Radiomics to predict response to neoadjuvant chemotherapy in rectal cancer: influence of simultaneous feature selection and classifier optimization

Radiomics to predict response to neoadjuvant chemotherapy in rectal cancer: influence of simultaneous feature selection and classifier optimization.

S. Rosati, C. M. Gianfreda, G. Balestra
Department of Electronics and Telecommunications
Politecnico di Torino
Torino, Italy
samanta.rosati@polito.it

V.Giannini, S. Mazzetti, D. Regge
Department of Surgical Sciences, University of Turin,
Torino, Italy
Department of Radiology, Candiolo Cancer Institute, FPO
- IRCCS, Candiolo (TO), Italy
Torino, Italy
valentina.giannini@polito.it

***Abstract*—According to the guidelines, patients with locally advanced colorectal cancer undergo neoadjuvant chemotherapy. However, response to therapy is reached only up to 30% of cases. Therefore, it would be important to predict response to therapy before treatment. In this study, we demonstrated that the simultaneous optimization of feature subset and classifier parameters on different imaging datasets (T2w, DWI and PET) could improve classification performance. On a dataset of 51 patients (21 responders, 30 non responders), we obtained an accuracy of 90%, 84% and 76% using three optimized SVM classifiers fed with selected features from PET, T2w and ADC images, respectively.**

Keywords—SVM optimization; response to chemoradiotherapy; rectal cancer; feature selection; Genetic algorithms.

I. INTRODUCTION

Colorectal cancer (CRC) is the third leading cause of cancer-related mortality in Western countries [1]. The standard of care for locally advanced rectal cancer (LARC) is neoadjuvant chemoradiotherapy (CRT), followed by surgical resection [2, 3]. However, pathological complete response (pCR) after CRT is reached only in a minority of cases (8-24%) [4]. Patients with pCR after CRT have a significantly improved prognosis, while non-responding patients should probably be addressed to other therapeutic strategies, without delaying surgery and avoiding unnecessary toxicity. However, at present no diagnostic examination can assess the likelihood of pCR before treatment, with the aim of better selecting eligible patients for CRT if the regimen is predicted to be successful. Recently, some radiomics features from magnetic resonance imaging (MRI) and 18F-FDG PET/CT have been proven useful predictors of tumour response [5–7]. However, despite some promising results, the literature on this topic needs to be expanded, exploring the relationships between different features, and evaluating different Features Selection (FS) and classification algorithms.

FS is a fundamental step when dealing with high-dimensional data, as it highlights those variables that are redundant or irrelevant for the system description. Moreover, it has been proven that FS increases the classification performances [8], due to the removal of variables introducing noise during the classifier construction and application. The algorithms proposed for FS can be mainly divided into two categories: filter and wrapper methods [9]. Filter methods perform FS independently of the learning algorithm. This means that variables are examined in order to identify those more relevant for describing the inner structure of the analysed dataset. Since each variable is considered independently during the selection procedure, groups of features having strong discriminatory power are ignored. Conversely, in wrapper methods, the selection of the feature subset is performed together with the estimation of its goodness in the learning task. The latter method can reach better performance since it allows to explore feature dependencies. On the other hand, it could be computationally very expensive and the obtained feature subset is optimized for the specific learning algorithm or classifier.

Besides, when classifiers are developed, it is compulsory to tune classifiers' parameters, since they strongly influence the classification performance [10]. Several approaches have been developed in this direction, e.g. grid search, random search, heuristic search [11]. However, fixed a specific set of parameter to be used for a classifier, different results might be obtained using different input feature subsets and vice-versa. For this reason the selection of the optimal feature subset should be conducted simultaneously to the optimization of the classifier parameters to reach higher performances. Since an exhaustive search of the best couple *feature subset-classifier parameters* is unfeasible in most real situations, heuristic search represents a convenient way to find a good compromise between reasonable computational time and sub-optimal solutions. In particular, Genetic Algorithms (GAs) have been applied for solving optimization problems connected to FS [8] and parameters

tuning [10] , but very poor applications can be found for the simultaneous optimization of both aspects.

The aim of this study is to evaluate the influence of simultaneous optimization of feature subset and classifier parameters on different image datasets (T2w, DWI and PET) in predicting response to CRT.

II. Materials and Methods

A. Population

The dataset was composed of patients with biopsy-confirmed stage III/IV locally advanced rectal cancer (LARC) that underwent axial MRI examination, including T2-weighted (T2w), diffusion weighted imaging (DWI), and fluoro-D-glucose (FDG) PET performed at our Institute prior to neoadjuvant chemo-radiotherapy treatment (CRT). After the completion of CRT, tumors were resected and evaluated by an experienced pathologist to assess histopathological tumor response (TRG) according to the Mandard scheme [12]. Signed informed consent was obtained from all participants before entering the study.

B. Reference standard

Patients were classified as responders (R+) or non-responders (R-) according to their pathological TRG as follows:

- R+ patients included TRG 1 that represents a complete regression and TRG 2 that represents a partial response;
- R- patients included TRG 3, defined as fibrosis outgrowing residual tumor, TRG 4, defined as residual tumor outgrowing fibrosis and TRG 5 that represents a complete non response.

C. Tumor segmentation

To segment tumors on MRI images a semi-automatic algorithm has been developed using C++ and ITK libraries [13]. The method consists of different steps: a) first a k-means algorithm is applied on both T2w and DWI images (k=3 for the T2w sequence and k=5 for the DWI image). Then, the segmentation mask is created taking into account all voxels belonging to the intersection between voxels classified in cluster 1 (i.e., the lower mean intensity value) in the T2w images and in cluster 4 and 5 (i.e., the 2 clusters with the highest intensity values) in DWI images. Finally, only the 2D biggest connected region is kept as the final region of interest, while other non-connected regions (i.e., noise, vessels, regions outside the tumor) are discarded. Once the automatic segmentation is performed, an experienced radiologist (more than 10 years of experience in interpreting abdominal MRI) manually reviewed the results of segmentation on both T2w and ADC maps to refine the masks.

Segmentation of tumors on PET images is obtained using the previously described automatic Adaptive Threshold Algorithm [14]. Briefly, this method consists of the following steps: a) a nuclear medicine physician draws a background area close to the lesion, b) then the algorithm iteratively determines a threshold value based on the percentage of the maximum intensity in the cross-section area of a sphere containing the tumor, c) finally, all masks are reviewed by an expert nuclear medicine physician.

D. Feature Extraction

From all voxels belonging to the largest 2D slice of the segmented masks in the T2w, ADC and PET images, the following radiomics features are extracted: a) 5 first order parameters, i.e. mean intensity, median intensity, 10^{th}, 25^{th} and 75^{th} percentile; b) 21 second order texture parameters derived from the Gray-Level Co-Occurrence matrix (GLCM) in which the intensities within each ROI were rescaled between the 1st and the 99th percentile using 64 bins. Four GLCMs are generated considering all directions of the 2D image (distance = 1 pixel) and then averaged to be rotationally invariant to the distribution of texture; c) for PET images metabolic volume, defined as the area of the segmented PET mask, and glycolytic volume, which is the product between metabolic volume and SUV_{mean}. The complete list of the features extracted for the three image datasets is reported in Table I.

E. Feature Selection and Classifier Optimization

GAs were used for the simultaneous optimization of feature subset and classifier parameters. In particular, a GA was implemented for each image dataset and the Support Vector Machine (SVM) was chosen for classification in all cases.

We codified each GA solution using a binary vector with the first part used for FS and the second part for parameters optimization. For FS, we set a number of bits equal to the total number of available features for the specific image dataset: we identified a feature included in the subset with the corresponding bit equal to "1". For parameters' optimization, the last four bits of each solution were used, two bits for the selection of the penalty term C and two bits for the choice of the kernel function. For the penalty term, the following values were explored: 1, 10, 50, 100; for the kernel we examined linear, Gaussian, polynomial of order 2 and polynomial of order 3 functions.

The fitness of each solution was measured according to the following equation:

$$fitness = 1 - \left(\frac{|spec+sens|}{2}\right) + 0.3 * |spec - sens| \qquad (1)$$

where *sens* and *spec* are the sensitivity and specificity obtained using a SVM classifier with parameters set according to the last four bits of the solution and trained using the feature subset specified by the first part of the solution. The SVM was validated using a k-fold cross-validation with k=5, using all available patients. The same 5 partitions of the dataset were used for calculating the fitness of each solution, in order to compare different solutions in equal conditions. In eq. (1), the first part of the formula aimed to maximize the classification accuracy, while the second part was a penalty term introduced to balance the performances for the two classes. In general, lower fitness values were associated to better solutions.

In order to widely explore as much as possible solutions, we started our GAs generating 400 solutions that were evolved for 5000 iterations. A further stopping condition was introduced to stop the algorithm if no improvements in the best fitness value were observed for 250 consecutive iterations, corresponding to

the 5% of total number of iterations. A crossover operator was implemented with 4 cut-points and probability equal to 0.9. The initial mutation probability was set to 0.2 and it was progressively reduced to 0.15 and 0.1 after 200 and 400 iterations respectively. The GA for each image dataset was repeated 50 times starting from the same initial population, thus obtaining 50 solutions for each dataset.

F. Performance Evaluation

For each image dataset, we selected the best 3 solutions according to their fitness values (i.e. solutions with the lowest fitness values) from the corresponding GA results. Accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated for the selected solutions. For this purpose, we used a SVM classifier having *C* value and kernel function specified by the last four bits of the solution and fed with the feature subset specified by the first part of the solution. The performances were evaluated using a 5-fold cross-validation. Moreover, a majority voting procedure was implemented across the classification results of the 3 best solutions for each kind of image, to try to reduce the classification errors [15].

In order to better evaluate the effect of simultaneous FS and classifier optimization, we compared the GA results with those reached by a SVM classifier trained with the whole set of features extracted for the 3 image datasets and validated using the 5-fold cross-validation. All possible combinations of *C* values and kernel function were tested in this situation, and the fitness value for each case was calculated according to eq. (1). Similarly to GA results, the best set of parameters was selected according to the fitness value and accuracy, sensitivity, specificity, PPV and NPV for the corresponding SVM were calculated as a comparison.

TABLE I. GA RESULTS FOR THE 3 BEST SOLUTIONS FOR EACH IMAGE DATASET. BLUE CELLS HIGHLIGHT SELECTED FEATURES, WHILE GRAY CELLS REPRESENT FEATURES NOT EXTRACTED.

	PET			T2			ADC		
	1st sol	2nd sol	3rd sol	1st sol	2nd sol	3rd sol	1st sol	2nd sol	3rd sol
Autocorrelation	0	0	0	0	1	1	1	0	0
Contrast	0	0	0	0	0	1	0	1	0
Correlation1	0	0	0	1	0	0	0	0	0
Correlation2	0	0	0	0	1	1	0	0	0
Cluster Prominence	0	0	0	0	0	0	0	1	0
Cluster Shade	0	0	0	0	0	0	1	1	1
Dissimilarity	0	1	0	0	0	1	0	0	0
Energy	1	1	1	1	0	0	1	1	0
Entropy	0	1	1	0	1	0	1	0	1
Homogeneity1	1	0	0	0	0	1	0	0	1
Homogeneity2	0	0	0	0	0	0	0	0	0
Maximum probability	0	0	0	0	0	0	0	0	0
Sum of squares: Variance	0	0	0	1	0	0	0	1	1
Sum average	0	0	0	1	0	0	0	1	1
Sum variance	0	0	0	1	1	1	1	1	0
Sum entropy	0	0	0	1	1	1	0	0	0
Difference variance	0	0	1	0	1	1	0	0	0
Difference entropy	0	0	0	0	0	1	0	1	0
Inf. measure of correlation1	0	0	0	1	1	1	1	0	0
Inf. measure of correlation2	0	0	0	1	1	1	1	0	1
Inverse diff. normalized	1	1	1	0	1	1	0	0	0
Inverse diff. moment normalized	0	0	0	1	0	1	0	1	0
Mean	0	1	1	1	0	0	0	0	0
Median	1	1	1	1	1	1	0	0	0
10th	0	1	1	1	1	1	0	1	1
25th	1	0	1	1	1	1	1	0	0
75th	0	1	1	0	0	0	0	0	0
SUVmax	0	0	1						
volMetab	0	0	0						
volGlic	1	1	1						
# of selected features	6	9	11	13	12	16	8	10	7
Fitness value	0.126	0.134	0.141	0.179	0.179	0.179	0.237	0.237	0.255

TABLE II. CLASSIFICATION RESULTS

		C value	Kernel function	acc	sens	spec	PPV	NPV
PET	**1st sol**	100	Poly3	0.882	0.867	0.905	0.929	0.826
	2nd sol	50	Poly2	0.882	0.900	0.857	0.900	0.857
	3rd sol	50	Poly2	0.863	0.867	0.857	0.897	0.818
	Maj. Voting	-	-	0.902	0.933	0.857	0.903	0.900
	All feat	10	Gauss	0.627	0.600	0.667	0.720	0.538
T2	**1st sol**	50	Poly3	0.843	0.867	0.810	0.867	0.810
	2nd sol	100	Poly3	0.843	0.867	0.810	0.867	0.810
	3rd sol	50	Poly2	0.843	0.867	0.810	0.867	0.810
	Maj. Voting	-	-	0.843	0.867	0.810	0.867	0.810
	All feat	100	Poly2	0.686	0.767	0.571	0.719	0.632
ADC	**1st sol**	100	Poly3	0.765	0.767	0.762	0.821	0.696
	2nd sol	100	Gauss	0.765	0.767	0.762	0.821	0.696
	3rd sol	100	Poly3	0.804	0.867	0.714	0.813	0.789
	Maj. Voting	-	-	0.765	0.800	0.714	0.800	0.714
	All feat	100	Gauss	0.608	0.633	0.571	0.679	0.522

III. RESULTS AND DISCUSSION

The final dataset included 51 patients (35 men and 16 women). Among them, 21 were pR+ (n=8 with TRG=1, n=13 with TRG=2), and 30 were pR- (n=16 with TRG=3, n=13 with TRG=4, n=1 with TRG=5).

Table I shows the features selected by the GAs in the three best solutions for each image dataset, and the fitness values reached by each solution. In general, features selected when using PET images are more reproducible than those selected when using MRI images. Indeed, considering the solutions with the least number of selected features, the GAs with PET features has 4 out of 6 features that were also selected by the other two solutions, while the GAs with T2w images has 7/12 and that with ADC features only 1 out of 7. This could indicate that some parameters extracted from PET images could better identify responder patients regardless the heterogeneity that could exists between different patients. This characteristic is also depict by the fact that the GAs fed with PET parameters reach lower values of fitness with fewer number of parameters.

Table II summarizes the classification performances obtained with the 3 best solutions, the majority voting of these solutions and the whole set of features for each image dataset. Moreover, the optimal SVM parameters for the GA solutions and using all features are reported. Analysing the optimized parameters found by the GAs for the SVM, it can be observed that very high values of the penalty term *C* are needed (50 or 100), associated with a polynomial kernel function in 8 out of 9 cases. Only the second solution of the ADC dataset uses a Gaussian kernel function. Conversely, employing the whole set of features the Gaussian function results best SVM kernel function in 2 out of 3 datasets, i.e. PET and ADC. Finally, it is important to notice that the linear kernel has never been chosen. One explanation could relies in the fact that response to therapy and texture features are not linearly dependent, therefore more complex kernels are needed to face the problem in order to provide reliable solutions.

Results show that feature selection is compulsory when dealing with such an high number of features in order to reach acceptable results. Indeed, when the optimized SVMs were fed

with all features the accuracy of all dataset was below 70%, while, when only selected features were used, the accuracy increased of 25%, 16% and 16% for PET, T2w and ADC dataset, respectively. This means that some variables removed by GAs are source of noise for the classification task.

In general, analysing the performances of the nine GA solutions, it is evident that the best imaging method for predicting patients' response to CRT is PET (accuracy above 86%) while ADC achieves the worst performances (accuracy below 80%).

Although the first two GA solutions obtained for PET have the same accuracy, they show different sensitivity and specificity values, meaning that they correctly recognize different patients. This behaviour allows for improving the classification performances using the majority voting, with whom higher accuracy is obtained due to the correct recognition of one more patient with respect to the first two solutions.

The three GA solutions obtained for T2w are equal in terms of performances, even if they uses different feature subsets (see Table I) and different SVM parameters. For this reason, no improvements were obtained using the majority voting, since the same patients were correctly classified by the 3 solutions. This means that there exists equivalent combinations of feature subset-classifier parameters.

Also for ADC it can be observed the presence of two different GA solutions (sol1 and sol2) with exactly equal performances. Conversely, the third solution has higher accuracy than the first two, even if its fitness value is higher. This solution obtained a very high sensitivity (86.7%) to the detriment of the specificity (71.4%), thus generating a marked unbalance between sensitivity and specificity, and consequently increasing the fitness value due to the increase of the penalty term. This behaviour represent the key point of the chosen fitness value, since for this kind of classification problems it is important to correctly classify both classes rather than reaching high sensitivities or specificities to the detriment of the other. When using parameters from ADC maps, the majority voting does not produce any accuracy improvement with respect to the first two solutions, but only a different distribution of patients correctly classified in the two classes.

IV. Conclusions

In this study we explore the potential role in predicting the response to CRT of texture parameters extracted from pretreatment MRI and PET images. These preliminary results, if confirmed on larger dataset, could be useful to personalize the oncological pathway for patients with rectal cancers.

Acknowledgment

This work was funded by FPRC 5 per mille 2015 Ministero della Salute.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," CA. Cancer J. Clin., vol. 65, no. 1, pp. 5–29, Jan. 2015.

[2] "NCCN Clinical Practice Guidelines in Oncology: Rectal Cancer, Version 1.2016 NCCN.org," 2016.

[3] R. Sauer et al., "Preoperative versus Postoperative Chemoradiotherapy for Rectal Cancer," N. Engl. J. Med., vol. 351, no. 17, pp. 1731–1740, Oct. 2004.

[4] M. Maas et al., "Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data," Lancet Oncol., vol. 11, no. 9, pp. 835–844, Sep. 2010.

[5] C. N. De Cecco et al., "Texture Analysis as Imaging Biomarker of Tumoral Response to Neoadjuvant Chemoradiotherapy in Rectal Cancer Patients Studied with 3-T Magnetic Resonance," Invest. Radiol., vol. 50, no. 4, pp. 239–245, Apr. 2015.

[6] D. Cusumano et al., "Fractal-based radiomic approach to predict complete pathological response after chemo-radiotherapy in rectal cancer," Radiol. Med., vol. 123, no. 4, pp. 286–295, Apr. 2018.

[7] P. Lovinfosse et al., "FDG PET/CT radiomics for predicting the outcome of locally advanced rectal cancer," Eur. J. Nucl. Med. Mol. Imaging, vol. 45, no. 3, pp. 365–375, Mar. 2018.

[8] V. Giannini, S. Rosati, C. Castagneri, L. Martincich, D. Regge, and G. Balestra, "Radiomics for pretreatment prediction of pathological response to neoadjuvant therapy using magnetic resonance imaging: Influence of feature selection," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 285–288.

[9] P. Senthil Kumar and D. Lopez, "A review on feature selection methods for high dimensional data," vol. 8, no. 2, pp. 669–672, 2016.

[10] S. Lessmann, R. Stahlbock, and S. F. Crone, "Genetic Algorithms for Support Vector Machine Model Selection," in The 2006 IEEE International Joint Conference on Neural Network Proceedings, 2006, pp. 3063–3069.

[11] A. Rojas-Dominguez, L. C. Padierna, J. M. Carpio Valadez, H. J. Puga-Soberanes, and H. J. Fraire, "Optimal Hyper-Parameter Tuning of SVM Classifiers With Application to Medical Diagnosis," IEEE Access, vol. 6, pp. 7164–7176, 2018.

[12] A.-M. Mandard et al., "Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations," Cancer, vol. 73, no. 11, pp. 2680–2686, Jun. 1994.

[13] R. J. Johnson, M. McCormick, and L. Ibanez, The ITK Software Guide Third Edition. Kitware Inc., 2013.

[14] M. Brambilla et al., "An Adaptive Thresholding Method for BTV Estimation Incorporating PET Reconstruction Parameters: A Multicenter Study of the Robustness and the Reliability," Comput. Math. Methods Med., vol. 2015, pp. 1–12, May 2015.

[15] V. Giannini, S. Rosati, D. Regge, and G. Balestra, "Specificity improvement of a CAD system for multiparametric MR prostate cancer using texture features and artificial neural networks," Health Technol. (Berl)., vol. 7, no. 1, pp. 71–80, Mar. 2017.