Radiotherapy and Oncology 188 (2023) 109774

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Deep learning based automated delineation of the intraprostatic gross tumour volume in PSMA-PET for patients with primary prostate cancer

Julius C. Holzschuh [a,b,c,*], Michael Mix [d], Juri Ruf [d], Tobias Hölscher [f], Jörg Kotzerke [g], Alexis Vrachimis [h], Paul Doolan [s], Harun Ilhan [i], Ioana M. Marinescu [a,b], Simon K.B. Spohn [a,b,j], Tobias Fechter [a,b,e], Dejan Kuhn [a,b,e], Peter Bronsert [k], Christian Gratzke [l], Radu Grosu [m,t], Sophia C. Kamran [n], Pedram Heidari [o], Thomas S.C. Ng [o,p,q], Arda Könik [p,q], Anca-Ligia Grosu [a,b], Constantinos Zamboglou [a,r]

[a] Department of Radiation Oncology, Medical Center - University of Freiburg; [b] German Cancer Consortium (DKTK), Partner Site Freiburg, Freiburg; [c] Faculty of Computer Science, Karlsruhe Institute of Technology, Karlsruhe; [d] Department of Nuclear Medicine, Medical Center - University of Freiburg; [e] Division of Medical Physics, Department of Radiation Oncology, Medical Center – University of Freiburg, Faculty of Medicine, Freiburg; [f] Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technical University Dresden; [g] Department of Nuclear Medicine, Faculty of Medicine and University Hospital Carl Gustav Carus, Dresden, Germany; [h] Department of Nuclear Medicine, German Oncology Center - University Hospital of the European University, Limassol, Cyprus; [i] Department of Nuclear Medicine, University Hospital - Ludwig-Maximilians-Universität, Munich; [j] Faculty of Medicine - University of Freiburg, Berta-Ottenstein-Programme; [k] Department of Pathology; [l] Department of Urology, Medical Center - University of Freiburg, Freiburg, Germany; [m] Cyber-Physical Systems Division, Institute of Computer Engineering and Faculty of Informatics, Technical University of Vienna, Vienna, Austria; [n] Department of Radiation Oncology, Massachusetts General Hospital - Harvard Medical School; [o] Division of Nuclear Medicine and Molecular Imaging, Massachusetts General Hospital - Harvard Medical School, Department of Radiology; [p] Joint Program in Nuclear Medicine, Brigham and Women's Hospital - Harvard Medical School; [q] Department of Imaging, Dana-Farber Cancer Institute - Harvard Medical School, Boston, USA; [r] German Oncology Center, European University of Cyprus; [s] Department of Radiation Oncology, German Oncology Center - University Hospital of the European University, Limassol, Cyprus; [t] Department of Computer Science, State University of New York at Stony Brook, NY, USA

ARTICLE INFO

Article history:
Received 18 April 2023
Received in revised form 17 June 2023
Accepted 22 June 2023
Available online 30 June 2023

Keywords:
PSMA-PET
Prostate
CNN
Machine Learning
Segmentation

ABSTRACT

Purpose: With the increased use of focal radiation dose escalation for primary prostate cancer (PCa), accurate delineation of gross tumor volume (GTV) in prostate-specific membrane antigen PET (PSMA-PET) becomes crucial. Manual approaches are time-consuming and observer dependent. The purpose of this study was to create a deep learning model for the accurate delineation of the intraprostatic GTV in PSMA-PET.

Methods: A 3D U-Net was trained on 128 different ^{18}F-PSMA-1007 PET images from three different institutions. Testing was done on 52 patients including one independent internal cohort (Freiburg: n = 19) and three independent external cohorts (Dresden: n = 14 ^{18}F-PSMA-1007, Boston: Massachusetts General Hospital (MGH): n = 9 ^{18}F-DCFPyL-PSMA and Dana-Farber Cancer Institute (DFCI): n = 10 ^{68}Ga-PSMA-11). Expert contours were generated in consensus using a validated technique. CNN predictions were compared to expert contours using Dice similarity coefficient (DSC). Co-registered whole-mount histology was used for the internal testing cohort to assess sensitivity/specificity.

Results: Median DSCs were Freiburg: 0.82 (IQR: 0.73–0.88), Dresden: 0.71 (IQR: 0.53–0.75), MGH: 0.80 (IQR: 0.64–0.83) and DFCI: 0.80 (IQR: 0.67–0.84), respectively. Median sensitivity for CNN and expert contours were 0.88 (IQR: 0.68–0.97) and 0.85 (IQR: 0.75–0.88) (p = 0.40), respectively. GTV volumes did not differ significantly (p > 0.1 for all comparisons). Median specificity of 0.83 (IQR: 0.57–0.97) and 0.88 (IQR: 0.69–0.98) were observed for CNN and expert contours (p = 0.014), respectively. CNN prediction took 3.81 seconds on average per patient.

Conclusion: The CNN was trained and tested on internal and external datasets as well as histopathology reference, achieving a fast GTV segmentation for three PSMA-PET tracers with high diagnostic accuracy comparable to manual experts.

Prostate Cancer (PCa) is one of the most frequently diagnosed malignancies in men worldwide [1]. External-beam radiotherapy is a crucial pillar in the treatment of localized PCa. For this modality, an accurate GTV delineation is required to allow novel treatment approaches like focal dose escalation with intensity-modulated radiotherapy, brachytherapy or stereotactic body radiotherapy [2,3].

* Corresponding author at: Department of Radiation Oncology, Medical Center - University of Freiburg, Robert Koch Straße, 379106 Freiburg, Germany.
E-mail address: julius.holzschuh@uniklinik-freiburg.de (J.C. Holzschuh).

https://doi.org/10.1016/j.radonc.2023.109774
0167-8140/

The emergence of prostate-specific membrane antigen positron emission tomography (PSMA-PET) has revolutionized PCa staging and management. Not only does PSMA-PET detect visceral and bone metastases but has also demonstrated significant potential for characterizing intraprostatic tumor lesions. Some studies even suggest higher sensitivities and comparable specificity of PSMA-PET for detecting intraprostatic lesions compared to prostate multiparametric magnetic resonance imaging (mpMRI) [4,5].

Although new manual delineation techniques on ^{18}F-PSMA-1007-PET have been proposed and validated for GTV segmentation [4], these approaches are time-consuming, require significant clinical expertise, and are observer-dependent. To overcome this issue, the use of automated software comes into mind.

Over the past decade machine learning has proven to be a valuable tool in the automation of complex tasks in medicine. Specifically convolutional neural networks (CNNs) have shown great results for image recognition.

First studies on CNN-based intraprostatic GTV delineation for PCa patients were conducted on ^{68}Ga-PSMA-PET [6,7]. To the best of our knowledge, this is the first work of describing a CNN trained on ^{18}F-PSMA-1007 PET/CT to segment intraprostatic GTV. Additionally, many modern deep-learning studies for medical imaging remain confined to validation on small internal cohorts whilst lacking external validation, histopathology reference, or data on real-world clinical utility [8].

Thus, in this work, a dataset of 180 patients with ^{18}F-PSMA-1007 PET/CT was used to train (n = 128), and to internally as well as externally test (n = 52) the CNN performance using different metrics and PSMA PET tracers as well as co-registered histological PCa information.

Materials and methods

Patients

In this study, 180 patients with primary PCa from six different centers (Freiburg: n = 96, Cyprus: n = 32, Munich: n = 19, Dresden: n = 14, Boston: MGH n = 9, Dana-Farber Cancer Institute n = 10) were retrospectively included. Training was conducted using 128 patients, with a further 52 for testing. An overview of different cohorts and patient characteristics is presented in Table 1. Inclusion criteria were biopsy-proven primary PCa who have not received prior treatment at the time of imaging. Approval or exemption from local ethics committees was obtained from all centers for this study.

Imaging (PET/CT)

An overview of scanner characteristics is provided in Table 2. CT-based attenuation correction was used. Standardized uptake values (SUV [g/ml]) were calculated to normalize images based on decay-corrected injected activity per kg body weight. All PSMA-PETs were performed as part of primary staging according to the clinical practice at the respective study centers.

Contouring of PSMA PET/CT

For expert GTVs, the consensus contour of two different readers board-certified in radiation oncology, diagnostic radiology and/or nuclear medicine was used with a level of experience in intraprostatic GTV contouring and PET image analysis of 3–7 years. For each patient a slice wise segmentation was done, using a windowing of $SUV_{min\text{-}max}$ of 0 to 10 for ^{18}F tracers and 0–5 for ^{68}Ga tracers and inverted grey scale as colormap as proposed by Zamboglou et al. [4 9]. Readers were instructed to delineate the suspicious avidity without applying any specific SUV_{max} criteria. Prostatic gland was manually contoured according to ESTRO-ACROP consensus guidelines [10].

For software Eclipse v15.1 (Varian Medical Systems, USA) and 3DSlicer v4.10 were used. Images from Boston cohorts were contoured locally. For the rest of the data, contouring was conducted in Freiburg.

Histopathology co-registration

Three-dimensional (3D) intra prostatic PCa distribution was obtained using histology information from prostatectomy specimens for 19 patients (Freiburg cohort). For co-registration, a previously established protocol was used [11]. Briefly, the resected prostate specimen underwent formalin fixation as well as an ex-vivo CT scan. A customized localizer and whole-mount step sections were used to cut the specimen into 4 mm slices. After Hematoxylin and eosin staining, slice-wise delineation of PCa tissue was performed by pathologists. Contours were then transferred to reg-

Table 1
Dataset characteristics.

Dataset (n = 180)	Freiburg	Cyprus	Munich	Dresden	Massachusetts General Hospital	Dana-Farber Cancer Institute
Mean Age in years (standard deviation)	69.3 (8.1)	69.2 (7.5)	67.2 (10.9)	70.4 (8)	74 (6.8)	70.8 (6.8)
Median iPSA in ng/ml (min–max)	14.6 (4.2–164)	10.2 (2.75–167)	10.4 (4.6 – 465)	16.5 (5–139)	18 (5–56.2)	11.9 (6.4–24.9)
ISUP						
1	5 (5%)	8(25%)	2 (10%)	1 (7%)	0	2 (20%)
2	24 (25%)	4(12.5%)	6 (32%)	2 (14%)	3 (33%)	3 (30%)
3	29 (30%)	8 (25%)	3 (16%)	4 (29%)	0	1 (10%)
4	21 (22%)	9 (28%)	6 (32%)	2 (14%)	5 (55%)	0
5	17 (18%)	3 (9%)	2 (10%)	3 (21%)	1 (11%)	4 (40%)
		-	-	-	-	-
unknown				2		
cT stage						
T1	-	11	-	4	4	
T2	48	10	12	4	2	3
T3	46	9	6	4	3	4
T4	2	-	1	-	-	-
Unknown	-	2	-	2	-	3
Training	77	32	19	-	-	-
Testing	19	-	-	14	9	10
With histological reference	19	-	-	-		

Table 2
Imaging characteristics.

Center	Freiburg	Munich	Dresden	Cyprus	Massachusetts General Hospital	Dana-Farber Cancer Institute
PET imaging system (Type, Manufacturer)	64-slice Vereos PET/CT and Gemini TF Big Bore (Philips Healthcare, USA)	GE Discovery 690 and Siemens (Biograph mCT and Biograph 64)	Biograph Vision 600, Siemens Healthcare GmbH	Discovery IQ2 PET/CT system (4 rings; 16 slices) of General Electric	Siemens mCT Flow 64 slice PET/CT	GE Discovery MI 4 ring
Tracer	^{18}F-PSMA-1007	^{18}F-PSMA-1007	^{18}F-PSMA-1007	^{18}F-PSMA-1007	^{18}F-DCFPyL	^{68}Ga-PSMA-11
Post injection time	2 hours	1 hour	ca. 90 min	2 hours	1 hour	1 hour
CT	120 kV, 100–400 mAs (dose modulation)	120 kV, 200–240 mAs	12 kV, 11 mAs	120 kV, 15–220 mAs (dose modulation)	120kv, 11mAs	120kv, 3 mAs, no forced diuresis
Contrast enhancement	mixed	+	-	-	-	-
Accreditation	EARL[1]	Phantom studies based on the National Electrical Manufacturers Association NU2-2001 standard were conducted to allow valid pooling of the results, and SUV conversion factors were calculated and implemented	Fulfilling EARL specifications	EARL	ACR[2]	ACR
Previous publications (if any)	[11,22,29]	-	[30]	[31,32]	-	-

[1] EANM Research Ltd.
[2] American College of Radiology.

istered ex-vivo CT and interpolated to create a 3D model. Manual co-registration was used to transfer GTV contours from ex-vivo CT to in-vivo CT. Hardware co-registration of hybrid PET/CT scanners resulted in the final alignment of in-vivo CT and PET scans. An example for co-registered histopathology can be seen in (Fig. 3).

Preprocessing

For PET imaging data, body weight-adapted SUVs were calculated and saved in nearly raw raster data format (nrrd). PET data were resampled to a voxel size of 2x2x2 mm3 with bspline interpolation [7]. Contours were resampled using nearest neighbour interpolation with SimpleITK v2.2.0 and plastimatch v1.8.0.

Inputs were cropped around the prostate to a size of 64x64x64 voxels, resulting in an approximate receptive field of 12.8 $\times$ 12.8 $\times$ 12.8 cm^3 for each voxel in the predicted GTV segmentation. Using value clipping (setting values above 15 to 15), SUV intensities were normalized to match an interval of [0, 15].

CNN

Architecture: A 3D U-Net variant was used [12],consisting of an encoder and decoder part with 3 layers each. A detailed description of the underlying architecture can be found in Supplementary Fig. 2. PET and prostate contour are used as input. Sigmoid function is used to create prediction probability values at the final layer. A threshold function with a threshold of 0.5 maps outputs to final GTV prediction. CNN was implemented in pytorch v1.10.

Training: For network optimization, the training dataset (n = 128) was further split randomly using 80%/20% split for training and validation. Prostate contour and normalized SUV-PET were used as input for the CNN. Manual GTVs were used for loss calculation while training. Hyperparameters were optimized using optuna [8]. Final model was trained using mini batch gradient descent with adaptive moment estimation, with a batch size of 8 and soft dice loss as loss function. While training, dice loss was calculated on the whole batch as a pseudo volume.

Evaluation

Volumetric Dice Similarity Coefficient (DSC) [13] and Hausdorff Distance (HD) [14] were assessed between expert GTVs and CNN GTVs at voxel level. Co-registered whole-mount histology was used as standard of reference to calculate sensitivity and specificity for the internal validation cohort as described previously [15]. For visual comparison between histopathology and GTVs, the prostate was divided into four equal segments for each slice in CT from PSMA-PET/CT images. A mean of 54 segments (range: 32–68) were analyzed per patient. Due to the lack of information on extra prostatic PCa tissue, only intraprostatic delineations were used for final calculations. As the dataset also contained patients with multiple intraprostatic GTVs, the CNN was trained to perform contouring of all intraprostatic GTVs within the same prostate. In instances where multiple intraprostatic GTVs were present, all of them were taken into consideration during the evaluation process. Reported values were then calculated on an individual patient basis, ensuring a patient-centric analysis.

Although a custom GPU accelerated implementation was used to calculate batched dice score while training, final evaluation metrics were calculated using MedPy v0.3.0 package for each individual patient to provide better comparability. In Freiburg plain Ubuntu with GPU acceleration was used whereas Boston used WSL without GPU acceleration.

External validation

External validation was performed in two cohorts from Boston (^{18}F-DCFPyL-PSMA and ^{68}Ga-PSMA-11 PET/CT images, respectively). The previously trained model was transferred to Boston and GTV predictions were made locally. The same methods used for internal testing were applied. No additional training of the model was performed on ^{18}F-DCFPyL-PSMA and ^{68}Ga-PSMA-11 PET/CT images.

Statistical analysis

Pairwise comparisons were performed using Wilcoxon single rank test with Bonferroni correction as the data did not show an underlying normal distribution (D'Agostino-Pearson test). For analysis, scipy library was used. Tests were performed two sided and significance level was set to 0.05. Figures were created using matplotlib. For non-gaussian distributed data, outlier identification was done based on visual inspection. For gaussian distributed data, grubbs test was used.

Results

Comparing CNN predictions (GTV-CNN) to expert contours (GTV-Expert) on the internal 18F-PSMA-1007 PET cohort, a median DSC of 0.82 (IQR: 0.73–0.88) and HD of 3.3 mm (IQR: 2.12–6.03) were observed. Sensitivity and specificity were calculated based on histopathology reference standard. Sensitivity of the CNN had no statistically significant differences (p = 0.401) to manual contours, with a median sensitivity of 0.88 (IQR: 0.68–0.97) and 0.85 (IQR: 0.75–0.88), respectively (Fig. 1). With a median specificity of 0.83 (IQR: 0.57–0.97) and 0.88 (IQR: 0.69–0.98) for CNN and expert contours, respectively, CNN specificity was slightly lower (p = 0.013) (Fig. 1). Histopathology-based sensitivity and specificity values showed concordance between the CNN and experts regarding challenging cases (lower quartile in Fig. 1.).

Absolute GTV volumes did not differ significantly (Expert vs. CNN: p = 0.148, CNN vs. Histology: p = 0.287, Expert vs. Histology: p = 0.676) with a median of 3.2 ml (IQR: 1.6 ml-5.2 ml) for GTV-CNN, 3.1 ml (IQR: 1.4–5.1 ml) for GTV-Expert and 2.3 ml (IQR: 0.9–4.3 ml) for whole mount histology as reference (Fig. 2).

For external testing, the Dresden dataset yielded a median DSC of 0.71 (IQR: 0.53–0.75) and a median HD of 7.81 mm (IQR: 5.85–11.39). Expert GTVs showed no significant (p = 0.135) differences to CNN predictions in terms of volumes with a median of 3.3 ml (IQR: 1.4––6.3 ml) for experts and a median volume of 4.4 ml (IQR: 1.5–7.3 ml) for CNN.

External testing on ^{18}F-DCFPyL tracer at the Massachusetts General Hospital showed a median DSC of 0.8 (IQR: 0.65–0.82) and a median HD of 4.12 mm (IQR: 3.00–7.00). Median absolute volumes were 5.98 (IQR: 4.95–10.76) for experts and 3.85 ml (IQR: 3.03–5.48) for CNN predictions (p = 0.0039).

External testing on ^{68}Ga tracer at the Dana Farber Cancer Institute showed a median DSC of 0.80 (IQR: 0.73–0.90) and HD of 9.53 mm (IQR: 4.61–11.79). Median volumes for experts were 7.72 ml (IQR: 2.89–21.77) and for CNN 6.6 ml (IQR: 2.41–13.02) (p = 0.19).

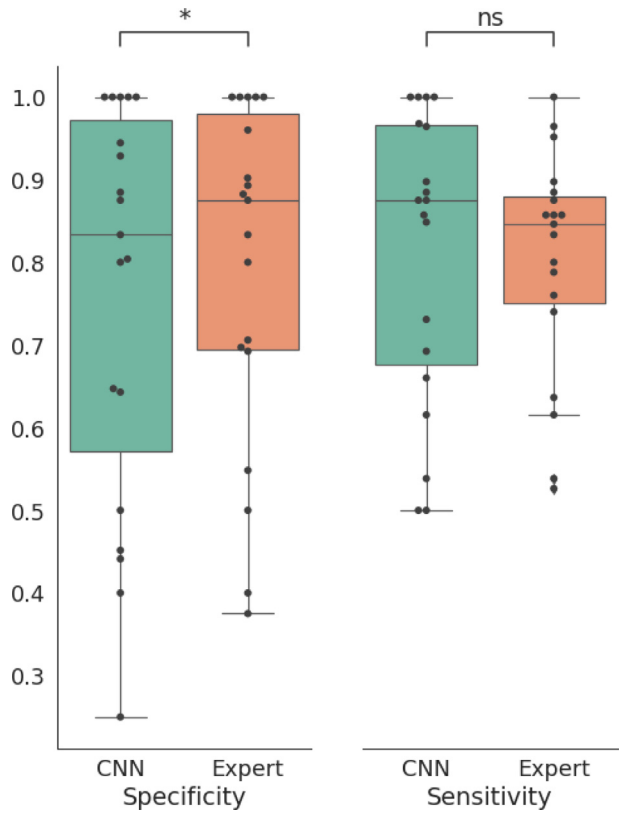

Fig. 1. Sensitivity and Specificity using histological whole mount as reference. Comparison of Specificity and Sensitivity visualized as boxplot using histopathological information of prostate specimen as reference. Experts have a slightly higher specificity, while the CNN has a slightly higher sensitivity. Legend: *: significance level of p = 0.05 for Wilcoxon single rank test was reached.

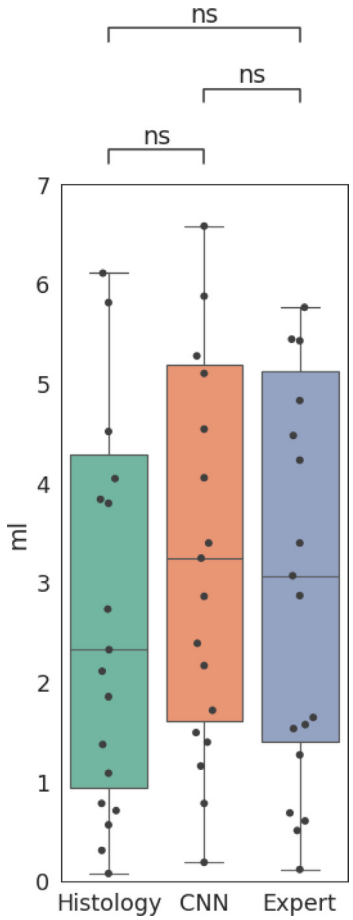

Fig. 2. Volume comparison Boxplot for comparison of PCa and GTV volumes. Significance levels of (p = 0.05) were not reached. Volumes do not differ significantly. Chosen scale doesn't include two outliers with volumes above 20 ml. Figure including outliers can be found in the supplementary.

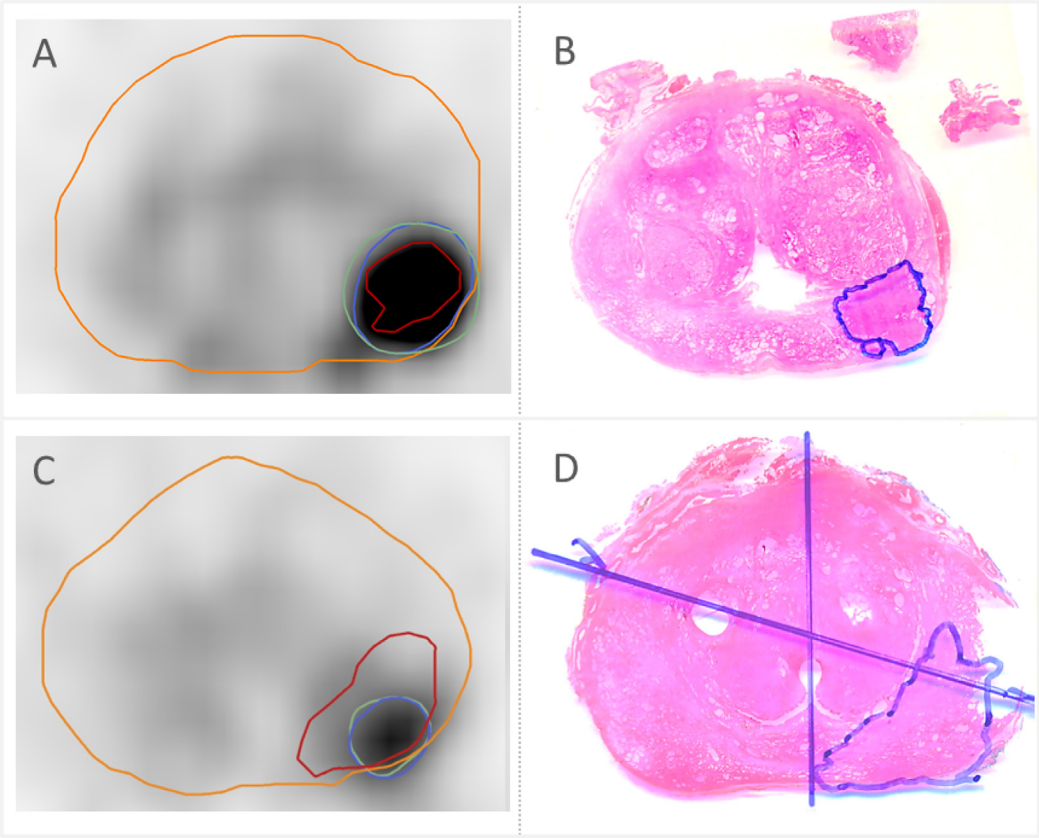

Fig. 3. Patient example A) and C) Axial PET image (windowing SUVmin-max: 0–10) of a patient from the internal testing cohort. Contours: orange = prostate contour by expert based on CT image, red = co-registered PCa in histology, green = CNN prediction, blue = expert contour.B) and C) Corresponding haematoxylin and eosin whole mount histology slide with marked PCa lesion in blue. Slight deformations resulting from interpolation and co-registration are observable when comparing histology-based PCa (A: red, B: blue) and (C: red, D: blue) in a side by side comparison.

Overall for the CNN, creating a single prediction, including reading from storage and writing the resulting file took 3.81 seconds with a standard deviation (SD) of 1.38 s for the whole training and testing dataset excluding Boston cohorts. About 95% of this time was taken up by relocation in the image tensor and writing the file. For Boston cohorts an average time of 3.65 s (SD: 0.84 s) was observed. Please see Fig. 4 for the prediction time for the CNN in the different datasets.

Discussion

Over the past decade, CNNs have shown to be a valuable tool for automating complex tasks in image recognition. This study aimed to provide a robust method for automated intraprostatic GTV segmentation. This task is critical for several diagnostic and therapeutic procedures such as PSMA-PET-based targeted biopsies [16], radiomic feature extraction [17], and focal radiotherapy dose escalation [18].

Our study used a CNN trained on a multicenter dataset of 128 patients. It was validated on internal as well as external datasets, and across multiple PSMA-targeted tracers. Further validation was performed using co-registered histopathology information as a standard of reference. Our results indicate the CNN performance is comparable to physician experts. While expert contouring can take approximately 5–10 minutes for each contour, our model was able to create them in a few seconds. Given the anticipated use of this method and the likely rise in the clinical volume of PCa patients, such time savings across a clinical workday can be substantial.

Comparing CNN contours to expert GTVs, a median DSC of 0.71 to 0.82 was measured. Overall DSC values compare well with previous studies. Draulans et al. [19] reported an interobserver variability with a DSC of 0.58 and 0.72 respectively, for ^{18}F-PSMA- and ^{68}Ga-PSMA PET. Spohn et al. [4] reported an interobserver agreement with a DSC of above 0.87 for ^{18}F-PSMA-1007 by using a validated contouring technique. Hence the results of our model lie well in the range of manual inter-observer variability.

The highest performance of the CNN was reached for the internal testing cohort. This is unsurprising as the biggest portion of patients in the training dataset were internal patients from the same center. Interestingly DSC was only slightly lower upon external testing on ^{18}F-DCFPyL and ^{68}Ga-PSMA-11, despite the different PET acquisition technique and overall tracer biodistribution.

^{18}F-DCFPyL has a statistically significant higher uptake in the urinary bladder [20], resulting in higher local SUVs. The same phenomenon has been described for ^{68}Ga-PSMA-11 [7]. As the network was trained solely on ^{18}F-PSMA-1007 PET images, the differentiation of bladder and GTV proposes a difficult challenge to the CNN as it introduces a distributional shift for training and testing data. To partially mitigate this issue, we implemented a cutoff SUV above 15, resulting in 106 out of 180 patients (58%) having a GTV exceeding this threshold.

Although data from Dresden used ^{18}F-1007 as tracer, the CNN performed worst on this part of the testing dataset. A potential reason for this might be the underlying reconstruction kernel. Nearly all other PET images were reconstructed using a variation of a soft

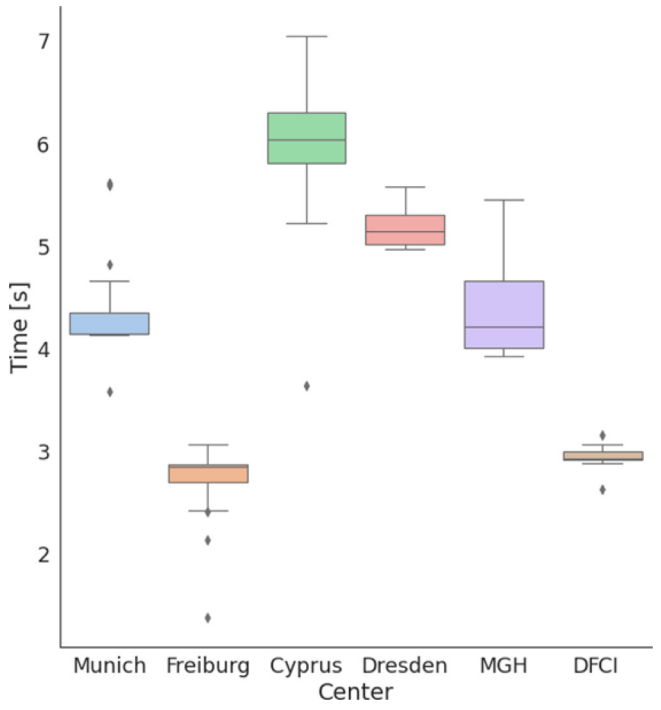

Fig. 4. Time comparison Boxplot for time measurements for creating a single prediction, including reading from storage and writing the resulting file. Training dataset was added to testing cohorts for this specific task. Measurements were taken on a different computer system without GPU acceleration for Massachusetts General Hospital (MGH) and Dana-Farber Cancer Institute (DFCI) cohorts.

gaussian kernel, but this part of the testing dataset did not use such an approach. This could also be seen in visual inspection as images had sharper edges and more noise (see supplementary Fig. 1). CNNs do not only rely on local intensity values but also feature maps that take shapes into account. Also, object texture seems to play a major role in the robustness of CNNs [21]. Hence, this domain shift might be the reason for decreased performance.

For the external datasets the CNN showed a poorer performance based on the HD metric (median 3.3–9.5 mm). The discrepancy between good DSC and moderate HD results might be explained by the fact that a single outlying voxel heavily affects the HD. Future studies should assess whether these outliers affect the clinical outcomes after focal dose-escalated radiotherapy on CNN-generated GTVs. Volumes of histological PCa, CNN predictions and expert contours did not differ significantly and are in concordance with other studies [22,4] and [19].

Quadrant based comparison between histology reference and the contours revealed slightly lower specificity for CNN than for experts, while the CNN had a higher sensitivity than experts. Significance levels were only reached for the comparison of the specificity. This difference in values is probably more of a neglectable result due to the sensitivity/specificity tradeoff. As the CNN outputs probability values in the last layer before binary discretization, a custom threshold could be used for further fine tuning the model's classification decision, allowing additional flexibility in adjusting the model's sensitivity and specificity based on specific application requirements or performance trade-offs. By setting a custom probability threshold, the user can define the minimum probability value that the output of the CNN must surpass to be classified as positive.

In the past several studies for ^{18}F-PSMA-1007 PET analyzed sensitivity and specificity with co-registered wholemount histology as reference [23 24 4]. Kuten et al. [23] observed a sensitivity of 100% and a specificity of 90.9%. Kesch et al. [24] reported a lower sensitivity of 71% and specificity of 81%. Spohn et al. [4] presented a sensitivity of 87% and specificity of 96%. In this study experts performed with a specificity of 88% and sensitivity of 85%. In an intra-individual comparison to experts, the CNN yielded a specificity of 83% and a sensitivity of 88%. One explanation for the difference in values might be the different approaches for co-registration and sensitivity/specificity calculations [15]. Also, differences in patient cohorts as well as observer variability seem to play a role as Spohn et al. [4] used the same in-house co-registration and analysis protocol as used for this study. Partial volume effects also make the detection of very small tumor lesions (around 2 mm) technically challenging since they can obscure small lesions [25 26]. Furthermore, the presence of visually undetectable lesions (due to low PSMA expression or small volume) introduces a certain level of uncertainty, resulting in lower sensitivity. In challenging cases where the specificity is below 50%, the lower values can primarily be attributed to two factors: overestimation of the GTV or the detection of PET positive regions that do not have a corresponding histopathological PCa correlate. As this work also included patients with small lesions, higher sensitivities of nearly 100%, as previously suggested [23], seem out of reach. Overall, our values for sensitivity and specificity based on co-registered wholemount histology lie well in the range of previously published results.

Bettermann et al. demonstrated that PSMA-PET and mpMRI offer complementary information when it comes to expert-driven manual delineation of prostate GTVs [34] Future studies should investigate whether this also holds true for GTVs generated using CNN-based methods.

Limitations of this study should be noted. A first limitation of our study can be seen in the size of the dataset used for testing (n = 52). Second, non-linear shrinkage of the prostate after prostatectomy can present alignment challenges between imaging and histopathology. To counteract this bias, evaluation was performed slice-wise and not for each individual voxel.

During volumetric analysis we also discovered that our testing dataset contained two very large intraprostatic GTVs (>20 ml). Although the model demonstrated satisfactory performance on those particular cases (DSC > 0.9, its overall accuracy when confronted with out-of-distribution cases remains unpredictable. Thus, we strongly suggest implementing the CNN model in a more supportive clinical use case rather than for fully autonomous decision-making. Despite efforts to maintain consistency in the training and testing cohorts, it is important to acknowledge that differences in clinical characteristics, particularly PSA and ISUP grades existed in our study. This highlights the generalizable performance of our CNN model given the known variability in SUV values of tumours among patients with varying PSA values and ISUP grades.

Although DSC is considered one of the most common scores for image segmentation, some pitfalls must be taken into consideration [27 28]. Especially small volumes relative to voxel size can be an issue. DSC is slightly biased toward single-object detection. Additionally, in silico segmentation metrics might not necessarily correlate with the clinical utility of the models as experts affect model performance [8]. Therefore, secondary metrics were considered and co-registered whole mount histopathology was used to evaluate sensitivity and specificity.

Conclusion

In this work, a CNN for automated PCa segmentation was trained on ^{18}F-1007-PSMA PET images and tested on ^{18}F-1007-, ^{18}F-DCFPyL- and ^{68}Ga-PSMA-11-PET as well as histopathology reference. The CNN performed much faster (seconds versus minutes)

compared to expected physician contouring for all tracers and predicted the GTVs with a comparable performance across all tracers considered. Future studies will seek to validate our CNN to guide PSMA-PET based diagnostic and therapeutic procedures. The trained CNN will be publicly available.

Funding

This work was funded by the German Federal Ministry of Education and Research and (ERA PerMed – PersoRad).

Author contributions

All authors read and approved the final manuscript.
Project idea: CZ, JH.
Project management: JH, CZ.
Image annotations: CZ, JH, IM, SS, TN, PH, SC.
CNN development, training and testing: JH, AK, DK, TF, RG.
Providing patient collectives: TH, JK, AV, HI, PB, CG, JR, PD, AG.
Data analysis, data interpretation: JH, CZ, MM.
Drafting of manuscript: JH, CZ.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2023.109774.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–49.
[2] Kerkmeijer LGW, Groen VH, Pos FJ, Haustermans K, Monninkhof EM, Smeenk RJ, et al. Focal boost to the intraprostatic tumor in external beam radiotherapy for patients with localized prostate cancer: Results from the FLAME randomized phase III trial. J Clin Oncol March 2021;39:787–96.
[3] Zamboglou C, Spohn SKB, Ruf J, Benndorf M, Gainey M, Kamps M, Jilg C, Gratzke C, Adebahr S, Schmidtmayer-Zamboglou B, und others. PSMA-PET–and MRI–Based Focal Dose Escalated Radiation Therapy of Primary Prostate Cancer: Planned Safety Analysis of a Nonrandomized 2-Armed Phase 2 Trial (ARO2020-01),“ *International Journal of Radiation Oncology* Biology* Physics*, 2022.
[4] Spohn SKB, Kramer M, Kiefer S, Bronsert P, Sigle A, Schultze-Seemann W, Jilg CA, Sprave T, Ceci L, Fassbender TF, und others. Comparison of manual and semi-automatic [18F] PSMA-1007 PET based contouring techniques for intraprostatic tumor delineation in patients with primary prostate cancer and validation with histopathology as standard of reference,“ *Frontiers in oncology* 2020; Bd. 10:600690.
[5] Eiber M, Weirich G, Holzapfel K, Souvatzoglou M, Haller B, Rauscher I, et al. Simultaneous 68Ga-PSMA HBED-CC PET/MRI improves the localization of primary prostate cancer. Eur Urol 2016;70:829–36.
[6] Matkovic LA, Wang T, Lei Y, Akin-Akintayo OO, Ojo OAA, Akintayo AA, Roper J, Bradley JD, Liu T, Schuster DM, und others. Prostate and dominant intraprostatic lesion segmentation on PET/CT using cascaded regional-net,“ *Physics in Medicine & Biology* 2021; Bd. 66:245006.
[7] Kostyszyn D, Fechter T, Bartl N, Grosu AL, Gratzke C, Sigle A, Mix M, Ruf J, Fassbender TF, Kiefer S, und others. Intraprostatic tumor segmentation on PSMA PET images in patients with primary prostate cancer with a convolutional neural network,“ *Journal of Nuclear Medicine* 2021; Bd. 62:823–828.
[8] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019.
[9] Zamboglou C, Fassbender TF, Steffan L, Schiller F, Fechter T, Carles M, et al. Validation of different PSMA-PET/CT-based contouring techniques for intraprostatic tumor definition using histopathology as standard of reference. Radiother Oncol : J Eur Soc Therap Radiol Oncol December 2019;141:208–13.
[10] Salembier C, Villeirs G, De Bari B, Hoskin P, Pieters BR, Van Vulpen M, Khoo V, Henry A, Bossi A, De Meerleer G, und others. ESTRO ACROP consensus guideline on CT-and MRI-based taret volume delineation for primary radiation therapy of localized prostate cancer,“ *Radiotherapy and Oncology* 2018; Bd. 127:49–61.
[11] Zamboglou C, Carles M, Fechter T, Kiefer S, Reichel K, Fassbender TF, Bronsert P, Koeber G, Schilling O, Ruf J, und others. Radiomic features from PSMA PET for non-invasive intraprostatic tumor discrimination and characterization in patients with intermediate-and high-risk prostate cancer-a comparison study with histology reference,“ *Theranostics* 2019; Bd. 9:2595.
[12] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, O. Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation,“ in *International conference on medical image computing and computer-assisted intervention*, 2016.
[13] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.
[14] Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 1993;15:850–63.
[15] Zamboglou C, Kramer M, Kiefer S, Bronsert P, Ceci L, Sigle A, Schultze-Seemann W, Jilg CA, Sprave T, Fassbender TF, und others. The impact of the co-registration technique and analysis methodology in comparison studies between advanced imaging modalities and whole-mount-histology reference in primary prostate cancer,“ *Scientific reports* 2021; Bd. 11:1–10.
[16] Zhang L-L, Li W-C, Xu Z, Jiang N, Zang S-M, Xu L-W, et al. (68)Ga-PSMA PET/CT targeted biopsy for the diagnosis of clinically significant prostate cancer compared with transrectal ultrasound guided biopsy: a prospective randomized single-centre study. Eur J Nucl Med Mol Imaging February 2021;48:483–92.
[17] Spohn SKB, Bettermann AS, Bamberg F, Benndorf M, Mix M, Nicolay NH, Fechter T, Hölscher T, Grosu R, Chiti A, und others. Radiomics in prostate cancer imaging for a personalized treatment approach-current aspects of methodology and a systematic review on validated studies,“ *Theranostics* 2021; Bd. 11:8027.
[18] Zamboglou C, Spohn SKB, Adebahr S, Huber M, Kirste S, Sprave T, Gratzke C, Chen RC, Carl EG, Weber WA, und others. PSMA-PET/MRI-based focal dose escalation in patients with primary prostate Cancer treated with stereotactic body radiation therapy (HypoFocal-SBRT): study protocol of a randomized, multicentric phase III trial,“ *Cancers* 2021; Bd. 13:5795.
[19] Draulans C, Pos F, Smeenk RJ, Kerkmeijer L, Vogel WV, Nagarajah J, Janssen M, Mai C, Heijmink S, van der Leest M, und others. 68Ga-PSMA-11 PET, 18F-PSMA-1007 PET, and MRI for gross tumor volume delineation in primary prostate cancer: intermodality and intertracer variability,“ *Practical Radiation Oncology* 2021; Bd. 11:202–211.
[20] Giesel FL, Will L, Lawal I, Lengana T, Kratochwil C, Vorster M, Neels O, Reyneke F, Haberkon U, Kopka K, und others. Intraindividual comparison of 18F-PSMA-1007 and 18F-DCFPyL PET/CT in the prospective evaluation of patients with newly diagnosed prostate carcinoma: a pilot study,“ *Journal of Nuclear Medicine* 2018; Bd. 59:1076–1080.
[21] Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness,“ *arXiv preprint arXiv:1811.12231,* 2018.
[22] Marinescu IM, Spohn SKB, Kiefer S, Bronsert P, Ceci L, Holzschuh J, Sigle A, Jilg CA, Rühle A, Sprave T, und others. Intraindividual Comparison Between [18F] PSMA-1007 PET/CT and Multiparametric MRI for Radiotherapy Planning in Primary Prostate Cancer Patients,“ *Frontiers in Oncology* 2022; Bd. 12.
[23] Kuten J, Fahoum I, Savin Z, Shamni O, Gitstein G, Hershkovitz D, et al. Head-to-head comparison of 68Ga-PSMA-11 with 18F-PSMA-1007 PET/CT in staging prostate cancer using histopathology and immunohistochemical analysis as a reference standard. J Nucl Med 2020;61:527–32.
[24] Kesch C, Vinsensia M, Radtke JP, Schlemmer HP, Heller M, Ellert E, Holland-Letz T, Duensing S, Grabe N, Afshar-Oromieh A, und others. Intraindividual comparison of 18F-PSMA-1007 PET/CT, multiparametric MRI, and radical prostatectomy specimens in patients with primary prostate cancer: a retrospective, proof-of-concept study,“ *Journal of Nuclear Medicine* 2017; Bd. 58:1805–1810.
[25] Zamboglou C, Bettermann AS, Gratzke C, Mix M, Ruf J, Kiefer S, et al. Uncovering the invisible-prevalence, characteristics, and radiomics feature-based detection of visually undetectable intraprostatic tumor lesions in (68) GaPSMA-11 PET images of patients with primary prostate cancer. Eur J Nucl Med Mol Imaging June 2021;48:1987–97.
[26] Trägårdh E, Simoulis A, Bjartell A, Jögi J. Tumor detection of 18F-PSMA-1007 in the prostate gland in patients with prostate cancer using prostatectomy specimens as reference method. J Nucl Med 2021;62:1735–40.
[27] Reinke A, Eisenmann M, Tizabi MD, Sudre CH, Rädsch T, Antonelli M, Arbel T, Bakas S, Cardoso MJ, Cheplygina V, und others. Common limitations of image processing metrics: A picture story,“ *arXiv preprint arXiv:2104.05642,* 2021.
[28] Maier-Hein L, Reinke A, Christodoulou E, Glocker B, Godau P, Isensee F, Kleesiek J, Kozubek M, Reyes M, Riegler MA, und others. Metrics reloaded: Pitfalls and recommendations for image analysis validation,“ *arXiv preprint arXiv:2206.01653,* 2022.

[29] Cardinale J, Martin R, Remde Y, Schäfer M, Hienzsch A, Hübner S, Zerges AM, Marx H, Hesse R, Weber K, und others. Procedures for the GMP-compliant production and quality control of [18F] PSMA-1007: a next generation radiofluorinated tracer for the detection of prostate cancer,'' *Pharmaceuticals* 2017; Bd. 10:77.

[30] Hoberück S, Löck S, Borkowetz A, Sommer U, Winzer R, Zöphel K, et al. Intraindividual comparison of [(68) Ga]-Ga-PSMA-11 and [(18)F]-F-PSMA-1007 in prostate cancer patients: a retrospective single-center analysis. EJNMMI Res October 2021;11:109.

[31] Vrachimis A, Ferentinos K, Demetriou E, Ioannides C, Zamboglou N. PET/CT imaging of prostate cancer in the era of small molecule prostate specific membrane antigen targeted tracers. Hell J Nucl Med September 2020;23:339–45.

[32] Tsechelidis I, Vrachimis A. PSMA PET in Imaging Prostate Cancer. Front Oncol 2022;12:831429.

[34] Bettermann AS, Zamboglou C, Kiefer S, Jilg CA, Spohn S, Kranz-Rudolph J, Fassbender TF, Bronsert P, Nicolay NH, Gratzke C, und others. [68Ga-] PSMA-11 PET/CT and multiparametric MRI for gross tumor volume delineation in a slice by slice analysis with whole mount histopathology as a reference standard–Implications for focal radiotherapy planning in primary prostate cancer'', *Radiotherapy and Oncology* 2019; Bd. 141:214–219.

Further Reading

[33] Wang T, Yang L, Schreibmann E, Roper J, Schuster DM, Bradley JD, Liu T, Jani AB, Yang X. Deep-learning-based extraprostatic nodal lesion segmentation on 18F-fluciclovine PET,'' in *Medical Imaging 2022: Biomedical Applications in Molecular, Structural, and Functional Imaging*, 2022.

[35] Hosny A, Bitterman DS, Guthier CV, Qian JM, Roberts H, Perni S, Saraf A, Peng LC, Pashtan I, Ye Z, und others. Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study,'' *The Lancet Digital Health* 2022; Bd. 4:e657–e666.

[36] Leung K, Ashrafinia S, Salehi Sadaghiani M, Dalaie P, Tulbah R, Yin Y, VanDenBerg R, Leal J, Gorin M, Du Y, Pomper M, Rowe S, Rahmim A. A fully automated deep-learning based method for lesion segmentation in 18F-DCFPyL PSMA PET images of patients with prostate cancer,'' *Journal of Nuclear Medicine* 2019; Bd. 60:399–399.

[37] Mortensen MA, Borrelli P, Poulsen MH, Gerke O, Enqvist O, Ulén J, Trägårdh E, Constantinescu C, Edenbrandt L, Lund L, und others. Artificial intelligence-based versus manual assessment of prostate cancer in the prostate gland: a method comparison study,'' *Clinical physiology and functional imaging* 2019; Bd. 39:399–406.