



MOX-Report No. 83/2023

**Dual adversarial deconfounding autoencoder for joint batch-effects
removal from multi-center and multi-scanner radiomics data**

Cavinato, L.; Massi, M.C.; Sollini, M.; Kirienko, M.; Ieva, F.

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

mox-dmat@polimi.it

<https://mox.polimi.it>

Dual Adversarial Deconfounding Autoencoder for joint batch-effects removal from multi-center and multi-scanner radiomics data

Lara Cavinato^{1,+,*}, Michela Carlotta Massi^{2,+}, Martina Sollini^{3,4}, Margarita Kirienko⁵, and Francesca Ieva^{1,2}

¹Department of Mathematics, Politecnico di Milano, Via Bonardi, 9, Milan, 20133, Italy

²Health Data Science Center, Human Technopole, Viale Rita Levi-Montalcini, 1, Milan, 20157, Italy

³Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini, 4, Pieve Emanuele, 20090, Italy

⁴Department of Nuclear Medicine, IRCCS Humanitas Research Hospital, Via Alessandro Manzoni, 56, Rozzano, 20089, Italy

⁵Fondazione IRCCS Istituto Nazionale dei Tumori, Via Giacomo Venezian, 1, Milan, 20133, Italy

*lara.cavinato@polimi.it

+these authors contributed equally to this work

ABSTRACT

Medical imaging represents the primary tool for investigating and monitoring several diseases, including cancer. The advances in quantitative image analysis have developed towards the extraction of biomarkers able to support clinical decisions. To produce robust results, multi-center studies are often set up. However, the imaging information must be denoised from confounding factors – known as batch-effect – like scanner-specific and center-specific influences. Moreover, in non-solid cancers, like lymphomas, effective biomarkers require an imaging-based representation of the disease that accounts for its multi-site spreading over the patient's body. In this work, we address the dual-factor deconfusion problem and we propose a deconfusion algorithm to harmonize the imaging information of patients affected by Hodgkin Lymphoma in a multi-center setting. We show that the proposed model successfully denoises data from domain-specific variability (p -value <0.001) while it coherently preserves the spatial relationship between imaging descriptions of peer lesions (p -value $=0$), which is a strong prognostic biomarker for tumor heterogeneity assessment. This harmonization step allows to significantly improve the performance in prognostic models with respect to state-of-the-art methods, enabling building exhaustive patient representations and delivering more accurate analyses (p -values <0.001 in training, p -values <0.05 in testing). This work lays the groundwork for performing large-scale and reproducible analyses on multi-center data that are urgently needed to convey the translation of imaging-based biomarkers into the clinical practice as effective prognostic tools. The code is available on GitHub at this [link](#).

Introduction

Hodgkin Lymphoma (HL) is a type of cancer that affects the lymphatic system, where lymphocytes proliferate uncontrollably in multiple lymph nodes and eventually in extranodal sites (e.g. spleen, bone, etc.). It is acknowledged as a curable disease thanks to its high rate of response to chemotherapy, often combined with radiotherapy. Still, a considerable percentage of patients do not respond to first-line treatments and the latest research has been devoting its efforts to discovering alternative and more efficient therapies, such as immunotherapy. Immunotherapy has indeed been approved for relapsing cases and has since represented a huge stride for patients, who are on average very young¹.

As the number of available therapies increases, treatment planning becomes more and more crucial, and personalized medicine is catching on in every aspect of medical practice to devise the optimal treatment for each patient. Nevertheless, such a tailored approach requires quantitative and informative data to input into powerful and transferrable models on which to rely decisions. On purpose, Positron Emission Tomography/Computed Tomography (PET/CT) radiomic analysis has been shown to be an insightful, non-invasive tool for histological prediction, prognostic assessment, and bone marrow involvement definition in Lymphoma². In brief, the radiomics framework entails the extraction of a high-dimensional vector description of the spatial gray levels' distribution of an image, the so-called radiomic features^{3,4}. Each of such features thus describes a statistical property of the image heterogeneity at different scales, which can inform several downstream analyses and modeling efforts.

As HL is a rare disease, studies performed at a single institution usually do not account for sufficient information to build

powerful enough models and derive general knowledge. Therefore, oftentimes multi-center cohorts need to be set up and large-scale studies have to be conducted, collecting data coming from different sources⁵. This raises a relevant issue, as radiomics features are known to be highly influenced by the image acquisition settings, the segmentation procedures, and the reconstruction parameters, jeopardizing the transferability and scalability of the results⁶⁻⁸. Typical exogenous confounding factors include both scanner characteristics, protocols and more general center-specific variabilities. These two factors must therefore be accounted for together when performing any type of analysis on multi-center data.

Moreover, the latest trend in radiomics is developing towards the extraction of more and more features, including first-order statistics, second- and higher-order statistics, and wavelet/frequency-derived indices. As the number of features rises, their pairwise correlation increases accordingly, and it becomes harder and harder to build effective models and disentangle the true signals of interest from technical artifacts, noise, and uninteresting biological variables. Here comes the need to properly reduce the dimensionality of radiomics vectors, transforming the features into low-dimensional vectors that keep the true informative signals while discarding domain-specific confounders.

While the above holds for many multi-center radiomics studies of (rare) diseases, when analyzing a hematological (like HL) or metastatic cancer, an additional level of complexity is added to the task of deconfounding and reducing radiomics features. In fact, different lesions can be found throughout the body of the patients. Despite the current approach for imaging-based quantitative assessment of most cancers, including HL, relies on the inspection of the bigger or hotter lesion, Sollini et al.⁹ have demonstrated how lesions are radiologically heterogeneous within patients in terms of radiomics description and how a prognostic classifier performs better when all tracer-avid lesions are considered. These findings align with the latest discoveries in the biological underpinnings of lymphomas. Some studies on solid cancers have previously described how both proximal and distant lesions deriving from the same primary tumors exhibit divergent patterns of both morphological and genetic heterogeneity¹⁰. Similarly, Tabanelli et al.¹¹ reported the same evolutionary crossroad between morphological heterogeneity and intra-clonal evolution in a case of high-grade B-cell lymphoma. Thus, morphological heterogeneity behaves as a surrogate of genetic heterogeneity, responsible for treatment inefficacies. It follows that all lesions' morphology must be taken into account, to exhaustively represent the disease in the prediction of cancer progression, therapy efficacy, and disease-free survival outcomes^{12,13}. This implies that any postprocessing (i.e. dimensionality reduction and/or deconfusion process) aimed at preparing radiomics features for patients' representations needs to keep the inter-lesion relationships within patients consistent, as here is where information of tumor morphological heterogeneity lies^{9,14}.

In light of the above, a robust post-image-acquisition method aimed to harmonize multi-lesion radiomics data from multi-center studies requires (i) to properly remove both scanner and center confounding effects, (ii) to treat features' collinearity and allow for simpler statistical modeling via proper dimensionality reduction, and (iii) to keep intra-patient heterogeneity consistent throughout the transformation. All this should be achieved while retaining all truly informative signals in the data, so as not to affect – and possibly improve - any potential downstream analysis.

Different strategies have been proposed in recent literature to minimize the batch-effects of radiomics variability, ranging from imaging-based to feature-based approaches¹⁵⁻¹⁷. Most of them aim to perform batch-specific standardization of images to disentangle the true signal from environment-related noise. Among these, the ComBat method was shown to be superior to other techniques, attracting attention in the radiomics field^{15,18,19}. Starting from its first conception, ComBat was improved over time by different independent researchers. One for all, Adamer et al. proposed a regularized solution of ComBat, namely ReComBat, computationally more efficient to facilitate the large-scale harmonization of data²⁰. However, it must be noted that ComBat and most of its derivative algorithms were developed in the computational biology domain, where usually only one main confounder (i.e. sequencing batch effect) needs to be removed. Indeed, to remove multiple confounders, they must be applied repeatedly, one factor at a time. As the context of radiomics studies oftentimes implies multiple confounders, Nested ComBat²¹ and its improved evolution from the same authors, OPNested Combat²², were recently proposed specifically to tackle multi-factor deconfusion. The latter applies ComBat iteratively on confounder-associated subsets of features, identifying the optimal order of factors to correct for. Notably, irrespective of the number of confounders removed from the data, ComBat-based methods rely on the hypothesis of normality of the features' errors, which might be unrealistic for radiomics data²². Moreover, none of the above methods perform dimensionality reduction and are thus typically followed by Principal Component Analysis (PCA) before the analysis. Additionally, to the best of our knowledge, none of them has neither explicitly addressed the problem of preserving inter-lesion relationships within patients, nor has been evaluated in their capability to improve prediction by exploiting heterogeneity information after deconfusion.

In this work, we propose a multi-factor deconfusion algorithm better suitable for the downstream analysis of multi-lesion/metastatic patients in multi-center studies, described in Figure 1. The algorithm builds upon the work of Dincer et al.²³, which, in the context of gene expression analysis, proposes an Adversarial Deconfounding AutoEncoder (AD-AE) model that requires no assumption on features' distribution and jointly performs dimensionality reduction and cleaning of the embeddings, enhancing the signal-to-noise ratio. Here, we exploit the rationale of this model for the context of multi-center PET/CT radiomics analysis, developing a Dual factor AD-AE (in the following, Dual AD-AE) model for the simultaneous removal of

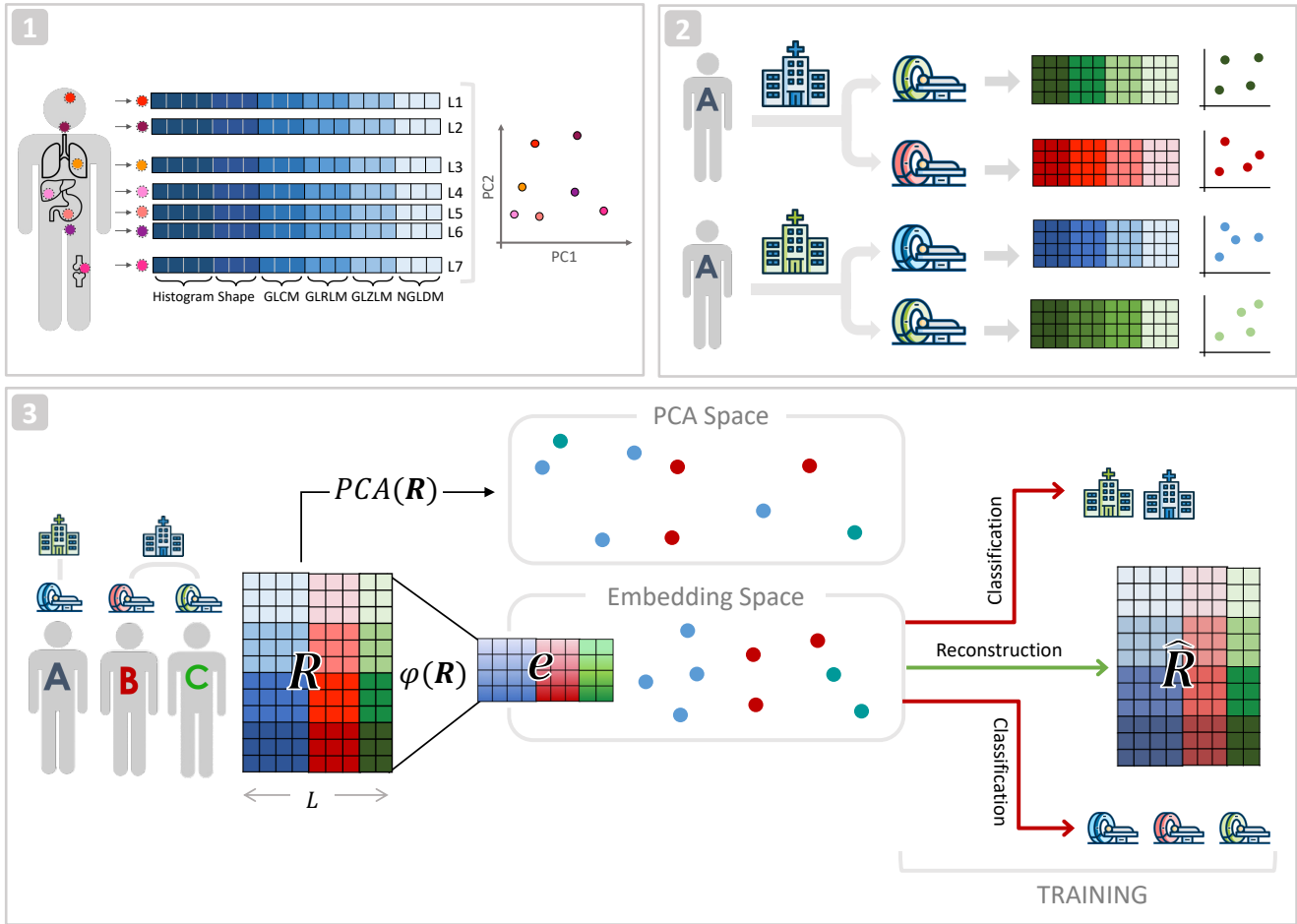


Figure 1. Graphical schema of multi-lesion multi-center radiomics studies confounding issues and Dual AD-AE solution: (Panel 1) In multi-lesion and/or metastatic tumor settings such as HL, patients can be modelled as clouds of points, where each point (L) is defined by the radiomic vector of a lesion (radiomic features can be semantically divided into histogram-derived, shape-derived, GLCM-derived, GLRLM-derived, GLZLM-derived and NGLDM-derived features). Among others, cloud heterogeneity is a valuable predictor of patients' outcome. (Panel 2) The radiomics features computed for the same patient (Patient A) can be highly affected by the center that collects the images and the scanner used, which are the two primary confounding factors. In turn, this affects the patient's cloud representation, biasing heterogeneity-based prediction and making patients with different center-scanner combinations hardly comparable. (Panel 3) To harmonize data from multi-center studies (matrix R of confounded radiomics features), Dual AD-AE embeds lesion into a lower dimensional space (Embedding Space) where exogenous confounding factors are removed and patients' clouds keep the predictive information – as opposed to directly performing PCA on the combined dataset, which results in biased patients' representations (Panel 3 top). To do that, the encoder of the Dual AD-AE (ϕ) is trained to transform the radiomics matrix R into the embedding matrix e by maximizing the reconstruction (from embeddings to the reconstructed version of the input \hat{R}) and simultaneously minimizing the prediction on both center and scanner confounding factors (Panel 3 right).

both center and scanner confounding effects (Figure 1). We evaluate the proposed model in terms of (1) its deconfusion power, (2) its ability to keep invariance of intra-lesion relationship with respect to original data - despite dimensionality reduction - (3) and its prognostic power. In experiments (1) and (3) we compare the results of Dual AD-AE to those of state-of-the-art ComBat-based approaches. In experiment (2), we propose a statistical test to access the consistency of the data transformation. We evaluate our proposed models on a multi-center dataset of HL patients in order to predict response to first-line chemotherapy, demonstrating that Dual AD-AE enables building exhaustive patient representations and delivering more accurate analyses, especially when trying to exploit the predictive power of intra-tumor heterogeneity.

Results

Data collection

Two centers were involved in the study; inclusion criteria were age ≥ 16 years old, newly diagnosed stage I-IV HL and baseline [^{18}F]FDG-PET/CT availability, and exclusion criteria were missing clinical/imaging/follow-up data; 128 HL patients were recruited and treated at IRCCS Humanitas Research Hospital (Institution 1), 78 at Fondazione IRCCS Istituto Nazionale dei Tumori (Institution 2). Personal information and clinical data were annotated for each patient in both hospitals and [^{18}F]FDG PET/CT imaging was inspected by experienced nuclear medicine physicians. Descriptive statistics of patients are available in Supplementary Table S1 and Supplementary Table S2 for Institution 1 and Supplementary Table S3 and Supplementary Table S4 for Institution 2. Of note, number of relapsing patients was 21 over 128 (16%) in Institution 1 dataset and 17 over 78 (22%) in Institution 2 dataset. All [^{18}F]FDG-avid lesions bigger than 64 voxels were segmented in each patient and radiomic features were extracted from each lesion using LIFEx software (www.lifexsoft.org²⁴). A total of 1340 and 794 lesions were collected at Institution 1 and Institution 2, respectively. Information about scanners' specification and acquisition settings is summarized in Supplementary Table S5 and Supplementary Table S6, while Imaging Biomarker Standardization Initiative (ISBI)-compliant standardization and data harmonization have been published elsewhere⁹. The study was approved by the local ethics committees at Institution 1 (n. 2595 on Jun16, 2020) and Institution 2 (code INT 212/20 on Sep28, 2020); given the observational retrospective design of the study, the signature of a specific informed consent was waived.

Experimental outline

As displayed in Figure 2 and further explained in the Methods, the proposed Dual AD-AE consists of (i) an autoencoder with multiple hidden layers and (ii) two adversary branches that predict the source of data, i.e., the center and the scanner. The rationale of this method is that penalizing the prediction performance of the adversaries while jointly maximizing the reconstruction accuracy of the autoencoder will result in lesion embeddings that keep as much as possible of the original signal while discarding solely the noise introduced exogenously by the two confounders.

As mentioned in the Introduction, to identify a robust post-image-acquisition method to harmonize multi-lesion radiomics data from multi-center studies, one needs to consider several aspects. Indeed, to propose Dual AD-AE as better suited to the task, we performed a series of experiments on different harmonization strategies (in the following, *modalities*).

We recall that the center confounding factor relates to the hospital's imaging facility, the clinical guidelines, and the personnel who segments and carries out the acquisition. On the other hand, the scanner confounding factor supplies information on the scanners' specifications and reconstruction parameters. The scanner variable may be intrinsically subordinated to the center variable, as usually different scanners are found in different centers. They may thus entail some extent of nesting nature and partially overlap in their confounding information.

For the sake of comparison with state-of-the-art approaches, we tested three major ComBat implementations, namely ComBat¹⁸, ReComBat²⁰ and OPNested ComBat²², comparing the results to quantify the improvements of our solution. Specifically, single-factor ComBat was applied twice in cascade (in both confounders' orders). These two ComBat-based models are namely ComBat-center-scanner and ComBat-scanner-center depending on the order of the batch-effects. The very same approach has been followed for ReComBat. OPNested ComBat was instead applied once on center and scanner effects at the same time, as it was specifically developed for multi-factor effect removal.

On these models and ours, we performed three different quantitative experiments. We tested the deconfusion power of the different modalities, comparing the proposed method to the state-of-the-art models (Experiment 1). Furthermore, considering that the Dual AD-AE encompasses dimensionality reduction as part of the deconfusion process, leading to a potentially detrimental transformation of intra-lesion relationships, we developed a novel test to assess this impact quantitatively (Experiment 2). Finally, we tested and compared all modalities on their ability to keep predictive information intact. We transformed the deconfounded features of each modality into different all-lesions patients' representations, to be fed into prognostic models and evaluated the performance of prognostic models (Experiment 3).

Experiment 1: Checking deconfusion power

To evaluate the strength of the confounders' effect, one can verify the predictability of the confounding variables (i.e. the center and the scanner) from the data. A high prediction performance denotes the presence of a strong confounder-related signal. Therefore, in order to quantify the effect of the deconfusion process, we compared the predictive power of cross-validated Logistic Regression models fed with the radiomics features before and after the application of the different modalities. Details on the analysis are provided in the Methods section. Accuracy was annotated for performance comparison through statistical tests. Table 1 shows the results for the Dual AD-AE, the two ComBat models, the two ReComBat models, and OPNested ComBat.

While radiomics, as expected, scored very high in predicting both the center and the scanner (as assessed by the univariate analysis in Supplementary Table 7), our embeddings showed evidence of deconfusion, comparable to state-of-the-art benchmarks.

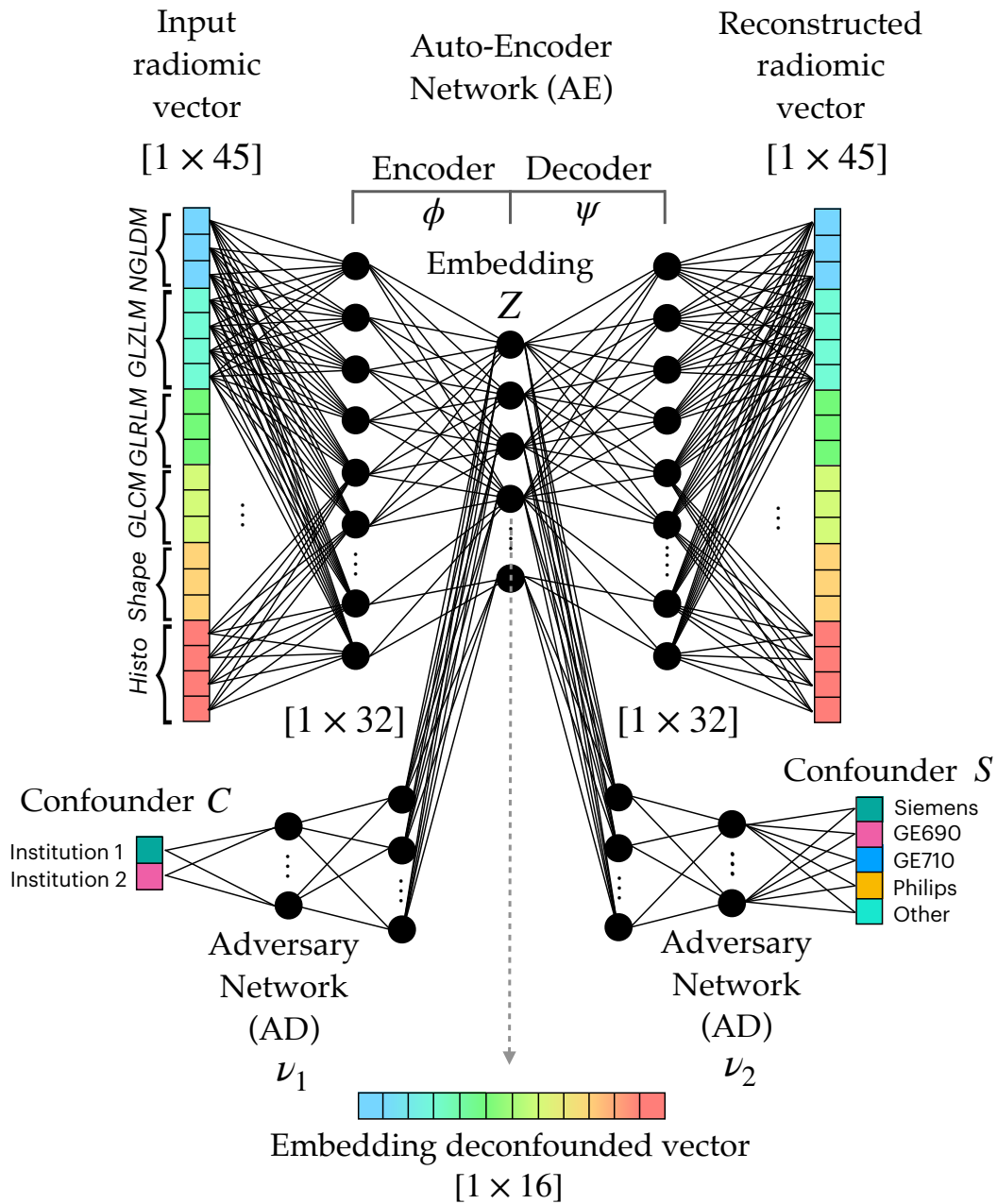


Figure 2. Architecture of the Dual Adversarial Deconfounding Autoencoder (AD-AE) model: the model is made of three parts: an autoencoder (encoder: ϕ , decoder: ψ), an adversary branch network predicting the center confounder (ν_1) and a parallel adversary branch network predicting the scanner confounder (ν_2). The network is trained by optimizing the input reconstruction task (autoencoder loss) and the deconfusion task (adversary losses) as in Equation 1 of Methods. The adversaries unlearn to predict the confounding factors, i.e. the center and the scanner.

Table 1. Experiment 1 results: comparison between the performance of the Logistic Regression models for predicting the two confounding factors: the center and the scanners. The modalities that have been evaluated are raw radiomic data, Dual AD-AE embedding, ComBat-standardized data (both with center-scanner order and with scanner-center order), ReComBat-standardized data (both with center-scanner order and with scanner-center order) and OPNested-standardized data (with scanner-center order). The Logistics Regression models are fitted on each of these modalities, in a cross-validated fashion. Values are annotated as mean \pm standard deviation. The models evaluate (1) the binary prediction of the center labels, and (2) the multi-class prediction of the scanner labels. The performances of the radiomics-based models are taken as reference, while the performances of the other modalities are analyzed in terms of decrease compared to the baseline models’ performance. Statistical tests have been performed and the models that are significantly different from radiomics are highlighted in bold.

		Accuracy CENTER	Accuracy SCANNER
Baseline	Radiomics	0.8559 \pm 0.0117	0.8617 \pm 0.0104
Embedding	Dual AD-AE	0.6251 \pm 0.0131	0.3308 \pm 0.0146
ComBat	ComBat-center-scanner	0.6220 \pm 0.0137	0.2968 \pm 0.0278
	ComBat-scanner-center	0.6236 \pm 0.0134	0.3006 \pm 0.0307
ReComBat	ReComBat-center-scanner	0.6228 \pm 0.0150	0.3009 \pm 0.0362
	ReComBat-scanner-center	0.6276 \pm 0.0124	0.2997 \pm 0.0359
Opnested	Opnested (scanner-center)	0.6239 \pm 0.0118	0.2967 \pm 0.0349

Both the Dual AD-AE and all the Combat-based modalities aligned to the same performance, outperforming the non-deconfounded radiomic vectors. Indeed, values highlighted in bold in Table 1 correspond to non-significantly different, yet lower than radiomics, performances. All modalities were thus equally powerful at the deconfusion task. Of note, the OPNested algorithm selected scanner-center as the optimal order, thus the two models are expected to perform similarly. Additionally, the proposed Dual AD-AE model showed a smaller standard deviation of the accuracy in predicting scanner type, supporting the robustness of the model.

Experiment 2: Cloud-shape invariance test

In multi-lesion and/or metastatic tumor settings such as HL, patients can be modeled as clouds of points¹⁴, where each point is defined by the radiomic vector – whether original, reduced, or deconfounded – of a lesion, and the shape of the cloud determines intra-patient tumor heterogeneity as the pairwise relationship between lesions²⁵. To ensure that the predictive information of the clouds’ shape has been preserved, Dual AD-AE embeddings must keep invariance with respect to the relative positions of lesions, despite the reduced dimensionality of the resulting vectors. That is, patient-wise lesions’ rankings and pairwise lesions’ distances should hold after deconfusion, under the hypothesis that they are not independently impacted by exogenous noise. Given this assumption, to test for the cloud shape-invariance of the feature transformations, we developed a quantitative method, introducing novel metrics called Point Cloud Semantic Drift (PCSD). As further detailed in the Methods section, PCSD quantifies the extent of the change in peer lesions’ distance rank order within a patient. Furthermore, to define a quantitative test of hypothesis for PCSD, we estimated an empirical null distribution of the PCSD values when point clouds are randomly transformed, inducing random neighbor swaps by injecting repeatedly Gaussian noise in subsets of the embeddings. The empirical p-value of the Dual ADAE transformation was then obtained from the Empirical Cumulative Distribution Function (ECDF) of this null PCSD distribution.

Figure 3 shows the results of the proposed method. The population distribution of PDSC from Dual AD-AE transformation is displayed alongside the Empirical Distribution Function (EDF) of 100 random transformations. From the visual inspection of the plots, the model produced PCSD values skewed toward zero, suggesting the shape-invariance of the clouds. Moreover, the empirical p-value was equal to zero, thus we can further sustain that Dual AD-AE successfully kept cloud-shape invariance and that the change in inter-lesion distance, which occurred during deconfusion, was significantly far from being random.

Experiment 3: Checking prognostic power

Despite its unsupervised nature, the proposed approach aims to enable the design of exhaustive patient representations to deliver accurate analyses for treatment planning on multi-center datasets. Here we provide an example of downstream analysis where to quantify the improvement in predicting the first-line chemotherapy outcome of patients affected by HL after correcting for confounding factors. To do this, we resorted to the use of three patients’ representations, encompassing both separately and jointly the location and the shape of the point clouds. In particular, we represented each patient as a point cloud and defined (1) a mean vector of all lesions of each patient (i.e. the centroid of the cloud), (2) a set of topological indexes describing

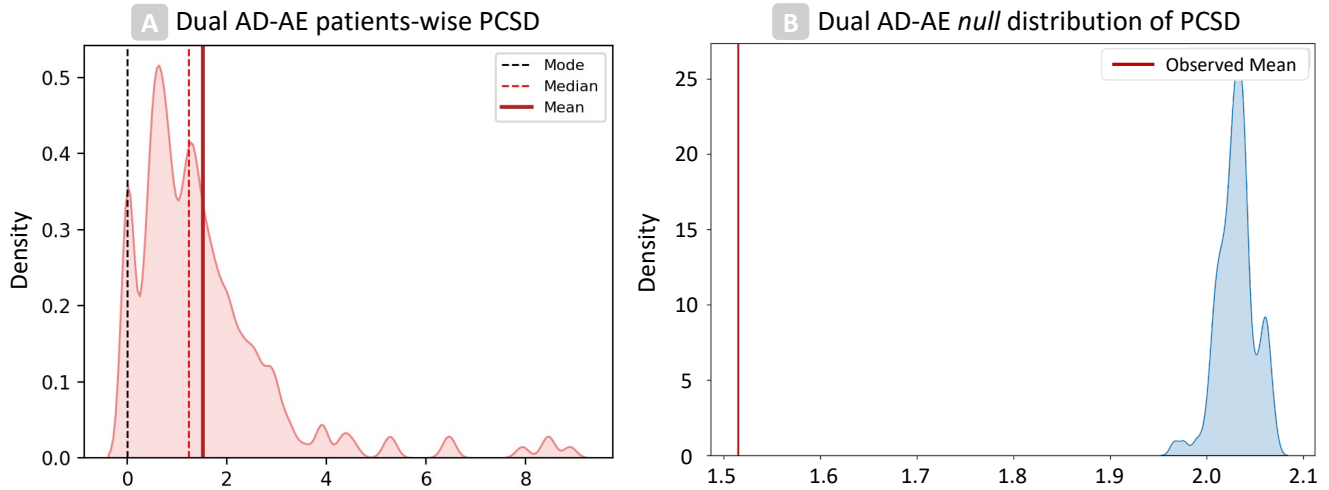


Figure 3. Results of patient-wise tests on PCSD for dual AD-AE embedding: (Panel A) The density plot displays the overall distribution of PCSDs in the population. (Panel B) The density plot shows the score of the Dual AD-AE results over a bootstrap random distribution. Fiducial values of the distribution are marked with vertical lines in the left plot and our model performances are displayed with a vertical red line in the right plot.

Table 2. Experiment 3 results: p-values of the tests comparing different survival models. The Dual AD-AE model is considered according to its three different patient representations: (1) the patient is described by the centroid of their point cloud, (2) the patient is described by the topological characteristics of their point cloud and (3) the patient is described by both the centroid and the topological characteristics of their point cloud. These are compared with the radiomics-based models, the ComBat-based models, the ReComBat-based models, and with the OPNested-based models. Comparisons are made upon the same patient representation: for instance, the Dual AD-AE model fed with centroid representation is compared to the other modalities which were fed with centroid representation as well, and so on. Significant values are highlighted in bold.

Comparison wrt	Centroid		Cloud description		Centroid + cloud description	
	P-value (train)	P-value (test)	P-value (train)	P-value (test)	P-value (train)	P-value (test)
Radiomics model	<<0.001	<<0.001	<<0.001	0.0001	<<0.001	<<0.001
ComBat-center-scanner model	<<0.001	0.1065	<<0.001	0.0003	<<0.001	0.0230
ComBat-scanner-center model	<<0.001	0.0473	<<0.001	0.0002	<<0.001	0.0160
ReComBat-center-scanner model	<<0.001	0.1381	<<0.001	0.0005	<<0.001	0.0368
ReComBat-scanner-center model	<<0.001	0.0676	<<0.001	0.0003	<<0.001	0.0274
OPNested ComBat model	<<0.001	0.0676	<<0.001	0.0003	<<0.001	0.0274

the structure/shape of the clouds (i.e. the mean and the standard deviation of the pairwise distances between lesions and the mean and the standard deviation of the distances between lesions and the cloud centroid) and (3) a representation including both the centroid vector and the cloud describing indexes. Further details are provided in the Methods section. These three representations were constructed from original radiomics features and their deconfounded versions with all the considered modalities (Dual AD-AE embeddings; ComBat-, ReComBat- and OPNested ComBat-transformed radiomics). Each of them was fed into a Cox proportional hazard model²⁶ to predict the time-varying response to therapy. Of note, vectors derived from the state-of-the-art algorithms needed to be reduced by PCA prior to being input into any model. The significance of the prognostic power of imaging information was assessed in terms of the Concordance Index (CI,²⁷). Performances on both the training and testing phases were produced by repeated sampling of 20 independent data splits. Supplementary Table S7 reports the means and standard deviations of the trials. For visual reference, Figure 4 displays the boxplots of the distributions of the performance indexes of the modalities, grouped by patients' representation strategy and deconfusion approach. Pairwise tests were performed between settings to be compared and can be appreciated in Table 2.

As displayed in Table 2, the model performance of the Dual AD-AE modality was significantly higher than radiomics', suggesting how the deconfusion step does also benefit the prediction and the signal-to-noise ratio. Of note, the patient representations including cloud topology descriptors (i.e. when using heterogeneity as a predictor) always achieved better performance than the benchmarks, being the most predictive and generalizable (i.e. test set performance) overall. From what centroid representation is concerned, ComBat-center-scanner, ReComBat-center-scanner, ReComBat-scanner-center, and OPNested

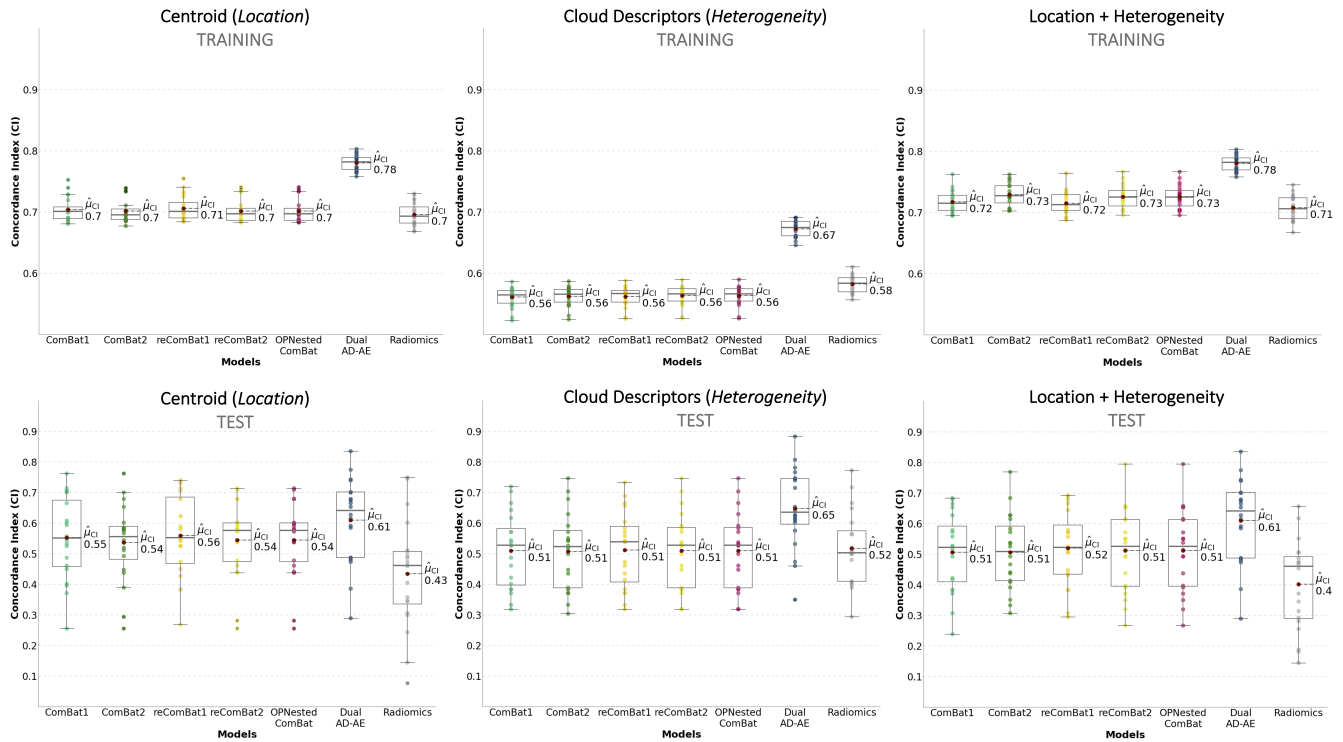


Figure 4. Experiment 3 results: the boxplots of the distributions of algorithms’ performances. The three different patients’ representation strategies are considered per modality and one representation is displayed per plot. The top row plots show training results while the bottom row plots show testing performances. All plots report on the y-axis the CIs of ComBat-center-scanner radiomics (light green, ComBat1 for short), ComBat-scanner-center radiomics (dark green, ComBat2 for short), ReComBat-center-scanner radiomics (dark yellow, reComBat1 for short), ReComBat-scanner-center radiomics (light yellow, reComBat2 for short), and OPNested ComBat radiomics (fuchsia), dual AD-AE embeddings (blue) and original radiomics (grey). The red dots highlight the mean CIs, which are also reported on the right of each respective boxplot.

ComBat scored similarly yet worse with respect to Dual AD-AE. We remind that the OPNested algorithm implemented the same sequence of ComBat-scanner-center, however, this latter model had lower performance, being significantly outperformed by our model.

Alternative to deconfusion: frailty Cox proportional hazards model

Deconfusion methods ultimately allow the effective modeling of patients’ representations across scanners and centers. However, instead of removing the confounding factor, an alternative yet well-established approach to model multi-source samples (i.e. multi-center/multi-scanner data, where we have dependence within groups) is the explicit modeling of the group-specific variability within the prediction model. For time-to-event data, this can be done via the frailty Cox proportional hazards model²⁸, which estimates center and/or scanner random effects together with the baseline hazard function. To verify whether this approach would make the deconfusion step irrelevant, the centroid representation derived from the raw radiomics features was reduced by PCA and fed into a frailty Cox proportional hazard model with center-specific and scanner-specific random intercepts. Unfortunately, this test dramatically failed due to a lack of model convergence. This result is motivated by the small sample size of the data at hand, combined with the high dimensionality of the radiomic variable (even after PCA) and the large number of censored patients, which did not allow the model to properly estimate the effects’ parameters neither on the training sets, on the testing sets, nor on the dataset as a whole.

Discussion

In this work, we developed a deconfusion algorithm to harmonize multi-center imaging data, with a particular focus on multi-lesion/metastatic cancers, like Hodgkin Lymphoma. The Dual AD-AE model performed dimensionality reduction of radiomic features while removing center- and scanner-related information simultaneously. The proposed approach was trained on a dataset of Hodgkin Lymphoma patients from two centers and outperformed the state-of-the-art methods in the task of

radiomic features harmonization, leading to higher prediction of response to first-line chemotherapy.

Three experiments were performed to evaluate the model's properties, raising some major points of discussion. First, the deconfounding power of the Dual AD-AE was granted. In fact, the accuracy of Logistic Regression models predicting the scanner and the center target variable sensibly decreased after deconfusion. The Dual AD-AE demonstrated a comparable deconfusion power with respect to ComBat-based models, showing no statistical differences in cross-validation. However, removing both confounding factors at the same time may uncover and discard inter-confounder relationships which may contribute to undesirable noise in the signal. Interestingly, the standard deviation of the accuracy of the Dual AD-AE model in predicting the scanner type was lower than other models, suggesting the robustness and stability of the proposed model. The ComBat (and ReComBat) algorithm applied twice showed variable results when changing the order of application. This inconsistency is not surprising, as it motivated the development of OPNested ComBat in the first place^{21,22}. In fact, despite the slight algorithmic differences between ComBat and OPNested, OPNested performed very similarly to ComBat-scanner-center.

Additionally, as the context of multi-lesion/metastatic data may benefit from the exploitation of intra-tumor heterogeneity as predictive information, we designed a novel metric (i.e. PCSD) and an associated empirical test to quantify the impact of the Dual AD-AE deconfusion and dimensionality reduction on intra-lesion relationships shaping the spatial conformation of patients' point clouds. Overall, the Dual AD-AE resulted in a significantly low PCSD value, rejecting the null hypothesis of no correlation between the original (raw) and the deconfounded clouds of lesions. On one hand, this was expected and desired as lesions of one patient share both the same center and scanner variability. That is, noise can be considered constant within a single patient and the relationship among peer lesions should in principle not be spoiled by center and scanner deconfusion. On the other hand, it might be possible that minor shifts could be appreciated in specific lesions, especially where massive non-linear transformations were needed to properly clear the data. This might be true for some patients lying on the far-right tail of the PCSD distribution. As proved by the test, such results do not translate into a detrimental data transformation, rather they show that a trade-off between deconfusion and cloud-shape invariance has to be tuned and rigorously assessed. On purpose, the PCSD metric can be exploited to highlight the presence of such additional sources of latent and interactive noise, that once removed would release the true predictive power of intra-lesion heterogeneity.

This point was further validated in the third experiment presented in this work, where we assessed the increase in the prognostic power of the deconfounded representation of patients in terms of response to therapy, against ComBat-based alternative approaches. In principle, a proper deconfusion allows the shape and location of the point clouds coming from different sources to be meaningfully compared. Thus, one can expect that predictive models built on these clouds' representation, that is lesions' characteristics and intra-tumor heterogeneity, benefit from the deconfusion process. In fact, Dual AD-AE embeddings showed significant improvements with respect to the baseline and the benchmarks, even though the gap between training performance and testing performance remains large and would necessitate some prevention strategies to overfitting. The results however testify how the proposed model can identify and remove the complex and potentially non-linear portion of confounders' noise that the competitors ignore. Moreover, it demonstrates the relevance of removing all confounders simultaneously when in presence of multiple factors of variability in the data.

A further particularly relevant result was the difference in performance when using heterogeneity (i.e. cloud describing indexes) as a predictor. While this cloud shape representation was merely a simple proof-of-concept example, Dual AD-AE embedding was seen to allow for a much better prediction than the baseline model and competitors. Conversely, ComBat-based and ReComBat-based benchmarks seemed to corrupt the heterogeneity signal to the point of achieving lower CI than the original radiomics features during training, and they grant just a very limited performance increase during testing. Additionally, to the best of our knowledge, none of the previous studies comparing deconfusion algorithms for radiomics data¹⁵⁻¹⁷ evaluated their impact on the predictive power of groups of lesions. Here, our proposed approach was the only deconfounding algorithm truly releasing the predictive power of heterogeneity, which became the most generalizable predictor.

This finding leads to two relevant considerations. Clinically speaking, it supports the hypothesis that intra-lesion heterogeneity does carry predictive information, once properly corrected for linear and non-linear confounders. Technically, it endorses the use of a more complex non-linear model like the AD-AE, that can uncover and remove explicit and latent types of noise effectively. Although not explicitly enforcing inter-lesion relationships consistency in the model we propose, so that it could be in principle applied as-is to single-lesion data, this result testifies in favor of its application (as opposed to the state-of-the-art) to contexts in which heterogeneity information is crucial for prediction.

Of course, training complex, non-linear, and heavily parametrized models such as the Dual AD-AE has higher computational, time, and memory demands compared to the simpler ComBat-based methods. Nevertheless, the latter algorithms rely on Gaussian distribution assumptions for estimating the parametric definitions of the statistical moments across batches (i.e. the mean and the variance across centers or scanners), prior to standardization. However, this strong hypothesis of underlying data structure may not always be appropriate for radiomics data, leading to underpowered and biased transformations. Conversely, we proposed a non-parametric algorithm removing linear and non-linear confounder-induced noise without any prior assumption. Furthermore, the Dual AD-AE was the only method that dealt with two confounders simultaneously. This permitted to reduce

the risk of ignoring the portion of noise induced by center and scanner interactions (for instance, if one center uses way more frequently a set of parameters for a specific scanner, compared to other centers). Moreover, thanks to its modular nature, one could easily extend the model to adversarially predict - that is, unlearn - more than two confounders. In fact, additional branches could be added, and the overall loss might be updated with the maximization of the corresponding accuracies. Further, the weighting parameters λ_i (with i being the number of adversary branches) enable defining the impact of each confounder, rebalancing the expected (or measured) relative effect of noising factors on the data. Both these aspects could hardly be integrated into the ComBat approach. Finally, as opposed to ComBat-based methods, Dual AD-AE performs dimensionality reduction together with cleaning of the embeddings. While this may affect the interpretability of the deconfounded data, we argue that radiomics features are not easily interpretable *per se*, and they usually need a dimensionality reduction (such as PCA) before modeling, as they are highly collinear.

As a final remark, disregarding the deconfounding algorithm employed, the two-step pipeline of removing confounding effects and then analyzing the corrected data has raised several critiques^{29,30}. Oppositely, the most sponsored solution when the confounder information is available is including it within the final prediction model. Nevertheless, we have shown in our last tentative experiment how a frailty CoxPH model (even if with only one confounder) does not converge when the sample size is small and the number of censored patients is high. This is quite common in multi-center studies of rare diseases.

A limitation of the present study is the lack of further data to test our proposed approach. However, no additional comparable data was available to the authors at the time of writing. Nevertheless, we believe that the comprehensive tests and benchmark studies performed on these cohorts represent a valuable proof-of-concept of the method’s potential. Moreover, data provenance may behave as a bias in our experiments. Yet, despite the data originated from two geographically close hospitals with standardized procedures and consistent image acquisition and feature extraction protocols, inherent heterogeneity and discrepancies in data values persisted, supporting the objective of implementing a feature-level harmonization³¹. This emphasizes the need for a comprehensive approach that combines image standardization, post-processing, and harmonization models to eliminate batch effects and achieve data consistency.

In conclusion, we provided a modular and effective approach for harmonizing imaging data coming from different sources. We proved that our approach could efficiently correct for multiple batch-related differences so that data appear as if they were acquired under a common set of conditions. This translates to higher prognostic performances, above all for what regards intra-tumor heterogeneity of multi-lesions/metastatic cancers. As it is well known that NN models such as the Dual AD-AE can benefit from Transfer Learning³² to aid the problem of suboptimal and/or overfitting parameters when training data is limited, we provide a tutorial to apply our method to new data, available on [GitHub](#). We currently share the weights of our pre-trained network on this study’s cohorts. Researchers might thus decide to use such weights to pre-train their Dual AD-AE model, “borrowing” information from additional samples without privacy concerns. This model-sharing framework could be pushed forward with the contribution of the scientific community sharing their fine-tuned parameters, paving the way for a virtuous cycle of open science. Insightful knowledge could be thus derived from more exhaustive models to optimally impact the clinical practice.

Methods

Dual Adversarial Deconfounding Autoencoders

Dual Adversarial Deconfounding AutoEncoder (Dual AD-AE) jointly tackles the denoising from both center- and scanner-related information. The architecture of the Dual AD-AE is described in Figure 2. The network consists of two parts: one autoencoder and an adversary branch. The autoencoder takes as input the radiomic vector associated with a lesion and performs the dimensionality reduction. It is made of one input layer (number of input nodes: $[1 \times 45]$) two hidden layers (number of first hidden layer nodes: $[1 \times 32]$, number of second hidden layer nodes: $[1 \times 16]$), and one output layer (number of output nodes: $[1 \times 45]$). The autoencoder represents the backbone of the model and, from its deepest layer, two adversary networks branch out for center and scanner predictions. Both adversary networks are made of two hidden layers (dimensions of the first hidden layer and the second hidden layer are $[1 \times 50]$ and $[1 \times 50]$ respectively) and one output layer ($[1 \times 2]$ for center prediction and $[1 \times 5]$ for the branch predicting the scanners).

The loss is then made of three terms, where the reconstruction error, the accuracy of the center classification, and the accuracy of the scanner classification sum up as in Equation 1:

$$\min_{\phi, \psi, \nu} E[|x - g_{\psi}(f_{\phi}(x))|_2^2 - \lambda_1 L(h_{v_1}(x), c) - \lambda_2 L(h_{v_2}(x), s)] \quad (1)$$

where v_1 is the center adversary branch, v_2 is the scanner adversary branch, λ_1 and λ_2 are weighting parameters and c and s are the true labels for center and scanner respectively. Of note, weighting parameters can be tuned to tailor the importance

of the tasks to be optimized. For instance, one could prioritize one confounding factor rather than the other, having *a priori* information about the latent variability of the specific case study data.

In our setting, hyperparameters were tuned according to grid search. The number of layers, the number of nodes, and weighting parameters were optimized based on the reconstruction error. The number of epochs was optimized according to early stopping strategy³³, i.e., iterations were stopped when no relevant improvements of the validation loss were recorded. The batch size was set to 128 and $\lambda_1 = \lambda_2 = 1$.

Benchmark state-of-the-art Methods

Among the methods proposed in the literature for imaging harmonization, ComBat has been repeatedly elected as the best approach such that different implementations and further improvements have been proposed in the last years.

ComBat was originally proposed by Johnson and Rabinovic¹⁸ for removing the batch-effect seen in genetics microarray analysis. The harmonization method consists of standardizing each batch according to its mean and variance. Specifically, the correction takes place at a specific location and scale (L/S), wherein the batch-related error is supposed to be present. L/S model states that the value Y for feature f from a sample j in a batch i follows the following formulation:

$$Y_{ijf} = a_f + X\beta_f + \gamma_{if} + \delta_{if}\epsilon_{ijf} \quad (2)$$

where a_f is the feature value, behaving as intercept; X is the design matrix and β_f is the features coefficients such that $X\beta_f$ is the observed variability; γ_{if} and δ_{if} are the additive and multiplicative batch effects respectively and ϵ_{ijf} the standard error. Accordingly, γ_{if} and δ_{if} can be estimated (either in parametric and non-parametric ways) from data, and Y_{ijf} can be corrected as:

$$Y_{ijf}^* = \frac{Y_{ijf} - \hat{a}_f - X\hat{\beta}_f - \hat{\gamma}_{if}}{\hat{\delta}_{if}} + \hat{a}_f + X\hat{\beta}_f \quad (3)$$

One of the main advantages of ComBat is being effective even with small batch sizes. Being $A = \tilde{X}^T \tilde{X}$ positive-definite, the optimization problem is strictly convex. However, when A happens to be singular the regression estimation does not exist and, if the system is underdetermined, ComBat is not guaranteed to bring out a unique solution. For this reason, Adamer et al²⁰ proposed a regularized solution of ComBat (ReComBat) computationally more efficient to facilitate the large-scale harmonization of data.

As to compare our method with the state of the art, we applied both ComBat and ReComBat models to our case study. We employed different pipelines to test their performance from different perspectives. ComBat was used for deconfounding the imaging data from the center and scanner information. The two ComBat models were applied in cascade to the data: (1) one label was used as a batch effect to be removed and (2) the obtained denoised vector was further deconfounded by the effect of the other label. We followed two different orders, namely ComBat-center-scanner and ComBat-scanner-center. The very same procedure was investigated by employing ReComBat implementation. Two different pipelines were thus derived, namely ReComBat-center-scanner and ReComBat-scanner-center.

As a matter of fact, applying ComBat or ReComBat in cascade to capture and remove the linear variability from more than one confounding factor may cause instabilities depending on the specific order of the harmonization steps. Very recently, Horng et al^{21,22} proposed an optimized procedure for sequentially harmonizing data from multiple batch effects, namely OPNested ComBat. Besides ComBat and ReComBat, we included OPNested as a benchmark model, to be tested in both deconfusion and predictive powers. On one hand, OPNested can show a more effective standardization procedure to compensate for the heterogeneity of diverse data sources, improving the generalization abilities of imaging data. On the other hand, higher harmonization performance might not imply higher predictive performance as it yet remains to be investigated whether and which latent factors have to be removed or smoothed.

Evaluating Point-Cloud shape consistency

We defined and implemented a novel approach to test the point-cloud shape consistency across transformation: the Point Cloud Semantic Drift (PCSD).

Before defining PCSD, let us introduce some necessary notation. Let $M_i(1) \dots M_i(K)$ be the scores associated with the ordered list L_i , where $M_i(1)$ is the best score, $M_i(2)$ is the second best, and so on. The best score can be the largest or the smallest depending on the context. Let $r^{L_i(A)}$ be the rank of A in the list L_i if element A is within the top k elements, and be it equal to $k + 1$ otherwise; $r^\delta(A)$ is defined likewise for a different list δ . The Spearman's footrule distance between L_i and any ordered list δ can be defined as:

$$S(\delta, L_i) = \sum_{t \in L_i \cup \delta} |r^\delta(t) - r^{L_i}(t)|. \quad (4)$$

Equation 4 is the sum of the absolute differences between the ranks of all the unique elements of the union of the two ordered lists. The smaller the value of the metric, the more similar the lists. To compute the Point Semantic Drift (PSD) for an arbitrary point t , we exploit a weighted version of S . We estimate the PSD as the weighted change in neighbor rankings, according to Equation 5.

$$PSD(\delta, L_i) = \sum_{t \in L_i \cup \delta} |M(r^\delta(t)) - M(r^{L_i}(t))| \times |r^\delta(t) - r^{L_i}(t)|. \quad (5)$$

PSD is the sum of penalties for moving an arbitrary element (data point) t of the list L_i from a position $r^\delta(t)$ to another position $r^{L_i}(t)$ within the same list (second term of the product) adjusted by the difference in scores between the two positions (first term). $M(r^\delta(t))$ and $M(r^{L_i}(t))$ are the normalized distances between t and all other points in the cloud, respectively after and prior to any transformation. This weighting scheme penalizes more the changes in the positions of very distant points, than the neighboring shifts of observations lying close in the original cloud. That is, higher weights are assigned to swaps between close-by and far-distant points, compared to changes among close neighbors. Such information can deeply inform the deconfusion-invariance trade-off.

Once computed the PSD for each point in the cloud C , the Point Cloud Semantic Drift is estimated as the average PSD_k of the K points in C :

$$PCSD = \frac{1}{K} \sum_{k=1}^K PSD_k \quad (6)$$

where K is the number of lesions in the patient under consideration.

In our setting, L_i corresponds to the set of lesions of patient i as described by the raw radiomic features (original set); δ corresponds to the set of lesions described by the transformed features after deconfusion (e.g. Dual AD-AE mode). PCSD thus accesses and quantifies the invariance of each cloud (patient) to the data transformation process.

Given that PCSD can take on values ranging from 0 to infinity, we need to establish a suitable test to assess the significance of the results obtained from our deconfounded point clouds. To accomplish this, we can build a null distribution of PCSD values ($PCSD_{null}$) which serves as an upper bound for the drift. That is, it represents the change in the cloud's shape that would occur if an arbitrary embedding function was employed, completely disregarding the initial data structure. Operationally, we randomly transform the original cloud by adding a random Gaussian noise with mean $\mu = 0$ and variance $\sigma = 3$ to a different subset of the lesions' vectors. We do this iteratively 100 times, computing the PCSD each time. Upon these values, we build the Empirical Cumulative Distribution Function. If the true PCSD value obtained from our deconfounded embeddings falls within the limits of the left tail of this empirical null distribution, significant evidence is obtained on the ability of our algorithm to maintain the original cloud structure. The empirical p-value is computed from the Empirical Cumulative Distribution Function by computing the ratio between the number of trials where the PCSD is lower than the computed real value and the total number of trials (100).

Experiments' implementation details

Three tests have been implemented to test for (1) deconfusion power, (2) transformation consistency, and (3) predictive power of the proposed algorithm compared to current literature models.

The deconfusion power has been tested by predicting the confounder(s) using the features under analysis. We employed a cross-validated Logistic Regression model, with 100 trials and replacement. Testing accuracy was annotated in each trial to compute the mean trend and the standard deviation of the performance of each modality. Additionally, to compare the performance of the models, given the normality of the data, we used a two-sided parametric t-test for paired samples and evaluated the improvements of the different harmonization strategies with respect to the pure radiomics description.

The point cloud shape consistency between radiomics data and transformed data was evaluated as described above. The PCSD was computed for each patient and a population test for testing the transformation consistency was carried out in the context of the Dual AD-AE.

Lastly, the predictive power of the imaging features has been evaluated with Cox proportional hazard survival models in a cross-validation fashion. Three patient representation strategies were implemented to summarize multi-lesion information

in a single vector object to be properly fed into the models. First, the centroid of each patient’s point cloud was computed as the mean profile of peer lesions belonging to them (“centroid representation”). Then, as a second patient representation, only the distribution of the lesions over the space was described and used as model input. For each patient, we computed the pairwise distances between all lesions in the patient and we calculated the mean and the standard deviation as an index for lesions’ variability. Moreover, we took the distances between every lesion of the patient and their centroid and kept the average and the standard deviation of these distances to quantify the lesions’ spreading from their center. Thus, the four indexes were exploited as “point cloud description representation” to be fed into the survival model. Finally, the two abovementioned representations were merged in a “complete representation” of the patient encompassing both the mean disease profile of patients and the variability of their lesions. For each of the modalities under testing, the three representations were computed and fed into Cox models. Additionally, raw radiomic, ComBat- and ReComBat-based standardized radiomic features were reduced using PCA. To result in a dimensionality comparable to the embeddings, we kept the first sixteen principal components, accounting for at least 90% of the variability. Training and testing sets are repeatedly split multiple times (20 splits) and c-index scores were reported to assess the improvements that the harmonization step brings in terms of prognostic power. To do this, given the normality of the data, one-sided parametric t-tests for paired samples were employed to establish the optimal harmonization strategy. Specifically, the Dual AD-AE embeddings’ performance was compared to ComBat-center-scanner, ComBat-scanner-center, ReComBat-center-scanner, ReComBat-scanner-center, and OPNested ComBat.

References

1. Mohty, R. *et al.* Latest advances in the management of classical hodgkin lymphoma: The era of novel therapies. *Blood Cancer J.* **11**, 126 (2021).
2. Rizzo, A. *et al.* The role of 18 f-fdg pet/ct radiomics in lymphoma. *Clin. Transl. Imaging* 1–10 (2021).
3. Afshar, P., Mohammadi, A., Plataniotis, K. N., Oikonomou, A. & Benali, H. From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. *IEEE Signal Process. Mag.* **36**, 132–160 (2019).
4. Scapicchio, C. *et al.* A deep look into radiomics. *La radiologia medica* **126**, 1296–1311 (2021).
5. Parmar, C., Barry, J. D., Hosny, A., Quackenbush, J. & Aerts, H. J. Data analysis strategies in medical imaging data science designs in medical imaging. *Clin. cancer research* **24**, 3492–3499 (2018).
6. Berenguer, R. *et al.* Radiomics of ct features may be nonreproducible and redundant: influence of ct acquisition parameters. *Radiology* **288**, 407–415 (2018).
7. Pavic, M. *et al.* Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol.* **57**, 1070–1074 (2018).
8. Pati, S. *et al.* Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the ivy glioblastoma atlas project (ivy gap) dataset. *Med. physics* **47**, 6039–6052 (2020).
9. Sollini, M. *et al.* Methodological framework for radiomics applications in hodgkin’s lymphoma. *Eur. journal hybrid imaging* **4**, 1–17 (2020).
10. Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci.* **105**, 13081–13086 (2008).
11. Tabanelli, V. *et al.* Evolutionary crossroads: morphological heterogeneity reflects divergent intra-clonal evolution in a case of high-grade b-cell lymphoma. *Haematologica* **105**, e432 (2020).
12. Sangaletti, S. *et al.* Intra-tumour heterogeneity of diffuse large b-cell lymphoma involves the induction of diversified stroma-tumour interfaces. *EBioMedicine* **61**, 103055 (2020).
13. Lavin, R. C. & Tan, S. Spatial relationships of intra-lesion heterogeneity in mycobacterium tuberculosis microenvironment, replication status, and drug efficacy. *PLoS Pathog.* **18**, e1010459 (2022).
14. Cavinato, L. *et al.* Imaging-based representation and stratification of intra-tumor heterogeneity via tree-edit distance. *Sci. reports* **12**, 19607 (2022).
15. Da-Ano, R., Visvikis, D. & Hatt, M. Harmonization strategies for multicenter radiomics investigations. *Phys. Medicine & Biol.* **65**, 24TR02 (2020).
16. Ligerio, M. *et al.* Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *Eur. radiology* **31**, 1460–1470 (2021).
17. Mali, S. A. *et al.* Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. *J. personalized medicine* **11**, 842 (2021).

18. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8**, 118–127 (2007).
19. Chen, C. *et al.* Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one* **6**, e17238 (2011).
20. Adamer, M. F. *et al.* recombata: batch-effect removal in large-scale multi-source gene-expression data integration. *Bioinforma. Adv.* **2**, vbac071 (2022).
21. Horng, H. *et al.* Generalized combat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. *Sci. reports* **12**, 1–12 (2022).
22. Horng, H. *et al.* Improved generalized combat methods for harmonization of radiomic features. *Sci. Reports* **12**, 19009 (2022).
23. Dincer, A. B., Janizek, J. D. & Lee, S.-I. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* **36**, i573–i582 (2020).
24. Nioche, C. *et al.* Lifex: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer research* **78**, 4786–4789 (2018).
25. Gil, D., Ramos, O. & Perez, R. Topological radiomics (topiomics): Early detection of genetic abnormalities in cancer treatment evolution. In *Extended Abstracts GEOMVAP 2019: Geometry, Topology, Algebra, and Applications; Women in Geometry and Topology*, 89–93 (Springer, 2021).
26. Lin, D. Y. & Wei, L.-J. The robust inference for the cox proportional hazards model. *J. Am. statistical Assoc.* **84**, 1074–1078 (1989).
27. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *Jama* **247**, 2543–2546 (1982).
28. Hougaard, P. Frailty models for survival data. *Lifetime data analysis* **1**, 255–273 (1995).
29. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29–39 (2016).
30. Zindler, T., Frieling, H., Neyazi, A., Bleich, S. & Friedel, E. Simulating combat: how batch correction can lead to the systematic introduction of false positive results in dna methylation microarray studies. *BMC bioinformatics* **21**, 1–15 (2020).
31. Cavinato, L. *et al.* Explainable domain transfer of distant supervised cancer subtyping model via imaging-based rules extraction. *Artif. intelligence medicine* **138**, 102522 (2023).
32. Tan, C. *et al.* A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III* **27**, 270–279 (Springer, 2018).
33. Yao, Y., Rosasco, L. & Caponnetto, A. On early stopping in gradient descent learning. *Constr. Approx.* **26**, 289–315 (2007).

Acknowledgements

We acknowledge all the personnel of the Nuclear Medicine Department of IRCCS Humanitas Clinical and Research Hospital and Fondazione IRCCS Istituto Nazionale dei Tumori for the assistance during the PET/CT scans, segmentation of lesions, extraction of radiomic features and retrieval of patients’ personal information from electronic health records.

Author contributions statement

L.C. conceived the pipeline, set up the case study, implemented the tests, analyzed the results, prepared the figures, and wrote the manuscript. M.M. conceived the pipeline, formalized the PCSD, implemented the tests, analyzed the results, prepared the figures, and wrote the manuscript. M.S. collected the data, segmented the lesions, and performed radiomics extraction for Institution 1 patients. M.K. collected the data, segmented the lesions, and performed radiomics extraction for Institution 2 patients. F.I. supervised the analyses and the manuscript drafting. All authors reviewed the manuscript.

Additional information

Data and Code availability The data that support the findings of this study are available from IRCCS Humanitas Research Hospital and Fondazione IRCCS Istituto Nazionale dei Tumori but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of both IRCCS Humanitas Research Hospital and Fondazione IRCCS Istituto Nazionale dei Tumori. Code is available at this [link](#).

Competing interests The author(s) declare no competing interests.

Dual Adversarial Deconfounding Autoencoder for joint batch-effects removal from multi-center and multi-scanner radiomics data: Supplementary materials

Lara Cavinato^{1,+,*}, Michela Carlotta Massi^{2,+}, Martina Sollini^{3,4}, Margarita Kirienko⁵, and Francesca Ieva^{1,2}

¹Department of Mathematics, Politecnico di Milano, Via Bonardi, 9, Milan, 20133, Italy

²Health Data Science Center, Human Technopole, Viale Rita Levi-Montalcini, 1, Milan, 20157, Italy

³Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini, 4, Pieve Emanuele, 20090, Italy

⁴Department of Nuclear Medicine, IRCCS Humanitas Research Hospital, Via Alessandro Manzoni, 56, Rozzano, 20089, Italy

⁵Fondazione IRCCS Istituto Nazionale dei Tumori, Via Giacomo Venezian, 1, Milan, 20133, Italy

*lara.cavinato@polimi.it

+these authors contributed equally to this work

ABSTRACT

Medical imaging represents the primary tool for investigating and monitoring several diseases, including cancer. The advances in quantitative image analysis have developed towards the extraction of biomarkers able to support clinical decisions. To produce robust results, multi-center studies are often set up. However, the imaging information must be denoised from confounding factors – known as batch-effect – like scanner-specific and center-specific influences. Moreover, in non-solid cancers, like lymphomas, effective biomarkers require an imaging-based representation of the disease that accounts for its multi-site spreading over the patient's body. In this work, we address the dual-factor deconfusion problem and we propose a deconfusion algorithm to harmonize the imaging information of patients affected by Hodgkin Lymphoma in a multi-center setting. We show that the proposed model successfully denoises data from domain-specific variability (p -value <0.001) while it coherently preserves the spatial relationship between imaging descriptions of peer lesions (p -value=0), which is a strong prognostic biomarker for tumor heterogeneity assessment. This harmonization step allows to significantly improve the performance in prognostic models with respect to state-of-the-art methods, enabling building exhaustive patient representations and delivering more accurate analyses (p -values <0.001 in training, p -values <0.05 in testing). This work lays the groundwork for performing large-scale and reproducible analyses on multi-center data that are urgently needed to convey the translation of imaging-based biomarkers into the clinical practice as effective prognostic tools. The code is available on GitHub at this [link](#).

1 Patients' characteristics

Table S 1. Patients' characteristics in Institution 1: variables are divided into categorical (number, percentage on the total) and numerical (mean, standard deviation). In the first group, they are listed the stage (four statuses), the sex (female F and male M), the presence of B symptoms like fever, sweats, weight loss (yes Y and no N), status of the disease (extranodal disease: yes Y and no N; bone disease: yes Y and no N), administration of radiotherapy (yes Y and no N), the outcome of interim PET (iPET, Deauville Score DS of the PET), end of treatment PET (EOT PET, Douville Score DS of the PET). Statistics are stratified by the treatment response, thus patients are divided into responders and non-responders.

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 S 2. Patients' characteristics in Institution 1: variables are divided into categorical (number, percentage on the total) and numerical (mean, standard deviation). Among the numerical variables, there are age, number of nodal lesions of the patients, number of extranodal lesions of the patients, and time to relapse (for censored patients, the time to last follow-up is taken). Statistics are stratified by the treatment response, thus patients are divided into responders and non-responders.

Numerical variables – mean (std deviation)		
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)
Time to relapse/follow-up [days]	1126.97 (704.94)	358.86 (322.854)

Table S 3. Patients' characteristics in Institution 2: variables are divided into categorical (number, percentage on the total) and numerical (mean, standard deviation). In the first group, they are listed the stage (four statuses), the sex (female F and male M), the presence of B symptoms like fever, sweats, weight loss (yes Y and no N), status of the disease (extranodal disease: yes Y and no N; bone disease: yes Y and no N), administration of radiotherapy (yes Y and no N), the outcome of interim PET (iPET, positive or negative), end of treatment PET (EOT PET, positive or negative). Statistics are stratified by the treatment response, thus patients are divided into responders and non-responders.

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 S 4. Patients' characteristics in Institution 2: variables are divided into categorical (number, percentage on the total) and numerical (mean, standard deviation). Among the numerical variables, there are age, number of nodal lesions of the patients, number of extranodal lesions of the patients, and time to relapse (for censored patients, the time to last follow-up is taken). Statistics are stratified by the treatment response, thus patients are divided into responders and non-responders.

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)
Time to relapse/follow-up [days]	1105.72 (546.490)	257.59 (167.17)

2 Scanners' specifications

Table S 5. Image acquisition protocols and scanner specification in Institution 1: 85 patients were scanned with Siemens Biograph scanner; 51 patients were scanned with General Electric Discovery 690 scanner; 5 were scanned with other unspecified scanners.

Institution 1	Biograph – Siemens		Discovery 690 – General Electric	
	PET	CT	PET	CT
Min/bed position (static/dynamic)	2.5 (static)	–	2 (static)	–
Crystal	LSO	–	LYSO	–
Reconstruction	Iterative	–	Iterative, TOF Sharp IR	–
Attenuation correction	On CT data	–	On CT data	–
Matrix (pixels)	128×128	512×512	256×256	512×512
Resolution (mm)	5.3×5.3	0.98×0.98	2.73×2.73	1.37×1.37
Slice thickness (mm)	2.0	4.0	3.27	3.27
Slices	–	6	–	64

Table S 6. Image acquisition protocols and scanner specification in Institution 2: 34 patients were scanned with General Electric Discovery 710 scanner; 38 patients were scanned with Philips Gemini scanner; 1 patient was scanned with other unspecified scanners.

Institution 2	Discovery 710 – General Electric		Gemini - Philips	
	PET	CT	PET	CT
Min/bed position (static/dynamic)	2 (static)	–	2 (static)	–
Crystal	LYSO	–	BGO	–
Reconstruction	VPFX	–	Iterative	–
Attenuation correction	On CT data	–	On CT data	–
Matrix (pixels)	192×192	512×512	169×169	512×512
Resolution (mm)	3.65×3.65	1.37×1.37	4×4	1.37×1.37
Slice thickness (mm)	3.27	3.75	4	4
Slices	–	64	–	64

3 Radiomics discrepancies

Table S 7. Descriptive Statistics and statistical comparisons of radiomics variables in terms of mean values and standard deviations for the two cohorts.

Variables	Institution 1		Institution 2		p-value
	Mean	Std_Dev	Mean	Std_Dev	
CONVENTIONAL_SUVmin	3.7737	1.8616	3.3597	1.6822	<< 0.01
CONVENTIONAL_SUVmean	6.1546	2.8542	5.5975	2.4896	<< 0.01
CONVENTIONAL_SUVstd	1.4715	0.7308	1.3788	0.6578	0.0054
CONVENTIONAL_SUVmax	10.2264	4.7974	9.4158	4.2288	<< 0.01
CONVENTIONAL_SUVpeak	2.3524	4.2680	5.9848	3.9711	<< 0.01
CONVENTIONAL_TLG	82.1013	230.5903	75.5440	149.7817	<< 0.01
HISTO_Skewness	0.5872	0.4147	0.6171	0.4045	0.1229
HISTO_Kurtosis	2.8807	0.8532	2.8957	0.8498	0.2132
HISTO_ExcessKurtosis	-0.1192	0.8532	-0.1042	0.8498	0.2132
HISTO_Entropy_log10	1.1423	0.2083	1.1269	0.1763	0.0067
HISTO_Entropy_log2	3.7946	0.6922	3.7434	0.5858	0.0067
HISTO_Energy	0.0944	0.0538	0.0942	0.0423	0.0073
SHAPE_Volume	11.6197	24.3370	12.6720	24.6858	<< 0.01
GLCM_Homogeneity	0.3335	0.1058	0.3327	0.0814	0.0873
GLCM_Energy	0.0191	0.0187	0.0167	0.0135	0.293
GLCM_Contrast	40.3799	42.1976	32.4835	29.6706	0.0437
GLCM_Correlation	0.3036	0.1836	0.3069	0.1315	0.9423
GLCM_Entropy_log10	1.9241	0.3304	1.9367	0.2623	0.2163
GLCM_Entropy_log2	6.3918	1.0975	6.4336	0.8715	0.2163
GLCM_Dissimilarity	4.4523	2.3504	4.1395	1.8138	0.0625
GLRLM_SRE	0.9483	0.0362	0.9505	0.0257	0.0204
GLRLM_LRE	1.2597	0.2396	1.2304	0.1672	0.3075
GLRLM_LGRE	0.0065	0.0132	0.0056	0.0059	<< 0.01
GLRLM_HGRE	516.1915	456.6894	419.2514	402.2381	<< 0.01
GLRLM_SRLGE	0.0059	0.0107	0.0052	0.0051	<< 0.01
GLRLM_SRHGE	496.7612	442.7954	400.0534	375.1861	<< 0.01
GLRLM_LRLGE	0.0102	0.0313	0.0075	0.0117	<< 0.01
GLRLM_LRHGE	607.9699	539.4094	518.3472	689.0347	<< 0.01
GLRLM_GLNU	22.8289	42.6938	19.5446	45.4589	0.5206
GLRLM_RLNU	245.5027	441.3303	194.1253	307.1768	0.0858
GLRLM_RP	0.9331	0.0454	0.9357	0.0337	0.0397
NGLDM_Coarseness	0.0365	0.0224	0.0368	0.0191	0.1317
NGLDM_Contrast	0.3887	0.3496	0.3412	0.2579	0.1648
NGLDM_Busyness	0.2764	0.5719	0.2360	0.4246	0.1134
GLZLM_SZE	0.6168	0.1427	0.6097	0.1149	0.0236
GLZLM_LZE	102.2394	772.5381	88.0518	1134.3479	0.0463
GLZLM_LGZE	0.0066	0.0133	0.0057	0.0057	<< 0.01
GLZLM_HGZE	510.9761	437.2528	411.2058	377.2074	<< 0.01
GLZLM_SZLGE	0.0033	0.0053	0.0032	0.0025	<< 0.01
GLZLM_SZHGE	353.6975	349.1995	276.7329	307.0125	<< 0.01
GLZLM_LZLGE	2.7491	30.4856	1.5488	19.6571	0.0014
GLZLM_LZHGE	12582.0082	42514.1309	12686.7792	93404.5481	0.0011
GLZLM_GLNU	7.4366	9.9953	6.4751	6.4384	0.6327
GLZLM_ZLNU	49.2334	87.4316	35.8290	47.5280	0.0509
GLZLM_ZP	0.4574	0.1997	0.4445	0.1594	0.0287

4 Results of experiment 3

Table S 8. Experiment 3 results: performance of the Cox proportional hazard models trained and tested in cross-validation using different patient representations. Each modality – i.e. radiomics, Dual AD-AE embeddings, ComBat-based standardization, ReComBat-based standardization, and OPNested-based standardization – is fed in the survival model according to three different patient representations: (1) the patient is described by the centroid of its point cloud (“centroid representation”), (2) the patient is described by the topological characteristics of its point cloud (“cloud description representation”) and (3) the patient is described by both the centroid and the topological characteristics of its point cloud (“complete representation”). Best values are highlighted in bold.

Models	Centroid representation		Cloud description representation		Complete representation	
	C-index (train)	C-index (test)	C-index (train)	C-index (test)	C-index (train)	C-index (test)
Radiomics	0.6962 ± 0.0190	0.4344 ± 0.1779	0.5828 ± 0.0141	0.5184 ± 0.1203	0.7077 ± 0.0207	0.4026 ± 0.1460
Dual AD-AE	0.7803 ± 0.0132	0.6100 ± 0.1382	0.6728 ± 0.0135	0.6481 ± 0.1305	0.7803 ± 0.0132	0.6100 ± 0.1382
ComBat-center-scanner	0.7041 ± 0.0201	0.5525 ± 0.1358	0.5611 ± 0.0162	0.5101 ± 0.1228	0.7173 ± 0.0180	0.5063 ± 0.1279
ComBat-scanner-center	0.7016 ± 0.0195	0.5366 ± 0.1266	0.5625 ± 0.0159	0.5075 ± 0.1262	0.7291 ± 0.0185	0.5066 ± 0.1234
ReComBat-center-scanner	0.7064 ± 0.021	0.5595 ± 0.1344	0.5621 ± 0.0161	0.5128 ± 0.1239	0.7153 ± 0.0186	0.5201 ± 0.1224
ReComBat-scanner-center	0.7019 ± 0.0194	0.5444 ± 0.1240	0.5635 ± 0.0160	0.5101 ± 0.1244	0.7260 ± 0.0192	0.5120 ± 0.1348
OPNested ComBat	0.7019 ± 0.0194	0.5444 ± 0.1240	0.5635 ± 0.0160	0.5101 ± 0.1244	0.7260 ± 0.0192	0.5120 ± 0.1348

5 Software

At this [link](#), the code for Dual AD-AE is distributed. We implemented the training of the dual AD-AE model, adhering to conventional practices involving several procedural steps. Following data preparation and model architecture definition, the training loop iterated through the training data. In each iteration, a batch of input data underwent encoding to derive latent representations, which were subsequently passed through the decoder to generate reconstructed data. The loss function was employed to quantify the dissimilarity between the input and reconstructed data, and backpropagation facilitated the computation of gradients pertaining to the model parameters. These gradients, in turn, drove the updating of model weights through the chosen optimizer. Periodic utilization of the validation set allowed for performance evaluation and the monitoring of training progression, facilitating early stopping to counteract overfitting.

The choice to employ a unique validation set with multiple epochs, rather than opt for cross-validation involving numerous train/test splits with fewer epochs, was underpinned by various considerations. These encompassed the size of our dataset, the inherent nature of the problem, and available computational resources. Although this approach harbored the potential for overfitting and a potentially less precise evaluation of model generalization, its implementation was both straightforward and computationally efficient, given its reliance on the complete dataset. It further provided a singular model framework endowed with a consistent validation set, thereby enabling continuous monitoring of performance dynamics.

However, a cross-validation approach can be interesting as an other training option, particularly maximizing the utility of limited data and furnishing a more robust assessment of generalization, thus we subsequently introduced an alternative training regimen for the AD-AE model. In this regimen, we implemented a cross-validation setup comprising 50 splits, each spanning 100 epochs.

MOX Technical Reports, last issues

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

- 81/2023** Buchwald, S.; Ciaramella, G; Salomon, J.; Sugny, D.
A SPIRED code for the reconstruction of spin distribution
- 82/2023** Pozzi, G.; Ciarletta, P.
Geometric control by active mechanics of epithelial gap closure
- 79/2023** Agosti, A.; Bardin, R.; Ciarletta, P.; Grasselli, M.
A diffuse interface model of tumour evolution under a finite elastic confinement
- 78/2023** Antonietti, P.F.; Bonizzoni, F.; Corti, M.; Dall'Olio, A.
Discontinuous Galerkin for the heterodimer model of prion dynamics in Parkinson's disease
- 77/2023** Fumagalli, I.; Corti, M.; Parolini, N.; Antonietti, P. F.
Polytopal discontinuous Galerkin discretization of brain multiphysics flow dynamics
- 76/2023** Ieva, F.; Galliani, G.; Secchi, P.
The impact of public transport on the diffusion of COVID-19 pandemic in Lombardy during 2020
- 74/2023** Pidò, S.; Pinoli, P.; Crovari, P.; Ieva, F.; Garzotto, F.; Ceri, S.
Ask Your Data—Supporting Data Science Processes by Combining AutoML and Conversational Interfaces
- 75/2023** Archetti, A.; Ieva, F.; Matteucci, M.
Scaling survival analysis in healthcare with federated survival forests: A comparative study on heart failure and breast cancer genomics
- 71/2023** Conni, G.; Piccardo, S.; Perotto, S.; Porta, G.M.; Icardi, M.
HiPhome: High order Projection-based HOMogEnisation for advection diffusion reaction problems
- 70/2023** Ragni, A.; Ippolito, D.; Masci, C.
Assessing the Impact of Hybrid Teaching on Students' Academic Performance via Multilevel Propensity Score-based techniques