Convolutional Neural Network Based Deep-learning Architecture for Prostate Cancer Detection on Multiparametric Magnetic Resonance Images

Yohannes K. Tsehay, Nathan S. Lay, Holger R. Roth, Xiaosong Wang, Jin Tae Kwak[a], Baris I. Turkbey, Peter A. Pinto[b], Brad J. Wood[c], and Ronald M. Summers[a]

[a]Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Department of Radiology and Imaging Science, National Institute of Health, Clinical Center, Bethesda, MD 20892
[b]Urologic Oncology Branch, National Cancer Institute, National Institute of Health, Bethesda, MD 20892, USA
[c]Center for Interventional Oncology, National Cancer Institute, National Institute of Health, Bethesda, MD 20892, USA

ABSTRACT

Prostate cancer (PCa) is the second most common cause of cancer related deaths in men. Multiparametric MRI (mpMRI) is the most accurate imaging method for PCa detection; however, it requires the expertise of experienced radiologists leading to inconsistency across readers of varying experience. To increase inter-reader agreement and sensitivity, we developed a computer-aided detection (CAD) system that can automatically detect lesions on mpMRI that readers can use as a reference. We investigated a convolutional neural network based deep-learing (DCNN) architecture to find an improved solution for PCa detection on mpMRI. We adopted a network architecture from a state-of-the-art edge detector that takes an image as an input and produces an image probability map. Two-fold cross validation along with a receiver operating characteristic (ROC) analysis and free-response ROC (FROC) were used to determine our deep-learning based prostate-CAD's (CAD_{DL}) performance. The efficacy was compared to an existing prostate CAD system that is based on hand-crafted features, which was evaluated on the same test-set. CAD_{DL} had an 86% detection rate at 20% false-positive rate while the top-down learning CAD had 80% detection rate at the same false-positive rate, which translated to 94% and 85% detection rate at 10 false-positives per patient on the FROC. A CNN based CAD is able to detect cancerous lesions on mpMRI of the prostate with results comparable to an existing prostate-CAD showing potential for further development.

Keywords: Computer-Aided Diagnosis, Holistically-nested Edge Detection, CNN, HED

1. INTRODUCTION

One in six men will develop prostate cancer (PCa) in their lifetime. An estimated 232,090 new cases and 30,350 deaths were expected for the year 2005, making PCa the second most common cause of cancer related deaths in men.[1] The state-of-the-art diagnostic method uses multiparameteric MR imaging (mpMRI) with transrectal ultrasound-guided biopsy.[2] MpMRI is the most accurate imaging method for PCa detection;[3] however, it requires the expertise of experienced radiologists, and as such, there is inconsistency across readers of varying experience.[4] To increase inter-reader agreement and sensitivity, we developed a computer-aided detection (CAD) system that can automatically detect lesions on mpMRI that readers can use as a reference. Many of the existing prostate CAD systems use hand-crafted features to differentiate between cancerous and normal tissue.[5–8] However, due to the recent success in deep convolutional neural netowrks (DCNN) architectures that can learn descriptive features from the given data,[9] we investigated a DCNN-based prostate-CAD (CAD_{DL}) to find an improved solution for PCa detection on mpMRI.

Further author information: (Send correspondence to Yohannes K Tsehay)
Yohannes K Tsehay: E-mail: yohannes.tsehay@nih.gov, Telephone: 1 301 402 5486
Nathan S Lay: E-mail: nathan.lay@nih.gov, Telephone: 1 301 451 8363

CAD_{DL} had a 0.897 area-under-the-curve (AUC) for a response operating characteristic analysis after training for 6 epochs, which translated to a 0.845 detection rate at 0.2 false-positive rate or 0.94 detection rate at 10 false positives per patient on the free response operating characteristic (FROC) curve. A competing support vector machine based CAD (CAD_{SVM}) that uses hand-crafted features had an inferior performance with 0.859 AUC translating to a 0.801 detection at the same 0.2 false positive rate or 0.85 detection for 10 false positives per patient. Qualitative analysis also showed that CAD_{DL} detects lesions with a higher confidence than CAD_{SVM}; albeit, it is at times prone to false positive detections arising from benign prostatic hyperplasia (BPH) presence and artifacts within the image.

Our results suggest that, DCNN architectures have the potential to improve prostate cancer detection on mpMRI. CAD_{DL} was able to learn discriminating features from the given data achieving competitive results. The free response operating characteristic analysis and generated prediction maps further elucidate its potential for clinical application. For future studies, we will attempt to develop a DCNN based prostate-CAD that has more applications in addition to tumor detection such as tumor size approximation and severity prediction.

2. MATERIALS AND METHODS

2.1 Patient Cohort

Our study cohort consisted of 52 patients with T2W, ADC, and B2000 MR images. For every patient we also had a prostate contour segmentation acquired using technique Ref. 10 and tumor annotations done by an expert radiologist. Figure 1 shows an example of the original image sequences and the regions of interest.

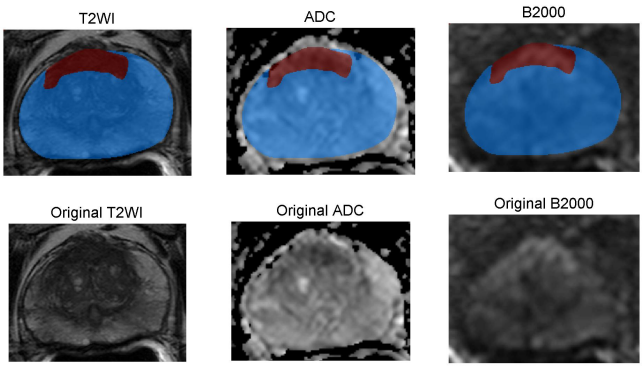

Figure 1. In the bottom row are the original T2W, ADC, and B2000 MR images, and the top row displays the tumor and prostate annotations overlaid on top of the original images.

The patients in the study presented with lesions of various severity and location. Biopsy results are used to characterize the lesions. Table 1 has all the biopsy points for the study cohort stratified by zone (peripheral or transition) and Gleason score(6-9), where the higher the score the more severe the lesion.

Table 1. Lesions stratified by location and severity

Zone	Benign	Gleason 6	Gleason 7	Gleason 8	Gleason 9	Total
Peripheral	24	5	19	10	3	61
Transition	15	8	21	13	7	64
Overall	39	13	40	23	10	125

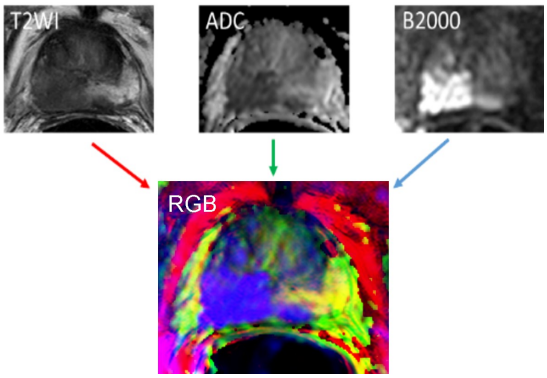

Figure 2. Generating a three channel RGB image from three mpMRIs.

2.2 Image Pre-processing Step

CNNs are mostly designed for natural images which usually have three channels. To satisfy this constraint, we generated a three channel RGB image from three MR sequences as shown in Figure 2. Most publicly available CNNs are further limited in the type of images they take as input into their architecture. Usually, the input must be an 8 bit image with intensity values between 0 and 255; however, medical images such as the mpMRIs that we used are 12-16 bit with intensity values extending to upwards of 4000. We performed histogram equalization for all images to minimize information loss during compression. The three image sequences were processed separately before combining them into a single RGB image as in Figure 2. Figure 3 shows the effects of using a histogram equalizing algorithm. The output from this algorithm has more contrast as compared to an output that uses linear compression according to Equation (1).

$$I' = \frac{255 \cdot I}{\max X}\,, \tag{1}$$

where I is the original mpMRI intensity value for a single pixel, I' is the compressed intensity value, and X is the input image. Lastly, prostate-masks were used to crop the images enabling the learning algorithm to ignore

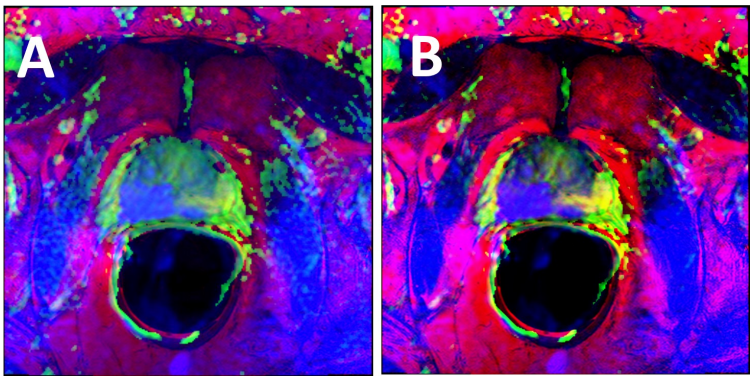

Figure 3. **(A)** is compressed in a linear fashion and without the use of histogram equalization. **(B)** Uses histogram equalization and shows more contrast within the image

regions outside the prostate.

2.3 CNN Architecture

We adopted the CNN architecture of Ref. 11, which takes a whole image as input and produces a probability map image. The architecture has multiple side outputs as shown in Figure 4. Each of the five convolutional

layers have their own loss function and learn in a deeply-supervised fashion; that is to say, the ground-truth is used in every layer for supervision to force each one to learn discriminating features. The hierarchy of features have different receptive field sizes, and the output from each layer can be combined for a final fused output. Inherent in its loss function is a balancing scheme that assigns weights for positive and negative loss outputs

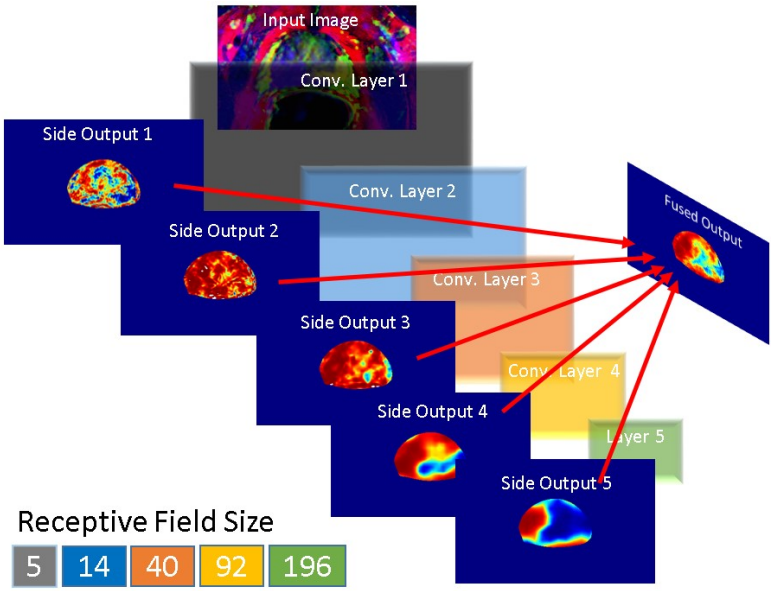

Figure 4. The adopted CNN architecture has 5 side outputs from 5 different convolutional layers each with a different receptive field. The five side outputs can be combined for a fused output.

according to Equation (2).

$$\ell = -\frac{Y^-}{Y} \sum_{i \in Y^+} \log Pr(y_i = 1|X) - \frac{Y^+}{Y} \sum_{i \in Y^-} \log Pr(y_i = 0|X)\,, \tag{2}$$

where $Y = \{y_i, i = 1, ..., |X|\}, y_i \in 0, 1$, Y^+ represent all pixels with positive labels, and Y^- all with negative labels. The loss becomes zero for images with no positive labels such as those image slices with no marked tumors present. This feature reduces our training data set quite dramatically since a lesion on average maintains visibility on 3 slices and a patient's scan can have up to 30 slices. Therefore, we adjusted the balancing weight such that the total number of positive and negative labels in the entire training data-set are used for its computation.

$$Y^{++} = \sum_{i}^{N} Y_i^+\,, and \tag{3}$$

$$Y^{--} = \sum_{i}^{N} Y_i^-\,, \tag{4}$$

where Y^{++} and Y^{--} are the total number of positive and negative labels in the entire training-set, and N is the total number of training examples. The resulting coefficients are essentially global weights that are applied at every iteration during training. Hyper-parameters such as the learning-rate, gamma, momentum, and weight-decay were set to $1e-8$, 0.1, 0.9, and 0.0002 respectively. During training, the weights were initialized using a pre-trained edge detection model, and Sigmoid-Cross-Entropy-Loss with stochastic gradient descent was used for optimization.

2.4 Analysis

2.4.1 Cross-validation

The patient cohort was divided equally into two sets which were used for training or testing depending on the fold. Due to the limited number of patients in the study cohort, a validation-set was left out of the experimental setup. Instead, the CAD was evaluated on the test-set at multiple learning stages to characterize its performance over multiple epochs.

2.4.2 ROC Analysis

ROC analysis is performed on 3D probability maps. The 3D probability maps are first formed for each case by taking the 2D probability map outputs from CAD_{DL} and stacking them. Detection rate is then determined by computing the 90th percentile of the probability scores within each cancerous lesion volume. If this exceeds some threshold, then the cancerous lesion is said to be *detected*. In other words, if at least 10% of the lesion has relatively high probability, it is considered detected. False positive rate is then determined by placing a 3mm x 3mm x 3mm grid on the prostate. Only cells that are inside the prostate are considered. If the 90th percentile of probability scores in a cell exceeds a threshold, then the cell is said to be a false positive. The rationale for the grid is that a prostate reader will likely not mark an imperceptibly small region (e.g. a single voxel) to be biopsied. Furthermore, the grid cells that are within 3mm to the boundary of a cancerous lesion are ignored owing to possible ambiguity in the annotations. The final ROC curve is an average of the curves from the two folds.

2.4.3 FROC Analysis

Like ROC analysis, FROC analysis is also performed on 3D probability maps. However, FROC analysis is compared against biopsy results rather than the cancerous lesion contours. First non-maximum suppression (NMS) is performed in 3D on the probability maps to produce a set of candidate detections. The NMS window used was $10\ mm \times 10\ mm \times\ 10mm$. The candidates roughly reflect local maxima in the probability map. Next, a bipartite graph is formed between detection candidates and ground truth biopsy points. An edge is placed between a detection candidate and ground truth if it is within $10\ mm$. Additionally the edges are weighted with the following heuristic

$$w = p \times 1.5^{g-6} \tag{5}$$

where w is the edge weight, p is the detection probability, and g is the recorded Gleason score ranging from 6-10. The special Gleason score $g = 0$ is used for benign biopsy points. The heuristic attempts to contrast detections paired with severe cancers to detections paired with less severe cancer or benign biopsies while also considering the detection probability. Finally a weighted maximum matching is determined on the bipartite graph. The matched detections then count toward detection rate. Non-matched detection candidates with at least one edge are not counted as these would be redundant detections for one of the biopsies. Non-matched detection candidates with no edge are counted as false positives. Biopsy points that are not matched are counted toward false negative rate. Similar to the final ROC curve, the FROC final plot is averaged over the two folds.

2.4.4 Comparison

The performance of our CAD_{DL} was compared to the published competing CAD_{SVM} by Ref. 6. CAD_{SVM} uses local binary pattern features extracted from T2WI and B2000 images as features to an SVM classifier. In contrast to CAD_{DL}, which was trained on a relatively small number of tumor contour segmentations, biopsy point examples from 108 patients were used to train CAD_{SVM}. The testing-sets from CAD_{DL} were used to compute curves and generate images for CAD_{SVM}.

2.4.5 Visualization

Insight Segmentation and Registration ToolKit (ITK) was used to render color probability maps for the CADs along with prostate and central gland outlines. Thresholds for rendering the images were determined from the ROC curves. A false-positive rate of 20% was chosen as the operating point and used to select a threshold probability. The raw and ground-truth images were generated using a MATLAB script.

3. RESULTS

The training loss plot for CAD_{DL} is shown in Figure 5. In the two folds, the CNN reduces the loss abruptly within the first few iterations and the learning process becomes gradual as the training continues. This trend is repeated in Table 2, where the increase in AUC becomes more incremental as the number of epochs increases plateauing at epoch 5. On the other hand, the detection rate at 0.1 false-postive rate (FPR) increases with increasing margin until it reaches maximums at epoch 5. All the models have a higher AUC and all except the model at epoch-1 have a higher detection rate at 0.1 FPR than CAD_{SVM}.

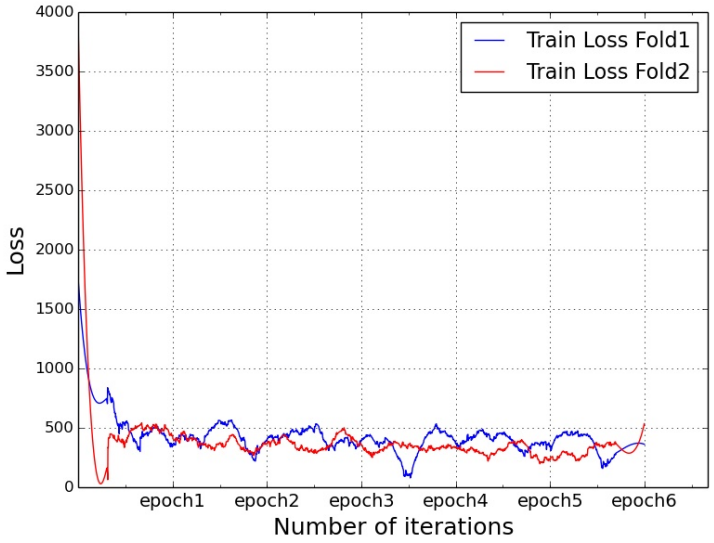

Figure 5. Training loss plot for the two folds in a two fold cross-validation. A filter was used to smooth out the noisy loss values for a final output.

Table 2. Detection rates at 0.1 false-positive rates (FPR) and AUCs for ROC curves of all models from CAD_{DL} and the competing CAD by Ref. 6

Model Name	Detection at 0.1 FPR	AUC
CAD_{DL} Epoch 1	0.597	0.876
CAD_{DL} Epoch 2	0.620	0.885
CAD_{DL} Epoch 3	0.654	0.891
CAD_{DL} Epoch 4	0.711	0.894
CAD_{DL} Epoch 5	**0.735**	**0.897**
CAD_{DL} Epoch 6	**0.735**	**0.897**
Kwak et. al	0.617	0.859

The ROC plots for four different models are presented in Figure 6. The first three are for CAD_{DL} models at different epochs, and the last curve is for CAD_{SVM}. As the model continues to train, CAD_{DL} curves move toward the y-axis, while simultaneously showing lower detection values for higher FPRs: at 0.3 FPR epoch-1 has the highest performance with 0.92 detection rate while epoch-6 has the highest performance at 0.1 FPR with 0.731 detection rate. Furthermore, the model at epoch-6 is either equivalent to or outperforms CAD_{SVM} at every FPR while the other two CAD_{DL} models have an inferior performances for low FPRs. A trend is apparent when inspecting the FROC curves for clinically significant lesions, which have a Gleason score of 7 or above. As

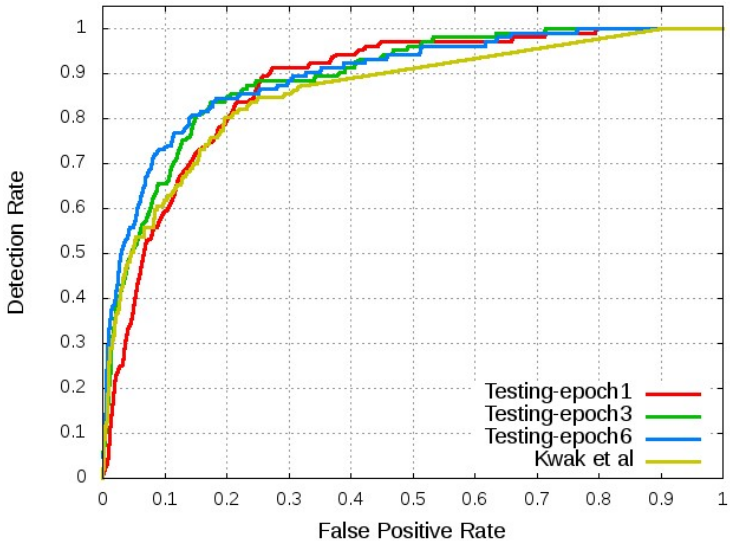

Figure 6. ROC curves for CAD_{DL} at the multiple training stages denoted by epochs and for competing CAD by Kwak et. al. (Ref. 6).

shown in Figure 7, with increasing epoch number the performance of CAD_{DL} increases. The model at epoch-6 shows the best performance with 0.94 detection rate at a 10 false-positives per patient. This model is either equivalent to or outperforms CAD_{SVM} for all FPRs while the epoch-1 model has an inferior performance with a 0.80 detection rate as compared to CAD_{SVM}'s 0.85 for 10 false-positives per patient.

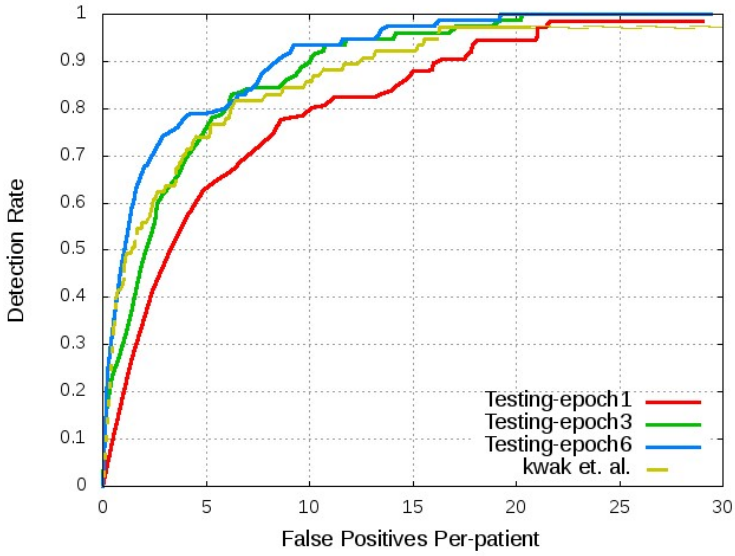

Figure 7. FROC curves for both CAD_{DL} and the competing CAD by Kwak et. al (Ref. 6) for clinically significant tumors (Gleason Score $\geq$ 7.)

Figures 8-11 give insight into the qualitative performance of the CADs. Figure 8 shows a detection of a transition zone lesion that appears suspicious on all MR sequences –lesions usually have low intensity on T2W and ADC images while appearing bright on B2000. CAD_{DL} detects it with high confidence while generating no false-positives. CAD_{SVM} has a true-positive detection as well, but presents false-positive detections visible

in the peripheral zone. Benign prostatic hyperplasia (BPH) can be erroneously identified as cancerous by both

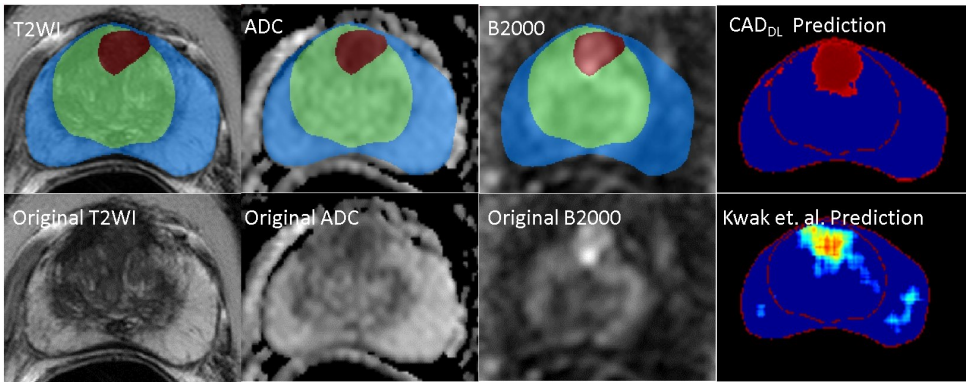

Figure 8. An example of a true-positive detection. The top row presents the prostate mask (blue), central gland (green), and ground-truth tumor (red) overlaid upon the T2W, ADC, and B2000 images respectively. The corresponding raw T2W, ADC, and B2000 images are shown in the bottom row. Prediction heat-maps from CAD_{DL} and the competing CAD by Kwak et. al (Ref. 6) are placed in the last column.

CADs as illustrated in Figure 9. However, in this instance only CAD_{DL} manages to generate high probability values for the peripheral zone lesion that cohabits this region of the prostate. Some lesions can be difficult to

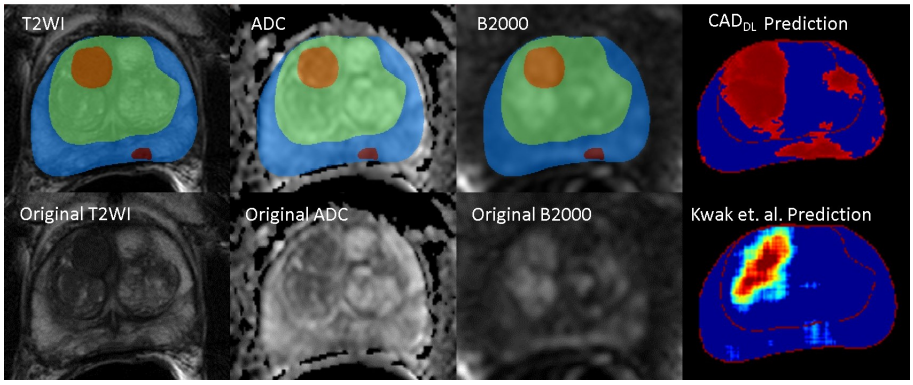

Figure 9. An example of a benign prostatic hyperplasia (BPH) and a lesion detection. The top row displays the prostate mask (blue), central gland (green), ground-truth tumor (red), and BPH (orange) overlaid upon the T2W, ADC, and B2000 images respectively. The corresponding raw T2W, ADC, and B2000 images are shown in the bottom row. Prediction heat-maps from CAD_{DL} and the competing CAD by Kwak et. al (Ref. 6) are in the last column.

identify for the CADs, such as the small peripheral zone tumor presented in Figure 10. In this example, both CADs incorrectly predict a large tumor in the transition zone where there are no marked regions of interest. Lastly, false predictions can also arise from an artifact within the image. Figure 11 presents an example where there is a blurry region in the peripheral zone of the T2W image. CAD_{DL} incorrectly predicts suspicion in this region with high probability while CAD_{SVM} remains largely unaffected by the low quality image.

4. DISCUSSION

In general, with more training iterations CAD_{DL}'s performance increases suggesting that the architecture is learning discriminating features (Figure 5). However, the learning progress slows down with increasing number of epochs reaching its maximum performance at epoch 5 (Table 2). Due to a limited data-set size, the CAD achieves its optimal performance after a small number of epochs. In addition, with more training the number of false-positives decreases, but at the same time, the number of true-positives decreases –albeit at a lower rate–

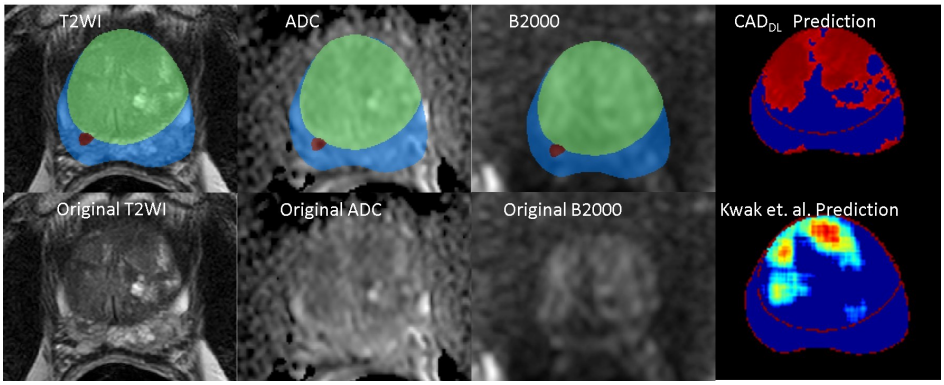

Figure 10. An example of a true-negative and a false-positive detection. The top row has the prostate mask (blue), central gland (green), and ground-truth tumor (red) displayed over T2W, ADC, and B2000 images respectively. The corresponding raw T2W, ADC, and B2000 images are shown in the bottom row. Prediction heat-maps from CAD_{DL} and the competing CAD by Kwak et. al (Ref. 6) are placed in the last column.

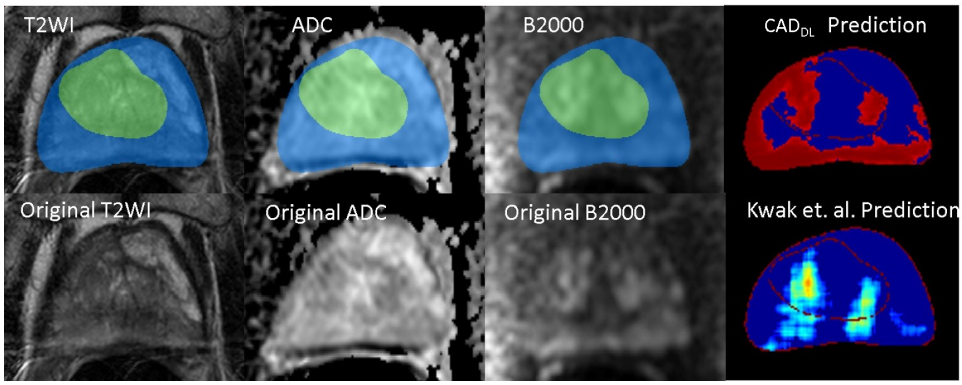

Figure 11. An example of a false-positive detection arising from what may be an artifact within the T2W image. The top row presents the prostate mask (blue), and central gland (green) overlaid T2W, ADC, and B2000 images respectively. The corresponding raw T2W, ADC, and B2000 images are shown in the bottom row. Prediction heat-maps from CAD_{DL} and the competing CAD by Kwak et. al (Ref. 6) are placed in the last column.

evidenced by the ROC curves that shift toward the y-axis while decreasing in detection rate for high false-positive rates (Figure 6). This suggests that the architecture is overfitting the training data resulting in missed detection of tumors found in the test-set that may be underrepresented in the training-set. A small data-set makes it likely that the training-set excludes some prostate lesion types that may be found in the test-set. Since CNNs learn important features from the data, the CAD will be unable to detect such lesion types because of the lack of corresponding features.

CAD_{DL} further shows potential for clinical application (Figure 7). It outperforms a published prostate-CAD with a margin of roughly 10% detection rate for 10 false-positives per patient despite a smaller training-set. If trained on a more extensive data-set, CAD_{DL} has the potential to perform at a higher detection rate for clinically significant lesions. Furthermore, the generated prediction maps have high probability values making them less ambiguous to interpret (Figure 8). False-positive predictions such as benign prostatic hyperplasia (BPH) (Figure 9) and those arising from artifacts in the image (Figure 11) exhibit high probability signatures as well; however, whether the CAD should be penalized for these types of erroneous predictions is debatable. Furthermore, the CAD_{DL} tends to miss small lesions (Figure 10) or overestimate the size of a lesion (Figure 9). This may be as a result of a lack of small-tumor example types in the training-set, or large receptive field sizes for the various layers (Figure 4) that are unable to capture relevant features for small lesions.

A small study cohort is one limitation in our study. We investigate CAD_{DL} performance over multiple epochs

instead of finding the best performing model using a validation-set. With a large enough data-set, it will be possible to split the data into training, validation, and testing sets to thoroughly characterize the performance of CAD_{DL}. A second limitation in our study is the loss of information during image compression to satisfy the constraints of using publicly available CNN architectures. Even though we apply a histogram equalization algorithm to convert the medical images from a 12-16 bit to 8 bit images, there remains a small amount of information loss. Despite these limitation, we hope that our experimental methods and results elucidate the challenges that accompany the application of DCNN architectures to medical images and ways to circumvent such challenges.

As a future study, we plan to investigate a weakly supervised learning scheme that can use the more ubiquitous database types containing biopsy points and histopathology images for training. Acquiring a large enough data-set with ground-truth tumor annotations is difficult because it required the expertise of experienced healthcare providers, who are preoccupied with patient care. Furthermore, our chosen CNN architecture can be modified for multi-label classification, which will allow us to identify lesions by severity in addition to detecting them. Lastly, we will attempt to develop an automatic segmentation tool for tumor size prediction.

5. CONCLUSION

We conclude that deep convolutional neural networks (DCNNs) have the potential to improve prostate cancer detection on multiparametric MR images (mpMRI). We demonstrated that our proposed DCNN-based prostate CAD (CAD_{DL}) is capable of learning lesion discriminating features from the given mpMR images. CAD_{DL} further achieved promising results when compared to a published prostate CAD and showed potential for clinical application. In the future, we hope to develop a CAD that has many more applications in addition to tumor detection such as severity and tumor size prediction.

ACKNOWLEDGMENTS

This work was supported by the Intramural Research Programs of the National Institutes of Health, Clinical Center and National Cancer Institute.

REFERENCES

[1] Jemal, A., Murray, T., Ward, E., Samuels, A., Tiwari, R. C., Ghafoor, A., Feuer, E. J., and Thun, M. J., "Cancer statistics, 2005," *CA: a cancer journal for clinicians* **55**(1), 10–30 (2005).

[2] Costa, D. N., Pedrosa, I., Donato Jr, F., Roehrborn, C. G., and Rofsky, N. M., "Mr imaging–transrectal us fusion for targeted prostate biopsies: implications for diagnosis and clinical management," *Radiographics* **35**(3), 696–708 (2015).

[3] Cornud, F., Delongchamps, N., Mozer, P., Beuvon, F., Schull, A., Muradyan, N., and Peyromaure, M., "Value of multiparametric mri in the work-up of prostate cancer," *Current urology reports* **13**(1), 82–92 (2012).

[4] Ruprecht, O., Weisser, P., Bodelle, B., Ackermann, H., and Vogl, T. J., "Mri of the prostate: interobserver agreement compared with histopathologic outcome after radical prostatectomy," *European journal of radiology* **81**(3), 456–460 (2012).

[5] Carlsson, S., Vickers, A. J., Roobol, M., Eastham, J., Scardino, P., Lilja, H., and Hugosson, J., "Prostate cancer screening: facts, statistics, and interpretation in response to the us preventive services task force review," *Journal of Clinical Oncology* **30**(21), 2581–2584 (2012).

[6] Kwak, J. T., Xu, S., Wood, B. J., Turkbey, B., Choyke, P. L., Pinto, P. A., Wang, S., and Summers, R. M., "Automated prostate cancer detection using t2-weighted and high-b-value diffusion-weighted magnetic resonance imaging," *Medical physics* **42**(5), 2368–2378 (2015).

[7] Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., and Huisman, H., "Computer-aided detection of prostate cancer in mri," *IEEE transactions on medical imaging* **33**(5), 1083–1092 (2014).

[8] Wang, S., Burtt, K., Turkbey, B., Choyke, P., and Summers, R. M., "Computer aided-diagnosis of prostate cancer on multiparametric mri: a technical review of current research," *BioMed research international* **2014** (2014).

[9] LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning," *Nature* **521**(7553), 436–444 (2015).

[10] Cheng, R., Turkbey, B., Gandler, W., Agarwal, H. K., Shah, V. P., Bokinsky, A., McCreedy, E., Wang, S., Sankineni, S., Bernardo, M., et al., "Atlas based aam and svm model for fully automatic mri prostate segmentation," in [*2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*], 2881–2585, IEEE (2014).

[11] Xie, S. and Tu, Z., "Holistically-nested edge detection," in [*Proceedings of the IEEE International Conference on Computer Vision*], 1395–1403 (2015).