THERAPY-AGNOSTIC PROGNOSTICATION OF PROSTATE CANCER VIA MR IMAGING AND CLINICAL DATA INTEGRATION

Haocheng Dai *Geoff Nelson* *Glen Morrell* *Jonathan Tward* *Sarang Joshi*

University of Utah

ABSTRACT

Prostate cancer is one of the most common malignancies and a leading cause of cancer-related mortality among men worldwide. Accurate prognostication is crucial for guiding clinical decision-making and optimizing patient outcomes. However, current methods often rely solely on clinical features and fail to integrate advanced imaging data effectively. In this study, we develop a progression prognostication model that combines clinical variables with magnetic resonance (MR) imaging analysis. Our model aims to provide therapy-agnostic predictions, enabling a more nuanced understanding of patient prognosis to facilitate treatment intensification decisions. We also analyze the impact of multimodal data integration and clinical variables within our framework to enhance prognostic accuracy and guide clinical decision-making.

***Index Terms*—** prostate cancer, multi-modality, random forest, prognostication, magnetic resonance (MR) imaging

1. INTRODUCTION

Prostate cancer is one of the most prevalent malignancies affecting men globally, posing a significant public health challenge. It is also a leading cause of cancer-related mortality among men, with incidence rates continuing to rise worldwide [1]. Beyond its physical impact, prostate cancer significantly affects the quality of life and psychological well-being of patients and their families, underscoring the need for effective prognostic tools. Uncertainty regarding the potential benefit of a particular therapy can hinder clinicians and patients from confidently selecting the optimal initial course of treatment, which may result in reduced patient adherence and increased stress for both patients and healthcare providers. Thus, accurate prognostication of prostate cancer is essential for optimizing patient management, guiding clinical decision-making, and ultimately improving patient outcomes.

Despite the critical importance of accurate prognostication, existing methods face several challenges, such as variability in clinical presentation and the complex nature of tumor progression. Many conventional prognostic approaches [2] are limited by their reliance on clinical features alone and fail to adequately integrate the wealth of information available from advanced imaging modalities. Wang et al. [3] identified potential multi-omics biomarkers for the early detection of prognostic recurrence in prostate adenocarcinoma (PRAD) patients and constructed multi-omics panels based on the intersected biomarkers from different models. Esteva et al. [4] introduced a multimodal deep learning approach for personalizing prostate cancer therapy by predicting long-term outcomes using clinical data and digital histopathology from micro-scale, demonstrating superior prognostication compared to traditional approaches, such as the National Comprehensive Cancer Network (NCCN) guidelines, which have repeatedly shown limited discriminatory performance.

In this study, we aim to address these limitations by developing a comprehensive prognostication model for prostate cancer that integrates clinical variables with macro perspective magnetic resonance (MR) imaging features, focusing on therapy-agnostic prediction of disease outcomes. By leveraging both clinical and imaging data, our model aims to provide a more nuanced understanding of prostate cancer prognosis, ultimately enhancing individualized patient care and enabling more informed clinical decisions. Furthermore, we investigate the contributions of different components within our framework, including multimodal data and various clinical variables, in order to provide a more comprehensive understanding of how to achieve accurate prognostication.

2. METHODS

Our framework integrates two machine learning models: a convolutional neural network (CNN) for MR image encoding and random forests for prognosis scoring. For each patient, we input their MR images (either T2 or diffusion-weighted) and clinical variables. The MR images are first resampled to achieve a voxel spacing of 0.3mm × 0.3mm and a slice thickness of 3mm using bilinear interpolation. After resampling, we apply symmetric central padding or cropping to standardize the volume to a size of $32 \times 128 \times 128$ voxels (corresponding to $96mm \times 96mm \times 96mm$ in physical space, a size can cover the whole prostate under most cases). This volume is then subjected to data augmentation, which includes random horizontal flipping along the transverse plane, followed by random rotation with angles uniformly sampled within $\pm 90^{\circ}$. The probability of applying these augmenta-

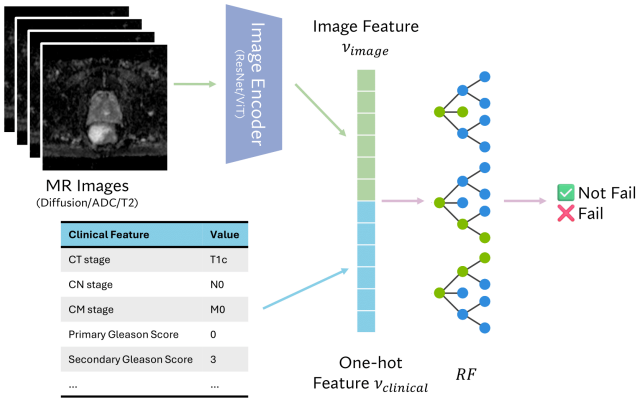

Fig. 1. Overview of the proposed framework.

tions is set to 80%. To reduce the influence of outlier values during normalization, we clip the top and bottom 1% of intensity values, after which we scale the pixel values to the range $[0, 1]$. Finally, intensity normalization is performed using the mean and standard deviation.

Instead of using a 3D CNN, we chose a standard 2D ResNet as our image encoding backbone because MRI is acquired in slices rather than as a continuous volume. The spacing between transverse slices is much larger than between coronal or sagittal planes, making linear interpolation between transverse slices unnecessary. We did not observe significant performance gains with a 3D architecture, which also comes with increased computational costs. To effectively leverage the 3D volumetric data, we treat the first dimension (of size 32) as the channel dimension and the remaining dimensions (of size 128) as the spatial dimensions for the ResNet convolutional layers. We trained the ResNet with two different objectives: standard binary classification and contrastive learning. In the contrastive learning approach, we generate two augmented versions of the same image and minimize the distance between their respective features using a contrastive loss. However, we found no substantial performance difference between the two types of image features between the conventional classification and contrastive learning strategies. Therefore, we opted for a standard binary classification task with binary cross-entropy loss to generate the image features. We did not rely solely on the binary classification results from CNNs because CNNs tend to overfit the training set in prostate cancer prognostication, even when using the smallest architectures, strong data augmentation, and early stopping techniques. By employing a multimodal framework, we can effectively incorporate detailed clinical variables alongside imaging features, providing a more nuanced and comprehensive understanding compared to using raw imaging data alone.

Our framework's clinical variables consist of categorical and numerical data. The categorical variables include Clinical AJCC-derived T/N/M stage, primary/secondary Gleason scores prior to the first definitive treatment, and International Society of Urological Pathology (ISUP) grade group at the first definitive treatment. These are encoded into one-hot vectors. The numerical variables include the percentage of positive biopsy cores, patient age at first definitive treatment, highest PSA level prior to first definitive treatment, and prostate width/depth/height. The numerical variables are standardized by subtracting the corresponding mean and scaling to unit variance. The one-hot encoded categorical vectors and the normalized numerical vectors are then concatenated to form the clinical variable feature.

We then prepare one random forest, RF, comprising 190 decision trees with a maximum depth of 5 and a maximum feature count of 1. The image features, v_{image}, are then concatenated with the clinical features, v_{clinical}, before sending through the random forest. The thresholded prediction will be treated as the final outcome. For more straight forward illustration, see Figure 1.

3. RESULTS

To validate the performance of our framework, we collected a dataset comprising 200 patients, with each patient having one T2-weighted MRI, one diffusion-weighted MRI (b1500), one apparent diffusion coefficient (ADC) map, and a set of clinical variables. We performed 60 random splits of the dataset, ensuring a balanced distribution of failed and non-failed classes within each split. For each split, 80% of the data was used for training, while the remaining 20% was reserved for testing and validation.

To evaluate the effectiveness of different imaging modalities, we use Figure 2 (a) to illustrate the distribution of AUC scores under different data splits, with ResNet-10 used as the image encoder for different MR modalities. The AUC scores range from 0.61 to 0.83 for diffusion-weighted MRI, 0.66 to 0.87 for the ADC map, and 0.61 to 0.85 for T2-weighted MRI. Among these, the ADC-based framework achieved the highest median AUC of 0.77, followed by 0.75 for diffusion-weighted MRI and 0.73 for T2-weighted MRI.

To further investigate the impact of different image encoders, we compared the ViT (tiny) and ResNet-10 models, as shown in Fig. 2 (b). Despite the minimum AUC scores being similar between the two image encoders, the mean and median AUC scores of ResNet are substantially higher than those of ViT, with ResNet outperforming ViT by 0.04 in median AUC. This trend remains consistent across different configurations of the two types of image encoders.

Next, we sought to understand the contribution of combining both image and clinical features in the random forest classification, as presented in Fig. 2 (c). Using only image features resulted in an average and median AUC score of approximately 0.60, which was only slightly better than using a trained linear projection matrix. In contrast, using clini-

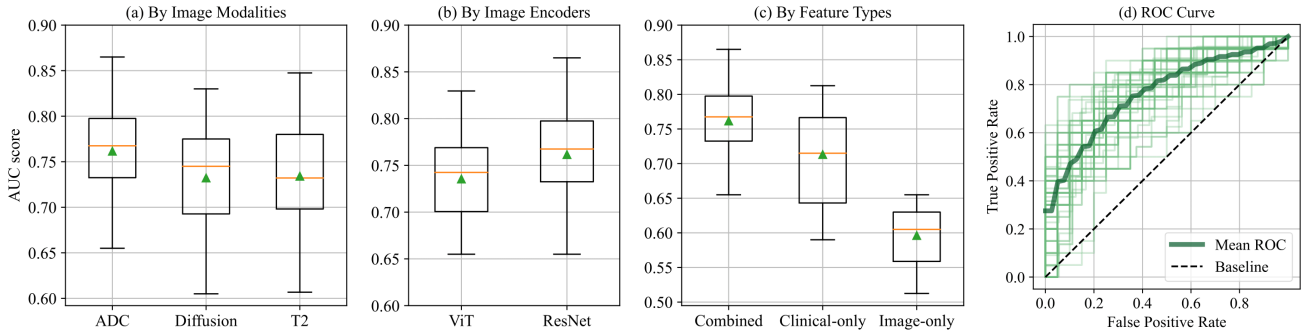

Fig. 2. Box plots of our framework's AUC scores by different image modalities (a), different image encoders (b), and different feature types used in random forest. (d) ROC curves under the different data splits. The water-shedded curve are the ROC of individual experiments under different data split.

cal features alone significantly improved the classification accuracy, with average and median AUC scores reaching 0.71. Notably, combining both image and clinical features led to further improvements in performance, surpassing the use of either feature type alone.

Fig. 2 (d) presents the ROC curves for individual data splits (represented by transparent curves) along with the average ROC curve (represented by the darker curve) across the experiments. In these experiments, we used the ADC map as input with ResNet-10 as the image encoder, and both clinical and image features were employed in the random forest classification. This combined approach consistently demonstrated superior performance across all data splits.

To better understand the significance of each clinical variable, Fig. 3 and Fig. 4 present two complementary perspectives. In Fig. 3, we use a model-agnostic approach to assess feature importance by evaluating the change in prediction accuracy when the values of a feature are randomly shuffled (permuted). Fig. 4 shows the mean SHAP values across all clinical variables, highlighting the average impact of each variable on the model's output magnitude. Only the top 10 most influential variables are visualized. Both evaluations indicate that the greatest PSA prior to the first definitive treatment (`greatest_psa1`) [1] and whether the secondary Gleason score at the first definitive treatment is 3 (`sgs1=3`) play critical roles in the final prediction. Additionally, alterations in the physical dimensions of the prostate (`prostate_width`, `prostate_depth`, and `prostate_max_axis`) can significantly degrade the performance of our framework.

To gain more insights into where the image encoder derives information for its decision-making process, in Fig. 5,

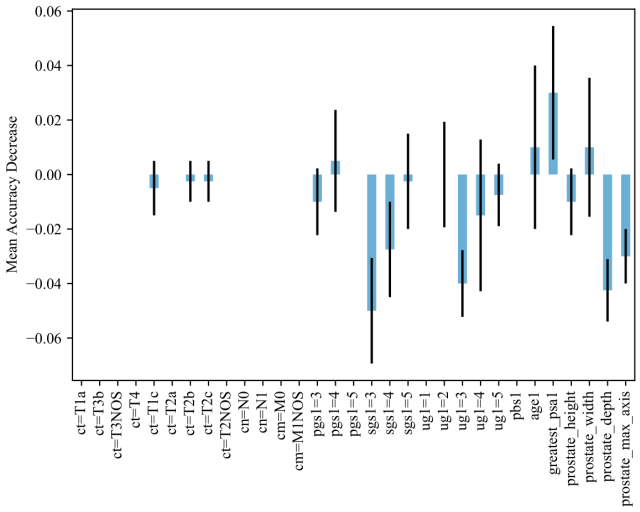

Fig. 3. Feature importance using permutation on full model.

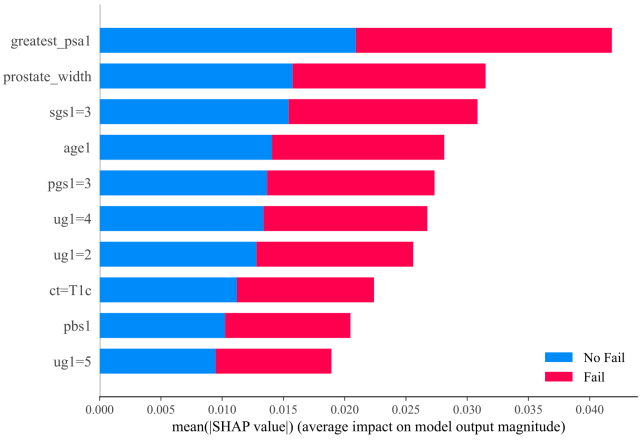

Fig. 4. Top 10 mean SHAP values of different features.

[1] `ct`: CT stage of record; `cm`: CM stage of record; `cn`: CN stage of record; `pgs1`: primary gleason score at first definitive treatment; `sgs1`: secondary gleason score at first definitive treatment; `ug1`: isup grade group at first definitive treatment; `pbs1`: percentage of positive biopsy cores at first definitive treatment; `age1`: age at first definitive treatment; `greatest_psa1`: greatest prostate-specific antigen (PSA) prior to first definitive treatment

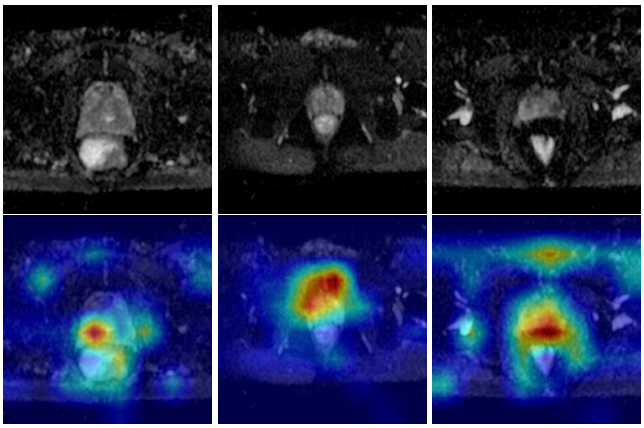

Fig. 5. GradCAM visualization on transversal MRI slice across different subjects. Top row: transversal prostate slice, bottom row: corresponding prostate slice overlaid with activation map.

we show the Grad-CAM activation maps corresponding to the most significant image features identified by SHAP values [5]. The activation maps align precisely with the prostate region, highlighting the ability of our framework to capture the most relevant features for the task.

3.1. Ablation Study

We conducted a series of ablation studies to evaluate the impact of the number of estimators, maximum depth, and maximum features on the random forest model's performance. We found that both the F1-score and accuracy increase monotonically with the number of estimators, plateauing when the number of estimators reaches 190. Regarding the maximum depth and maximum features, we found that optimal performance was achieved when these parameters were set to 5 and 1, respectively.

Additionally, we explored the use of alternative classification models within the same framework. Our results indicate that the random forest consistently outperforms other binary classification models, including SVM, AdaBoost, and a single decision tree. We attribute the random forest's significant performance margin to its ensemble nature, which is particularly beneficial given the relatively small size of our dataset.

4. CONCLUSIONS

In this study, we investigated effective methods for prostate cancer prognostication from multiple perspectives. By integrating macro-level MR images with carefully selected clinical variables, we achieved prognostication performance that is equivalent to or surpasses conventional risk classifiers and commercially available genomic classifiers. Our analysis revealed that when combined with imaging analyses, the highest PSA level prior to the first definitive treatment and the secondary Gleason score at the time of treatment are significantly inform the prognostication model. Furthermore, we demonstrated that MR image features alone are insufficient for accurate prediction; we can only achieve a reliable prognostic outcome by combining imaging features with clinical data with the current methodology.

5. ACKNOWLEDGMENTS

This work is supported by Huntsman Cancer Institute, University of Utah. This study was conducted under IRB_00173764 EXEMPTION UMBRELLA: Secondary and Molecular Analyses of Genitourinary Cancers.

6. REFERENCES

[1] Christina Fitzmaurice, Christine Allen, Ryan M Barber, Lars Barregard, Zulfiqar A Bhutta, Hermann Brenner, Daniel J Dicker, Odgerel Chimed-Orchir, Rakhi Dandona, Lalit Dandona, et al., "Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study," *JAMA oncology*, vol. 3, no. 4, pp. 524–548, 2017.

[2] Mohammad Hassan Houshdar Tehrani, Mohammadreza Gholibeikian, Abdolhamid Bamoniri, and Bi Bi Fatemeh Mirjalili, "Cancer treatment by caryophyllaceae-type cyclopeptides," *Frontiers in Endocrinology*, vol. 11, pp. 600856, 2021.

[3] Tzu Hao Wang, Cheng Yang Lee, Tzong-Yi Lee, Hsien Da Huang, Justin Bo-Kai Hsu, and Tzu Hao Chang, "Biomarker identification through multiomics data analysis of prostate cancer prognostication using a deep learning model and similarity network fusion," *Cancers*, vol. 13, no. 11, pp. 2528–, 2021.

[4] Andre Esteva, Jean Feng, Douwe van der Wal, Shih-Cheng Huang, Jeffry P Simko, Sandy DeVries, Emmalyn Chen, Edward M Schaeffer, Todd M Morgan, Yilun Sun, et al., "Prostate cancer therapy personalization via multi-modal deep learning on randomized phase iii clinical trials," *NPJ digital medicine*, vol. 5, no. 1, pp. 71, 2022.

[5] Scott M Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., pp. 4765–4774. Curran Associates, Inc., 2017.