




Textural Feature Based Segmentation: A Repeatable and Accurate Segmentation Approach for Tumors in PET Images

Elisabeth Pfaehler¹ , Liesbet Mesotten^{2,3}, Gem Kramer⁴,
Michiel Thomeer^{2,5}, Karolien Vanhove^{2,6}, Johan de Jong¹,
Peter Adriaensens⁷, Otto S. Hoekstra⁴, and Ronald Boellaard^{1,4}

- ¹ Department of Nuclear Medicine and Molecular Imaging, Medical Imaging Center, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
e. a. g. pfaehler@umcg.nl
- ² Faculty of Medicine and Life Sciences, Hasselt University, Agoralaan Building D, 3590 Diepenbeek, Belgium
- ³ Department of Nuclear Medicine, Ziekenhuis Oost Limburg, Schiepse Bos 6, 3600 Genk, Belgium
- ⁴ Department of Radiology and Nuclear Medicine, VU University Medical Center, Amsterdam, The Netherlands
- ⁵ Department of Respiratory Medicine, Ziekenhuis Oost Limburg, Schiepse Bos 6, 3600 Genk, Belgium
- ⁶ Department of Respiratory Medicine, AZ Vesalius Hospital, Hazelereik 51, 3700 Tongeren, Belgium
- ⁷ Institute for Materials Research (IMO) - Division Chemistry, Hasselt University, Agoralaan Building D, 3590 Diepenbeek, Belgium

Abstract. In oncology, Positron Emission Tomography (PET) is frequently performed for cancer staging and treatment monitoring. Metabolic active tumor volume (MATV) as well as total MATV (TMATV - including primary tumor, lymph nodes and metastasis) derived from PET images have been identified as prognostic factor or for evaluating treatment efficacy in cancer patients. To this end a segmentation approach with high precision and repeatability is important. Moreover, to derive TMATV, a reliable segmentation of the primary tumor as well as all metastasis is essential. However, the implementation of a repeatable and accurate segmentation algorithm remains a challenge. In this work, we propose an artificial intelligence based segmentation method based on textural features (TF) extracted from the PET image. From a large number of textural features, the most important features for the segmentation task were selected. The selected features are used for training a random forest classifier to identify voxels as tumor or background. The algorithm is trained, validated and tested using a lung cancer PET/CT dataset and, additionally, applied on a fully independent test-retest dataset. The approach is especially designed for accurate and repeatable segmentation of primary tumors and metastasis in order to derive TMATV. The segmentation results are compared with conventional segmentation approaches in terms of accuracy and repeatability. In summary, the TF segmentation proposed in this study provided better repeatability and accuracy than conventional segmentation approaches. Moreover, segmentations were

accurate for both primary tumors and metastasis and the proposed algorithm is therefore a good candidate for PET tumor segmentation.

Keywords: Tumor segmentation · PET · Textural feature segmentation · Repeatability · Artificial intelligence

1 Introduction

Positron Emission Tomography is widely used in oncology for cancer diagnosis and treatment monitoring [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 or, in case of malignant lymphoma: all involved sites) is one important metric for the evaluation of therapy response. For a correct diagnosis, segmentation accuracy is crucial. For treatment follow-up, it is essential that observed differences in MATV/TMATV are caused by biological changes in the underlying tumor tissue and not by segmentation errors. Therefore, it is important that a segmentation algorithm yields accurate as well as repeatable segmentation results. Moreover, it is of interest to use a segmentation approach relying on PET information only as MATV may not be equal to and is fundamentally different from anatomical tumor size which can be extracted from a CT image. However, the implementation of such an algorithm is not trivial due to the challenges coming with PET images among them factors regarding the low signal-to-noise ratio, low spatial resolution, and partial volume effects [3]. Especially for smaller lesions, the partial volume effect can reduce the apparent tumor uptake making the lesion therefore difficult to detect and segment. Additionally, the image quality of a PET image depends highly on the scanner type, as well as on image acquisition and reconstruction. A segmentation algorithm leading to good results for images of one institution might not work for images of another institution. However, the EARL accreditation program aims to address this problem by harmonizing images across institutions.

In recent years, machine learning (ML) based segmentations such as Convolutional Neural Networks or classifiers classifying each voxel as tumor or background have shown very promising results for various segmentation tasks [4]. However, in PET imaging, few studies use advanced ML based segmentation approaches for metabolic active tumor segmentation. Even more, most studies combine the information of PET and low-dose CT images in order to get reliable segmentation results [5–7] but the image quality of low-dose CT is not optimal for segmentation purposes. Moreover, as stated above, MATV is not the same as anatomical tumor size and a segmentation based on PET information only may be more suitable to measure MATV, i.e. the metabolic active parts of the tumors. Therefore, it is of interest to develop segmentation approaches that rely on PET information only. Additionally, ML based segmentation approaches were so far only trained and applied on primary tumors, while for the calculation of TMATV, also an accurate and repeatable segmentation of metastasis and lymph nodes is important.

In this work, we present a textural feature (TF) based segmentation method designed especially for the accurate and repeatable segmentation of primary tumors and

metastasis. Moreover, our aim was to develop an explainable algorithm that uses hand-crafted features which can provide additional knowledge about tumor characteristics to physicians. The results of the TF segmentation approach is compared with conventional segmentation algorithms used in the clinic.

2 Materials and Methods

2.1 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 scientific research.

Two datasets acquired at two hospitals were included in this study with both datasets providing similar image quality as they were following the recommendations of the EARL accreditation program [8]. All images were converted to standardized uptake value (SUV) in order to normalize the images for differences in injected tracer dose and patient weight. Before the start of the segmentation process, a random bounding box was drawn around every tumor.

Training and Testing Dataset. For training, validating, and testing the algorithm, 96 images of patients with NSCLC Stage III and IV were included (Ziekenhuis Oost Limburg, Belgium). Patients fasted at least six hours before scan start and were scanned 60 min 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 $4 \times 4 \times 4$ mm using the vendor provided BLOB-OS-TOF algorithm. More details about the patient cohort can be found in previous studies [9]. The images were split randomly in training, validating, and testing sets, where 56 images were used for training, 14 images for validation, and 26 images for independent testing. All 451 lesions of the patients (primary tumors, lymph nodes, and metastasis) were included.

Test-Retest Dataset. For test-retest evaluation, we analyzed ten fully independent PET/CT scans of patients with Stage III and IV NSCLC. Images were acquired on a Gemini TF PET/CT scanner (Philips Healthcare, Cleveland, OH, USA) at a different institution (Amsterdam University Medical Center). These ten patients underwent two whole-body PET/CT scans on two consecutive days. Patient fasting time, time between tracer injection and scan start, as well as image reconstruction was the same as in the previous described dataset. More information about the patient cohort can be found in previous work [10]. A total of 28 lesions were included in the analysis.

Ground Truth Segmentations. The ground truth segmentations were obtained by applying an automatic segmentation which identified all voxels with a SUV above 2.5 as tumor (here after SUV2.5). An expert medical physicist adjusted all segmentations manually. 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 [11].

2.2 Segmentation Approaches

Textural Feature Segmentation (TF). In this segmentation approach, textural features were used for the voxel-wise segmentation of the tumor. As first step, every voxel was regarded as center of a scanning window. For each scanning window, textural features were calculated. Initially, scanning windows of size 3, 5, and 7 were used. For every view (axial, sagittal, coronal) a separate segmentation was performed. This means for e.g. the axial view textural features of scanning windows with size $3 \times 3 \times 1$, $5 \times 5 \times 1$ etc. were calculated. Calculated features included statistical (e.g. mean/kurtosis), intensity, as well as features describing the heterogeneity of a region (textural features). Before the calculation of textural features, the intensity values in the bounding box were discretized using a fixed bin number of 64. In the discretization step, every voxel intensity value is transformed to an integer value between 1 and 64. This step is required in order to calculate a large number of textural features. In total, 264 features were calculated for each voxel (88 for each neighborhood size). All steps before feature calculation are illustrated in Fig. 1.

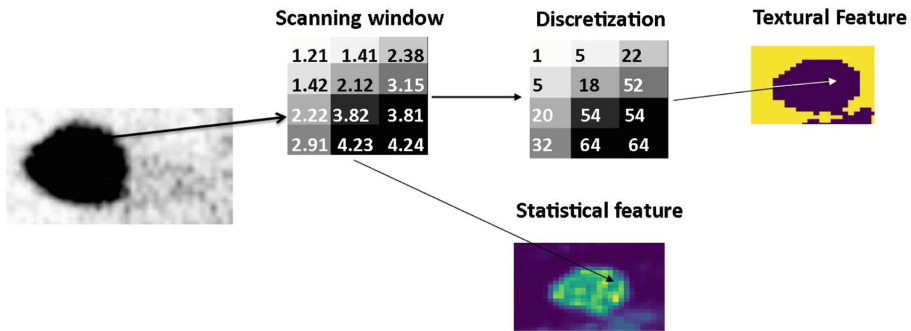


Fig. 1. Illustration of steps prior to feature selection. First, for each voxel a scanning window is defined. Directly from this scanning window, statistical features are calculated. For the calculation of textural features, the scanning window is discretized to contain only a limited number of discretized values. From the discretized scanning window, textural features are calculated.

In order to reduce the number of calculated features, the most useful features for the segmentation task were obtained using the feature importance obtained by a random forest. For each orientation, a separate feature selection was applied. This led to five representative features: energy, total energy, perc90, maximum, and mean for scanning window of size 3 and additionally maximum for size 5. The number of features resulting in the best accuracy was derived by experiments. Examples of these feature images are displayed in Fig. 2.

For classifying each voxel as tumor or background, a random forest was trained. The optimal hyper-parameter for the random forest were determined using a grid search with varying several parameters. This led to a random forest consisting of 1000

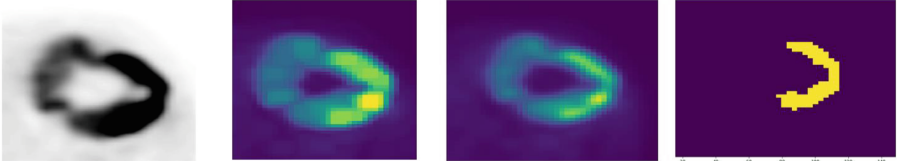


Fig. 2. Original PET image and example feature images of one tumor: Energy (left), Mean (middle), and Kurtosis (left). Energy and mean were selected as representative feature for the segmentation task, while kurtosis was not. All features are calculated from a $3 \times 3 \times 1$ scanning window.

decision trees, a tree depth of 100, minimum samples split of 5 with using bootstrap samples for tree building. After classification, the probability prediction images of the three orientations were stored. The probability images contain information how certain the algorithm is with its decision. In our algorithm, we used the majority vote of the probability images as final segmentation. Hereby, all three probability images are added and all voxels with a summed probability of more than 1.8 (and not 1.5 as it would be in a classical majority vote setting) were included in the tumor mask. This more strict threshold was chosen in order to include only voxels with a confident prediction certainty and deal in this way with the fuzziness of the tumor border in PET images.

Different Configurations of TF Approach. For the TF segmentation, we especially investigated the impact on using the combination of three 2D neighborhoods with the use of one 3D neighborhood as described in previous papers [4, 6]. For this purpose, we extracted textural features of cubic 3D neighborhoods of size 3, 5, and 7. Feature selection and classification were performed as described above. As for the 3D approach, there is only one classifier used (and not three for the different views), also no majority vote was necessary for the construction of the final segmentation.

Moreover, we also compared the use of a ‘classical’ majority vote approach (with a threshold of 1.5) with the more strict MV approach described above in terms of accuracy and repeatability.

Conventional Segmentation Algorithm. The accuracy and reproducibility of the TF segmentation was compared with two threshold based and 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 based on four frequently used thresholding approaches were included in the comparison. The underlying segmentation algorithm are described in previous work [12]. 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

2.3 Evaluation of Segmentation Algorithm

For the evaluation of the implemented segmentation algorithm, the approaches were compared in terms of accuracy and repeatability. The data analysis was performed in Python 3.6.2 using the packages numpy and scipy.

Accuracy of Segmentation Approaches. In order to determine segmentation accuracy, the Jaccard Coefficient (JC) between ground truth and performed segmentation was calculated. The JC is a measure for the overlap of the two segmentations: It is the ratio between the intersection and the union of two labels:

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

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

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:

$$TRT\% = \frac{|\mathbf{vol}_{\text{Day1}} - \mathbf{vol}_{\text{Day2}}|}{(\mathbf{vol}_{\text{Day1}} + \mathbf{vol}_{\text{Day2}})/2} * 100$$

The %TRT gives a measure for the proportional differences in segmented volume between the two consecutive scans.

JC values and TRT% were compared across segmentations 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 0.05 were considered as statistically significant. In order to correct for multiple comparisons, the p-values were corrected using the Benjamini-Hochberg correction.

3 Results

3.1 Comparison with Different Configurations of TF Algorithm

As displayed in Fig. 3 the TF approach using the classical majority vote resulted in lower accuracy when compared with the approach proposed in this work. The classical MV resulted in an underestimation in the majority of lesions. Moreover, it missed more lesions than the more strict majority vote approach.

The main segmentation differences between the two majority vote thresholds were observed at the tumor border. Two examples where the more strict threshold resulted in

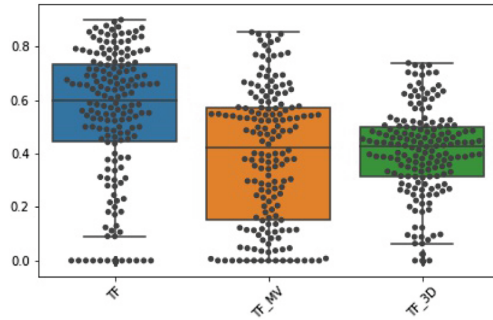


Fig. 3. JC values for different configurations of the TF segmentation approach: On the left (TF) the approach proposed in this work, Middle (TF_MV): The approach using a ‘normal’ majority vote (MV) approach, Right (TF_3D): Combining three dimensional neighborhoods

more true positive classified voxels are illustrated in Fig. 4. The use of 3D neighborhoods led also to a drop in segmentation accuracy when compared with the approach proposed in this work. Hereby, overestimations were observed in all cases. The differences in JC values between the approach proposed in this study and the two comparable approaches were found to be significant (p -value < 0.05).

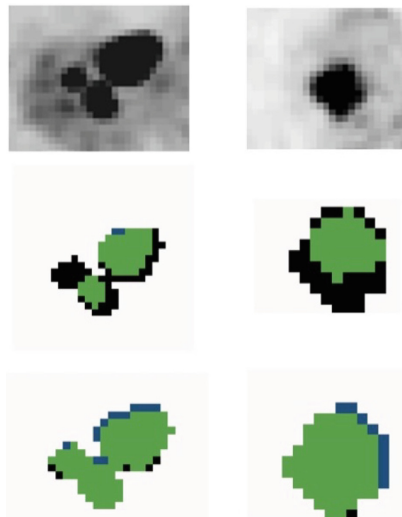


Fig. 4. Two segmentation results of random forest (one patient left, one patient right): Original PET image (Upper row), results with original majority vote approach (middle row) and with combining the probability images of the random forest (lower row). Green: true positives, White: true negative, Blue: False positives, Black: False positives (Color figure online)

At the same time, the approach proposed in this work resulted in better repeatability than the other two configurations of the TF algorithm as displayed in Fig. 5. The 3D approach resulted in the lowest repeatability. However, the differences in TRT% coefficient were not significant.

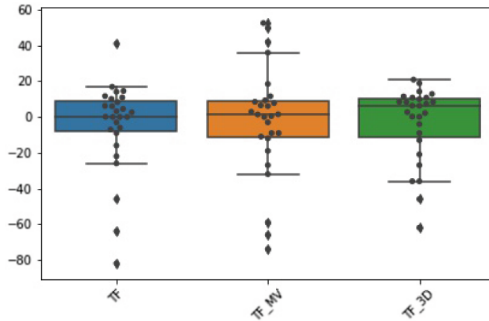


Fig. 5. Test-retest coefficient %TRT for different configurations of the TF segmentation: On the left (TF) the approach proposed in this paper, Middle (TF_MV): The approach using a ‘normal’ majority vote (MV) approach, Right (TF_3D): Combining three dimensional neighborhoods

3.2 Comparison with Conventional Segmentation Approaches

The boxplots of JC values for the testing as well as the external TRT dataset are displayed in Fig. 6. TF and MV2 segmentation resulted in general in the highest JC values. Significant differences in JC values were only observed between TF and SUV4, TF and 41%SUV_{MAX}, MV2 and SUV4, as well as MV2 and 41%SUV_{MAX} segmentation for the testing data. In the test-retest dataset, only TF and 41%SUV_{MAX}, as well as MV2 and 41%SUV_{MAX} segmentation resulted in significant differences.

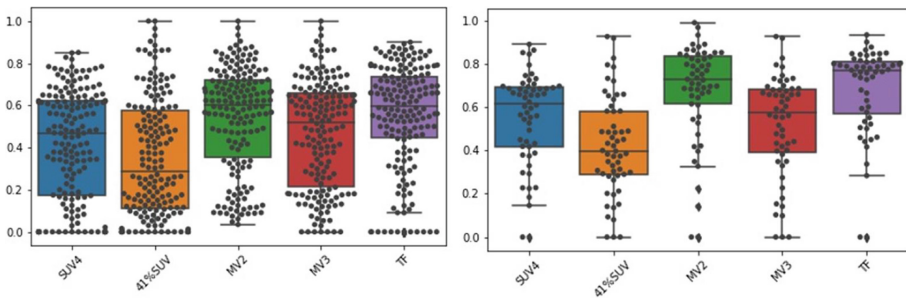


Fig. 6. Jaccard Coefficient (JC) values for both datasets: JC values for the testing set (left figure) and the test-retest dataset (right figure) for the different segmentation algorithm included in the study (SUV4: Standardized Uptake Value 4, 41%SUV_{MAX}, MV2: Majority Vote 2, MV3: Majority Vote 3, TF: Textural Feature based approach).

As displayed in Fig. 7, the accuracy of the segmentation was dependent on the lesion size. Bigger lesions resulted in higher JC values than smaller lesions. However, also for smaller lesions the TF and MV2 approach resulted in a better accuracy than the conventional segmentation approaches.

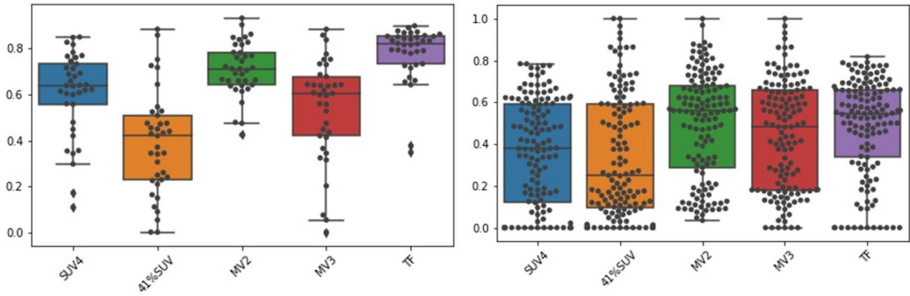


Fig. 7. Jaccard Coefficient (JC) values dependent on lesion size: JC values for bigger (left figure) and smaller (right figure) lesions for all segmentation approaches included in the study (SUV4: Standardized Uptake Value 4, 41%SUVMAX, MV2: Majority Vote 2, MV3: Majority Vote 3, TF: Textural Feature based approach).

All approaches missed some small lesions completely. This was the case when the tumors were located close to another high uptake region such as the kidney. Hereby, the kidney was incorrectly identified as tumor and the tumors were completely missed. A similar scenario was observed for two bigger lesions, for which all approaches resulted in a JC value below 0.5. This was the case when the tumors were located close to the heart which was incorrectly included in the segmentation.

Figure 8 displays the TRT-coefficients for all segmentation algorithm. TF and MV2 segmentation yield lower mean, and standard deviation of TRT% values than the other segmentation approaches. After applying the Benjamini-Hochberg correction, the differences in TRT were not significantly different. In the majority of the cases, a high TRT% came in combination with low JC values and large percentage volume differences. The lesion size did not influence the repeatability of the segmentations.

4 Discussion and Conclusion

The segmentation approach proposed in this work, outperformed conventional segmentation algorithm regarding segmentation accuracy and repeatability. Its performance was similar to a majority vote based approach. Therefore, the proposed segmentation approach is suitable for the segmentation of all lesions in PET images.

The segmentation of smaller lesions remains also for this approach a challenging task. One reason for this effect might be that with decreasing tumor size, small misclassifications have a higher impact on accuracy metrics. Smaller lesions also come with a lower tumor-to-background ratio and are therefore more difficult to detect. Moreover, some of the metastasis are also located close to other high-uptake regions

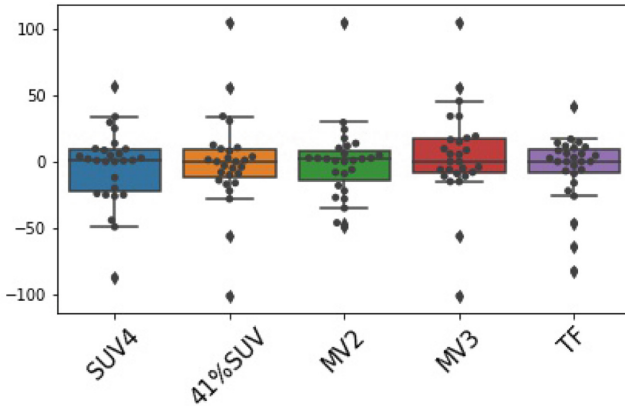


Fig. 8. Test-Retest Coefficient (TRT%) for all segmentation approaches: If the TRT% is close to 0, the repeatability of the segmentations is excellent. Abbreviations of the segmentation algorithm: MV3: Majority Vote 3, MV2: Majority Vote 2, TF: Textural Feature based approach, Max41: 41% SUV_{MAX} , SUV4: Standardized Uptake Value 4)

(such as the kidney) what opposes a special challenge to a segmentation algorithm. The TF based approach achieves in general a higher accuracy than similar approaches proposed in other studies [3, 4]. An important difference between our method and other published algorithm is that our approach relies 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 [5, 6].

The selected features are simple statistic measurements describing tumor uptake. This is due to the fact that they were determined for primary tumors, metastasis, and lymph nodes. More complex textural features are selected when only primary tumors are included in the feature selection step. This is likely due to the fact that for primary tumors texture and background are similar. However, it is a strength of our algorithm that it yields accurate and repeatable segmentation results for all lesions in a patient. Nonetheless, including the feature selection in the segmentation step might improve the selected features and performance of the segmentation algorithm.

In future work, we plan to compare the results of the proposed segmentation algorithm with the results of a CNN. However, as CNNs require a large amount of training data and act more like a black-box, we wanted to focus in this work on an explainable machine learning based segmentation approach that can also be used with little amount of training data. We developed this approach for the segmentation of MATV in PET images, but this approach will likely also yield good results when applied on MR or contrast-enhanced CT images. In future studies, we also plan to use this approach in order to understand changes in tumor tissue e.g. before and after radiotherapy.

In summary, we demonstrate in this work that our proposed ML based segmentation has not only the potential to accurately segment lesions but also to result in repeatable segmentations. Therefore, the proposed segmentation approach is suitable for the segmentation of tumors in PET images.

Acknowledgements. 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.

Disclosure of Conflicts of Interest. The authors have no relevant conflicts of interest to disclose.

Ethical Approval. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

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.

References

1. Volpi, S., Ali, J.M., Tasker, A., et al.: The role of positron emission tomography in the diagnosis, staging and response assessment of non-small cell lung cancer. *Ann. Transl. Med.* **6**, 95–95 (2018). <https://doi.org/10.21037/atm.2018.01.25>
2. Griffith, L.K.: Use of PET/CT scanning in cancer patients: technical and practical considerations. *Proc. (Bayl. Univ. Med. Cent.)* **18**, 321–330 (2005). <https://doi.org/10.1080/08998280.2005.11928089>
3. Hatt, M., Laurent, B., Ouahabi, A., et al.: The first MICCAI challenge on PET tumor segmentation. *Med. Image Anal.* **44**, 177–195 (2018). <https://doi.org/10.1016/j.media.2017.12.007>
4. Markel, D., Caldwell, C., Alasti, H., et al.: Automatic segmentation of lung carcinoma using 3D texture features in 18-FDG PET/CT. *Int. J. Mol. Imaging* **2013**, 1–13 (2013). <https://doi.org/10.1155/2013/980769>
5. Zhong, Z., Kim, Y., Zhou, L., et al.: 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. In: *Proceedings - International Symposium on Biomedical Imaging*, pp. 228–231, April 2018. <https://doi.org/10.1109/ISBI.2018.8363561>
6. 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* **28**, 374–383 (2009). <https://doi.org/10.1109/TMI.2008.2004425>
7. Yu, H., Caldwell, C., Mah, K., 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.* **75**, 618–625 (2009). <https://doi.org/10.1016/j.ijrobp.2009.04.043>
8. Aide, N., Lanson, C., Veit-Haibach, P., et al.: EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. *Eur. J. Nucl. Med. Mol. Imaging* **44**, 17–31 (2017). <https://doi.org/10.1007/s00259-017-3740-2>
9. Vanhove, K., Mesotten, L., Heylen, M., et al.: Prognostic value of total lesion glycolysis and metabolic active tumor volume in non-small cell lung cancer. *Cancer Treat Res Commun.* **15**, 7–12 (2018). <https://doi.org/10.1016/j.ctarc.2017.11.005>

10. Kramer, G.M., Frings, V., Hoetjes, N., et al.: Repeatability of quantitative whole-body 18F-FDG PET/CT uptake measures as function of uptake interval and lesion selection in non-small cell lung cancer patients. *J. Nucl. Med.* **57**, 1343–1349 (2016). <https://doi.org/10.2967/jnumed.115.170225>
11. van Baardwijk, A., Bosmans, G., Boersma, L., 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.* **68**, 771–778 (2007). <https://doi.org/10.1016/j.ijrobp.2006.12.067>
12. Kolinger, G.D., Vázquez García, D., Kramer, G.M., et al.: Repeatability of [18F]FDG PET/CT total metabolic active tumour volume and total tumour burden in NSCLC patients. *EJNMMI Res.* **9**, 14 (2019). <https://doi.org/10.1186/s13550-019-0481-1>