An innovative radiomics approach to predict response to chemotherapy of liver metastases based on CT images

An innovative radiomics approach to predict response to chemotherapy of liver metastases based on CT images

Valentina Giannini, Arianna Defeudis, Samanta Rosati, Giovanni Cappello, Simone Mazzetti, Jovana Panic, Daniele Regge, and Gabriella Balestra, *Member, IEEE*

***Abstract*—Liver metastases (mts) from colorectal cancer (CRC) can have different responses to chemotherapy in the same patient. The aim of this study is to develop and validate a machine learning algorithm to predict response of individual liver mts. 22 radiomic features (RF) were computed on pre-treatment portal CT scans following a manual segmentation of mts. RFs were extracted from 7x7 Region of Interests (ROIs) that moved across the image by step of 2 pixels. Liver mts were classified as non-responder (R-) if their largest diameter increased more than 3 mm after 3 months of treatment and responder (R+), otherwise. Features selection (FS) was performed by a genetic algorithm and classification by a Support Vector Machine (SVM) classifier. Sensitivity, specificity, negative (NPV) and positive (PPV) predictive values were evaluated for all lesions in the training and validation sets, separately. On the training set, we obtained sensitivity of 86%, specificity of 67%, PPV of 89% and NPV of 61%, while, on the validation set, we reached a sensitivity of 73%, specificity of 47%, PPV of 64% and NPV of 57%. Specificity was biased by the low number of R- lesions on the validation set. The promising results obtained in the validation dataset should be extended to a larger cohort of patient to further validate our method.**

***Clinical Relevance*— to personalize treatment of patients with metastastic colorectal cancer, based on the likelihood of response to chemotherapy of each liver metastasis.**

I. Introduction

Colorectal cancers (CRC) are highly heterogeneous and frequently harbor mutations that render them refractory to common treatments. This heterogeneity can lead to variations in individual response, different responses between the primary tumor and the metastatic lesions or among different metastatic lesions in the same patient [1].

Response to treatment is commonly evaluated using the Response Evaluation Criteria In Solid Tumours (RECIST), which measure changes in the longest axial tumor diameters after chemotherapy (nCT) of the largest metastasis (mts), up to 2 per organ [2]. However, there are some limitations in the RECIST criteria. First, this diameter might not be representative of the behavior of all mts, since the RECIST only provides a per-patient evaluation. Moreover, RECIST assesses tumor response after treatment has been completed, while it would be useful to predict the response to treatment and to adapt it based on the anticipated treatment response [2]. Radiomics-based biomarkers have shown success in predicting response to nCT for different tumor types [3]–[11], including liver mts. However, most of previous studies computed radiomics features from the whole segmented tumor, without considering the intrinsic heterogeneity of the mts [12]. In the present study, we assessed the usefulness of partitioning each liver mts into smaller regions of interest (ROI) that will be individually classified by a machine learning algorithm as responder or non-responder. Each liver mts, will be subsequently classified based on the behavior of its 7x7 ROIs.

II. Materials and Methods

A. Patients and reference standard

We retrospectively evaluated 95 patients with a newly diagnosed stage IV CRC and having at least one measurable secondary liver lesion as defined by the RECIST 1.1 Criteria (greater diameter ≥ 10 mm). All patients were enrolled in a clinical trial at the Candiolo Cancer Institute (FPO-IRCC) from 2012 to 2018 and underwent a CT exam with contrast injection within 2 weeks from the start of the first line treatment. A resident radiologist, with 5 year of experience in reading CT exams, manually segmented all liver mts with a maximum diameter ≥ 10 mm using an open-source software (ITK-snap) on the portal phase of the baseline CT exam. For each patient a maximum number of 10 mts were selected (excluding confluent/subdiaphragmatic mts, or those containing large vessels).

For each segmented mts, the radiologist measured the longest diameter at baseline and after 12 weeks of nCT. Mts were accordingly classified as non-responder (R-) if their diameter increased more than 3 mm and responder (R+) if their diameter decreased more than 3 mm or remained stable (±3 mm). This cut-off was chosen based on a preliminary study, in which we demonstrated that 95% confidence interval on the difference between means of diameters of liver mts in CT exams measured by two radiologists was 3 mm.

The study was approved by the local Ethics Committee, in accordance with the Helsinki Declaration; signed informed consent to use and analyze imaging data was obtained from all participants before entering the study.

B. Dataset construction

Radiomics features (RF) were extracted from a 7x7 ROI that moved across the image by step of 2 pixels. From each ROI fully included in the tumor mask, 22 RFs were computed on

V.G., A.D., S.M., D.R., are with with Candiolo Cancer Institute, FPO - IRCCS, Strada Provinciale 142, km 3.95, Candiolo (TO), Italy and University of Turin, Department of Surgical Science, Torino, Italy (corresponding author email: valentina.giannini@ircc.it).

J.P., and G.C. are with Candiolo Cancer Institute, FPO -IRCCS, Strada Provinciale 142, km 3.95, Candiolo (TO), Italy.

S.R and G.B. are with Polytechnic of Turin, Department of Electronics and Telecommunications, Torino, Italy.

the CT scan acquired before nCT, including: a) 4 first order features (mean, standard deviation, skewness and kurtosis), and b) 18 texture features computed from the grey level co-occurrence matrix (GLCM). GLCM were computed with the following parameters: distance=1 pixel, number of bins= 32, intensities histogram of each ROI rescaled between the minimum and maximum value of each ROI.

ROIs were classified as R+ or R- based on the classification of the lesion to whom they belong.

Patients were divided into a construction dataset and a validation set. A normalization step based on the min-max scaling was performed on both datasets using the same minimum and maximum value, i.e., those of the construction set. Then, the construction set was subsequently divided into a training and a test set using the self-organizing maps (SOM) [13]. To this scope, 2 different 5x5 SOM were constructed to cluster R+ and R- ROIs into similar subgroups. The training set was composed of 3000 R+ ROIs and 3000 R- ROIs randomly selected from each SOM obtained using R+ and R- ROIs separately and according to the following equation (1)

$$\sum_{i=1}^{N} \frac{n_i * 3000}{ROI_{tot}}, \quad (1)$$

where N is the total number of neurons, n_i is the number of elements of the i^{th} neuron, and ROI_{tot} is the total number of R+/R- ROIs in the construction set. The test set was composed of the remaining ROIs.

C. Features selection and classification

To perform feature selection and classifier optimization we used a genetic algorithm (GA) [14],[15]. To this scope, we codified each GA solution as a binary vector composed of 26 bits, 22 representing each RF ("0" in a given position identified a feature not selected whereas a "1" labeled a feature included in the final subset), and 4 bits for the optimization of the parameters of a Support Vector Machine (SVM) classifier: 2 bits were allocated for the kernel type (linear, gaussian, or polynomial of 2nd or 3rd order) and 2 bits were used for the choice of the box Constrain of the SVM (C=1,10,50 or 100).

The goodness of each solution explored by the algorithm was evaluated by a fitness function that measured the ability of the corresponding feature subset fed into the optimized SVM classifier able to classify ROI as R+ or R-. To avoid overfitting, the SVM was trained using only the ROIs belonging to the training set, and subsequently classified all ROIs of the construction set, i.e., training and test set. For each mts, we evaluated the percentage of ROI classified as R+ and we computed the Receiver Operating Characteristic (ROC) curve. Then, we selected the cut-off of the ROC curve that optimized sensitivity and specificity, i.e., the Youden Index, and we classified each mts as R+ if its percentage of R+ ROIs was higher than the cut-off point, and R- otherwise. Finally, we computed the fitness value of the current solution as:

$$fitness = 1 - \frac{sens + spec}{2} \quad (2)$$

The algorithm started with an initial population of 40 randomly generated solutions. Then, the roulette wheel selection [16] was applied to select the 50% of solutions to be used as parents of the next generation: the probability of each solution to be selected was inversely proportional to its fitness value. Starting from parents' solutions, a set of newborns' solutions was generated applying a 4-point crossover operator with probability equal to 1, and the mutation operator in which bits of the solutions were complemented with a probability equal to 0.3. Finally, parents' and newborns' solutions were pooled together, and 40 solutions were randomly extracted and used to restart the algorithms. During the GA evolution, the best current solution was stored, that was the one with the lowest fitness value in the actual population. This loop was iterated until 500 iterations were reached or no change of the best current solution occurred for 100 consecutive iterations.

To consider the random component of GA, the entire algorithm was repeated 5 times starting from the same initial population of random solutions. The solution having the lowest value of fitness was finally selected.

D. Statistical analysis

The best features subset and SVM parameters were used to develop the final SVM that was used to evaluate the performance on both the construction and the validation set. The latter was never seen by the classifier during the training phase. For each lesion the percentage of ROI classified as R+ was computed and the corresponding ROC curve was constructed. Each lesion having a percentage of R+ ROIs higher than the value represented by the Youden Index derived from the ROC curve of the construction set was considered as R+, while those having a percentage lower than the selected threshold was considered R-. The same threshold was used for both the training and the construction sets. Confusion matrices were created and the corresponding sensitivity, specificity, positive predictive values (PPV) and negative predictive values (NPV) were computed. Results of our algorithm were compared to those obtained by a stepwise backward regression model that used features extracted from the whole tumor rather than from smaller ROIs. A per-patient

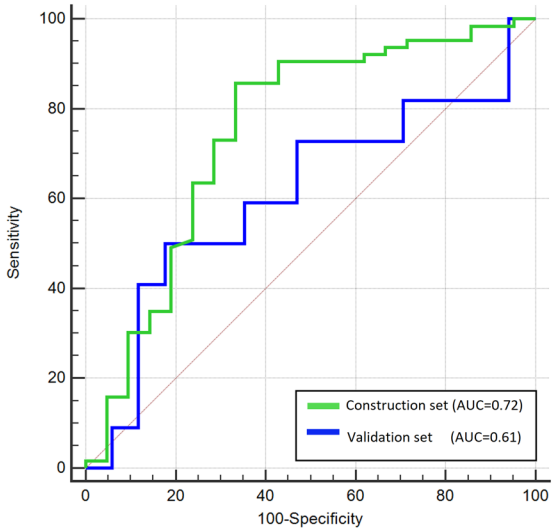

Figure 1: ROC curves for both the construction and the validation sets computed using the percentage of ROIs classified as R+ for each metastasis.

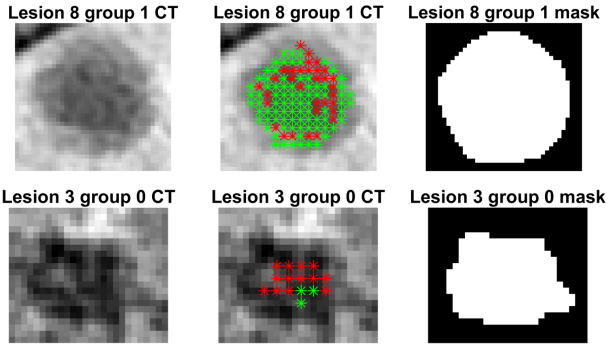

Figure 2: Examples of the classification of one slice of two different metastases: one R+ (first row) and 1 R- (second row). The first column represents the CT image, the second column shows the classification of each ROIs: in green are ROIs classified as R+ and in red ROIs classified as R- (only the center of each ROI is represented). The third column represents the manual mask used as reference standard.

analysis was also performed, in which a patient was defined either R+ or R- if the majority of his metastasis were classified as R+ or R-, respectively. Patients having an even number of mts either after classification or in the reference standard were discarded from this analysis. Statistics analysis was performed with Medcalc ® software.

III. RESULTS

A. Patients

24 patients matched the inclusion criteria. Among them 17% (4/24) had all R- lesions, 67% (16/24) had all R+ lesions and 17% (4/24) showed a mixed response (both R+ and R- lesions). A total of 123 lesions were included: 38 R- and 85 R+. The construction set was composed of 16 patients (2 with all R- mts, 2 with mixed response and 12 with all R+ mts) trying to balance patients with a few and a large number of mts. A total of 84 mts were included in the construction set (21 R- and 63 R+).

B. Per-lesion analysis

Figure 1 shows the ROC curves computed using the percentage of ROIs classified as R+. The best cut-off point to maximize sensitivity and specificity computed on the construction set was equal to 0.35. Using this cut-off on both the construction and the test sets, we obtained the per-lesion results showed in TABLE I. Our algorithm reached an accuracy of 80.9% (95%CI:70.9-88.7%) in the construction set and 61.5% (95%CI:44.6-76.6%) in the validation set. These results were higher than those obtained by the logistic regression that used features extracted from the whole tumour and that reached accuracy of 75.0% (95%CI:64.3-83.8%) in the construction set and 56.4% (95%CI:39.6-72.2%) in the validation set. Two examples of classification in the test set are shown in Figure 2.

C. Per-patient analysis

One patient from the construction set was discarded from this analysis, since he had 2 R- mts that were classified one as R- and one as R+. Among the remaining 23 patients, 8 were R- and 15 were R+. Per-patient sensitivity was 80.0%(12/15;95%CI:51.9-95.7%), specificity was 62.5%(5/8;95%CI:24.5-91.5%), PPV was 80.0%(12/15;95%CI:61.2-91.0%) and NPV was 62.5%(5/8;34.6-84.0%). 4/6 patients with only one R+ mts were correctly classified. No one had only one R- mts.

IV. DISCUSSION

In our study, we demonstrated the feasibility of developing a radiomics model able to predict the likelihood of response of individual mts in patients with mCRC, using an innovative approach. Indeed, usually radiomics features are developed from the whole tumor, either 2D slices or 3D volume, without considering the heterogeneity within the same tumor. Conversely, the presented model extracted quantitative features from smaller ROIs inside the tumor and classified each mts according to the behavior of these ROIs. Using, this approach the accuracy was higher than using a logistic regression model developed with features extracted from the whole 3D tumor. Moreover, since each ROI is characterized by a percentage of ROIs classified as R+, it would be possible to adjust the cut-off value of the ROC curve to maximize either sensitivity or specificity. From a clinical point of view, in this setting, the most important metric is sensitivity, indeed the main objective is to avoid treatment in case of R- mts, and however we should keep treating R+ lesions. However, even if the performance obtained in the validation set could be considered good and promising, they could be adjusted by choosing a cut-off that optimizes sensitivity.

The prognostic and predictive values of radiomics features in metastatic CRC have previously been exploited [9], [10],

TABLE I: RESULTS OF THE PER-LESION ANALYSIS ON NOTH THE CONSTRUCTION AND THE TEST SETS. VALUES ARE EXPRESSED IN PERCENTAGE (95% CONFIDENCE INTERVALS).

	Construction set				**Test set**			
	Sensitivity	***Specificity***	***NPV***	***PPV***	***Sensitivity***	***Specificity***	***NPV***	***PPV***
ROI SVM	85.7 (74.6-93.3)	66.7 (43.0-85.4)	60.9 (44.2-75.4)	88.5 (80.7-93.4)	72.7 (49.8-89.3)	47.1 (22.3-72.2)	57.1 (36.3-75.7)	64 (51.5-74.9)
Lostistic regression	50.0 (31.3-68.7)	88.9 (77.4-95.8)	76.2 (68.8-82.2)	71.4 (51.0-82.2)	50.0 (29.1-70.9)	66.7 (38.4-88.2)	45.4 (32.8-58.8)	70.6 (51.4-84.5)

NPV= negative predictive value; PPV= positive predictive value.

[17]–[19]. However, none of previous studies developed and validated a machine learning model able to predict the response of each mts. Most studies that analyzed single mts compared differences between R+ and R- mts on a single dataset, while we validated our results on an independent dataset. To the best of our knowledge, only Ahn et al. [20], demonstrated their findings on a validation cohort of 90 patients, showing that some texture features were independently associated with response to chemotherapy of the largest hepatic mts. Conversely, we evaluated the results of all mts. Detecting if one or more mts will respond differently than others, could help providing a tool toward personalized medicine. Indeed, if we will be able to detect whether a patient has a mixed response, we could treat the not responding mts more specifically, i.e. metastasectomy.

This study has some limitations. First, the total number of lesions in the validation dataset should be improved to better generalize our promising preliminary results. Second, our method did not reach a high specificity, especially in the validation dataset. This could be because the number of R- mts was 1/3 the number of R+ mts, meaning that the classifier had much more information about R+ mts than about R- mts.

In conclusion, we developed an innovative method to predict response to chemotherapy of liver mts. However, the promising results obtained in the validation dataset of this study should be extended to a larger cohort of patient to further validate our method.

ACKNOWLEDGMENT

This work was funded by FONDAZIONE AIRC under 5 per Mille 2018 - ID. 21091 program – P.I. Bardelli Alberto, G.L. Regge Daniele.

REFERENCES

[1] C. J. A. Punt, M. Koopman, and L. Vermeulen, “From tumour heterogeneity to advances in precision treatment of colorectal cancer,” *Nature Reviews Clinical Oncology*. 2017.

[2] R. C. J. Beckers *et al.*, “Advanced imaging to predict response to chemotherapy in colorectal liver metastases – a systematic review,” *HPB*, vol. 20, no. 2, pp. 120–127, Feb. 2018.

[3] V. Giannini, S. Mazzetti, A. Marmo, F. Montemurro, D. Regge, and L. Martincich, “A computer-aided diagnosis (CAD) scheme for pretreatment prediction of pathological response to neoadjuvant therapy using dynamic contrast-enhanced MRI texture features,” *Br. J. Radiol.*, vol. 90, no. 1077, 2017.

[4] 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 *Proceedings - International Symposium on Biomedical Imaging*, 2018, vol. 2018-April.

[5] V. Giannini *et al.*, “Predicting locally advanced rectal cancer response to neoadjuvant therapy with 18 F-FDG PET and MRI radiomics features,” *Eur. J. Nucl. Med. Mol. Imaging*, 2019.

[6] S. Trebeschi *et al.*, “Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers,” *Ann. Oncol.*, 2019.

[7] E. H. Cain, A. Saha, M. R. Harowicz, J. R. Marks, P. K. Marcom, and M. A. Mazurowski, “Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set,” *Breast Cancer Res. Treat.*, vol. 173, no. 2, pp. 455–463, Jan. 2019.

[8] K. Bera, V. Velcheti, and A. Madabhushi, “Novel Quantitative Imaging for Predicting Response to Therapy: Techniques and Clinical Applications,” *Am. Soc. Clin. Oncol. Educ. B.*, no. 38, pp. 1008–1018, May 2018.

[9] S. X. Rao *et al.*, “CT texture analysis in colorectal liver metastases: A better way than size and volume measurements to assess response to chemotherapy?,” *United Eur. Gastroenterol. J.*, 2016.

[10] R. C. J. Beckers *et al.*, “CT texture analysis in colorectal liver metastases and the surrounding liver parenchyma and its potential as an imaging biomarker of disease aggressiveness, response and survival,” *Eur. J. Radiol.*, 2018.

[11] R. Klaassen *et al.*, “Feasibility of CT radiomics to predict treatment response of individual liver metastases in esophagogastric cancer patients,” *PLoS One*, 2018.

[12] J. Liu, H. Dang, and X. W. Wang, “The significance of intertumor and intratumor heterogeneity in liver cancer,” *Experimental and Molecular Medicine*. 2018.

[13] T. Kohonen, “Self-organized formation of topologically correct feature maps,” *Biol. Cybern.*, 1982.

[14] S. Rosati, C. M. Gianfreda, G. Balestra, V. Giannini, S. Mazzetti, and D. Regge, “Radiomics to predict response to neoadjuvant chemotherapy in rectal cancer: Influence of simultaneous feature selection and classifier optimization,” in *2018 IEEE Life Sciences Conference, LSC 2018*, 2018.

[15] G. De Leonardis *et al.*, “Human Activity Recognition by Wearable Sensors : Comparison of different classifiers for real-time applications,” in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2018, pp. 1–6.

[16] W. Pedrycz, A. Sillitti, and G. Succi, “Computational intelligence: An introduction,” *Studies in Computational Intelligence*, vol. 617. pp. 13–31, 2016.

[17] M. G. Lubner *et al.*, “CT textural analysis of hepatic metastatic colorectal cancer: pre-treatment tumor heterogeneity correlates with pathology and clinical outcomes,” *Abdom. Imaging*, vol. 40, no. 7, pp. 2331–2337, Oct. 2015.

[18] A. L. Simpson *et al.*, “Computed Tomography Image Texture: A Noninvasive Prognostic Marker of Hepatic Recurrence After Hepatectomy for Metastatic Colorectal Cancer,” *Ann. Surg. Oncol.*, 2017.

[19] A. Saini *et al.*, “Radiogenomics and radiomics in liver cancers,” *Diagnostics*. 2019.

[20] S. J. Ahn, J. H. Kim, S. J. Park, and J. K. Han, “Prediction of the therapeutic response after FOLFOX and FOLFIRI treatment for patients with liver metastasis from colorectal cancer using computerized CT texture analysis,” *Eur. J. Radiol.*, 2016.