BIOPSY-GUIDED LEARNING WITH DEEP CONVOLUTIONAL NEURAL NETWORKS FOR PROSTATE CANCER DETECTION ON MULTIPARAMETRIC MRI

Yohannes Tsehay[a], *Nathan Lay*[a], *Xiaosong Wang*[a], *Jin Tae Kwak*[a], *Baris Turkbey*[b],
Peter Choyke[b], *Peter Pinto*[b], *Brad 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 highly prevalent and is the second most common cause of cancer-related deaths in men. Multiparametric MRI (mpMRI) is robust in detecting PCa. We developed a weakly supervised computer-aided detection (CAD) system that uses biopsy points to learn to identify PCa on mpMRI. Our CAD system, which is based on a deep convolutional neural network architecture, yielded an area under the curve (AUC) of 0.903±0.009 on a receiver operation characteristic (ROC) curve computed on 10 different models in a 10 fold cross-validation. 9 of the 10 ROCs were statistically significantly different from a competing support vector machine based CAD, which yielded a 0.86 AUC when tested on the same dataset (α = 0.05). Furthermore, our CAD system proved to be more robust in detecting high-grade transition zone lesions.

Index Terms— Biopsy Database, Prostate, Holistically-nested Edge Detection, Computer-Aided Detection, Prostate-CAD, Radiology

1. INTRODUCTION

In 2015, an estimated 220,800 new cases and 27,540 deaths were recorded making PCa the second most common cause of cancer related deaths in men[1]. Multiparametric MRI (mpMRI) is the most accurate imaging method for PCa detection [2]; however, it requires the expertise of experienced radiologists. Moreover, there is inconsistency across readers of varying experience [3]. State-of-the-art diagnostic method uses transrectal ultrasound (TRUS)-MRI fusion-guided biopsy[4], where MRI registered TRUS imaging is used to direct the biopsy needle. In order to less frequently overlook clinically significant lesions, it is important that MRI based detections have a high sensitivity. To increase inter-reader consistency and sensitivity, we developed a computer-aided detection (CAD) system that can automatically detect lesions on mpMRI that readers can use as a reference.

Due to the recent success in deep-learning algorithms that can directly learn discriminative features from the given data [5], we investigated a deep convolutional neural network (DCNN) to find an improved solution for PCa detection on mpMRI. Many of the publicly available DCNN architectures utilize fully-supervised learning in which every pixel is given a ground-truth label. However, a database with pixel-wise annotation for PCa is difficult to find and making one is time consuming and expensive because it requires the expertise of healthcare providers. Therefore, we developed a CAD system that learns in a weakly supervised fashion where biopsy points are used as ground-truth labels. Our approach can be quite useful because biopsy point databases are much easier to acquire. Furthermore, TRUS-MRI fusion-guided biopsy's accuracy is limited due to inherent MR and TRUS image registration errors. Thus, a CAD system that is robust against ambiguity resulting from weakly labeled training examples is essential to apply biopsy databases for an automatic cancer detection tasks.

We performed a detailed analysis of our CAD system's performance and compared with an existing prostate CAD system that uses hand-crafted features to differentiate between cancerous and normal prostate tissue. Overall, our CAD system had a higher performance with a 0.903±0.009 area under the receiver operation characteristic (ROC) curve computed on 10 different models in a 10 fold cross-validation. The competing CAD with ROC 0.86 resulted in 9 out of 10 statistically significantly different outcomes when compared to the 10 models from our proposed CAD using ROCkit (α = 0.05). Furthermore, our CAD had a better detection rate for high-grade transition zone lesions. Lesions in this region are known to be difficult to detect using traditional techniques found in [6, 7, 8].

2. MATERIALS AND METHODS

2.1. Patient Cohort

The patient cohort consisted of 196 patients. Each patient underwent mpMR scanning, and each one has T2W, ADC, and B2000 axial image volumes of the prostate. Reference standard segmentations of the prostate and transition zone were provided by an expert radiologist Fig.1. The patients then underwent TRUS-MRI fusion-guided biopsies. The PCa Gleason score distribution of these biopsies with respect to peripheral and transition zone is given in Table 1.

2.2. Image Pre-processing

Publicly available DCNN architectures usually work with images limited to 8 bit channels with intensity values between 0 and 255, but mpMRI images may have 12-16 bit channels with intensity ranges of 0 to upwards of 4,000. To minimize information loss, a histogram equalization algorithm was used to map pixel intensity values from the mpMRI sequences to values between 0 and 255. The

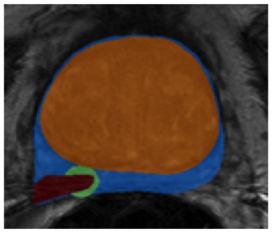

Fig. 1. Prostate mask in blue, transition zone in orange, lesion in red and biopsy point with a 5mm radius in green are displayed over a T2W image of the prostate.

Table 1. Biopsy points for all patients stratified by Gleason score and zone.

Zone	Benign	G6	G7	G8	G9	Total
Peripheral	127	20	35	18	4	204
Transition	163	22	28	15	13	241
Overall	290	42	63	33	17	445

compressed images were then combined into a three-channel RGB image as shown in Fig. 2.

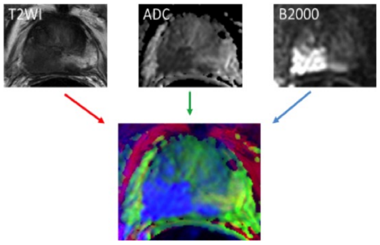

Fig. 2. RGB image generated using compressed mpMRI images.

2.3. Ground-truth Annotation

All the patients had undergone TRUS-MRI fusion-guided biopsy, and our database contains the x, y and z coordinates and a corresponding Gleason scores. Furthermore, our database has 52 patients with contour annotations performed by an experienced radiologist. We simply label any voxel within 5mm radius around the biopsy point as cancerous following the rationale that an average prostate lesion has a diameter of 10mm [9] to generate ground-truth segmentations from biopsy points for the remaining 144 patients lacking contour annotations. All lesion types are given a positive label because the TRUS-MRI fusion-guided biopsy follows a radiologist's recommendation, which makes the biopsy area a region of interest despite a benign biopsy result. This is further justified because MR to ultrasound image registration is erroneous and can lead to an inaccurate TRUS-MRI fusion-guided biopsy results. Fig. 1 shows an example of a peripheral zone tumor and the biopsy point.

2.4. CNN Architecture

We adopted the same CNN architecture as [10], which is composed of 5 convolutional layers. It is an end-to-end network that takes full images as input and outputs the corresponding prediction maps. Each convolutional layer has its own loss function which is calculated using the ground-truth image allowing the network to learn in a deeply-supervised fashion. Furthermore, each convolutional layer has side-outputs that are subsequently combined for a final fused output. Thus, the final output is a product of highly refined, hierarchical features. The hyper-parameters used are as follows: learning rate was set to 1.0e-8, weight decay was set to 0.0002, and momentum was set to 0.9.

2.5. Analysis

2.5.1. Cross-validation

The 52 cases with contour segmentations are set aside for the testing set, and the remaining 144 patients with biopsy points are used to train the CNN models. A 10 fold cross-validation is used for the training stage, and the training set is split into train and validation sets (90% for training and 10% for validation).

2.5.2. ROC analysis

ROC analysis was computed on 3D probability maps: if the CAD probability scores lie within a contour segmentation volume and the 90$^{\text{th}}$ percentile of the probability scores are greater than a threshold, then the lesion is said to be detected. The CAD probability map volume is divided into 3x3 voxel cells, and if 90$^{\text{th}}$ percentile of probability scores within a cell is greater than a probability threshold and more than 3mm away from the hand drawn segmentation volume, it is considered to be a false positive detection. The ROC curves from each model in the 10 fold cross-validation experiment are averaged to get a single curve, and the 95% confidence interval is calculated and plotted alongside the average curve as shown in Fig. 3.

2.5.3. FROC analysis

Free-response ROC (FROC) analysis was also done in three dimensions. It uses information from the biopsy point database and a number of scores acquired from a probability map using non-maximum suppression (NMS) with a 10mm x 10mm x 10mm window. These local maxima points acquired with NMS are paired with the nearest biopsy point, and if the distance between the NMS point and the nearest biopsy point is greater than 10mm, it is considered a false positive detection. For cases where an NMS point is within a range of multiple biopsy points with different Gleason scores, the FROC tool is designed to consider tumor grade and the probability score; such that, it pairs high probability values with more severe tumor grades. The FROC curves from each fold in a 10 fold cross-validation are averaged to generate a single curve for each tumor type as shown in Fig. 4 and Fig. 5. Fig. 5 further includes 95% confidence interval plots.

2.5.4. Compare with a Competing CAD system

Our proposed DCNN based prostate CAD (CAD$_{DCNN}$) system was compared to a published SVM based CAD (CAD$_{SVM}$) by [6]. CAD$_{SVM}$ uses local binary pattern features extracted from T2W and b-2000 images as features to an SVM classifier. Similar to our method, it was trained on examples of biopsies. It was run on the 52 cases with tumor contour segmentations, which are also the test-set for the CAD$_{DCNN}$. The resulting probability maps were used to generate ROC and FROC plots. ROCkit was used to compute p-values to quantify the difference between the 10 CAD$_{DCNN}$ models

Table 2. Training set stratified by zone and Gleason score

Zone	Benign	G6	G7	G8	G9	Total
Peripheral	103	15	16	8	1	143
Transition	148	14	7	2	6	177
Overall	251	29	23	10	7	320

and the single CAD$_{SVM}$ model which were all run on the same testing set. Probability map thresholds of 0.95 and 0.5 were applied to CAD$_{DCNN}$ and CAD$_{SVM}$ respectively to generate images for qualitative analysis.

3. RESULTS

3.1. Training Data

Table 2 presents all the lesions for our training set stratified by Gleason score and zone. There were only a few training examples for Gleason 9, and there were more lesions in the transition zone than peripheral zone.

3.2. Quantitative Results

At 0.2 false-positive rate CAD$_{DCNN}$'s average performance was 0.86 and CAD$_{SVM}$'s was 0.81 as shown in Fig. 3. The performance achieved by CAD$_{SVM}$ at 0.2 false positive rate lay outside the 95% confidence interval. The area under the curve for the two ROC plots were 0.903±0.009 and 0.86 respectively. In comparing the 10 models of CAD$_{DCNN}$ with the single model of CAD$_{SVM}$ using ROCkit, 9 out of 10 of the models were statistically significantly different ($\alpha = 0.05$).

CAD$_{DCNN}$ outperformed CAD$_{SVM}$ at Gleason 8 and 9 tumor detection showing the best performance for Gleason 8 at 96% detection rate at 6 false positives per-patient (Fig. 4). However, the latter detected tumors with Gleason scores 6 and 7 at a higher rate: for an average of 13 false positives per patient, it had a 100% detection rate for both tumor types as shown in Fig. 4. CAD$_{DCNN}$ especially performed better for high-grade tumors (Gleason 8 and 9) located in the transition zone as shown in Fig. 5. It has a 94% detection rate for an average of 2 false positives per patient. At the same false positive rate the performance of CAD$_{SVM}$ lay outside the 95% confidence interval with an 80% detection rate. Fig. 5 also presents peripheral zone FROC curves for the high-grade lesions for the two CADs, and the two CADs are comparable.

3.3. Qualitative Results

Fig. 6 and 7 show rendered probability maps and the original images alongside prostate, central gland, biopsy and lesion segmentations. Fig. 6 has two lesions: one in the transition zone with a Gleason score of 6 and another in the peripheral zone with a benign biopsy result. Both CAD$_{DCNN}$ and CAD$_{SVM}$ detect the transition zone lesion but fail to detect the peripheral zone lesion. Benign prostate hyperplasia (BPH) detection is shown in Fig. 7.

4. DISCUSSION

Overall, the CAD$_{DCNN}$ performed both qualitatively and quantitatively better than the CAD$_{SVM}$. The advantages of CAD$_{DCNN}$ were especially apparent for high-grade lesions found in the transition zone. This is most likely because transition zone lesions were well represented in the training dataset. The good transition zone performance achieved here may prove clinically beneficial since traditional techniques struggle to detects these lesion types. The training data distribution shown Table 2 explains why our CAD system performed comparably for Gleason 9 and peripheral zone lesions. Having more training examples for the underrepresented lesion types can potentially increase the CAD's performance.

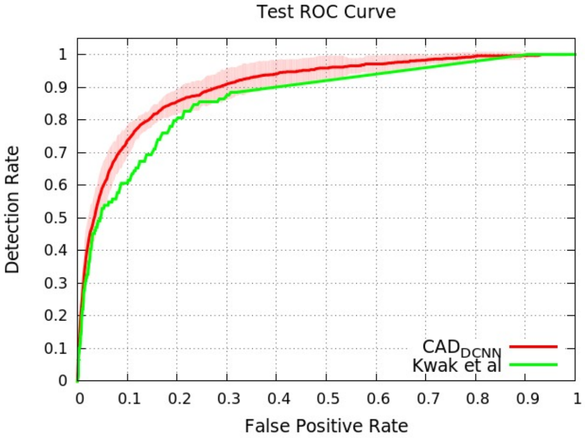

Fig. 3. ROC curves our proposed CAD (CAD$_{DCNN}$) and a previously published CAD by Kwak et al. The shaded region is the 95% confidence interval of CAD$_{DCNN}$.

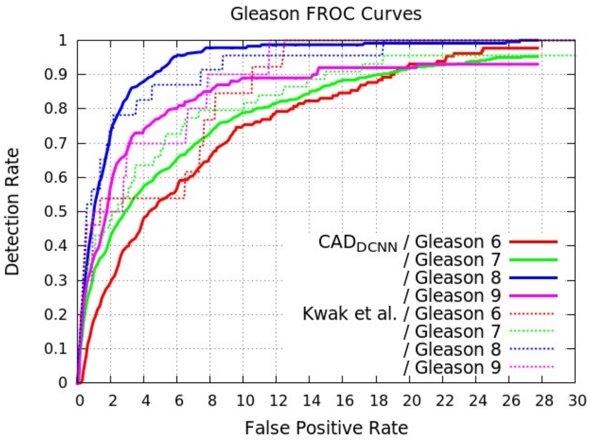

Fig. 4. Per-Gleason FROC curves for our proposed prostate CAD (CAD$_{DCNN}$) along with the results from a competing CAD by Kwak et al.

The probability maps by CAD$_{DCNN}$ are clean and easy to read as shown in Fig. 6, and 7. The CAD can miss lesions as illustrated in Fig. 6; however in this instance, the biopsy result is benign. This seems to be a recurring theme, and it presents a limitation in our data. The discrepancies between annotations and biopsy results can be as a result of a missed biopsy or an inaccurate annotation by the expert. Nonetheless, the CAD's ROC curve is affected by these types of ambiguous regions. The presence of BPH as in Fig. 7 can also impact the CADs performance. Whether the CAD should be penalized for detecting BPHs is debatable.

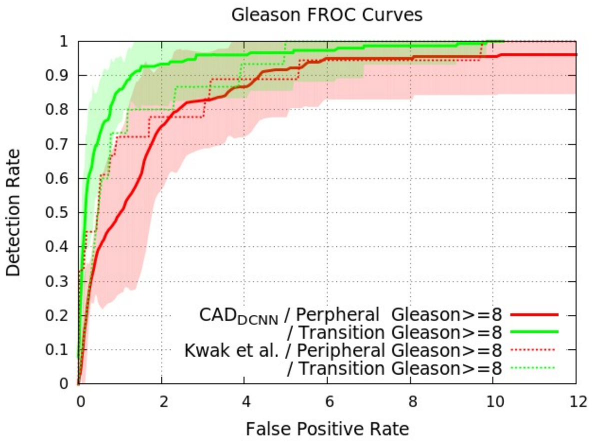

Fig. 5. FROC curves for high-grade lesions found in the peripheral and transition zone plotted separately. The filled-in regions represent the 95% confidence interval for our proposed prostate CAD (CAD$_{DCNN}$).

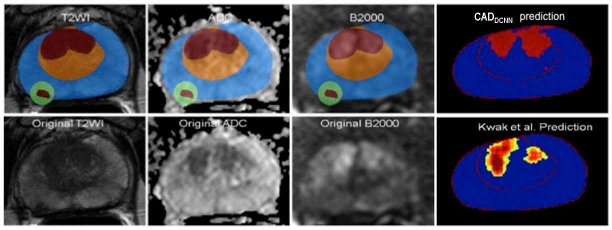

Fig. 6. Transition zone lesion detection and missed peripheral zone lesion. The proposed prostate CAD (CAD$_{DCNN}$) and Kwak et al. CAD outputs are displayed in the last column. Prostate mask (blue), central gland (orange), tumor segmentation (red), and biopsy point contour (green) are shown in the top row. The bottom row shows the original mpMRIs.

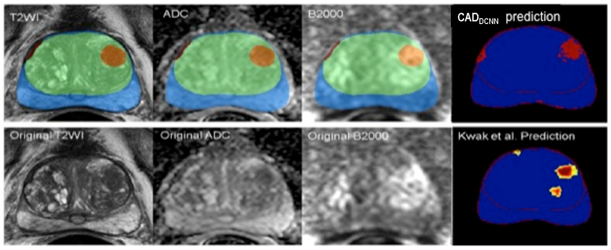

Fig. 7. Benign prostatic hyperplasia (BPH) detection by both CADs (to is for proposed prostate CAD (CAD$_{DCNN}$)) is displayed in the last column. Prostate mask (blue), central gland (green), tumor segmentation (red), and BPH (orange) are shown in the top row. The bottom row shows the original mpMRIs

5. CONCLUSION

Despite a limited ground-truth annotation resulting from the use of biopsy points as the reference standard, our CAD system performs better than an existing CAD. It proved to be more effective at detecting high-grade lesions found in the transition zone, a feat that has proven to be difficult for traditional CADs. Our proposed CAD demonstrated the potential of applying a weakly labeled image data for a supervised learning task of prostate cancer detection on mpMRI.

6. ACKNOWLEDGMENTS

This work was supported by the Intramural Research Programs of the National Institutes of Health, Clinical Center and National Cancer Institute. It also utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov)

7. REFERENCES

[1] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal, “Cancer statistics, 2015,” *CA: a cancer journal for clinicians*, vol. 65, no. 1, pp. 5–29, 2015.

[2] F Cornud, NB Delongchamps, P Mozer, F Beuvon, A Schull, N Muradyan, and M Peyromaure, “Value of multiparametric mri in the work-up of prostate cancer,” *Current urology reports*, vol. 13, no. 1, pp. 82–92, 2012.

[3] Oliver Ruprecht, Philipp Weisser, Boris Bodelle, Hanns Ackermann, and Thomas J Vogl, “Mri of the prostate: interobserver agreement compared with histopathologic outcome after radical prostatectomy,” *European journal of radiology*, vol. 81, no. 3, pp. 456–460, 2012.

[4] Daniel N Costa, Ivan Pedrosa, Francisco Donato Jr, Claus G Roehrborn, and Neil M Rofsky, “Mr imaging–transrectal us fusion for targeted prostate biopsies: implications for diagnosis and clinical management,” *Radiographics*, vol. 35, no. 3, pp. 696–708, 2015.

[5] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] Jin Tae Kwak, Sheng Xu, Bradford J Wood, Baris Turkbey, Peter L Choyke, Peter A Pinto, Shijun Wang, and Ronald M Summers, “Automated prostate cancer detection using t2-weighted and high-b-value diffusion-weighted magnetic resonance imaging,” *Medical physics*, vol. 42, no. 5, pp. 2368–2378, 2015.

[7] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman, “Computer-aided detection of prostate cancer in mri,” *IEEE transactions on medical imaging*, vol. 33, no. 5, pp. 1083–1092, 2014.

[8] Shijun Wang, Karen Burtt, Baris Turkbey, Peter Choyke, and Ronald M Summers, “Computer aided-diagnosis of prostate cancer on multiparametric mri: a technical review of current research,” *BioMed research international*, vol. 2014, 2014.

[9] Tineke Wolters, Monique J Roobol, Pim J van Leeuwen, Roderick CN van den Bergh, Robert F Hoedemaeker, Geert JLH van Leenders, Fritz H Schröder, and Theodorus H van der Kwast, “A critical analysis of the tumor volume threshold for clinically insignificant prostate cancer using a data set of a randomized screening trial,” *The Journal of urology*, vol. 185, no. 1, pp. 121–125, 2011.

[10] Saining Xie and Zhuowen Tu, “Holistically-nested edge detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1395–1403.