L, Wollmer P, Trägårdh E. The use of a proposed updated EARL harmonization of ^{18}F-FDG PET-CT in patients with lymphoma yields significant differences in Deauville score compared with current EARL recommendations. EJNMMI Res. 2019;9:65.
14. Genc M, Yildirim N, Coskun N, Ozdemir E, Turkolmez S. The variation of quantitative parameters and Deauville scores with different reconstruction algorithms in FDG PET/CT imaging of lymphoma patients. Revista Española de Medicina Nuclear e Imagen Molecular (English Edition). 2023;42(6):388–92.
15. Ruppert AS, Dixon JG, Salles G, Wall A, Cunningham D, Poeschel V, et al. International prognostic indices in diffuse large B-cell lymphoma: a comparison of IPI, R-IPI, and NCCN-IPI. Blood. 2020;135:2041–8.
16. Meignan M, Cottereau A-S, Specht L, Mikhaeel NG. Total tumor burden in lymphoma — an evolving strong prognostic parameter. Br J Radiol. 2021;94:20210448.
17. El-Galaly TC, Villa D, Cheah CY, Gormsen LC. Pre-treatment total metabolic tumour volumes in lymphoma: does quantity matter? Br J Haematol. 2022;197:139–55.
18. Cottereau A-S, Meignan M, Nioche C, Capobianco N, Clerc J, Chartier L, et al. Risk stratification in diffuse large B-cell lymphoma using lesion dissemination and metabolic tumor burden calculated from baseline PET/CT†. Ann Oncol. 2021;32:404–11.
19. Alderuccio JP, Kuker RA, Barreto-Coelho P, Martinez BM, Miao F, Kwon D, et al. Prognostic value of presalvage metabolic tumor volume in patients with relapsed/refractory diffuse large B-cell lymphoma. Leuk Lymphoma. 2022;63:43–53.
20. Barrington SF, Meignan M. Time to prepare for risk adaptation in lymphoma by standardizing measurement of metabolic tumor burden. J Nucl Med. 2019;60:1096–102.
21. Hasani N, Paravastu SS, Farhadi F, Yousefirizi F, Morris MA, Rahmim A, et al. Artificial intelligence in lymphoma PET imaging: a scoping review (current trends and future directions). PET Clin. 2022;17:145–74.
22. Veziroglu EM, Farhadi F, Hasani N, Nikpanah M, Roschewski M, Summers RM, et al. Role of artificial intelligence in PET/CT imaging for management of lymphoma. Semin Nucl Med. 2023;53:426–48.
23. Burggraaff CN, Rahman F, Kaßner I, Pieplenbosch S, Barrington SF, Jauw YWS, et al. Optimizing workflows for fast and reliable metabolic tumor volume measurements in diffuse large B cell lymphoma. Mol Imaging Biol. 2020;22:1102–10.
24. Weisman AJ, Kieler MW, Perlman S, Hutchings M, Jeraj R, Kostakoglu L, et al. Comparison of 11 automated PET segmentation methods in lymphoma. Phys Med Biol. 2020;65:235019.
25. Huang L, Denœux T, Tonnelet D, Decazes P, Ruan S. Deep PET/CT fusion with Dempster-Shafer theory for lymphoma segmentation. Machine Learning in Medical Imaging. Springer International Publishing; 2021. p. 30–9.
26. Berthon B, Spezi E, Galavis P, Shepherd T, Apte A, Hatt M, et al. Toward a standard for the evaluation of PET — auto-segmentation methods following the recommendations of AAPM task group No. 211: Requirements and implementation. Med Phys. 2017;44:4098–111.
27. Ilyas H, Mikhaeel NG, Dunn JT, Rahman F, Møller H, Smith D, et al. Defining the optimal method for measuring baseline metabolic tumour volume in diffuse large B cell lymphoma. Eur J Nucl Med Mol Imaging. 2018;45:1142–54.
28. Hu H, Decazes P, Vera P, Li H, Ruan S. Detection and segmentation of lymphomas in 3D PET images via clustering with entropy-based optimization strategy. Int J Comput Assist Radiol Surg. 2019;14:1715–24.
29. Weisman AJ, Kim J, Lee I, McCarten KM, Kessel S, Schwartz CL, et al. Automated quantification of baseline imaging PET metrics on FDG PET/CT images of pediatric Hodgkin lymphoma patients. EJNMMI Phys. 2020;7:76.
30. Weisman AJ, Kieler MW, Perlman SB, Hutchings M, Jeraj R, Kostakoglu L, et al. Convolutional neural networks for automated PET/CT detection of diseased lymph node burden in patients with lymphoma. Radiol Artif Intell. 2020;2:e200016.
31. Blanc-Durand P, Jégou S, Kanoun S, Berriolo-Riedinger A, Bodet-Milin C, Kraeber-Bodéré F, et al. 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. 2021;48:1362–70.
32. Shi T, Jiang H, Wang M, Diao Z, Zhang G, Yao YD. Metabolic anomaly appearance aware U-Net for automatic lymphoma segmentation in whole-body PET/CT scans. IEEE J Biomed Health Inform. 2023.
33. Yousefirizi F, Holloway C, Alexander A, Tonseth P, Uribe C, Rahmim A. Tumor segmentation of multi-centric whole-body PET/CT images from different cancers using a 3D convolutional neural network. J Nucl Med. 2022;63:2517–2517.
34. Jemaa S, Fredrickson J, Carano RAD, Nielsen T, de Crespigny A, Bengtsson T. Tumor segmentation and feature extraction from whole-body FDG-PET/CT using cascaded 2D and 3D convolutional neural networks. J Digit Imaging. 2020;33:888–94.
35. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn. 2021;110:457–506.
36. Gatidis S, Hepp T, Früh M, La Fougère C, Nikolaou K, Pfannenberg C, Schölkopf B, Küstner T, Cyran C, Rubin D. A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions. Sci Data. 2022;9(1):601. Available from: https://wiki.cancerimagingarchive.net/x/LwKPBQ.
37. Zhang M, Levine S, Finn C. Memo: Test time robustness via adaptation and augmentation. Adv Neural Inf Process Syst. 2022;35:38629–42.
38. Matsunaga K, Hamada A, Minagawa A, Koga H. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. arXiv preprint arXiv:1703.03108. 2017;
39. Jin H, Li Z, Tong R, Lin L. A deep 3D residual CNN for false-positive reduction in pulmonary nodule detection. Med Phys. 2018;45:2097–107.
40. Jha AK, Bradshaw TJ, Buvat I, Hatt M, Prabhat KC, Liu C, Obuchowski NF, Saboury B, Slomka PJ, Sunderland JJ, Wahl RL. Nuclear medicine and artificial intelligence: best

practices for evaluation (the RELAINCE guidelines). J Nucl Med. 2022;63(9):1288–99.

41. Saboury B, Bradshaw T, Boellaard R, Buvat I, Dutta J, Hatt M, et al. Artificial intelligence in nuclear medicine: opportunities, challenges, and responsibilities toward a trustworthy ecosystem. J Nucl Med. 2023;64:188–96.
42. Gatidis S, Früh M, Fabritius M, Gu S, Nikolaou K, La Fougère C, Ye J, He J, Peng Y, Bi L. The autoPET challenge: Towards fully automated lesion segmentation in oncologic PET/CT imaging. preprint at Research Square (Nature Portfolio). 2023.
43. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26:1045–57.
44. Gatidis S, Kuestner T. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions [Internet]. The Cancer Imaging Archive; 2022. Available from: https://wiki.cancerimagingarchive.net/x/LwKPBQ.
45. Shrestha A, Watkins A, Carlos U. RT-Utils: a minimal Python library to facilitate the creation and manipulation of DICOM RTStructs. GitHub; 2022. Available from: https://github.com/qurit/rt-utils/tree/main.
46. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412. 2017 Oct 25. cs.LG]. 2017. Available from: http://arxiv.org/abs/1710.09412.
47. Lee C-Y, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. In: Lebanon G, Vishwanathan SVN, editors. Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics. San Diego: PMLR; 2015. p. 562–70.
48. Graziani M, Lompech T, Müller H, Depeursinge A, Andrearczyk V. On the scale invariance in state of the art CNNs trained on ImageNet. Mach Learn Knowl Extr. 2021;3:374–91.
49. Kim B, Ye JC. Mumford-Shah loss functional for image segmentation with deep learning. IEEE Trans Image Process. 2020;29:1856–66.
50. Yousefirizi F, Shiri I, Joo HO, Bloise I, Martineau P, Wilson D, et al. Semi-supervised learning towards automated segmentation of PET images with limited annotations: application to lymphoma patients [Internet]. arXiv [physics.med-ph]. 2022. Available from: http://arxiv.org/abs/2212.09908.
51. Yousefirizi F, Ahamed S, Joo HO, Bloise I, Saboury B, Rahmim A. Semi-supervised and unsupervised convolutional neural networks for automated lesion segmentation in PET imaging of lymphoma. J Nucl Med. 2022;63:3351.
52. Yousefirizi F, Dubljevic N, Ahamed S, Bloise I, Gowdy C, Joo HO, et al. Convolutional neural network with a hybrid loss function for fully automated segmentation of lymphoma lesions in FDG PET images. Medical Imaging 2022: Image Processing. SPIE; 2022. p. 214–20.
53. Huang L, Ruan S, Decazes P, Denœux T. Lymphoma segmentation from 3D PET-CT images using a deep evidential network. Int J Approx Reason. 2022;149:39–60.
54. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing; 2022. p. 272–84.
55. Hadjiiski L, Cha K, Chan H-P, Drukker K, Morra L, Näppi JJ, et al. AAPM task group report 273: recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Med Phys. 2023;50:e1-24.
56. Bradshaw T, Boellaard R, Dutta J, Jha A, Jacobs P, Li Q, et al. Pitfalls in the development of artificial intelligence algorithms in nuclear medicine and how to avoid them. J Nucl Med. 2022;63:2724–2724.
57. Bradshaw TJ, Boellaard R, Dutta J, Jha AK, Jacobs P, Li Q, et al. Nuclear medicine and artificial intelligence: best practices for algorithm development. J Nucl Med [Internet]. 2021; Available from: https://doi.org/10.2967/jnumed.121.262567.
58. Yousefirizi F, Bloise I, Martineau P, Wilson D, Benard F, Bradshaw TB, et al. Reproducibility of a semiautomatic gradient-based segmentation approach for lymphoma PET. In: EANM abstract book, a supplement of the European Journal of Nuclear Medicine and Molecular Imaging (EJNMMI). Springer Science+Business Media; 2021.
59. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging. 2004;23:903–21.
60. Andrearczyk V, Oreiller V, Abobakr M, Akhavanallaf A, Balermpas P, Boughdad S, et al. Overview of the HECKTOR challenge at MICCAI 2022: automatic head and neck tumor segmentation and outcome prediction in PET/CT. Head Neck Tumor Chall. 2022;2023(13626):1–30.
61. Yousefirizi F, Jha AK, Brosch-Lenz J, Saboury B, Rahmim A. Toward high-throughput artificial intelligence-based segmentation in oncological PET imaging. PET Clin. 2021;16:577–96.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Fereshteh Yousefirizi[1] · **Ivan S. Klyuzhin**[1] · **Joo Hyun O**[2] · **Sara Harsini**[3] · **Xin Tie**[4] · **Isaac Shiri**[5] · **Muheon Shin**[4] · **Changhee Lee**[4] · **Steve Y. Cho**[4] · **Tyler J. Bradshaw**[4] · **Habib Zaidi**[5,6,7,8] · **François Bénard**[3,9] · **Laurie H. Sehn**[3,10] · **Kerry J. Savage**[3,10] · **Christian Steidl**[3,11] · **Carlos F. Uribe**[1,3,9] · **Arman Rahmim**[1,9,12,13]

✉ Fereshteh Yousefirizi
frizi@bccrc.ca

[1] Department of Integrative Oncology, BC Cancer Research Institute, 675 West 10Th Avenue, Vancouver, BC V5Z 1L3, Canada

[2] College of Medicine, Seoul St. Mary's Hospital, The Catholic University of Korea, Seoul, Republic of Korea

[3] BC Cancer, Vancouver, BC, Canada

[4] Department of Radiology, University of WI–Madison, Madison, WI, USA

[5] Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva, Switzerland

[6] University Medical Center Groningen, University of Groningen, Groningen, Netherlands

[7] Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark

[8] University Research and Innovation Center, Óbuda University, Budapest, Hungary

[9] Department of Radiology, University of British Columbia, Vancouver, BC, Canada

[10] Centre for Lymphoid Cancer, BC Cancer, Vancouver, Canada

[11] Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, Canada

[12] Departments of Physics and Biomedical Engineering, University of British Columbia, Vancouver, BC, Canada

[13] Department of Biomedical Engineering, University of British Columbia, Vancouver, Canada