Virtual biopsy in prostate cancer: can machine learning distinguish low and high aggressive tumors on MRI?

Virtual biopsy in prostate cancer: can machine learning distinguish low and high aggressive tumors on MRI?

Giulia Nicoletti, Davide Barra, Arianna Defeudis, Simone Mazzetti, Marco Gatti, Riccardo Faletti, Filippo Russo, Daniele Regge, and Valentina Giannini

***Abstract*— In the last decades, MRI was proven a useful tool for the diagnosis and characterization of Prostate Cancer (PCa). In the literature, many studies focused on characterizing PCa aggressiveness, but a few have distinguished between low-aggressive (Gleason Grade Group (GG) <=2) and high-aggressive (GG>=3) PCas based on biparametric MRI (bpMRI). In this study, 108 PCas were collected from two different centers and were divided into training, testing, and validation set. From Apparent Diffusion Coefficient (ADC) map and T2-Weighted Image (T2WI), we extracted texture features, both 3D and 2D, and we implemented three different methods of Feature Selection (FS): Minimum Redundance Maximum Relevance (MRMR), Affinity Propagation (AP), and Genetic Algorithm (GA). From the resulting subsets of predictors, we trained Support Vector Machine (SVM), Decision Tree, and Ensemble Learning classifiers on the training set, and we evaluated their prediction ability on the testing set. Then, for each FS method, we chose the best classifier, based on both training and testing performances, and we further assessed their generalization capability on the validation set. Between the three best models, a Decision Tree was trained using only two features extracted from the ADC map and selected by MRMR, achieving, on the validation set, an Area Under the ROC (AUC) equal to 81%, with sensitivity and specificity of 77% and 93%, respectively.**

***Clinical Relevance*— Our best model demonstrated to be able to distinguish low-aggressive from high-aggressive PCas with high accuracy. Potentially, this approach could help clinician to noninvasively distinguish between PCas that might need active treatment and those that could potentially benefit from active surveillance, avoiding biopsy-related complications.**

I. Introduction

Prostate Cancer (PCa) is the most frequently diagnosed cancer in men, alone accounting for 26% of new cancer diagnoses [1]. The Gleason Score (GS) is the gold standard for PCa risk stratification [2]. Based on GS, PCa is usually divided into three risk classes that refer to the probability of tumor progression: low risk (GS <7), intermediate risk (GS=7), and high risk (GS >7) [2]. Specifically, the intermediate-risk group shows a heterogeneous behavior, conferring to GS 4+3 a worst prognosis [3] and an increased risk of mortality [4]. With the aim of better defining the clinical distinction between GS 3+4 and 4+3, the 2014 International Society of Urological Pathology (ISUP) approved a new grading system, that limits the PCa grades in five Grade Groups (GG): 1 (GS 2-6), 2 (GS 3+4), 3 (GS 4+3), 4 (GS 8), 5 (GS 9-10) [5]. Currently, the GG is assigned by analyzing a biopsy sample. This method, however, does not ensure a complete representation of the characteristics of the tumor. Indeed, the GS calculated from a biopsy is often not the same as the GS calculated from the resected tumor, with implication on selecting the best treatment option on men diagnosed with PCa [6]. Thus, the introduction of PCa grading system based on imaging might solve the problem of the sampling site dependence and would also prevent the patient from invasive procedures.

Multiparametric Magnetic Resonance Imaging (mpMRI) has been widely used to detect and characterize PCa, also using Computer-Aided Diagnosis (CAD) systems [7]. However, mpMRI, involving Dynamic Contrast Enhanced (DCE) sequence, is a time consuming and invasive examination. To overcome these drawbacks, in the last few years, biparametric MRI (bpMRI), i.e., involving only T2-Weighted Image (T2WI) and Diffusion Weighted Image (DWI), without endorectal coil, has been proposed as a fast and noninvasive alternative to mpMRI. In literature, several studies assessed the association between PCa aggressiveness and textural features [8], [9] extracted from bpMRI, but only a small number of researchers focused on the characterization of PCa aggressiveness through bpMRI without endorectal coil [10]–[12], and, therefore, there is still a need for a multicenter validated study.

For all these reasons, the aim of this study is to provide a noninvasive method to distinguish between high-aggressive (GG>=3) and low-aggressive (GG<=2) PCa, based on texture features extracted from bpMRI examinations.

II. Materials and Methods

A. Patient cohort

This was a multi-center retrospective study. Patients were enrolled from the Candiolo Cancer Institute FPO-IRCCS (center A) and the AOU Città della Salute e della Scienza di Torino (center B). Study inclusion criteria were: 1) biopsy-proven PCa and 2) bpMRI examination of the prostate without endorectal coil, including T2WI and DWI. Exclusion criteria were: 1) presence of any image artifacts in the MRI and 2) patients who underwent Trans-Urethral Resection of the Prostate (TURP). The local ethics committee approved this retrospective study.

B. Magnetic Resonance Image acquisition

BpMRI in cohort A were collected with a 1.5 T MRI scanner (Optima MR450w, GE Healthcare, Milwaukee, WI, USA) using a 32-channel phased-array coil. The specific parameters set for T2WI were: slice thickness of 3 mm;

G.N., D.B., A.F., S.M., M.G., R.F., D.R., and V.G. are with the Department of Surgical Science, University of Turin, Via Genova 3, 10126 Turin, Italy (corresponding author e-mail: giulia.nicoletti@unito.it).

F.R. and D.R. are with the Department of Radiology at the Candiolo Cancer Institute, FPO, IRCCS, Strada Provinciale 142 km 3.95, 10060 Candiolo, Turin, Italy.

TR/TE/FA of 4640ms/102ms/160°; Field Of View (FOV) of 16×16 cm2; acquisition matrix of 256×192 with a reconstruction matrix of 512×512. DWI acquisition parameters were: slice thickness of 3 mm; TR/TE/FA of 6600ms/min/90°; acquisition matrix of 128×128 with a reconstruction matrix of 256×256; b-values equal to 50 and 1000 s/mm2. BpMRI of cohort B were collected with a 1.5 T MRI scanner (Achieva, Philips Medical System, Eindhoven, The Netherlands) using a 32-phased array coil. The specific parameters set for T2WI were: slice thickness of 3 mm; TR/TE/FA of 4570ms/100ms/90°; FOV of 18×18 cm2; acquisition matrix of 256×204 with a reconstruction matrix of 384×384. DWI was obtained with the same protocol except for TR/TE/FA of 4061/74/90; acquisition matrix of 64×63 with a reconstruction matrix of 96×96; b-values equal to 50, 1000, and 1700 s/mm2. ADC maps were created by fitting the DWI with the mono-exponential model.

C. Histopathology and Reference Standard

ADC map and T2WI were cropped and resampled in order to obtain the same FOV and spacing. A radiologist with more than 5 years of experience manually segmented all tumors based on the information provided by the biopsy report. The segmented mask was obtained using ITK Snap 3.8 that allows to overlay the mask on different MRI sequences. The radiologist carefully checked that the segmentation mask included the tumor area on both the ADC map and the T2WI. Therefore, for each tumor we obtained one segmentation, equal for the two MRI sequences. This single-mask approach was preferred instead of using two separate masks, one for ADC map and the other for T2WI, as it simulates the output obtained from a hypothetical CAD system which automatically provides a single PCa segmentation, from bpMRI.

As a reference standard we used the GS obtained after prostatectomy, when available, or after a targeted biopsy. A dedicated pathologist examined the Hematoxylin-Eosin-stained slides and recorded the GS. Then, we converted GS in GG (1 (GS 2-6), 2 (GS 3+4), 3 (GS 4+3), 4 (GS 8), 5 (GS 9-10)) and defined low-aggressive tumors those with GG <=2 and high-aggressive PCas those with GG>=3.

D. Dataset division

The partition of a dataset in construction set and validation set is a crucial step. Indeed, we wanted to balance the number of cases in each class, but at the same time we wanted to have a cohort representative of the target population. To cope with this aspect, we decided to divide the dataset in a construction set and a validation set, based on tumor volumes and aggressiveness. More specifically, we split lesions into the two classes of aggressiveness. Each of them was sorted in ascending order of volume and divided into four equinumerous groups. Next, we randomly selected the same number of lesions from each group, in order to create a construction set composed of the 75% of the entire dataset. Remaining dataset samples (25%) were included in the validation set, used to evaluate the generalization capability of the best models. From the construction set, we further randomly chose the 70% of the high-aggressive cases and the 70% of low-aggressive ones to create the training set, using the leftover 30% for the testing set.

E. Feature extraction

Both 3D and 2D radiomic features were extracted, using the open-source python package Pyradiomics [13], to be compliant with the Image Biomarker Standardization Initiative (IBSI) [14], thus ensuring reproducibility. As suggested by IBSI, we resampled all images in order to obtain a rotationally invariant voxel for the calculation of 3D features. For this reason, after the application of a Gaussian filter, with sigma equal to the dimension of the pixel, we decided to interpolate all images along the three dimensions with the same spacing (0.5 mm for x, y, z). The pixel range was set between the 1st and 99th percentiles, removing all other pixels from the mask. Then, texture features were calculated from Gray Level Cooccurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Neighboring Gray Tone Difference Matrix (NGTDM), and Gray Level Dependence Matrix (GLDM). GLCM and GLRLM were computed with a fixed number of bins (32) and their features were calculated on the mean of matrices which were computed in all spatial directions (13 and 4 respectively for 3D and 2D). Note that, in a preliminary phase we tested different number of bins (32 and 64), but the results, in terms of feature values and classifier performances, were comparable. For this reason, here, we report the results obtained with bin value of 32, randomly chosen between the two. The values of the remaining parameters were set by default using the Pyradiomics package [13].

F. Feature selection and classifier construction

The Feature Selection (FS) phase allows to decrease the computational time and complexity of the model, while increasing its simplicity and interpretability. In particular, we used the following three FS methods, and we evaluated the efficacy of the resulting feature subsets for predicting the two classes of PCa aggressiveness.

Minimum Redundance Maximum Relevance (MRMR) algorithm was used to assign an importance score to each feature. Then, features with nonzero score were selected.

Affinity Propagation (AP) was computed to perform feature clustering. Then, the exemplar of each cluster was chosen and included in the optimal feature subset. For the AP algorithm, we set the value of the damping factor (0.5), the maximum number of total iterations (200), and the maximum number of iterations for which clusters do not change (20).

Genetic Algorithm (GA) was used as a search and optimization tool. A Support Vector Machine (SVM) classifier was trained, on the training set, with the feature subset selected by the chromosome. Then, in order to evaluate the goodness of the solution, the fitness value was calculated as follow:

$$\text{Fitness} = 1 - (se + sp)/2 \quad (1)$$

where *se* and *sp* are, respectively, sensitivity and specificity that the SVM model obtained on the testing set. To dichotomize the SVM output probabilities, the cut-off value was binary encoded, from zero to one in steps of 1/31 in the final five bits of each solution, and was optimized by the GA. For the solution, we used a binary codification, i.e., 0 for unselected feature and 1 for selected one. Therefore, the number of genes was set equal to the number of variables plus

the five additional bits for the cut-off optimization. In all, GA was performed three times for each dataset, in order to evaluate the three different SVM kernels (linear, polynomial, and Gaussian). For each of these repetitions, we chose the best solution, i.e., that with the lower fitness value. Other parameters set were the number of individuals (500), the number of iterations (2500), the number of parents (80% of the number of individuals), the number of repetitions (5), the mutation probability (20%), and the crossover probability (100%).

The feature subsets obtained from MRMR and AP algorithms were used to train the following classifiers: SVM (with linear, polynomial, and Gaussian kernel), ensemble learning classifiers (AdaBoost and RobustBoost), and Decision Tree. During the training phase, to dichotomize the output, we identified the best cut-off by maximizing the Youden Index. The features selected by GA were evaluated with a SVM classifier, characterized by the corresponding kernel and optimized cut-off of the chosen solution. Lastly, we chose the best classifier for each FS method by calculating the average of the accuracy obtained on the training set and on the testing set. In case of models with equal mean accuracy, we chose the model with higher value of accuracy, Negative Predictive Value (NPV) or sensitivity, on the testing set. Then, we evaluated the ability of the selected best models to generalize on the validation set. All algorithms were implemented in MATLAB ®2020b, except for the feature extraction, that was implemented in Python 3.8.

III. Results

A. Patient characteristics and dataset partition

A total of 108 PCas were included in the dataset (Table 1), 55 of which were low-aggressive lesions and the remaining 53 high-aggressive. 81 PCas (75%) were randomly assigned to the construction set (57 for training and 24 for testing) and the remaining 27 to the validation set. The method used to partition the dataset allowed to obtain a balance of the two classes (low-aggressive/high-aggressive) in training (29/28), testing (12/12), and validation set (14/13).

B. Feature extraction

In total, we computed the following 76 features, both from ADC map and T2WI, and both 3D and 2D: 1) ROI volume (mm3), 2) 24 features from GLCM, 3) 14 features from GLDM, 4) 16 features from GLRLM, 5) 16 features from GLSZM, and 6) 5 features from NGTDM.

C. Feature selection and classification

Regarding the three different FS methods, the number of features that were selected by AP (2.8±0.5) and MRMR (9.5±8.2) was always smaller than that by GA (29.4±7.4). Specifically, considering the three groups of features used to train the best models, no feature was found in common among all three subsets.

In Table 2, we reported the performances of the chosen three best classifiers, on training, testing, and validation sets. In particular, all other models trained on AP feature subsets achieved a lower value of the train-test mean accuracy (from 59.4% to 76.6%) than the chosen one (77.5%), except for a model that obtained the same value of accuracy on both training and testing, but a lower NPV on the latter (69.2% vs 72.7%). All MRMR-derived models obtained a lower mean accuracy (from 54.2% to 77.1%) than the best one (78.3%), except for a case in which the classifier achieved a higher mean accuracy (78.6%) but with a lower accuracy (62.5% vs 70.8%) and NPV (63.6% vs 77.8%) on the testing set. Regarding GA models, all classifiers achieved a lower value of mean accuracy (from 65.1% to 76.8%) than the chosen best model (77.6%), except for one that obtained a higher mean accuracy (89.9%), but, on the testing set, the same accuracy (83.3%), and a lower value of sensitivity (66.7% vs 83.3%) and NPV (75% vs 83.3%). Thus, looking at the results of the three best classifiers, MRMR and AP models achieved high performances on the training set and slightly lower performances on the testing set. This may suggest a possible overfitting, but observing the performances on the validation set, both models showed a good ability to generalize (AUC of 81.3% [95%CI:0.65-0.98] and 81.9% [95%CI:0.65-0.98], respectively). In contrast, GA model obtained higher results on the testing set than on the training set. However, the performances on the validation set (AUC of 87.4% [95%CI:0.73-1.00]) confirmed its generalization capability.

TABLE I. Clinical characteristics of the dataset

Grade Group	TRAIN n(%n/N) (center A/B)	TEST n(%n/N) (center A/B)	VAL n(%n/N) (center A/B)	TOT. n(%n/N) (center A/B)
1	7(12) (7/0)	4(17) (4/0)	4(15) (4/0)	15(14) (15/0)
2	22(39) (18/4)	8(33) (7/1)	10(37) (9/1)	40(37) (34/6)
3	14(25) (9/5)	7(30) (2/5)	9(33) (4/5)	30(28) (15/15)
4	10(17) (5/5)	3(12) (3/0)	4(15) (2/2)	17(16) (10/7)
5	4(7) (2/2)	2(8) (2/0)	0(0) (0/0)	6(5) (4/2)
TOT N (center A/B)	57 (41/16)	24 (18/6)	27 (19/8)	108 (78/30)

IV. Discussion and Conclusion

With our three best classifiers, we obtained an accuracy greater than 70% and 74% on testing set and validation set, respectively, and a NPV greater than 72% on both sets. Interestingly, after choosing the best models, we noticed that all three were trained with 2D features derived from ADC maps. This result is consistent with other studies, where, to predict the PCa aggressiveness, researchers decided to focus only on the ADC map [12], as it is considered highly relevant to differentiate low/high GG. Comparing the three chosen classifiers, the Decision Tree seems to be the best one: it achieved results comparable to those of the other two models but using a lower number of features, with a simpler model structure, and, therefore, with the advantage of being easier to interpret and reproduce.

In the literature, the distinction between PCa with GG<=2 and GG>=3 based on bpMRI remains challenging. Chaddad et al. [10] proposed a Joint Intensity Matrix (JIM), a method to compute the joint intensity distribution between ADC map and T2WI. They combined JIM and GLCM features to train a Random Forest classifier and obtained an AUC of 65% on the testing set. Jensen et al. [11] extracted histogram and textural features from T2WI and DWI and trained a KNN classifier. Performing a three-fold cross validation, they reached an AUC

TABLE II. CLASSIFIER PERFORMANCES

FS (n)	Dataset, classifier	AUC train test val	Acc train test val	Se train test val	Sp train test val	NPV train test val	PPV train test val
MRMR (1GLRLM, 1GLSZM)	2D ADC, Decision Tree	91.7 69.8 81.3	85.9 70.8 85.2	92.8 83.3 76.9	79.3 58.3 92.9	92.0 77.8 81.3	81.2 66.7 90.9
AP (1GLCM, 1GLRLM, 1GLDM)	2D ADC, SVM (polyno-mial)	92.7 71.5 81.9	84.2 70.8 74.1	85.7 75.0 69.2	82.8 66.7 78.6	85.7 72.7 73.3	82.8 69.2 75.0
GA (13GLCM, 3GLRLM, 6GLDM, 7GLSZM, 1NGTDM)	2D ADC, SVM (linear)	82.9 81.3 87.4	71.9 83.3 81.5	64.3 83.3 69.2	79.3 83.3 92.9	69.7 83.3 76.5	75.0 83.3 90.0

The performances are reported in percentage. In parentheses (n) the number of features of the five texture matrices used by each model in the final training. Acc, Accuracy; AP, Affinity Propagation; AUC, Area Under the ROC Curve; FS, feature selection; GA, Genetic Algorithm; MRMR, Minimum Redundance Maximum Relevance; NPV, Negative Predictive Value; PPV, Positive Predicted Value; Se, Sensitivity; Sp, Specificity; SVM, Support Vector Machine.

of 96% and 83% on peripheral and transitional PCas, respectively. Bernatz et al. [12] trained a Random Forest classifier, with PIRADS and Maximum3D diameter as features extracted from ADC images and obtained a mean AUC of 76% on a 100-fold Cross-Validation (CV). In a comparison with these three studies, our work achieves, in one case, better performances [10], while, in the others, comparable results [11], [12] but with the advantage of being a multicentric study, providing more generalizable findings. In addition, unlike two of these studies [11], [12], we decided to not use CV. The use of CV means that the performances obtained are an average of the results of multiple models, but our aim was to create a single classifier, which, in future, could be used in a CAD system. For this reason, we demonstrated the robustness of our classifiers on two different sets of new samples (testing set and validation set). Furthermore, as in the case of Bernatz et al. [12], our approach is IBSI compliant, an important requirement to ensure, in future development, an easy reproducibility of our predictors.

This study has some limitations. First, we used an internal validation set. Images derived from the two centers, A and B, were qualitatively different, with different acquisition parameters, and, for this reason, we decided to use images of both centers in the construction set. A broad validation set is still needed, in order to demonstrate that the models are able to generalize on datasets not used in the construction set. To overcome this issue, we are planning to reformulate the dataset division in order to use the center B only as validation set, but also to include imaging from a third center in order to externally validate the three best models presented in this paper. Second, the spacing value used to interpolate images before extracting the features was equal for both ADC map and T2WI. Therefore, spacing was greater than that of the original T2WI pixel, while it was smaller than that of the ADC map. For this reason, in future we will resample images with a spacing bigger than 0.5mm, in order to include the entire pixel of the ADC map.

In conclusion, in this study we evaluated different machine learning techniques to create a classifier able to distinguish low- from high- aggressive PCas, and we demonstrated the ability of our best models to generalize on new samples. Using bpMRI without contrast and endorectal coil, our approach can help clinicians to find PCa that could be monitored with active surveillance in a fast and noninvasive way.

ACKNOWLEDGMENT

This work has received funding from the Fondazione AIRC under IG2017 - ID.20398 project – P.I. Regge Daniele and from the European Union's Horizon 2020 research and innovation programme under grant agreement no 952159.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer Statistics, 2021," *CA. Cancer J. Clin.*, vol. 71, no. 1, pp. 7–33, 2021, doi: https://doi.org/10.3322/caac.21654.

[2] N. Mottet *et al.*, "EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer," in *European urology*, Arnhem, The Netherlands: EAU Guidelines Office.

[3] T. Y. Chan, A. W. Partin, P. C. Walsh, and J. I. Epstein, "Prognostic significance of Gleason score 3+4 versus Gleason score 4+3 tumor at radical prostatectomy.," *Urology*, vol. 56, no. 5, pp. 823–827, Nov. 2000, doi: 10.1016/s0090-4295(00)00753-6.

[4] J. L. Wright *et al.*, "Prostate cancer specific mortality and Gleason 7 disease differences in prostate cancer outcomes between cases with Gleason 4 + 3 and Gleason 3 + 4 tumors in a population based cohort.," *J. Urol.*, vol. 182, no. 6, pp. 2702–2707, Dec. 2009, doi: 10.1016/j.juro.2009.08.026.

[5] J. I. Epstein, L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, and P. A. Humphrey, "The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System.," *Am. J. Surg. Pathol.*, vol. 40, no. 2, pp. 244–252, Feb. 2016, doi: 10.1097/PAS.0000000000000530.

[6] I. T. Köksal, F. Ozcan, T. C. Kadioglu, T. Esen, I. Kiliçaslan, and M. Tunç, "Discrepancy between Gleason scores of biopsy and radical prostatectomy specimens.," *Eur. Urol.*, vol. 37, no. 6, pp. 670–674, Jun. 2000, doi: 10.1159/000020216.

[7] R. Suarez-Ibarrola *et al.*, "Artificial Intelligence in Magnetic Resonance Imaging-based Prostate Cancer Diagnosis: Where Do We Stand in 2021?," *Eur. Urol. Focus*, Mar. 2021, doi: 10.1016/j.euf.2021.03.020.

[8] A. Vignati *et al.*, "Texture features on T2-weighted magnetic resonance imaging: new potential biomarkers for prostate cancer aggressiveness.," *Phys. Med. Biol.*, vol. 60, no. 7, pp. 2685–2701, Apr. 2015, doi: 10.1088/0031-9155/60/7/2685.

[9] G. Nketiah *et al.*, "T2-weighted MRI-derived textural features reflect prostate cancer aggressiveness: preliminary results.," *Eur. Radiol.*, vol. 27, no. 7, pp. 3050–3059, Jul. 2017, doi: 10.1007/s00330-016-4663-1.

[10] A. Chaddad, M. J. Kucharczyk, and T. Niazi, "Multimodal Radiomic Features for the Predicting Gleason Score of Prostate Cancer.," *Cancers (Basel).*, vol. 10, no. 8, Jul. 2018, doi: 10.3390/cancers10080249.

[11] C. Jensen, J. Carl, L. Boesen, N. C. Langkilde, and L. R. Østergaard, "Assessment of prostate cancer prognostic Gleason grade group using zonal-specific features extracted from biparametric MRI using a KNN classifier.," *J. Appl. Clin. Med. Phys.*, vol. 20, no. 2, pp. 146–153, Feb. 2019, doi: 10.1002/acm2.12542.

[12] S. Bernatz *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.," *Eur. Radiol.*, vol. 30, no. 12, pp. 6757–6769, Dec. 2020, doi: 10.1007/s00330-020-07064-5.

[13] J. J. M. van Griethuysen *et al.*, "Computational Radiomics System to Decode the Radiographic Phenotype.," *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[14] A. Zwanenburg *et al.*, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020, doi: 10.1148/radiol.2020191145.