Enhancing Trust in Clinically Significant Prostate Cancer Prediction with Multiple Magnetic Resonance Imaging Modalities

Benjamin Ng
University of Waterloo, Canada

BENJAMIN.NG@UWATERLOO.CA

Chi-en Amy Tai
University of Waterloo, Canada

AMY.TAI@UWATERLOO.CA

E. Zhixuan Zeng
University of Waterloo, Canada

EMILYZHIXUAN.ZENG@UWATERLOO.CA

Alexander Wong
University of Waterloo, Canada

ALEXANDER.WONG@UWATERLOO.CA

Abstract

In the United States, prostate cancer is the second leading cause of deaths in males with a predicted 35,250 deaths in 2024. However, most diagnoses are non-lethal and deemed clinically insignificant which means that the patient will likely not be impacted by the cancer over their lifetime. As a result, numerous research studies have explored the accuracy of predicting clinical significance of prostate cancer based on magnetic resonance imaging (MRI) modalities and deep neural networks. Despite their high performance, these models are not trusted by most clinical scientists as they are trained solely on a single modality whereas clinical scientists often use multiple magnetic resonance imaging modalities during their diagnosis. In this paper, we investigate combining multiple MRI modalities to train a deep learning model to enhance trust in the models for clinically significant prostate cancer prediction. The promising performance and proposed training pipeline showcase the benefits of incorporating multiple MRI modalities for enhanced trust and accuracy.

Keywords: prostate cancer, MRI, trust, explainability, deep learning

Data and Code Availability We use the SPIE-AAPM-NCI PROSTATEx Challenges, PROSTATEx_masks, and The Cancer Imaging Archive (TCIA) datasets (Litjens et al., 2014, 2017; Cuocolo et al., 2021; Clark et al., 2013). The data is available online and is available to other researchers. The code is available at https://github.com/catai9/multiple-modality-prostate-cancer-prediction.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

In the United States, prostate cancer is the second leading cause of deaths in males with a predicted 35,250 deaths in 2024 (National Cancer Institute). However, most diagnoses are non-lethal and deemed clinically insignificant which means that the patient will likely not be impacted by the cancer over their lifetime (Shaw et al., 2014). Thus, clinicians must distinguish between clinically significant and insignificant cancers to provide essential treatment when necessary while also avoiding overdiagnosis and overtreatment.

In the past decade, advances in machine learning, image classification, and semantic segmentation have made it possible to integrate deep learning into the clinical workflow for cancer diagnosis. Numerous research studies have explored the accuracy of predicting the clinical significance of prostate cancer based on magnetic resonance imaging (MRI) modalities and deep neural networks (Li et al., 2022). Examples include Yoo et al. (2019) which utilized diffusion-weighted magnetic resonance imaging (DWI) for cancer classification, and Tai and Wong (2024), which focuses on the T2-weighted (T2w) modality. Wang et al. (2018) utilized both T2w and Apparent Diffusion Coefficient (ADC) images to localize the prostate and detect clinically significant prostate cancer, but did not use DWI.

Despite their high performance, these models are not trusted by most clinical scientists. One reason is the lack of transparency and explanations provided by the deep learning models (Bluemke et al., 2020; of Radiology , ESR; Jia et al., 2020). Additionally, clinical scientists often use multiple magnetic resonance imaging modalities for their diagnosis (Steiger

and Thoeny, 2016), whereas most prediction models only consider a single or limited set of modalities.

In this paper, we investigate combining multiple MRI modalities to train a deep learning model to enhance trust in the models for clinically significant prostate cancer prediction. We further analyze the results using Explainable AI (XAI) techniques, which are analyzed and verified by a clinical scientist.

2. Methodology

In this study, we first use XAI to gain insight into a high-performing model that was recently published (Tai and Wong, 2024), which we will refer to as the Tai and Wong Model. Building on these insights, we then guide model improvements and conduct comparative analysis of single versus multi-modal MRI approaches for machine learning.

2.1. Explainable AI

To validate a model's performance, Grad-Cam++ implemented with M3D MedCam (Gotkowski et al., 2020) was utilized to generate 3D classification and segmentation attention maps. The attention maps were extracted from the last applicable layer (Chattopadhay et al., 2018).

To generate the summed attention maps, leave-one-out cross-validation was first used to organize each patient prostate MRI volume into one of the four output categories: true positive, true negative, false positive or false negative. Then, for every output category, a summed attention map was generated by passing each volume in that category through the model and summing the resulting attention maps. Finally, the outputs were visualized using the jet colormap, as seen in Figure 2.

2.2. Model Training

As transfer learning was shown to perform best in this task (Tai et al., 2023), the model employed for training was the MONAI ResNet-34 (Cardoso et al., 2022) with initial weights from breast cancer grade prediction (Tai et al., 2023). In this study, we also use the cohort of 200 patients from the SPIE-AAPM-NCI PROSTATEx Challenges, PROSTATEx_masks, and The Cancer Imaging Archive (TCIA) datasets for comparison consistency (Litjens et al., 2014, 2017; Cuocolo et al., 2021; Clark et al., 2013). Model training was conducted with a learning rate of 0.001 over 40 epochs, with no data augmentation or transformations applied.

Leave-one-out cross-validation (Cheng et al., 2017) was employed to split the dataset, ensuring that each sample served once as the validation set while the remaining data were used for training, thus maximizing the use of the limited dataset. The cross-entropy loss function (Goodfellow, 2016) was utilized to compute the error between the predicted and actual labels. The model output is binary, with a prediction of 1 indicating a clinically significant case of prostate cancer and 0 indicating the clinically insignificant prostate cancer. Model predictions were compared with expert medical evaluations to assess the model's accuracy.

2.3. Multiple MRI Modality Training

The three MRI modalities studied were DWI with a b-value of 800 (referred to as DWIb3), T2w, and ADC. For each modality, the 3D volumes were preprocessed by standardizing to 12 slices and cropping each slice to focus on the prostate region. A composite 3D volume was then constructed by stacking the corresponding slices from the three modalities. The process of preprocessing and training this 3-modality model is shown in Figure 1.

3. Results

3.1. XAI on the Tai and Wong Model

This model was trained using only the T2-weighted (T2w) modality and achieved a leave-one-out cross-validation accuracy of 97.5% (Tai and Wong, 2024). However, following consultation with a medical professional and further analysis using M3D MedCam, this model was determined to be suboptimal, likely due to overfitting on the volumetric data. This was identified after creating a summed attention map for Grad-Cam++ for each of the patients in the output categories (true positive, true negative, false positive and false negative). The expected result is that the highlighting would be primarily around the prostate mask region, i.e., indicating that the model is looking at the prostate to determine clinical significance. However, as demonstrated in Figure 2, the highlighting was sporadic and not centered around the prostate region.
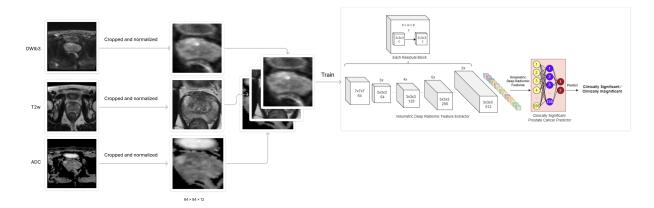
Figure 1: Preprocessing and training flow chart for 3-modality prostate cancer clinical significance model with the volumetric deep radiomic features and model portion adapted from (Tai and Wong, 2024).
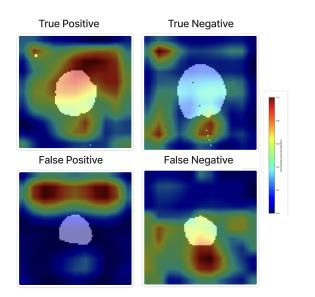

Figure 2: Summed Grad-CAM++ attention map and prostate mask for the Tai and Wong model (Tai and Wong, 2024) (used to identify unintended behaviour of T2w-only model and potential overfitting).

(a) Example Slice   (b) Grad-Cam++ overlay

Figure 3: Example slice for the single modality (T2w) model trained on images with masked prostate region with the associated Grad-Cam++ result.

3.2. Guided Model Improvement

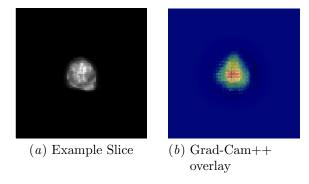The XAI insights into the Tai and Wong model guided model improvement through explicit focusing around the prostate. Specifically we preprocessed the T2w modality patient volumes and masked areas outside of the prostate region. As a result, only the prostate region was visible and the rest of the image is black as demonstrated in the left side of Figure 3. We then trained a model using these processed patient volumes. Although improving the outcome of the XAI (Figure 3 right), this adjustment resulted in a reduced accuracy of 65%. This lower performance is attributed to the poor resolution of the images which led to insufficient information provided for model training.
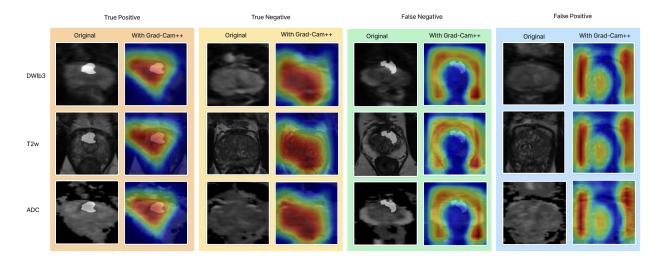
Figure 4: Example patient original (source image) slice with lesion mask (white highlighted region) vs Grad-Cam++ attention map for the 3-modality trained model.

Table 1: Accuracy (Acc.), Sensitivity (Sen.), Specificity (Spec.), and F1 score comparison for (1) the Tai and Wong model (trained only on T2w) (Tai et al., 2023), (2) the single modality model (masked to the prostate region), and (3) the 3-modality model with cropping to the prostate region.

Model	Acc.	Sen.	Spec.	F1 score
(1)	97.5%	94.3%	99.2%	96.3%
(2)	65.0%	0.0%	100.0%	0.0%
(3)	85.0%	84.3%	85.4%	79.7%

3.3. 3-Modality Model Results

The 3-modality model involved combining the three modalities and cropping the slices around the prostate. As shown in Table 1, this model performed superior to the single modality masked. The 3-modality model demonstrated a leave-one-out cross-validation accuracy of 85% in accurately identifying cases of clinically significant prostate cancer. Though the performance was not as impressive as the Tai and Wong model, the 3-modality model had better and more interpretable XAI results. Figure 4 illustrates the model's prediction results, classified into four distinct categories.

For the true positive outcome, the highlighted regions coincided with the lesion mask (which represents the area where the clinically significant prostate should be identified), while the false negative outcome did not. This result aligns with expected behaviour as the model is expected to focus on the lesion mask within the prostate region to make accurate predictions. Conversely, in true negative cases, the model correctly identified the absence of lesions by examining the central region of the image where the lesion mask would be present. This observation also supports the conclusion that the model is focusing on the right regions.

4. Conclusion

This paper leverages XAI to provide two insights about prostate model training enhancements: cropping to focus around the prostate region and using multiple modalities. Initially, XAI showed issues in the Tai and Wong model because the attention map did not focus on the prostate. However, by addressing these issues through the aforementioned methods, we achieved more interpretable XAI results. This outcome demonstrates that combining multiple MRI modalities in deep learning model training has the potential to foster greater trust in clinically significant prostate cancer prediction. Furthermore, the strong accuracy of the multi-modality training pipeline highlights the effectiveness of this approach. Future work includes expanding the study with other prostate cancer datasets and studying the XAI results for other high-performing models in the cancer domain to analyze their explainability.

References

David A Bluemke, Linda Moy, Miriam A Bredella, Birgit B Ertl-Wagner, Kathryn J Fowler, Vicky J Goh, Elkan F Halpern, Christopher P Hess, Mark L Schiebler, and Clifford R Weiss. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the radiology editorial board, 2020.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

Hao Cheng, Dorian J Garrick, and Rohan L Fernando. Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *Journal of animal science and biotechnology*, 8:1–5, 2017.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013.

Renato Cuocolo, Arnaldo Stanzione, Anna Castaldo, Davide Raffaele De Lucia, and Massimo Imbriaco. Quality control and whole-gland, zonal and lesion annotations for the prostatex challenge public dataset. *European Journal of Radiology*, 138: 109647, 2021. doi: 10.1016/j.ejrad.2021.109647.

Ian Goodfellow. Deep learning, 2016.

Karol Gotkowski, Camila Gonzalez, Andreas Bucher, and Anirban Mukhopadhyay. M3d-cam: A pytorch library to generate 3d data attention maps for medical deep learning. *arXiv preprint arXiv:2007.00453*, 2020.

Xun Jia, Lei Ren, and Jing Cai. Clinical implementation of ai technologies will require interpretable ai models. *Medical physics*, (1):1–4, 2020.

Huanye Li, Chau Hung Lee, David Chia, Zhiping Lin, Weimin Huang, and Cher Heng Tan. Machine learning in prostate mri for prostate cancer: current status and future opportunities. *Diagnostics*, 12(2):289, 2022.

Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Computer-aided detection of prostate cancer in mri. *IEEE Transactions on Medical Imaging*, 33(5):1083–1092, 2014. doi: 10.1109/TMI.2014.2303821.

Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Prostatex challenge data [data set], 2017.

National Cancer Institute. Seer cancer stat facts: Prostate cancer. URL https://seer.cancer.gov/statfacts/html/prost.html. Accessed: 2024-09-03.

European Society of Radiology (ESR) communications@ myesr. org Neri Emanuele de Souza Nandita Brady Adrian Bayarri Angel Alberich Becker Christoph D. Coppola Francesca Visser Jacob. What the radiologist should know about artificial intelligence–an esr white paper. *Insights into imaging*, 10(1):44, 2019.

GL Shaw, BC Thomas, SN Dawson, G Srivastava, SL Vowler, VJ Gnanapragasam, NC Shah, AY Warren, and DE Neal. Identification of pathologically insignificant prostate cancer is not accurate in unscreened men. *British journal of cancer*, 110(10):2405–2411, 2014.

Philipp Steiger and Harriet C Thoeny. Prostate mri based on pi-rads version 2: how we review and report. *Cancer Imaging*, 16(1):9, 2016.

Chi-en Amy Tai and Alexander Wong. Enhancing clinically significant prostate cancer prediction in t2-weighted images through transfer learning from breast cancer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Women in Computer Vision (WiCV), 2024. URL https://arxiv.org/abs/2405.07869.

Chi-en Amy Tai, Hayden Gunraj, Nedim Hodzic, Nic Flanagan, Ali Sabri, and Alexander Wong. Enhancing clinical support for breast cancer with deep

learning models using synthetic correlated diffusion imaging. In *International Workshop on Applications of Medical AI*, pages 83–93. Springer, 2023.

Zhiwei Wang, Chaoyue Liu, Danpeng Cheng, Liang Wang, Xin Yang, and Kwang-Ting Cheng. Automated detection of clinically significant prostate cancer in mp-mri images based on an end-to-end deep neural network. *IEEE transactions on medical imaging*, 37(5):1127–1139, 2018.

Sunghwan Yoo, Isha Gujrathi, Masoom A Haider, and Farzad Khalvati. Prostate cancer detection using deep convolutional neural networks. *Scientific reports*, 9(1):19518, 2019.