Abdominal Radiology (2019) 44:2030–2039
https://doi.org/10.1007/s00261-018-1824-5

SPECIAL SECTION: RADIOGENOMICS

Deep transfer learning-based prostate cancer classification using 3 Tesla multi-parametric MRI

Xinran Zhong[1,2] · Ruiming Cao[1,3] · Sepideh Shakeri[1] · Fabien Scalzo[4] · Yeejin Lee[1] · Dieter R. Enzmann[1] · Holden H. Wu[1,2] · Steven S. Raman[1] · Kyunghyun Sung[1,2]

Published online: 20 November 2018

Abstract

Purpose The purpose of the study was to propose a deep transfer learning (DTL)-based model to distinguish indolent from clinically significant prostate cancer (PCa) lesions and to compare the DTL-based model with a deep learning (DL) model without transfer learning and PIRADS v2 score on 3 Tesla multi-parametric MRI (3T mp-MRI) with whole-mount histopathology (WMHP) validation.

Methods With IRB approval, 140 patients with 3T mp-MRI and WMHP comprised the study cohort. The DTL-based model was trained on 169 lesions in 110 arbitrarily selected patients and tested on the remaining 47 lesions in 30 patients. We compared the DTL-based model with the same DL model architecture trained from scratch and the classification based on PIRADS v2 score with a threshold of 4 using accuracy, sensitivity, specificity, and area under curve (AUC). Bootstrapping with 2000 resamples was performed to estimate the 95% confidence interval (CI) for AUC.

Results After training on 169 lesions in 110 patients, the AUC of discriminating indolent from clinically significant PCa lesions of the DTL-based model, DL model without transfer learning and PIRADS v2 score $\geq$ 4 were 0.726 (CI [0.575, 0.876]), 0.687 (CI [0.532, 0.843]), and 0.711 (CI [0.575, 0.847]), respectively, in the testing set. The DTL-based model achieved higher AUC compared to the DL model without transfer learning and PIRADS v2 score $\geq$ 4 in discriminating clinically significant lesions in the testing set.

Conclusion The DeLong test indicated that the DTL-based model achieved comparable AUC compared to the classification based on PIRADS v2 score ($p = 0.89$).

Keywords Multi-parametric MRI · Clinically significant lesion classification · Prostate cancer · Whole-mount histopathology · PIRADS v2 score · Deep learning

Introduction

Prostate cancer (PCa) is the most common solid organ malignancy and the second leading cause of cancer-related deaths in men in the United States [1]. Detection and grading of PCa are important for patient prognosis and treatment since small low-grade lesions may undergo observation, whereas larger, higher grade lesions would be treated aggressively with surgery or radiation [2]. The histological primary or secondary Gleason grade of 4 has generally been shown to be predictive of less favorable outcomes such as increased rates of cancer progression and mortality [3]. As a result, a lesion with Gleason Score (GS) $\geq$ 7 is defined as clinically significant lesion with higher rates of adverse outcomes [4].

✉ Xinran Zhong
XZhong@mednet.ucla.edu

[1] Department of Radiological Sciences, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

[2] Physics and Biology in Medicine IDP, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

[3] Department of Computer Science, School of Engineering, University of California, Los Angeles, Los Angeles, CA, USA

[4] Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

For the past 25 years, the standard diagnostic method of PCa diagnosis has been elevation of prostate-specific antigen (PSA) followed by transrectal ultrasound (TRUS)-guided biopsy, which has resulted in decreased PCa mortality by 20–30% [5], but with significant diagnostic errors in undersampling and understaging PCa and resulting in overtreatment related morbidity such as incontinence and impotence [6, 7].

Over the past decade, 3 Tesla multi-parametric MR imaging (3T mp-MRI), mainly consisting of 2D or 3D T_2 weighted (T2) imaging, high b-value diffusion weighted imaging (DWI), and high temporal resolution dynamic contrast-enhanced (DCE) imaging, has become the dominant non-invasive diagnostic tool for diagnosing and grading PCa [8]. 3T mp-MRI enables detection of 50% of all PCa lesions and 80% of clinically significant lesions [9]. However, one of the main limitations of the broader applications of mp-MRI is its requirement for training and expertise for image interpretation by subspecialized radiologists, which may lead to high inter- and intra-reader variability in diagnosis [10], even with standardized guidelines such as Prostate Imaging Reporting and Data System (PIRADS v2) [11]. These guidelines which evaluate mainly qualitative parameters for T2, DWI, and DCE enable more standardized and reproducible reporting of clinically significant PCa foci, but are still subject to limitations such as interobserver variability [12] and improved but limited correlation with final pathology [13]. Methods to standardize image interpretation would be desirable.

Of currently available techniques, deep neural networks have demonstrated superior capabilities in non-medical imaging domains for extracting multi-level abstraction from raw images directly with little human intervention [14]. Application in medical imaging domains remains challenging in part due to large amount of labeled data required to train a reliable neural network. In addition, the labeling process is tedious, requires expertise, and can be expensive. Another technique, deep transfer learning (DTL) using fine-tuned pre-trained convolutional neural network (CNN) alleviates the large labeled data requirement and has been successfully applied in medical domains [15].

In this study, we develop the DTL-based prostate cancer classification model and compare the classification performance of the DTL-based model with deep learning (DL) model without transfer learning and standard PIRADS v2 radiologist interpreted score to distinguish clinically significant from indolent PCa lesions on a curated 3T mp-MRI dataset with whole-mount histopathology (WMHP) correlation.

Materials and methods

Study population and MR imaging technique

This retrospective study was approved by the institutional review board (IRB) and was compliant with the 1996 Health Insurance Portability and Accountability Act. Between December 2010 and June 2016, a standardized 3T mp-MRI protocol consisting of T2, DWI, and DCE imaging was performed. The initial study cohort consisted of 154 patients scanned prior to planned robotic radical prostatectomy with biopsy-proven prostate cancer. Out of 154 patients, 14 were excluded due to lack of pathology. The final study cohort comprised of 140 patients (age 43–80 years and weight 59.0–133.8 kg, and PSA = 7.9 ± 12.5 ng/dl) with 216 lesions identified on mp-MRI and correlated to thin section WMHP. 41% of the patients had MRI scans before undergoing MRI-targeted fusion biopsy (n = 58), and the rest had MRI scans after biopsy for surgical planning and local staging (n = 82). The scans were interpreted by one of the three expert readers and each MR-detected lesion was scored by PIRADS v2 components.

3T mp-MRI was performed on a variety of scanners (Trio, Verio, Prisma or Skyra, Siemens Healthineers, Erlangen Germany) using a pelvic phase-array coil with or without the endorectal coil. Each scan used a standard mp-MRI scanning protocol including 3D axial T2 images using Sampling Perfection with Application optimized Contrasts using different flip angle Evolution (SPACE) sequence, echo-planar imaging DWI sequence, and DCE images using Time-resolved angiography With Interleaved Stochastic Trajectories (TWIST). In this study, we used T2 SPACE images and apparent diffusion coefficient (ADC) images calculated from DWI. The echo time and repetition time of the T2 SPACE was 2200/200 ms, and the echo train length was 88. With a 17 cm FOV and matrix size of 256 × 230, we acquired and reconstructed T2 SPACE images with 0.66 mm in-plane resolution and 1.5 mm through plane resolution. For DWI acquisition, we used echo time and repetition time of 4800 and 80 ms. With FOV of 21 × 26 cm and matrix of 94 × 160, DWI images were reconstructed with in-plane resolution of 1.6 mm and a slice thickness of 3.6 mm. The ADC images were calculated from four b values 0, 100, 400, and 800 s/mm^2.

The ground-truth of this study was lesions detected by genitourinary (GU) pathologist on post robotic-assisted laparoscopic prostatectomy WMHP, blinded to all MRI information. At a separate matching session, a team of one GU radiologist and one GU pathologist matched each previously reported lesion on 3T mp-MRI and each previously reported lesion on WMHP classifying MRI lesions as true or false positives. After matching, the T2 SPACE and ADC images were imported into a commercially available image processing program OsiriX (Pixmeo SARL, Bernex, Switzerland) and a region of interest (ROI) was drawn around each true and false positive lesion on T2 SPACE and ADC images. Each ROI was marked as either clinically significant lesion ($GS \geq 7$) or indolent lesion ($GS \leq 6$ or false positive) [4]. In total, we marked 111 indolent lesions and 105 clinically significant lesions, consisting of 45 false positive lesions (21%), 66 GS 3 + 3 lesions (31%), 66 GS 3 + 4 lesions (31%), 23 GS 4 + 3 lesions (11%) and 16 GS > 7 lesions (7%). Representative lesion segmentation and corresponding label definition are shown in Fig. 1a.

General Workflow

Our general workflow is shown in Fig. 1a. The input data were T2 SPACE and ADC images with each lesion contoured on both sequences using OsiriX. After proper preprocessing, the image patches enclosing the lesion were generated as the input to the proposed DTL based model. A predicted probability of clinically significant PCa lesion was estimated through the model and was evaluated to determine the final prediction. The zone information of the prostate lesion (either peripheral zone or transition zone) was reviewed by the multidisciplinary team and added to the DTL-based model as a separate feature input.

Pre-processing

During pre-processing, the square image patches of T2 SPACE and ADC enclosing the lesion were created based on the lesion contour (shown in Fig. 1b). To prepare for further data augmentation, we cropped the image patches with a certain margin. Based on each lesion contour, a rectangle ROI was generated (orange box), and the ROI was expanded to a square ROI (yellow box) which has the same center and a side length of 1.4 times of the longer side of the rectangle ROI. The image patch was cropped based on the square ROI. The side lengths of cropped T2 SPACE image patches ranged from 17 pixels to 92 pixels with a mean of 35 pixels, and this T2 pixel size was recorded and fed into the DTL-based model as a separate feature as well.

Each image patch was normalized to pixel intensity of 0–255 before fed into the DTL based model. The study population included both scans with and without the endorectal coil, and an additional image normalization of the T2 SPACE images was implemented for patients with the endorectal coil to account for high signal variation at the interface of coil and tissue. The image normalization of T2 SPACE was based on the maximum value within a rough contour of normal prostate tissue defined on each patient. For quantitative ADC images, we set an empirical intensity upper bound of 4000 to filter extreme values before normalization to maintain the distribution of ADC values within and among cases. One representative example with and without the endorectal coil is shown in Fig. 1b. The T2 SPACE and ADC image patches were resized to 32 by 32 pixel before fed into CNNs.

DTL-based model

The DTL-based model structure is described in Fig. 1c. We utilized ResNet [16], a state-of-the-art variant of CNN performing well on similar tasks [17]. The ResNet model we utilized consists of 19 convolutional layers (blue) and 9 building blocks (dark green). To fuse the two ResNet models, features (f_{T2} and f_{ADC}) from the last average pooling layer (green) with two key features of T2 SPACE image patch size and the zone information of the lesion were concatenated. The two ResNet models were initialized with weights trained from CIFAR10 dataset [18], and the weights of the last convolutional layer of the two ResNet models together with the added fully connected layer were fine-tuned using back-propagation during the training process. In this study, we implemented the deep learning models using Caffe framework (University of California, AI Research, Berkeley CA) [19].

Model evaluation and comparison

We selected 169 lesions (83 clinically significant and 86 indolent) in 110 patients as training set and the remaining 47 lesions (22 clinically significant and 25 indolent) in 30 patients as testing set from 105 clinically significant and 111 indolent lesions. The training and testing set data splitting was based per patient so that all lesions from the same patient would be in either the training or testing set. As described by Shin et al., proper data augmentation on the training data with large variety is crucial for applying deep learning algorithm in medical image domain [20]. We applied the same random flipping, cropping and slightly shearing transformation to each T2 SPACE and ADC image pair to increase the variety of the training data and the robustness of the model. Random contrast adjustment was only applied to T2 SPACE images. An example of T2 SPACE and ADC images after data augmentation is shown in Fig. 2. After data augmentation, 2574 clinically significant and 2580 indolent lesions were used as the input of the training process. The numbers of samples after data augmentation in both classes were almost equal so that the prediction bias during training introduced by class imbalance can be avoided. The DTL-based model was trained on the augmented training set and evaluated on the testing set.

Additionally, DTL-based models with T2 SPACE images (DTL_{T2}) alone, ADC images (DTL_{ADC}) alone and

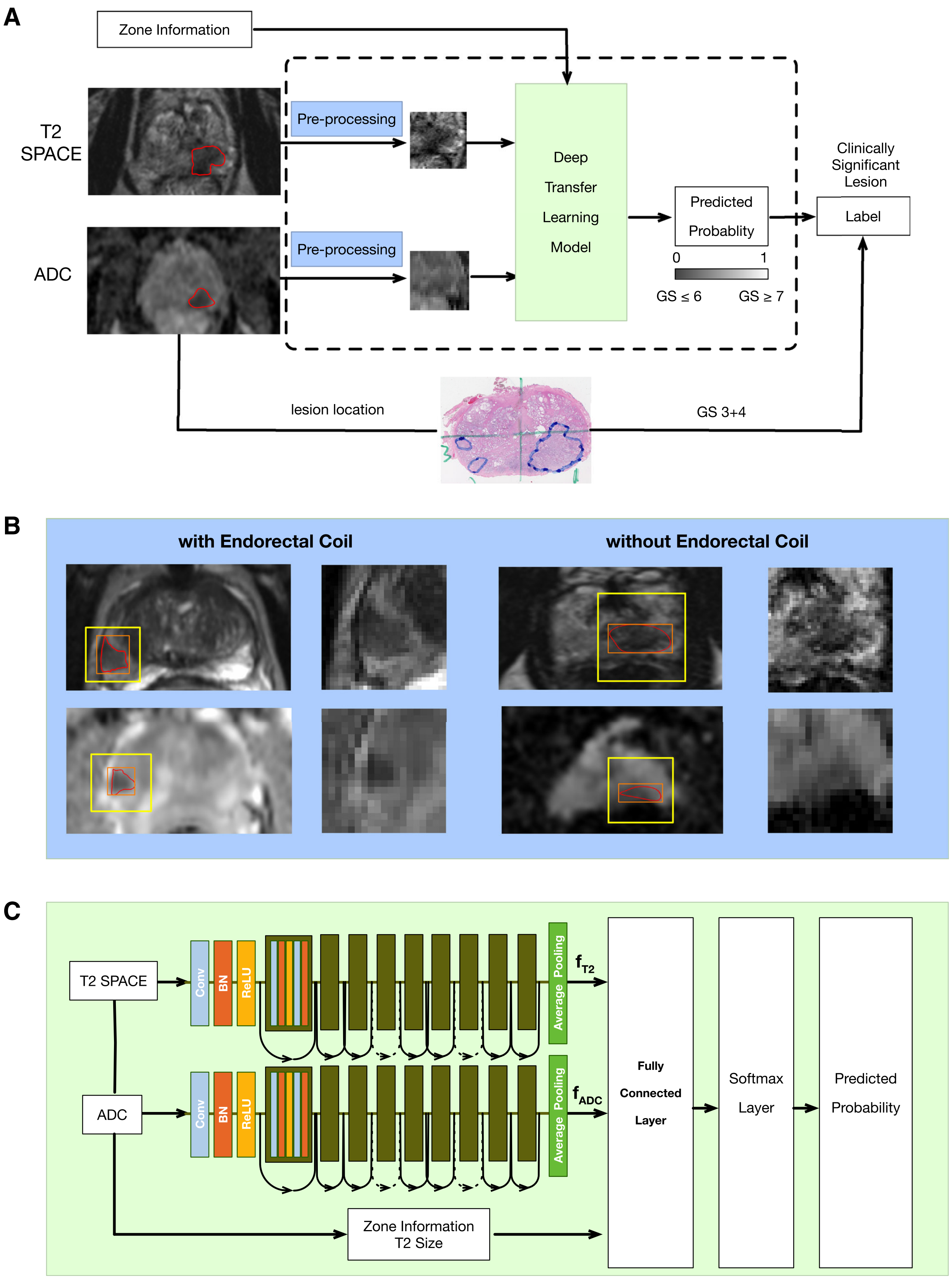

Fig. 1 Summary of the overall workflow (**a**), preprocessing (**b**) and deep transfer learning-based model (**c**)

combined T2 SPACE and ADC images ($\mathrm{DTL}_{\mathrm{T2+ADC}}$) without added features were individually evaluated to determine the contribution of each component of the model. In $\mathrm{DTL}_{\mathrm{T2}}$ and $\mathrm{DTL}_{\mathrm{ADC}}$ model, one ResNet model trained on CIFAR10 was fined-tuned, and the input to the fully connected layer was f_{T2} or $\mathrm{f}_{\mathrm{ADC}}$ only. Likewise, $\mathrm{DTL}_{\mathrm{T2+ADC}}$ model architecture was similar to DTL based model architecture with only f_{T2} and $\mathrm{f}_{\mathrm{ADC}}$ concatenated and fed to the fully connected layer. DL model without transfer learning was implemented with the same architecture as DTL-based model but trained from scratch with MSRA initialization [21]. DL model was evaluated to illustrate the contribution of transfer learning. All these models were also trained on the augmented training set and evaluated on the testing set to validate our proposed DTL-based model architecture.

The PIRADS v2 score for each lesion was extracted from the final radiology report interpreted by a subspecialized radiology fellow and an experienced prostate subspecialized genitourinary radiologist. We compared the performance of our DTL-based model and the performance of expert reader PIRADS v2 score $\geq$ 4 in the task of distinguishing clinically significant PCa lesions from indolent lesions on the same testing set. With a threshold of 4, lesion with PIRADS v2 score smaller than 4 was considered indolent lesion, whereas lesion with PIRADS v2 score larger than or equal to 4 was considered clinically significant lesion. The PIRADS v2 score served as an expert reader baseline to compare with results from our proposed DTL-based model.

Statistical analysis

The classification performance was quantitatively evaluated by accuracy, sensitivity, specificity, and area under curve (AUC) of receiver operating characteristics (ROC) curve using 47 testing lesions from 30 cases. ROC curves and precision-recall curves were also shown to give a more thorough description of the classification performance. The threshold to calculate accuracy, sensitivity, and specificity was picked based on the best accuracy. Those analyses were performed on Matlab 2014a (MathWorks, Natick, MA). Bootstrapping with 2000 resamples was performed to estimate the 95% confidence interval (CI) for AUC and DeLong test was used to compare the AUC of the DTL-based model and other models as well as PIRADS v2 score. The AUC-related analysis was implemented using pROC toolbox [22] in R package [23].

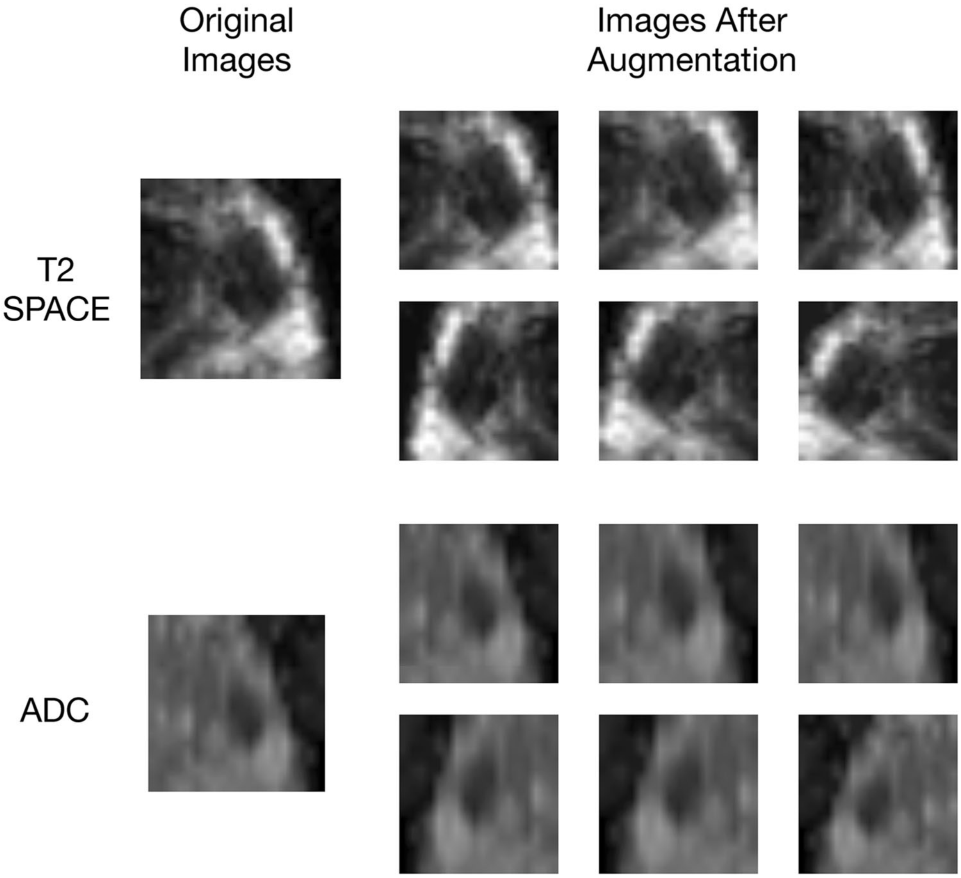

Fig. 2 Representative examples of the data augmentation. The left column shows the image patches before data augmentation, and the right column shows six examples of the image patches after data augmentation for T2 SPACE and ADC images

Results

DTL-based model evaluation

With GS from WMHP as the ground truth, the DTL-based model achieved an accuracy of 0.723 in distinguishing indolent from clinically significant PCa lesion. The corresponding sensitivity and specificity were 0.636 and 0.800, respectively. The AUC of the ROC curve was 0.726 (CI [0.575, 0.876]). Representative examples of successful and failed predictions of the DTL-based model are depicted in Fig. 3. The false positive (Fig. 3b) and false negative (Fig. 3d) examples were all from scans with the endorectal coil, and the T2 SPACE contrast after pre-processing was suboptimal. Also, those examples illustrated how similar lesions appear on 3T mp-MRI and showed challenges for both algorithms and human to properly score the lesions.

DTL-based model comparison with other models

To properly evaluate the contribution from T2 SPACE images, ADC images and two added features to the classification performance, we trained and compared the models with each image separately (DTL_{T2} and DTL_{ADC}) and the model with two images combined (DTL_{T2+ADC}). The experiments results are summarized in Table 1. The resulting ROC and precision-recall curve are shown in Fig. 4. DTL_{T2} achieved the least accuracy of 0.617 among all the models, indicating currently T2 SPACE images provided less valuable information compare to ADC images regarding the classification task. DTL_{T2}, DTL_{ADC}, and DTL_{T2+ADC} models all achieved a higher sensitivity of 0.773 compared to the DTL-based model but with less specificity, which illustrated that the added zone information and T2 size features help to reduce false positive prediction, increasing the specificity. Overall, these experiment results validated the DTL-based model structure design as DTL_{T2}, DTL_{ADC}, and DTL_{T2+ADC} experiments generated less accurate prediction compared to DTL-based model regarding accuracy and AUC. The *p* values of DeLong test between DTL-based model and DTL_{T2}, DTL_{ADC}, DTL_{T2+ADC} were 0.14, 0.30, and 0.74, respectively. Although the difference was not statistically significant due to the limited size for testing, the general trend indicated the relative significance of each component.

To evaluate the contribution of transfer learning, we also compared the DTL-based model with the DL model without transfer learning using the same model structure with MSRA initialization. Our DTL-based model achieved higher accuracy and AUC compared to non-transfer leaning DL (Table 1), which achieved an accuracy of 0.702 and AUC of 0.687 ($p = 0.70$). This experiment showed that transfer learning helped to improve the testing accuracy with a limited training set. Additionally, both DL model and DTL-based model provided predictions with higher specificity and lower sensitivity compared to DTL_{T2}, DTL_{ADC}, and DTL_{T2+ADC} models, confirming the two added feature could reduce false positive prediction. The corresponding ROC and precision-recall curves are shown in Fig. 4.

DTL-based model comparison with PIRADS v2 score expert reader performance

In detection of clinically significant from indolent PCa, the accuracy of the expert radiologist assigned PIRADS v2 score ≥ 4 was 0.708 with AUC of 0.765 in all lesions (both training and testing set), and was 0.660 with AUC of 0.711 (CI [0.575, 0.874]) in the testing set. The performance with PIRADS v2 score ≥ 4 is summarized in Table 1. DeLong test showed that the DTL-based model generated comparable performance with PIRADS v2 score ($p = 0.89$).

The ROC and precision-recall curve comparison between the DTL-based model and PIRADS v2 score are shown in Fig. 5. With estimation favoring specificity (left side of the curve), the proposed DTL-based model outperformed PIRADS v2 score expert radiologist interpretation. While with estimation favoring sensitivity (right side of the curve), the expert reader PIRADS v2 score outperformed proposed DTL-based model. This observation is consistent with performance in Table 1, where the best sensitivity and specificity were 0.636 and 0.800 for DTL-based model and 0.864 and 0.480 for PIRADS v2 score. The proposed DTL-based model tends to reduce the overdiagnosis with higher specificity compared to PIRADS v2 score.

Discussion

The DTL-based model was developed and compared with the performance of a DL model without transfer learning and an expert radiologist detecting clinically significant PCa using the PIRADS v2 score to distinguish clinically significant from indolent PCa lesions using 3T mp-MRI with the highest available reference standard, histopathological grading on WMHP. Each discrete prostate lesion was first identified on 3T mp-MRI and contoured on OsiriX based on the appearance, and small image patches enclosing the lesion of T2 SPACE and ADC images were used as input to the DTL-based model with labeling based on WMHP GS. The proposed DTL-based model outperformed DL model without transfer learning and achieved comparable performance compared to PIRADS v2 score

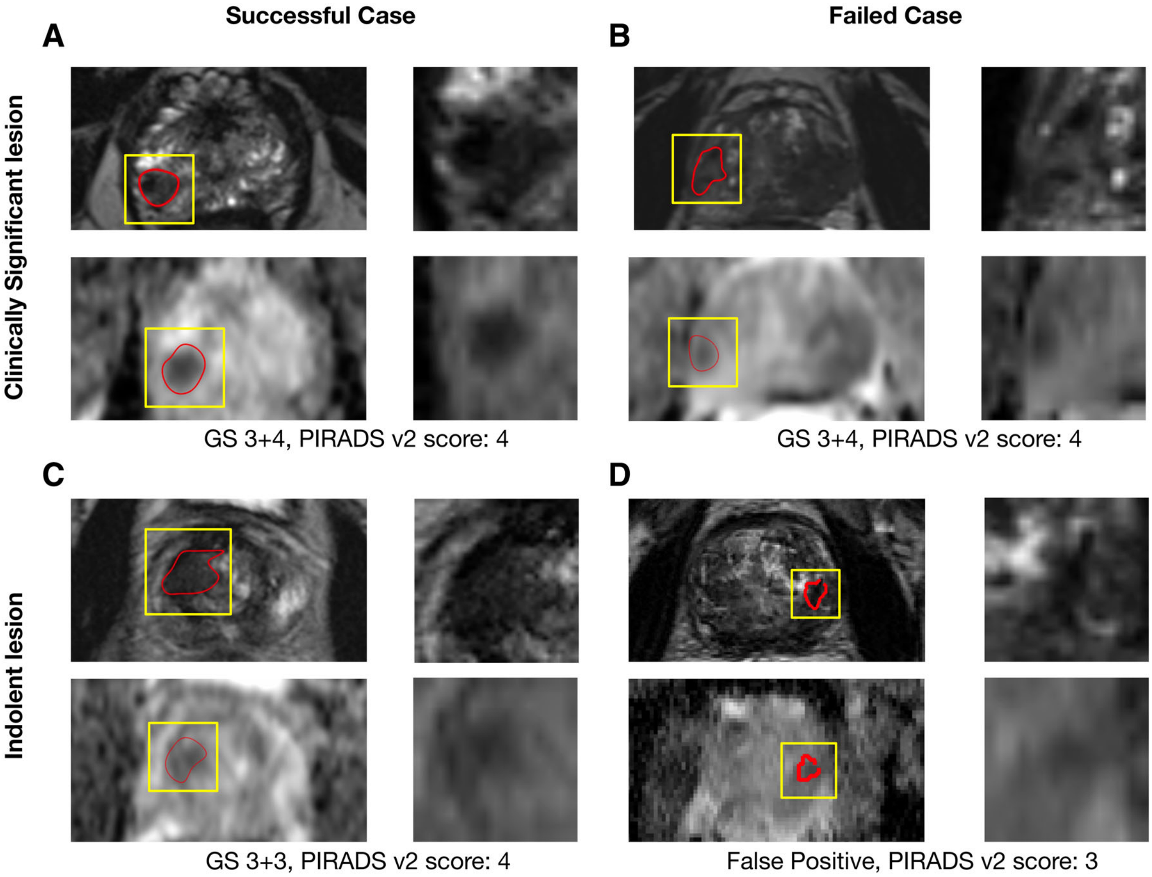
Successful Case
Failed Case
A
B
Clinically Significant lesion
GS 3+4, PIRADS v2 score: 4
GS 3+4, PIRADS v2 score: 4
C
D
Indolent lesion
GS 3+3, PIRADS v2 score: 4
False Positive, PIRADS v2 score: 3

Fig. 3 Representative cases for successful and failed prediction cases based on the DTL-based model on clinically significant lesion (**a**, **b**) and indolent lesion (**c**, **d**). The corresponding GS from WMHP and PIRADS v2 score for each lesion are also shown under the images

Table 1 Performance summary on the testing set (n = 47)

Performance	AUC (95% CI)	Accuracy	Sensitivity	Specificity
DTL_{T2}	0.580 (0.412–0.748)	0.617	0.773	0.480
DTL_{ADC}	0.620 (0.455–0.785)	0.638	0.773	0.520
$\text{DTL}_{\text{T2+ADC}}$	0.713 (0.563–0.863)	0.702	0.773	0.640
DL	0.687 (0.532–0.843)	0.702	0.636	0.760
DTL	**0.726 (0.575–0.876)**	**0.723**	0.636	**0.800**
PIRADS v2 score	0.711 (0.575–0.847)	0.660	**0.864**	0.480

The best performances among all methods were emboldened

using single slice 2D image patches from T2W and ADC images on our 47 lesions testing data.

One advantage of our method is that it does not rely on the detailed contour defined by the radiologists, because the information used to generate the image patch was mainly the location and a rough size of the lesion. We further improved the robustness of the model to lesion ROI definition by adding random cropping into the data augmentation. In this way, the potential inconsistency of ROI definition from either lesion detection algorithms or radiologists will have little influence on the model performance.

A prior study has shown great potential to use DTL to distinguish clinically significant lesion [17] using publicly available prostate MRI data. Here, we have chosen to use the ResNet trained on CIFAR10 data to minimize potential overfitting and added T2 size and zone information as additional features to the DTL-based model to improve the overall classification performance. Also, our model implementation and evaluation were based on labeling information referenced by the WMHP and included more practical situations, such as cases with and without the endorectal coil.

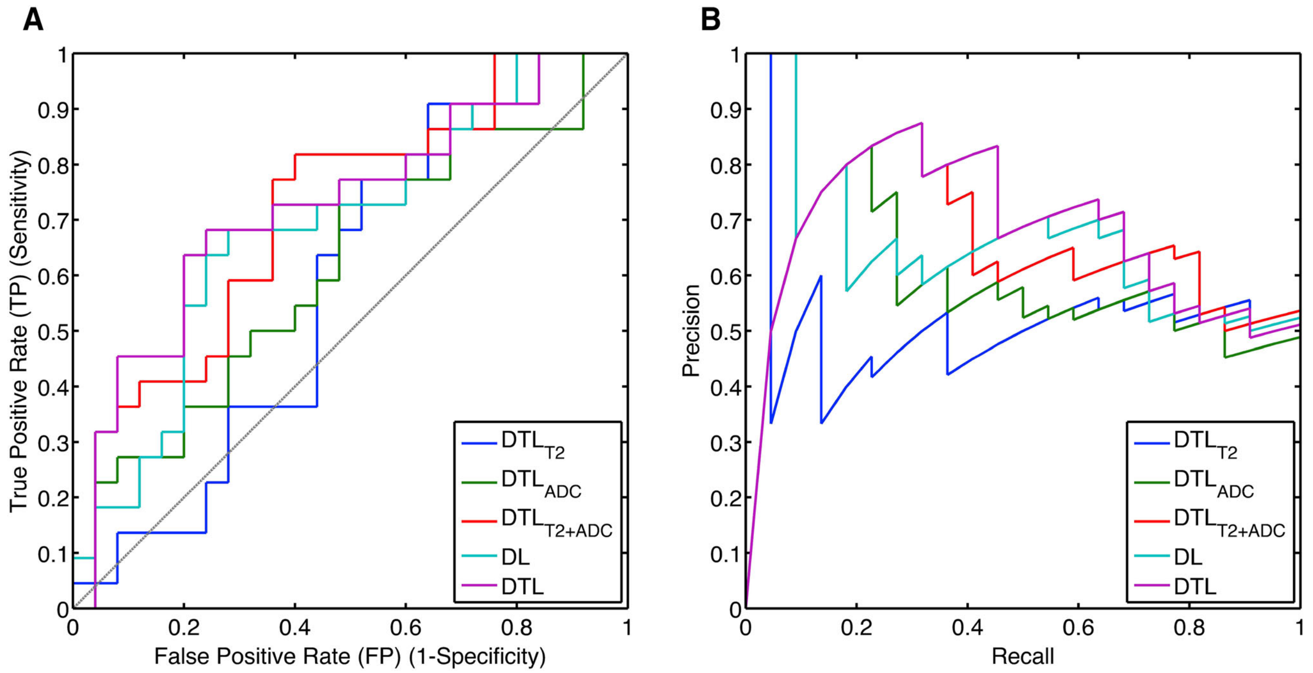

Fig. 4 Receiver operating characteristics curve (**a**) and precision recall curve (**b**) comparison between DTL, $\mathrm{DTL}_{\mathrm{T2}}$, $\mathrm{DTL}_{\mathrm{ADC}}$, $\mathrm{DTL}_{\mathrm{T2+ADC}}$, and DL models on 47 testing lesions, validating the DTL-based model architecture

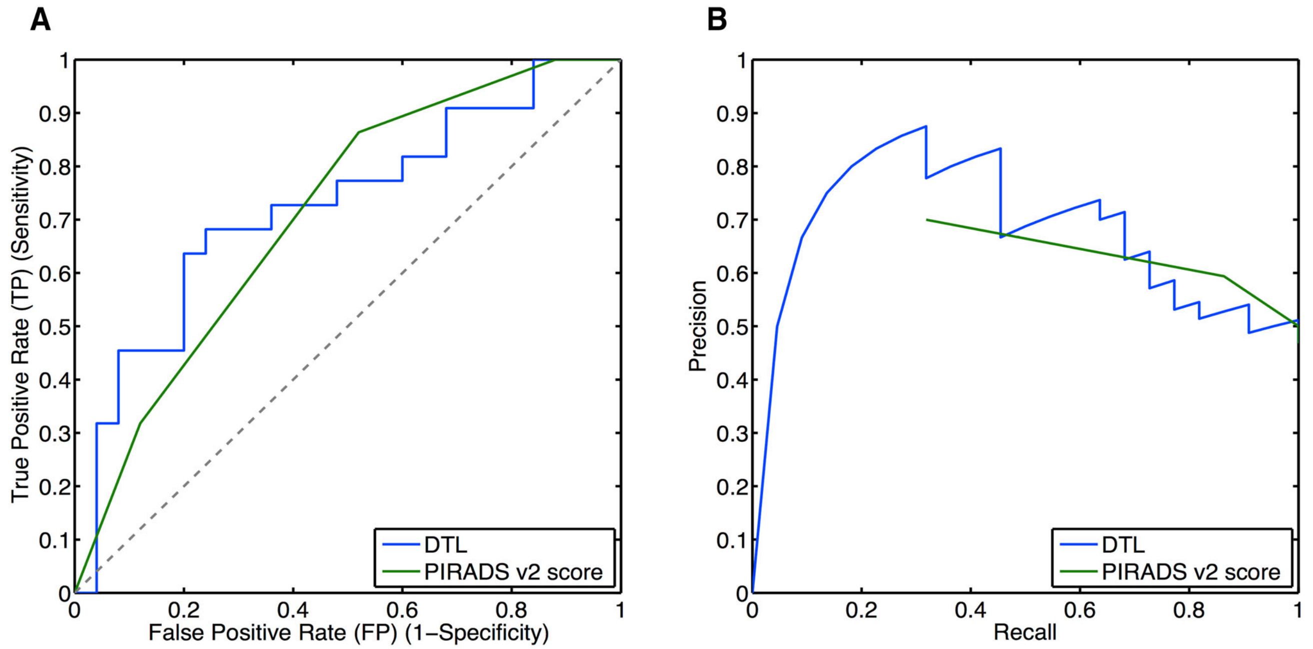

Fig. 5 Receiver operating characteristics curve (**a**) and precision recall curve (**b**) comparison between the proposed DTL-based model and the expert reader PIRADS v2 score (n = 47)

Our study has compared the DTL based prostate cancer classification to PIRADS v2 score. Although PIRADS v2 score was not originally designed for this task, studies have shown that lesions on 3T mp-MRI with higher PIRADS v2 score correlate with PCa lesions with higher Gleason Score on WMHP [24]. We picked a threshold four, which achieved the best accuracy for PIRADS v2 score in the testing set. All prostate mp-MRI cases were interpreted by expert genitourinary radiologists who have more than 10 years of experience with approximately 500 prostate MRI cases per year. We believe the PIRADS v2 scores in the study would be close to the upper limit of the prostate

MRI interpretation, a relatively good approximation of human performance.

We distinguished clinically significant lesions from indolent lesions and false positive lesions rather than normal tissue, as shown in Fig. 3. The task is challenging because differences on MRI images between these lesions are sometimes not visually apparent. Although there is no universally accepted definition of the clinically significant lesion, we used one of the most common definitions, GS > 6 [4], as our working definition of the clinically significant lesion to emphasize any differences between low- and intermediate/high-grade prostate lesions. We have also compared our proposed DTL-based model with a traditional machine learning method as another baseline. We vectorized the pre-processed T2 SPACE and ADC images as the input and trained a random forest model [25] using the same training cohort before data augmentation. On the testing set, the random forest model provided a prediction with AUC of 0.66. This result confirmed the well-known challenges of not using detailed lesion contours and indicated that deep learning based methods can alleviate the requirements of handcrafted imaging features to achieve relatively satisfactory results.

The proposed DTL-based model has achieved comparable results compared to PIRADS v2 score, but there still exist potential improvements in the performance, as the model has not fully utilized the imaging information from 3T mp-MRI. For example, the current input was from single 2D segmentation, and the feature extracted from 3D region of interest may provide a more comprehensive depiction of the tissue character. Also, we did not take DCE-MRI into consideration because DCE-MRI was not available for some of the cases. As the cases accumulate, we would only include cases with K^{trans} maps and add K^{trans} maps into our model. Another potential improvement is to apply registration between T2 SPACE and ADC images for each case so that the two images can be fed into the same model, reducing the model complexity and potential overfitting problem.

Our study has several limitations. One limitation of this study is the small sample size for testing because of the limited available labeled data. Even with the transfer learning, the model requires sufficient training data to produce a more generalized model, resulting in a limited testing set size. The evaluation on a larger testing cohort will be conducted in the future as we acquire more labeled data. With a larger testing set, a more statistically powerful evaluation of the DTL method can be made. Moreover, our current evaluation was performed on one random splitting cohort. Although it is commonly used in deep learning-related papers due to computation limitation, it might give prediction with a certain bias. We included confidence interval of the AUC from bootstrapping resampling to give a more thorough evaluation. Another limitation of our study is that we included a manual segmentation of the prostate to assist the normalization for the cases with the endorectal coil (55 out of 140 cases were scanned with the endorectal coil). We also included contrast adjustment in data augmentation process to improve the robustness to T2 contrast. Although this method was proved to achieve similar contrast for all T2 SPACE images, the manual segmentation of the prostate could be burdensome, and a more systematic way to find the normalization threshold is preferable and still under investigation. In the evaluation of $\mathrm{DTL}_{\mathrm{T2}}$ and $\mathrm{DTL}_{\mathrm{ADC}}$, the features from T2 SPACE images contribute less than features from ADC in this classification task. This observation is consistent with radiologists' experience, but the performance of $\mathrm{DTL}_{\mathrm{T2}}$ can be improved with well-designed pre-processing. Another limitation is that the system requires the lesion detection as the input to define an image patch. Although a prostate lesion detection and classification system is possible by combining other lesion detection systems [26] with our classification system, it may be preferable to design one integrated system that could complete the pixel-level lesion detection and classification task together.

Conclusion

We implemented and evaluated a deep transfer learning (DTL)-based model to differentiate between clinically significant (GS ≥ 7) and indolent PCa lesions (GS ≤ 6 and false positive) using 3T mp-MRI with WMHP correlation. The proposed DTL-based model outperformed the DL-based model without transfer learning, confirming the contribution of transfer learning. The DTL-based model performance generated comparable performance to the expert reader PIRADS v2 score ($p = 0.89$), showing great potential to augment PCa for non-experts. This model would need to be validated in much larger datasets to further evaluate its clinical utility.

Acknowledgements The study was supported by Siemens Medical Solutions and by funding from the Integrated Diagnostics program of the Departments of Radiological Sciences & Pathology & Laboratory Medicine, David Geffen School of Medicine at UCLA.

Compliance with ethical standards

Conflict of interest All the authors declare no conflict of interest.

Ethics 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. Since the study involved purely retrospective analysis of previously acquired data, the Institutional Review Board waived the need for additional informed consent.

References

1. Siegel RL, Miller KD, Jemal A (2018) Cancer statistics, 2018. CA Cancer J Clin 68:7–30.
2. Ahmed HU, Akin O, Coleman JA, Crane S, Emberton M, Goldenberg L, Hricak H, Kattan MW, Kurhanewicz J, Moore CM, Parker C, Polascik TJ, Scardino P, van As N, Villers A (2012) Transatlantic Consensus Group on active surveillance and focal therapy for prostate cancer. BJU Int 109:1636–1647.
3. Lavery HJ, Droller MJ (2012) Do gleason patterns 3 and 4 prostate cancer represent separate disease states? J Urol 188:1667–1675.
4. Felker ER, Raman SS, Margolis DJ, Lu DSK, Shaheen N, Natarajan S, Sharma D, Huang J, Dorey F, Marks LS (2017) Risk Stratification Among Men With Prostate Imaging Reporting and Data System version 2 Category 3 Transition Zone Lesions: Is Biopsy Always Necessary? Am J Roentgenol 209:1272–1277.
5. Schröder FH, Hugosson J, Roobol MJ, Tammela TLJ, Ciatto S, Nelen V, Kwiatkowski M, Lujan M, Lilja H, Zappa M, Denis LJ, Recker F, Berenguer A, Määttänen L, Bangma CH, Aus G, Villers A, Rebillard X, van der Kwast T, Blijenberg BG, Moss SM, de Koning HJ, Auvinen A (2009) Screening and Prostate-Cancer Mortality in a Randomized European Study. N Engl J Med 360:1320–1328.
6. Caster JM, Falchook AD, Hendrix LH, Chen RC (2015) Risk of Pathologic Upgrading or Locally Advanced Disease in Early Prostate Cancer Patients Based on Biopsy Gleason Score and PSA: A Population-Based Study of Modern Patients. Int J Radiat Oncol Biol Phys 92:244–51.
7. Cohen MS, Hanley RS, Kurteva T, Ruthazer R, Silverman ML, Sorcini A, Hamawy K, Roth RA, Tuerk I, Libertino JA (2008) Comparing the Gleason Prostate Biopsy and Gleason Prostatectomy Grading System: The Lahey Clinic Medical Center Experience and an International Meta-Analysis. Eur Urol 54:371–381.
8. Hoeks CMA, Barentsz JO, Hambrock T, Yakar D, Somford DM, Heijmink SWTPJ, Scheenen TWJ, Vos PC, Huisman H, van Oort IM, Witjes JA, Heerschap A, Fütterer JJ (2011) Prostate cancer: multiparametric MR imaging for detection, localization, and staging. Radiology 261:46–66.
9. Tan N, Margolis DJ, Lu DY, King KG, Huang J, Reiter RE, Raman SS (2015) Characteristics of detected and missed prostate cancer foci on 3-T multiparametric MRI using an endorectal coil correlated with whole-mount thin-section histopathology. Am J Roentgenol 205:W87–W92.
10. Litjens GJS, Barentsz JO, Karssemeijer N, Huisman HJ (2015) Clinical evaluation of a computer-aided diagnosis system for determining cancer aggressiveness in prostate MRI. Eur Radiol 3187–3199.
11. ACM (2015) Pi-Rads Prostate Imaging - Reporting and Data System. Am. Coll. Radiol.
12. Giannarini G, Girometti R, Sioletic S, Rossanese M, Palumbo V, Calandriello M, Crestani A, Zuiani C, Ficarra V (2018) Inter-reader agreement of Prostate Imaging Reporting and Data System version 2 in detecting prostate cancer on 3 Tesla multiparametric MRI: A prospective study on patients referred to radical prostatectomy. Eur Urol Suppl 17:e893.
13. Vaché T, Bratan F, Mège-Lechevallier F, Roche S, Rabilloud M, Rouvière O (2014) Characterization of Prostate Lesions as Benign or Malignant at Multiparametric MR Imaging: Comparison of Three Scoring Systems in Patients Treated with Radical Prostatectomy. Radiology 272:446–455.
14. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444.
15. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J (2016) Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Trans Med Imaging 35:1299–1312.
16. Wu S, Zhong S, Liu Y (2017) Deep residual learning for image steganalysis. Multimed Tools Appl 1–17.
17. Le MH, Chen J, Wang L, Wang Z, Liu W, Cheng K-T (Tim), Yang X (2017) Automated diagnosis of prostate cancer in multiparametric MRI based on multimodal convolutional neural networks. Phys Med Biol 62:6497–6514.
18. Krizhevsky A (2009) Learning Multiple Layers of Features from Tiny Images.
19. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional Architecture for Fast Feature Embedding.
20. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM (2016) Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning.
21. He K, Zhang X, Ren S, Sun J (2015) Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.
22. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77.
23. R Core Team (2014) R: A language and environment for statistical computing. R Found Stat Comput Vienna, Austria 2014.
24. Zhao C, Gao G, Fang D, Li F, Yang X, Wang H, He Q, Wang X (2016) The efficiency of multiparametric magnetic resonance imaging (mpMRI) using PI-RADS Version 2 in the diagnosis of clinically significant prostate cancer. https://doi.org/10.1016/j.clinimag.2016.04.010
25. Breiman L (2001) Random forests. Mach Learn 45:5–32.
26. Ishioka J, Matsuoka Y, Uehara S, Yasuda Y, Kijima T, Yoshida S, Yokoyama M, Saito K, Kihara K, Numao N, Kimura T, Kudo K, Kumazawa I, Fujii Y (2018) Computer-aided diagnosis of prostate cancer on magnetic resonance imaging using a convolutional neural network algorithm. BJU Int 122:411–417.