






Performance of Deep CNN and Radiologists in Prostate Cancer Classification: A Comparative Pilot Study

Piotr Sobecki^(✉) , Rafał Józwiak , and Ihor Mykhalevych 

National Information Processing Institute, Warsaw, Poland
{piotr.sobecki, rafal.jozwiak, ihor.mykhalevych}@opi.org.pl

Abstract. In recent years multiple deep-learning solutions have emerged that aim to assist radiologists in prostate cancer (PCa) diagnosis. Most of the studies however do not compare the diagnostic accuracy of the developed models to that of radiology specialists but simply report the model performance on the reference datasets. This makes it hard to infer the potential benefits and applicability of proposed methods in diagnostic workflows. In this paper, we investigate the effects of using pre-trained models in the differentiation of clinically significant PCa (csPCa) on mpMRI and report the results of conducted multi-reader multi-case pilot study involving human experts. The study aims to compare the performance of deep learning models with six radiologists varying in diagnostic experience. A subset of the ProstateX Challenge dataset counting 32 prostate lesions was used to evaluate the diagnostic accuracy of models and human raters using ROC analysis. Deep neural networks were found to achieve comparable performance to experienced readers in the diagnosis of csPCa. Results confirm the potential of deep neural networks in enhancing the cognitive abilities of radiologists in PCa assessment.

Keywords: Deep learning · Prostate Cancer · Computer Aided Diagnosis

1 Introduction

In light of the increasing incidence rate of prostate cancer (PCa) over the previous years [2], there is a global focus on providing modern solutions that can address this growing health issue. Noninvasive diagnostics based on multiparametric magnetic resonance imaging (mpMRI) became essential in clinical decision-making as it enables more accurate risk stratification and therefore plays an important role in selecting patients for biopsy and direct targeting of lesions [6, 11].

Radiological assessment of the prostate gland involves the interpretation and reporting of mpMRI examinations according to the established global standards. The current version of the standardized prostate MRI assessment Prostate Imaging-Reporting and Data System (PI-RADS v2.1) [8], provides an approach

to the interpretation and reporting of PCa examinations. The system assumes the evaluation of each mpMRI sequence separately. Each lesion assessment category is established on a 5-point scale according to the assessment algorithm involving previously scored sequences. Introducing the standard in diagnostic practice improved the diagnostic accuracy of performed examinations and improved the availability of the method. Low specificity, however, remains a considerable aspect of MRI assessment and clinically significant (cs) PCa differentiation, potentially leading to unnecessary biopsies. Because of the assessment complexity and steep learning curve, it is mainly the case for radiologists with low experience in prostate MRI reporting [10].

Recently, solutions based on machine learning have achieved promising results in applications in PCa diagnostics. A PCa classification challenge held in 2017 (ProstateX) provided a way of comparing tools of automatic PCa differentiation based on mpMRI [1]. 71 competing methods were evaluated on the lesion classification task. The area under the receiver operating characteristic curve (AUC) of submitted models ranged between 0.45 to 0.87 AUC. The top three scoring teams achieved results of $AUC = 0.84$ and 0.87 . We used the ProstateX dataset to develop and validate the deep convolutional neural network (CNN) model that achieved $AUC = 0.84$ on the test ProstateX dataset [7].

Narrative Review by Twilt et. al. [9] presents an overview of recently proposed tools (between 2018 and 2022) that have been suggested to aid in the diagnosis of PCa. Overall deep learning (DL) solutions achieve the highest performance on PCa detection and diagnosis tasks. Computational models show the potential in enhancing the diagnostic processes and increasing the specificity of mpMRI assessment. However, only a limited number of studies validated the results in clinical workflows - 85% of them report only stand-alone model diagnostic accuracy [9]. It remains a question of how the diagnostic accuracy of DL solutions relates to that of radiology experts and what could be expected from the integration of computational models in diagnostic workflows.

The objective of our study was to evaluate the diagnostic accuracy of radiologists with various levels of diagnostic experience in comparison to the proposed DL solution for csPCa differentiation.

2 Methods

The retrospective study design involved the assessment of 32 suspicious lesions by six radiologists and the deep CNN model in a multi-case multi-reader (MCMR) setting.

2.1 Dataset

A group of cases from a publicly available database of annotated mpMRI data were used in the study [3]. Complete mpMRI data (T2W, DCE, DWI, and ADC sequences) were included for all cases in the database. We have selected a thirty-two lesion dataset diversified according to its clinical significance based on the results of a histopathological evaluation.

The selected dataset contained:

- 14 PZ lesions (7 cs and 7 not cs),
- 11 TZ lesions (5 cs and 6 not cs),
- 7 AS lesions (4 cs and 3 not cs).

2.2 Radiological Assessment Study

The study was carried out by a group of specialists with diversified expertise. Six radiologists involved in the study had practical experience in PCa diagnosis based on the mpMRI (three specialists with diagnostic experience of one to five years; three specialists with more than ten years of diagnostic experience, and at least five years of experience using the PI-RADS standard). Those groups of experts are referred to in the paper as experienced and inexperienced raters. The participating experts did not interact with each other during the assessment phase.

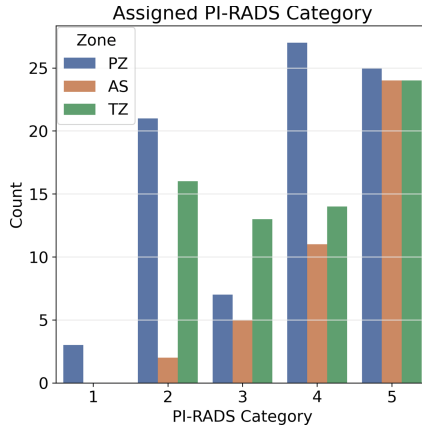


Fig. 1. PI-RADS score evaluations for lesions in the dataset. Only three assessments assigned lesions to the PI-RADS 1 category (only PZ lesions).

Experts participating in the study were not involved in the dataset selection, development of the study methodology, and experiment results analysis.

The results of the assessments are presented in the Fig. 1. Even though the dataset was balanced (close to an equal rate of cs and non-cs lesions) the distribution of assigned scores did not reflect that.

2.3 Deep CNN Model

The model evaluated in this study was a multi-modal deep CNN network of VGG-inspired architecture, adapted to the input sequence resolution and

problem complexity. Introduced modifications reduced the number of trainable parameters. The developed model architecture design reflected a PI-RADS category assessment algorithm based on lesion zonal location. This was done by integration of output routing and using complex loss function for optimization. Resulting predictions were assigned using two subnetworks designed to base predictions on T2W and DWI (subnetwork for TZ, AS, and SV lesions) and on DWI/ADC and DCE (subnetwork for PZ lesions). The model has been evaluated on the test reference dataset and resulted in a score corresponding to the state-of-the-art results (AUC = 0.84) [1]. Model architecture and analysis of achieved diagnostic accuracy have been described in the previously published study[7].

CNN predictions were made on unseen samples using 5-fold cross-validation and collecting validation split classification results.

2.4 Probability Mapping

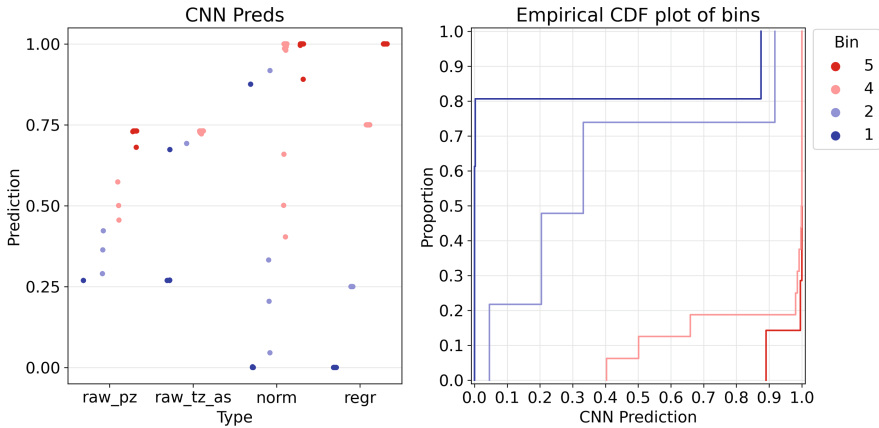


Fig. 2. The left figure shows the results of a mapping of raw CNN predictions to bins corresponding to PI-RADS categories. Separate raw predictions for PZ (raw_pz) and TZ and AS lesions (raw_tz_as) show different response characteristics resulting from separate network optimization processes. The right figure presents the empirical cumulative distribution function (ECDF) of bins in relation to normalized CNN output.

It could be argued that continuous predictions resulting from softmax output layers can produce superior AUC results in comparison to the ordinal estimations made by human experts using the Likert scale. Therefore, to further evaluate the diagnostic characteristics of the proposed model, we have mapped the raw continuous CNN predictions to bins corresponding to PI-RADS scores (Fig. 2). Bin discretization involved several steps that mapped the raw CNN predictions to ordinal categories. We have used the mode of PI-RADS assessments for lesions resulting from a radiological assessment study as ground truth for labels.

First, continuous outputs resulting from PZ and TZ/AS sub-networks were normalized to the range of [0,1]. Leave-one-out cross-validated estimates of bins were obtained for each lesion using two ordinal regression models[4] (separate each subnetwork predictions). This resulted in the mapping of continuous network output to the Likert scale that reflected the PI-RADS category characteristics based on the CNN prediction (Fig. 2).

We have mapped the 5-point Likert scale of manually assigned PI-RADS categories and automatically estimated bins to the [0, 0.25, 0.5, 0.75, and 1] probability values to perform the ROC analysis.

2.5 Statistical Analysis

We compare the model performance with that of experienced and inexperienced radiology specialists using assessments collected during the retrospective study. To evaluate the differences, we have used the Area under the Receiver operating characteristic Curve (AUC) as a measure of diagnostic accuracy. Extensive simulations using bootstrap re-sampling (1000 tests) [5] were conducted to construct 95% confidence intervals and perform hypothesis testing in various scenarios. We have conducted separate experiments to compare the diagnostic characteristics in relation to the combinations of assessment methods (CNN, human raters), lesion location (PZ, TZ, and AS), and examiner experience. An alpha of 0.05 was used as the cutoff for statistical significance (we additionally report test results with an alpha of 0.1 due to the small dataset sample size).

3 Results

The following section presents the results of the comparison of diagnostic accuracy between inexperienced, experienced radiologists and model predictions. Additionally, we report the change in diagnostic accuracy resulting from the integration of human and CNN assessments.

3.1 Results of Raw CNN Predictions

The results achieved by the CNN model (AUC = 0.83, CI [0.80, 0.88]) demonstrated superior diagnostic accuracy in comparison with both:

experienced (AUC = 0.80, $p > .1$, CI [0.74, 0.86]) and
 inexperienced (AUC = 0.71, $p < .1$, CI [0.63, 0.80])

specialists in the evaluation of lesions' clinical significance using the PI-RADS v2.1 standard.

The lowest diagnostic accuracy has been observed for AS lesions, where CNN solution provides higher quality estimations in comparison both to experienced and inexperienced radiologists (AUC = 0.79 vs. AUC = 0.64 vs. AUC = 0.59). In the case of other lesion locations, the differences between the neural network and the experienced radiology specialists were less pronounced: PZ (AUC = 0.88 vs.

AUC = 0.85 vs. AUC = 0.72) and TZ (AUC = 0.89 vs. AUC = 0.84 vs. AUC = 0.78). Differences in diagnostic accuracy dependent on lesion location were however not statistically significant.

Following analyses were performed using binned CNN predictions.

3.2 CNN Performance Compared to Human Raters

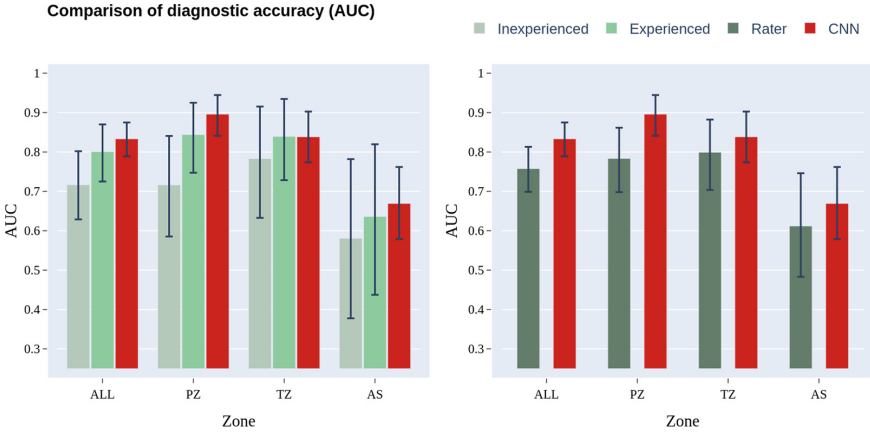


Fig. 3. Diagnostic accuracy of inexperienced, experienced (left), and all (right) assessments in comparison to the CNN predictions, expressed in established AUC values and 95% confidence intervals.

The Fig. 3 presents the results of a comparison of diagnostic accuracy measured in AUC between inexperienced and experienced raters in comparison to model predictions restricted to ordinal categories. Overall, CNN achieved superior ($p < .1$) diagnostic accuracy (AUC = 0.83, CI [0.79, 0.87]) in comparison to all (AUC = 0.76, CI [0.70, 0.81]) and inexperienced (AUC = 0.72, CI [0.63, 0.80]) rater assessments.

Differences were statistically significant for PZ lesion evaluation when considering assessments of all (CNN AUC = 0.90 vs AUC = 0.78, $p < .05$) and inexperienced raters (AUC = 0.71, $p < .05$). There were no statistically significant differences found in diagnostic accuracy between assessments of experienced raters and CNN predictions.

3.3 Diagnostic Accuracy of Combined Assessment

To investigate the potential change in diagnostic accuracy by integration of computer-aided assessment we analyzed the potential results of combining human and automatic predictions. Integrated predictions were obtained by computing average expert and binned CNN predictions on the lesion level and mapping those back to the Likert scale. This allowed investigation of the potential gain in diagnostic accuracy in computer-aided PCa diagnosis.

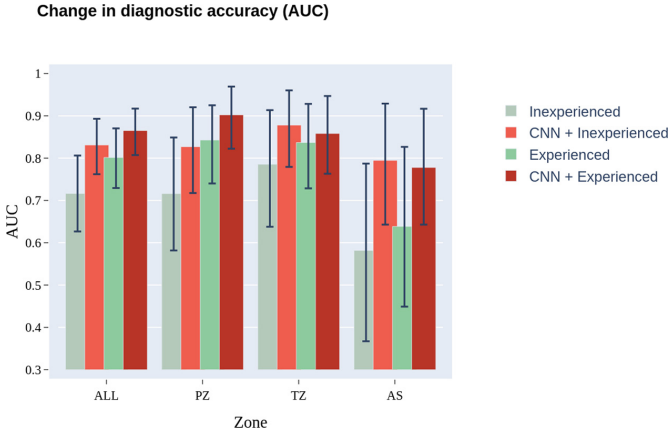


Fig. 4. Diagnostic accuracy for assessments of inexperienced and experienced radiologists compared to combined predictions expressed in established AUC values and 95% confidence intervals.

A positive diagnostic accuracy change has been observed after combining the model predictions with expert assessments in all tested settings (Fig. 4). Integration of CNN with rater predictions resulted in an overall increase of diagnostic accuracy by 0.09 AUC (CNN+rater AUC = 0.85, $p < .05$, CI [0.80, 0.89])

4 Discussion and Conclusion

In this study, we investigated the performance of the deep CNN model for PCa diagnosis on mpMRI data in comparison to human raters in an MRMC study setting on a subset of the reference dataset.

The results suggest that the proposed model outperformed inexperienced radiologists and achieved diagnostic accuracy similar to that of experienced raters. The achieved results are promising, yet decisive conclusions cannot be drawn confidently given the study design and small sample size used for validation.

Our study had several limitations. First of all, the dataset size used for validation in the conducted study was limited by the availability of readers. The study involved a substantial number of readers, however, the analysis has been performed in subgroups defined based on radiologist experience, which limited the number of assessments considered in hypotheses testing. The modest sample size resulted in wide 95% CIs constructed using bootstrap simulations and therefore affected the power of performed statistical tests. Furthermore, the study design was far from the clinical setting and based on the evaluation of selected single lesions. Finally, we could not evaluate the stability of the model performance on external data.

Although promising, results need confirmation in further, more extensive studies.

References

1. Armato, S.G., et al.: PROSTATEx challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *J. Med. Imaging* **5**(4), 044501 (2018)
2. Carioli, G., et al.: European cancer mortality predictions for the year 2020 with a focus on prostate cancer. *Ann. Oncol.* **31**(5), 650–658 (2020)
3. Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H.: SPIE-AAPM PROSTATEx challenge data (2017). <https://doi.org/10.7937/K9TCIA.2017.MURS5CL>, <https://wiki.cancerimagingarchive.net/x/iIFpAQ>
4. Liu, Q., Shepherd, B.E., Li, C., Harrell, F.E., Jr.: Modeling continuous response variables using ordinal regression. *Stat. Med.* **36**(27), 4316–4335 (2017)
5. MacKinnon, J.G.: Bootstrap hypothesis testing. *Handb. Comput. Econometrics* **183**, 213 (2009)
6. Mottet, N., et al.: EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer-2020 update. part 1: screening, diagnosis, and local treatment with curative intent. *Eur. Urol.* **79**(2), 243–262 (2021)
7. Sobeci, P., Józwiak, R., Sklinda, K., Przelaskowski, A.: Effect of domain knowledge encoding in CNN model architecture-a prostate cancer study using mpMRI images. *PeerJ* **9**, e11006 (2021)
8. Turkbey, B., et al.: Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *Eur. Urol.* **76**(3), 340–351 (2019)
9. Twilt, J.J., van Leeuwen, K.G., Huisman, H.J., Fütterer, J.J., de Rooij, M.: Artificial intelligence based algorithms for prostate cancer classification and detection on magnetic resonance imaging: a narrative review. *Diagnostics* **11**(6), 959 (2021)
10. Westphalen, A.C., et al.: Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: experience of the society of abdominal radiology prostate cancer disease-focused panel. *Radiology* **296**(1), 76 (2020)
11. Witherspoon, L., Breau, R.H., Lavallée, L.T.: Evidence-based approach to active surveillance of prostate cancer. *World J. Urol.* **38**(3), 555–562 (2020)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

