Training Robust Radiomics-based Machine Learning Classifiers for Prediction of Prostate Cancer Disease Aggressiveness

Ana Carolina Rodrigues[1,2]
anacarolina.rodrigues@research.fchampalimaud.org
Inês Domingues[3,4]
inesdomingues@gmail.com
Nickolas Papanikolaou[1]
nickolas.papanikolaou@research.fchampalimaud.org

[1] Champalimaud Research, Champalimaud Foundation, Lisbon, Portugal
[2] Faculty of Medicine, University of Porto, Porto, Portugal.
[3] Instituto Politécnico de Coimbra, Instituto Superior de Engenharia, Coimbra, Portugal
[4] Centro de Investigação do Instituto Português de Oncologia do Porto (CI-IPOP): Grupo de Física Médica, Radiobiologia e Protecção Radiológica, Porto, Portugal

Abstract

Over the past decade, there has been growing evidence that artificial intelligence and radiomics may be helpful in the prediction of clinical outcomes in the entire prostate cancer disease continuum, such as the prediction of disease aggressiveness. However, the radiomics pipeline's dependence on segmentation masks has made it challenging to build machine-learning algorithms robust to inter- and intra-radiologist segmentation variability. With the goal of getting insight into the best methodology to build models that are robust to this heterogeneity, two radiologists were asked to draw whole prostate gland segmentations on T2W and DWI MRI examinations, and the resulting radiomic features calculated were used in several model training approaches: training with purely stable radiomic features according to their intraclass correlation coefficient (ICC); training independently with features extracted from each radiologist's mask; training with the feature average between both radiologists; extracting radiomic features from the intersection or union of the two masks; and creating a heterogeneous dataset by randomly selecting one of the radiologists' masks for each patient. The classifier trained with this last resampled dataset presented with the lowest generalization error, suggesting that training with heterogeneous data leads to the development of the most robust classifiers. On the contrary, removing features with low ICC resulted in the highest generalization error.

1 Introduction

In 2020, prostate cancer was the second most frequent cancer in men worldwide and ranked 5th in terms of mortality [10]. An accurate determination of clinical significance is essential for ascertaining the most appropriate treatment options and ensuring the best clinical outcome. With this purpose, artificial Intelligence and, in particular, radiomics have been reported to be predictive of prostate cancer disease aggressiveness [1, 2, 3, 4, 6, 7, 12]. However, a major limitation of this analysis is the tight link between the computed radiomic features and the delineated volume of interest (VOI) from where they have been extracted. The delineation of the VOI suffers from inter- and intra-radiologist variability[9] (Figure 1), which inevitably leads to feature value changes[13]. Hence, when these are used for model training, a lack of robustness is often found.

Thus, the purpose of this work was to find the best approach to train robust classifiers to minor differences in segmentation margins.

2 Methodology

Dataset: 181 patients with T2W, DW, and ADC exams from the SPIE-AAPM-NCI PROSTATEx challenge [5]. Manual segmentations of the whole prostate gland were performed by two radiologists on T2W and DWI. Radiomic features extracted using PyRadiomics[11] and used to train machine learning classifiers to predict clinical significance.

Seven approaches were compared: (1) Train only with ICC stable features; (2) Train with features extracted from masks drawn by radiologist 1; (3) Train with features extracted from masks drawn by radiologist 2; (4) Train with the feature average; (5) Train with features extracted from the masks' intersection; (6) Train with features extracted from the masks' union; (7) Train with a randomly resampled dataset.

All classifiers were tested on the same hold-out test sets and their performance was statistically compared.

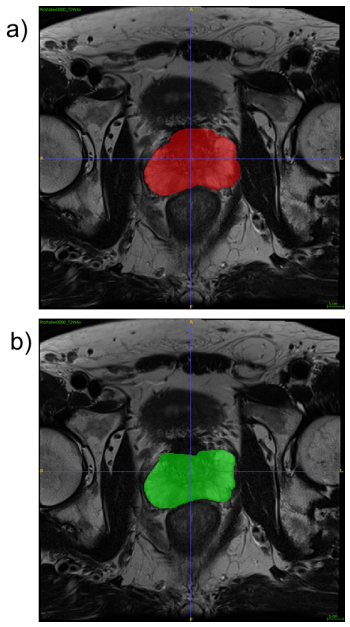

Figure 1: An example of the segmentation variability in the masks drawn by radiologist 1, (a), and radiologist 2, (b), on patient Prostatex0000.

A more comprehensive description of the methodology can be found at [8], as this work has already been published.

3 Results

Table 1 shows the performance of each model on the three hold-out test sets. The resampledRad dataset consistently produced the highest performance (approach 7). While the lowest-performing models were obtained by approach 1, which consisted of training classifiers only with features robust to segmentation differences.

Figure 2 shows ROC curve for the resampledRad classifier. The decision threshold chosen for this classifier was 0.32, which ensured a minimum 0.9 sensitivity on the train set.

The five features with the highest impact on model output were consistent on the three hold-out test sets. Thus, as an example, the shap analysis for the predictions of the resampledRad classifier on the resampledRad test set is displayed in Figure 3. The feature DWI_gradient_glcm_Imc1 is inversely associated with a clinically significant output, while the remaining four features are directly associated with it.

4 Discussion

In the current study, we attempted to answer the well-known issue of inter-reader variability introduced into the radiomics pipeline in the segmentation stage (RQ I). Our results revealed that the removal of unstable features through ICC, a technique currently recommended by radiomics guidelines and evaluated in the radiomics quality score, proved to produce the classifiers with the least ability to generalize to hold-out data. On the other hand, training classifiers with a radiomics dataset annotated by different radiologists, proved to be the highest performing across all

Table 1: Classification performance on three different hold-out test sets. The highest value per column is highlighted in **bold**.

a)

Training data	rad1 Hold-out test-set performance				
	F2	CohensKappa	AUC	Sensitivity	Specificity
stableRad1	0.6696	0.2042	0.7625	0.7895	0.4615
Rad1	0.7353	0.4636	0.7827	0.7895	0.7179
Rad2	0.7143	0.5110	**0.8219**	0.7368	**0.7949**
avgRad	0.6881	0.2758	0.7584	0.7895	0.5385
unionRad	0.7522	0.3073	0.7814	**0.8947**	0.4872
intersectionRad	0.6364	0.1863	0.6802	0.7368	0.4872
resampledRad	**0.7767**	**0.5055**	0.8198	0.8421	0.7179

b)

Training data	rad2 Hold-out test-set performance				
	F2	CohensKappa	AUC	Sensitivity	Specificity
stableRad1	0.6757	0.2275	0.7605	0.7895	0.4872
Rad1	0.7353	0.4636	0.7794	0.7895	0.7179
Rad2	0.6250	0.4208	0.8151	0.6316	**0.7949**
avgRad	0.7075	0.3525	0.7901	0.7895	0.6154
unionRad	0.7522	0.3073	0.7659	**0.8947**	0.4872
intersectionRad	0.6522	0.1371	0.6356	0.7895	0.3846
resampledRad	**0.7843**	**0.5351**	**0.8381**	0.8421	0.7436

c)

Training data	resampledRad Hold-out test-set performance				
	F2	CohensKappa	AUC	Sensitivity	Specificity
stableRad1	0.6696	0.2042	0.7537	0.7895	0.4615
Rad1	0.7353	0.4636	0.7841	0.7895	0.7179
Rad2	0.6633	0.4358	**0.8192**	0.6842	**0.7692**
avgRad	0.6944	0.3008	0.7746	0.7895	0.5641
unionRad	0.7522	0.3073	0.7827	**0.8947**	0.4872
intersectionRad	0.6522	0.1371	0.6430	0.7895	0.3846
resampledRad	**0.7767**	**0.5055**	**0.8192**	0.8421	0.7179

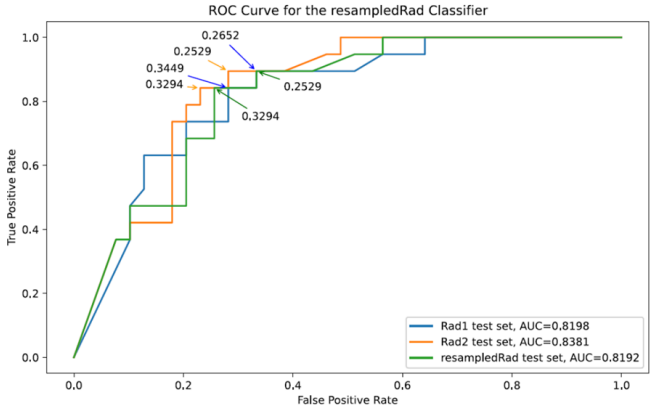

Figure 2: Receiver Operator Characteristics Curve for the resampledRad classifier when applied to the rad1, rad2 and resampledRad hold-out test sets, respectively in blue, orange and green. Some of the probability decision thresholds are included as annotations.

hold-out test sets. Supporting the hypothesis that the more heterogeneous the training data the more generalizable the results may be on unseen data. This classifier also performed similarly on the different hold-out test sets, indicating its robustness to radiologists with different years of experience. This was further confirmed by the performance on the resampledRad test set, which simulates a real-world clinical environment, where a deployed model would be used by several physicians.

5 Conclusions

Heterogeneous radiomics datasets where segmentation masks come from more than one radiologist produced classifiers with the highest generalization power. These results are extremely relevant for the clinical translation of AI models.

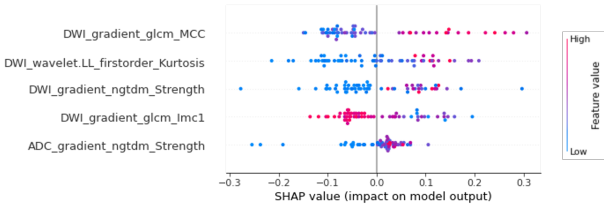

Figure 3: Five features with the highest impact on resampledRad model output according to a SHAP analysis.

References

[1] Simon Bernatz, Jörg Ackermann, Philipp Mandel, Benjamin Kaltenbach, Yauheniya Zhdanovich, Patrick N Harter, Claudia Döring, Renate Hammerstingl, Boris Bodelle, Kevin Smith, et al. Comparison of machine learning algorithms to predict clinically significant prostate cancer of the peripheral zone with multiparametric mri using clinical assessment categories and radiomic features. *European radiology*, 30(12):6757–6769, 2020.

[2] Giuseppe Cutaia, Giuseppe La Tona, Albert Comelli, Federica Vernuccio, Francesco Agnello, Cesare Gagliardo, Leonardo Salvaggio, Natale Quartuccio, Letterio Sturiale, Alessandro Stefano, et al. Radiomics and prostate mri: current role and future applications. *Journal of Imaging*, 7(2):34, 2021.

[3] Lixin Gong, Min Xu, Mengjie Fang, Jian Zou, Shudong Yang, Xinyi Yu, Dandan Xu, Lijuan Zhou, Hailin Li, Bingxi He, et al. Noninvasive prediction of high-grade prostate cancer via biparametric mri radiomics. *Journal of Magnetic Resonance Imaging*, 52(4):1102–1109, 2020.

[4] Simone Giovanni Gugliandolo, Matteo Pepa, Lars Johannes Isaksson, Giulia Marvaso, Sara Raimondi, Francesca Botta, Sara Gandini, Delia Ciardo, Stefania Volpe, Giulia Riva, et al. Mri-based radiomics signature for localized prostate cancer: a new clinical tool for cancer aggressiveness prediction? sub-study of prospective phase ii trial on ultra-hypofractionated radiotherapy (airc ig-13218). *European Radiology*, 31(2):716–728, 2021.

[5] Geert Litjens Oscar Debats Jelle Barentsz Nico Karssemeijer and Henkjan Huisman. "prostatex challenge data", the cancer imaging archive (2017). doi: 10.7937/K9TCIA.2017.MURS5CL.

[6] Tianping Li, Linna Sun, Qinghe Li, Xunrong Luo, Mingfang Luo, Haizhu Xie, and Peiyuan Wang. Development and validation of a radiomics nomogram for predicting clinically significant prostate cancer in pi-rads 3 lesions. *Frontiers in oncology*, 11, 2021.

[7] Federico Midiri, Federica Vernuccio, Pierpaolo Purpura, Pierpaolo Alongi, and Tommaso Vincenzo Bartolotta. Multiparametric mri and radiomics in prostate cancer: A review of the current literature. *Diagnostics*, 11(10):1829, 2021.

[8] Ana Rodrigues, Nuno Rodrigues, João Santinha, Maria V Lisitskaya, Aycan Uysal, Celso Matos, Inês Domingues, and Nickolas Papanikolaou. Value of handcrafted and deep radiomic features towards training robust machine learning classifiers for prediction of prostate cancer disease aggressiveness. *Scientific Reports*, 13(1):6206, 2023.

[9] P Steenbergen, K Haustermans, F Pos, R Oyen, S Heijmink, L De Wever, R Kalisvaart, J Teertstra, L Van den Bergh, and U Van der Heide. Prostate tumor delineation using multiparametric mri: Inter observer variability and pathology validation. *Radiotherapy and Oncology*, 111:S53–S54, 2014.

[10] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[11] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[12] Piotr Woźnicki, Niklas Westhoff, Thomas Huber, Philipp Riffel, Matthias F Froelich, Eva Gresser, Jost von Hardenberg, Alexander Mühlberg, Maurice Stephan Michel, Stefan O Schoenberg, et al. Multiparametric mri for prostate cancer characterization: Combined use of radiomics model with pi-rads and clinical parameters. *Cancers*, 12(7):1767, 2020.

[13] Binsheng Zhao. Understanding sources of variation to improve the reproducibility of radiomics. *Frontiers in oncology*, page 826, 2021.