Artificial Intelligence In Medicine 138 (2023) 102522

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Explainable domain transfer of distant supervised cancer subtyping model via imaging-based rules extraction

Lara Cavinato [a,*], Noemi Gozzi [c,1], Martina Sollini [b,c], Margarita Kirienko [d], Carmelo Carlo-Stella [b,e], Chiara Rusconi [f], Arturo Chiti [b,c], Francesca Ieva [a,g]

[a] Department of Mathematics, Politecnico di Milano, Via Bonardi 9, Milan, 20133, Italy
[b] Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy
[c] Department of Nuclear Medicine, IRCCS Humanitas Research Hospital, Milan, Italy
[d] Fondazione IRCCS Istituto Nazionale dei Tumori, Giacomo Venezian 1, Milan 20133, Italy
[e] Oncology and Hematology Unit, IRCCS Humanitas Research Hospital, Milan, Italy
[f] Division of Hematology and Stem Cell Transplantation, Fondazione IRCCS Istituto Nazionale dei Tumori, Via Giacomo Venezian 1, Milan 20133, Italy
[g] Health Data Science Center, Human Technopole, Milan, Italy

ARTICLE INFO

Keywords:
Cancer subtyping
Explainability
Image Clustering
Radiomics
Rule extraction
Domain transfer

ABSTRACT

Image texture analysis has for decades represented a promising opportunity for cancer assessment and disease progression evaluation, evolving in a discipline, i.e., radiomics. However, the road to a complete translation into clinical practice is still hampered by intrinsic limitations. As purely supervised classification models fail in devising robust imaging-based biomarkers for prognosis, cancer subtyping approaches would benefit from the employment of distant supervision, for instance exploiting survival/recurrence information. In this work, we assessed, tested, and validated the domain-generality of our previously proposed Distant Supervised Cancer Subtyping model on Hodgkin Lymphoma. We evaluate the model performance on two independent datasets coming from two hospitals, comparing and analyzing the results. Although successful and consistent, the comparison confirmed the instability of radiomics due to an across-center lack of reproducibility, leading to explainable results in one center and poor interpretability in the other. We thus propose a Random Forest-based Explainable Transfer Model for testing the domain-invariance of imaging biomarkers extracted from retrospective cancer subtyping. In doing so, we tested the predictive ability of cancer subtyping in a validation and perspective setting, which led to successful results and supported the domain-generality of the proposed approach. On the other hand, the extraction of decision rules enables to draw of risk factors and robust biomarkers to inform clinical decisions. This work shows the potentialities of the Distant Supervised Cancer Subtyping model to be further evaluated in larger multi-center datasets, to reliably translate radiomics into medical practice. The code is available at this GitHub repository.

1. Introduction

Cancer subtyping typifies the process of stratifying patients into classes of different risks. It is currently the trending approach in literature for targeting personalized medicine and steering treatment decisions in oncological research [1,2]. Several methodological strategies have been explored, ranging from supervised, semi-supervised, and unsupervised learning models on both structured and unstructured data, above all genomics [3–5]. Furthermore, imaging data analysis - in the form of radiomic features [6] - is known to be a non-invasive surrogate of tumor biological underpinnings, extracted from routinely acquired exams. In fact, throughout machine learning literature, imaging-based cancer subtyping has started catching on and several associations have been found between imaging/radiomics data and molecular cancer subtypes, hormone receptor status, and cancer severity [7,8]. However, traditional supervised approaches as currently exploited in clinical literature have unveiled the limitations of the radiomics framework [9]. First, high dimensional data calls for massive feature selections which mostly require, as well as classification models, multiple and balanced data. Poor repeatability and reproducibility of the results are indeed due

* Corresponding author.
E-mail address: lara.cavinato@polimi.it (L. Cavinato).
[1] Department of Health Sciences and Technology, ETH Zurich, Universitatstrasse 2, Zurich 8090, Switzerland (present address).

https://doi.org/10.1016/j.artmed.2023.102522
Received 31 August 2022; Received in revised form 19 January 2023; Accepted 23 February 2023
Available online 4 March 2023

to imbalance and scarcity of data. This can hardly be overcome: in fact, the number of samples is limited to the number of cases, few when dealing with rare diseases; the 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 of patients; finally, the variability of the reconstruction parameters, acquisition settings, and scanners is due to the lack of standardization in clinical practice. For these reasons, the current paradigm of radiomics has shifted towards more complex and non-fully supervised strategies for patient stratification and imaging-based risk factor identification.

Several Image Clustering (IC) techniques have been proposed to match imaging features to clinical cancer subtypes and to quantify its prognostic association with survival and recurrence-free survival rates [10,11]. The most up-to-date IC approaches for survival risk prediction in medical imaging adopt unsupervised or semi-supervised deep learning solutions [12,13]. [14] developed an unsupervised encoder with Cox loss to compress clinical, mRNA, microRNA expression data, and histopathology Whole Slide Images (WSIs) to perform cancer subtyping. Similarly, [15] performed a prognostic analysis of histopathological images of hepatocellular carcinoma using a pre-trained CNN to extract latent features; they kept the features significant at Cox analysis and applied an SVM model for stratification. Moreover, [16] proposed a pipeline consisting of learning the image latent representation from survival CNN, a dimensionality reduction step, and the clustering evaluation. Finally, in the framework of stochastic gradient variational inference, [17] proposed a deep probabilistic approach to retrieve clusters driven by latent variables and survival information. All such approaches extract imaging representation features from somewhat trained CNNs and need to apply an a posteriori supervised feature selection procedure, to either reduce the data dimensionality or to keep only survival-informative variables. Therefore, the fragmented nature of these pipelines prevents them from explainable assessing the imaging capability of devising risk factors in a perspective way. Moreover, deep embeddings, unlike radiomic vectors, do not entail standardized and interpretable features.

In our previous work [18], we leveraged a Distant Supervision (DS) approach to perform Cancer Subtyping (CS) of Hodgkin Lymphoma patients according to their radiomic phenotype. Specifically, DS is a particular case of weekly supervision where some higher-level labels are used to perform the classification task [19]. This approach often allows for making the training more efficient and, here, permits boosting an unsupervised model. The prognostic reliability of the detected subpopulations, the scalable performance, and the interpretability of the model was shown to be the main advantages of this approach. In fact, the characterization of the groups emerging by subtyping the population may enhance the clinical interpretation of radiomic features in terms of both cancer severity and therapy response. We indeed provided a tool for reversing the paradigm of interpreting the biological meaning of higher-order radiomic variables.

Based on these considerations, in this work, we explored the robustness of the Distant Supervised Cancer Subtyping (DS-CS) model to the domain shift, in particular concerning the across-center variability of the scans. Specifically, we compare the results obtained on two datasets coming from different hospitals to discuss the concordance of findings (Section 2.4). The consistency of the results suggests the DS-CS domain-generality, however, interpretability appears to be bounded by the informative content of data. Interestingly, the prognostic power of imaging biomarkers improves with a borrowing strength strategy (Section 3.1). Upon such findings, as a second contribution, we use a classification model to exploit the domain transfer in a perspective way. Specifically, we propose a Random Forest-based Explainable Transfer Model to transfer the cancer subtyping policy from one setting to the other (Section 2.5). We extract agnostic imaging-based rules that are shown to be both robust and prognostic (see Section 3.3). Although such results cannot be considered definitive, we believe that this work provides interesting insights for testing the robustness of Distant Supervised Cancer Sub typing in identifying imaging cancer subtypes in an agnostic and perspective way.

2. Methods and analyses

This section exposes the analytical pipeline. Before illustrating the models, in Sections 2.1 and 2.2 we describe the data collection and the harmonization process to provide an overview of the datasets summary information. The DS-CS model is described in Section 2.3 since it represents the basic block on which the contributions of this work are built upon.

As in Fig. 1, we then explain the analytical workflow to support the claims of the present work. In Section 2.4, we conduct a robust reproducibility analysis and compare the DS-CS model in different settings: two different single-center datasets and one multi-center dataset. Results are assessed in terms of cancer subtypes characterization, i.e., the group-wise probability to recur and between-groups discrimination power of radiomic features.

As will be pointed out in Section 3.1, the intrinsic limitations of radiomic features are overcome with distant supervision, however, the identification of domain-general prognostic biomarkers could be bounded by the biases of the labels. Pertinently, in Section 2.5 we introduce a model for explainable transferring the DS-CS model onto different domains and deduce imaging-based perspective rules to be applied in other settings/domains. Performance evaluation, improvements, and interpretation of results follow this section.

All models have been implemented in MATLAB [20] as well as evaluation results. However, we performed survival analysis and extracted explainability rules using R [21].

2.1. Data collection

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

ICH enrolled 128 patients in the study as they met the inclusion criteria. They were diagnosed with Hodgkin Lymphoma and were treated and followed up at the center. Pre-treatment [^{18}F]FDG PET/CT imaging 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 the [^{18}F]FDG-avid lesions were located and semi-automatically segmented by an expert nuclear medicine physician (M. S.) with a 40 % of SUVmax threshold. The LIFEx software was used for segmentation, as well as for imaging harmonization and feature extraction as explained in Section 2.2 (www.lifexsoft.org, [22]). In total, 1340 lesions were collected and quantitatively assessed. Survival and recurrence-free survival information were also registered. Chemotherapy starting dates, dates of *ad interim* PET (iPET), and End Of Treatment (EOT) PET were collected to extract temporal information of therapy pathways. The 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 the first-line treatment (refractory) and patients who 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. Patient information is made available in Table 1 for categorical variables and Table 2 for numerical variables.

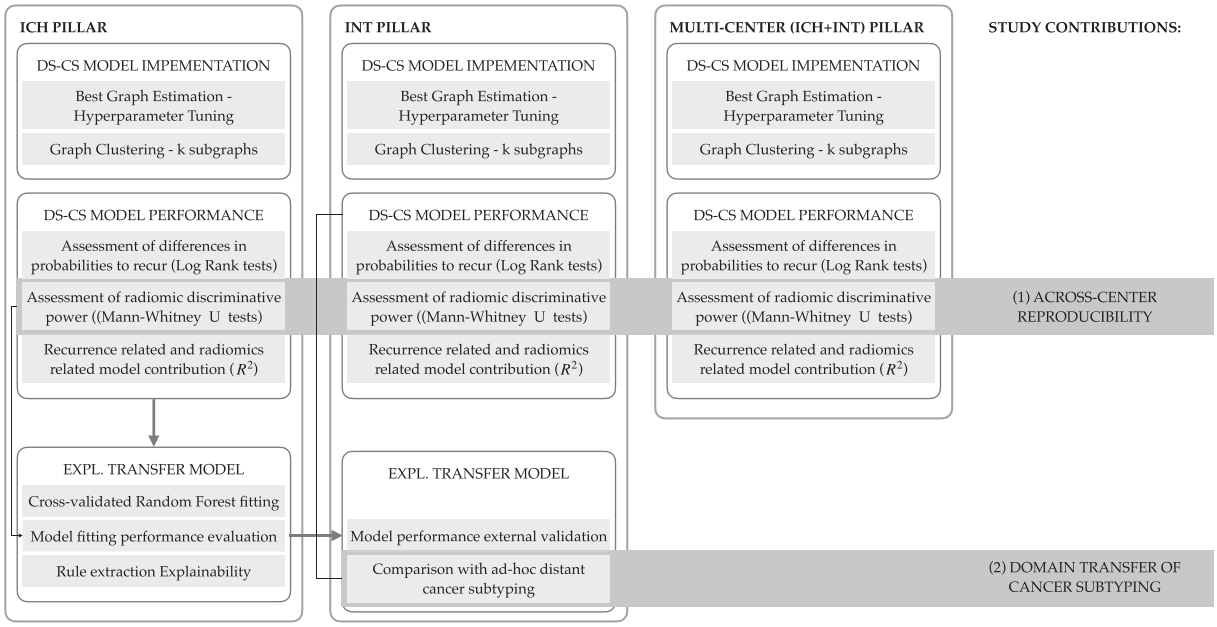

Fig. 1. Methodological workflow: objective (1) provides a comparison between the Distant Supervised Cancer Subtyping model applied to different domains. In particular: DS-CS model implementation on the ICH pillar is presented in Section 2.4.1; DS-CS model performance on the ICH pillar is presented in Section 3.1.1; DS-CS model implementation on the INT pillar is presented in Section 2.4.2; DS-CS model performance on ICH pillar is presented in Section 3.1.2; DS-CS model implementation on ICH + INT pillar is presented in Section 2.4.3; DS-CS model performance on ICH + INT pillar is presented in Section 3.1.3; overall across-center reproducibility is discussed in Section 3.1.4. Objective (2) describes the domain transfer of the DS-CS model via Explainable Transfer Model training (Sections 2.5 and 3.2); validation performance on testing dataset and robustness concerning domain-shift are displayed in Section 3.3. Rule extraction explainability is shown in Section 3.4.

Table 1
Humanitas Research Hospital (ICH) patients' categorical characteristics.

Categorical variables – N (%)		Responders ($N = 107$)	Non-responders ($N = 21$)
Stage	I	9 (8 %)	0 (0 %)
	II	57 (53 %)	11 (52 %)
	III	12 (11 %)	2 (10 %)
	IV	30 (28 %)	8 (38 %)
Sex	F	62 (58 %)	14 (67 %)
	M	45 (42 %)	7 (33 %)
B symptoms	N	60 (56 %)	7 (33 %)
	Y	47 (44 %)	14 (67 %)
Extranodal disease	N	74 (69 %)	11 (52 %)
	Y	33 (31 %)	10 (48 %)
Bone disease	N	80 (75 %)	18 (86 %)
	Y	27 (25 %)	3 (14 %)
Radiotherapy	N	38 (35 %)	17 (81 %)
	Y	69 (65 %)	4 (19 %)
iPET	DS1	82 (77 %)	10 (48 %)
	DS2	12 (11 %)	2 (9 %)
	DS3	11 (10 %)	1 (5 %)
	DS4	2 (2 %)	5 (24 %)
	DS5	0 (0 %)	3 (14 %)
PET EOT	DS1	77 (72 %)	13 (62 %)
	DS2	11 (10 %)	3 (14 %)
	DS3	10 (9 %)	1 (5 %)
	DS4	3 (3 %)	1 (5 %)
	DS5	6 (6 %)	3 (14 %)

Table 2
Humanitas Research Hospital (ICH) patients' numerical characteristics.

Numerical variables – mean (std deviation)	Responders (N = 107)	Non Responders (N = 21)
Age	39.252 (15.875)	40.143 (15.963)
# Nodal lesions	6.673 (4.813)	6.619 (6.184)
# Extranodal lesions	1.916 (5.750)	3.857 (10.256)
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 (z-score)	0.028 (0.520)	0.455 (0.963)
Std. dev. volume (z-score)	0.529 (0.833)	1.270 (1.702)
Minimum volume (z-score)	−0.307 (0.141)	−0.326 (0.084)
Maximum volume (z-score)	1.157 (2.136)	2.582 (3.352)
Time to relapse [days]	1126.97 (704.94)	358.86 (322.854)

The same criteria were used to enroll patients and analyze images at INT. Most of the patients were diagnosed at the center and information about those patients for whom this was not the case was retrieved and properly annotated. [^{18}F]FDG PET/CT images of 76 Hodgkin Lymphoma patients (794 lesions) were analyzed by an expert nuclear medicine physician (M.K.) using LIFEx software. Clinical data about demographics, chemotherapy cycle length, radiotherapy treatment, and follow-up were collected. Both iPET and EOT PET were defined as positive in presence of an area of [^{18}F]FDG uptake higher than the background as defined by a Deauville Score of DS4 or DS5. DS3 or lower was consistent with a negative exam [23]. Information about the specific Deauville Score of each patient was available only for the ICH dataset. For 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 3 for categorical variables and Table 4 for numerical variables. In Supplementary Tables 1 and 2, scanner specifications are detailed for both centers. Moreover, in Section 2 of Supplementary Materials, descriptive statistics of disease-free-survival times according to clinical variables are explored.

2.2. Harmonization and patient representation

For comparison purposes, a harmonization step was required both from clinical and imaging points of view. First, all clinical and personal information was processed with a strategy of compliance with the less rich dataset. That is, response to treatment and cancer progression were

Table 3
National Cancer Institute (INT) patients' categorical characteristics.

Categorical variables – N (%)		Responders ($N = 59$)	Non-responders ($N = 17$)
Stage	I	1 (2 %)	0 (0 %)
	II	31 (52 %)	4 (23 %)
	III	6 (10 %)	1 (6 %)
	IV	21 (36 %)	12 (71 %)
Sex	F	34 (58 %)	8 (47 %)
	M	25 (42 %)	9 (53 %)
B symptoms	N	35 (59 %)	4 (23 %)
	Y	24 (41 %)	13 (77 %)
Extranodal disease	N	39 (65 %)	7 (41 %)
	Y	20 (45 %)	10 (59 %)
Bone disease	N	44 (75 %)	13 (77 %)
	Y	15 (25 %)	4 (23 %)
Radiotherapy	N	20 (45 %)	14 (82 %)
	Y	39 (65 %)	3 (18 %)
iPET	Negative	55 (93 %)	8 (47 %)
	Positive	4 (7 %)	9 (53 %)
PET EOT	Negative	59 (100 %)	0 (0 %)
	Positive	0 (0 %)	17 (100 %)

Table 4
National Cancer Institute (INT) patients' numerical characteristics.

Numerical variables – mean (std deviation)		
Age	36.478 (13.915)	42.867 (17.868)
# Nodal lesions	7.271 (5.499)	9.706 (6.362)
# Extranodal 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 (z-score)	0.075 (0.542)	0.176 (0.784)
Std. dev. volume (z-score)	0.625 (1.030)	0.931 (1.394)
Minimum volume (z-score)	−0.331 (0.090)	−0.358 (0.087)
Maximum volume (z-score)	1.312 (2.453)	2.304 (3.368)
Time to relapse [days]	1105.72 (546.490)	257.59 (167.17)

flagged by a dichotomous variable, survival information was neglected and times to events were computed.

Moreover, as LIFEx software participates in the Image Biomarker Standardization Initiative, consistent imaging harmonization and feature extraction were implemented. We performed homogeneous gray levels discretization with a Fixed Bin Number (FBN) of 64 and we rescaled pixels' intensities according to absolute resampling bounds (min = 0 SUV; max = 20 SUV) to account for the tissue-specific variability. No spatial resampling and further image pre-processing were implemented for managing the voxel volume dependency of features. However, only ROIs bigger than 64 voxels were considered adequate for radiomics analysis and kept. From Regions Of Interest (ROIs), i.e., lesions, radiomic description was computed. The radiomic signature consisted of 45 radiomic features including conventional (e.g., intensity-based indexes), first (e.g., histogram-based and shape indexes), second (e.g., GLCM-derived indexes), and higher order (e.g., GLRLM-, GLZLM-, and NGLDM-derived indexes) statistics. The features' definitions are described in the LIFEx 4.9 manual [22]. These groups are known to entail different texture information and can thus be seen as four different imaging views of the tumor. Additionally, to normalize features to remove the so-called batch, or center, effect, we performed z-score normalization to each radiomic variable separately for each hospital. *Z*-score normalization was proven to outperform other standardization methods [24]. Specifically, z-transform reshapes each variable to fit a Normal distribution $\mu = 0$ and $\sigma = 1$ and applies the following transformation:

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

where x_i is the un-normalized variable value, $\bar{x}$ the variable mean and σ the variable standard deviation. In principle, each radiomic variable was thus disentangled from the center effect.

Given the multi-lesion nature of the patients, a single radiomic vector was built as described by [18]. Specifically, lesions' radiomic features were averaged patient-wise to obtain the lesions' mean radiomic profile/phenotype of patients. Additionally, some variables were added and/or transformed to enrich the disease description. The number of total lesions, number of nodal and extranodal lesions, and dispersions of all, nodal and extranodal lesions within a patient were computed as a proxy of tumor spread and heterogeneity. In this way, each patient i was described by a standardized vector X_i entailing all the tumor information, including the four quantitative radiomic views and the one clinical description of the disease, i.e., the six qualitative variables. A total of five types of features, i.e., five views, was thus accounted for in the imaging-based disease representation X_i of patients. After harmonization, the ICH dataset contained 128 patients described by 61 variables and the INT dataset held 76 patients described by the same 61 variables.

2.3. Distant Supervised Cancer Subtyping (DS-CS) model

The DS-CS model takes as input the imaging-based disease representation X_i of each patient. The pipeline is built through two methodological steps and one interpretation step, as depicted in Fig. 2. First, the patient-to-patient graph describing the population under analysis is computed. Specifically, the affinity matrix needs to be estimated. Of course, the algorithm's hyperparameters must be optimized. Consequently, the graph is segmented according to spectral clustering, that is, homogeneous sub-populations of nodes with similar properties are devised and clustered apart. To validate the subtyping procedure, sub-populations of patients need to be characterized with clinical variables, endogenous and exogenous to the model building.

2.3.1. Patient-to-patient graph estimation

The patient-to-patient similarity graph is obtained by minimizing the following objective function as suggested by [25]:

$$\begin{aligned} \min_{w;S} & \sum_{k=1}^{m}\left(-\sum_{i=1}^{n}\delta_i\left(X_i^k w^k - log\sum_{j\in R_i} exp\left(X_j^k w^k\right)\right)\right) \\ & +\lambda\sum_{k\neq j}\left\|X^k w^k - X^j w^j\right\|_2^2 + \eta\sum_{k=1}^{m}\left\|w^k\right\| \\ +\min_{S}\gamma & \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\left\|X_i - X_j\right\|^2 + \left\|X_i w - X_j w\right\|^2\right) S_{i,j} + \mu S_{i,j}^2 \\ & s.t.\sum_{j}^{n} S_{i,j} = 1; S_i \geqslant 0; i = 1, 2, \ldots, n. \end{aligned} \tag{1}$$

The loss function (1) allows estimating the distance between patients in terms of both imaging-based disease representation and disease-free survival prediction. Specifically, it develops in four terms. We recall that each patient is described by a set of radiomic and qualitative features X_i as described in Section 2.2 and their disease-free survival information $\{\delta_i, T_i\}$, where δ_i is the censoring variable indicating if the recurrence has taken place or not and T_i the time to recurrence. Moreover, as previously stated, the X_i imaging-based disease representation is a vector containing features of $m = 5$ different natures (or *views*), whose contribution may be weighted differently in the following model.

The first term in (1) implements the distant supervision of the model. Specifically, the negative partial log-likelihood of the Cox (survival) model is computed, where X_i^k is the set of variables belonging to $\mathrm{k-th}$ view of $\mathrm{i-th}$ patient and R_i is the set of patients observed alive almost at time T_i. Additionally, n is the total number of patients, and m is the number of imaging views. In other words, per each patient, the survival risk is obtained as a linear combination of the risks associated with every

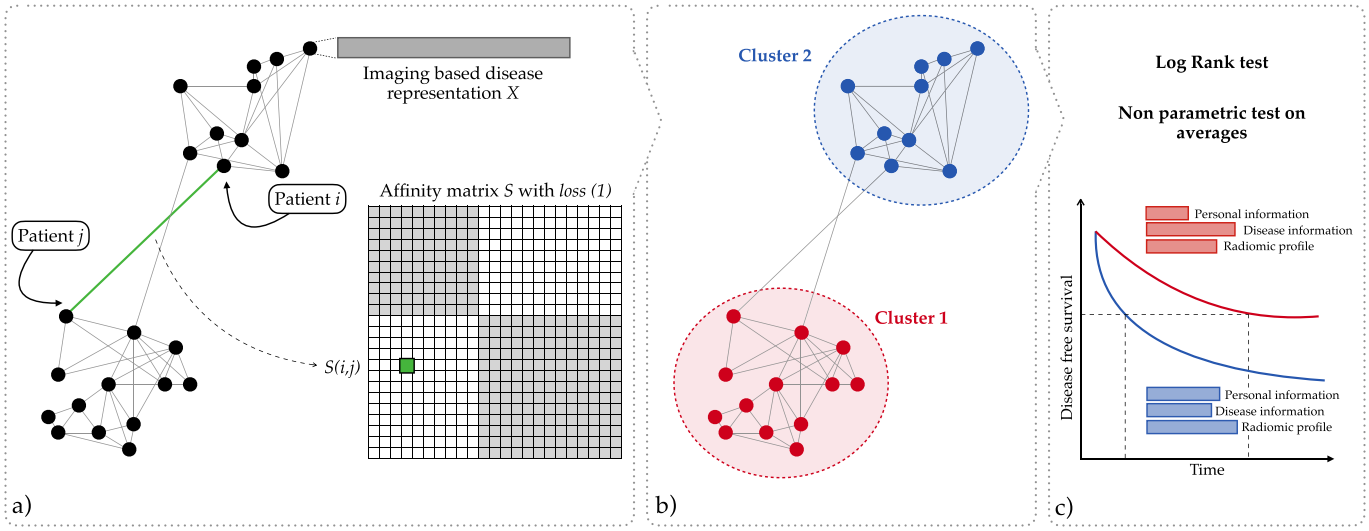

Fig. 2. The Distant-Supervised Cancer Subtyping pipeline: a) computation and optimization of patient-to-patient similarity graph via minimization of the loss function (1) as described in Sections 2.3.1 and 2.3.2 respectively; b) clustering/segmentation of patient-to-patient graph into subgraphs, i.e., clusters of nodes (Section 2.3.3); c) cluster-wise characterization and comparison in terms of survival probabilities and imaging variables.

input feature $X_j^k \cdot w^k$. The loss function computes the distance between the i − th patient's survival risk and the risks of the rest of the patients in the population and minimizes it by estimating the risk coefficients w. In this way, patients at similar disease-free survival risks will be closer in the graph than patients at different risks. The recurrence probability is thus the distant label for supervising the model.

The second and the third terms in (1) carry out the L2 and L1 regularization over the input features. In fact, given the high dimensionality of the vector X_i, some feature selection may be needed. λ drives the regularization between the views, shrinking the contributions of views with similar information. The sparsity parameter η removes non-informative features, penalizing their contribution to the overall model.

Finally, the fourth term in (1) minimizes the distance between imaging-based disease representations of patients (i.e., the five-view vector). It computes the pairwise difference between the input vectors of every pair of patients (X_i, X_j) and their survival probabilities $(X_i \cdot w, X_j \cdot w)$ estimated in the first term. In this way, the overall graph affinity matrix $S \in \Re^{n \times n}$ is learnt. The entries $S_{i,j}$ of the affinity matrix constitute the recurrence-informed similarity between each pair of patients. Finally, the parameter γ represents the learning rate and μ is a information-complexity trade-off parameter.

According to this formulation, a two-fold objective is pursued: the survival analysis with the computation of risks w given S and the estimation of the similarity graph S given the risks w. As proposed by [25], an alternating optimization algorithm was used to solve the corresponding problem.

2.3.2. Hyperparameters optimization

Hyperparameters are optimized according to grid search. In fact, λ (the co-regularization parameter), η (the L1 penalization parameter), and γ (the learning rate) contribute to modulating the distance between patients. Since the only supervision involved in the estimation is represented by the Cox survival loss, a concordance index with survival information may be exploited. Specifically, we select the parameter values among a range experimentally, estimate the graph's affinity matrix, and compute the Harrell's concordance index (c-index) of the estimated survival risks [26]. The values that maximize Harrell's c-index are selected as the optimal ones.

2.3.3. Spectral clustering

To segment the population graph, the spectral clustering algorithm is used as it represents an efficient strategy for the clustering of heterogeneous disease expression data [27]. Specifically, spectral clustering uses information from the spectrum of the graph affinity matrix to cluster nodes. The affinity matrix indeed entails the information about the quantitative pairwise relationship between nodes and encodes such interaction in a way that it can be exploited for clustering purposes. Resulting clusters, or classes, are then intended as groups of patients with similar properties, in terms of both imaging-based disease phenotype and survival expectation. Accordingly, each class corresponds to a cancer subtype.

To choose the number of clusters k, the eigengap heuristic is followed. Once the Laplacian of the graph, either normalized or non-normalized, is computed, k is equal to the number of its null eigenvalues [28]. If k leads to obtaining subgraphs with <10 nodes, every small subgraph is merged with its closest subgraph and k is accordingly decreased.

2.4. Across-center reproducibility

Having described the data and the DS-CS model, we proceed to detail the assessment of across-center reproducibility. This represents the first step towards the evaluation of DS-CS domain-invariance.

2.4.1. DS-CS model on ICH data

The DS-CS model was first applied to the ICH dataset. The loss function was optimized and tuned on data for estimating the patient-to-patient similarity graph. The optimal choice was 0.1 for γ, meaning that convergence requires several iterations to be guaranteed, whereas regularization parameters were set to 0.4 and 0.01 for η and λ respectively. Since $\eta \neq 0$, the less informative features were indeed removed from the model. The value of $\lambda \sim 0$ suggests a disagreement between imaging views, that is, they provide different perspectives of imaging information which need to be exploited in full. The spectral clustering procedure devised $k = 2$ classes of patients (nodes), bringing to the separation of different cancer imaging phenotypes with different prognoses (results will be discussed in Section 3.1.1).

2.4.2. DS-CS model on INT data

For comparison and qualitative assessment purposes, the very same procedure was applied to and optimized for the INT dataset. Optimal parameters for γ, η and λ were found as described in Section 2.3.2. Their values were set to 0.1, 0.4, and 0.02 respectively, being in line with the

Table 5
Discrimination power of radiomic variables in stratifying low-risk and high-risk patients in the three datasets (ICH, INT, multi-center). The table presents the p-values of the Mann-Whitney *U* tests for the difference in imaging variables' distributions in each model. Significance is marked with a "." if 0.05 < p-value < 0.1, with "*" if$0.01 < p-value < 0.05, with "**" if 0.001 < p-value < 0.01, and with "***" if p-value < 0.001.

Variable	P-values on ICH dataset	P-values on INT dataset	P-values on multi-center dataset	Variable	P-values on ICH dataset	P-values on INT dataset	P-values on multi-center dataset
Stage	0.0098 **	0.0026 **	0.0000 ***	GLCM Contrast	0.0328 *	0.295	0.0612 .
Sex	0.3503	0.3869	0.4478	GLCM Correlation	0.0099 **	0.840	0.0935 .
Age	0.9176	0.1265	0.2906	GLCM Entropy log1	0.0480 *	0.626	0.3539
B Symptoms	0.0000 ***	0.0014 **	0.0000 ***	GLCM Entropy log2	0.0480 *	0.626	0.3539
Extranodal disease	0.0111 *	0.0753 .	0.0002 ***	GLCM Dissimilarity	0.0546 .	0.386	0.1052
Bone disease	0.1767	0.6932	0.6338	GLRLM SRLGE	0.2018	0.949	0.3490
Radiotherapy	0.0000 ***	0.0000 ***	0.0000 ***	GLRLM LRE	0.1700	0.824	0.3466
# nodal lesions	0.0547 .	0.1087	0.0288 *	GLRLM LGRE	0.0882 .	0.369	0.1795
# extranodal lesions	0.0032 **	0.3415	0.0005 ***	GLRLM HGRE	0.0086 **	0.330	0.0503 .
Dispersion nodal	0.1226	0.0359 *	0.2131	GLRLM SRLGE	0.0836 .	0.357	0.1689
Dispersion extranodal	0.0045 **	0.8894	0.0008 ***	GLRLM SRHGE	0.0092 **	0.330	0.0532 .
Dispersion all	0.0047 **	0.0557 .	0.0046 **	GLRLM LRLGE	0.1087	0.519	0.233
Volume mean	0.1214	0.4658	0.0388 *	GLRLM LRHGE	0.0094 **	0.285	0.0535 .
Volume std	0.0019 **	0.1662	0.0025 **	GLRLM GLNU	0.0087 **	0.253	0.0481 *
Volume min	0.0010 **	0.1933	0.0064 **	GLRLM RLNU	0.0001 **	0.115	0.0040 **
Volume max	0.0003 ***	0.0970 .	0.0003 ***	GLRLM RP	0.2127	0.991	0.3442
Conventional SUVmin	0.0123 *	0.3523	0.1052	NGLDM Coarseness	0.0043 **	0.136	0.0978 .
Conventional SUVmean	0.0155 *	0.3204	0.0632 *	NGLDM Contrast	0.1931	0.352	0.2105
Conventional SUVstd	0.0048 **	0.3634	0.0245 *	NGLDM Busyness	0.1965	0.695	0.4669
Conventional SUVmax	0.0021 **	0.2857	0.0157 *	GLZLM SZE	0.0074 **	0.485	0.0383 *
Conventional SUVpeak	0.0002 ***	0.4097	0.0353 *	GLZLM LZE	0.2439	0.719	0.3948
Conventional TLG (mL)	0.0049 **	0.3634	0.0173 *	GLZLM LGZE	0.0533 .	0.352	0.1363
HISTO Skewness	0.0067 **	0.7998	0.2062	GLZLM HGZE	0.0058 **	0.290	0.0423 *
HISTO Kurtosis	0.0463 *	0.8161	0.1032	GLZLM SZLGE	0.0521 .	0.216	0.1057
HISTO ExcessKurtosis	0.0463 *	0.8161	0.1032	GLZLM SZHGE	0.0044 **	0.295	0.0360 *
HISTO Entropy log10	0.0185 *	0.6419	0.0707 .	GLZLM LZLGE	0.8165	0.924	0.9706
HISTO Entropy log2	0.0185 *	0.6419	0.0707 .	GLZLM LZHGE	0.0007 **	0.147	0.0099 **
HISTO Energy Uniformity	0.0509 .	0.7036	0.1826	GLZLM GLNU	0.0036 **	0.125	0.0211 *
SHAPE Volume (mL)	0.0403 *	0.4658	0.0347 *	GLZLM ZLNU	0.0001 **	0.330	0.0023 **
GLCM Homogeneity	0.1327	0.6495	0.2376	GLZLM ZP	0.2358	0.634	0.1873
GLCM Energy	0.1700	0.857	0.8858				

ICH model. Similarly to the ICH dataset, $k = 2$ groups of different cancer subtypes were identified by spectral clustering (see Section 3.1.2) and could be compared with ICH results (see Section 3.1.4).

2.4.3. DS-CS model on ICH + INT data

As an additional level of analysis, the two datasets have been merged and the DS-CS pipeline was run on the multi-center dataset to evaluate the results irrespectively to the provenience of the observations. As will be pointed out in Section 3.1, we anticipate that the DS-CS ICH model brought to high discrimination power of imaging features while DS-CS INT model did not. For such reason, we have investigated whether such power could be used to improve the DS-CS INT model, inflating the variability of the data coming from a different population, i.e., ICH center. The values of the parameters that led to the higher c-index performance were 0.1 for the learning rate (γ), 0.5 for L1-penalization (η), and 0.01 for L2-regularization (λ). Accordingly, convergence was guaranteed, non-informative features were deleted and a mild shrinkage between views was accounted for. Similarly to ICH and INT cases, the spectral clustering procedure on the multi-center dataset resulted in $k = 2$ clusters of patients exhibiting different cancer subtypes. Results are described in Section 3.1.3 and compared with the ones of DS-CS ICH model and DS-CS INT model in Section 3.1.4.

2.5. Explainable transfer model

As it will be further discussed in Section 3.1, the application of DS-CS models to different datasets did produce consistent cancer subtyping policies. However, it was not possible to validate the across-domain agreement of the imaging biomarkers because of the DS-CS INT model lack of interpretability and the retrospective nature of the model. Generally speaking, it is a literature open problem to perspectively apply a retrospective model to new observations. Here, we propose an Explainable Transfer Model (ETM), that is, a perspective and interpretable approach for transferring any retrospective and unsupervised model, e.g. the DS-CS model, in a validation setting. In our context, we implemented the domain transfer of the DS-CS model via ETM to test DS-CS domain-generality.

2.5.1. Training

To build the ETM, ICH decision rules have been extracted from DS-CS ICH model with the scope of exploiting its prognostic power and its robustness. A Random Forest (RF) 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 the ICH dataset, considering only those features that were significant at univariate testing in DS-CS ICH model (see Mann-Whitney U tests in Section 3.1.4 and Table 5). The training performance of the transfer model is discussed in Section 3.2.

2.5.2. Testing

Upon model training, it was applied to the INT dataset and performance has been evaluated in terms of survival differences and radiomics prognostic power. A new set of labels resulted from the model transferring, which led to grouping patients into two risk classes, one with a poorer and one with a milder prognosis. The new labels were compared with the one resulting from the DS-CS INT model described in Section 2.4.2 and improvements were evaluated in terms of interpretability (Section 3.3).

2.5.3. Explainability

As for interpretability, decision rules were extracted from ETM and were interpreted to identify clinical and radiomic features as risk factors and/or biomarkers. Every rule of every tree split was annotated and kept when common enough in the forest to be relevant; similar rules were then post-treated and aggregated to define a stable, interpretable, and unique set of elementary rules driving the decisions making [29,30]. The algorithm was first trained in a cross-validation fashion to estimate the optimal hyperparameter p_0 used to select the number of relevant rules to extract. Specifically, p_0 represents the proportion of RF's trees in which a rule must appear 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 n most relevant decision rules. Section 3.4 details the findings.

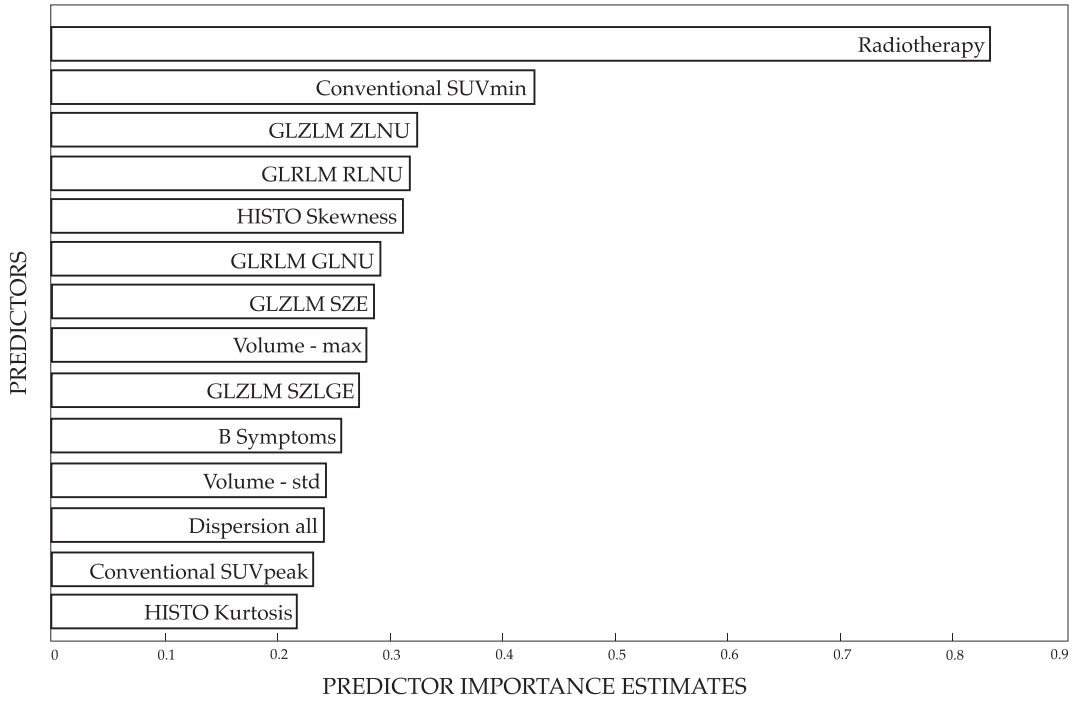

Fig. 3. Feature importance plot of the ETM: the most relevant features in the Random Forest-based model are presented with a descending order The importance of the features was computed according to Out-of-bag (OOB) permuted predictor delta error [33].

3. Results

Section 3 of Supplementary Materials describes the limitations of performing a traditional machine learning-based radiomics model following current literature guidelines. Supervised frameworks are proven to fail in discriminating tumor subtypes, laying the foundation for this work's objectives. In fact, the retrospective nature of the models and the limited variability of observations prevent translating the traditional radiomic approach into clinical practice. Additionally, the clinical conclusions that may derive are tightly dependent on the observed dataset – and their labels – and result in poor validation performance. It happens quite frequently in literature to find inconsistencies and lack of consensus in radiomics literature, even concerning same cancer [31]. The very same questions have been inquired about the DS-CS model.

3.1. Across-center reproducibility

In this section, we first present the results of each DS-CS model separately and, then, discuss the concordance of the findings.

3.1.1. DS-CS model on ICH data

As displayed in Section 4 of Supplementary Materials (Fig. 3), two groups were identified and the Kaplan-Meier curves of groups' probability to recur were built for each group. The p-value of the Log Rank test between the two Kaplan-Meier curves resulted to be $\ll 0.01$, showing strong significance. Moreover, the Hazard Ratio was computed as the ratio between the risk of recurrence in group 1 and the risk of recurrence in group 2. According to Hazard Ratio (0.2176, IC 95 %: 0.1202–0.3937), group 1 was characterized by a better prognosis with almost no recurrence experienced, while group 2 contained patients with a poorer prognosis, who were instead more likely to recur. Clinical and radiomic features were used to interpret the risk classes, emerging significantly different in several cases. To compare the average imaging description of one class with respect to the other class, two-sided non parametric tests on averages (Mann-Whitney U tests) were used and p-values lower than the threshold of 0.1 were considered significant (see Table 5, first column).

3.1.2. DS-CS model on INT data

As displayed in Section 4 of Supplementary Materials (Fig. 3), two groups were obtained and tested to be significantly different in terms of prognosis (p-value of the Log Rank test between the two Kaplan-Meier curves $\ll 0.01$). As emerged from Hazard Ratio (0.0627, IC 95 %: 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. Mann-Whitney U tests were performed to evaluate differences between the two groups and to characterize the cancer subtyping policy (see Table 5, second column). The clustering characterization was interpreted as a rule for describing cancer subtypes. Such characterization could thus be compared with the one coming from the ICH dataset for repeatability purposes.

3.1.3. DS-CS model on ICH + INT data

Similarly to ICH and INT cases, the cancer subtyping model resulted successfully on the multi-center dataset. Two Kaplan-Meier curves were computed for the patients belonging to the two risk classes and the Log Rank test led to significant results (p-value of the Log Rank test $\ll 0.01$). From the Hazard Ratio assessment (0.1117, IC 95 %: 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. Mann-Whitney U tests on variables were performed to compare the two groups, resulting to be significant in almost all cases (see Table 5, third column).

3.1.4. Comparison between models

The three models brought a significant classification of patients with different prognoses as survival curves were tested to be different in all cases. The Hazard Ratios were consistent in all three cases ($HR_{ICH} < 1$, $HR_{INT} < 1$, $HR_{ICH+INT} < 1$), suggesting the coherency of the cancer subtyping policies. In each dataset, two - severe and mild - classes of risks were obtained and could thus be compared. In principle, all three classes of mild cancer should present the same radiomic characterization while all three classes of severe cancer should display a similar imaging phenotype. To investigate this point, tests on input features were performed and compared in the three cases. Results are displayed in Table 5. For each of the three datasets - namely ICH, INT, and multi-center ICH + INT datasets - we list the p-values of the univariate tests performed on every variable. Significance is marked with a "·" if $0.05 <$ p-value < 0.1, with "*" if $0.01 <$ p-value < 0.05, with "**" if $0.001 <$ p-value < 0.01, and "***" if p-value < 0.001.

45/61 features were significant in ICH dataset, 7/61 in INT dataset and 34/61 in multi-center dataset. 33/61 features were significant in both the ICH dataset and multi-center dataset while 6/61 in both the INT dataset and multi-center dataset. Features significant in both the ICH dataset and INT dataset preserved significance in the multi-center dataset and showed consistency. Specifically, these were 6/61: Stage, B Symptoms, Extranodal disease, Radiotherapy, Dispersion of all lesions, and Volume (i.e., lesions' maximum value). Most of the features that resulted significant in the DS-CS ICH model but not in the DS-CS INT model (27/45) were strong enough to remain significant in the multi-center DS-CS ICH + INT model. Such variables were equally found among first-order, second-order, and higher-order radiomic features, as well as qualitative disease information like volume and number of nodal and extranodal lesions. The remaining features did not hold significance in the multi-center DS-CS ICH + INT model, being overshadowed by INT data noise. Of course, variables that were not significant in DS-CS ICH model nor DS-CS INT model remained not significant in the multi-center case (14/61). These include Sex, Age, Bone disease, and 11 radiomic features.

As the patient-to-patient similarity graph was estimated by minimizing patients' differences both in terms of imaging-based disease representation and disease-free-survival probabilities (distant supervision), we evaluated the imaging-survival balance in the graph estimation in each of the three models via Logistic Regression. Specifically, we fed the radiomic features into a Logistic Regression to predict the cancer subtypes (clustering labels). The pseudo – R2 of the model was computed as the ratio between the log-likelihood of the intercept model, i.e., the one with no features, and the log-likelihood of the full model, with all radiomic features. The pseudo – R2 thus quantifies the improvement offered by the full model over the intercept model [32] and can be intended as the capacity of radiomics to explain the cancer subtyping. As expected from tests' significance, the pseudo – R2 statistics was 65 % (p-value $\ll 0.01$) in the DS-CS ICH model and 46 % (p-value $= 0.045$) in the DS-CS INT model. In fact, features that are not significant in univariate testing are less likely to be predictive in a multivariate setting. In the multi-center DS-CS ICH + INT model, the informative content of radiomic data rose thanks to the higher variability, suggesting the strength of having multiple - even if domain-shifted - data. The pseudo – R2 statistics of the Logistic Regression was 70 % (p-value $\ll 0.01$), testifying the preponderant role of radiomics.

3.2. Explainable transfer model training performance

The cross-validated ETM was successfully trained on the ICH dataset and intentionally let overfit. As expected from univariate testing and pseudo−R2 values, the model was able to capture all the variability entailed in the data ready to be exploited to classify new observations into risk classes. Since radiomics contribution was high in the DS-CS ICH model, a purely radiomics-based model was informative enough to

perfectly fit the data into cluster labels. Besides accuracy, which resulted to be 97.66 %, other more relevant performance evaluation criteria were 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 reveal a model highly overfitting the training data, with the clear aim of obtaining an interpretable and predictive mirror of the cancer subtyping model. Contrary to common machine learning best practices, here we 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 of the two groups was significant (p-value < 0.01), suggesting that the ETM-based classification was indeed associated with cancer prognosis. The classification devised one group with fast-relapsing patients and one with long- or non-relapsing patients (Hazard Ratio = 0.2230, IC: 0.1227–0.4054). Being the fitting robust enough to be intended as a Rule Extractor of the DS-CS model, the ETM was worth to be applied on the INT dataset to test domain-generality of DS-CS and to deduce robust risk factors and imaging biomarkers.

3.3. Explainable transfer model testing performance

The ETM model was transferred to the INT dataset. The resulting classification identified two risk classes. Having borrowed the information about radiomic variability and cutoffs from the ICH dataset, classes of patients are expected by construction to have characteristics of imaging phenotypes similar to ICH groups. Furthermore, the obtained classification showed an appreciable agreement with the stratification performed by the ad hoc DS-CS INT model (see Section 2.4.2). In fact, the concordance index between the two reached 0.7.

The two groups resulting from the Explainable Transfer Model classification were compared in terms of recurrence probabilities (Log Rank test on groups' Kaplan-Meier curves), leading to significant discrimination between a better prognosis and a poorer prognosis class (p-value = 0.0105). As highlighted by the Hazard Ratio (0.2496, IC 95 %: 0.1240–0.5026), group 1 was characterized by a higher disease-free life expectancy than group 2. Moreover, unlike CS-DS INT model (see Sections 2.4.2 and 3.1.2), the Explainable Transfer Model led to 38 radiomic variables being significantly different between the two groups.

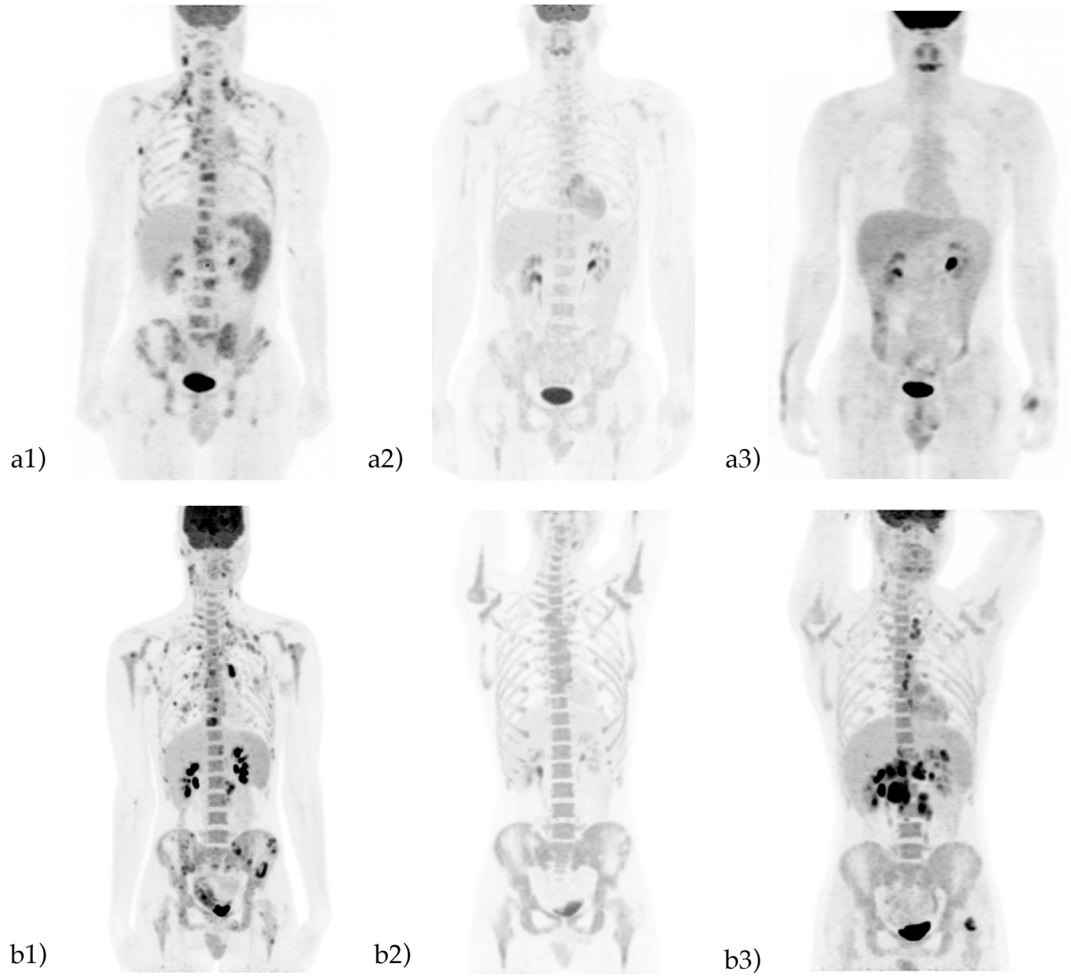

Fig. 4. Example of two patients with HL and different outcomes. a) [^{18}F]FDG PET/CT in a 35-year-old male with HL presenting B symptoms. Baseline MIP image (a1) shows [^{18}F]FDG uptake in lymph nodes, lungs, spleen, and bones; accordingly, the patient was staged as a stage IV HL. Interim PET/CT (a2) obtained after two cycles of chemotherapy, shows a complete metabolic response (Deauville Score 1) demonstrating the disappearance of all sites of pathological [^{18}F]FDG uptake and a mild diffuse bone marrow hypermetabolism, as typically observed shortly after chemotherapy (B). End-Of-Treatment MIP (a3) was negative confirming the previous finding. The patient had no evidence of disease at the last follow-up, 17 months after the end of chemotherapy; b) [^{18}F]FDG PET/CT in a 19-year-old male with HL presenting B symptoms. Baseline MIP image (b1) shows [^{18}F]FDG uptake in lymph nodes, lungs, spleen, and bones; accordingly, the patient was staged as a stage IV HL. Interim PET/CT (b2) obtained after two cycles of chemotherapy, shows a complete metabolic response (Deauville Score 1) demonstrating the disappearance of all sites of pathological [^{18}F]FDG uptake and a diffuse moderate bone marrow hypermetabolism, as typically observed shortly after chemotherapy (B). End-Of-Treatment imaging was negative, but the disease relapsed 12 months after the end of chemotherapy as confirmed by MIP image (b3) which shows [^{18}F]FDG uptake in lymph nodes, lungs, and bones. At the last follow-up, the patient was alive with evidence of disease.

These include conventional, first and second-order texture statistics. It follows coherently that the pseudo – R2 statistics of the Logistic Regression resulted to be 63 % (p-value < 0.01), attesting to the radiomics contribution in the cancer subtyping policy. Of course, these features were consistent with the evaluation of CS-DS ICH model and CS-DS ICH + INT model.

3.4. Explainability of the extracted rules

As to interpret the rules extracted by the ETM, we first want to look at the feature importance plot. As displayed in Fig. 3, the ranking of the Random Forest predictors has been computed based on their importance and the top relevant ones were plotted. We selected the first variables which presented higher absolute importance, for a total of 18 features. Most of them (13/18) were found among those features that showed significance also in the DS-CS INT model or held significance in the DS-CS ICH + INT model (see Table 5). 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 Random Forest.

In line with the importance plot, different clinical and radiomic features were found in the common rules set. The lists of common rules can be assessed in Appendix A. The lists' order should not be intended as consecutive, but all rules are rather as important (and frequent) as the others and are divided according to the type of variable they refer to. Moreover, the thresholds of continuous features (that is, all features except for Radiotherapy and B Symptoms) refer to the z-standardized feature values and can be interpreted in terms of quantiles of the cumulative distribution function.

Among the extracted rules, radiotherapy was a strong factor to predict the cancer subtypes: absence of radiotherapy treatment led to a higher probability of incurring into cancer relapsing subtype (Pr = 0.909). Although such finding was already known in clinical practice more severe patients are often treated with both chemotherapy and radiotherapy - it is worth noticing that radiotherapy was frequently observed together with dispersion and radiomic variables. The absence of radiotherapy and values of lesions dispersion higher than 67 % of the samples (z = −0.431) brought a higher probability of recurrence (Pr = 0.979). Moreover, when considered together with values of GLRLM Run Length Non-Uniformity higher than 77 % of the samples (z = −0.744), the probability of recurring cancer subtype without radiotherapy rose 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.

Other relevant rules also testify to the same point. In fact, other clinical variables were fundamental for cancer prognosis, i.e., the presence of B symptoms [34], related to a 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 samples (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 samples (z = −0.439). Indeed, huge differences in lesions' volume within the same patient are proxies of intra-patient heterogeneity. All the other decision rules account for radiomic variables, in particular 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 led to worse tumor progression when assuming high values compared 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, stronger results were found in the correspondence of rules exploiting higher-order radiomic features, supporting the prognostic value of radiomics.

4. Discussion

Clinical practice has for long relied on purely visual inspection of images for diagnosis, treatment planning, and follow-up. However, some crucial information might not be caught, affecting clinical decisions. For instance, as shown in Fig. 4, apparently identical patients may develop diseases with different outcomes. For these patients, visual analysis of the medical imaging would result in the same treatment approach, producing ineffective results.

On purpose, imaging-based cancer subtyping promises to be a reliable tool for tumor evolution prediction, especially when informed by survival/recurrence information. However, its robustness and domain-generality are yet to be explored. In this work, we intended to address this question and provide a domain-transfer framework for the DS-CS model. The first aim of this work was to compare the model tuned in different settings, as to discuss the consistency of the results. Cancer subtyping on the ICH dataset was appreciably driven by imaging tumor characterization, as most of the features resulted to be significantly different in the two groups. It follows that distant supervision helped in properly modulating the information already entailed in the imaging data. On contrary, although DS-CS INT model was successfully carried out and showed consistency with DS-CS ICH model, very few variables emerged significant at testing. Also, the pseudo – R2 statistics of the Logistic Regression between radiomic variables and cancer subtypes (clustering labels) supported this point. In fact, the pseudo – R2 metric was considered as a quantification of the radiomic features' predictive power in the cancer subtyping model. The low percentage of the contribution of radiomics in the model suggests that the classification of INT data was mainly dragged by the survival estimation (first term of the loss function), whereas radiomic variables played a limited role. In other words, the supervision overshadowed the imaging information, and no prognostic risk factors could be extracted to inform the clinical practice in a perspective way. This issue was overcome with the DS-CS ICH + INT model. In fact, merging the two datasets improved the interpretability of the model. Inflating INT data with the variability of a different source of information (ICH) increased the imaging features significance without affecting the performances, in a borrowing strength strategy [35].

Furthermore, comparing radiomic variable significance across models enabled us to acknowledge variables that are agnostic to imaging acquisition settings and texture extraction parameters. As expected, agnostic variables were mainly related to clinical and qualitative information about the disease, i.e., the stage, the B symptoms, the extra-nodal disease status, radiotherapy, and – more importantly – the dispersion of lesions. Such disease information was proven to be agnostic to the center of provenience and could be acknowledged as robust in a prospective study. We recall that dispersion is the variability over the lesions' radiomic features of a patient. Here, this variability was computed as the average distance between radiomic variables of peer lesions, i.e., belonging to the same patient. Accordingly, it can be intended as a proxy of biological tumor heterogeneity that has been previously shown to cause treatments' inefficacy and relapsing [36,37]. Indeed, this definition of tumor dispersion emerges robust as it aggregates the imaging information in a standardized way [38,39].

As a second point of contribution, we extracted imaging-based rules via Random Forest to explainable perform the domain transfer of the DS-CS model and ensure its repeatability. We built the DS-CS model on one dataset (ICH) and transferred the knowledge to a domain-shifted dataset (INT). Of note, in previous literature – and in Section 3 of Supplementary Materials – 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 [40,41]. Nevertheless, our results demonstrated the domain-generality of the DS-CS model, being robust throughout different centers/domains. The proposed Explainable Transfer Model was in fact successful in devising groups of at-different-risk patients with

significantly different time-to-recurrence probabilities in the testing domain (INT). That is, it allowed us to exploit the imaging information in the data to correctly predict the cancer subtypes. Also, the ICH-informed cancer subtyping showed agreement with the ad hoc DS-CS INT model. We remind that the DS-CS model trained on INT did lead to a significant and consistent classification of patients with different prognoses, but the prognostic interpretation of radiomic features was limited by the informative content of data. The fact that the purely radiomics-based classification model showed concordance with the ground truth strengthens the domain-generality of the transferring. In fact, the INT dataset did contain information, although it was masked by radiomic well-known limitations and such information could not be appreciated. ICH-informed model behaved as a magnifying glass and enabled the extraction of radiomic-based knowledge from noisy data. As Hodgkin Lymphoma, like several other tumor diseases, is a rare condition, this approach would support decisions in those cases where only a few observations are available. Also, the aggregated information coming from other sources may aid the evaluation/assessment. We may acknowledge that the ETM was used to provide a proof-of-concept for transferring the subtyping from one domain to another. Of course, under a federated learning perspective, it would be necessary to further test it in other cohorts, to enrich the generality of the information it carries.

At explainability analysis, both clinical and imaging variables emerged as relevant risk factors. The extracted rules have, on one hand, confirmed the prognostic power of known qualitative factors such as tumor volume, radiotherapy, and the presence of B symptoms; on the other hand, they brought out the benefit of accounting for imaging-based tumor heterogeneity measures to consistently improve the cancer subtyping. In fact, several 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 (although we showed and discussed only the more common ones among the tree splits of the Random Forest). In line with these findings, recent literature has sharpened its focus towards repeatability and reproducibility of radiomics in multi-center studies [42–44]. Although sensitive to all the above-mentioned acquisition criteria, far from a few lower and higher order radiomic features were proven to be robust and agnostic.

In conclusion, the proposed approach showed the reliability of the proposed Distant Supervised Cancer Subtyping model and discussed the domain-generality of its transfer to different domains. Of course, these findings could be further confirmed by collecting more data from many different centers/hospitals and integrating the harmonized information to build more informative and agnostic decision models. As only highly anonymous and aggregated data is 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 can be estimated from a higher number of patients and more robust rules can be derived from the procedure. In this direction, an additional point of improvement could be acknowledged: other harmonization strategies could be implemented. Among these, a grouping strategy would be suitable to exploit the hierarchical nature of a multi-center dataset. For instance, frailty loss instead of Cox loss could be implemented to consider the nesting levels (patients into centers) in the graph estimation phase of the algorithm.

Ultimately, alternative patient representation strategies could be considered. Our approach currently relies on a weighting strategy between patients' lesions to end up with an easy-to-handle vector representation where the dispersion indexes account for the multi-level structure of observations. 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's relevant information that lets exploiting the reliability of the matrix data. Of course, additional information, including other sources 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, other weighting strategies and/or lesion selection approaches could be explored to exhaustively represent the complexity of the tumors.

5. Conclusions

In this work, we discussed the potentialities of our previously proposed Distant Supervised Cancer Subtyping model. Robustness and domain-generality were investigated in a multi-center setting of Hodgkin Lymphoma patients. The model was applied and evaluated in two one-center datasets and one multi-center dataset. We quantified and compared findings when considering diverse populations, acquisition protocols, and operator variability, remarking on the limitations of a retrospective approach. Cancer subtypes were coherently found in the three cases, although radiomic predictive power was controversial. To transfer the cancer sub-typing model to different domains, we employed an Explainable Transfer Model. This allowed us to confirm the transferable properties of the model and to extract decision rules to be interpreted in a perspective way. This work provides preliminary yet robust evidence of the reliability of Distant Supervised Cancer Subtyping in properly highlighting cancer subtypes, ready to inform clinical practice.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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.

Appendix A. Common rules sets

Verbose common rules set are listed and divided into radiotherapy informed by radiomics features (Common rules set 1), clinical features (Common rules set 2), and stand-alone radiomic features (Common rules set 3). The thresholds of continuous features (that is, all features except for Radiotherapy and B Symptoms) refer to the z-standardized feature values and can be interpreted in terms of quantiles of the cumulative distribution function.

Common rules set 1 Radiotherapy and Radiomics

Rule 1:
if *Radiotherapy* $= 0$ **then** $Pr = 0.909$
else if *Radiotherapy* $= 1$ **then** $Pr = 0.47$
Rule 2:
if *(Radiotherapy* $= 0$ *& Dispersion all* ≥ -0.431*)* **then** $Pr = 0.979$
else if *(Radiotherapy* $= 1$ *& Dispersion all* < -0.431*)* **then** $Pr = 0.481$
Rule 3:
if *(Radiotherapy* $= 0$ *& GLRLM RLNU* > -0.744*)* **then** $Pr = 0.959$
else if *(Radiotherapy* $= 1$ *& GLRLM RLNU* ≤ -0.744*)* **then** $Pr = 0.481$

Common rules set 2 Clinical variables

Rule 1:
if *B symptoms* $= 0$ **then** $Pr = 0.478$
else if *B symptoms* $= 1$ **then** $Pr = 0.869$
Rule 2:
if *Volume (min)* < -0.0889 **then** $Pr = 0.810$
else if *Volume (min)* ≥ -0.0889 **then** $Pr = 0.523$
Rule 3:
if *Volume (max)* < -0.439 **then** $Pr = 0.490$
else if *Volume (max)* $\geq s -0.439$ **then** $Pr = 0.779$

Common rules set 3 Radiomics

Rule 1:
if *GLZLM ZLNU* < -0.326 **then** $Pr = 0.471$
else if *GLZLM ZLNU* ≥ -0.326 **then** $Pr = 0.792$
Rule 2:
if *GLCM Correlation* < -0.482 **then** $Pr = 0.490$
else if *GLCM Correlation* ≥ -0.482 **then** $Pr = 0.779$
Rule 3:
if *GLZLM LZHGE* < -0.0908 **then** $Pr = 0.562$
else if *GLZLM LZHGE* ≥ -0.0908 **then** $Pr = 0.897$
Rule 3:
if *SUV Peak* < -1.05 **then** $Pr = 0.490$
else if *SUV Peak* ≥ -1.05 **then** $Pr = 0.779$
Rule 3:
if *GLRLM GLNU* < -0.158 **then** $Pr = 0.545$
else if *GLRLM GLNU* ≥ -0.158 **then** $Pr = 0.843$

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2023.102522.

References

[1] Santos C, Sanz-Pamplona R, Nadal E, Grasselli J, Pernas S, Dienstmann R, Moreno V, Tabernero J, Salazar R. Intrinsic cancer subtypes-next steps into personalized medicine. Cell Oncol 2015;38(1):3–16.
[2] Vliet MHVan, Ubels J, Ridder JDe. Method for identifying gene expression signatures, uS Patent App. 16/500,379. Jun. 3 2021.
[3] Menyha´rt O. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Comput. Struct. Biotechnol. J. 2021;19:949.
[4] Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, Shi B. Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res 2015;5(10):2929.
[5] Szymiczek A, Lone A, Akbari MR. Molecular intrinsic versus clinical subtyping in breast cancer: a comprehensive review. Clin Genet 2021;99(5):613–37.
[6] Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016;278(2):563.
[7] Wu M, Ma J. Association between imaging characteristics and different molecular subtypes of breast cancer. Acad Radiol 2017;24(4):426–34.
[8] Ab Mumin N, Hamid MTR, Wong JHD, Rahmat K, Ng KH. Magnetic resonance imaging phenotypes of breast cancer molecular subtypes: a systematic review. Acad Radiol 2022;29:S89–106.
[9] Taha B, Boley D, Sun J, Chen C. Potential and limitations of radiomics in neuro-oncology. J Clin Neurosci 2021;90:206–11.
[10] Wu J, Cui Y, Sun X, Cao G, Li B, Ikeda DM, Kurian AW, Li R. Un- supervised clustering of quantitative image phenotypes reveals breast cancer subtypes with distinct prognoses and molecular pathways. Clin Cancer Res 2017;23(13):3334–42.
[11] Thongprayoon C, Mao MA, Keddis MT, Kattah AG, Chong GY, Pattharanitima P, Nissaisorakarn V, Garg AK, Erickson SB, Dillon JJ, Garovic VD, Cheungpasitporn W. Hypernatremia subgroups among hospitalized patients by machine learning consensus clustering with different patient survival. J Nephrol 2021:1–9.
[12] Raza K, Singh NK. A tour of unsupervised deep learning for medical image analysis. Curr Med Imaging 2021;17(9):1059–77.
[13] Ay D, Tastan O. Identifying cross-cancer similar patients via a semi- supervised deep clustering approach. bioRxiv; 2021. 2020–11.
[14] Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics 2019;35(14):i446–54.
[15] Lu L, Daigle Jr BJ. Prognostic analysis of histopathological images us- ing pre-trained convolutional neural networks: application to hepatocellular carcinoma. PeerJ 2020;8:e8668.
[16] Marinos G, Symvoulidis C, Kyriazis D. Micsurv: medical image clustering for survival risk group identification. In: 2021 4th international conference on bio-engineering for smart technologies (BioSMART). IEEE; 2021. p. 1–4.

[17] Manduchi L, Marcinkevics R, Massi MC, Weikert T, Sauter A, Gotta V, Vasella F, Neidert MC, Pfister M, Stieltjes B, Vogt JE, Mu ̈ller T. A deep variational approach to clustering survival data. arXiv preprint arXiv:2106.05763; 2021.
[18] Cavinato L, Gozzi N, Sollini M, Carlo-Stella C, Chiti A, Ieva F. 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. p. 2155–8.
[19] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision, CS224N project report, Stanford 1 (12)2009; 2009.
[20] MATLAB. Version 9.11.0 (R2021b). Natick, Massachusetts: The MathWorks Inc; 2021.
[21] R.C.Team. R: a language and environment for statistical computing. 2013.
[22] Nioche C, Orlhac F, Boughdad S, Reuzé S, Goya-Outi J, Robert C, Pellot-Barakat C, Soussan M, Frouin F, Buvat I. Lifex: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer Res 2018;78(16):4786–9.
[23] Barrington SF, Mikhaeel NG, Kostakoglu L, Meignan M, Hutch-ings M, Schwartz LH, Zucca E, Fisher RI, Trotman J, Hoekstra OS, Hicks RJ, O'Doherty MJ, Hustinx R, Biggi A, Mu ̈eller SP, Che- son BD. Role of imaging in the staging and response assessment of lymphoma: consensus of the international conference on malignant lymphomas imaging working group. J. Clin. Oncol. 2014;32(27):3048.
[24] Haga A, Takahashi W, Aoki S, Nawa K, Yamashita H, Abe O, Nakagawa K. Standardization of imaging features for radiomics analysis. J Med Investig 2019;66 (1.2):35–7.
[25] Liu C, Cao W, Wu S, Shen W, Jiang D, Yu Z, et al. Supervised graph clustering for cancer subtyping based on survival analysis and integration of multi-omic tumor data. IEEE/ACM Trans Comput Biol Bioinform 2020;19(2):1193–202.
[26] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982;247(18):2543–6.
[27] Ng AY, Jordan MI, Weiss Y. On spectral clustering: analysis and an algorithm. In: Advances in neural information processing systems; 2002. p. 849–56.
[28] Von Luxburg U. A tutorial on spectral clustering. Stat Comput 2007;17(4): 395–416.
[29] Bé nard C, Biau G, Veiga SDa, Scornet E. Sirus: stable and interpretable rule set for classification. Electron J Stat 2021;15(1):427–505.
[30] Bé nard C, Biau G, Veiga S, Scornet E. Interpretable random forests via rule extraction. In: International conference on artificial intelligence and statistics, PMLR; 2021. p. 937–45.
[31] Sollini M, Cozzi L, Ninatti G, Antunovic L, Cavinato L, Chiti A, Kirienko M. Pet/ct radiomics in breast cancer: mind the step. Methods 2021;188:122–32.
[32] McFadden D. Conditional logit analysis of qualitative choice behavior. 1973.
[33] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[34] Shanbhag S, Ambinder RF. Hodgkin lymphoma: a review and update on recent progress. CA Cancer J Clin 2018;68(2):116–32.
[35] Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. Stat Med 1996;15(24):2733–49.
[36] Greaves M, Maley CC. Clonal evolution in cancer. Nature 2012;481(7381):306–13.
[37] Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature 2013;501(7467):328–37.
[38] Sollini M, Kirienko M, Cavinato L, Ricci F, Biroli M, Ieva F, Calderoni L, Tabacchi E, Nanni C, Zinzani PL, Fanti S, Guidetti A, Alessi A, Corradini P, Seregni E, Carlo-Stella C, Chiti A. Methodological framework for radiomics applications in hodgkin's lymphoma. Eur J Hybrid Imaging 2020;4:1–17.
[39] Sollini M, Bartoli F, Cavinato L, Ieva F, Ragni A, Marciano A, Zanca R, Galli L, Paiar F, Pasqualetti F, Erba PA. [18f] fmch pet/ct biomarkers and similarity analysis to refine the definition of oligometastatic prostate cancer. EJNMMI Res 2021;11(1):1–10.
[40] Nardone V, Reginelli A, Guida C, Belfiore MP, Biondi M, Mormile M, Buonamici FB, Di Giorgio E, Spadafora M, Tini P, Grassi R, Pirtoli L, Correale P, Cappabianca S, Grassi R. Delta-radiomics increases multicentre reproducibility: a phantom study. Med Oncol 2020;37(5):1–7.
[41] Crandall JP, Fraum TJ, Lee M, Jiang L, Grigsby P, Wahl RL. Repeatability of 18f-fdg pet radiomic features in cervical cancer. J Nucl Med 2021;62(5):707–15.
[42] Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol Biol Phys 2018;102(4): 1143–58.
[43] Mali SA, Ibrahim A, Woodruff HC, Andrearczyk V, Mu ̈ller H, Primakov S, Salahuddin Z, Chatterjee A, Lambin P. Making radiomics more re- producible across scanner and imaging protocol variations: a review of harmonization methods. J Personalized Med 2021;11(9):842.
[44] Da-Ano R, Visvikis D, Hatt M. Harmonization strategies for multicenter radiomics investigations. Phys Med Biol 2020;65(24):24TR02.