MOX-Report No. 25/2022

Perspective transfer model building via imaging-based rules extraction from retrospective cancer subtyping in Hodgkin Lymphoma

Cavinato, L; Gozzi, N.; Sollini, M; Kirienko, M; Carlo-Stella, C; Rusconi, C; Chiti, A; Ieva, F.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Perspective transfer model building via imaging-based rules extraction from retrospective cancer subtyping in Hodgkin Lymphoma

Lara Cavinato, *Student Member, IEEE*, Noemi Gozzi, *Student Member, IEEE*, Martina Sollini, Margarita Kirienko, Carmelo Carlo-Stella, Chiara Rusconi, Arturo Chiti and Francesca Ieva

***Abstract*— Image texture analysis has for decades represented a promising opportunity for cancer assessment and disease progression evaluation, evolving over time in a discipline, i.e., radiomics. However, the road for a complete translation into clinical practice is still hampered by intrinsic limitations. As purely supervised classification models fails in devising univocal imaging-based differences in tumors with different prognosis, cancer subtyping approaches would benefit from the employment of distant supervision, for instance exploiting survival/recurrence information. In this work, we transfer our previous model for Hodgkin Lymphoma subtyping to a multi-center study case. We evaluated model performance in two independent datasets coming from two hospitals, comparing and analyzing the results. Our preliminary data confirmed the instability of radiomics due to across-center lack of reproducibility, leading to meaningful results in one center and poorer performance in the other. We then learnt stratification rules from the first dataset via Random Forest and leveraged those rules to transfer the stratification policy onto the second dataset. In this way, on the one hand, we tested the stratification ability of cancer subtyping in a validation setting and, on the other hand, enriched the noisier dataset with valuable information, in a borrowing strength fashion. The transfer of the model resulted successful. Moreover, having extracted decision rules for cancer subtyping, we were able to draw up risk factors to be considered in clinics. The work shows the potentialities of the proposed pipeline to be further evaluated in larger multi-center datasets, with the goal of translating radiomics into clinical practice.**

***Index Terms*— Cancer subtyping, Clinical guidelines, Explainability, Hodgkin Lymphoma, Imaging Clustering, Radiomics, Rule extraction, Survival Clustering, Transfer model.**

I. Introduction

Cancer subtyping and patients stratification are currently the trending approaches in literature about personalized medicine and tuning of treatment pathways in oncological research. Several methodological strategies have indeed been explored, ranging from supervised, semisupervised and unsupervised learning models on both structured and unstructured data, above all genomics [1]–[3]. In [4], we proposed a recurrence-informed supervised graph clustering for stratifying Hodgkin Lymphoma patients according to their radiomic phenotype and clinical characteristics as retrieved by daily practice. We leveraged distant supervision approach [5], training the imaging-based model to learn recurrence probabilities with the scope of devising significant risk classes. Beside the prognostic reliability of detected sub-populations and the scalable performance, the main advantage of this approach regards the interpretability of the model building. In fact, groups characterization enables the radiomic features clinical interpretation in terms of both cancer severity and therapy response. We indeed provided a tool for reversing the biological interpretation paradigm of higher order radiomic variables, robust when properly trained on a real case study.

In this context, the retrospective nature of the model and the limited variability of observations coming from one single hospital prevent from translating such approach into clinical practice. In fact, the clustering of the estimated patient-to-patient graph into homogeneous sub-populations requires the availability of both medical, imaging and cancer evolution data and serves for insightful descriptive purposes. Additionally, the clinical conclusions that may derive from this pipeline are tightly dependent on the observed data set as far as it is not evaluated in other settings. It happens in fact quite frequently in literature to find inconsistencies and lack of consensus about the application of radiomics framework to clinics even with regard to the same cancer. This is due to non uniform experimental protocols, scanner variability, *ad hoc* performance evaluation and instability of results [6]–[8]. The intrinsic limitations of radiomics research along with the complexity of cancer prognosis have since long bounded the performance of classification-based models in the field. Pertinently, deducing perspective rules from distant-supervised stratification in different datasets - rather than classification strategies - might pave the way to single out the agnostic predictive power of radiomics as to translate it to clinical practice and overcome across-center reproducibility.

Here, we extend our previously proposed framework with a two-fold purpose: 1) we compare results on two different datasets coming from different hospitals to discuss concordance of findings and limitations of retrospective approach when considering diverse populations, acquisition protocols and operator-variability; 2) we deduce general knowledge from one setting and transfer a robust rule-based model to the other, in order to enrich it and validate the guidelines in a perspective way for the clinics. This work is intended as a proof of concept for employing the reliability of recurrence-specific supervised graph clustering approach in properly stratifying imaging cancer subtypes in an agnostic and perspective way.

II. Related works

Several associations between imaging characteristics and different molecular subtypes, hormone receptor status and cancer severity have been found [9] [10]. Given the hypothesis that imaging subtypes provide a proxy for established histopathological or molecular subtypes, it is widely agreed that they may help in stratifying patients. Accordingly, a number of unsupervised Image Clustering (IC) techniques has

L. Cavinato and F. Ieva are with the Department of Mathematics, Politecnico di Milano, Via Edoardo Bonardi, 9, 20133 Milan, Italy (e-mail: {lara.cavinato,francesca.ieva}@polimi.it)

N. Gozzi was with the Department of Nuclear Medicine, IRCCS Humanitas Research Hospital, Milan, Italy and is now with the Department of Health Sciences and Technology, ETH Zürich, Universitätstrasse 2, 8092 Zürich, Switzerland (e-mail: noemi.gozzi@hest.ethz.ch)

M. Sollini and A. Chiti are with the Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy and the Department of Nuclear Medicine, IRCCS Humanitas Research Hospital, Milan, Italy (e-mail: {martina.sollini,arturo.chiti}@hunimed.eu)

M. Kirienko is with the Fondazione IRCCS Istituto Nazionale dei Tumori, Via Giacomo Venezian, 1, 20133 Milan, Italy (e-mail: margarita.kirienko@istitutotumori.mi.it

C. Carlo-Stella is with the Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy and the Oncology and Hematology Unit, IRCCS Humanitas Research Hospital, Milan, Italy (e-mail: carmelo.carlostella@hunimed.eu)

C. Rusconi is with the Division of Hematology and Stem Cell Transplantation, Fondazione IRCCS Istituto Nazionale dei Tumori, Via Giacomo Venezian, 1, 20133 Milan, Italy (e-mail: chiara.rusconi@istitutotumori.mi.it

been proposed to identify imaging and clinical subtypes as to evaluate their prognostic capacity of predicting recurrence-free survival [11] [12].

The most up-to-date approaches of IC for survival risk prediction in medical imaging adopt unsupervised or semi-supervised deep learning approaches [13] [14]. In [15], an unsupervised encoder with Cox loss was developed to compress clinical, mRNA, microRNA expression data and histopathology Whole Slide Images (WSIs) and to perform clinically-relevant cancer subtyping. Similarly, [16] performed prognostic analysis of histopathological images of hepatocellular carcinoma using pre-trained CNN to extract latent features; they kept those features significant at Cox analysis and applied SVM model for stratification. Moreover, in [17], the authors proposed a pipeline consisting of learning the image latent representation from survival CNN, a dimensionality reduction step, and the clustering evaluation. Such approaches extract imaging representation features from somewhat trained CNNs and need to apply a feature selection procedure in order to either reduce the data dimensionality or to keep only survival-informative variables. On the contrary, we here may want to exploit the imaging radiomic representation as is, as to explainably assess its capability in devising relevant groups of patients. Additionally, we intend the disease-free-survival information to be part of the learning process of patient-to-patient similarities, overcoming the above-mentioned fragmented procedure. In the framework of stochastic gradient variational inference, [18] proposed a deep probabilistic approach to retrieve clusters driven by latent variables and survival information. However, difficulty in explaining radiomic variables and retrospective nature of the method still represent preventing issues for its clinical translation.

The most comparable work has been proposed in [19] where authors employed an ensemble of clustering methods and performed consensus via Harrell's C-index computation. They implemented tree-based risk model approaches to identify interactions between clinico-genomic features for colorectal from genome expressions. However, they did not externally validate the survival risk rules, preventing the translation into clinical practice.

III. Methodological pipeline

The methodological study workflow is depicted in Figure 1. Prior to models illustration, we describe the data collection and harmonization process as to provide an overview of the datasets' summary information (Section III-A). Moreover, in Section III-B we perform a classical ML-based radiomics model following current literature guidelines. Several limitations occur, laying the foundation for this work objectives. Afterwards, Section III-C opens the proposed analytical workflow and tackles the first aim of the present work. Although distant supervised cancer subtyping has been ascertained to be the paradigm shift needed for a translation of radiomics into clinical practice, across-center reproducibility problems still hold. We thus implement the model separately on two different single center datatsets and one multi-center dataset as to assess agreement of results in terms of clusters' probability to recur and radiomics discrimination power. Additionally, we recall that patient-to-patient graph is estimated by minimizing differences both in patients' radiomic description and in patients' Cox-based disease-free-survival functions (for further detail and mathematical formulation see [4], [20]). As a third point of comparison, we are interested in assessing the survival-radiomics balance in the graph estimation in each of the three models via logistic regression. Finally, Section III-D faces the second contribution of this work. As distant supervised cancer subtyping suffers from dependency on dataset specification and retrospective nature, Section III-D describes a rule-based perspective model implementation and its explainability. Performance evaluation, improvements and interpretation of results follow in Section IV.

TABLE I
HUMANITAS RESEARCH HOSPITAL (ICH) PATIENTS' CATEGORICAL CHARACTERISTICS

ICH		RESPONDERS (N=107)	NON RESPONDERS (N=21)
Stage	I	7,4%	0%
	II	53,4%	52,4%
	III	11,2%	9,5%
	IV	28%	38,1%
Sex	F	57,9%	66,6%
	M	42,1%	33,4%
B Symptoms	N	56,1%	33,4%
	Y	43,9%	66,6%
Extranodal	N	69,2%	52,4%
	Y	30,8%	47,6%
Bone	N	74,8%	85,7%
	Y	25,2%	14,3%
Radiotherapy	N	35,5%	81%
	Y	64,5	19%
iPET	DS1	76,6%	47,6%
	DS2	11,2%	9,5%
	DS3	10,3%	4,8%
	DS4	1,9%	23,8%
	DS5	0%	14,3%
PET EOT	DS1	71,9%	61,8%
	DS2	10,3%	14,3%
	DS3	9,3%	4,8%
	DS4	2,9%	4,8%
	DS5	5,6%	14,3%

TABLE II
HUMANITAS RESEARCH HOSPITAL (ICH) PATIENTS' NUMERICAL CHARACTERISTICS

ICH	RESPONDERS (N=107)		NON RESPONDERS (N=21)	
	Mean	SD	Mean	SD
Age	39.252	15.875	40.143	15.963
# Nodal lesions	6.673	4.813	6.619	6.184
# Extranoodal lesions	1.916	5.750	3.857	10.258
Dispersion of nodal lesions	0.967	0.441	1.169	0.564
Dispersion of extranodal lesions	0.827	1.652	1.882	4.383
Dispersion of all lesions	0.931	0.409	1.352	0.714
Mean volume (std)	0.028	0.520	0.455	0.963
SD volume (std)	0.529	0.833	1.270	1.702
Minimum volume (std)	-0.307	0.141	-0.326	0.084
Maximum volume (std)	1.157	2.136	2.582	3.352
Time to relapse [days]	1126.97	704.94	358.86	322.854

A. Data Collection and Harmonization

Data were collected from two hospital of Milan area, Humanitas Research Hospital (ICH - Istituto Clinico Humanitas) and the Italian National Cancer Institute (INT - Istituto Nazionale dei Tumori). The study was performed in accordance with the Declaration of Helsinki and approved by the local ethics committees. In view of the observational retrospective study design, the signature of a specific informed consent and the legal requirements of clinical trials were waived.

At Humanitas Research Hospital, 128 patients were enrolled in the study as they met inclusion criteria. They were diagnosed with Hodgkin Lymphoma and were treated and followed up at the center. Pre-treatment [18F]FDG PET/CT was available for all patients. Personal and clinical information regarding demography, therapy, follow-up and qualitative disease information was collected from Digital Medical Records per each patient. In addition, all [18]FDG-avid

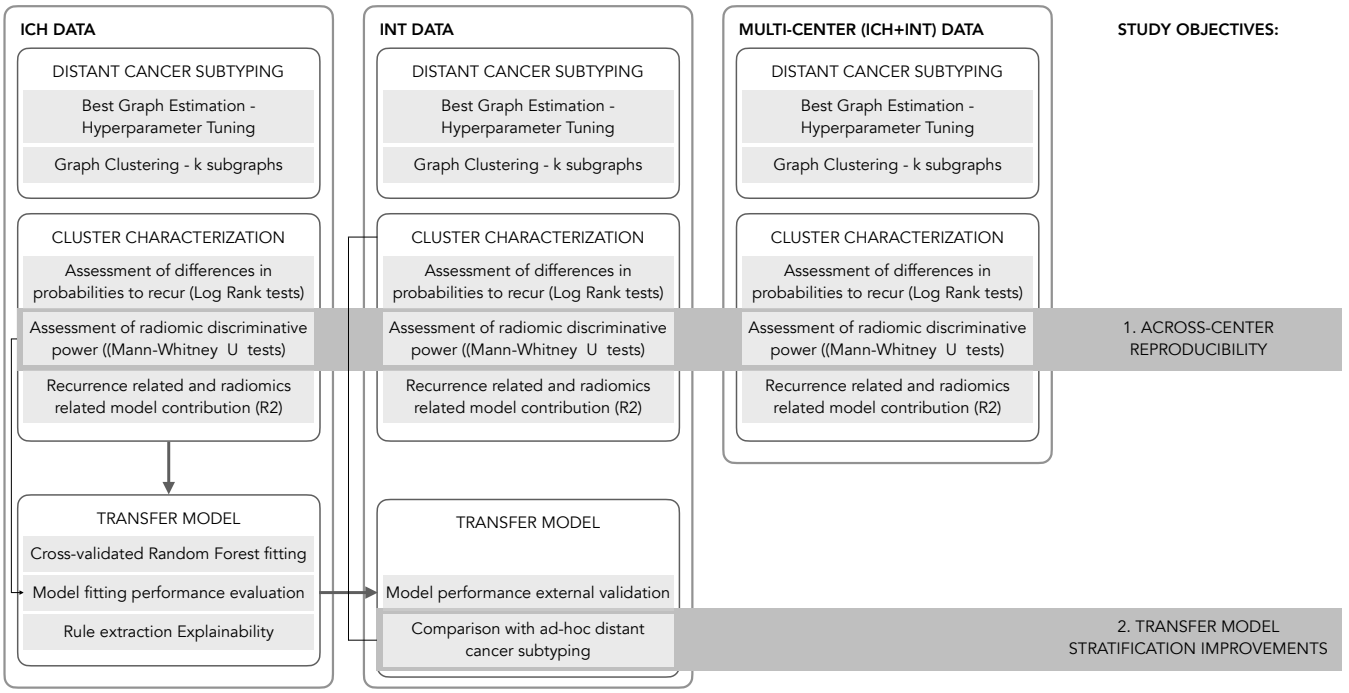

Fig. 1. Methodological study workflow: objective (1) provides a comparison between distant cancer subtyping model as trained on different datasets; objective (2) describes the knowledge transfer from the more informative setting to the less valuable one and performs improvements assessment.

lesions bigger than 64 voxels were located and semi-automatically segmented by expert nuclear medicine physicians (M.S.). From regions of interest, radiomic description was computed with LifeX software (www.lifexsoft.org, [21]), consisting of 45 radiomic features including conventional, first, second and higher order statistics. A total of 1340 lesions was collected and quantitatively assessed. Survival and recurrence free survival times were also registered along with censoring information. Chemotherapy starting dates, dates of *ad interim* PET (iPET) and End Of Treatment (EOT) PET were collected as to extract temporal information of therapy pathways. Radiotherapy date was also made available when performed. For what treatment efficacy and recurrence/relapse are concerned, response to therapy was monitored over time, with checkpoints at the end of first line of chemotherapy (iPET), at the end of all chemotherapy cycles (EOT PET) and at the time of the last follow up (LFU). Patients were defined as responders and non responders which included patients who progressed during or early after first-line treatment (refractory) and patients who have eventually relapsed within the observation period (recurrent/relapsing). Additionally, survival information at the time of the last follow up was collected, yet only one patient experienced the event. Patients information is made available in Table I for categorical variables and II for numerical variables. In Appendix I, descriptive statistics of disease-free-survival times according to clinical-informed stratification is explored.

The same criteria were used to enroll patients and analyse images at National Cancer Institute. Most of patients were diagnosed at the center and information about those patients for whom this was not the case was retrieved and properly annotated. Briefly, [18F]FDG PET/CT images of 76 Hodgkin Lymphoma patients (794 lesions) were analysed by expert nuclear medicine physicians (M.K.) using LifeX software. The radiomic descriptions (45 radiomic feature) were obtained for all regions of interest. Clinical data about demographics, chemotherapy cycles length, radiotherapy treatment and follow-up was collected. Both iPET and EOT PET were defined as positive in presence of an area of [18F]FDG uptake higher than background as defined by the Deauville score (DS) DS4 and DS5. DS3 or lower was consistent with a negative exam [22]. Information about the specific DS of each patient was available only for the ICH dataset. At INT, no distinction was made if non responding patients at LFU were refractory or relapsing, however time to recurrence allowed to retrieve such information when compared to chemotherapy cycles duration. No survival information was collected. Patients information is made available in Table III for categorical variables and IV for numerical variables. In Appendix I, descriptive statistics of disease-free-survival times according to clinically-informed stratification is explored.

TABLE III
NATIONAL CANCER INSTITUTE (INT) PATIENTS' CATEGORICAL CHARACTERISTICS

INT		RESPONDERS (N=59)	NON RESPONDERS (N=17)
Stage	1	1 ,8%	0%
	2	52,4%	23,5%
	3	10,2%	5,9%
	4	35,6%	70,6%
Sex	F	57,6%	47%
	M	42,4%	53%
B Symptoms	N	59,3%	23,5%
	Y	40,7%	76,5%
Extranodal	N	66,1%	41,1%
	Y	44,9%	58,9%
Bone	N	74,6%	76,5%
	Y	25,4%	23,7%
Radiotherapy	N	44,9%	82,3%
	Y	66,1%	17,7%
iPET	Negative	93,2%	47%
	Positive	6,8%	53%
PET EOT	Negative	100%	0%
	Positive	0%	100%

TABLE IV
NATIONAL CANCER INSTITUTE (INT) PATIENTS' NUMERICAL CHARACTERISTICS

INT	RESPONDERS (N=59)		NON RESPONDERS (N=17)	
	Mean	SD	Mean	SD
Age	36.478	13.915	42.867	17.868
# Nodal lesions	7.271	5.499	9.706	6.362
# Extranoodal lesions	2.288	5.789	3.706	7.355
Dispersion of nodal lesions	0.900	0.463	1.405	2.049
Dispersion of extranodal lesions	0.747	1.636	1.938	3.425
Dispersion of all lesions	0.900	0.443	1.406	1.886
Mean volume (std)	0.075	0.542	0.176	0.784
SD volume (std)	0.625	1.030	0.931	1.394
Minimum volume (std)	-0.331	0.090	-0.358	0.087
Maximum volume (std)	1.312	2.453	2.304	3.368
Time to relapse [days]	1105.72	546.490	257.59	167.17

As our aim was to assess across-center imaging variability and transfer the cancer subtyping policy from one to the other, an harmonization step was required. First, all clinical and personal information was processed with a strategy of compliance to the less rich dataset. That is, response to treatment and cancer progression were flagged by a dichotomous variable, survival information was neglected and times to events were computed. Additionally, some variables were added or transformed as to enrich the disease description and obtain a single vector patient representation as described in [4]. In fact, number of total lesions, number of nodal and extranodal lesions and dispersions of all, nodal and extranodal lesions within a patient as a proxy of tumor heterogeneity were computed and analyzed. In addition, radiomic features were averaged patient-wise such that every patient was described by its lesions' mean radiomic profile/phenotype. After harmonization, Humanitas Research Hospital (ICH) dataset contained 128 patients described by 61 variables and National Cancer Institute (INT) dataset held 76 patients described by the same 61 variables.

B. ML-based model limitations

Classical machine learning algorithms have been evaluated for radiomics-based recurrence prediction. As to remove redundancy and collinearity among variables, correlation between clinical features was performed and brought to removal of stage ($pcc > 80\%$ with extranodal and bone disease statuses) while correlation between radiomic features was assessed and lead to keep 15/45 features. The selection was made according to domain knowledge, robustness with respect to the tumor segmentation method [23] and matrix frequency. No relevant correlation between clinical and radiomic features was found. Variance Inflation Factor was implemented for additional correlation removal, leading to a set of 14 features. In order to capture lesions' imaging variability within multi-lesion patients, statistic moments (i.e., mean, min, max, std) were computed and used to build the vector-based patient representation. A total of 68 variables were considered.

Prior to be fed into the model for relapsing vs non- relapsing patients' classification, the predictive power of features was assessed with both univariate and multivariate feature selection methods. ANOVA, Pearson Correlation, Spearman Correlation, Mutual Information and Univariate Logit were considered by assessing the relationship between each variable and target, i.e., relapse. A reasonable agreement was found among different techniques and 28 variables were robust in at least 4/5 methods. Being robust throughout univariate selection, such features were additionally screened with multivariate selection. Forward Selection, Backward Selection, Bidirectional Selection, Recursive Feature Elimination, Ridge Regression, LASSO and Elastic Net were considered. Wrapper methods and penalized logistics separately showed agreements in selecting features, yet quite a discordance between each other: 11 of them resulted robust to 5/7 techniques. Interestingly, these included volume and both lower and higher order radiomics features. The flow of the feature selection process and a complete list of the selected features can be found in Appendix II.

On selected features, we trained and tuned different models for imbalanced data based on CARTs. The best model resulted to be an easy ensemble classifier with Random Forest base estimator with 50 trees and replacement. Results on cross-validation on ICH dataset were good, reaching 0.76 ± 0.035 accuracy with 0.67 ± 0.179 of sensitivity, leading to 0.72 ± 0.082 of AUC. Due to across-center imaging variability, performance in the external INT test set dropped to 0.62 accuracy with 0.41 sensitivity.

This preliminary yet rigorous assessment brings up the well-known limitation of radiomic framework as it is currently presented and exploited in clinical literature [6]. First, high dimensional data call for massive feature selection approaches which mostly require, as well as classification models, several and balanced data. Poor repeatability and reproducibility of the results are indeed due to datasets' intrinsic limitations, imbalance and scarsity of data. Unfortunately, such issues could hardly be overcome: in fact, the number of samples is limited to the number of cases, few when dealing with rare diseases like Hodgkin Lymphoma; number of minority class observations is limited to the number of patients who do not heal and eventually recur, which is a small percentage over the total patients; moreover, variability in the reconstruction parameters, acquisition settings and scanners is given by the lack of standardization of the *status quo* clinical practice.

For these reasons, the current paradigm of radiomics should shift towards more complex and unsupervised strategies for patient stratification and imaging based-risk factor identification. A priori knowledge, temporal information, data structural similarities and more informative representation should be exploited for pivoting a more effective radiomic research. Pertinently, distant supervised cancer subtyping allows for partially consider all these factors, overcoming radiomics intrinsic limitation.

C. Across-center reproducibility of Cancer Subtyping Models

1) S2GC on ICH data: Following the pipeline in [4], upon datasets processing and harmonization, the cancer subtyping model (S2GC [20]) was applied on ICH dataset in order to reproduce the results. Loss function was optimized for estimating the patient-to-patient similarity graph, prompt to be clustered into risk classes of patients according to spectral clustering algorithm. Clustering was performed on eigenvectors of the graph laplacian matrix, normalized with symmetric method [24]. As pointed out in our previous paper, such procedure brings to the slicing of patients' diseases into cancer imaging phenotypes with different prognosis, exploiting all the variability of the data. Hyper-parameters were set as in [4] and two groups were identified and tested for significance at Kaplan-Meier estimates (p-value of the Log Rank test $<< 0.01$). According to Hazard Ratio (0.2176, IC: 0.1202-0.3937), group 1 was characterized by a better prognosis with almost no recurrence experienced, while group 2 contained patients with poorer prognosis, who were instead more likely to recur. Clinical and radiomic features were used for interpreting the clusters' risks, emerging as significant in several cases. To compare average imaging description of one cluster with respect to the other, two sided non parametric tests on averages (Mann-Whitney U tests) were used and p-values lower than the threshold of 0.05 were considered significant.

2) S2GC on INT data: For comparison and qualitative assessment purposes, the very same procedure was applied to and optimized also in INT dataset. Similarly to ICH dataset, two significant groups

(p-value of the Log Rank test $<< 0.01$) were obtained and tested as significantly different. In particular, as emerged from Hazard Ratio (0.0627, IC: 0.0321-0.1223), group 1 featured those patients with a better prognosis with no events of recurrence, while group 2 was populated by patients with poorer prognosis and a higher chance of recurrence. Tests were again performed to evaluate differences between the two populations and to characterize INT risk stratification policy. In fact, clustering interpretation was hereby interpreted as rules for patient slicing, to be compared with the criteria built on ICH dataset for repeatability purposes.

3) S2GC on ICH+INT data: As an additional level of analysis, the two datasets have been merged and survival clustering pipeline was run on the multi-center dataset as to evaluate the significance of variables irrespectively to the provenience of observations. In fact, we anticipate that the ICH model brought to high stratification power of radiomic features while INT model did not. For such reason we have investigated whether such power holds when noising the ICH one-center data with data coming from different populations, i.e., INT center. Similarly to ICH and INT cases, the survival clustering procedure on the multi-center dataset resulted successful and the Kaplan-Meier curves of patients belonging to the two obtained risk classes were tested different (p-value of the Log Rank test $<< 0.01$). From Hazard Ratio assessment (0.1117, IC: 0.0732-0.1705), it was clear how group 1 was again related to non recurrent patients and group 2 to recurrent and bad prognosis cases. Non parametric tests on radiomic variables were performed to compare the two populations, resulting significant - as we will see in Section IV-A - in almost all cases.

The three models have brought to significant and comparable results and each have led to the stratification of the populations into two - severe and mild - risk classes. Results describing shared features between the study cases could thus be discussed, assessing the significance of radiomics discrimination power.

D. Transfer Model Building and Evaluation

Beside *ad hoc* model comparison, generalization ability of S2GC stratification method was assessed with interpretable predictive model. ICH decision rules have been extracted from III-C.1 stratification and a transfer model was implemented, exploiting the borrowing strength strategy. Specifically, a Random Forest of 100 trees, cross-validated with Out-of-Bag prediction, with a minimum leaf size of 5 and empirical prior was used for rule extraction. The model was trained on ICH dataset, where those features which resulted significant at univariate testing (Mann-Whitney U tests) on stratified populations have been considered. Performance are discussed in Section IV-B. Upon model training, it was applied on INT dataset and performance has been evaluated in terms of prognosis stratification power. A new set of labels resulted from rule transferring, leading to grouping patients in two risk classes, ideally with poorer and milder prognosis respectively. Kaplan-Meier estimates of survival curves of the groups were compared with the Log Rank test and stratification power of radiomics was assessed in terms of non parametric univariate test significance (if Mann-Whitney U test p-value < 0.05). The resulting stratification was compared with the *ad hoc* INT cancer subtyping model described in Section III-C.2 and improvements were evaluated (Section IV-C).

Finally, the interpretation of rule extraction step was performed according to the explainability analysis of the Random Forest model as to highlight the role of specific clinical and radiomic variables as risk factors. In addition to feature importance assessment, common rule set was in fact estimated from the Random Forest model according to [25] [26]. Every rule of every tree split was annotated and kept when common enough in the forest to be relevant to the model; similar rules were than post-treated and aggregated as to define a stable, interpretable and unique set of elementary rules pivoting the model decisions. The algorithm was first trained in a cross-validation fashion in order to estimate the optimal hyperparameter $p0$ used to select the number of relevant rules to extract. Specifically, $p0$ represents the proportion of forest's trees in which a selected rule must appear in order to be defined as relevant, and is estimated according to a performance-stability trade-off. The algorithm was then run on the trained Random Forest to retrieve the k most relevant decision rules. Section IV-D details the findings.

IV. Results

A. Agreement analysis between S2GC models

The significance of clinical and radiomic variables was first compared in the two single center datasets (models III-C.1 and III-C.2), leading to some observations. First, ICH stratification was appreciably driven by radiomic tumor characterization, as the majority of features resulted significant at the slicing. On contrary, although INT patients stratification was successfully carried out, very few variables emerged as significant at testing. As a matter of facts, the $pseudo - R^2$ statistics of the logistic regression between radiomic variables (independent variables) and stratification labels (dependent variable, namely group 1 and group 2) was 65% (p-value $<< 0.01$) in the ICH model and 46% (p-value = 0.045) in the INT model. Indeed, the $pseudo-R^2$ is the ratio between the log-likelihood of the intercept model (as a total sum of squares) and the log-likelihood of the full model (as the sum of squared errors), suggesting the level of improvement over the intercept model offered by the full model [27]. In other words, the $pseudo-R^2$ metrics is here considered as a proxy for the variability of the estimated graph explained by the differences in the radiomic features with respect to the total estimation model. The low percentage of the contribution of radiomics in the model suggests how stratification on INT data has been mainly dragged by minimization of recurrence time periods in the Cox related part of the loss function, whereas radiomic variables played a limited role. Of course, this might not be acceptable when tackling the problem from a predictive point of view, aiming to extract general information that could inform clinical practice in a perspective way. We remind that INT dataset included PET scans performed at INT and other centers, with different acquisition procedures. Consequently, INT situation represents the perfect case study for transfer knowledge, as stratification criteria could be guided and enriched with information coming from ICH rules, exploiting the borrowing strength strategy [28]. Additionally, comparing radiomic variable significance across models enables to acknowledge variables which are agnostic with respect to imaging acquisition settings and texture extraction parameters. In the multi-center model (III-C.3), these differences in the radiomic predictive power appeared to settle, suggesting the strength of higher variability. In fact, the $pseudo-R^2$ statistics of the logistic regression between radiomic variables (independent variables) and stratification labels (dependent variable) was 70% (p-value $<< 0.01$), testifying the preponderant role of radiomics.

A complete list of Mann-Whitney U tests p-values for each radiomic feature in the three datasets can be found in Appendix III. Few variables (5/61) appeared to be significant in both one-center datasets, holding their significance when applying the whole framework in the multi-center dataset of patients. This was expected as intrinsic limitations of radiomics often leads to discordant results and lack of literature consensus [29].

The majority of features that resulted significant in the ICH (27/45) model but not in the INT one were strong enough to remain significant in the third multi-center model. Such variables were equally

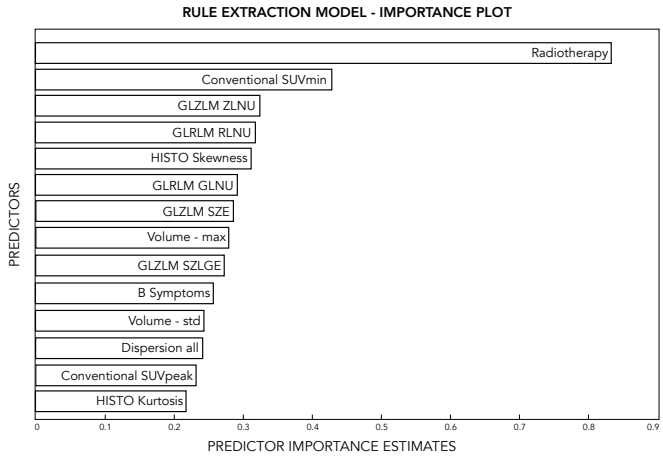

Fig. 2. Random Forest model explainability analysis: feature importance plot with descending order.

found among first order, second order and higher order features, as well as qualitative disease information like volume and counts of nodal and extranodal lesions. The remaining features significant in the ICH model (15/45) did not hold significance in the multi-center dataset. Of course, variables that were not significant in ICH dataset nor in INT dataset remained not significant in the multi-center case (14/61). Only one variable - the dispersion of nodal lesions - resulted significant in INT model and not in ICH model yet was not strong enough to remain significant in the multi-center model.

B. Rule Extraction Performance

The cross-validated Random Forest was successfully trained on ICH dataset. As expected from previous consideration, the model was able to capture all the variability entailed in the data and exploit such information to classify observations into predefined risk classes. In fact, radiomics played a fundamental role in the stratification algorithm of S2GC, showing again predictive power under the classification perspective. Beside accuracy, which resulted to be 97.66%, other more relevant performance evaluation criteria was found to be widely satisfactory: sensitivity and specificity were respectively 98.82% and 95.34%, while F-measure was 98.24%. We remark that such performance values reveals a model highly overfitting the training data, with the clear aim of obtaining an interpretable and predictive mirror of the retrospective cancer subtyping model. Contrary to common ML best practice, we here want to discard generability to appreciate the peculiar intrinsic structure of the model we are mimicking. Also the Log Rank test on the Kaplan-Meier curves was significant (p-value of the Log Rank test < 0.01), such that the splitting into two groups of different prognosis phenotypes devised a group with fast-relapsing patients and one with long- or non-relapsing patients (HR: 0.2230, IC: 0.1227-0.4054). Being fitting and modeling robust enough to be intended as a subtyping Rule Extractor, the Random Forest model is worth to be applied and tested in the INT scenario, and further analyzed in order to explainably define such rules.

C. Transferred Stratification Performance

The stratification rules were thus transferred to INT dataset applying the Random Forest model to it. The resulting stratification brought out two risk classes of patients having characteristics of imaging phenotypes similar to ICH risk groups, from which we borrowed the information about radiomic variability and cutoffs. The stratification based on information transfer was quite similar to the one obtained with the *ad hoc* cancer subtyping algorithm as seen in Section III-C.2. In fact, the concordance index between the two reached 0.6. We remind that the S2GC approach on INT did lead to a significant patients' prognosis stratification, but no quantitative radiomic information could be elected as relevant risk factors on which to eventually rely the prognosis, because of non significance of tests. Only qualitative disease information and primarily time to event information accounted for the majority of the model stratification power. The fact that the purely radiomics-based classification model showed concordance and coherence with the ground truth strengthens the reliability of the transferring. Generally speaking, models involving radiomics, as any other data analysis model, benefit from high dimensional datasets. However, this numerosity often comes at the expense of informative variability from which we can extract valuable knowledge. Particular attention needs in fact to be payed when trading off between data dimentionality and consistent inclusion criteria. In the context of our analysis, INT dataset did contain information, although it was masked by radiomic well-known constrains and could not be appreciated. ICH-informed model behaved as a magnifying glass and enabled the extraction of radiomic-based knowledge from noisy data.

The two groups resulting from the transferred stratification model were compared in terms of Kaplan-Meier estimates of survival curves, leading to a significant discrimination between the better prognosis and the poorer prognosis class (p-value = 0.0105), as highlighted by the Hazard Ratio (0.2496 (0.1240-0.5026)) and displayed in Figure 3. Irrespectively to the *ad hoc* procedure that brought to the significance of only 6 variables (none of whose was purely related to radiomics), the transfer stratification model lead to 38 significant variables, including conventional, first- and second-order texture statistics. It follows coherently that the $pseudo - R^2$ statistics of the logistic regression between radiomic variables (independent variables) and stratification labels (dependent variable) resulted to be 63% (p-value < 0.01), attesting the role played by radiomics.

D. Rule Extraction Explainability

As to explain and interpret the rules extracted by the transfer model, we first may want to look at feature importance plot. As displayed in Figure 2, the ranking of the Random Forest predictors has been computed basing on their importance in the model and the top relevant ones were plotted. We selected the first variables which presented significantly higher absolute importance, leading to a total of 18 considered features. The majority of them (13/18) were found among the variables that showed to be significant at S2GC model in both one-center datasets or significant in ICH dataset and holding in multi-center dataset. However, few variables (5/18) emerged even if they did not hold significance from ICH to multi-center datasets. Of interest, the most important factor that dragged the classification was radiotherapy, followed by conventional and second order radiomic features. Volume and dispersion of lesions were relevant as well. Intuitively, the most relevant features were the ones driving the decisions throughout the trees of the forests.

In line with the importance plot, different clinical and radiomic features were leveraged by common Random Forest tree splits. The list of such rules can be assessed in Figure 4. Among the extracted rules, radiotherapy was the first key factor to be considered when determining the recurrence outcome of patients: absence of radiotherapy treatment lead to the higher probability of incurring into tumor relapsing ($Pr = 0.909$). Although such finding was already known in clinical practice - in fact, more severe patients are often

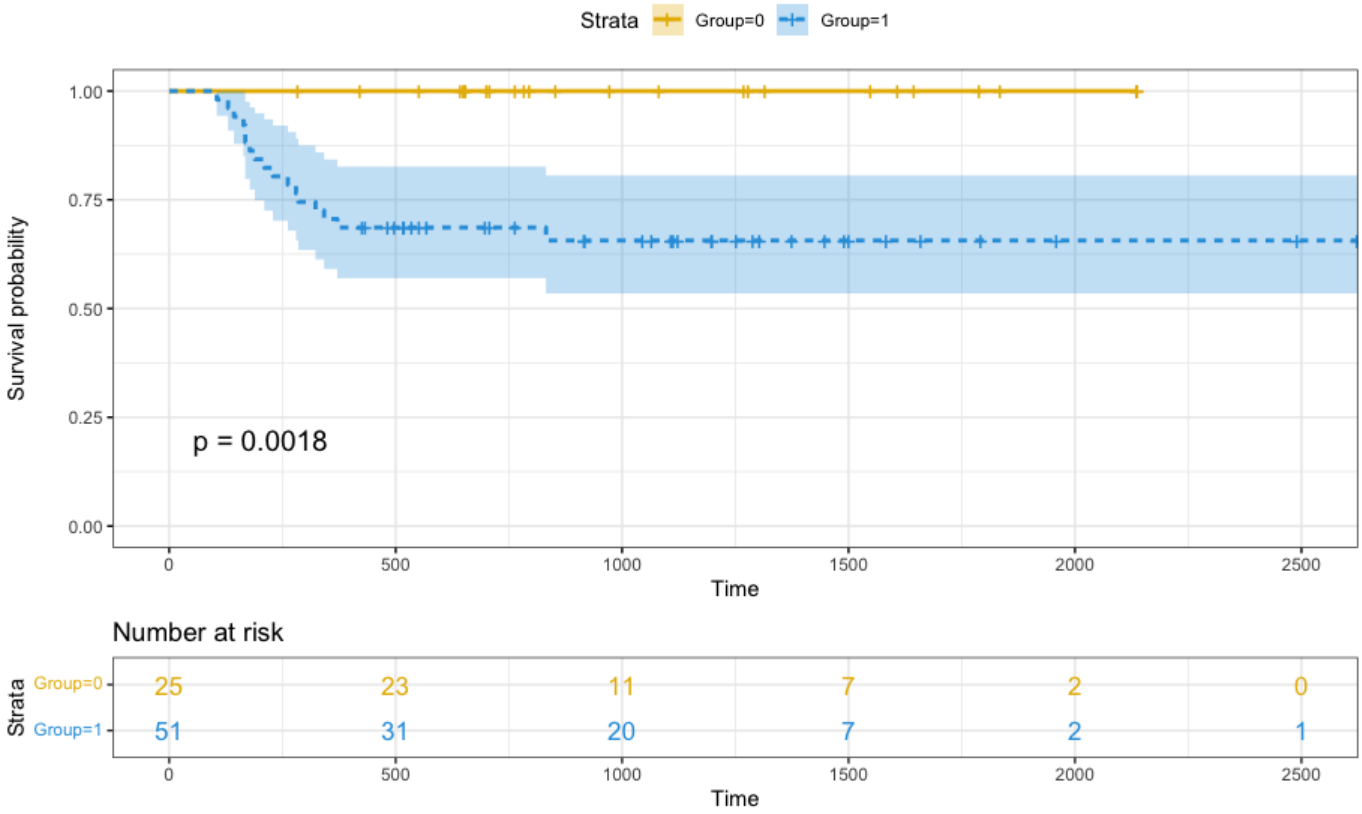

Fig. 3. National Cancer Institute patients' stratification into risk classes with different survival estimates according to Random Forest model.

treated with both chemotherapy and radiotherapy - it is worth to notice that it was frequently observed together with dispersion and radiomic variables. The absence of radiotherapy and values of lesions dispersion higher than 67% of the sample ($z = -0.431$) brought to higher probability, i.e., certainty, of recurrence ($Pr = 0.979$). Moreover, when considered together with values of *GLRLM Run Length Non-Uniformity* higher than 77% of the sample ($z = -0.744$), the probability of recurrence of patients without radiotherapy rise to $Pr = 0.959$. It follows that clinical information about patients demography and therapeutic pathways are solid markers for patients disease progression, yet their power is deeply increased when taking the imaging and heterogeneity information into account as well.

The other relevant rules also testify the same point. In fact, other few clinical variables were fundamental for prognosis, i.e., the presence of B symptoms, related to higher probability of recurrence ($Pr = 0.869$), the volume of the patient's smallest lesion, leading to poorer outcomes ($Pr = 0.81$) when lower than 81% of the sample ($z = -0.0889$) and the volume of the patient's biggest lesion, leading to poorer outcomes ($Pr = 0.779$) when higher of the 66% of the sample ($z = -0.439$). Indeed, huge differences in lesions' volume within the same patients are proxies of intra-patient heterogeneity. All the other decision splits account for radiomic descriptors values, in particular for lesions' heterogeneity measures. *Conventional SUV Peak*, *GLZLM Zone Length Non-Uniformity*, *GLCM Correlation*, *GLZLM Long Zone High Gray-level Emphasis* and *GLRLM Gray Level Non-Uniformity* lead to worse tumor progression outcomes when assume high values with respect to the population distribution ($Pr = 0.779$ with $z = -1.05$, $Pr = 0.792$ with $z = -0.326$, $Pr = 0.779$ with $z = -0.482$, $Pr = 0.897$ with $z = -0.0908$, $Pr = 0.843$ with $z = -0.158$ respectively). Interestingly, higher probabilities were found in correspondence of rules exploiting higher order radiomic features, supporting the prognostic value of radiomics.

V. Discussion

Survival- or recurrence-specific supervised graph clustering has found evidence of being a reliable tool for patients stratification and tumor evolution prediction. Although it exploits retrospective cancer data to perform insightful cancer subtyping, it could help in defining decision rules and imaging-based guidelines in a perspective sense, as far as imaging characteristics are used to build the model. In this work, we intended to address this matter and provide a proof of concept of the potentialities of such approach. We built the retrospective model on a very rich and informative dataset of one hospital and transferred the deduced knowledge to a smaller and noisy dataset of a different hospital. Of note, in previous literature such transferring has been shown often unreliable and unstable due to the limitation of radiomics, which is known to be dependent on operators, i.e., the segmentation of regions of interest, acquisition settings, scanner characteristics and other independent factors [30] [31]. Nevertheless, our results demonstrated the possibility of singling out agnostic features that remain robust throughout different centers and radiomics inconsistencies. Interestingly, variables that showed a significance in all three datasets were mainly related to clinical and qualitative information about the disease, i.e., the stage, the B symptoms, the extranodal disease status, radiotherapy and dispersion of lesions. Such disease information thus appeared agnostic with respect to the center of provenience and could be acknowledged as robust in a perspective study. We recall that dispersion is the extent to which the distribution of lesions within a patient is stretched or squeezed. Here, this variability is computed as the patient-wise spreading of lesions in the radiomics space, i.e., the average distance between radiomic variables of peer lesions. Of note, the dispersion

COMMON RULES SET

RADIOTHERAPY AND RADIOMICS

if Radiotherapy = 0 then Pr = 0.909 (n=55)
if Radiotherapy = 1 then Pr = 0.479 (n=73)

if Radiotherapy = 0 & Dispersion all >= -0.431* then Pr = 0.979 (n=47)
if Radiotherapy = 1 & Dispersion all < -0.431* then Pr = 0.481 (n=81)

if Radiotherapy = 0 & GLRLM RLNU >= -0.744* then Pr = 0.959 (n=49)
if Radiotherapy = 1 & GLRLM RLNU < -0.744* then Pr = 0.481 (n=79)

CLINICAL VARIABLES

if B Symptoms = 0 then Pr = 0.478 (n=67)
if B Symptoms = 1 then Pr = 0.869 (n=61)

if Volume - min < -0.0889* then Pr = 0.81 (n=63)
if Volume - min >= -0.0889* then Pr = 0.523 (n=65)

if Volume - max < -0.439* then Pr = 0.49 (n=51)
if Volume - max >= -0.439* then Pr = 0.779 (n=77)

RADIOMICS

if GLZLM ZLNU < -0.326* then Pr = 0.471 (n=51)
if GLZLM ZLNU >= -0.326* then Pr = 0.792 (n=77)

if GLCM Correlation < -0.482* then Pr = 0.49 (n=51)
if GLCM Correlation >= -0.482* then Pr = 0.779 (n=77)

if GLZLM LZHGE < -0.0908* then Pr = 0.562 (n=89)
if GLZLM LZHGE >= -0.0908* then Pr = 0.897 (n=39)

if CONVENTIONAL SUVpeak < -1.05* then Pr = 0.49 (n=51)
if CONVENTIONAL SUVpeak >= -1.05* then Pr = 0.779 (n=77)

if GLRLM GLNU < -0.158* then Pr = 0.545 (n=77)
if GLRLM GLNU >= -0.158* then Pr = 0.843 (n=51)

Fig. 4. Verbose common rules set divided into radiotherapy informed by radiomics features, clinical features and stand alone radiomic features. (*) thresholds refer to z-standardized variable values.

was robust and agnostic with respect to the acquisition settings as it aggregates the imaging information in a standardized way [32] [33].

The strategy of borrowing the strength and knowledge transfer from one set of data to a less informative one has been successful in devising groups of at-different-risk patients with significantly different time-to-recurrence curves. As Hodgkin Lymphoma, like several other tumor diseases, is a rare condition, this approach could support decisions in those cases where only few observations are available and the aggregated information coming from other sources may aid the evaluation/assessment.

Of interest, among risk factors, both clinical and imaging variables have emerged as relevant. Indeed, rules have, on one hand, confirmed the prognostic power of known qualitative factors as tumor volume, radiotherapy and the presence of B symptoms; on the other hand, tumor heterogeneity measures have appeared to consistently aid the recurrence probability estimation. In fact, a number of radiomic features - conventional, first-, second- and higher order features - significantly rose the precision of clinical variables in estimating the probability of relapsing. Several of them were exploited in the decision making, however we showed and discussed the more common ones among the trees splits of the Random Forest.

The mean intensity value in the higher intensity 1 mL volume sphere, i.e., *Conventional SUV Peak*, was found higher in recurrent patients. The maximum uptake of PET radiotracer is indeed a sign of more aggressive diseases [34], [35]. The linear dependency of grey-levels in Grey Level Co-occurrence Matrix, i.e., *GLCM Correlation*, appeared slightly higher in worse prognosis group, meaning that lower-order heterogeneity measures do contribute to define the intra-lesion variability of tumor phenotypes [36]. Such variability comes often with and is underlined by higher-order heterogeneity assessments which strengthen the characterization of lesions. Gray-Level Non-Uniformity for runs, i.e., *GLRLM GLNU*, and Run Length Non-Uniformity, i.e., *GLRLM RLNU*, represent the non-uniformity of the grey-levels or the length of the homogeneous runs. Higher values lead to higher probability of recurrence since, as expected, intrinsic variability of lesions' uptake is related to heterogeneous thus severer tumors [37]. Analogously, the Gray-Level Zone Length Non-Uniformity, that is the non-uniformity of the length of the homogeneous zones, was again found higher in recurrent patients. The less uniform the homogeneous runs, the less uniform also the homogeneous zone, which are the 3D extensions of runs [21]. This means that more aggressive lesions, even if heterogeneous, do not exhibit all grey levels with equal proportions, rather they show some gray values preferably than others. Specifically, high values of grey levels were found more often in worse lesions and cancers with respect to low values. In fact, Long-Zone High Gray-level Emphasis represents the distribution of the long homogeneous zones with high grey-levels and contributed positively in defining relapsing patients among decision rules. In line with this findings, recent literature has sharpened its focus towards repeatability and reproducibility of radiomics in multi-center studies [38]. Although sensitive to all above-mentioned acquisition criteria, far from a few lower- and higher-order radiomic features appeared to be robust and agnostic.

Beside quantitative radiomics-based measures of intra-lesions variability, intra-patient variability assessment was decisive as well. The patient-wise dispersion of lesions' radiomic profiles increased the reliability of recurrence probability estimation, being affected by its value. Multi-lesions tumors are known to exhibit heterogeneity over lesions. They entail biological, both genetic and epigenetic, aberration that make sub-populations of cells evolve and acquire mutations, conferring resistance to specific therapies and leading to treatments' inefficacy and relapsing [39] [40]. A tumor evolves changing its molecular characterization over time while spreads throughout the patient's body such that metastases have a molecular fingerprint different from primary tumor. Although this heterogeneity has only been explored and evaluated in terms of biological characterization, texture-based comparison among lesions may represent the non-invasive and easy-to-retrieve counterpart of the same information. In this sense, lesions-wise radiomic features' dispersion could be a reliable index for reassuming intra-patient heterogeneity [32], [33].

Although promising, the proposed approach positions itself as a proof of concept and should tackle some current limitations that may prevent the immediate translation into a perspective clinical study. It would be desirable to collect data from many different centers/hospital to harmonize and integrate available information and to build more informative and agnostic decision models. As highly anonymous and aggregated data are needed, this step might not represent a bottleneck from the privacy point of view, which is often an issue when sharing medical data. Collected datasets could thus update the current decision rules with an online-updating framework as new observations become gradually available, in a federated fashion. Larger graphs could be estimated from a higher number of patients and more robust rules could be derived from the procedure. In this direction, an additional point of improvement could be acknowledged: a grouping strategy would be desirable to exploit the hierarchical nature of a multi-center dataset as to automatically consider the nesting levels in the graph estimation phase of the algorithm. Accordingly, loss function could be revised and the grouping term could be appended and minimized.

Ultimately, alternative patient representation implementation could

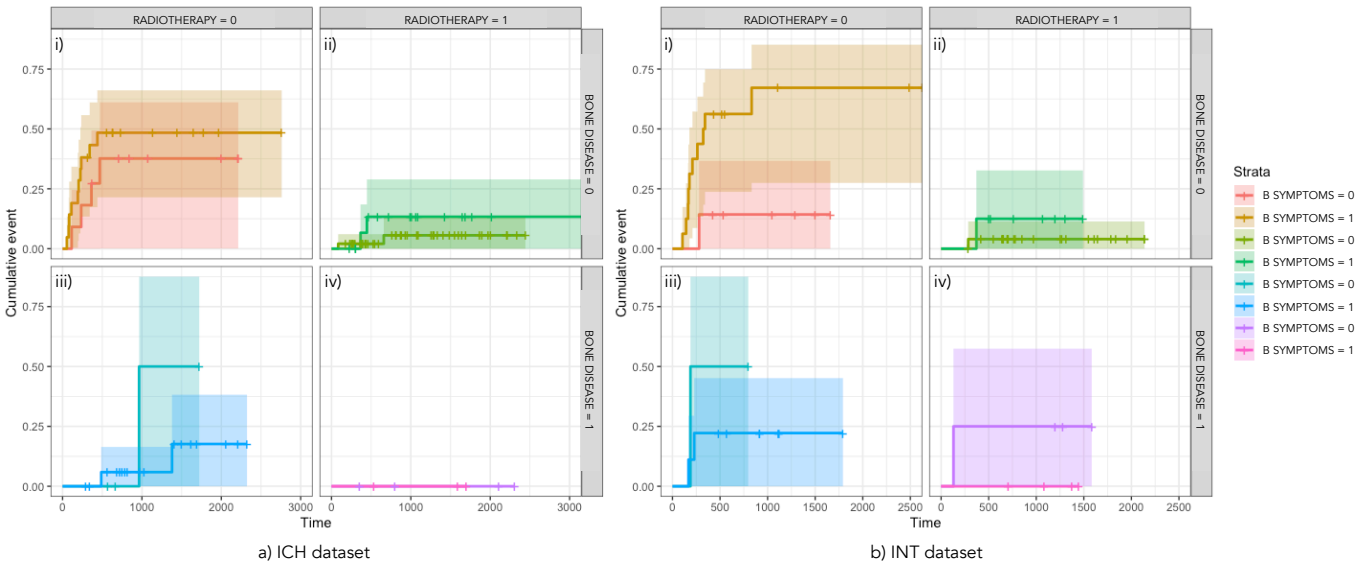

Fig. 5. Cumulative disease-free-survival curves in a) ICH dataset and b) INT dataset. In both cases, four types of patients are considered in the box, namely i) not having bone lesions and not undergoing radiotherapy, ii) not having bone lesions and undergoing radiotherapy, iii) having bone lesions and not undergoing radiotherapy and vi) having bone lesions and undergoing radiotherapy. Each group is divided into subjects presenting and not presenting B symptoms.

be implemented and properly compared. Our approach currently relies on a weighting strategy between patient's lesions as to end up with an easy-to-handle vector representation where the dispersion and the counting indexes account for the multi-level structure of samples. On one hand, the employed wide data format - as the transformation of the long data format - has been shown to entail an exhaustive summary of the patient relevant information that let exploit the reliability of the matrix data. Of course, additional information, including other source of data such as genomics and blood analysis, could be included in the vector to better describe the cancer assessment from a multi-omic point of view. On the other hand, lesions' observations could be re-arranged in an object-based representation of the patients with higher complexity as to manipulate the least the raw data.

VI. Conclusions

In this work, we exploited recurrence-specific graph clustering model for Hodgkin Lymphoma subtyping. The model was applied and evaluated in three different settings, two one-center datasets and one multi-center dataset. We quantified and compared findings when considering diverse populations, acquisition protocols and operator-variability, remarking the limitations of a retrospective approach. In order to extract relevant insights in a perspective way, we employed a interpretable predictive model in order to generalize and transfer the deduced knowledge from a more informative setting to a less valuable one. This work provided a preliminary yet robust evidence of the reliability of recurrence-specific supervised graph clustering approach in properly stratifying cancer subtypes in a perspective way.

Appendix I
Descriptive Survival Statistics

An exploratory survival analysis has been conducted on clinical information of both datasets, i.e. ICH and INT. Specifically, collected data included the annotation of clinical-relevant risk factors, such as lesions location in bone or extra nodal tissue and the onset of B symptoms significant to the prognosis and staging of the disease (namely fever, drenching night sweats and heavy body weight loss). We explored the influence of these factors in the disease-free-survival probability and found consistent results in the two centers.

In Figures 5, we represented the curves of cumulative events in different groups of patients, stratified according to clinical risk factors: ICH patients and INT patients are displayed respectively in plot a), on the left, and plot b), on the right. In both plots, the upper left box i) shows patients who did not exhibit bone lesions and did not undergo radiotherapy, differentiated among who presented (orange line) and not presented B (red line) symptoms; the upper right box ii) presents those subjects not exhibiting bone disease who underwent radiotherapy, divided into groups with (green) and without (light green) B symptoms; the lower left box iii) features patients with bone disease, no radiotherapy treatment, who did (blue) or did not (light blue) manifest B symptoms; finally, lower right box iv) presents groups with (pink) and without (purple) B symptoms who had bone disease and were treated with radiotherapy.

In both centers, patients who underwent also radiotherapy were less likely to recur with respect to patients who were treated with chemotherapy only. Of course, even though radiotherapy appeared to be a preventing factor from recurrence, disease-free-survival probability was also dependent on the stage of the tumor and the health status of the subjects. Patients with bone lesions seemed to not have worse prognosis in matter of recurrence while patients with B symptoms exhibited in every of the four cases a higher probability to recur and thus a poorer prognosis.

Appendix II
Feature Selection in ML Methods

In Figure 6, the flow of the feature selection process is display. We first carried out a correlation-based skimming of the radiomic variables. According to pair-wise correlation coefficients, 7/45 features did not correlate with any particular feature while the remaining 38/45 features formed 8 uncorrelated groups of redundant variables (colorized feature sets in Figure 6). Only one feature per group (highlighted in bold) was selected according to domain knowledge, robustness and matrix frequency, as to scale back the number of covariates. Additionally, *GLCM Entropy log2* was removed as to

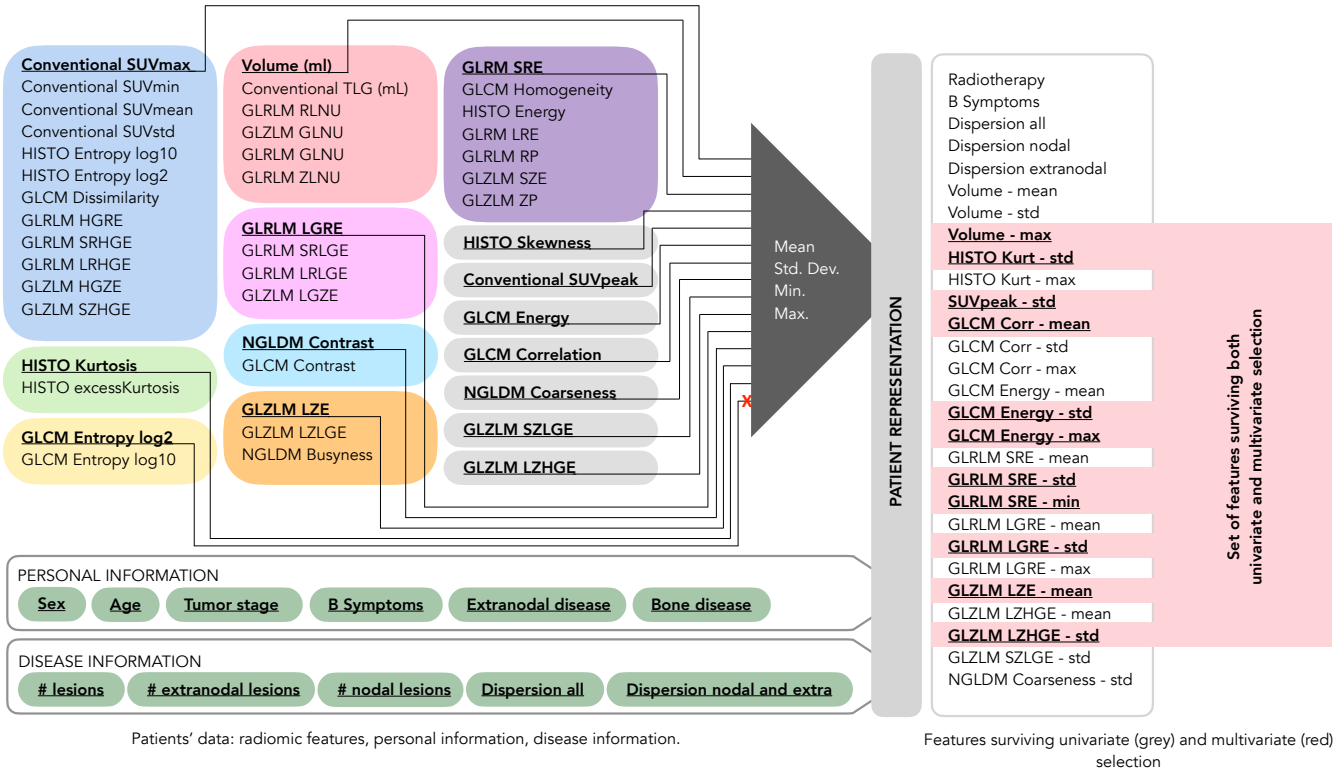

Fig. 6. Patient representation and feature selection process of ML-based model: a correlation criterion was used to select uncorrelated features; personal information, disease information and basic statistics of non-redundant radiomic features formed the patient representation. The grey box contains features robust at univariate selection and red box the ones robust at multivariate selection.

TABLE V

DISCRIMINATION POWER OF RADIOMIC VARIABLES IN STRATIFYING LOW-RISK AND HIGH-RISK PATIENTS IN THE THREE DATASETS (ICH, INT, MULTI-CENTER)

Variables	P-values on ICH dataset	P-values on INT dataset	P-values on multi-center dataset	Variables	P-values on ICH dataset	P-values on INT dataset	P-values on multi-center dataset
Stage	0.0098 **	0.0026 **	0.0000 ***	**GLCM Energy**	0.1700	0.8574	0.8858
Sex	0.3503	0.3869	0.4478	**GLCM Contrast**	0.0328 *	0.2954	0.0612 .
Age	0.9176	0.1265	0.2906	**GLCM Correlation**	0.0099 **	0.8408	0.0935 .
B Symptoms	0.0000 ***	0.0014 **	0.0000 ***	**GLCM Entropy log10**	0.0480 *	0.6268	0.3539
Extranodal disease	0.0111 *	0.0753 .	0.0002 ***	**GLCM Entropy log2**	0.0480 *	0.6268	0.3539
Bone disease	0.1767	0.6932	0.6338	**GLCM Dissimilarity**	0.0546 .	0.3861	0.1052
Radiotherapy	0.0000 ***	0.0000 ***	0.0000 ***	**GLRLM SRE**	0.2018	0.9494	0.3490
# nodal lesions	0.0547 .	0.1078	0.0288 *	**GLRLM LRE**	0.1700	0.8243	0.3466
# extranodal lesions	0.0032 **	0.3415	0.0005 ***	**GLRLM LGRE**	0.0882 .	0.3690	0.1795
Dispersion nodal	0.1226	0.0359 *	0.2131	**GLRLM HGRE**	0.0086 **	0.3309	0.0503 .
Dispersion extra	0.0045 **	0.8894	0.0008 ***	**GLRLM SRLGE**	0.0836 .	0.3578	0.1689
Dispersion all	0.0047 **	0.0557 .	0.0046 **	**GLRLM SRHGE**	0.0092 **	0.3309	0.0532 .
Volume mean	0.1214	0.4658	0.0388 *	**GLRLM LRLGE**	0.1087	0.5191	0.2339
Volume std	0.0019 **	0.1662	0.0025 **	**GLRLM LRHGE**	0.0094 **	0.2857	0.0535 .
Volume min	0.0010 **	0.1933	0.0064 **	**GLRLM GLNU**	0.0087 **	0.2537	0.0481 *
Volume max	0.0003 ***	0.0970 .	0.0003 ***	**GLRLM RLNU**	0.0001 ***	0.1153	0.0040 **
Conventional SUVmin	0.0123 *	0.3523	0.1052	**GLRLM RP**	0.2127	0.9916	0.3442
Conventional SUVmean	0.0155 *	0.3204	0.0632 *	**NGLDM Coarseness**	0.0043 **	0.1361	0.0978 .
Conventional SUVstd	0.0048 **	0.3634	0.0245 *	**NGLDM Contrast**	0.1931	0.3523	0.2105
Conventional SUVmax	0.0021 **	0.2857	0.0157 *	**NGLDM Busyness**	0.1965	0.6957	0.4669
Conventional SUVpeak	0.0002 ***	0.4097	0.0353 *	**GLZLM SZE**	0.0074 **	0.4854	0.0383 *
Conventional TLG (mL)	0.0049 **	0.3634	0.0173 *	**GLZLM LZE**	0.2439	0.7193	0.3948
HISTO Skewness	0.0067 **	0.7998	0.2062	**GLZLM LGZE**	0.0533 .	0.3523	0.1363
HISTO Kurtosis	0.0463 *	0.8161	0.1032	**GLZLM HGZE**	0.0058 **	0.2905	0.0423 *
HISTO ExcessKurtosis	0.0463 *	0.8161	0.1032	**GLZLM SZLGE**	0.0521 .	0.2162	0.1057
HISTO Entropy log10	0.0185 *	0.6419	0.0707 .	**GLZLM SZHGE**	0.0044 **	0.2954	0.0360 *
HISTO Entropy log2	0.0185 *	0.6419	0.0707 .	**GLZLM LZLGE**	0.8165	0.9242	0.9706
HISTO Energy Uniformity	0.0509 .	0.7036	0.1826	**GLZLM LZHGE**	0.0007 ***	0.1476	0.0099 **
SHAPE Volume (mL)	0.0403 *	0.4658	0.0347 *	**GLZLM GLNU**	0.0036 **	0.1254	0.0211 *
GLCM Homogeneity	0.1327	0.6495	0.2376	**GLZLM ZLNU**	0.0001 ***	0.3309	0.0023 **
				GLZLM ZP	0.2358	0.6343	0.1873

reduce the Variance Inflation Factors. Beside radiomic features, personal and qualitative disease information were considered. As to exhaustively represent the patients, the patient-wise distributions of uncorrelated radiomic features have been considered. In fact, each

patient had a variable number of lesions with corresponding radiomic description. As to consider the variability within the patient, the mean value, the standard deviation, the minimum value and the maximum value of each radiomic variable have been computed per each subject. In this way, the vector based patient representation is composed by 6 patient-related covariates, 6 disease-related covariates, 56 imaging-related covariates, for a total of 68 covariates. These features were tested to be relevant in the stratification of patients through different methods: univariate selection included ANOVA, Pearson Correlation, Spearman Correlation, Mutual Information and Univariate Logit while multivariate selection included wrapper methods (Forward Selection, Backward Selection, Bidirectional Selection and Recursive Feature Elimination) and penalized logistics (Ridge Regression, LASSO and Elastic Net). Variables in the grey box are the ones that survived at univariate reduction phase (being significant at 4/5 methods). Variables in the red box are the ones that hold their significance at multivariate reduction phase (being robust at 5/7 techniques).

Appendix III
Discrimination power of radiomic variables

In Table V, we list the specification of the Mann-Whitney U tests of radiomic variables in patient stratification. For each of the three dataset - namely ICH, INT and multi-center datasets - we display the p-values of the univariate tests performed on every radiomic variable. Significance is marked with a "." if $0.05 < p - value < 0.1$, "*" if $0.01 < p - value < 0.05$, "**" if $0.001 < p - value < 0.01$ and "***" if $p - value < 0.001$.

Stage, *B Symptoms*, *Extranodal disease*, *Radiotherapy*, *Dispersion of all lesions* and *Volume* (lesions' maximum value) were significant in all datasets. 45/65 features were significant in ICH dataset, 7/65 in INT dataset and 35/65 in multi-center dataset. 13 features were not significant in any of the datasets, including *Sex*, *Age* and 11 radiomic features. 33 features were significant both in ICH dataset and multi-center dataset.

Acknowledgment

We acknowledge all the personnel of Medicine Department for the assistance during the PET/CT scans, segmentation of lesions, extraction of radiomic features and retrieval of patients' personal information from EHR. We particularly thank dr. Matteo Biroli (ICH), dr. Fabrizia Gelardi (ICH), dr Francesca Ricci (ICH), dr. Ettore Seregni (INT) and dr. Paolo Corradini (INT) for their support.

References

[1] O. Menyhárt and B. Győrffy, "Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis," *Computational and Structural Biotechnology Journal*, vol. 19, p. 949, 2021.

[2] X. Dai, T. Li, Z. Bai, Y. Yang, X. Liu, J. Zhan, and B. Shi, "Breast cancer intrinsic subtype classification, clinical use and future trends," *American journal of cancer research*, vol. 5, no. 10, p. 2929, 2015.

[3] A. Szymiczek, A. Lone, and M. R. Akbari, "Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review," *Clinical Genetics*, vol. 99, no. 5, pp. 613–637, 2021.

[4] L. Cavinato, N. Gozzi, M. Sollini, C. Carlo-Stella, A. Chiti, and F. Ieva, "Recurrence-specific supervised graph clustering for subtyping hodgkin lymphoma radiomic phenotypes," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 2155–2158.

[5] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.

[6] B. Taha, D. Boley, J. Sun, and C. Chen, "Potential and limitations of radiomics in neuro-oncology," *Journal of Clinical Neuroscience*, vol. 90, pp. 206–211, 2021.

[7] Y. Chang, K. Lafata, C. Wang, X. Duan, R. Geng, Z. Yang, and F.-F. Yin, "Digital phantoms for characterizing inconsistencies among radiomics extraction toolboxes," *Biomedical Physics & Engineering Express*, vol. 6, no. 2, p. 025016, 2020.

[8] D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A. K. Jones, and L. Court, "Measuring ct scanner variability of radiomics features," *Investigative radiology*, vol. 50, no. 11, p. 757, 2015.

[9] M. Wu and J. Ma, "Association between imaging characteristics and different molecular subtypes of breast cancer," *Academic radiology*, vol. 24, no. 4, pp. 426–434, 2017.

[10] C. Alili, E. Pages, F. C. Doyon, H. Perrochia, I. Millet, and P. Taourel, "Correlation between mr imaging–prognosis factors and molecular classification of breast cancers," *Diagnostic and interventional imaging*, vol. 95, no. 2, pp. 235–242, 2014.

[11] J. Wu, Y. Cui, X. Sun, G. Cao, B. Li, D. M. Ikeda, A. W. Kurian, and R. Li, "Unsupervised clustering of quantitative image phenotypes reveals breast cancer subtypes with distinct prognoses and molecular pathways," *Clinical Cancer Research*, vol. 23, no. 13, pp. 3334–3342, 2017.

[12] C. Thongprayoon, M. A. Mao, M. T. Keddis, A. G. Kattah, G. Y. Chong, P. Pattharanitima, V. Nissaisorakarn, A. K. Garg, S. B. Erickson, J. J. Dillon *et al.*, "Hypernatremia subgroups among hospitalized patients by machine learning consensus clustering with different patient survival," *Journal of Nephrology*, pp. 1–9, 2021.

[13] K. Raza and N. K. Singh, "A tour of unsupervised deep learning for medical image analysis," *Current Medical Imaging*, vol. 17, no. 9, pp. 1059–1077, 2021.

[14] D. Ay and O. Tastan, "Identifying cross-cancer similar patients via a semi-supervised deep clustering approach," *bioRxiv*, pp. 2020–11, 2021.

[15] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, 2019.

[16] L. Lu and B. J. Daigle Jr, "Prognostic analysis of histopathological images using pre-trained convolutional neural networks: application to hepatocellular carcinoma," *PeerJ*, vol. 8, p. e8668, 2020.

[17] G. Marinos, C. Symvoulidis, and D. Kyriazis, "Micsurv: Medical image clustering for survival risk group identification," in *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)*. IEEE, 2021, pp. 1–4.

[18] L. Manduchi, R. Marcinkevičs, M. C. Massi, T. Weikert, A. Sauter, V. Gotta, T. Müller, F. Vasella, M. C. Neidert, M. Pfister *et al.*, "A deep variational approach to clustering survival data," *arXiv preprint arXiv:2106.05763*, 2021.

[19] Y. Wei, N. Papachristou, S. Mueller, W. H. Chang, and A. G. Lai, "Application of ensemble clustering and survival tree analysis for identifying prognostic clinicogenomic features in patients with colorectal cancer from the 100,000 genomes project," *BMC research notes*, vol. 14, no. 1, pp. 1–7, 2021.

[20] C. Liu, C. Wenming, S. Wu, W. Shen, D. Jiang, Z. Yu, and H. San Wong, "Supervised graph clustering for cancer subtyping based on survival analysis and integration of multi-omic tumor data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[21] C. Nioche, F. Orlhac, S. Boughdad, S. Reuzé, J. Goya-Outi, C. Robert, C. Pellot-Barakat, M. Soussan, F. Frouin, and I. Buvat, "Lifex: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity," *Cancer research*, vol. 78, no. 16, pp. 4786–4789, 2018.

[22] S. F. Barrington, N. G. Mikhaeel, L. Kostakoglu, M. Meignan, M. Hutchings, S. P. Müeller, L. H. Schwartz, E. Zucca, R. I. Fisher, J. Trotman *et al.*, "Role of imaging in the staging and response assessment of lymphoma: consensus of the international conference on malignant lymphomas imaging working group," *Journal of clinical oncology*, vol. 32, no. 27, p. 3048, 2014.

[23] F. Orlhac, M. Soussan, J.-A. Maisonobe, C. A. Garcia, B. Vanderlinden, and I. Buvat, "Tumor texture analysis in 18f-fdg pet: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis," *Journal of Nuclear Medicine*, vol. 55, no. 3, pp. 414–422, 2014.

[24] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[25] C. Bénard, G. Biau, S. Veiga, and E. Scornet, "Interpretable random forests via rule extraction," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 937–945.

[26] C. Bénard, G. Biau, S. Da Veiga, and E. Scornet, "Sirus: Stable and interpretable rule set for classification," *Electronic Journal of Statistics*, vol. 15, no. 1, pp. 427–505, 2021.

[27] D. McFadden *et al.*, "Conditional logit analysis of qualitative choice behavior," 1973.
[28] J. P. Higgins and A. Whitehead, "Borrowing strength from external trials in a meta-analysis," *Statistics in medicine*, vol. 15, no. 24, pp. 2733–2749, 1996.
[29] S. S. Yip and H. J. Aerts, "Applications and limitations of radiomics," *Physics in Medicine & Biology*, vol. 61, no. 13, p. R150, 2016.
[30] V. Nardone, A. Reginelli, C. Guida, M. P. Belfiore, M. Biondi, M. Mormile, F. B. Buonamici, E. Di Giorgio, M. Spadafora, P. Tini *et al.*, "Delta-radiomics increases multicentre reproducibility: a phantom study," *Medical Oncology*, vol. 37, no. 5, pp. 1–7, 2020.
[31] J. P. Crandall, T. J. Fraum, M. Lee, L. Jiang, P. Grigsby, and R. L. Wahl, "Repeatability of 18f-fdg pet radiomic features in cervical cancer," *Journal of Nuclear Medicine*, vol. 62, no. 5, pp. 707–715, 2021.
[32] M. Sollini, M. Kirienko, L. Cavinato, F. Ricci, M. Biroli, F. Ieva, L. Calderoni, E. Tabacchi, C. Nanni, P. L. Zinzani *et al.*, "Methodological framework for radiomics applications in hodgkin's lymphoma," *European Journal of Hybrid Imaging*, vol. 4, pp. 1–17, 2020.
[33] M. Sollini, F. Bartoli, L. Cavinato, F. Ieva, A. Ragni, A. Marciano, R. Zanca, L. Galli, F. Paiar, F. Pasqualetti *et al.*, "[18f] fmch pet/ct biomarkers and similarity analysis to refine the definition of oligometastatic prostate cancer," *EJNMMI research*, vol. 11, no. 1, pp. 1–10, 2021.
[34] L. Deantonio, M. E. Milia, T. Cena, G. Sacchetti, C. Perotti, M. Brambilla, L. Turri, and M. Krengli, "Anal cancer fdg-pet standard uptake value: correlation with tumor characteristics, treatment response and survival," *La radiologia medica*, vol. 121, no. 1, pp. 54–59, 2016.
[35] S. G. Ahn, J. T. Park, H. M. Lee, H. W. Lee, T. J. Jeon, K. Han, S. A. Lee, S. M. Dong, Y. H. Ryu, E. J. Son *et al.*, "Standardized uptake value of 18 f-fluorodeoxyglucose positron emission tomography for prediction of tumor recurrence in breast cancer beyond tumor burden," *Breast Cancer Research*, vol. 16, no. 6, pp. 1–10, 2014.
[36] F. Davnall, C. S. Yip, G. Ljungqvist, M. Selmi, F. Ng, B. Sanghera, B. Ganeshan, K. A. Miles, G. J. Cook, and V. Goh, "Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice?" *Insights into imaging*, vol. 3, no. 6, pp. 573–589, 2012.
[37] S. H. Moon, J. Kim, J.-G. Joung, H. Cha, W.-Y. Park, J. S. Ahn, M.-J. Ahn, K. Park, J. Y. Choi, K.-H. Lee *et al.*, "Correlations between metabolic texture features, genetic heterogeneity, and mutation burden in patients with lung cancer," *European journal of nuclear medicine and molecular imaging*, vol. 46, no. 2, pp. 446–454, 2019.
[38] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and reproducibility of radiomic features: a systematic review," *International Journal of Radiation Oncology* Biology* Physics*, vol. 102, no. 4, pp. 1143–1158, 2018.
[39] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, no. 7381, pp. 306–313, 2012.
[40] C. E. Meacham and S. J. Morrison, "Tumour heterogeneity and cancer cell plasticity," *Nature*, vol. 501, no. 7467, pp. 328–337, 2013.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

23/2022 Masci, C.; Ieva, F.; Paganoni, A.M.
A multinomial mixed-effects model with discrete random effects for modelling dependence across response categories

24/2022 Cappozzo, A.; McCrory, C.; Robinson, O.; Freni Sterrantino, A.; Sacerdote, C.; Krogh, V.; Par
A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events

22/2022 Regazzoni, F.; Pagani, S.; Quarteroni, A.
Universal Solution Manifold Networks (USM-Nets): non-intrusive mesh-free surrogate models for problems in variable domains

21/2022 Cappozzo, A.; Ieva, F.; Fiorito, G.
A general framework for penalized mixed-effects multitask learning with applications on DNA methylation surrogate biomarkers creation

20/2022 Clementi, L.; Gregorio, C; Savarè, L.; Ieva, F; Santambrogio, M.D.; Sangalli, L.M.
A Functional Data Analysis Approach to Left Ventricular Remodeling Assessment

19/2022 Lupo Pasini, M.; Perotto, S.
Hierarchical model reduction driven by machine learning for parametric advection-diffusion-reaction problems in the presence of noisy data

18/2022 Bennati, L; Vergara, C; Giambruno, V; Fumagalli, I; Corno, A.F; Quarteroni, A; Puppini, G; L
An image-based computational fluid dynamics study of mitral regurgitation in presence of prolapse

17/2022 Regazzoni, F.
Stabilization of staggered time discretization schemes for 0D-3D fluid-structure interaction problems

14/2022 Zappon, E.; Manzoni, A.; Quarteroni A.
Efficient and certified solution of parametrized one-way coupled problems through DEIM-based data projection across non-conforming interfaces

16/2022 G. Ciaramella, T. Vanzan
Substructured Two-grid and Multi-grid Domain Decomposition Methods