Repeatability of Two Artificial Intelligence Approaches for Tumor Segmentation in PET

Elisabeth Pfaehler (e.a.g.pfaehler@umcg.nl)
Universitair Medisch Centrum Groningen https://orcid.org/0000-0002-6160-3011

Liesbet Mesotten
Faculty of Medical and Life Science, Hasselt University

Gem Kramer
Department of Radiology and Nuclear Medicine, VUMC Amsterdam

Michiel Thomeer
Faculty of Medicine and Life Sciences, Hasselt University

Karolien Vanhove
Faculty of Medicine and Life Science, Hasselt University

Johan de Jong
Department of Nuclear Medicine and Molecular Imaging, University Medical Center Groningen

Peter Adriaensens
Department of respiratory medicine, Hasselt University

Otto S. Hoekstra
Department of Radiology and Nuclear Medicine, VUMC Amsterdam

Ronald Boellaard
Department of Radiology and Nuclear Medicine, VUMC Amsterdam

Original research

Keywords: repeatability, textural segmentation, convolutional neural network, tumor segmentation PET

DOI: https://doi.org/10.21203/rs.3.rs-54809/v1

Abstract

Background: Positron Emission Tomography (PET) is routinely used for cancer staging and treatment follow up. Metabolic active tumor volume (MATV) as well as total MATV (TMATV - including primary tumor, lymph nodes and metastasis) and/or total lesion glycolysis (TLG) derived from PET images have been identified as prognostic factor or for the evaluation of treatment efficacy in cancer patients. To this end a segmentation approach with high precision and repeatability is important. However, the implementation of a repeatable and accurate segmentation algorithm remains an ongoing challenge.

Methods: In this study, we compare two artificial intelligence (AI) based segmentation methods with conventional segmentation approaches in terms of repeatability. One segmentation approach is based on a textural feature (TF) segmentation approach designed for accurate and repeatable segmentation of primary tumors and metastasis. Moreover, a Convolutional Neural Network (CNN) is trained. The algorithms are trained, validated and tested using a lung cancer PET dataset. The segmentation accuracy of both segmentation approaches is compared using the jaccard coefficient (JC). Additionally, the approaches are applied on a fully independent test-retest dataset. The repeatability of the methods is compared with the repeatability of two majority vote (MV2, MV3) approaches, 41%SUV_{MAX}, and a SUV>4 segmentation (SUV4). Repeatability is assessed with test-retest coefficients (TRT%) and intraclass correlation coefficient (ICC). A TRT% of 0 indicates perfect repeatability and an ICC>0.9 was regarded as representing excellent repeatability.

Results: The accuracy of the segmentations with the reference segmentation was good (JC median TF: 0.7, CNN: 0.73) Both segmentation approaches outperformed together with the MV2 approach the other conventional segmentation methods in terms of test-retest coefficient (TRT% mean: TF: 13.0%, CNN: 13.9%, MV2: 14.1%, MV3: 28.1%, 41%SUV_{MAX}: 28.1%, SUV4: 18.1%) and ICC (TF: 0.98, MV2: 0.97, CNN: 0.99, MV3: 0.73, SUV4: 0.81, and 41%SUV_{MAX}: 0.68).

Conclusion: The AI based segmentation approaches used in this study provided better repeatability than conventional segmentation approaches. Moreover, both algorithms lead to accurate segmentations for both primary tumors as well as metastasis and are therefore good candidates for PET tumor segmentation.

Introduction

Positron Emission Tomography in combination with Computed Tomography (PET/CT) using the tracer fluorodeoxyglucose (FDG) is an important imaging modality for cancer diagnosis, tumor staging, prognosis or treatment follow-up [1, 2]. The volume of the segmented tumor in the PET image, also known as metabolic active tumor volume (MATV) as well as the total MATV (TMATV – including metastasis and lymph nodes), is one important metric for the evaluation of therapy response [3]. Observed differences in MATV/TMATV have to be due to biological changes in the tumor tissue and not to segmentation errors. Therefore, a repeatable segmentation is of outermost importance. Hereby, a repeatable segmentation refers to a segmentation algorithm leading to comparable results when applied on two consecutive PET/CT images of the same patient under the same physiological conditions. The implementation of a repeatable segmentation algorithm is not trivial due to the challenges coming with PET images. Among them are factors regarding the image quality, e.g. the low signal-to-noise ratio, low spatial resolution, and partial volume effects. Especially for smaller lesions, the partial volume effect can reduce the apparent tumor uptake making the lesion therefore difficult to detect and segment.

Up to now, a manual segmentation by an expert or (if available) the consensus segmentation of several experts are considered as gold standard. However, manual segmentations have several drawbacks, e.g. they are time-consuming, non-reproducible and come with a high inter-observer variability [4–6]. A recent study also demonstrated that even the consensus of several observers results in a low repeatability [7].

To overcome the limitations of manual segmentations and to increase repeatability, a large number of (semi-) automatic segmentation methods have been proposed. The most basic and frequently used ones are simple fixed thresholding algorithm defining voxels with an intensity value above a certain threshold as part of the tumor [8]. Also adaptive and iterative thresholding algorithm are available which are adapting the threshold according to the actual image characteristics [9]. However, all thresholding approaches are dependent on the scanner type, reconstruction algorithm, as well as image noise and have therefore limitations [10].

Therefore, more robust segmentation algorithm have been developed aiming to improve segmentation accuracy and repeatability. Developed methods include methods using the statistical properties of the image as well as learning-based methods [11, 12].Nevertheless, most of these approaches have only been tested on limited datasets and are not publically available. Therefore, the only (semi-) automated segmentation methods used in the clinic are thresholding approaches.

Due to the mentioned limitations of available segmentation algorithm, there is the need for new, more robust segmentation approaches. Artificial intelligence (AI) based segmentations such as Convolutional Neural Networks (CNN) have shown very promising results for various segmentation tasks [13] and yield great promise also for the segmentation of tumors in PET images. However, only a few studies use AI based segmentation approaches for metabolic active tumor segmentation in PET images. Even more, most studies combine the information of PET and CT images in order to get reliable segmentation results [14] or use some post-processing for an improvement of CNN segmentations [15]. Classifiers classifying each voxel as tumor or non-tumor using textural features of voxel neighborhoods have been used for the segmentation of e.g. lung carcinoma or head-and-neck cancer [16–18]. All of these studies combine the information of PET and CT images. In many cases the PET/CT is performed with a low-dose CT which is not of optimal image quality to be used for segmentation purposes. Therefore, it is of interest to develop AI based PET segmentation approaches that rely on PET information only. Additionally, in previous papers segmentation approaches were only applied on primary tumors, while for the calculation of TMATV, also an accurate and repeatable segmentation of metastasis and lymph nodes is important. This task is especially challenging due to the small size of metastasis, different tumor-to background ratios and different locations of the metastasis in the body.

While several studies already reported on the segmentation accuracy of AI based segmentation algorithm, to the best of our knowledge, no study reported yet on the repeatability of those algorithms. In this study, we investigate the repeatability of two AI approaches especially built to segment primary tumors and metastasis accurately and repeatable. We focus hereby on the segmentation task and do not consider lesion detection. This study includes a textural feature based segmentation approach, as well as a 3D CNN. All algorithm are trained, validated, and tested on a dataset of Non-Small-Cell-Lung-Cancer (NSCLC) patients. As second step, the algorithm are applied to a fully independent test-retest dataset of ten NSCLC patients scanned on two consecutive days. The repeatability of the AI segmentation approaches are compared with conventional segmentation algorithms used in the clinic.

Materials And Methods

Datasets

The study was registered at clinical trials.gov (NCT02024113) and was approved by the Medical Ethics Review Committee of the Amsterdam UMC and registered in the Dutch trial register (trialregister.nl, NTR3508). All patients gave informed consent for study participation and use of their data for (retrospective) scientific research. Two datasets acquired at two institutions were included in this study with both datasets following the recommendations of the EARL accreditation program [19, 20]. All images were converted to Standardized Uptake Value (SUV) units before the segmentation process started in order to normalize the images for differences in injected tracer dose and patient weight. The focus of this paper lies on the segmentation process and not on lesion detection. Therefore, before the start of the segmentation process, a large bounding box was drawn around every lesion including also a large number of non-tumor voxels as illustrated in Fig. 1. The bounding box was drawn randomly such that the tumor was not always appearing in the middle but on different locations in the box. This step was performed in order to avoid that the CNN remembers the location of the object instead of other, more important characteristics.

Training and testing dataset

For training, validating, and testing the segmentation approaches, 96 images of patients with NSCLC Stage III - IV were included. Patients fasted at least six hours before scan start and were scanned 60 minutes after tracer injection. All images were acquired on a Gemini TF Big Bore (Philips Healthcare, Cleveland, OH, USA). For attenuation correction, a low dose CT was performed. All images were reconstructed to a voxel size of 4 × 4 × 4 mm using the vendor provided BLOB-OS-TOF algorithm. More details about the patient cohort can be found in previous studies [21]. The images were split randomly in training, validating, and testing sets, where 56 images (286 lesion) were used for training, 14 images (98 lesions) for validation, and 26 images (171 lesions) for independent testing.

Test-Retest dataset

For a fully independent test-retest evaluation, ten PET/CT scans of patients with Stage III and IV NSCLC were analyzed. These ten patients underwent two whole-body PET/CT scans on two consecutive days. Images were acquired on a Gemini TF PET/CT scanner (Philips Healthcare, Cleveland, OH, USA) at a different institution (Amsterdam University Medical Center). Patient fasting time, time between tracer injection and scan start, as well as reconstruction algorithm and voxel size were the same as in the previous described dataset. A total of 28 lesions were included in the analysis.

Reference segmentations

The reference segmentations used for training, validating, and testing the algorithm, were obtained by applying an automatic segmentation which identified all voxels with a SUV above 2.5 as tumor (here after SUV2.5). The segmentations were manually adjusted by an expert medical physicist (RB) with more than twenty years of experience in PET tumor segmentation. This approach was chosen as it has been demonstrated that the manual adaption of a (semi-) automatic algorithm is more robust than a pure manual segmentation [22].

Segmentation Algorithm

All segmentation algorithm were implemented in Python 3.6 using the libraries keras and scikit-learn.

Convolutional Neural Network (CNN)

A 3D CNN following the U-Net architecture proposed by Ronneberger et al. [23] was implemented with the keras library. U-net is one of the most famous and most frequently used CNN architectures for biomedical image segmentation as it was especially designed for scenarios where only a small number of training examples are available. More details about the architecture and the used configuration can be found in the supplemental material.

In order to increase the amount of training data and to avoid over-fitting, data augmentation was performed. This included rotations within − 20 to 20 degrees, shifting in width and height direction within 20% of the side length, a rescaling of the images within 25%, intensity stretching, as well as adding Gaussian noise to the image.

For training, testing, and applying the CNN, the dataset was divided into smaller (< = 12.8 ml) and bigger tumors. The threshold was chosen by experiments, as this threshold let to the best performance. For each tumor size, one separate CNN was trained. The split of the dataset by lesion size was performed as this led to more accurate and repeatable segmentations (illustrated in supplemental material Sect. 2.1). For training, the tumor size was determined by calculating the volume of the ground truth mask. For testing and applying the CNN, an initial guess of the tumor size was performed using the majority vote (MV) segmentation of four established threshold approaches (see supplemental material, Sect. 3). The MV segmentation was chosen for this task as it resulted in previous work in the most accurate segmentation when compared with manual segmentations [7] and is easy to implement.

Textural feature segmentation (TF)

In this segmentation approach, textural features of voxel neighborhoods were used for the voxel-wise segmentation of the tumor. For every view (axial, sagittal, coronal) a separate segmentation was performed and the majority vote of the three views was regarded as final segmentation. The workflow of the TF segmentation for one view is illustrated in Fig. 2. As illustrated, every voxel was regarded as center of a scanning window. For each scanning window, statistical and textural features were calculated using the open-source software pyradiomics [24]. The feature space was then reduced by selecting the most important features for the segmentation task which were identified by a random forest.

Next, a random forest classifier was trained to classify each voxel as tumor or non-tumor. The trained random forest was then applied to the testing dataset. The probability images of the three orientations are combined in order to obtain the final classification. A probability image contains information about how certain the classifier is that it made the right decision. Hereby, all voxels with a summed probability of more than 1.8 were included in the tumor mask. A more detailed description of the algorithm can be found in the supplemental and in Pfaehler et *al*. [25].

In order to evaluate how well the AI based segmentations are matching the reference segmentation which was used for training, the segmentation results and the reference segmentation were compared in terms of accuracy.

Conventional segmentation algorithm

The repeatability of the AI based segmentations were compared with two established segmentation algorithm:

- 41%SUV_{MAX} : all voxels with intensity values higher than 41% of the maximal SUV value (SUV_{MAX}) are regarded as tumor
- SUV4: all voxels with a SUV higher than 4 are included in the segmentation

Moreover, two majority vote (MV) approaches combining four frequently used thresholding approaches were included in the comparison. Both MV approaches have been demonstrated in previous work to be more repeatable than conventional approaches. The underlying segmentation algorithm are explained in the supplemental Sect. 3 and are also described in previous work [7]. The two MV segmentation methods include:

- MV2: the consensus of at least two of the approaches
- MV3: the consensus of at least three of the approaches

Evaluation Of Segmentation Algorithm

For the evaluation of the segmentation algorithm, several metrics were combined. The data analysis was performed in Python 3.6.2 using the packages numpy and scipy.

Accordance of AI segmentation and reference segmentation

In order to determine the accordance of AI and reference segmentation, the Jaccard Coefficient (JC) was calculated. The JC is defined as the ratio between the intersection and the union of two labels and gives an indication about the overlap of the two labels:

$$\mathrm{JC} = \frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{A} \cup \mathrm{B}}$$

A JC of 1 indicates perfect overlap, while a JC of 0 indicates that there is no overlap at all.

Furthermore, as the JC does not contain information about volume differences, the percentage MATV differences of performed and reference segmentation were calculated: $\frac{MATV_{SEGM}}{MATV_{REF}}$. A percentage volume difference above 1 indicates an over- and a percentage volume difference below 1 an under-estimation. A percentage difference of 1 represents a perfect alignment. Finally, the distance of mass (barycenter distance) of the segmentations was calculated. Hereby, a barycenter distance close to 0 indicates perfect agreement.

Repeatability evaluation

The repeatability of the segmentation approaches was evaluated by comparing the differences of segmented volume across days. For this purpose, the percentage Test-Retest difference (%TRT) was calculated:

$$TRT\% = \frac{|vol_{Day1} - vol_{Day2}|}{(vol_{Day1} + vol_{Day2})/2} * 100$$

The %TRT gives a measure for the proportional differences in segmented volume between the two consecutive scans. Moreover, the repeatability coefficient (RC) which is defined as 1.96 × standard deviation(TRT%) was calculated. Additionally, intraclass correlation coefficients (ICC) were calculated using a two-way mixed model with single measures checking for agreement. An ICC between 0.9 and 1 indicates excellent and an ICC between 0.75 and 0.9 indicates good repeatability [26]. If a lesion was completely missed by one segmentation approach, it was discarded from the analysis in order to analyze the same dataset for all segmentation approaches.

The accuracy metrics of the AI based segmentations as well as the TRT% of all approaches were compared using the Friedman test. The Friedman test is a non-parametric test which does not assume a normal distribution of the data or independency of observations. It compares the rank of each data point instead of only comparing mean or median values. This means that if a segmentation algorithm results consistently in more accurate results, it will be ranked higher even though its mean or median might be lower. As the Friedman test only contains information if there was a significant difference in the data, a Nemenyi test was performed in order to assess which methods resulted in significant differences. P-values below 0.01 were considered as statistically significant. A Benjamini-Hochberg correction was applied in order to correct for multiple comparisons.

Results

Accordance reference – AI based segmentation

Figure 3 displays the JC values between AI based and reference segmentation for the testing and test-retest dataset. In both datasets, both approaches resulted in similar accuracies which were not significantly different ($p > 0.01$). In the testing dataset, both approaches yielded good JC values (TF: median: 0.7, 25th percentile: 0.59, 75th percentile: 0.79, CNN: median: 0.73, 25th percentile: 0.47, 75th percentile: 0.82) indicating a good accordance with the reference segmentations. Volume differences and barycenter distances are listed in Table 1. The CNN yields less underestimations and more overestimations of tumor volume (higher percentage volume differences (25th /75th percentile: 0.83/1.34)). While the TF approach resulted in more underestimations of tumor volume (25th /75th percentile: 0.59/0.83). The barycentric distances of the TF approach were lower than the barycentric distances of the CNN. The corresponding values for the test-retest dataset can be found in Supplemental Table S1.

Table 1
Abbreviations as used in the text

Abbreviations	
PET	Positron Emission Tomography
MATV	Metabolic Active Tumor Volume
TMATV	Total MATV
TLG	Total Lesion Glycolysis
TF	Textural Feature
CNN	Convolutional Neural Network
JC	Jaccard Coefficient
TRT	Test-retest coefficient
ICC	Intraclass correlation coefficient
FDG	fluorodeoxyglucose
CT	Computed Tomography
AI	Artificial Intelligence
NSCLC	Non-Small Cell Lung Cancer
SUV	Standardized Uptake Value
MV	Majority Vote

In general, the accuracy of the segmentations depended on the lesion size as illustrated in Fig. 4. Segmentations of bigger tumors resulted in better accuracy than segmentations of smaller lesions. For larger lesions, the CNN resulted in a median JC value of 0.79, while the TF approach yielded a median JC of 0.86. For both approaches, the percentage volume differences were close to 1. Here, the TF approach resulted in lower percentage volume differences than the CNN and therefore more underestimations.

For smaller lesions, the CNN resulted in a median value of 0.69 which was higher than the median of the TF approach (0.66). The median percentage volume differences of the CNN was 1.02 (25th /75th percentile: 0.81/1.40) indicating that the CNN resulted more often in overestimations for smaller than for larger lesions. While the TF approach yielded in the majority of the cases percentage volume differences below 0.7 and therefore also for smaller lesions more underestimations. Quartile values as well as corresponding percentage volume differences and barycentric distances for smaller and bigger lesions are listed in Table 2.

Table 2
Percentage volume differences and barycentric distances for TF and CNN

	Percentage volume difference **median (25th /75th quartile)**	**Barycentric distance** **median (25th /75th quartile)**
TF	0.70 (0.59/0.79)	0.37 (0.24, 0.66)
CNN	0.99 (0.83/1.34)	0.50 (0.21, 1.23)

Table 3
Accuracy metrics for smaller and bigger lesions

	JC bigger **median (25th /75th quar)**	**Perc vol. diff bigger** **median (25th /75th quar)**	**Barycentric distance** **median (25th /75th quartile)**	**JC** **median (25th /75th quar)maller**	**Perc vol. diff smaller** **med (25th /75th quar)**	**Barycentric distance** **median (25th /75th quartile)**
TF	0.86 (0.77/0.89)	0.86 (0.77/0.89)	0.37 (0.24/0.92)	0.66 (0.55/0.75)	0.67 (0.55/0.76)	0.38 (0.23/0.63)
CNN	0.79 (0.72/0.87)	0.97 (0.87/1.1)	0.58 (0.24/1.32)	0.69 (0.43/0.81)	1.02 (0.81/1.40)	0.48 (0.19/1.14)

As displayed in Fig. 4, TF and CNN resulted in three cases in JC values around/below 0.4 for bigger lesions. In both cases, the tumors were located close to the heart which was incorrectly included in the segmentation. Therefore, the tumor was highly overestimated. A similar effect was observed for smaller lesions: The CNN missed some of the smaller lesions completely while this was not the case for the TF based approach. All lesions that were completely missed were located close to the kidney which was wrongly identified as tumor. The TF approach also identified the kidney regions as tumors but also detected the tumors.

Repeatability

Figure 5 displays the TRT-coefficients for all segmentation algorithm. Two lesions were completely missed by the CNN and therefore discarded from the analysis.

CNN-based segmentations outperformed the other approaches regarding TRT% with an absolute mean value of 13.9% and a standard deviation of 16%. TF and MV2 segmentation yielded with absolute mean values of 13.0% and 14.1% and standard deviations of 17% and 21% similar values to the CNN. MV3, 41%SUVMAX, and SUV4 segmentations yielded mean values of 28.1%, 28.1%, and 18.1%, and standard deviations of 50%, 51%, and 26%. The corresponding repeatability coefficients can be found in supplemental table S2. After applying the Benjamini-Hochberg correction, the differences in TRT% were not significantly different.

The CNN resulted in 3 out of 28 cases in a TRT% of more than 10%, while the conventional methods resulted in 12 (MV2, SUV4, 41%SUV_{MAX}), or 13 cases in a TRT% higher than 10% (MV3). The TF segmentation resulted in 8 cases in a TRT% of more than 10%.

TF, CNN, and MV2 approach yielded similar ICC values (TF: 0.98, MV2: 0.97, CNN: 0.99) indicating a very good repeatability. MV3, SUV4, and 41%SUVMAX resulted in ICC of 0.73, 0.81, and 0.68, respectively. The lesion size did not influence the repeatability of the segmentations.

Summary of the results

In summary, CNN and TF segmentation resulted in a better repeatability when compared with conventional approaches. Furthermore, both approaches resulted in a good accuracy when compared with the reference segmentations. The observed differences between the AI based methods were neither for accuracy nor for repeatability significant. Therefore, our results suggest that both methods are equally good candidates for the segmentation of tumors in PET images and are more powerful than conventional approaches in terms of repeatability.

Discussion

In this paper, we evaluated two AI based segmentation approaches in terms of repeatability and analyzed their accordance with the reference segmentation. Both approaches resulted in a good accuracy when compared with the reference segmentation. The differences in performance between both AI approaches were small and statistically non-significant.

The segmentation of smaller lesions remains also for these two approaches a challenging task. One reason for this effect might be that with decreasing tumor size, small misclassifications have a higher impact on accuracy metrics as illustrated in supplemental table S3. Smaller lesions also come with a lower tumor-to-background ratio and are therefore more difficult to detect what might be the reason that the CNN missed some smaller lesions completely. Moreover, some of the metastasis are also located close to other high-uptake regions (such as the kidney) what opposes a special challenge to a segmentation algorithm. Especially for the CNN, the different locations of the metastasis and therefore the differences in background tissue yield a more challenging learning task than the segmentation of one type of primary tumor.

In terms of accuracy and precision, the CNN trained and tested in this study was comparable with previous CNNs designed for the segmentation of primary tumors in PET images. An important difference between our methods and other published algorithm is that our approaches rely on the PET image information only and can therefore also be used when only a low-dose CT is acquired aside of the PET image [14, 16]. Previous studies reported on low segmentation performance when only using the PET image for segmentation.

When the tumor was located close to another high uptake region such as the heart or the kidney, both segmentation approaches regarded also the high-uptake region as tumor. The automatic segmentation methods included in this study are mainly intensity driven and are therefore also not capable of distinguishing between one and another high-uptake region when they are close to each other. For these cases it is likely that a human interaction will always remain necessary as was mentioned in previous studies [27]. However, in future studies we will investigate if these segmentation approaches might also be used for lesion detection.

Also when compared with previous studies, CNN, and TF approach outperformed other (semi-) automatic segmentation methods. Frings et *al*. reported a TRT% repeatability coefficient of 44.4–71.1 for all lesions included in their analysis when using different threshold-based segmentation approaches with background correction [28]. The AI based segmentation methods yielded with repeatability coefficients of 31.36 (CNN) and 33.36 (TF) lower repeatability coefficients. For images acquired under the same conditions as in this study (i.e. 60 minutes time between tracer injection and scan start and EARL-compliant reconstructions), Kolinger et *al*. found with repeatability coefficients between 43–56 higher repeatability coefficients than the ones of the AI based segmentations [7]. However, Kolinger et *al*. also reported lower repeatability coefficients for MV3 and 41%SUV_{MAX} segmentation approaches. The reason for this might be that Kolinger et *al*. compared the repeatability for MATV of all lesions (TMATV), while we compared the repeatability of MATV lesion by lesion. A discrepancy in the segmentation of one lesion, especially if the lesion is small, might have less impact on the overall repeatability when including all lesions.

A disadvantage of AI based segmentation approaches is the need of reliable training data. The lack of reasonable training data is one drawback making the clinical implementation of AI based segmentation algorithm challenging. However, the MV2 approach used in this study was found to result in accurate and robust segmentations in a previous study [7]. Moreover, in our study it also outperformed the conventional segmentation approaches in terms of repeatability without depending on training data. Especially for tasks where segmentation accuracy is important such as radiotherapy planning, the MV2 is a good candidate for clinical use. Yet, regardless method used, the final segmentation should always be supervised. In terms of repeatability, especially the CNN segmentation outperformed the MV2 approach and is the method of choice when segmentation repeatability is important such as for the evaluation of treatment response, i.e. precision may be more important that accuracy for those clinical applications.

One limitation of this study is that the ground truth segmentations were delineated by one, yet experienced, observer while the consensus of three expert segmentations is considered as gold standard. To account for this, the segmentation was initiated with a semi-automated delineation method, an approach known to reduce observer variability. Of note, for the test-retest study the same lesions were delineated by 5 observers in a previous study (7) and it was shown that even the consensus contour of these observers was less repeatable than those seen with any of the automated approaches. Finally, in our repeatability study we included the AI based approaches as well as several conventional methods and this repeatability study showed that our trained AI approaches provided very good results, even if the ground truth segmentations used during training of the AI methods would have been suboptimal.

Another limitation is the small dataset used for repeatability analysis. However, the collection of test-retest scans is unfortunately limited due to the patient burden coming with consecutive scans of the same patients. Future studies, especially studies using data from different centers should confirm our findings.

Conclusion

In this paper, we compared the repeatability of AI based segmentation algorithm with conventional segmentation approaches. Our results illustrate the advantage of AI based segmentation approaches: Both approaches resulted in a good accuracy when compared with the reference segmentation and a high repeatability. Together with a majority vote approach (combining the results of four conventional segmentation approaches) the proposed segmentation methods were superior to the other segmentation algorithms included in this study in terms of repeatability. This study demonstrates that AI based segmentations have not only the potential to accurately segment lesions but also result in more repeatable segmentations.

Ethics approval and consent to participate:

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.

Consent for publication:

Not applicable

Availability of data and materials:

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests:

The authors declare that they have no competing interests

Financial support

This work is part of the research program STRaTeGy with project number 14929, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). This study was financed by the Dutch Cancer Society, POINTING project, grant 10034.

Authors' contributions:

EP implemented the segmentation methods, analyzed the data, and wrote the manuscript. LM, GK, MT, KV contributed to the patient inclusion, data acquisition, and manuscript revision. JdJ contributed to manuscript revision. OSH contributed to the patient inclusion, data acquisition, and manuscript revision. RB designed and managed the study, developed software for image processing, and wrote the manuscript.

Acknowledgments:

We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

References

1. Volpi S, Ali JM, Tasker A, Peryt A, Aresu G, Coonar AS. The role of positron emission tomography in the diagnosis, staging and response assessment of non-small cell lung cancer. Ann Transl Med. 2018;6:95–5.
2. Griffeth LK. Use of PET/CT scanning in cancer patients: technical and practical considerations. Proc. (Bayl. Univ. Med. Cent). 2005;18:321–30.
3. Hammerschmidt S, Wirtz H. Lung Cancer. Dtsch. Aerzteblatt Online. 2009.
4. Vorwerk H, Beckmann G, Bremer M, Degen M, Dietl B, Fietkau R, et al. The delineation of target volumes for radiotherapy of lung cancer patients. Radiother Oncol. 2009;91:455–60.
5. Johansson J, Alakurtti K, Joutsa J, Tohka J, Ruotsalainen U, Rinne JO. Comparison of manual and automatic techniques for substriatal segmentation in 11C-raclopride high-resolution PET studies. Nucl Med Commun. 2016;37:1074–87.
6. Hatt M, Lee JA, Schmidtlein CR, Naqa I, El, Caldwell C, De Bernardi E, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. Med Phys. 2017;44:e1–42.
7. Kolinger GD, Vállez García D, Kramer GM, Frings V, Smit EF, de Langen AJ, et al. Repeatability of [18F]FDG PET/CT total metabolic active tumour volume and total tumour burden in NSCLC patients. EJNMMI Res. 2019;9:14.
8. Schinagl DAX, Vogel WV, Hoffmann AL, van Dalen JA, Oyen WJ, Kaanders JHAM. Comparison of Five Segmentation Tools for 18F-Fluoro-Deoxy-Glucose–Positron Emission Tomography–Based

Target Volume Definition in Head and Neck Cancer. Int J Radiat Oncol. 2007;69:1282–9.

9. Jentzen W, Freudenberg L, Eising EG, Heinze M, Brandau W, Bockisch A. Segmentation of PET volumes by iterative image thresholding. J Nucl Med. 2007;48:108–14.
10. Nestle U, Kremp S, Schaefer-Schuler A, Sebastian-Welsch C, Hellwig D, Rübe C, et al. Comparison of different methods for delineation of ^{18}F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-Small cell lung cancer. J Nucl Med. 2005;46:1342–8.
11. Halt M, Le Rest CC, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. IEEE Trans Med Imaging. 2009;28:881–93.
12. Foster B, Bagci U, Mansoor A, Xu Z, Mollura DJ. A review on segmentation of positron emission tomography images. Comput Biol Med Elsevier. 2014;50:76–96.
13. Zhang Y, Oikonomou A, Wong A, Haider MA, Khalvati F. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. Sci Rep Nature Publishing Group. 2017;7:46349.
14. Zhong Z, Kim Y, Zhou L, Plichta K, Allen B, Buatti J, et al. 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. Proc. - Int. Symp. Biomed. Imaging. IEEE; 2018;2018-April:228–31.
15. Blanc-Durand P, Van Der Gucht A, Schaefer N, Itti E, Prior JO. Automatic lesion detection and segmentation of18F-FET PET in gliomas: A full 3D U-Net convolutional neural network study. PLoS One. 2018;13:1–11.
16. Yu H, Caldwell C, Mah K, Mozeg D. Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning. IEEE Trans Med Imaging. 2009;28:374–83.
17. Yu H, Caldwell C, Mah K, Poon I, Balogh J, MacKenzie R, et al. Automated Radiation Targeting in Head-and-Neck Cancer Using Region-Based Texture Analysis of PET and CT Images. Int J Radiat Oncol Biol Phys. 2009;75:618–25.
18. Markel D, Caldwell C, Alasti H, Soliman H, Ung Y, Lee J, et al. Automatic Segmentation of Lung Carcinoma Using 3D Texture Features in 18-FDG PET/CT. Int J Mol Imaging. 2013;2013:1–13.
19. Aide N, Lasnon C, Veit-Haibach P, Sera T, Sattler B, Boellaard R. EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. Eur J Nucl Med Mol Imaging. 2017;44:17–31.
20. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging. 2015;42:328–54.
21. Vanhove K, Mesotten L, Heylen M, Derwael R, Louis E, Adriaensens P, et al. Prognostic value of total lesion glycolysis and metabolic active tumor volume in non-small cell lung cancer. Cancer Treat Res Commun. 2018;15:7–12.
22. van Baardwijk A, Bosmans G, Boersma L, Buijsen J, Wanders S, Hochstenbag M, et al. PET-CT-Based Auto-Contouring in Non-Small-Cell Lung Cancer Correlates With Pathology and Reduces Interobserver

Variability in the Delineation of the Primary Tumor and Involved Nodal Volumes. Int J Radiat Oncol Biol Phys. 2007;68:771–8.

23. Convolutional Networks for Biomedical Image Segmentation
U-Net
Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.
24. Griethuysen JJM Van, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. 2017;77:104–8.
25. Pfaehler E, Mesotten L, Kramer G, Thomeer M, Vanhove K, de Jong J, et al. Textural Feature Based Segmentation: A Repeatable and Accurate Segmentation Approach for Tumors in PET Images. 2020. p. 3–14.
26. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15:155–63.
27. Hatt M, Laurent B, Ouahabi A, Fayad H, Tan S, Li L, et al. The first MICCAI challenge on PET tumor segmentation. Med Image Anal Elsevier BV. 2018;44:177–95.
28. Frings V, de Langen AJ, Smit EF, van Velden FHP, Hoekstra OS, van Tinteren H, et al. Repeatability of Metabolically Active Volume Measurements with 18F-FDG and 18F-FLT PET in Non-Small Cell Lung Cancer. J Nucl Med. 2010;51:1870–7.
29. **b)a)Figures**.

Figures

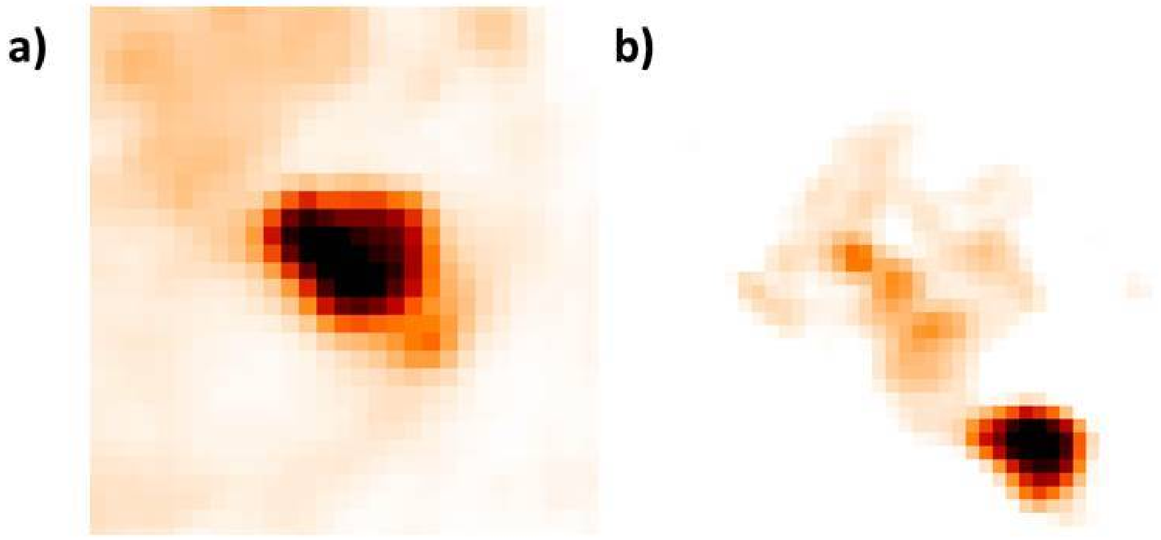

Figure 1

Two examples of a bounding box: A large bounding box is drawn around each lesion so that it also includes a large amount of background. For each lesion, the lesion is placed in a different position in the bounding box such that the CNN is not learning mainly the position of the lesion in the bounding box

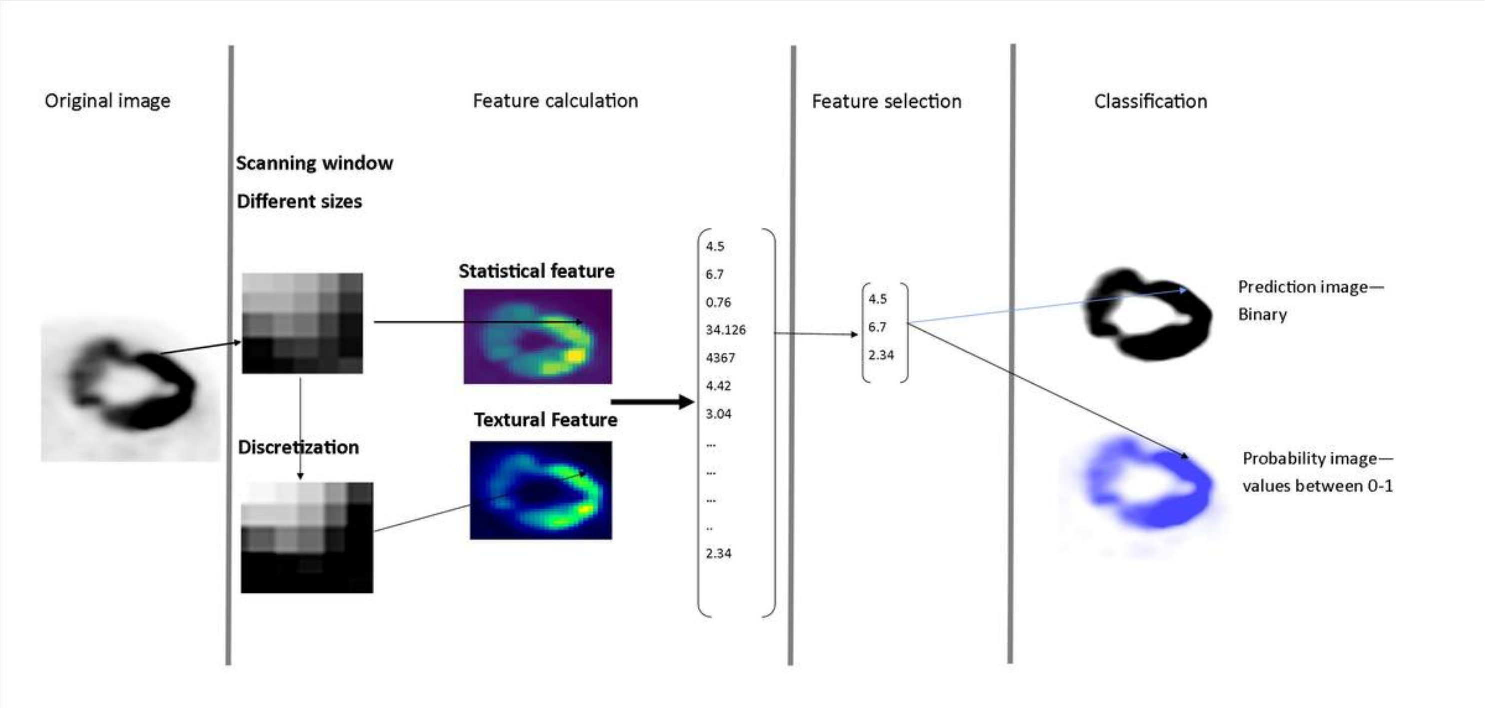

Figure 2

Workflow of the Textural feature based segmentation for the axial view

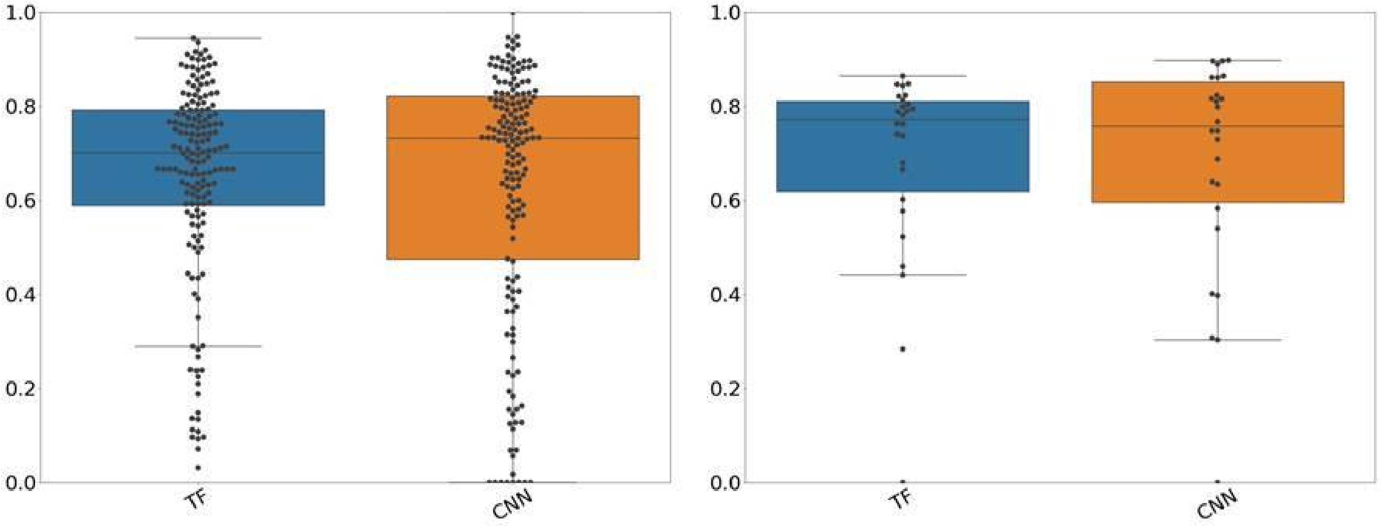

Figure 3

Jaccard Coefficient (JC) values for both datasets: JC values for the testing set (left figure) and the test-retest dataset (right figure) for the AI based segmentation algorithm included in the study

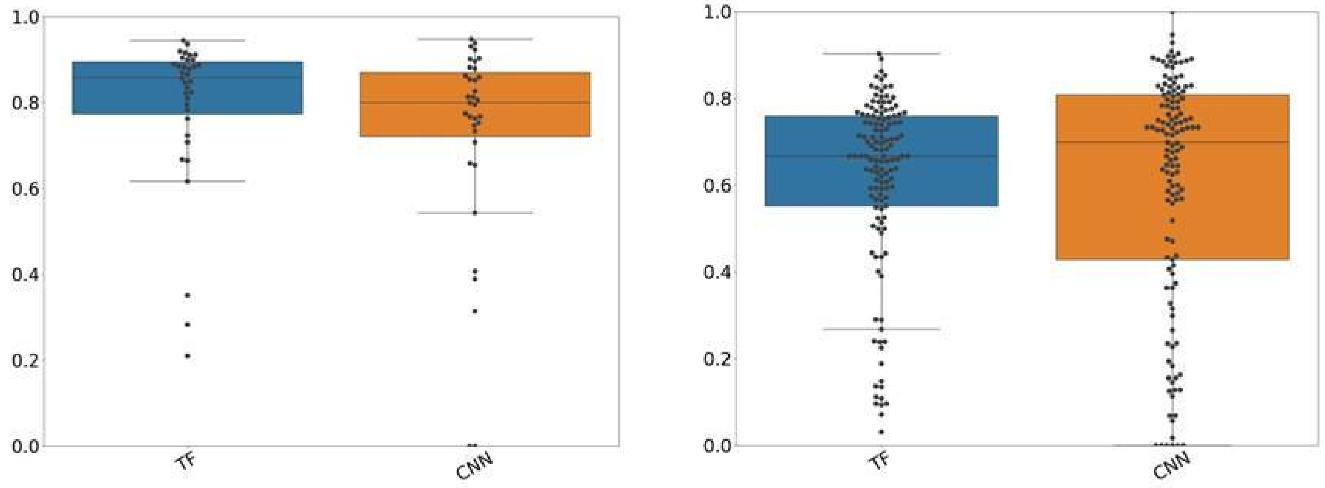

Figure 4

Jaccard Coefficient (JC) values dependent on lesion size: JC values for bigger (left figure) and smaller (right figure) lesions for both AI based segmentation approaches

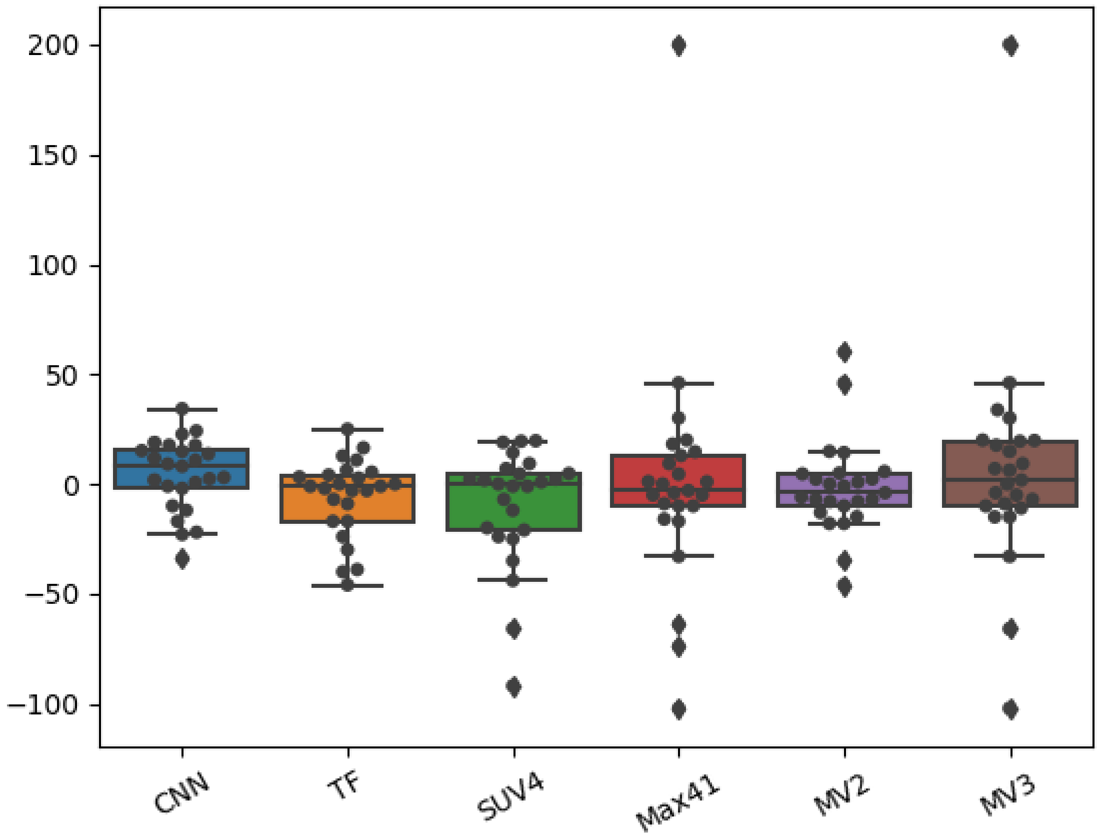

Figure 5

Test-Retest Coefficient (TRT%) for all segmentation approaches: If the TRT% is close to 0, the repeatability of the segmentations is excellent. Abbreviations of the segmentation algorithm: SUV4: Standardized Uptake Value 4, 41%SUVMAX, MV2: Majority Vote 2, MV3: Majority Vote 3, TF: Textural Feature based approach, CNN: Convolutional Neural Network

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementalv2.docx