SCIENTIFIC REPORTS
nature research

OPEN

Prostate Cancer Nodal Staging: Using Deep Learning to Predict ^{68}Ga-PSMA-Positivity from CT Imaging Alone

A. Hartenstein[1], F. Lübbe[1], A. D. J. Baur[1], M. M. Rudolph[1], C. Furth[2], W. Brenner[2], H. Amthauer[2], B. Hamm[1], M. Makowski[1,4,5] & T. Penzkofer[1,3,5*]

Lymphatic spread determines treatment decisions in prostate cancer (PCa) patients. 68Ga-PSMA-PET/CT can be performed, although cost remains high and availability is limited. Therefore, computed tomography (CT) continues to be the most used modality for PCa staging. We assessed if convolutional neural networks (CNNs) can be trained to determine 68Ga-PSMA-PET/CT-lymph node status from CT alone. In 549 patients with 68Ga-PSMA PET/CT imaging, 2616 lymph nodes were segmented. Using PET as a reference standard, three CNNs were trained. Training sets balanced for infiltration status, lymph node location and additionally, masked images, were used for training. CNNs were evaluated using a separate test set and performance was compared to radiologists' assessments and random forest classifiers. Heatmaps maps were used to identify the performance determining image regions. The CNNs performed with an Area-Under-the-Curve of 0.95 (status balanced) and 0.86 (location balanced, masked), compared to an AUC of 0.81 of experienced radiologists. Interestingly, CNNs used anatomical surroundings to increase their performance, "learning" the infiltration probabilities of anatomical locations. In conclusion, CNNs have the potential to build a well performing CT-based biomarker for lymph node metastases in PCa, with different types of class balancing strongly affecting CNN performance.

Prostate cancer (PCa) is the most common malignant cancer in men worldwide, and the second most common cause of cancer related death in men[1]. Patients with intermediate or high-risk PCa undergo regular staging examinations in order to determine if the tumor has spread beyond the prostate. As treatment success is highly dependent on the presence of systemic spread[2,3], staging procedures with high sensitivity and specificity are necessary.

Standard of care imaging for PCa staging typically includes contrast-enhanced computed tomography (CT) and Technetium-99m-methylene diphosphonate bone scans[4,5]. Despite the continued recommendation of CT in staging, it has been shown that predicting lymph node infiltration (LNI) with CT scans is not very reliable[6,7], with one study reporting a sensitivity and specificity of only 42% and 82%[8]. This low performance is most likely due to the limited morphological criteria used to define a lymph node as positive for infiltration, with size being the most relevant[9]. A threshold of 8–10 mm is often used despite the fact that 80% of lymph node metastases are less than 8 mm in the short axis[10]. Further criteria, such as status of hilum fat, nodal shape, and enhancement characteristics are used to aid diagnosis, but it remains difficult to exclude LNI in large benign hyperplastic nodes or detect it in small nodes below the size threshold[11].

In 2012 imaging agents binding to Prostate Specific Membrane Antigen (PSMA) were introduced, leading to the development of PSMA PET/CT[8]. PSMA, an integral membrane glycoprotein expressed 100–1000 fold on membranes of PCa cells compared to prostate cells, has been shown to correlate with aggressive disease, disease

[1]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Department of Radiology, Augustenburger Platz 1, 13353, Berlin, Germany. [2]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Department of Nuclear Medicine, Charitéplatz 1, 13353, Berlin, Germany. [3]Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Str. 2, 10178, Berlin, Germany. [4]Institute for Diagnostic and Interventional Radiology, Klinikum rechts der Isar der Technischen Universität München, Ismaninger Straße 22, D-81675, München, Germany. [5]These authors contributed equally: M. Makowski and T. Penzkofer. *email: tobias.penzkofer@charite.de

recurrence, and metastasis[12–14], and radio-tracer targeting of PSMA in conjunction with CT has been shown in a systematic review and meta-analysis of 5 studies to predict LNI with a sensitivity and specificity of 80% and 97% respectively[15]. PSMA PET/CT has been used to detect PCa in the prostate, soft tissue, and bone, and has been shown to detect LNI in nodes even under 10 mm in size, with one study reporting a 60% detection rate for nodes between 2–5 mm[16,17].

Even though PSMA PET/CT has proven to be very valuable in PCa staging, it remains of limited availability and hybrid imaging such as PET/CT is associated with high costs. The goal of this study was to evaluate – using 68Ga-PSMA PET/CT as a reference standard – if it is possible to elucidate the status of lymph nodes based on contrast-enhanced CT images alone using deep learning in the form of convolutional neural networks (CNNs).

Materials and Methods

Imaging datasets. Inclusion criteria for this retrospective study was the availability of a 68Ga-PSMA PET/CT examination with parallel contrast-enhanced CT examination performed between September 2013 and April 2017. All patients had histopathologically verified prostate cancer that warranted staging examinations. Exclusion criteria were non-contrast or low-dose only CT examination, insufficient image quality, and follow-up studies (only the first 68Ga-PSMA PET/CT of each patient was included). Of 738 patients, 549 patients (68.7 ± 7.54 [45–87] years, PSA 20.9 ± 94.6 [0–1423] ng/ml) fulfilled our inclusion criteria. The study was approved by the Charité Ethics Committee, and due to the retrospective design, the need for informed written consent was waived by the same review board, in accordance with institutional guidelines and regulations. The study was performed in accordance with the Declaration of Helsinki.

All patients had received 68Ga-PSMA PET/CT examinations for clinical purposes during the course of treatment. A standard 68Ge/68Ga generator (Eckert and Ziegler Radiopharma GmbH, Berlin, Germany) was used for 68Ga production, and PSMA- HBED-CC (ABX GmbH, Radeberg, Germany) labelling with 68Ga was performed according to the previously described method[18]. All PET/CT images were acquired using a Gemini Astonish TF 16 PET/CT scanner (Phillips Medical Systems, Best, The Netherlands) after intravenous injection of 68Ga-PSMA-HBED-CC[19] using 3-D acquisition mode for all PET scans.

Semi-automated manual three dimensional segmentation of lymph nodes was performed using the MITK software suite (MITK v. 2016.3.0, DKFZ, Heidelberg, Germany)[20]. Using the PSMA PET image as ground truth, a label of positive or negative for tumor infiltration was generated for each lymph node in consensus of two radiologists experienced in hybrid imaging, correlated with SUVmax. Figure 1 shows an example of a 68Ga-PSMA PET/CT full body scan and two selected lymph nodes, one positive and one negative for infiltration. In addition to the tumor infiltration label, the position of each lymph node in the body was manually assigned a categorical variable from a set of 9 possible categories (inguinal, iliacal (including obturator fossa), perirectal, (ascending) retroperitoneal, axillary, mediastinal, supra or infraclavicular, and cervical).

Patient collective and dataset generation. A final set of 549 patients fulfilled the inclusion criteria. An average of 4.72 ± 0.77 (SD) lymph nodes were segmented and labelled in each patient resulting in a total of 2,616 labelled lymph nodes, with 431 of these labelled as positive for infiltration. Figure 2 shows how these images were used to generate test and training datasets, and is explained as follows. A set of 130 lymph nodes was set aside for testing all CNNs and experts. This test set was created by taking 15% of the available positive nodes (65 nodes) and matching with 65 randomly selected negative nodes to create a 50:50 class balanced set. The remaining 366 positive nodes were matched with 366 randomly selected negative nodes to create a 50:50 class balanced set referred to as the 'status balanced' training set, with a total of 732 lymph nodes. The majority of lymph nodes in the status balanced and test dataset were in the inguinal region (32%), followed by the iliacal region (23%), and retroperitoneal region (19%). Figure 3a shows anatomical distribution by training set. To investigate effects of anatomical localization on classification results, the same 366 positive nodes used to create the status balanced set were sorted by anatomical category and matched to randomly selected negative nodes from within the same anatomical category, thus creating a 50:50 class balanced set with 548 lymph nodes, referred to as the 'location balanced' training set.

Neural network training. Images were resampled to an isotropic resolution of $1 \times 1 \times 1\,mm^3$. A volume of $80 \times 80 \times 80\,mm^3$ was cropped around the lymph node, centered at the center point of the manual lymph node segmentation. Image augmentation was performed online during model training, while only non-augmented images were provided to the model during validation and testing. A total of four random augmentations were performed: brightness was augmented by a factor between 0.5 and 1.5, after which images were rotated between ± 180 degrees, translated by a maximum of 5 voxels in the x, y and/or z axis, and finally flipped across the sagittal or axial plane or both. In order to ensure that no 'black borders' (i.e. areas with no image data due to rotation and shifting during augmentation) would be fed to the model, images were again cropped to a final volume of $48 \times 48 \times 48\,mm^3$. Finally, a single axial central slice was provided to the model as input.

Networks received two-dimensional images of 48×48 voxels and output a binary prediction whether or not the single lymph node displayed contained tumor or not. A final network architecture with 16 convolutional layers and three densely connected layers, inspired by the success of similar architectures by the Visual Geometry Group (VGGNet)[21], was selected using k-fold validation with $k = 10$. Figure 4 shows the architecture used by all CNNs. CNNs were not pre-trained. Batch normalization was performed after every layer, with rectified linear units (ReLU) used as the activation function. The output of the convolutional layers was fed to a fully connected feed forward network with 3 hidden layers. Adam optimization was used to update network weights[22], with parameters for alpha, beta1, beta2 and epsilon set at 0.0001, 0.9, 0.999 and 1e-08.

Three separate CNNs were trained. All models shared identical network architecture and were distinguished by the dataset used to train them: the status balanced model received status balanced CT images and

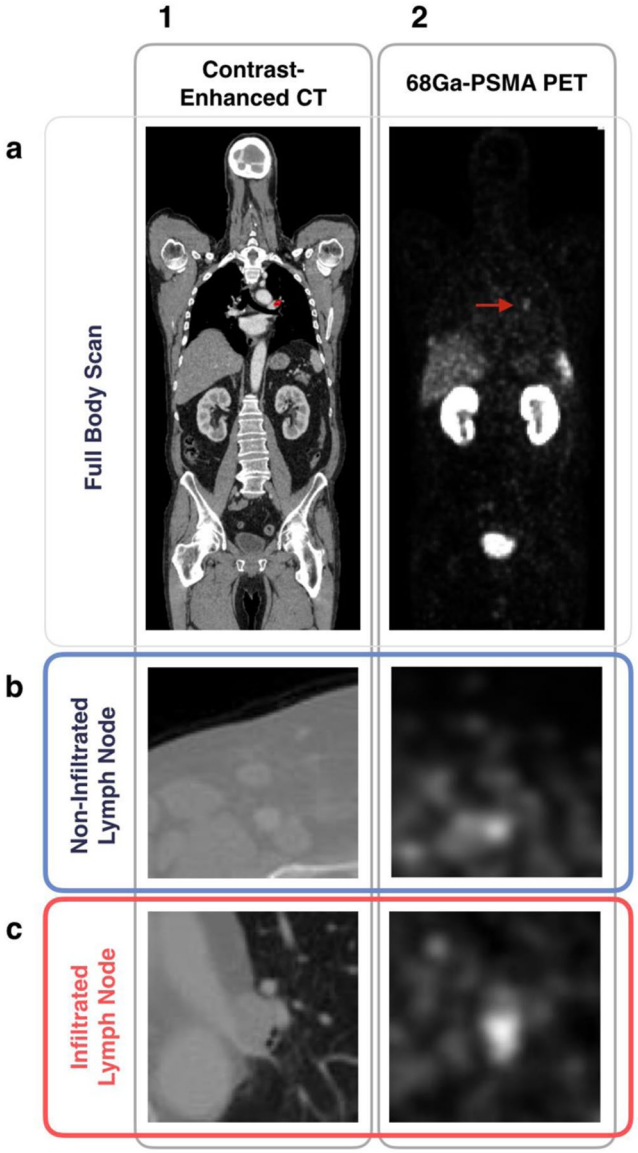

Figure 1. Generation of Labelled Dataset. (**a**) Imaging of a single patient with (1) a contrast-enhanced CT scan and (2) a 68Ga-PSMA PET scan. An average of 4.72 ± 0.77 lymph nodes were selected and semi-automatically segmented for each patient. A single lymph node positive for infiltration by PCa can be seen in the mediastinal region outlined in red in the CT image in (a1), and demarcated by a red arrow in PET scan in (a2). Using the 68Ga-PSMA PET/CT as our reference standard, a label for infiltration status by prostate cancer (either positive or negative) was assigned on a per lymph node basis. (**b**) An example of a negative 68Ga-PSMA PET/CT image pair in which the centered lymph node does not exceed background. (**c**) An example of a positive image pair.

segmentations, the location balanced model received location balanced CT images, and the xMask model received status balanced CT images multiplied by their corresponding segmentation mask. All models were implemented in Keras and Tensorflow (v. 1.10.1) and run on a Nvidia TITAN Xp graphics card (NVIDIA Titan Xp, Rev A1, Santa Clara, CA, United States). Heatmaps were generated using the Innvestigate (v. 1.0.2) package[23] using the PatternAttribution method[24].

Random forests. In order to validate neural network performance, random forests were generated to predict nodal infiltration status taking only nodal volume in mm^3 and nodal anatomical location into account. Two random forests were trained for each of the training sets used (status balanced, location balanced). Anatomical location was encoded as a one hot vector. Random forests were implemented using the sklearn python package[25] with maximum depth set at 5 to prevent overfitting to the training data.

Study readers. Two radiologists, with at least 5 years of experience in urogenital imaging, were presented with all test CT images ($n = 130$). Radiologists were presented an $80 \times 80 \times 80\,mm^3$ volume centered on the lymph node in question at $1 \times 1 \times 1\,mm^3$ resolution, and were asked to categorize the likelihood of lymph node

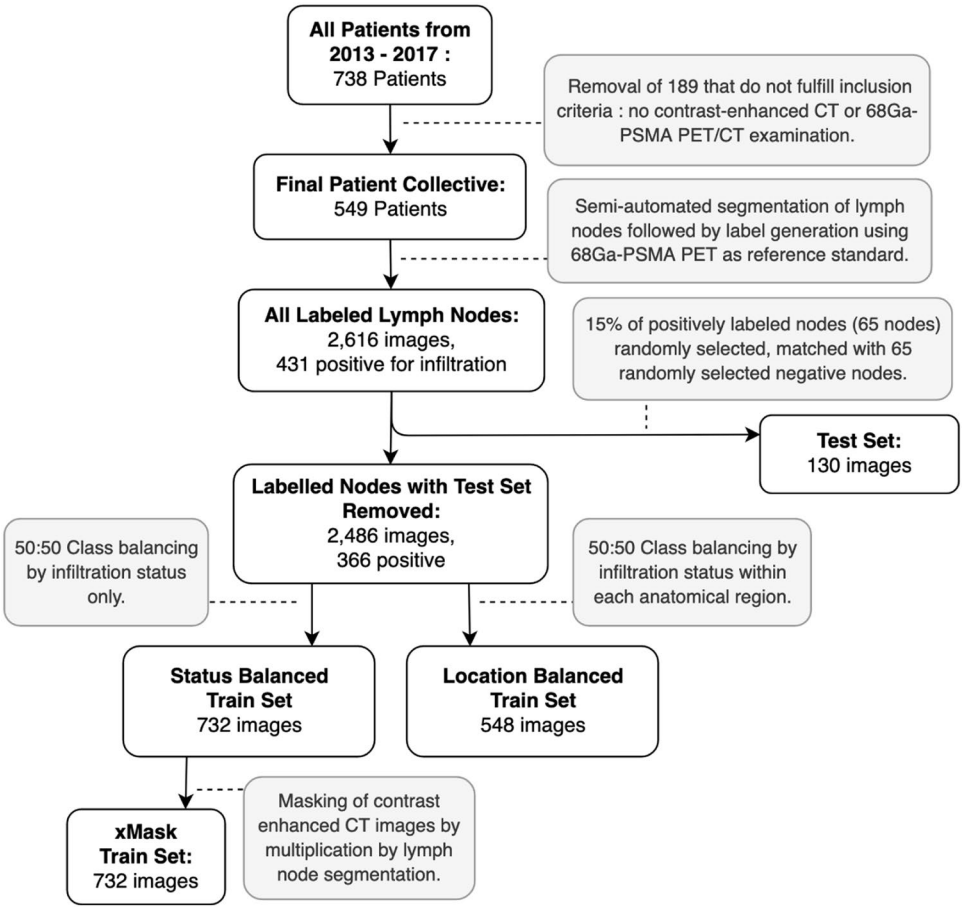

Figure 2. Dataset Generation Flowchart. Diagram describing generation of train and test datasets. Three train datasets shown, (status balanced, location balanced and xMask) were used to train three distinct neural networks. All neural networks and experts were tested and compared using a separate test set of 130 images, which was withheld from the neural networks during training. 50:50 class balancing was performed by taking all available infiltrated lymph nodes and randomly selecting an equally sized set of non-infiltrated lymph nodes, either from all available non-infiltrated nodes or from nodes within the same location category.

infiltration by tumor from the following four categories: very unlikely, unlikely, likely, and very likely. Neither the segmentation, 68Ga-PSMA PET/CT images, nor label were provided.

Statistical analysis. Model performance was evaluated for each CNN on the independent test set ($n = 130$) using the area under curve (AUC) of the receiver operating characteristic (ROC) curve. AUCs and confidence intervals were calculated using the pROC package in R[26], with confidence intervals computed using the bootstrap method with 10,000 stratified replicates. To allow for model comparison, the optimal threshold at which to consider CNN output as positive was set by maximizing Youden's index (sensitivity + specificity − 1), from which binary predictions were generated. Accuracy, sensitivity, specificity, PPV and NPV were calculated using the binary predictions. For study readers, the four categories were simplified to a dichotomous prediction of likely/unlikely. AUC for each radiologist is equivalent to the average of specificity and sensitivity[27]. McNemar's test was applied to all pairs of CNNs and experts. Results were considered statistically significant at a reduced $P < 0.005$ level to correct for multiple comparison. All variables are given as mean along with standard deviation and range where applicable.

Results

Evaluation of CNN classifiers and experts. The best performing Neural Network was trained using the status balanced training set, with an AUC of 0.955 (95% CI from 0.923–0.987). The CNNs trained with datasets where implicit frequency data was stripped using 50:50 class balancing by location category (the location balanced training set) or masking by the segmentation masks (xMask) performed comparably well, with an AUC of 0.858 (95% CI from 0.793–0.922) and 0.863 (95% CI from 0.804–0.923), respectively. Setting the sensitivity at 90% for all CNN models, the specificities of status balanced, location balanced, and xMask models was 88%, 52%, and 55%, respectively. Figure 5a shows ROC curves of all CNNs. Figure 5b shows histograms of CNN classification performance. Table 1 presents classification performance.

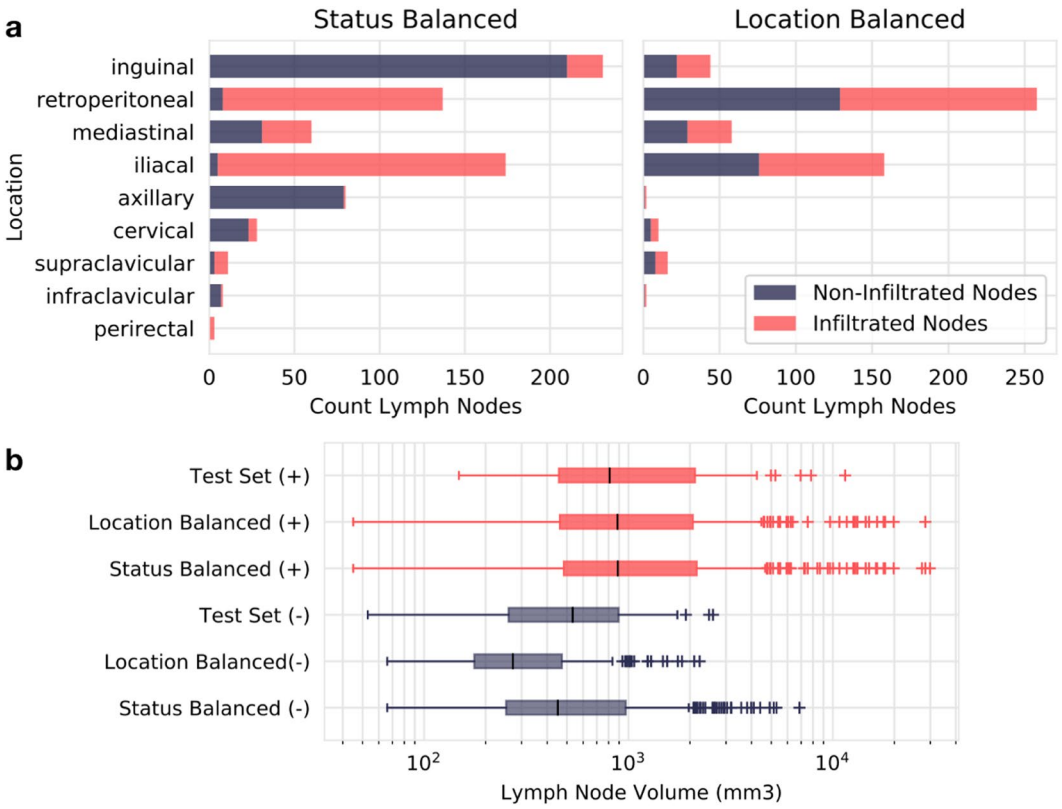

Figure 3. Dataset Regional and Volume Distributions. (**a**) The final distribution of lymph node images by location and infiltration status for the two training sets, referred to as 'status balanced' with 732 images and 'location balanced' with 548 images. (**b**) Boxplots depicting volume distribution for the location and status balanced training sets and test set grouped by infiltration status. Due to considerable overlap of the two distributions, size or volume is not a powerful indicator of infiltration.

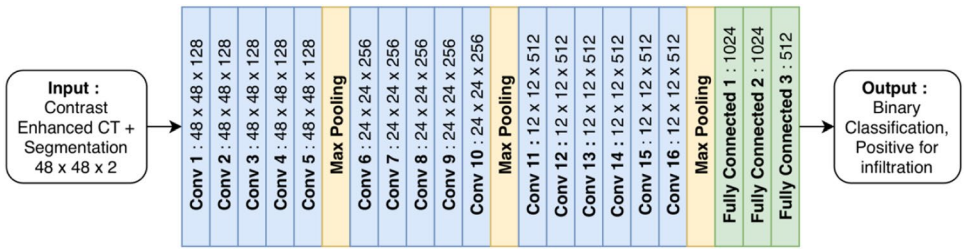

Figure 4. Convolutional neural network architecture. All three CNNs developed shared a common architecture and differed by the data used for training. CNNs received 2D contrast-enhanced CT images and segmentation masks as input, with input images augmented randomly during training. All convolutional layers used a kernel size of 3×3. A rectified linear unit (ReLU) activation function followed by batch normalization was performed at every layer. Adam optimization was used to update network weights, with parameters for alpha, beta1, beta2 and epsilon set at 0.0001, 0.9, 0.999 and 1e-08. Training was continued for 50 epochs.

The experienced uroradiologists achieved an average AUC, sensitivity, specificity and accuracy of 0.81, 65%, 96% and 81% respectively. The first radiologist performed with a calculated AUC of 0.86, while the second radiologist achieved a calculated AUC of 0.75. All differences in error rate between CNNs and expert readers was not statistically significant using McNemar's test and p set at a reduced 0.005.

The random forest trained with the status balanced training set achieved an AUC, sensitivity, specificity and accuracy of 0.900, 84%, 95% and 90% respectively on the test set. The random forest trained with the location balanced set performed significantly worse with an AUC, sensitivity, specificity and accuracy of 0.654, 70%, 60% and 65% respectively on the test set.

Use of heatmaps to explain differences in performance. Using heatmaps, we sought to elucidate how deep learning achieves a high classification performance. Examples of heatmaps are shown in Figs. 6 and 7. It appears that the CNNs are able to learn features within the lymph node and more surprisingly, outside the

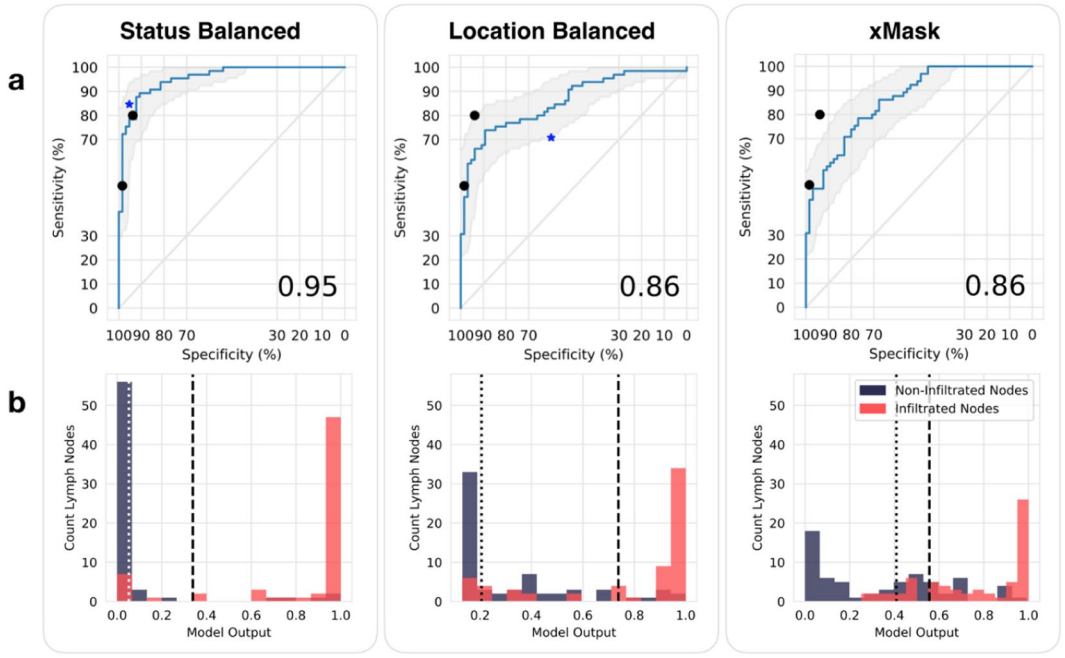

Figure 5. Classification performance. (**a**) Shown are the ROC curves for the three trained CNNs on the separate test set (n = 130) with 95% confidence interval of the sensitivity at given specificities in shaded gray. Displayed in the lower right hand corner is the corresponding AUC. Classification by individual radiologists on the same test set are displayed as black dots. Blue stars show random forest performance on the separate test set using the corresponding training dataset (status or location balanced). (**b**) Histograms of CNN model classification performance on the test set. The threshold that maximizes Youden's index is shown as a dashed line. The threshold which corresponds to a 90% sensitivity is shown as a dotted line. Infiltrated nodes (red bars) to the right of the given threshold are 'true positive', while those to the left are 'false negative': non-infiltrated nodes (blue) to the left are true negative, to the right are false positive.

Classifier	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1 Score
CNN: Status Balanced	0.95	89	86	92	91	86	88
CNN: Location Balanced	0.86	80	72	89	87	76	78
CNN: xMask	0.86	76	76	76	76	76	76
RF: Status Balanced	0.90	90	84	95	94	86	89
RF: Location Balanced	0.65	65	70	60	63	67	67
Expert 1	0.86	86	80	93	92	82	85
Expert 2	0.75	74	50	98	97	66	66

Table 1. Classification performance. Classification results are displayed in percentages. The optimal threshold for the three CNNs was selected by maximizing Youden's Index. RF: Random Forest.

boundaries of the lymph node (such as the aorta or air/skin borders), that correlate with lymph node infiltration status. It is critical to note that our best performing model, trained on status balanced data, appears to rely on features outside of the lymph node in question. This can be most clearly seen on images of inguinal or mediastinal lymph nodes, where areas of skin/air border (often found in the inguinal region) or lung/mediastinum border contribute heavily to final classification output, and the lymph node centered in the image is not highlighted. Heatmaps from the same CNN show that the lymph node itself is more important in true positive considerations, suggesting that 'inguinality', i.e. features of the inguinal region are important considerations in a negative infiltration status. Heatmaps generated can also be diffuse, with CNN attention displayed in many regions of the image but not particularly focused on the lymph node or surrounding region.

Discussion

In this study we trained and tested three CNNs that predict metastatic infiltration of lymph nodes by PCa using contrast-enhanced CT images and assessed their performance versus that of experienced human readers. The CNNs performed at the same level of two expert radiologists.

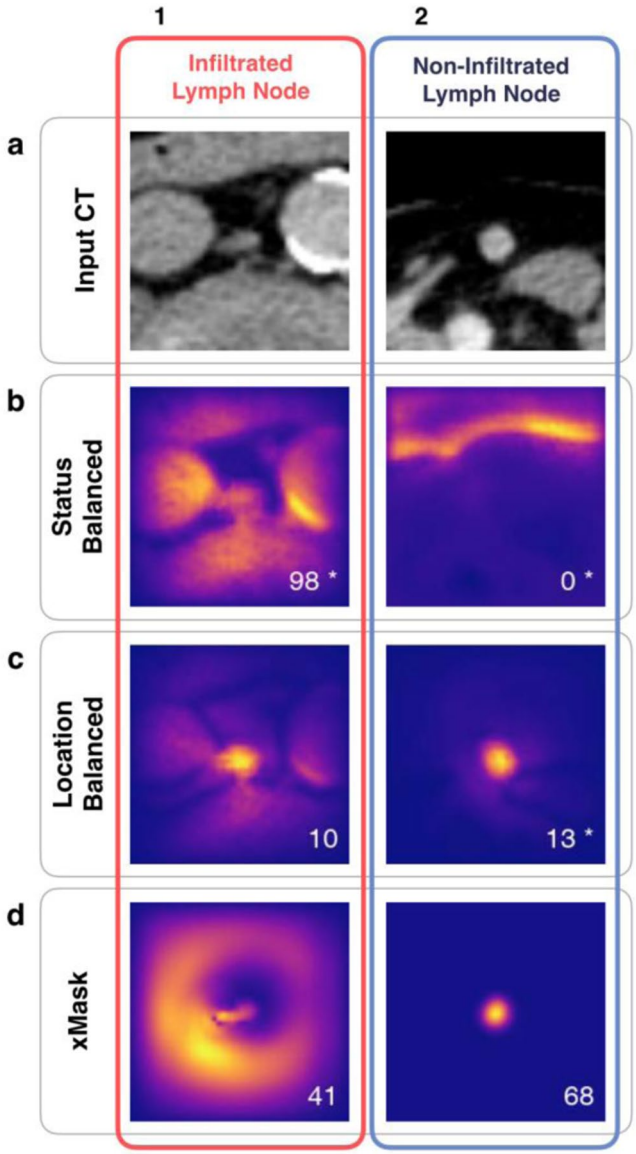

Figure 6. Heatmaps display neural network attention. (**a**) Contrast-enhanced CT images for two lymph nodes that were used as input to generate all heatmaps displayed, with (1) a retroperitoneal lymph node positive for infiltration by PCa, and (2) an inguinal lymph node negative for infiltration. In (**b–d**) heatmaps for the lymph nodes shown in (**a**), produced by three CNNs trained with status balanced training data, location balanced training data, or masked input data, respectively. CNN output, a pseudo probability score that the lymph node was classified as positive for tumor infiltration, is shown in the bottom right of each heatmap in (**b–d**). Stars signify true output predictions (either true positives for the lymph node in column 1 or true negative for column (2), with thresholds set by optimizing Youden's index for each CNN, set at 34, 73 and 54 for b,c, and d respectively. Within heatmaps, light colors represent areas that contribute to output prediction, while dark regions contribute little to output prediction. CNNs often highlight regions within the lymph node that expert radiologists recognize as important for infiltration status, such as nodal center density and contrast enhancement. In true positive images it appears that high central density is the most relevant parameter in designating a 'positive' label. We postulate that the 'halo' surrounding the lymph node in the xMask CNN (d1), depicts the CNN attention to size. Heatmaps produced by the CNN trained with status balanced data highlight anatomical regions which aid in classification of lymph nodes, often demarcating the air-skin border seen in images of inguinal lymph nodes, as in b2.

Current attempts at detecting lymph node metastases in PCa by radiological reading have been shown to be suboptimal, with a sensitivity and specificity shown in one study to be 42% and 82%, respectively[6–8]. Size is often the most relevant diagnostic criteria, with nodes greater than 10 mm deemed as suspicious and all below as benign[9,10]. Other criteria are difficult to quantify and are highly dependent on reader experience. Thus, the use of quantitative or algorithmic methods to detect LNI is desired. By radiomic analysis, in which a host of quantitative features are extracted from images and analyzed for statistical correlations, it has been suggested

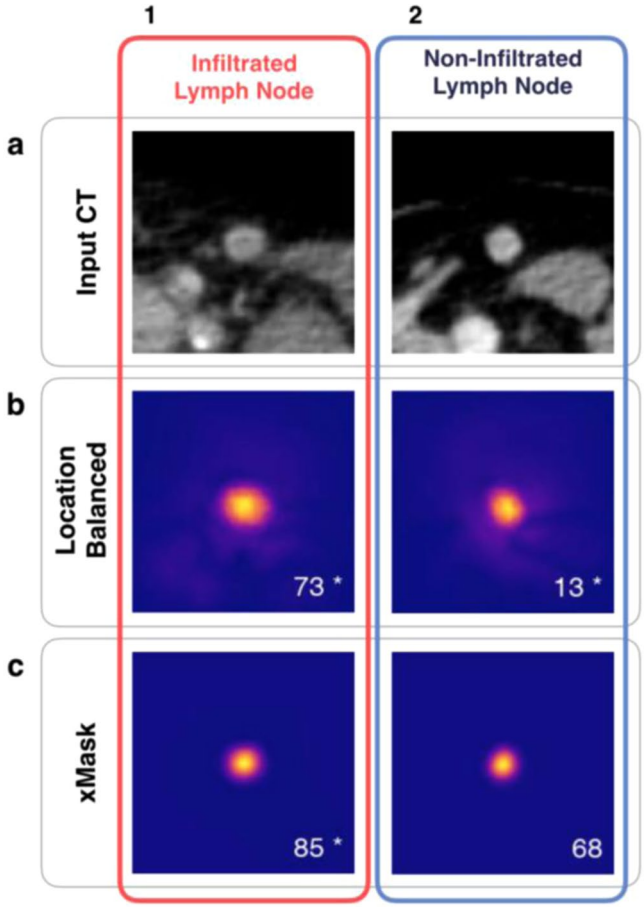

Figure 7. Limitations of heatmaps as tool to explain black box predictions. (**a**) Contrast-enhanced CT images for two inguinal lymph nodes that were used as input to generate all heatmaps displayed, with (1) a lymph node positive for infiltration by PCa, and (2) a lymph node negative for infiltration. In (**b,c**) heatmaps produced by two CNNs trained with location balanced training data, or masked input data, respectively. Beyond verifying that the lymph node is important for classification, heatmaps provide little additional information as to why classification output was either true positive (b1,c1), true negative (b2), or false positive (c2).

that a 7.5–Hounsfield CT density threshold could act as a surrogate parameter to differentiate LNI from benign processes[28], with 89% of non-infiltrated LNs below this threshold and 92% infiltrated LN above, though this study used many different cancer types and a mix of PET tracers as a standard of reference. Deep learning, in which optimization algorithms are used to train neural network models in classification tasks, have shown mixed success in detecting LNI. It has been previously found that CNNs are able to predict SUVmax in a PET scan using CT images of lymph nodes with a moderate accuracy, with an AUC of 0.85[29]. A number of studies predicting mediastinal LNI by lung cancer and breast cancer have been performed, with one study finding an AUC of 0.76[30] and another study classifying LNI in axillary lymph nodes by breast cancer achieving an AUC of 0.84[31]. CNNs were also found to classify head and neck tumor extranodal extension with an AUC of 0.91 using 3D CT images[32]. It has also been shown feasible to identify tumor infiltrated lymph nodes in MRI using deep learning[33]. To our knowledge, no study using deep learning to identify metastases of PCa into the lymphatic system by CT has been performed so far.

Generation of heatmaps is an attempt to explain how deep learning models reach classification decisions on a per-image basis, and represents a growing field of research known as 'explainability'. Each heatmap can be interpreted as displaying CNN attention; regions of an input image that influenced the classification decision are demarcated. From the heatmaps produced in our study, it becomes clear that identical CNN architectures learn different methods to solve the same problem, depending on which data is used for training. It appears that the CNN trained with status balanced data learned not only to recognize features of the lymph node in question, but also to recognize anatomical features surrounding the lymph node. Using these anatomical features, it appears that the status balanced CNN implicitly learned frequency of infiltration in different anatomical regions and used these frequencies or probabilities to improve output prediction. For example, in the status balanced dataset, 91% of inguinal lymph nodes were negative (see Fig. 3a). Thus, labeling all inguinal lymph nodes as negative is highly rewarded during the training process, and recognizing 'inguinality' aided in achieving high classification accuracy. Indeed, the air/skin border found in inguinal lymph nodes was often well demarcated in heatmaps, as seen in Fig. 6. However, it is unclear to what extent such anatomical features influenced classification; the CNN trained

with status balanced data did classify some inguinal lymph nodes as positive, and some retroperitoneal lymph nodes (of which 94% were positive in the status balanced dataset) as negative, as shown in Fig. 5b. The fact that learning anatomical features within the image (as proxy for anatomical location) greatly improves classification performance in the status balanced dataset is underscored by the high performance of the random forest trained on the this dataset; using nodal volume and location alone, high classification performance was achieved (AUC 0.90). Thus, our best performing neural network is most likely essentially useless on external datasets not sharing the anatomical bias found in the status balanced dataset.

We created two additional CNNs to eliminate anatomical clues within images in an attempt to force neural network attention to the lymph node. First, we created a new training dataset created by balancing positive and negative lymph nodes within each location category. By doing so we eliminated the possibility of learning infiltration frequency at each anatomical location. While it is clear from generated heatmaps that the CNN trained with this location balanced set did focus more on the lymph node and not on anatomical features, it was not able to achieve the same classification performance as the status trained CNN. However, a random forest receiving nodal volume and location information trained on this location balanced dataset performed poorly, considerably worse than the CNN (AUC 0.677 vs 0.858). This leads us to believe that the neural network is indeed focusing on features within the lymph node to perform classification. Secondly, a new CNN was provided images created by multiplying the CT image by the manually generated segmentation (xMask), thus setting all values outside of the lymph node to zero. This removed all contextual information, such as location in the body or presence of neighboring structures. The resulting performance was similar to the location balanced CNN. Interestingly, heatmaps created by the xMask CNN often showed a diffuse halo like pattern of attention outside of the lymph borders, which we postulate may be the CNNs attention to size. We cannot definitively state that any of the CNNs developed are able to determine nodal size due to intrinsic limitations of heatmaps as an explainability tool and the black box nature of neural networks, which often created very similar looking heatmaps (see Fig. 7). Regardless, size alone is a poor predictor of infiltration, as can be intuited by the considerable overlap of volume distributions for lymph nodes positive and negative for infiltration (see Fig. 3b) and shown quantitatively by the poor performance of the random forests trained with location balanced data.

There are a number of limitations to our study. It is important to note that the usage of PSMA PET/CT is an imperfect method of label generation. In comparison to the gold standard for detecting LN metastases, namely histopathological analysis after extended pelvic lymph node dissection (PLND)[34], PSMA PET/CT was found to have a sensitivity of 80% and specificity of 97% in a systematic review and meta-analysis[15,16,35,36]. Due to the high specificity, it is unlikely that our models were trained with large numbers of false positive lymph nodes. In addition, we relied on manual detection of segmentation of lymph nodes, and we do not perform lymph node detection. The tendency to select easily definable and large lymph nodes for analysis led to a large amount of inguinal lymph nodes being included in our dataset, a limitation we sought to overcome by various means of class balancing.

The obvious attention to anatomical features demonstrated by our best performing CNN raises a number of issues in the implementation of deep learning in the medical field. Deep learning models are able to learn frequencies and summary statistics, known as biases, within datasets, which can lead to high classification performance based upon undesirable features. This problem is distinct from overfitting to the training dataset, and instead points to the need for a more rigorous explainability of deep learning models. Our results represent a moderate success in the use of saliency maps (heatmaps), as through this instance-based analysis of CNN attention, we were able to determine that our best performing model was using anatomical features of the lymph node environment in addition to features within the lymph node. We were able to compensate for anatomical variations in infiltration frequency because we had collected coarse data on anatomical location. However, not only does class balancing at ever higher levels of abstraction encroach on the notion of 'automated feature generation', it is not feasible in the medical field due to lack of knowledge of what constitutes a relevant category. The lack of explainability methods for deep learning models is also a limitation. Our use of heatmaps, known as an attribution method, of which there are several, is problematic not just because of inconsistencies in implementation and performance[24], but the underpinning assumption that individual pixels in an input image should be the primary unit of relevance for classification.

Current deep learning systems can perform remarkably well and will most likely continue to improve with larger datasets and access to more contextual information, such as blood serum values and genomic data. Our results show that CNNs are capable of classifying lymphatic infiltration by PCa on contrast-enhanced CT scans alone as compared to the 68Ga-PSMA PET/CT reference standard. Anatomical context influences the performance of CNNs and should be carefully considered when building such imaging based biomarkers.

Received: 29 July 2019; Accepted: 11 February 2020;
Published online: 25 February 2020

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J. Clin.* **68**, 7–30, https://doi.org/10.3322/caac.21442 (2018).
2. Oderda, M., Joniau, S., Spahn, M. & Gontero, P. Debulking surgery in the setting of very high-risk prostate cancer scenarios. *BJU Int.* **110**, E192–E198, https://doi.org/10.1111/j.1464-410X.2012.10942.x (2012).
3. Luchini, C. *et al.* Extranodal extension of lymph node metastasis influences recurrence in prostate cancer: a systematic review and meta-analysis. *Sci. Rep.* **7**, 2374, https://doi.org/10.1038/s41598-017-02577-4 (2017).
4. Carroll, P. R. *et al.* NCCN Guidelines Insights: Prostate Cancer Early Detection, Version 2.2016. *J. Natl Compr. Canc Netw.* **14**, 509–519 (2016).
5. Mottet, N. *et al.* EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur. Urol.* **71**, 618–629, https://doi.org/10.1016/j.eururo.2016.08.003 (2017).

6. Engeler, C. E., Wasserman, N. F. & Zhang, G. Preoperative assessment of prostatic carcinoma by computerized tomography: Weaknessnes and new perspectives. *Urol.* **40**, 346–350, https://doi.org/10.1016/0090-4295(92)90386-B (1992).
7. Flanigan, R. C. *et al.* Limited efficacy of preoperative computed tomographic scanning for the evaluation of lymph node metastasis in patients before radical prostatectomy. *Urol.* **48**, 428–432, https://doi.org/10.1016/S0090-4295(96)00161-6 (1996).
8. Hövels, A. M. *et al.* The diagnostic accuracy of CT and MRI in the staging of pelvic lymph nodes in patients with prostate cancer: a meta-analysis. *Clin. Radiol.* **63**, 387–395, https://doi.org/10.1016/j.crad.2007.05.022 (2008).
9. Maurer, T. *et al.* Diagnostic Efficacy of (68)Gallium-PSMA Positron Emission Tomography Compared to Conventional Imaging for Lymph Node Staging of 130 Consecutive Patients with Intermediate to High Risk Prostate Cancer. *J. Urol.* **195**, 1436–1443, https://doi.org/10.1016/j.juro.2015.12.025 (2016).
10. Heesakkers, R. A. M. *et al.* MRI with a lymph-node-specific contrast agent as an alternative to CT scan and lymph-node dissection in patients with prostate cancer: a prospective multicohort study. *Lancet Oncol.* **9**, 850–856, https://doi.org/10.1016/S1470-2045(08)70203-1 (2008).
11. Gillessen, S. *et al.* Management of Patients with Advanced Prostate Cancer: The Report of the Advanced Prostate Cancer Consensus Conference APCCC 2017. *Eur. Urol.* **73**, 178–211, https://doi.org/10.1016/j.eururo.2017.06.002 (2018).
12. Silver, D. A., Pellicer, I., Fair, W. R., Heston, W. D. & Cordon-Cardo, C. Prostate-specific membrane antigen expression in normal and malignant human tissues. *Clin. Cancer Res.* **3**, 81–85 (1997).
13. Bostwick, D. G., Pacelli, A., Blute, M., Roche, P. & Murphy, G. P. Prostate specific membrane antigen expression in prostatic intraepithelial neoplasia and adenocarcinoma. *Cancer* **82**, 2256–2261, 10.1002/(SICI)1097-0142(19980601)82:11<2256::AID-CNCR22>3.0.CO;2-S (1998).
14. Perner, S. *et al.* Prostate-specific membrane antigen expression as a predictor of prostate cancer progression. *Hum. Pathol.* **38**, 696–701, https://doi.org/10.1016/j.humpath.2006.11.012 (2007).
15. Perera, M. *et al.* Gallium-68 Prostate-specific Membrane Antigen Positron Emission Tomography in Advanced Prostate Cancer—Updated Diagnostic Utility, Sensitivity, Specificity, and Distribution of Prostate-specific Membrane Antigen-avid Lesions: A Systematic Review and Meta-analysis. *Eur. Urol.* https://doi.org/10.1016/j.eururo.2019.01.049 (2019).
16. Leeuwen, P. J. V. *et al.* Prospective evaluation of 68Gallium-prostate-specific membrane antigen positron emission tomography/computed tomography for preoperative lymph node staging in prostate cancer. *BJU Int.* **119**, 209–215, https://doi.org/10.1111/bju.13540 (2017).
17. Hofman, M. S., Hicks, R. J., Maurer, T. & Eiber, M. Prostate-specific Membrane Antigen PET: Clinical Utility in Prostate Cancer, Normal Patterns, Pearls, and Pitfalls. *Radiographics* **38**, 200–217, https://doi.org/10.1148/rg.2018170108 (2018).
18. Afshar-Oromieh, A. *et al.* PET imaging with a [68Ga]gallium-labelled PSMA ligand for the diagnosis of prostate cancer: biodistribution in humans and first evaluation of tumour lesions. *Eur. J. Nucl. Med. Mol. Imaging* **40**, 486–495, https://doi.org/10.1007/s00259-012-2298-2 (2013).
19. Surti, S. *et al.* Performance of Philips Gemini TF PET/CT scanner with special consideration for its time-of-flight imaging capabilities. *J. Nucl. Med.* **48**, 471–480 (2007).
20. Wolf, I. *et al.* The medical imaging interaction toolkit. *Med. Image Anal.* **9**, 594–604, https://doi.org/10.1016/j.media.2005.04.005 (2005).
21. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (2014).
22. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2014).
23. Alber, M. *et al.* iNNvestigate neural networks! *arXiv:1808.04260 [cs, stat]* (2018).
24. Kindermans, P.-J. *et al.* Learning how to explain neural networks: PatternNet and PatternAttribution. *arXiv:1705.05598 [cs, stat]* (2017).
25. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
26. pROC: Display and Analyze ROC Curves v. 1.13.0 (2018).
27. Haenssle, H. A. *et al.* Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **29**, 1836–1842, https://doi.org/10.1093/annonc/mdy166 (2018).
28. Giesel, F. L. *et al.* Correlation Between SUVmax and CT Radiomic Analysis Using Lymph Node Density in PET/CT-Based Lymph Node Staging. *J. Nucl. Med.* **58**, 282–287, https://doi.org/10.2967/jnumed.116.179648 (2017).
29. Shaish, H. *et al.* Prediction of Lymph Node Maximum Standardized Uptake Value in Patients With Cancer Using a 3D Convolutional Neural Network: A Proof-of-Concept Study. *American Journal of Roentgenology*, 1–7, https://doi.org/10.2214/AJR.18.20094 (2018).
30. Beig, N. *et al.* Perinodular and Intranodular Radiomic Features on Lung CT Images Distinguish Adenocarcinomas from Granulomas. *Radiology*, 180910, https://doi.org/10.1148/radiol.2018180910 (2018).
31. Ha, R. *et al.* Axillary Lymph Node Evaluation Utilizing Convolutional Neural Networks Using MRI Dataset. *J Digit Imaging*, https://doi.org/10.1007/s10278-018-0086-7 (2018).
32. Kann, B. H. *et al.* Pretreatment Identification of Head and Neck Cancer Nodal Metastasis and Extranodal Extension Using Deep Learning Neural Networks. *Sci. Rep.* **8**, 14036, https://doi.org/10.1038/s41598-018-32441-y (2018).
33. Lu, Y. *et al.* Identification of Metastatic Lymph Nodes in MR Imaging with Faster Region-Based Convolutional Neural Networks. *Cancer Res.* **78**, 5135–5143, https://doi.org/10.1158/0008-5472.CAN-18-0494 (2018).
34. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part II: Treatment of Relapsing, Metastatic, and Castration-Resistant Prostate Cancer. *Eur. Urol.* **71**, 630–642, https://doi.org/10.1016/j.eururo.2016.08.002 (2017).
35. Hijazi, S. *et al.* Pelvic lymph node dissection for nodal oligometastatic prostate cancer detected by 68Ga-PSMA-positron emission tomography/computerized tomography. *Prostate* **75**, 1934–1940, https://doi.org/10.1002/pros.23091 (2015).
36. Jilg, C. A. *et al.* Diagnostic Accuracy of Ga-68-HBED-CC-PSMA-Ligand-PET/CT before Salvage Lymph Node Dissection for Recurrent Prostate Cancer. *Theranostics* **7**, 1770, https://doi.org/10.7150/thno.18421 (2017).

Acknowledgements

We acknowledge support from the German Research Foundation (DFG) and the Open Access Publication Funds of Charité – Universitätsmedizin Berlin.

Author contributions

All authors contributed to the study design. A.H., F.L., A.B., M.M.R., C.F., H.A., W.B. collected and curated the data. A.H. and T.P. wrote the manuscript. All authors critically revised and reviewed the manuscript.

Competing interests

A.H., F.L., C.F., M.M., W.B. and M.M.R. declare no competing interest with respect to the relation to the work described. A.B. received fee as a speaker for Bayer, and Bender Gruppe, both outside of the current work. H.A. reports grants from Sirtex Medical, Bayer and GE Healthcare; lecture and/or travel fees from Sirtex Medical, GE Healthcare, Novartis, Eisai and Terumo. All outside the submitted work. B.H. declares the following competing interests: Grant money from the following companies or nonprofit organizations to the Dept of Radiology: Abbott, AbbVie, Ablative Solutions, Accovion, Achaogen Inc., Actelion Pharmaceuticals, ADIR,

Aesculap, AGO, AIF Arbeitsgemeinschaft industrieller Forschungsvereinigungen, AIO: Arbeitsgemeinschaft Internistische Onkologie, Alexion Pharmaceuticals, Amgen, AO Foundation, Arena Pharmaceuticals, ARMO Biosciences, Inc., art photonics GmbG Berlin, ASR Advanced sleep research, Astellas, AstraZeneca, BARD, Bayer Healthcare, Bayer Schering Pharma, Bayer Vital, BBraun (Sponsoring a workshop), Berlin-Brandenburger Centrum für Regenerative Therapien (BCRT), Berliner Krebsgesellschaft, Biotronik, Bioven, BMBF, Boehring Ingelheimer, Boston Biomedical Inc., BRACCO Group, Brainsgate, Bristol-Myers Squibb, Cascadian Therapeutics, Inc., Celgene, CELLACT Pharma, Celldex, Therapeutics, CeloNova BioSciences, Charité research organisatin GmbH, Chiltern, CLOVIS ONCOLOGY, INC., Covance, CUBIST, CureVac AG, Tübingen, Curis, Daiichi, DC Devices, Inc. USA, Delcath Systems, Dermira Inc. Deutsche Krebshilfe, Deutsche Rheuma Liga, DFG, DSM Nutritional Products AG, Dt. Stiftung für Herzforschung, Dynavax, Eisai Ltd., European Knowledge Centre, Mosquito Way, Hatfield, Eli Lilly and Company Ltd. EORTC, Epizyme, INC., Essex Pharma, EU Programmes, Euroscreen S.A., Fibrex Medical Inc., Foxused Ultrasound Surgery Foundation, Fraunhofer Gesellschaft, Galena Biopharma, Galmed Research and Development Ltd., Ganymed, GE, Genentech Inc., GETNE (Grupo Espanol de Tumores Neuroendocrinos), Gilead Sciences, Inc, Glaxo Smith Kline, Glycotope GmbH, Berlin, Goethe Uni Frankfurt, Guerbet, Guidant Europe NV, Halozyme, Hewlett Packard GmbH, Holaira Inc. ICON (CRO), Idera Pharmaceuticals, Inc., Ignyta, Inc. Immunomedics Inc., Immunocore, Incyte, INC Research, Innate Pharma, InSightec Ltd., Inspiremd, inVentiv Health Clinical UK Ltd., Inventivhealth, IOMEDICO, IONIS, IPSEN Pharma, IQVIA, ISA Therapeutics, Isis Pharmaceuticals Inc., ITM Solucin GmbH, Jansen, Kantar Health Gmbh (CRO), Kartos Therapeutics, Inc., Karyopharm Therapeutics, Inc., Kendle/MorphoSys Ag, Kite Pharma, Kli Fo Berlin Mitte, La Roche, Land Berlin, Lilly GmbH, Lion Biotechnology, Lombard Medical, Loxo Oncology, Inc., LSK BioPartners; USA, Lundbeck GmbH, Lux Biosciences, LYSARC, MacroGenics, MagForce, MedImmune Inc., Medpace, Medpace Germany GmbH (CRO), MedPass (CRO), Medronic, Merck, Merromack Pharmaceuticals Inc., MeVis Medical Solutions AG, Millennium Pharmaceuticals Inc., Mologen, Monika Kutzner Stiftung, MSD Sharp, NeoVacs SA, Newlink Genentics Corporation, Nexus Oncology, NIH, Novartis, novocure, Nuvisan, Ockham oncology, OHIRC Kanada, Orion Corporation Orion Pharma, Parexel CRO Service, Perceptive, Pfizer GmbH, Pharma Mar, Pharmaceutical Research Associates GmbH (PRA), Pharmacyclics Inc., Philipps, PIQUR Therapeutics Ltd., Pluristem, PneumRX.Inc, Portola Pharmaceuticals, PPD (CRO), PRAint, Premier-research, Provectus Biopharmaceuticals, Inc., psi-cro, Pulmonx International Sàrl, Quintiles GmbH, Regeneron Pharmaceuticals, Inc., Respicardia, Roche, Samsung, Sanofi, sanofis-aventis S.A., Schumacher GmbH (Sponsoring a workshop), Seattle Genetics, Servier (CRO), SGS Life Science Services (CRO), Shore Human Genetic Therapies, Siemens, Silena Therapeutics, Spectranetics GmbH, Spectrum Pharmaceuticals, St. Jude Medical, Stiftung Wolfgang Schulze, Symphogen, Taiho Oncology, Inc., Taiho Pharmaceutical Co., TauRx Therapeutics Ltd., Terumo Medical Corporation, Tesaro, tetec-ag, TEVA, Theorem, Theradex, Threshold Pharmaceuticals Inc., TNS Healthcare GmbH, Toshiba, UCB Pharma, Uni München, VDI/VDE, Vertex Pharmaceuticals Incorporated, winicker-norimed, Wyeth Pharma, Xcovery Holding Company, Zukunftsfond Berlin (TSB). TP receives grant support from the Berlin Institute of Health within the Clinician Scientist Programme. TP declares no additional conflict of interest with respect to the relation to the work described. Outside of the current work there are institutional relationship with the following entities (no personal payments to TP): research support from Siemens Healthcare and Philips Healthcare, clinical trials with AGO, Aprea AB, Astellas Pharma Global Inc., AstraZeneca, Celgene, Genmab A/S, Incyte Coporation, Lion Biotechnologies, Inc., Millennium Pharmaceuticals, Inc., Morphotec Inc., MSD, Tesaro Inc., and Roche.

Additional information

Correspondence and requests for materials should be addressed to T.P.