Computer Vision Aided Diagnosis for Prostate Cancer

Yen Hann Yoo Jessie Tanaboriboon

6.8300 Advances in Computer Vision | Final Project

yyoo@mit.edu phakjira@mit.edu

Abstract

The diagnosis of prostate cancer heavily relies on complex bi-parametric Magnetic Resonance Imaging (MRI), which suffers from low inter-reader agreement and suboptimal interpretation. This study examined the performance of pre-trained Deep Learning models, specifically VGG19, in detecting prostate cancer from biparametric MRI scans. It compared two MRI modalities, Apparent Diffusion Coefficient (ADC) and Transverse Relaxation Time Weighted Images (T2W), on a dataset of 1500 annotated case studies. Despite employing preprocessing, reweighting, and data augmentation techniques, the performance between the two modalities was similar, albeit slightly favoring T2W. The study suggests further exploration of different model architectures, 3D models, and vision transformers for improved prostate cancer detection.

1. Introduction

Prostate cancer, a leading cause of cancer-related deaths in men globally, presents challenges in its diagnosis using prostate biparametric Magnetic Resonance Imaging (MRI). The reliance on multiple imaging modalities, such as ADC and T2W, coupled with the heterogeneous nature of the disease, results in low inter-reader agreement (less than 50%) and sub-optimal interpretation (Westphalen et al., 2020).

Past studies have explored 3D imaging and multi-modal approaches to enhance interpretation, often by reformulating the problem into a 2D context through lesion center slicing (Liu et al., 2017). However, these methods require lesion location identification, which is complex and poses limitations. Our project aims to address this limitation by investigating the possibility of cancer detection without prior knowledge of the lesion location. Additionally, we seek to explore the feasibility of utilizing a single imaging modality for detecting clinically significant prostate cancer, thereby streamlining the decision-making process.

To achieve these objectives, we conducted experiments using popular pre-trained Deep Learning models for image classification, comparing the performance of ADC and T2W modalities. Our project aims to determine if either modality outperforms the other in prostate cancer detection. By tackling these challenges, we strive to improve the accuracy and reliability of prostate cancer diagnoses while reducing barriers for less experienced clinicians, ultimately benefiting patient outcomes.

2. Methodology

Our experiment aims to test the hypothesis that ADC and T2W modalities will demonstrate similar performance in classifying whether a patient's imaging shows clinically significant indications of cancer using transfer learning. However, we anticipate that using a single modality may not perform as well as using multiple modalities. Nonetheless, it could still be adequate for initial screening, with the understanding that further analysis by a medical professional would be necessary.

Dataset:

The dataset for this study was obtained from the Prostate Imaging: Cancer AI challenge (PI-CAI) hosted by Grand Challenge (Saha et al., 2022). It consists of 1500 case studies involving 1476 unique patients who underwent parametric MRI scans of the prostate between 2012 and 2021. Each case study includes five different modalities of MRI images (ADC, COR, HBV, SAG, and T2W), each with varying dimensions and depths. The dataset also provides clinical reports indicating whether the patient has clinically significant prostate cancer or not, with a prevalence of 28.3% in the dataset.

Annotations by a radiologist expert are available for the ADC and T2W modalities, allowing for the identification of the location of clinically significant prostate cancer through masking. The dataset includes two sets of annotations, one for the ADC modality and another for the T2W modality. The focus of this study is on comparing the ADC and T2W modalities, while the remaining three modalities are not considered due to the lack of annotations.

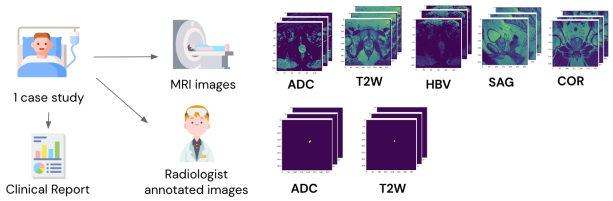

Figure 1. Data available for one patient

Preprocessing:
Each modality in a case study is represented by an array with the shape (Height, Width, Depth), where "Height" and "Width" denote the dimensions of the image, and "Depth" represents the number of images in that case study. The "Depth" dimension captures the MRI scans at different depths within the human body, providing a comprehensive view of the prostate region.

To ensure consistency, the images are resized to a standardized size of (128, 128, Depth). This resizing process maintains the depth dimension, preserving important information captured in the original images. For the development of a 2D model, each case's image slice along the depth is treated as an independent input. A dataset comprising 1,290 patients with both ADC and T2W annotations was compiled, resulting in 27,342 inputs for the ADC model and 29,328 inputs for the T2W model. The input data shape is (27,342, 128, 128) for the ADC model and (29,328, 128, 128) for the T2W model. To comply with model architectures requiring three channels, the first channel was replicated three times, resulting in input shapes of (27,342, 128, 128, 3) for the ADC model and (29,328, 128, 128, 3) for the T2W model.

Additionally, the ADC and T2W annotated images were converted into binary labels, indicating the presence or absence of a tumor for each depth. Consequently, the label shapes are (27,342,) for the ADC model and (29,328,) for the T2W model. These preprocessing steps enable the prediction of tumor presence based on ADC and T2W imaging data.

The dataset is divided into three distinct sets: training, validation, and testing, to facilitate model development and evaluation. The allocation percentages for each set are as follows: [1] Training Set: 70% of the Patients. A portion of the training set is used for model fine-tuning during the training process. [2] Validation Set: 15% of the Patients. This set is utilized for model comparison and selection. [3] Testing Set: 15% of the Patients. This independent set is employed to assess the model's performance on unseen data, providing an unbiased evaluation of its generalization capabilities.

Model:
Through the annotated image to binary label mapping outlined above, the distributions of 1's and 0's were found to be skewed. For ADC, 26,326 and 1,016 images were labeled as 0's and 1's respectively. On the other hand, for T2W, 28,051 and 1,277 images were labeled as 0's and 1's respectively. The overall input and output setup is depicted in the following figure:

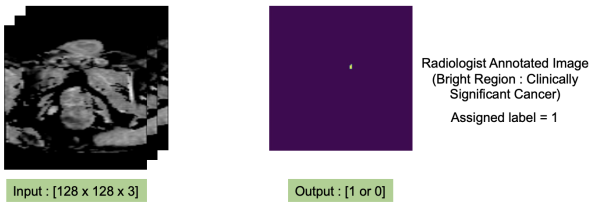

Figure 2. Input and output setup for the model

Given the skewed output distribution, accuracy was deemed an ineffective evaluation metric, favoring recall and precision instead. These metrics offer a more balanced performance assessment in the context of imbalanced classes. The focus was on patient-level assessments, classifying a patient's study as '1' if any image indicated cancer presence. This approach was aimed at reducing false negatives, as the cost of missing a potential cancer diagnosis outweighs the risk of false positives (i.e., a high recall is preferred). Furthermore, two output sets: raw (depth) and patient-level were outputted, depicted in the figure below.

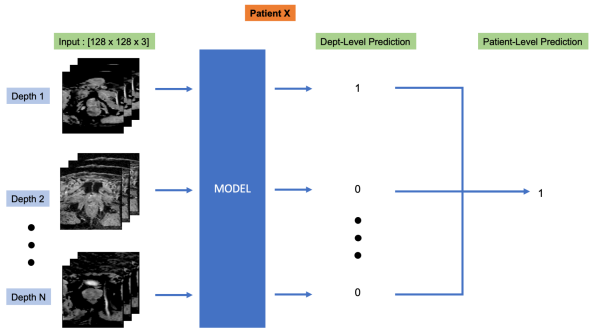

Figure 3. Example: Patient X originally has N channels in their study. Each channel is repeated (concatenated) 3 times depth-wise and used to generate raw level predictions and mapped back to the patient level.

This project considered two popular pre-trained models, VGG19 and ResNet50, for medical image analysis due to their deep architectures and performance in computer vision tasks. However, given computational constraints and the larger number of parameters in ResNet50, VGG19 was chosen for further experimentation in various setups.

3. Experimental Results & Discussion

To conduct the experiments, a VGG19 (Simonyan & Zisserman, 2015) model with pre-trained ImageNet weights was utilized as the base model for extracting embeddings. Additional dense layers were added in succession to the base model to enhance its performance. The experimental setups performed in this study are as follows:

[1] Headless VGG19 (No Reweighting): A simple model without reweighting. Due to class imbalance and pre-trained weights from ImageNet, performance may be limited for MRI data without additional training.
[2] Headless VGG19 with Reweighting: Addressing data imbalance by incorporating reweighting of the loss function. Equal contribution from all classes helps mitigate the effects of imbalanced data.
[3] Retraining Last Block of VGG19 with Reweighting: Retraining the last block of VGG19 to adapt it for our task. This approach leverages useful embeddings from earlier layers while accounting for data imbalance through class weighting.
[4] Retraining Entire VGG19 with Reweighting: Exploring retraining the entire VGG19 model to adapt it fully to our MRI data. However, overfitting may occur due to limited data availability.
[5] Headless VGG19 with Reweighting and Data Augmentation: Addressing overfitting through data augmentation techniques, which includes random flipping, rotation, zooming, contrast adjustments, and translation. Augmented data enhances diversity and quantity, improving the model's generalization. Applied to headless VGG19 model with reweighting.
[6] Retraining Last Block of VGG19 with Reweighting and Data Augmentation: Combining data augmentation with retraining the last block of VGG19. Leveraging augmented data and fine-tuned last block layers to enhance performance on our MRI dataset.

The test results are effectively visualized through the following graphs:

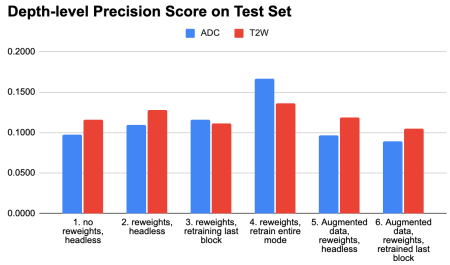

Figure 3. Depth-level Precision Score by Model

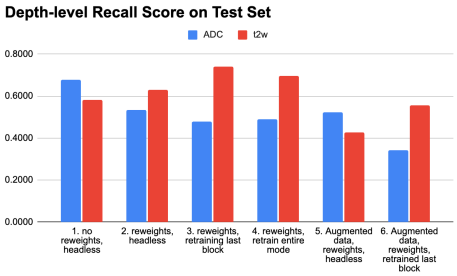

Figure 4. Depth-level Recall Score by Model

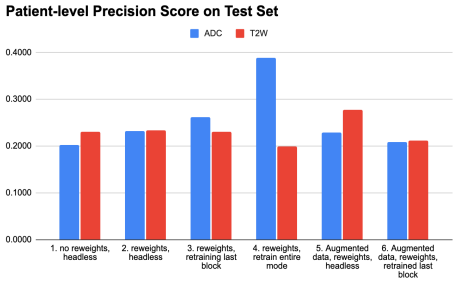

Figure 5. Patient-level Precision Score by Model

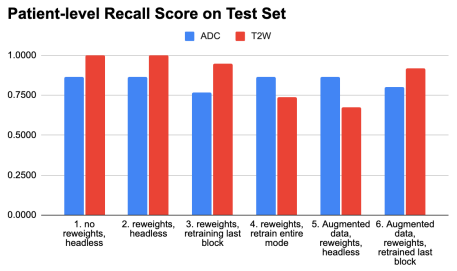

Figure 6. Patient-level Recall Score by Model

Results Analysis:
Due to computational limitations as both members exhausted Colab Pro units, a baseline model with the same architecture outlined above incorporating both T2W and ADC modalities could not be developed. This limitation also prohibited exploration of more complex models such as vision transformers and nnUNet.

[1] Observation: Low precision, high recall. In general, the precision is low (between 0.1 to 0.4) whilst the recall is high. This is due to the chosen probability threshold between 0.05 and 0.10 based on the validation set. Our observations revealed that increasing the threshold beyond 0.10 resulted in a significant decrease in recall, while only marginally improving precision. Therefore, we opted to use this relatively low threshold. This also implies that the model might be struggling to distinguish between positive (cancerous) and negative

(non-cancerous) cases - the decision boundary that the model has learned might not be well-defined.

Even so, in healthcare it is preferable to be conservative and thus a high recall is preferred over precision. With this reality, a probability threshold of 0.05 or 0.10 was selected for experimental setups. Nevertheless, the fact that a very low probability threshold results in high recall but low precision indicates that many of the true positive predictions are accompanied by a significant number of false positives - the model is being over conservative by frequently predicting the positive class. In doing so, it is incorrectly classifying many negative examples as positive. The reasons for this could be attributed to the severity of the class imbalance, such that reweighting alone was insufficient to mitigate its issues, or the model's inability to learn features from the input images to distinguish between the two classes.

[2] Observation: Data augmentation demonstrated no significant performance improvements, occasionally even detrimenting model performance. This is because medical imaging often contains very specific features of interest that are crucial for diagnosis. In this dataset, these would be the very localized and small cancerous regions within the prostate, as shown by the annotated images. Unlike in natural images, these features are often not invariant to transformations like flips, rotations, or zooms. To demonstrate this idea, a cat for example, can be upside down or sideways and still be recognizably a cat. This isn't always the case in medical imaging. A particular pattern or structure may have a very specific orientation that is meaningful. The structure of certain tissues, the orientation of cells, or the shape and orientation of a tumor can all provide crucial information. For instance, by augmenting through zooming out, the small cancerous region may be even harder to detect and thus hinders the model's learning abilities. Therefore, care must be exercised on the augmentation type applied to images.

[3] Observation: Depth-level T2W performed better than ADC. Based on the depth-level precision and recall metrics, in general, the VGG19 model fine-tuned on the T2W dataset performed slightly better. A plausible reason for this is the fact that there were more training samples with T2W than with ADC - 29,328 vs. 27,342 respectively. Furthermore, the nature of the data itself in T2W and ADC images can contribute to this difference. T2W images often provide high contrast and detailed anatomical structures, as seen in Figure 7, which may make it easier for the model to identify relevant features for cancer detection, improving both precision and recall. On the other hand, ADC images, derived from diffusion-weighted imaging, provide information on the diffusion of water molecules within tissues (Koh & Collins, 2007).

However, the information provided by ADC maps is often less intuitive and harder to interpret visually compared to T2W images, but nonetheless complementary to T2W. The signal in ADC maps is generally less distinct, and the maps are more susceptible to noise and artifacts. With deep learning models, these characteristics may make it more challenging for it to extract meaningful features from ADC images as compared to T2W images. This doesn't mean ADC images are less valuable, but rather that they might require more specialized approaches or models specifically designed or trained to interpret this type of data.

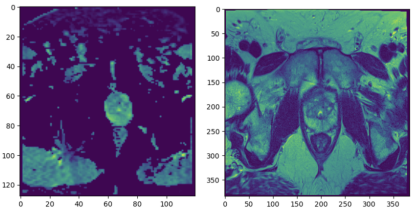

Figure 7. Sample ADC Image (Left), T2W Image (Right)

[4] Observation: Recall for the T2W model, only fine-tuned on the last block is higher than the entirely retrained model. The reverse is true for ADC. In healthcare, the emphasis is on recall over precision for a conservative approach. Nonetheless, the aforementioned recall metric discrepancies between T2W and ADC can be attributed to the different data characteristics. T2W images, rich in spatial detail and contrast, demand fewer parameters for feature learning, making last-block fine-tuning effective and full model fine-tuning prone to overfitting. In contrast, ADC images, with less spatial detail and contrast, necessitate more extensive fine-tuning to capture relevant features, explaining why full model fine-tuning outperformed in this case.

GradCAM Outputs:

For interpretability purposes and to better understand what features the model was learning, a GradCAM (Gradient Class Activation Map) was adopted to visualize the learned features from the final convolutional layer of the model. An example is shown in Figure 8 for one of the depths from patient study 5, which was flagged with clinically significant cancer, for the reweighted and fine-tuned last block model.

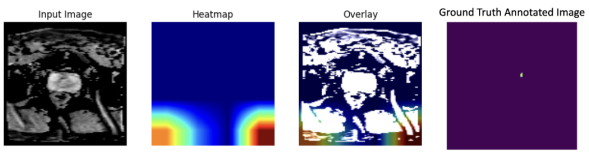

Figure 8. Example of GradCAM Result 1 (ADC)

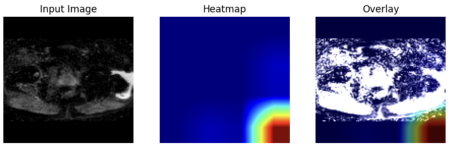

Figure 9. Example of GradCAM Result 2 (ADC)

The model underperformance can be attributed to its difficulty in accurately identifying the small cancerous regions, as reflected in Figure 8. As the cancerous regions are relatively small and not well defined, the features that distinguish cancerous cells from healthy ones might be highly complex or subtle, thus making them difficult for the model to learn and accurately detect. Moreover, in Figure 9, the model is focusing on the borders of the image and thus learning spurious features. For improvements, one would have to conduct additional preprocessing steps such as zooming in on images with localized cancerous regions to enlarge these features and facilitate better learning. Additionally, cropping images with dark borders surrounding the prostate gland as seen in Figure 9 is needed to avoid spurious feature learning.

Baseline Comparison:

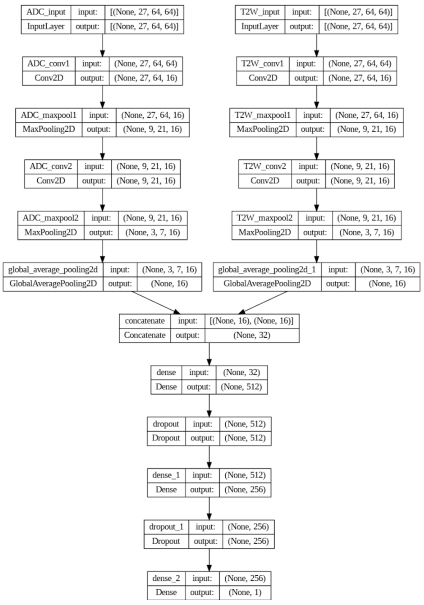

Figure 10. Baseline model, inputs: ADC and T2W

A baseline model, shown in Figure 10, consists of two branches - one for ADC, the other for T2W. Output embeddings are concatenated and fed through dense layers for classification. The baseline model's precision and recall on the test set were 0.297 and 0.900 respectively for a probability threshold of 0.05. Although similar in performance to earlier models, the baseline's architecture is far simpler and exhibits more potential for further development. The baseline model is a 3D input convolutional architecture which enables it to learn more meaningful features. Furthermore, in earlier models, feature extraction needed to be performed on a single modality, and the model needed to learn all relevant features from that single source. In the two-branch model, the burden of feature extraction is shared across two different modalities, which might make the task easier, hence resulting in comparable testing performances.

Additionally, we also conducted experiments using the baseline architecture but with a single branch for each modality (i.e., a model per modality). The ADC model achieved a test set precision of 0.284 and recall of 0.857, whilst the T2W model achieved a test set precision of 0.314 and recall of 0.786. It is evident that using only one modality results in a reduction in performance compared to using both modalities. However, the performance of the ADC model is comparable to the combined modality model. Therefore, in order to thoroughly evaluate the sufficiency of using only one modality, further experimentation is necessary, as outlined in the conclusion.

4. Conclusion & Further Works

In conclusion, our project aimed to improve prostate cancer diagnoses by comparing the performance of ADC and T2W imaging modalities using deep learning models. We found that both modalities showed similar performances, with T2W slightly outperforming ADC. However, the precision was generally low while the recall was high, indicating challenges in distinguishing between positive and negative cases.

Data augmentation also did not improve the model's performance, suggesting limited invariance of medical imaging features to transformations. Further exploration is needed, including comparing different model architectures such as UNET, a commonly employed model architecture for medical image segmentation tasks, transitioning to 3D models as the spatial information can yield a more comprehensive analysis of the cancerous regions, and exploring Vision Transformers with attention mechanisms to attend to specific patches where the cancerous regions exists, which may prove beneficial.

Other potential extensions include cancer grade classification (0 to 5 on the ISUP scale), which can improve treatment planning. These future investigations have the potential to enhance our understanding of model capabilities and limitations, ultimately improving the accuracy and reliability of prostate cancer diagnoses.

References

Koh, D.-M., & Collins, D. J. (2007). Diffusion-weighted MRI in the body: Applications and challenges in oncology. *American Journal of Roentgenology*, *188*(6), 1622–1635. https://doi.org/10.2214/ajr.06.1403

Liu, S., Zheng, H., Feng, Y., & Li, W. (2017). Prostate cancer diagnosis using deep learning with 3D multiparametric MRI. *SPIE Proceedings*. https://doi.org/10.1117/12.2277121

Saha, A., Twilt, J. J., Bosma, J. S., Ginneken, B. van, Yakar, D., Elschot, M., Veltman, J., Fütterer, J., Rooij, M. de, & Huisman, H. (2022). *Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge (Study Protocol)*. https://doi.org/10.5281/zenodo.6522364

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. https://doi.org/ https://doi.org/10.48550/arXiv.1409.1556 Focus to learn more

Westphalen, A. C., McCulloch, C. E., Anaokar, J. M., Arora, S., Barashi, N. S., Barentsz, J. O., Bathala, T. K., Bittencourt, L. K., Booker, M. T., Braxton, V. G., Carroll, P. R., Casalino, D. D., Chang, S. D., Coakley, F. V., Dhatt, R., Eberhardt, S. C., Foster, B. R., Froemming, A. T., Fütterer, J. J., … Rosenkrantz, A. B. (2020). 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), 76–84. https://doi.org/10.1148/radiol.2020190646

Individual Contributions

- **Preprocessing:** Jessie Tanaboriboon
- **ADC Models:** Yen Hann Yoo
- **T2W Models:** Jessie Tanaoboriboon
- **GradCam Analysis**: Yen Hann Yoo
- **Baseline Model:** Yen Hann Yoo and Jessie Tanaboriboon
- **Report & Presentation:** Yen Hann Yoo and Jessie Tanaboriboon