Physical and Engineering Sciences in Medicine
https://doi.org/10.1007/s13246-024-01408-x

SCIENTIFIC PAPER

Semi-supervised learning towards automated segmentation of PET images with limited annotations: application to lymphoma patients

Fereshteh Yousefirizi[1] · Isaac Shiri[2] · Joo Hyun O[3] · Ingrid Bloise[4] · Patrick Martineau[4] · Don Wilson[4,5] · François Bénard[4] · Laurie H. Sehn[4,6] · Kerry J. Savage[4,6] · Habib Zaidi[2,7,8,9] · Carlos F. Uribe[1,4,5] · Arman Rahmim[1,4,5,10]

Received: 31 May 2023 / Accepted: 18 February 2024
© Australasian College of Physical Scientists and Engineers in Medicine 2024

Abstract

Manual segmentation poses a time-consuming challenge for disease quantification, therapy evaluation, treatment planning, and outcome prediction. Convolutional neural networks (CNNs) hold promise in accurately identifying tumor locations and boundaries in PET scans. However, a major hurdle is the extensive amount of supervised and annotated data necessary for training. To overcome this limitation, this study explores semi-supervised approaches utilizing unlabeled data, specifically focusing on PET images of diffuse large B-cell lymphoma (DLBCL) and primary mediastinal large B-cell lymphoma (PMBCL) obtained from two centers. We considered 2-[^{18}F]FDG PET images of 292 patients PMBCL (n = 104) and DLBCL (n = 188) (n = 232 for training and validation, and n = 60 for external testing). We harnessed classical wisdom embedded in traditional segmentation methods, such as the fuzzy clustering loss function (FCM), to tailor the training strategy for a 3D U-Net model, incorporating both supervised and unsupervised learning approaches. Various supervision levels were explored, including fully supervised methods with labeled FCM and unified focal/Dice loss, unsupervised methods with robust FCM (RFCM) and Mumford-Shah (MS) loss, and semi-supervised methods combining FCM with supervised Dice loss (MS + Dice) or labeled FCM (RFCM + FCM). The unified loss function yielded higher Dice scores (0.73 ± 0.11; 95% CI 0.67–0.8) than Dice loss (p value < 0.01). Among the semi-supervised approaches, RFCM + αFCM ($\alpha = 0.3$) showed the best performance, with Dice score of 0.68 ± 0.10 (95% CI 0.45–0.77), outperforming MS + αDice for any supervision level (any α) ($p < 0.01$). Another semi-supervised approach with MS + αDice ($\alpha = 0.2$) achieved Dice score of 0.59 ± 0.09 (95% CI 0.44–0.76) surpassing other supervision levels ($p < 0.01$). Given the time-consuming nature of manual delineations and the inconsistencies they may introduce, semi-supervised approaches hold promise for automating medical imaging segmentation workflows.

Keywords PET · Segmentation · Lymphoma · Quantification · Unsupervised · Semi-supervised learning · Fuzzy clustering

✉ Fereshteh Yousefirizi
frizi@bccrc.ca

1 Department of Integrative Oncology, BC Cancer Research Institute, Vancouver, BC, Canada

2 Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva, Switzerland

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

4 BC Cancer, Vancouver, BC, Canada

5 Department of Radiology, University of British Columbia, Vancouver, Canada

6 Centre for Lymphoid Cancer, BC Cancer, Vancouver, Canada

7 University Medical Center Groningen, University of Groningens, Groningen, Netherlands

8 Department of Nuclear Medicine, University of Southern Denmark, Vancouver, Odense, Denmark

9 University Research and Innovation Center, Óbuda University, Budapest, Hungary

10 Departments of Physics and Biomedical Engineering, University of British Columbia, Vancouver, Canada

Published online: 21 March 2024

Introduction

Precise measurement of disease burden is crucial for improving therapy response assessment and outcome prediction in lymphoma positron emission tomography (PET) scans [1]. The prognostic power of total metabolic tumor volume (TMTV), assessed through comprehensive whole-body ^{18}F-fluorodeoxyglucose 2-[^{18}F]FDG PET scans, has undergone thorough validation in the domain of lymphoma [2–12]. However, the automated and accurate segmentation of lesions from whole-body PET images represents a significant obstacle that must be overcome to facilitate the widespread utilization of quantitative imaging biomarkers in clinical settings, such as radiomics analysis. This step is essential for computing the total metabolic tumor volume (TMTV) and conducting analyses on individual lesions.

Despite the considerable heterogeneity in lesion characteristics such as location, size, and contrast, simple thresholding methods continue to be widely used in clinical workflows, particularly for cancer types like lymphoma [1, 13, 14]. Integrating the statistical distinctions among uptake regions and the surrounding tissues, advanced segmentation techniques have been developed, including active contour models [15], region growing, and clustering algorithms such as Gaussian mixture models (GMM) [16] or fuzzy C-means (FCM) [17]. Recently Cui et al. [18] suggested a technique for lesion segmentation from PET images, incorporating definition density peak clustering to segment the lesion and normal tissue in 2D in an unsupervised manner. These techniques aim to enhance segmentation accuracy beyond basic methods.

Furthermore, in conventional variational segmentation approaches, an energy function such as Mumford-Shah [19] is minimized to classify image voxels (pixels) into distinct classes without the need for ground truth or supervision [19–21]. While these techniques have been extensively employed for medical image segmentation, they often come with computational complexity and limited capabilities in semantic segmentation, often necessitating user input to define parameters or scanner settings [22]. AI has the potential to quantify the disease burden by segmenting the TMTV of lymphoma lesions [1, 23–26]. Current segmentation approaches have predominantly emphasized AI-based techniques, often disregarding the potential benefits offered by conventional approaches [27]. The effectiveness of supervised AI-based techniques improves proportionally with the expansion of labeled training data [28, 29]. Consequently, insufficient labeled data may hinder performance expectations, making semi-supervised approaches beneficial in such cases.

The implementation of advanced supervised learning methods for tumor segmentation in PET scans faces challenges due to the need for precise and consistent ground truth annotations in an adequate training dataset. Ground truth in medical image segmentation refers to the precise boundary of the object of interest, typically determined through histopathological analysis of an excised tumor. However, obtaining ground truth data is not always feasible [30]. As a substitute, the consensus derived from multiple manual segmentations by different experts is often used as an alternative form of ground truth (with a single expert frequently delineating a given tumor in practice) [31, 32]. Nonetheless, intra- and inter-observer variabilities introduce reproducibility challenges in ground truth generation [33] which, in turn, impact the performance of supervised learning approaches. Unsupervised segmentation techniques can effectively mitigate the impact of uncertainty and inconsistency inherent in ground truths during the learning phase.

These limitations serve as a driving force for exploring advanced AI techniques that can be trained with varying levels of supervision, aiming to alleviate the manual ground truth labeling and annotation. Different levels of supervision can be employed when training a segmentation model, ranging from pixel/voxel-level annotations in supervised learning [34–36], to image-level or imprecise annotations in weakly-supervised learning [37], and even to no annotations in unsupervised learning [38–41]. Shi et al. [42] proposed an unsupervised image generation approach, utilizing anatomical-metabolic consistency representations obtained from co-aligned PET/CT scans, to enhance the accuracy of lymphoma segmentation in PET/CT images, addressing limitations posed by insufficient annotated data and tumor heterogeneity. Lian et al. [43] introduced an unsupervised method PET-CT image segmentation. Their approach employed a belief function to effectively represent uncertain image information and employed an adaptive distance metric to incorporate spatial information into the segmentation process. Joint unsupervised learning was also suggested as an approach for segmentation, wherein images are progressively clustered and deep representations are learned using a convolutional neural network. Additionally, the combination of joint unsupervised learning with clustering techniques like k-means was recommended for medical image segmentation [40].

The choice of loss function plays a pivotal role in defining the optimization problem and directly impacts the convergence of the segmentation model during training. Kim et al. [44] proposed a loss function based on the Mumford-Shah (MS) functional, which is designed for unsupervised segmentation. They demonstrated that the discrete implementation of the MS loss function can be regarded as a k-means clustering approach with total variational regularization that effectively suppresses noise in the membership function [44]. By utilizing the MS loss, they were able to train

the segmentation network without relying on ground truth labels. Additionally, the MS loss term can be incorporated into dice or cross-entropy losses as a regularized function in supervised approaches, thereby aiding the network in enhancing its segmentation performance [44]. Considering the inherent suitability of fuzzy clustering methods to the low-resolution characteristics of nuclear medicine imaging, recent advancements have proposed loss functions based on Fuzzy C-Means (FCM) [45] enabling their application in supervised, semi-supervised, and unsupervised segmentation scenarios.

In this paper, we explore the efficacy of semi-supervised approaches in the context of lymphoma lesion segmentation within PET scans of DLBCL and PMBCL patients. Specifically, our contributions are twofold: (i) Investigating the effectiveness of different semi-supervised approaches. (ii) Assessing the adaptability of semi-supervised techniques for segmentation of multiple tumor (TMTV) in PET scans. The subsequent sections delve into the methods in details, followed by presenting the results, discussion and conclusions.

Methods

PET scans

Table 1 summarizes the data used in this study. PET images (n = 292) included baseline and interim scans of patients with DLBCL (n = 90) from two different centers: BC Cancer (BCC) with (n = 86) cases of limited stage diagnosed after 2005, and St. Mary's Hospital (SM) with (n = 102) cases that were diagnosed after 2014 and stages varied from I to IV, and patients with PMBCL (n = 104) from BC Cancer. All patients were managed according to the departmental protocol (BCC and SM) which states a 6 h fast and sampled blood glucose of < 200 ng/dL prior to the injection of 300–400 MBq [^{18}F]FDG followed by a 60-min uptake phase.

Ground truth segmentation

The ground truth volumes of interest (VOI) were delineated by experienced nuclear medicine physicians using a built in-house semi-automatic workflow for MIM (MIM Software, USA), where lesions were drawn utilizing the software's gradient-based segmentation tools (PETedge and PETedge+), designated into different body parts (neck, chest, abdomen and pelvis, muscles, bones, central nervous system and other). As previously demonstrated [46] this workflow has shown reproducibility for lesion segmentation and helps to reduce the inter-observer variability.

Preprocessing and data augmentation

PET images were adjusted to a standardized resolution of $4 \times 4 \times 2$ mm^3 using trilinear interpolation and the corresponding segmentation masks have been resized using nearest neighbor interpolation. A slice thickness of 2 mm was specifically chosen to preserve fine image details that might otherwise be lost if interpolated at a larger voxel size. This resolution is approximately close to the weighted average voxel spacing of our dataset. The choice of a non-isotropic voxel size ($4 \times 4 \times 2$ mm^3) was made to balance computational efficiency with the preservation of the object shape to our segmentation objectives. We have selected 4mm based on the in-plane resolution that we have in our dataset. PET image intensities can exhibit considerable variability within and between images. After SUV conversion, PET SUV range was transformed from [0, 30] SUV to [0, 1] to capture a broader range of intensities. To mitigate these intensity differences, we implemented Z-score normalization independently for each scan. This normalization process involved computing the mean and standard deviation solely based on voxels with non-zero intensities corresponding to the body region.

To increase the diversity in lesion size and shape, we incorporated scaling with a random factor and elastic

Table 1 Multi-center dataset information from different lymphoma types (all the presented cases were annotated)

Center	Lymphoma type	Matrix size	Voxel spacing (mm^3)	Average injected radioactivity (MBq)	Scanner models
BC Cancer, Canada (BCC)	PMBCL	168 × 168 (n = 20) 192 × 192 (n = 84)	4.06 × 4.06 × 2 (n = 20) 4.06 × 4.06 × 3.27 (n = 119)	347.5 ± 52.6	GE (Discovery D600 and D690)
BC Cancer, Canada (BCC)	DLBCL (stage I to II)	192 × 192 (n = 86)	3.65 × 3.65 × 3.27 (n = 86)	335.9 ± 50.8	
St. Mary's Hospital, South Korea (SM)	DLBCL (stages I to IV)	168 × 168 (n = 27) 192 × 192 (n = 15)	3.65 × 3.65 × 3.27 (n = 15) 3.65 × 3.65 × 5 (n = 27)	252.0 ± 48.1	GE (Discovery 710)
St. Mary's Hospital South Korea (SM)	DLBCL (stages I to IV)	168 × 168 (n = 60)	4.07 × 4.07 × 5 (n = 60)	240.5 ± 47	Siemens (Biograph40 TruePoint)

deformations as data augmentation techniques. These augmentations were employed to assist the model in learning the wide range of lesion sizes and shapes encountered. Several augmentation strategies were employed to increase the complexity of the training data, including: (i) spatial and intensity transformations, such as rotation in random directions (less than 25° i.e. The rotations were performed in random directions with the angle uniformly sampled from the range of [0, 25] degrees), (ii) scaling with a random factor within the range of 0.8 to 1.2, (iii) elastic deformations, and (iv) gamma corrections, where the parameter γ was randomly sampled from a uniform distribution spanning 0.8 to 1.2.

Segmentation network architecture

The 3D U-Net architecture [47] consists of standard convolutional blocks, which include a 3 × 3 × 3 convolution operation, a normalization layer, and a ReLU activation function. We utilized residual blocks, accompanied by a concurrent spatial and channel squeeze and excitation module, denoted as SE normalization (depicted as blue blocks in Fig. 1). The SE module gives weight to the feature maps, so that the network can emphasize its attention adaptively. The SE module "squeezes" along the spatial domain and "excites" along the channels, enabling the model to emphasize meaningful features while suppressing weaker ones. We implemented SE normalization layers with a fixed reduction ratio (r = 2), which controls the size of the bottleneck within the SE normalization layers. Additionally, we used instance normalization to reduce memory consumption [48].

Within the network architecture, we used learnable downsampling blocks (green blocks in Fig. 1) that consist of a 3 × 3 × 3 strided convolutional layer, instance norm, ReLU activation, and the SE module. In the decoder, upsampling blocks using 3 × 3 × 3 transposed convolutions was used (yellow blocks in Fig. 1). To further enhance the network receptive field, we incorporated a downsampling block with a kernel size of 7 × 7 × 7 immediately after the input. Additionally, the model output, generated by the last convolutional layer, is passed through a sigmoid activation function with a kernel size of 1 × 1 × 1. Considering the relatively large size of PET images, we adopted a strategy of training the 3D model using randomly extracted patches with the size of 128 × 128 × 64 voxels, and they do not necessary include any parts of a tumor. During training, we employed a batch size of 2, allowing for efficient processing of manageable subsets of the data while still capturing the essential information required for accurate segmentation. We applied a threshold of 0.5 to all prediction maps across all approaches in this study.

The model underwent 400 epochs of training, utilizing the Adam optimizer [41] with β1 = 0.9 and β2 = 0.99 for the exponential decay rates of moment estimates. To optimize the learning process, we applied a cosine annealing schedule, gradually reducing the learning rate from a maximum value of $lr_{max} = 10^{-4}$ to a minimum value of $lr_{min} = 10^{-6}$ every 25 epochs. This adjustment was made at each epoch to ensure the model ability to converge effectively. All models were implemented using Python with PyTorch library. We trained and tested all models on NVIDIA V100 GPUs.

Unsupervised learning

Fuzzy clustering-based loss functions

Unsupervised learning techniques have emerged as valuable tools in addressing the limitations posed by the

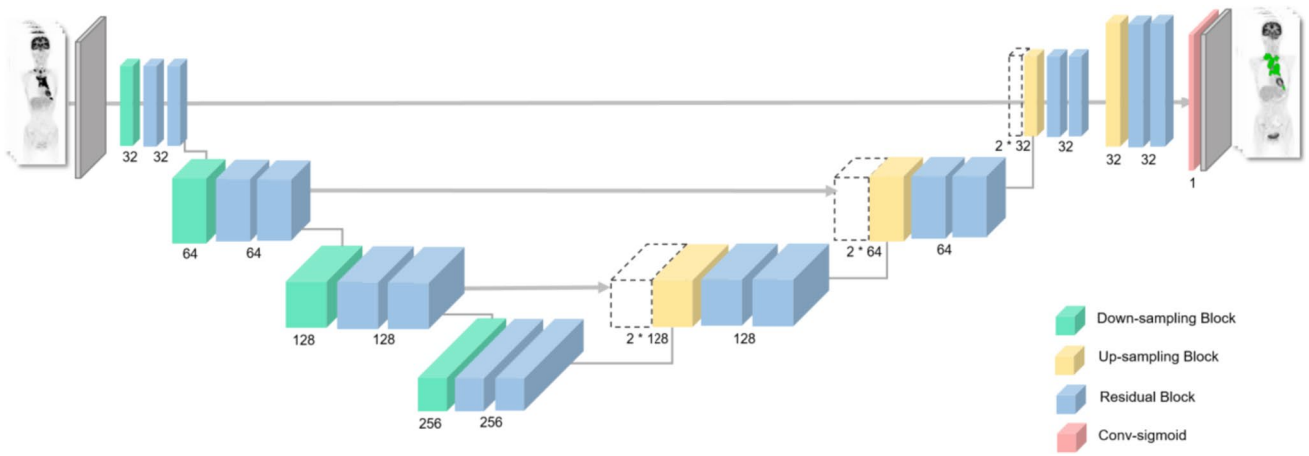

Fig. 1 Encoder-Decoder Network with residual blocks. The number of output channels is depicted under blocks of each group. Max pooling operations in the encoder of the network are replaced by learnable downsampling blocks (green blocks). The upsampling blocks in the decoder of the network were implemented by a 3×3×3 transposed convolution instead (yellow blocks)

scarcity and heterogeneity of labeled data in medical imaging. In the realm of AI loss function design, inspiration can be drawn from conventional segmentation methods, which serve as a foundation for both supervised and unsupervised approaches. These traditional clustering-based techniques, such as k-means, FCM, and GMM, leverage objective functions to cluster voxels based on their intensity statistics, enabling segmentation of medical images. Among these techniques, FCM stands out due to its simplicity, robustness, and computational efficiency, making it widely employed in both supervised and unsupervised learning tasks. However, clustering methods like FCM primarily focus on intensity information and often neglect crucial spatial details, making them prone to errors caused by image noise and artifacts. To overcome this limitation, various adaptations of the FCM approach have been proposed. These modifications aim to incorporate spatial constraints and contextual information, allowing for the influence of neighboring voxel labels on the labeling of a given voxel. By integrating spatial information, these enhanced versions of FCM enhance segmentation accuracy and robustness, addressing the challenges posed by noisy and artifact-ridden medical images [49–54]. The inclusion of a Markov-random-field (MRF) based regularization term in the Robust FCM (RFCM) enables the model to effectively account for variations in membership functions within local neighborhoods (equation 1) [55]:

$$J_{RFCM} = \sum_{j\in\Omega}\sum_{k=1}^{C} u_{jk}^{q}\left\|y_j - v_k\right\|^2 + \beta \sum_{j\in\Omega}\sum_{k=1}^{C} u_{jk}^{q} \sum_{j\in N_j}\sum_{m\epsilon M_k} u_{lm}^{q} \tag{1}$$

where the membership function u_{jk} represents the fuzzy assignment of the jth voxel to the kth class. The class centroid v_k is defined and y_j denotes the voxel value at location j, C indicates the total number of classes while Ω represent the spatial domain of the image. The fuzzy overlap between clusters is regulated by parameter q and the second term in the above expression corresponds the $J_{spatial}$. Here, N_j represents the set of neighboring voxels of voxel j, and M_k is a set containing all class numbers except k. The spatial smoothness term is weighted by β to control its impact [45]. To solve this optimization problem, the method of Lagrange multipliers is employed to enforce the given constraint, and partial derivatives with respect to v_k and u_{jk} are computed. Building upon the definition of the Fuzzy C-Means (FCM) objective function, a novel FCM loss function is proposed by Chen et al. [45] for both supervised and unsupervised training scenarios. By utilizing the principles and concepts of the FCM algorithm, the suggested FCM loss function serves as a valuable tool in training models for various tasks. Its application in both supervised and unsupervised settings demonstrates its versatility and effectiveness in capturing and leveraging the inherent characteristics of the data for accurate and reliable segmentation.

The membership functions, u, in the proposed method are modeled based on the objective function of RFCM. To achieve this, the softmax output of the last layer is utilized ($f(y;\theta)$) [45]. By employing the softmax function, the membership values are obtained, reflecting the degree of association of each voxel with different classes. This approach leverages the principles of RFCM to effectively assign membership values and enable accurate and robust segmentation results (equation 2):

$$\begin{aligned} L_{RFCM}(y;\theta) &= \sum_{j\in\Omega}\sum_{k=1}^{C} f_{jk}^{q}(y;\theta)\left\|y_j - v_k\right\|^2 \\ &+ \beta \sum_{j\in\Omega}\sum_{k=1}^{C} f_{jk}^{q}(y;\theta) \sum_{j\in N_j}\sum_{m\epsilon M_k} f_{lm}^{q}(y;\theta) \end{aligned} \tag{2}$$

where $f_{jk}(y;\theta)$ is the kth channel softmax output of the CNN at location j, and the class mean v_k is defined as follows (equation 3):

$$v_k = \frac{\sum_{j\epsilon\Omega} f_{jk}^{q}(y;\theta) y_j}{\sum_{j\epsilon\Omega} f_{jk}^{q}(y;\theta)} \tag{3}$$

Mumford-Shah loss function

The MS loss function also helps the network utilize unlabeled images as elements of the training data (equation 4).

$$L_{MS} = \sum_{k=1}^{C}\sum_{j\epsilon\Omega} f_{jk}\left\|y_j - v_k\right\|^2 + \eta \sum_{k=1}^{C}\sum_{j\epsilon\Omega}\left|\nabla f_{jk}\right| \tag{4}$$

where f_{jk} is the softmax output of CNN, considering that $\sum_{k=1}^{c} f_{jk} = 1$ and $\left|\nabla f_{jk}\right|$ is the approximation of total variant of f_{jk} and can be approximated by $\nabla f_{jk} = f_{(j+1)k} - f_{jk}$. The average voxel intensity is shown by v_k here as well. By considering $\eta = 0$, the L_{MS} loss function is changed to a FCM loss with $q = 1$ [45].

Semi-supervised learning

The concept of combining a weighted supervised loss with an unsupervised loss forms the foundation of semi-supervised learning techniques, proposed by Kim et al. [44] and Chen et al. [44]. In their work, they suggested the idea of incorporating a weighted supervised loss alongside the unsupervised loss, specifically in scenarios where labeled data is available. This approach allows for leveraging both labeled and unlabeled data to enhance the training process and improve the performance of the model (equation 5):

$$L^{\alpha}_{semi}(y, g;\theta) = L_{unsupervised}(y;\theta) + \alpha L_{supervised}(y, g;\theta) \quad (5)$$

where α is a weighting parameter that controls the strength of the supervised term, and g denotes ground truth. Setting a small value for α in the network training prioritizes the characterization of intensity distributions rather than relying heavily on the ground truth of the annotated training dataset. This approach allows the network to learn and capture the underlying patterns and variations in the intensity distributions present in the data, leading to a more robust and flexible model.

Semi-supervised learning based on Mumford-Shah approach

Kim et al. [44] proposed a semi-supervised learning approach by combining Mumford-Shah (MS) loss and cross entropy (CE) loss. The MS loss was utilized as the unsupervised loss, while the CE loss served as the supervised loss in their framework. In this work, we explored the use of Dice loss as an alternative supervised term in the semi-supervised learning approach (equation 6):

$$L^{\alpha}_{semi-MS}(y, g;\theta) = L_{MS}(y;\theta) + \alpha L_{Dice}(y, g;\theta) \quad (6)$$

The combination of CE and Dice for the supervised part of the loss function was also explored in a recent study by our team [56].

Semi-supervised learning based on Fuzzy clustering approach

Incorporating a supervised loss function based on the FCM objective function, such as Label-based FCM [45]) allows for the design of a semi-supervised loss function that accounts for the inherent fuzziness in FCM classification. This addresses the incompatibility between Dice loss or cross-entropy and the fuzzy nature of FCM, enabling the development of a more comprehensive and effective loss function for semi-supervised learning (equation 7):

$$L_{FCM}(y;\theta) = \sum_{j\in\Omega}\sum_{k=1}^{C} f^{q}_{jk}(y;\theta)\left\|g_{jk} - \mu_k\right\|^2 \quad (7)$$

where g_{jk} represents the ground truth label at location j for k^{th} class, while μ_k denotes the class mean computed within the ground truth image g and can be defined as a constant ($\mu_k = 1$). The degree of fuzzy overlap between the softmax channels is controlled by the parameter q in equation 8:

$$L^{\alpha}_{semi-FCM}(y, g;\theta) = L_{RFCM}(y;\theta) + \alpha L_{FCM}(y, g;\theta) \quad (8)$$

Supervised learning

When employing deep neural networks for supervised segmentation, common choices for loss functions include cross-entropy loss, Dice loss, or a combination of the two. The variations in segmentation performance observed with different loss functions emphasize the importance of selecting an appropriate loss function, as it directly impacts the robustness and convergence of the segmentation model [57].

Loss functions used for medical image segmentation can be classified into three main categories: distribution-based losses (e.g., cross-entropy, Focal loss [58]), region-based losses (e.g., Dice coefficient), and boundary-based losses [44, 59] (e.g., MS). Additionally, combinations of these loss functions are often employed. Studies have indicated that the best performance is typically achieved with combined loss functions, such as the summation of cross-entropy and Dice similarity or the combination of Focal and Dice loss, known as the Unified Focal loss [60].

Training strategies

In total we conducted our study on 292 labeled PET images. We developed the segmentation model on 232 cases (including 104 PMBCL cases from BCC center and 188 DLBCL cases from SM and BCC centers) and experimented with the supervision level for training on the 232 cases. As depicted in the first 5 rows of Table 2, all 232 cases were utilized for both supervised and unsupervised training approaches, with and without consideration of their corresponding labels, respectively. For semi-supervised training, we used the 60 annotated cases and considered the remaining 172 cases as unannotated. For all the above-mentioned training strategies, 60 cases from SM center were considered as the external test set. In the annotated set in the above mentioned semi-supervised training approaches, we utilized the same number of cases from our three study cohorts (20 PMBCL cases from BCC, 20 DLBCL cases from BCC, and 20 DLBCL cases from SM).

To evaluate the effect of domain shift and the amount of performance drop due to generalizability of the trained model, we ran another training strategy for semi-supervised approaches (experiment II versus the above-mentioned strategy that we can consider it as experiment I now). In this experiment, we conducted two semi-supervised approaches including the training and external testing on 122 DLBCL case (42 cases from SM center and 80 cases from BCC). We considered only 30 annotated cases (10 from SM and 20 from BCC) and applied the trained semi-supervised model to 30 cases from BCC center as the external test set.

Table 2 Summary of quantitative image segmentation performance metrics (Mean ± SD)

Methods	Training			External test (n = 60)		
	#annotated	# unannotated	Hyper-parameters	DSC	Relative error of TMTV	Absolute error (mL)
Supervised (Unified Focal)	232	0	$\lambda = 0.5, \delta = 0.6, \gamma = 0.5$	**0.73 ± 0.11**	0.34 ± 0.23	134.36 ± 170.38
Supervised (Dice)			–	0.67 ± 0.10	**0.31 ± 0.21**	135.820 ± 189.28
Supervised (FCM)			$q = 2$	0.71 ± 0.01	0.32 ± 0.21	**132.33 ± 165.83**
Unsupervised (MS)	0	232	$\eta = 10^{-6}$	0.28 ± 0.04	0.60 ± 0.15	346.19 ± 441.31
Unsupervised (RFCM)			$q = 2, \beta = 0.0016$	**0.41 ± 0.06**	**0.41 ± 0.22**	**251.23 ± 351.85**
Semi-supervised (MS + Dice)	60	172	$\alpha = 0.1$	0.52 ± 0.08	0.32 ± 0.21	187.61 ± 276.59
			$\alpha = 0.2$	**0. 59 ± 0.09**	**0.30 ± 0.19**	**156.63 ± 223.85**
			$\alpha = 0.3$	0.53 ± 0.08	0.31 ± 0.21	181.71 ± 266.11
			$\alpha = 0.4$	0.53 ± 0.09	0.31 ± 0.21	172.68 ± 255.24
			$\alpha = 0.5$	0.51 ± 0.08	0.32 ± 0.21	190.79 ± 281.34
			$\alpha = 0.6$	0.39 ± 0.06	0.43 ± 0.22	259.43 ± 359.24
			$\alpha = 0.7$	0.34 ± 0.05	0.50 ± 0.19	297.29 ± 395.72
Semi-supervised (RFCM + FCM)			$\alpha = 0.1$	0.39 ± 0.06	0.43 ± 0.22	260.03 ± 361.73
			$\alpha = 0.2$	0.60 ± 0.09	0.31 ± 0.22	153.11 ± 218.53
			$\alpha = 0.3$	**0.68 ± 0.10**	**0.30 ± 0.19**	**131.15 ± 184.15**
			$\alpha = 0.4$	0.65 ± 0.01	0.31 ± 0.20	140.62 ± 196.70
			$\alpha = 0.5$	0.64 ± 0.01	0.31 ± 0.19	143.32 ± 200.06
			$\alpha = 0.6$	0.59 ± 0.09	0.30 ± 0.20	158.48 ± 227.29
			$\alpha = 0.7$	0.58 ± 0.09	0.30 ± 0.20	158.85 ± 230.05

Best performances in supervised, unsupervised, as well as MS + Dice and RFCM + FCM semi-supervised methods are shown in bold. For all the semi-supervised (MS + Dice) approaches $\eta = 10^{-6}$ was selected and for all the Semi-supervised (RFCM + FCM) approaches, $\beta = 0.0016$ was selected. Also for semi-supervised (RFCM + FCM) approaches, $q = 2$ was selected (TMTV: Total Metabolic Tumor Volume, DSC: Dice similarity coefficient)

Quantitative evaluation and statistical analysis

We evaluated the segmentation performance based on Dice similarity coefficient (DSC), formulated as follows, based on the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (FN) (equation 9):

$$DSC = \frac{2 \times TP}{(TP + FP) + (TP + FN)} \tag{9}$$

Following the recommended RELAINCE guideline framework for AI-based algorithm evaluation [31], it is important to assess the segmentation technique using the figure of merit specific to this segmentation task; i.e. TMTV quantification and radiomics analysis as aligned with the ultimate goals of segmentation. Besides TMTV, we computed PET metrics that are clinically relevant, such as SUVmax, SUVmean, and SUVmedian. Additionally, we extracted first-order (FO) radiomics features, including percentiles (10th and 90th), energy, interquartile range, kurtosis, mean absolute deviation, range, robust mean absolute deviation, root mean squared, total energy, and variance. These feature extractions were carried out following the image biomarker standardization initiative, utilizing LIFEx [61].

To evaluate the performance, we calculated the mean relative error compared to manual segmentation, employing Relative error (%) = ((predicted mask − ground truth)/(ground truth)) × 100. We also considered the absolute and relative error for TMTV prediction by the suggested approaches. We compared the different approaches using Wilcoxon signed rank test (a non-parametric statistical hypothesis test), and reported the mean ± SD and 95% confidence interval (CI) for the quantitative metrics. Furthermore, we benchmarked our approaches with varying levels of supervision against state-of-the-art segmentation techniques for lymphoma. Additionally, we applied our approaches to the publicly available dataset (autoPET [62, 67 and 68]).

Results

Qualitative analysis

For visual inspection and qualitative analysis of the segmentation results, Fig. 2 illustrates the 2D axial views of lesion

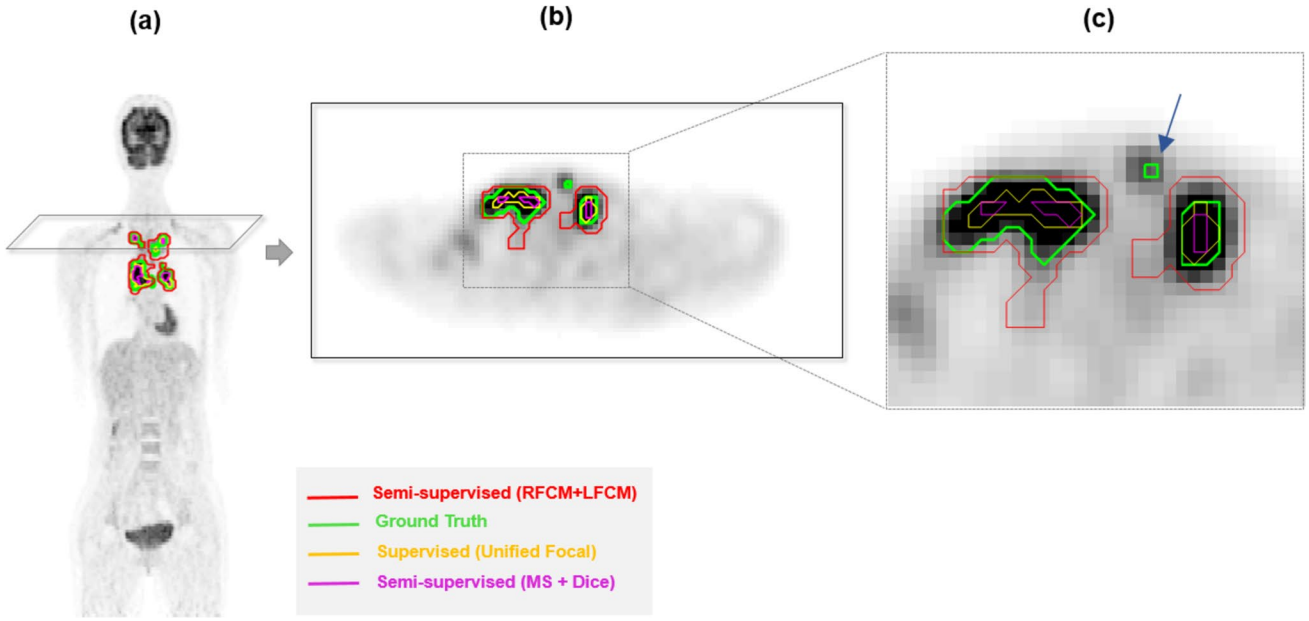

Fig. 2 Segmentation results achieved by the different levels of supervision in supervised and semi-supervised approaches. **a** A coronal view of a DLBCL patient, **b** is the axial view of the segmented lesions and **c** is the zoomed area of the segmented lesions. Visual inspection of compassion between the segmentation approaches with different supervision levels. Segmentation models with any supervision level could not segment the small lesion (blue arrow)

segmentation results of a patient from the external test set (SK) along with their zoomed version. As it is shown in this example, the predicted segmentation mask by the different supervised approaches are in good agreement with manual segmentations of lymphoma lesions presenting with different sizes, locations, textures, and contrast. The semi-supervised approach based on FCM, i.e., RFCM + αFCM, shows better qualitative and quantitative performance than MS + αDice. Also, Fig. 3 shows the segmented lesions of the patient by the same segmentation model with different levels of supervision; false positive regions segmented by unsupervised and semi-supervised (MS + αDice) approaches are shown in this figure. However, unsupervised methods based on RFCM and MS losses cannot capture the lesion area correctly.

Figure 4 shows a few representative outliers segmented by our techniques to show the wrong predicted areas by our technique. In these examples, the investigated approaches with various levels of supervision failed to properly segment the lymphoma lesions. This figure illustrates how false positives (left) and missed lesions (right) were caused by the low uptake of lesions (right) and high uptake in the background (left). The presence of nearby tissues with a relatively high uptake that could be mistaken for the tumor is another factor that can cause errors in the model predictions (see Fig. 4).

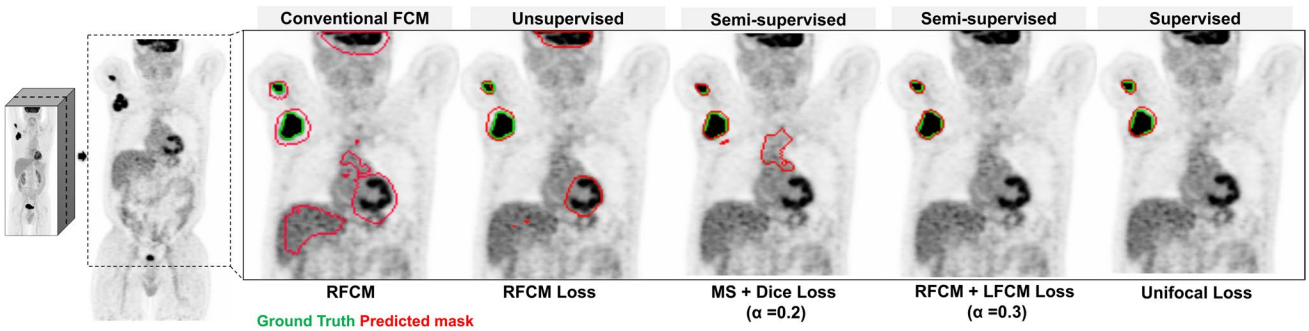

Fig. 3 Segmentation results on the coronal view achieved by the different levels of supervision in unsupervised, supervised and semi-supervised AI approaches for segmentation along with the conventional segmentation based on RFCM. (RFCM: Robust FCM)

Fig. 4 Axial views of ground truth and 3D U-net trained by different levels of supervision on different cases where failure was observed resulting in outliers. This figure visually demonstrates the impact of false positives (left) and missed lesions (right), attributing them to low uptake in lesions (right) and high uptake in the background (left)

In Fig. 5, we visualize the probability map predictions obtained from our network using a semi-supervised learning approach. These predictions are generated through the utilization of FCM losses (RFCM + αFCM), and they are superimposed on the axial slice of a PET scan. Four cases of different probability threshold settings are displayed. The range of probability maps is shown on the image for each cluster.

Quantitative analysis

We compared our FCM-based methodologies under different levels of supervision, employing supervised Unified focal and Dice losses, unsupervised MS loss (Kim et al. [44]) and semi-supervised MS + Dice loss (Table 2). For the quantitative analysis of our suggested approaches, the segmentation performance of the techniques in terms of Dice score with different supervision level are presented in Table 2. Figure 6 compares the Dice coefficients of the various networks to demonstrate the significance of the differences using the signed ranked test, where a p value < 0.001 is regarded as significant. The supervised approach with Unified focal loss function yielded the highest Dice score [mean ± standard deviation (SD)] of 0.73 ± 0.11; 95% CI 0.67–0.8) compared to Dice loss (p value < 0.01). There is no significant difference between the segmentation performance of supervised Unified focal and FCM losses. Supervised approach with FCM provided the performance with Dice score of (0.71 ± 0.01; 95% CI 0.62–0.81). The semi-supervised approach by RFCM and FCM loss with $\alpha = 0.3$ showed the best performance among the semi-supervised approaches with Dice score (0.68 ± 0.10; 95% CI 0.45–0.77) (p value < 0.01). The best performer among MS + αDice semi-supervised approaches with $\alpha = 0.2$ showed Dice score of (0.59 ± 0.09; 95% CI 0.44–0.76) ($p < 0.01$). It was observed that the unsupervised approach with MS loss showed the lowest performance. With the exception of few non-significant differences, as shown in Fig. 6, most differences are significant and are shown in blue.

Table 2 also summarizes the performance of unsupervised, semi-supervised, and supervised approaches based on the absolute error and relative error for TMTV prediction. The minimum relative error and the minimum absolute error for the supervised techniques are respectively related to the supervised with Dice loss) and supervised with FCM loss, but their differences are not statistically significant (p value > 0.01). The lower absolute and relative error among the unsupervised approaches are related to unsupervised segmentation with RFCM loss (p value < 0.01) that is also aligned with the performance of the unsupervised approaches in terms of DSC. MS + αDice semi-supervised approach with $\alpha = 0.2$ is also the best performer in terms of absolute and relative TMTV error (p value < 0.01). RFCM+FCM with $\alpha = 0.3$ showed the lowest absolute and relative error.

For every segmented lesion, the results of the image-derived PET metrics are shown in Fig. 7, along with first-order and shape features percent relative error for various approaches with various levels of supervision and loss functions. This reveals that if the lesion is segmented by varying levels of supervision and losses, the segmented region contains the maximum value of the lesion, and the relative error RE for SUV_{max} is less than 1%. In any case, it is seen that errors for unsupervised approach by MS loss are high. Unsupervised with RFCM and semi-supervised approaches with MS+αDice ($\alpha = 0.6$ and 0.7) have higher relative errors in the SUV and FO based radiomics features compared to other techniques. The percent relative error of SUV_{mean} was less than 10% in techniques with some levels of supervision, including semi and supervised methods with both MS +

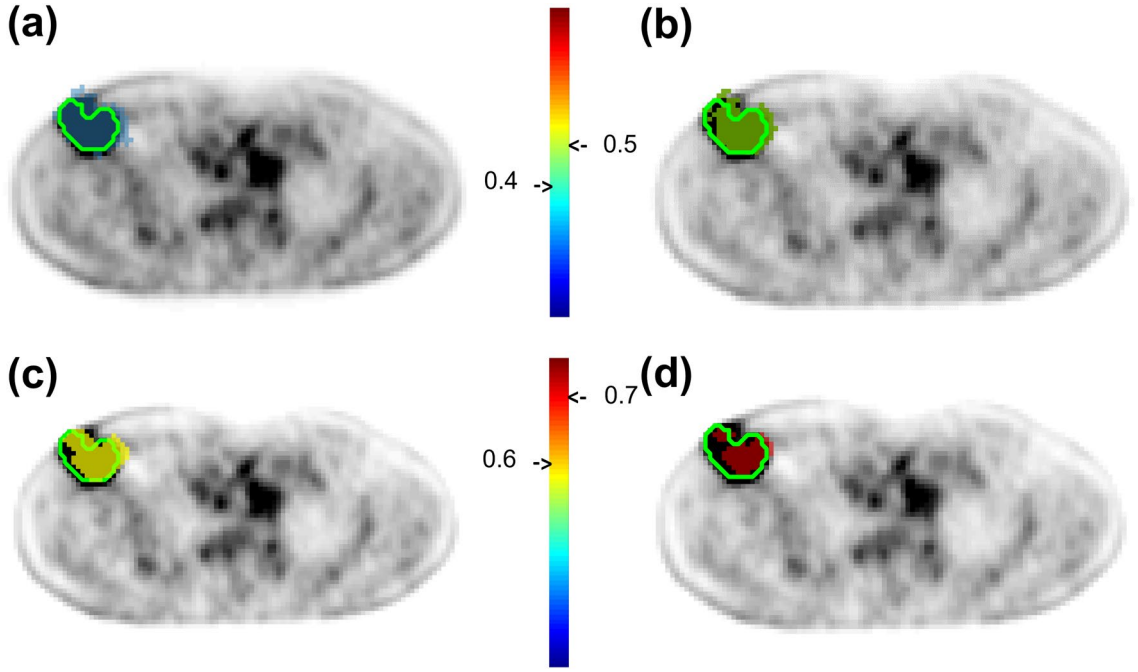

Fig. 5 Examples of the output of the semi-supervised (RFCM + FCM). Displayed on the axial view of slices extracted from the patient PET 3D volume in the dataset, we present outcome probabilities. The ground truth tumor area is represented by the green line, while various colors indicate different tumor probability regions. Four customized output examples are showcased, demonstrating the impact of different probability threshold settings, as indicated by the color bars on the right side of the images. In (**a**) only the areas containing pixels with tumor probabilities above 0.4 are displayed, above 0.5 in (**b**), above 0.6 in (**c**) and above 0.7 in (**d**)

αDice and RFCM + αFCM losses. We also observe that with other radiomics features, errors are less than < 10–20%, with

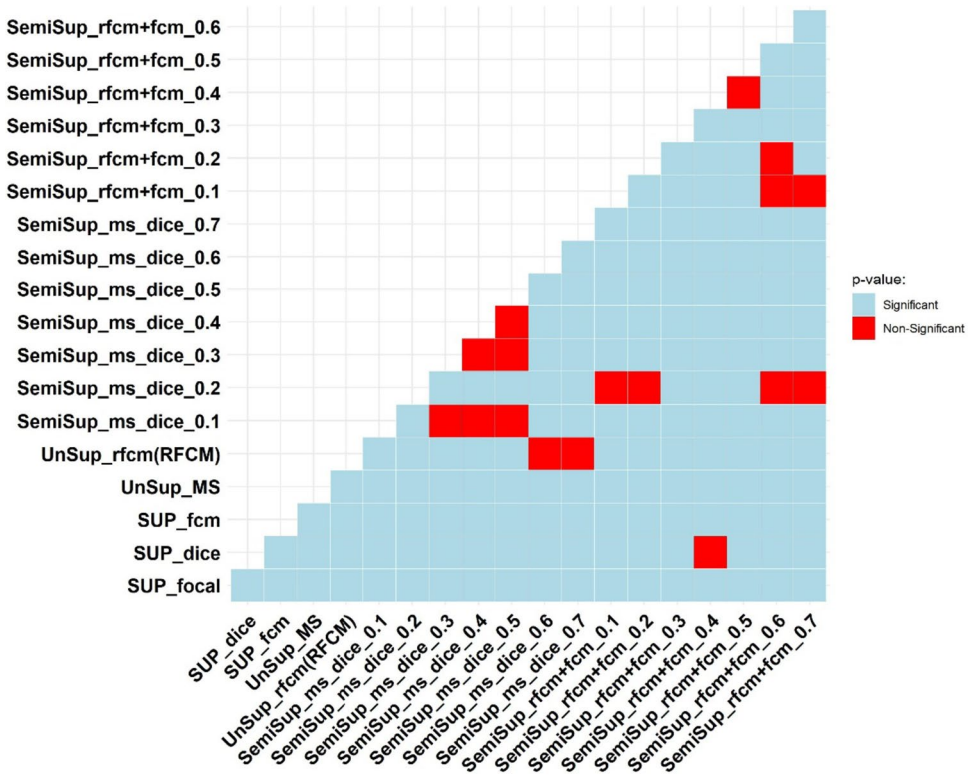

Fig. 6 Comparison of performance of different supervision levels and loss functions (p values) in terms of Dice coefficient (p value < 0.001 used as significant). For example we did not observe statistically significant differences between supervised approaches with Unified focal and FCM losses, semi-supervised (MS + αDice, α = 0.7) and unsupervised (MS), unsupervised (RFCM) and semi-supervised (MS + αDice, α = 0.7) approaches. The significance of the different performances of the semi-supervised approaches cannot be concluded amongst all α values

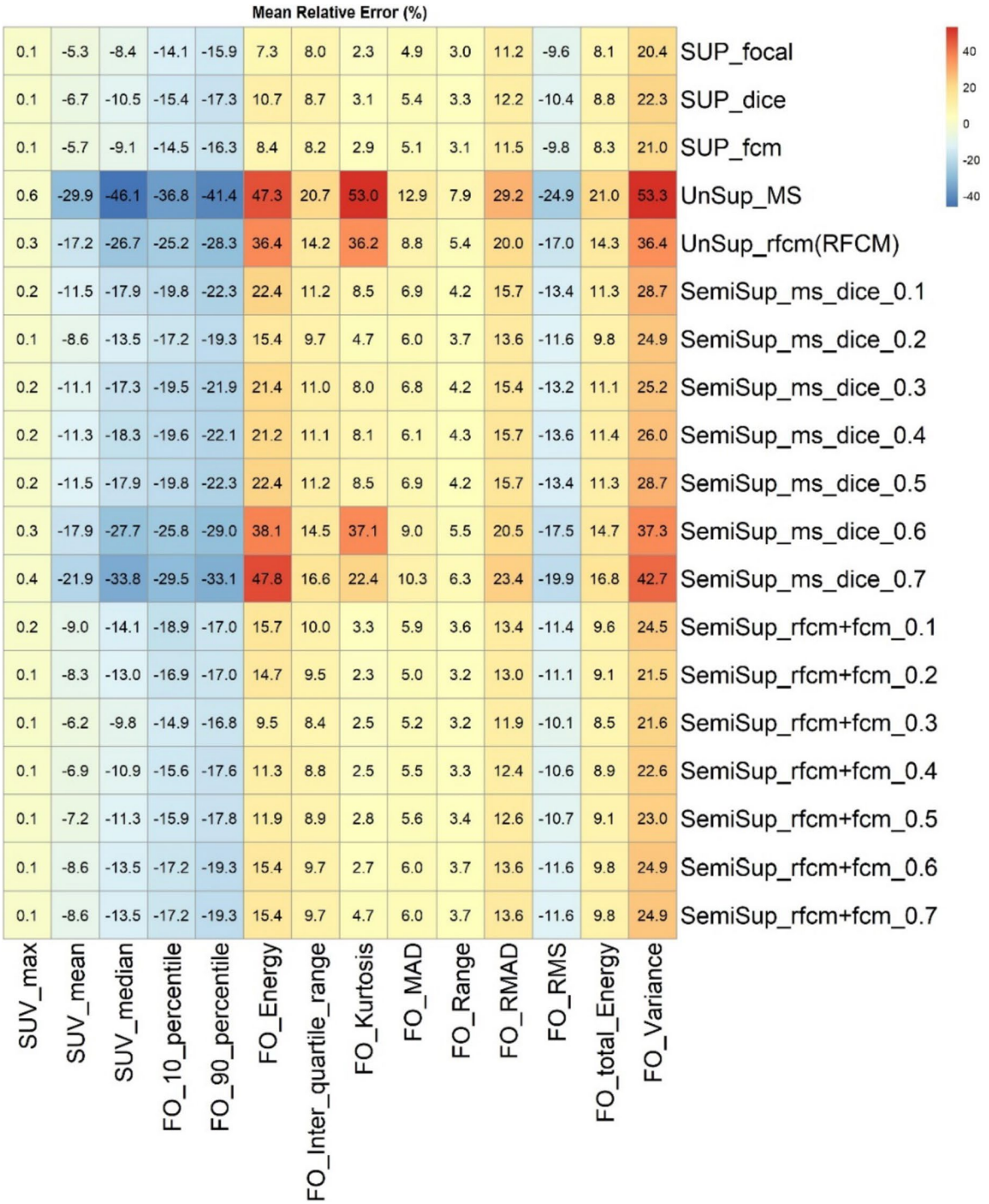

Fig. 7 Mean relative error percentage (MRE %) of radiomic features for the different levels of supervision and different loss functions

the exception of FO-variance. In supervised and FCM-based semi-supervised approaches compared to unsupervised and MS-based semi-supervised techniques, the mean relative errors of SUV-based and FO-based features are relatively lower.

The correlation (Spearman) between the ground truth TMTV and the predicted TMTV with the best semi-supervised approaches, i.e. MS + Dice with $\alpha = 0.2$ and RFCM + FCM with $\alpha = 0.3$, were calculated for $n = 60$ external test cases (both correlations = 0.96, $p < 0.0001$) and are shown in Fig. 8a and b. The volume errors (absolute difference

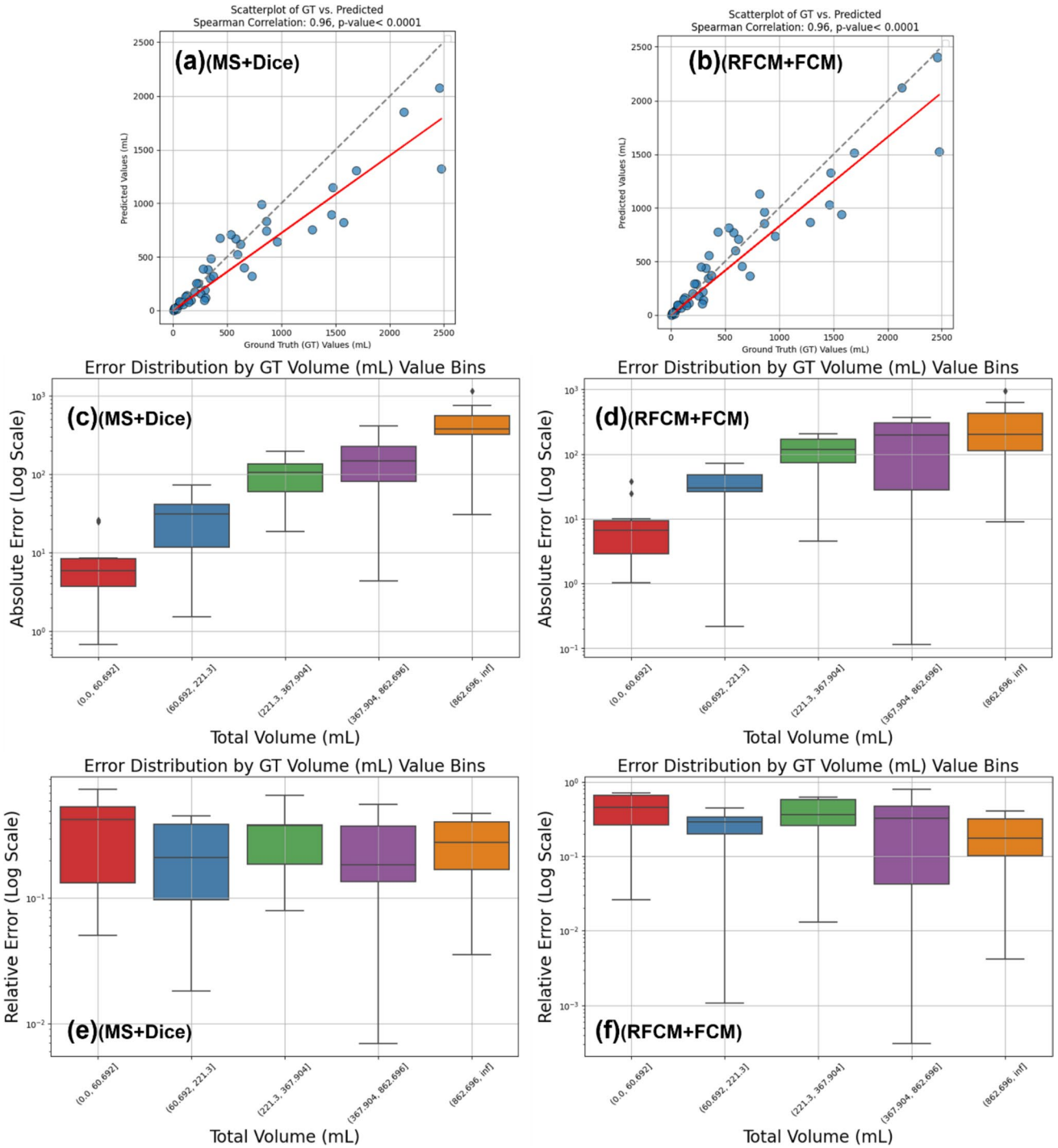

Fig. 8 The correlation between the ground truth total metabolic tumor volume (TMTV) and the predicted TMTV with the best semi-supervised approaches: **a** MS + Dice with $\alpha = 0.2$ and **b** RFCM + FCM with $\alpha = 0.3$. The volume errors (absolute difference between predicted and ground-truth TMTV) are shown in **c** and **d** respectively. The relative errors (absolute difference relative to ground-truth value) of two semi-supervised approaches are shown in **e** and **f**. (The n = 60 test cases)

between predicted and ground-truth TMTV) were calculated in 5 volume bins and their distribution are shown in (c) and (d). Additionally, the relative errors (absolute difference relative to ground-truth value) distribution in the same

5 volume bins were calculated and shown in (e) and (f), respectively. The correlation analysis and the absolute and relative errors distributions for the supervised and unsupervised approaches are shown in Figs. S1 and S2.

We considered the impact of α in $L^{\alpha}_{semi-FCM}(q=2,\beta=0.0016)$ and $L^{\alpha}_{semi-MS}\left(\eta=10^{-6}\right)$ on the performance of the lesion segmentation. As shown in Table 2, the performance of semi-supervised approach $L^{\alpha}_{semi-FCM}$ (RFCM + αFCM) was consistently higher than $L^{\alpha}_{semi-MS}$ (MS + αDice) for the different α values of (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7) that we extensively examined in Fig. 9.

Comparison to state-of-the-art techniques

We employed state-of-the-art approaches at various levels of supervision on our external test dataset. The training was also conducted on our training dataset, utilizing the same data splitting we used for our own approaches. Additionally, we leveraged the trained model shared by Blanch-Durand et al. [63]. The results of these comparisons are presented in Table 3. For comparison with unsupervised approaches, we applied the unsupervised method by Kim et al. [44], which utilizes the Mumford-Shah loss function, into Table 2. To align with state-of-the-art methods sharing a similar spirit, we have now included their suggested unsupervised approach with an MS + cross-entropy loss function (Table 3). Furthermore, for a holistic evaluation, we incorporated the FCM approach, a conventional unsupervised segmentation method, given that our approaches are inspired by FCM.

Discussion

Accurate tumor segmentation in PET images faces several challenges, encompassing issues such as limited spatial resolution, blurred boundaries due to the partial volume effect [64] leading to underestimated activity in small lesions

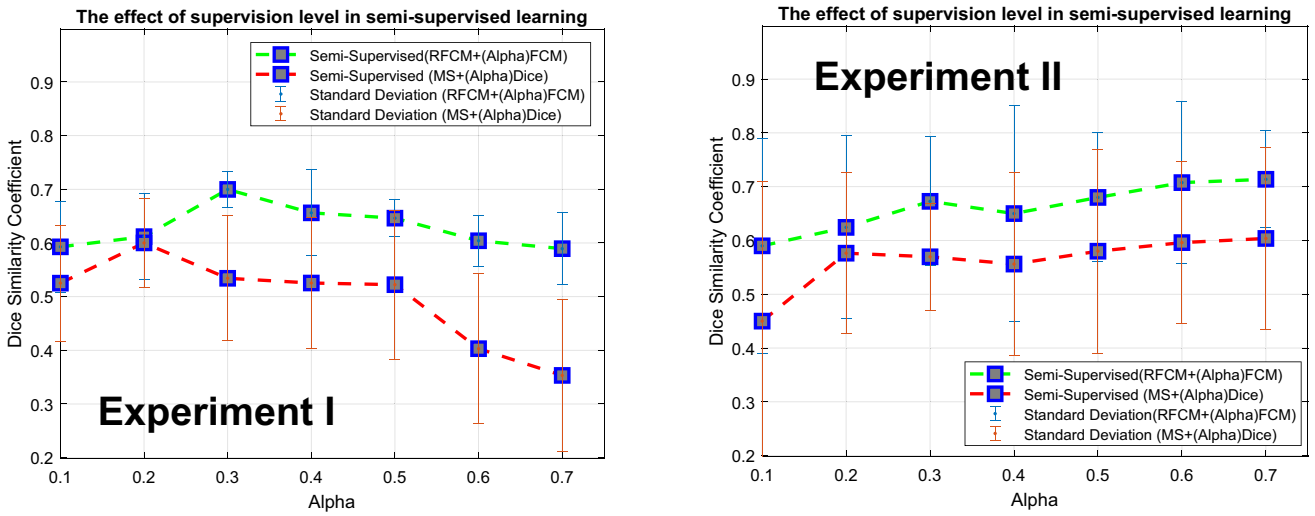

Fig. 9 The effect of supervision level by changing α on the segmentation performance (Dice score) in two semi-supervised approaches: RFCM + αFCM and MS + αDice for experiment I and II

Table 3 Comparison of the supervised approach with state-of-the-art (SOTA) approaches

Approach	Supervision level	SOTA approach	Dice score trained model	Relative error of TMTV	Dice score training from the scratch	Relative error of TMTV
AI	Supervised	Blanch-Durand et al. [63]	0.36 ± 0.14	0.78 ± 1.01	0.47 ± 0.20	0.56 ± 0.90
AI	Semi-supervised (MS + Cross-Entropy loss)	Kim et al. [44]	Trained model is Not shared	Trained model is Not shared	0.34 ± 0.18	0.82 ± 1.81
Conventional	Un-supervised	FCM	–	–	No training (0.24 ± 0.11)	0.61 ± 1.26

[30]. The diverse characteristics of lesions further complicate segmentation, especially in lymphoma patients with heavy disease burdens. Conventional voxel classification methods encounter difficulties with the fuzzy boundaries of tumors, prompting the exploration of fuzzy clustering techniques like FCM. However, their application is hindered by time-consuming processes and the need for user interaction. The development of an automated segmentation approach requires addressing ground truth inconsistencies arising from inter- and intra-observer variabilities and the lack of standardized segmentation approaches. Additionally, addressing the domain shift between the training dataset and real-world scenarios is crucial for achieving generalization in AI models.

In this study, we applied two semi-supervised approaches for tumor segmentation in PET scans. Semi-supervised learning approaches were implemented, integrating two loss functions designed for unsupervised learning based on the FCM cost function and Mumford-Shah formulation. In addition to their inherent noise-suppressing capabilities [65] and higher accuracy for tumor segmentation tasks [30], CNNs have much shorter prediction times than traditional segmentation methods like FCM. The FCM loss function is well-suited for training deep networks in tumor segmentation due to its ability to incorporate the classical FCM objective function, which allows for consideration of fuzzy edges without the need for supervision. Previous studies have demonstrated the effectiveness of the FCM loss in tumor segmentation, particularly in SPECT/CT images [45]. The fuzzy clustering loss function can serve as both a supervised and unsupervised loss function, depending on how the desired output is defined within the loss and with adaptive modifications made to the training process.

In Fig. 5, we present the probability map predictions generated by our network using a semi-supervised learning approach with Fuzzy C-means (FCM) losses (RFCM + αFCM). These predictions are overlaid on the axial slice of a PET scan. Increasing the probability percentage shrinks the predicted area from around to the inside of the tumor. The parameter α allows us to regulate the degree of supervision in semi-supervised approaches. For small value of α, training the network is focused on intensity distribution characteristics of the image rather than the ground truth labels of annotated training data. We showed that combined unsupervised RFCM and supervised FCM (RFCM + αFCM), performed better compared to integration of unsupervised MS loss and supervised Dice loss (MS + αDice). RFCM + αFCM with $\alpha = 0.3$ showed the best performance compared to the semi-supervised approach based on MS loss (p value < 0.01) with the percent relative error (RE%) of SUV_{max} quantification less than 1% (Fig. 7). Table 2 provides a comprehensive summary of unsupervised, semi-supervised, and supervised approaches, showcasing their performance in TMTV prediction through absolute and relative error metrics. Notably, the supervised techniques and the best performer of semi-supervised approaches in terms of DSC, exhibited minimal relative and absolute errors.

As Fig. 9 shows, in experiment II, training and test data are from DLBCL cases. Besides the training data was mainly composed of data from BCC center. By increasing α (from $\alpha = 0.1$ to $\alpha=0.7$) the impact of supervised loss was increased and the segmentation performance on test data (also from BCC) improved. In experiment I, the reduction in performance when we increase α (Fig. 9), can be explained by "domain shift". As the weight of the supervised term grows, more domain shift issues arise in the model; since the model was trained on data that were mostly from BCC center and the supervised learning does not generalize well on test data from SM. In other words, since most of the data used for training and testing are not drawn from the same distribution (center), increasing the weight of the supervised term decreases the segmentation performance. The performance drop in experiment I was higher in MS based semi-supervised approach (MS + αDice) compared to FCM based (RFCM + α FCM). Experiment I is close to the real-world scenario that we mainly apply trained models on unseen data from external centers. While, in the case of having test data from a center with limited contribution to training data, the segmentation performance was decreased due to domain shift phenomena.

Table 3 presents a comparison with state-of-the-art supervised approaches. The results indicate that domain shifts adversely impact the performance of the trained model, and starting training from scratch does not yield improved results compared to those reported in the original study but it was the best performer among state-of-the-art techniques (Table 3). We also compared our results with those obtained from unsupervised/semi-supervised methodologies utilizing Mumford-Shah loss, as well as conventional techniques such as FCM (Tables 2, 3). Achieving a fair comparison with conventional techniques necessitates the inclusion of pre/post-processing steps [66].

We also applied the trained supervised, semi-supervised and unsupervised models to the lymphoma cases of the benchmark dataset of autoPET which is publicly available [6–8] and the results are provided in the supplementary material (Table S1). The results showed the performance to drop (due to the domain shift) for models with different supervision levels. The best performer was supervised (unified focal loss), unsupervised (RFCM loss) that are similar to the results of our external test set. However, for the semi-supervised approaches, the best performers are with different supervision levels compared to our external test dataset: (MS + Dice) with $\alpha = 0.4$ and (RFCM + FCM) with $q = 2$ and $\alpha = 0.2$ in terms of both dice and relative

error of TMTV, but their differences were not significant in terms of TMTV relative error (p value > 0.01).The limitations in this study include the small size of the lymphoma lesions in two of our cohorts (DLBCL and PMBCL from BC Cancer) that constitute difficult cases for segmentation task. These cohorts include scans for limited stage (< III) and interim scans that mostly include small size lymphoma lesions. In addition, some of the labeled cases from the SM center were segmented by thresholding techniques (40%), and this could increase ground truth inconsistencies, and as such, we removed them from this study, and they are not included in the dataset in Table 1.

In summary, our investigation thoroughly explores the valuable prospect of harnessing unsupervised and semi-supervised approaches in the context of widely available but unlabeled PET data. We considered the potential of semi-supervised methods as a robust alternative when dealing with a scarcity of annotated data or ground truth inconsistencies. The study underscores the efficacy of a loss function within a semi-supervised framework, particularly in scenarios where both unsupervised and supervised components target the same category, such as region, boundary, or distribution. Notably, our findings demonstrate that the semi-supervised learning paradigm, specifically employing FCM loss (RFCM + αFCM), outperforms supervised approaches trained on a limited set of labeled data in terms of both Dice score and the relative error in TMTV prediction. This highlights the promising role of semi-supervised methods in addressing challenges associated with manual delineations, observer variability, and inconsistent ground truth annotations. In essence, our study illuminates the potential of semi-supervised approaches to revolutionize and streamline segmentation workflows in medical imaging, offering a more efficient and reliable avenue for lymphoma lesion characterization.

Conclusion

Given the wide availability of unlabeled PET data, it is possible to leverage the need for high-quality annotated data using unsupervised or semi-supervised approaches. To this end, we evaluated two semi-supervised approaches for 3D segmentation of lymphoma lesions. Our study showed that a semi-supervised approach with a well-designed loss function could be a great alternative when having access to only a limited amount of annotated data or when having ground truth inconsistencies. Specifically, a semi-supervised method that combines an unsupervised loss function with a supervised loss from the same category (region, boundary, or distribution) can achieve promising results. Compared to supervised approaches trained on a smaller amount of labeled data, semi-supervised learning via FCM loss (RFCM + αFCM) demonstrated improved performance in terms of Dice and TMTV prediction as well as a number of radiomics features (FO and shape features). We showed the level of supervision and the choice of loss function affect the accuracy of lesion segmentation and subsequent analysis of PET metrics and radiomics features. Semi-supervised methods hold great promise for automating segmentation workflows, addressing the challenges posed by time-consuming manual delineations performed by experts and the inherent variability among observers, which often lead to inconsistent ground truth annotations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13246-024-01408-x.

Author contributions Fereshteh Yousefirizi and Arman Rahmim contributed to the study conception, design, and preparation of the manuscript. They also revised the manuscript based on input from co-authors. Isaac Shiri and Habib Zaidi provided support in the evaluation design, implementation, and text editing. Joo Hyun O, Laurie H. Sehn, Kerry J. Savage, and Carlos F. Uribe contributed to the provision of patient data, clinical study design, and overall support. Joo Hyun O, Ingrid Blioise, Patrick Martineau, Don Wilson, Francois Benard, and Carlos Uribe assisted with manual delineations and text revisions.

Funding This research was supported by the Canadian Institutes of Health Research (CIHR) Project Grant PJT-173231, in part through computational resources and services provided by Microsoft for Health, and the Swiss National Science Foundation under Grant SNRF 320030_176052.

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 and Seoul St. Mary's Hospital.

Declarations

Competing interests Carlos Uribe and Arman Rahmim are co-founders of Ascinta Technologies Inc.

Ethical approval The retrospective study was conducted with the following ethics numbers: H19-01611 for the PMBCL study and H19-01866 for the DLBCL study at BC Cancer with the Ethics Committee and the Ethics Approval Number of UBC BC Cancer Research Ethics Board H19-001611, as well as KC11EISI0293 for the DLBCL study at Seoul St. Mary's Hospital. This study was performed in accordance with the approved guidelines of Seoul St. Mary's Hospital's institutional review board approved on 10 November 2020 (KC20RASI0867). Since these were the retrospective studies, the ethics committee of the hospital waived the requirement for obtaining informed consents from patients.

References

1. Hasani N, Paravastu SS, Farhadi F et al (2022) Artificial intelligence in lymphoma PET imaging: a scoping review (current trends and future directions). PET Clin 17:145–174

2. Cottereau A-S, Lanic H, Mareschal S et al (2016) 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 22:3801–3809
3. Kostakoglu L, Martelli M, Sehn LH, Belada D (2017) 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 130:824
4. Vercellino L, Cottereau A-S, Casasnovas O et al (2020) High total metabolic tumor volume at baseline predicts survival independent of response to therapy. Blood 135:1396–1405
5. Ceriani L, Martelli M, Zinzani PL et al (2015) Utility of baseline 18FDG-PET/CT functional parameters in defining prognosis of primary mediastinal (thymic) large B-cell lymphoma. Blood 126:950–956
6. Ceriani L, Milan L, Martelli M et al (2018) Metabolic heterogeneity on baseline 18FDG-PET/CT scan is a predictor of outcome in primary mediastinal B-cell lymphoma. Blood 132:179–186
7. Cottereau A-S, Versari A, Loft A et al (2018) Prognostic value of baseline metabolic tumor volume in early-stage Hodgkin lymphoma in the standard arm of the H10 trial. Blood 131:1456–1463
8. Mikhaeel NG, Smith D, Dunn JT et al (2016) 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 43:1209–1219
9. Song M-K, Yang D-H, Lee G-W et al (2016) 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 42:1–6
10. Sasanelli M, Meignan M, Haioun C et al (2014) Pretherapy metabolic tumour volume is an independent predictor of outcome in patients with diffuse large B-cell lymphoma. Eur J Nucl Med Mol Imaging 41:2017–2022
11. Toledano MN, Desbordes P, Banjar A et al (2018) 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 45:680–688
12. Chang C-C, Cho S-F, Chuang Y-W et al (2017) Prognostic significance of total metabolic tumor volume on 18F-fluorodeoxyglucose positron emission tomography/computed tomography in patients with diffuse large B-cell lymphoma receiving rituximab-containing chemotherapy. Oncotarget 8:99587–99600
13. Eude F, Toledano MN, Vera P et al (2021) Reproducibility of baseline tumour metabolic volume measurements in diffuse large B-cell lymphoma: is there a superior method? Metabolites 11:72. https://doi.org/10.3390/metabo11020072
14. Barrington SF, Zwezerijnen BGJC, de Vet HCW et al (2021) Automated segmentation of baseline metabolic total tumor burden in diffuse large B-cell lymphoma: which method is most successful? A study on behalf of the PETRA consortium. J Nucl Med 62:332–337
15. Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour models. Int J Comput Vis 1:321–331
16. Sanjay-Gopal S, Hebert TJ (1998) Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm. IEEE Trans Image Process 7:1014–1028
17. Bezdek JC, Ehrlich R, Full W (1984) FCM: The fuzzy c-means clustering algorithm. Comput Geosci 10:191–203
18. Cui R, Chen Z, Wu J et al (2021) A multiprocessing scheme for PET image pre-screening, noise reduction, segmentation and lesion partitioning. IEEE J Biomed Health Inform 25:1699–1711
19. Mumford D, Shah J (1989) Optimal approximations by piecewise smooth functions and associated variational problems. Commun Pure Appl Math 42:577–685
20. Vese LA, Chan TF (2002) A multiphase level set framework for image segmentation using the Mumford and Shah model. Int J Comput Vis 50:271–293
21. Liu S, Li J (2006) Automatic medical image segmentation using gradient and intensity combined level set method. Conf Proc IEEE Eng Med Biol Soc 2006:3118–3121
22. Zaidi H, El Naqa I (2010) PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur J Nucl Med Mol Imaging 37:2165–2187
23. Weisman AJ, Kieler MW, Perlman SB et al (2020) Convolutional neural networks for automated PET/CT detection of diseased lymph node burden in patients with lymphoma. Radiol Artif Intell 2:e200016
24. Blanc-Durand P, Van Der Gucht A, Schaefer N et al (2018) Automatic lesion detection and segmentation of 18F-FET PET in gliomas: a full 3D U-Net convolutional neural network study. PLoS ONE 13:e0195798
25. Yousefirizi F, Dubljevic N, Ahamed S et al (2022) Convolutional neural network with a hybrid loss function for fully automated segmentation of lymphoma lesions in FDG PET images. In: Medical imaging 2022: image processing. SPIE, pp 214–220
26. Yousefirizi F, Jha A, Ahamed S et al (2022) A novel loss function for improved deep learning-based segmentation: implications for TMTV computation. J Nucl Med 63:2588–2588
27. Coudray N, Moreira AL, Sakellaropoulos T, et al (2017) Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. medRxiv
28. Sun, Shrivastava, Singh (2017) Revisiting unreasonable effectiveness of data in deep learning era. Proc Estonian Acad Sci Biol Ecol
29. Willemink MJ, Koszek WA, Hardell C et al (2020) Preparing medical imaging data for machine learning. Radiology 295:4–15
30. Hatt M, Lee JA, Schmidtlein CR et al (2017) Classification and evaluation strategies of auto-segmentation approaches for PET: report of AAPM task group No. 211. Med Phys 44:e1–e42
31. Jha AK, Bradshaw TJ, Buvat I et al (2022) Nuclear medicine and artificial intelligence: best practices for evaluation (the RELAINCE guidelines). J Nucl Med. https://doi.org/10.2967/jnumed.121.263239
32. Bradshaw TJ, Boellaard R, Dutta J et al (2021) Nuclear medicine and artificial intelligence: best practices for algorithm development. J Nucl Med. https://doi.org/10.2967/jnumed.121.262567
33. Hatt M, Rest CC-L, van Baardwijk A et al (2011) Impact of tumor size and tracer uptake heterogeneity in 18F-FDG PET and CT non-small cell lung cancer tumor delineation. J Nucl Med 52:1690–1697
34. Cheplygina V, de Bruijne M, Pluim JPW (2019) Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal 54:280–296
35. Zhou Y, Wang Y, Tang P, et al (2019) Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In: 2019 IEEE winter conference on applications of computer vision (WACV). ieeexplore.ieee.org, pp 121–140
36. Afshari S, BenTaieb A, MiriKharaji Z, Hamarneh G (2019) Weakly supervised fully convolutional network for PET lesion segmentation. In: Medical imaging 2019: image processing. SPIE, pp 394–400
37. Hu Y, Modat M, Gibson E et al (2018) Weakly-supervised convolutional neural networks for multimodal image registration. Med Image Anal 49:1–13
38. Kamnitsas K, Baumgartner C, Ledig C et al (2017) Unsupervised domain adaptation in brain lesion segmentation with adversarial Networks. In: Information processing in medical imaging. Springer, New York, pp 597–609
39. Moriya T, Oda H, Mitarai M et al (2019) Unsupervised segmentation of micro-CT images of lung cancer specimen using deep

generative models. In: Medical image computing and computer assisted intervention—MICCAI 2019. Springer, New York, pp 240–248
40. Moriya T, Roth HR, Nakamura S, et al (2018) Unsupervised segmentation of 3D medical images based on clustering and deep representation learning. In: Medical imaging 2018: biomedical applications in molecular, structural, and functional imaging. SPIE, pp 483–489
41. Yousefirizi F, Jha AK, Brosch-Lenz J et al (2021) Toward high-throughput artificial intelligence-based segmentation in oncological PET imaging. PET Clin 16:577–596
42. Shi T, Jiang H, Wang M et al (2023) Metabolic anomaly appearance aware U-Net for automatic lymphoma segmentation in whole-body PET/CT scans. IEEE J Biomed Health Inform 1–12
43. Lian C, Li H, Vera P, Ruan S (2018) Unsupervised co-segmentation of tumor in PET-CT images using belief functions based fusion. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). ieeexplore.ieee.org, pp 220–223
44. Kim B, Ye JC (2020) Mumford-Shah loss functional for image segmentation with deep learning. IEEE Trans Image Process 29:1856–1866
45. Chen J, Li Y, Luna LP et al (2021) Learning fuzzy clustering for SPECT/CT segmentation via convolutional neural networks. Med Phys 48:3860–3877
46. Yousefirizi F, Bloise I, Martineau P, et al (2021) Reproducibility of a semi-automatic 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, New York
47. Çiçek Ö, Abdulkadir A, Lienkamp SS et al (2016) 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical image computing and computer-assisted intervention—MICCAI 2016. Springer, New York, pp 424–432
48. Iantsen A, Ferreira M, Lucia F et al (2021) Convolutional neural networks for PET functional volume fully automatic segmentation: development and validation in a multi-center setting. Eur J Nucl Med Mol Imaging 48:3444–3456
49. Pham DL (2001) Spatial models for fuzzy clustering. Comput Vis Image Underst 84:285–297
50. Ahmed MN, Yamany SM, Mohamed N et al (2002) A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data. IEEE Trans Med Imaging 21:193–199
51. Cai W, Chen S, Zhang D (2007) Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. Pattern Recognit 40:825–838
52. Chen S, Zhang D (2004) Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure. IEEE Trans Syst Man Cybern B Cybern 34:1907–1916
53. Chuang K-S, Tzeng H-L, Chen S et al (2006) Fuzzy c-means clustering with spatial information for image segmentation. Comput Med Imaging Graph 30:9–15
54. Wang X-Y, Bu J (2010) A fast and robust image segmentation using FCM with spatial information. Digit Signal Process 20:1173–1182
55. Geman S, Geman D (1984) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6:721–741
56. Yousefirizi F, KlyuzhinJooHyun ISO et al (2024) TMTV-Net: fully automated total metabolic tumor volume segmentation in lymphoma PET/CT images—a multi-center generalizability analysis. Eur J Nucl Med Mol Imaging. https://doi.org/10.1007/s00259-024-06616-x
57. Ma J, Chen J, Ng M et al (2021) Loss odyssey in medical image segmentation. Med Image Anal 71:102035
58. Lin TY, Goyal P, Girshick R, He K (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
59. Kervadec H, Bouchtiba J, Desrosiers C (2019) Boundary loss for highly unbalanced segmentation. on medical imaging
60. Yeung M, Sala E, Schönlieb C-B, Rundo L (2022) Unified Focal loss: generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Comput Med Imaging Graph 95:102026
61. Nioche C, Orlhac F, Boughdad S et al (2018) LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer Res 78:4786–4789
62. Gatidis S, Früh M, Fabritius M et al (2023) The autoPET challenge: towards fully automated lesion segmentation in oncologic PET/CT imaging
63. Blanc-Durand P, Jégou S, Kanoun S et al (2021) 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 48:1362–1370
64. Soret M, Bacharach SL, Buvat I (2007) Partial-volume effect in PET tumor imaging. J Nucl Med 48:932–945
65. Roy P, Ghosh S, Bhattacharya S, Pal U (2018) Effects of degradations on deep neural network architectures. arXiv:1807.10108
66. Yousefirizi F, Klyuzhin I, Girum K et al (2023) Federated testing of AI techniques: towards sharing of implementations, not just code. J Nucl Med 64:P1482–P1482
67. Clark K, Vendt B, Smith K et al (2013) The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26:1045–1057
68. Gatidis S, Hepp T, Früh M et al (2022) A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. Sci Data 9:601

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.