Neural Network Training for Cross-Protocol Radiomic Feature Standardization in Computed Tomography

Vincent Andrearczyk[a], Adrien Depeursinge[a,b], Henning Müller[a,c]

[a] *Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland*
[b] *Biomedical Imaging Group (BIG), Ecole polytechnique fédérale de Lausanne (EPFL), Switzerland*
[c] *University of Geneva (UNIGE), Switzerland*

Abstract

Radiomics has shown promising results in several medical studies, yet it suffers from a limited discrimination and informative capability as well as a high variation and correlation with the tomographic scanner types, pixel spacing, acquisition protocol and reconstruction parameters. This paper proposes and compares two methods to transform quantitative image features in order to improve their stability across varying image acquisition parameters while preserving the texture discrimination abilities. In this way, variations in extracted features are representative of true physio-pathological tissue changes in the scanned patients. A first approach is based on a two-layer neural network that can learn a non-linear standardization transformation of various types of features including hand-crafted and deep features. Second, domain adversarial training is explored to increase the invariance of the transformed features to the scanner of origin. The generalization of the proposed approach to unseen textures and unseen scanners is demonstrated by a set of experiments using a publicly available CT texture phantom dataset scanned with various imaging devices and parameters.

Keywords: quantitative imaging, radiomics, deep learning, standardization, domain adversarial

*Corresponding author
Email address: vincent.andrearczyk@hevs.ch (Vincent Andrearczyk)

1. Introduction

Radiomics aims at extracting and analyzing large amounts of quantitative image features (e.g. volume, shape, intensity and texture) from medical images including Computed Tomography (CT), Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI) to improve decision-support, mostly in cancer treatment. The number of related papers has followed an exponential growth since the first publications in 2010 [1–3]. In the context of oncology, radiomics allows establishing complex links between tumoral Regions Of Interest (ROI) and clinical endpoints, such as diagnostic (presence and type of cancer [4]), prognostic (information on overall survival and recurrence [3, 4]) or predictive (treatment responses and benefits [2]) analyses. Various organs and cancer types have been analyzed with radiomics including lungs [3–10], liver [11], breast [11, 12], head-and-neck [3] and brain gliomas [13]. Radiomics generally refers to an interlinked sequence of processes including image acquisition and reconstruction, ROI segmentation, quantitative feature extraction and analysis. This study focuses on the impact of the first two processes, namely acquisition and reconstruction, on the values of quantitative features. The standardization of features can also impact the segmentation of tumor regions when the latter involves the spatial clustering of features.

Uncovering disease characteristics or predicting a response to treatment relies on the fact that the extracted features describe the patients' biomarkers (physio-pathological effects) independently from the image acquisition device or protocol. Quantitative features extracted from a ROI in scans of the same person acquired in different hospitals should ideally be identical (without considering temporal variations due to disease evolution). Scanning protocols and machines are frequently changed over time and vary across hospitals while even the same scanners can be configured in different ways and software on the scanners is regularly updated without knowing details of the impact of updates (such as noise reduction algorithms) on the produced images. While the difficulty for clinicians to take these variations into account can be limited thanks to their experience and knowledge, radiomics biomarkers such as texture features lack this abstraction level and can be strongly impacted by these changes [7]. Several studies have shown a high variability and dependence of texture radiomic features across scans, limiting their interpretability and comparison [5, 7, 8, 14, 15]. Yet, little attention has been devoted to reducing this variation and many radiomic studies are based on very clean data from a single scanner type and often with the exact same protocol, which is not realistic in standard clinical situations.

The influence of image processing and of feature extraction algorithms and implementation on the feature variation is tackled by the Image Biomarker Standardization Initiative (IBSI) [16]. Various studies have evaluated the reproducibility and stability of texture features and the influence of scanner variation and reconstruction settings [5, 6, 8, 14, 17–20]. These studies generally aim at selecting stable and repeatable texture features for a given task with test-retest and inter-rater reliability analysis, without proposing a method to standardize unstable features. The main limitations of such studies are their lack of generalization as the reproducibility is valid for only one scanner and one task as well as the questionable assumption that the analyzed body part appearance has not changed between acquisitions. A study of bias and variability of texture features across synthetically simulated scans with various image acquisition settings and reconstruction algorithms was proposed in [15]. The simulations included modifying the slice thickness, in-plane pixel size, dose, the Task-Transfer Function (TTF) and Noise Power Spectrum (NPS). The extracted features were compared with those from an original phantom scan from which the synthesized scans are computed. The results show that image acquisition and reconstruction conditions lead to substantial bias and variability of the texture features.

Texture phantom images allow evaluating the variation of features extracted from different scanners and with varying protocols of an unchanged body. It avoids repeatedly exposing a patient to radiation and tiring protocols [7] and only presents slight differences in positioning between scans. Recent stability analyses studied the use of phantom volumes, similar to those used in this paper, to ensure the similarity of the scanned body between consecutive scans and across multiple scanners [7, 21]. CT images were pre-processed in [10] by resampling and filtering to standardize image pixel sizes, resulting in a reduced variability of radiomic features. Another phantom study was proposed in [22] to evaluate intensity and texture features across varying CT acquisitions of the same phantom. Again, the same conclusion was drawn, claiming that quantitative changes may be primarily due to acquisition variability rather than from real physio-pathological effects.

Finally, an excellent systematic review of the repeatability and reproducibility of radiomic features with and without phantom studies was recently presented in [23]. We refer to this work for more details on the mentioned analyses and a more exhaustive literature review.

Recently, a study was performed on the dependency of deep features from image pixel sizes across CT scanners [24]. The features were extracted from pre-trained Convolutional Neural Networks (CNN), mainly VGG networks of different depths. A normalization method was proposed based on a holistic assumption of quadratic and cubic proportionality between the features and the pixel size, with limited success in removing the dependency of some of the deep features. Another study of deep learning for CT texture classification was performed in [25], in the context of image quality after reconstruction of CT images with reduced radiation doses. This study used a phantom dataset but considered a classification accuracy rather than a standardization of features presented in this paper. By focusing on the accuracy, classification may achieve an excellent class recognition, although the extracted features may be non-informative for a radiomics task (e.g. average of Hounsfield Units in the phantom dataset in [7]) and highly correlated with scanner parameters. With motivations similar to ours, yet without the use of phantom volumes that ensure the stability of the measured body to isolate the variability due to scanners, a simple harmonization method named ComBat was recently used in [11] to standardize radiomic texture features.

The adequacy of deep learning for texture analysis and medical imaging was extensively demonstrated in various studies [13, 26–29]. Besides this, the complementarity of deep and radiomic features was demonstrated in [4, 30] for the prediction of patient survival in the context of lung cancer and in [12] for breast cancer detection. This paper is therefore not dedicated to yet another illustration of the informativeness and generalization of these features in a classic radiomics task. Our goal is to demonstrate that the performance and reliability of hand-crafted and deep descriptors can be further improved by using phantom images to learn a feature stability transformation allowing robust generalization to unknown textures and unknown scanners. The obtained features are robust to changes in the acquisition and reconstruction methods. This allows oncologists to better evaluate and compare patient biomarkers over time and across scanners and hospitals, while predictive models based on the standardized features will achieve better generalization. An overview of the proposed approach is illustrated in Figure 1.

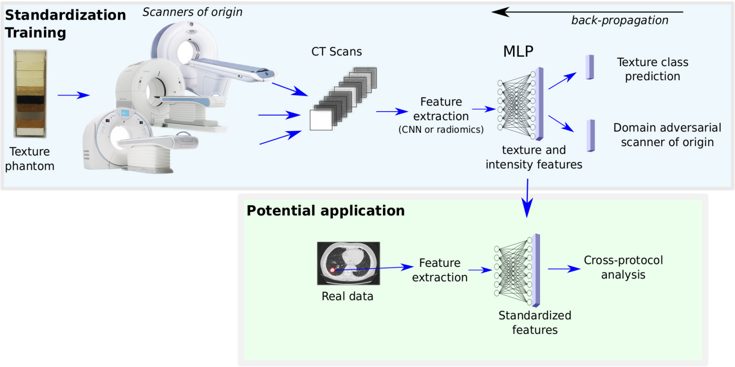


Figure 1: Overview of the proposed standardization of visual features extracted from CT scans of various scanners.

This work expands our previous study [31] in which a first standardization method was proposed using neural network training on phantom CT scans. The main extensions proposed in this paper are summarized as follows: (a) Experiments are extended to the analysis of a generalization to unknown scanners. The networks are trained to standardize features from a set of scans using the phantom images. The resulting features extracted from new scans acquired and reconstructed differently become also more stable. (b) Domain adversarial training is proposed to avoid that the extracted features contain information about the domain of origin, i.e. the scanner and protocol. (c) Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction methods are used for visualization of the learned features to support the analyses. (d) The transfer learning of the CNNs is evaluated by comparing pre-trained networks with those trained from scratch.

2. Material and Methods

2.1. General Overview

For a given ROI in an image, a stack of slices or a volume I, we extract a feature vector $\boldsymbol{g} = g(I)$. The function $g(I)$ is generally a set of classic radiomic features, e.g. first, second and higher order statistics of pixel values (see Section 2.4). These radiomic features were shown to strongly vary for a same phantom texture scanned across different scanners with varying protocols [5, 8, 14]. More discriminative features can also be extracted with CNNs.

Using a phantom of texture volumes, we train neural networks on top of the image features to classify slices from a subset of textures from 17 scans acquired with different scanners and protocols and reconstructed with different algorithms. In this way, hidden layers converge to similar values for each texture type, where the extracted features become standardized for the considered set of scanners. We can then test whether this standardization generalizes to another set of textures, implying a reduced variability of the features across scanners essential to robust clinical analyses. With a perfect standardization, the obtained features should be nearly identical across scans and informatively characterize the textures (thus the features should be unable to identify the scanner type). The training process is therefore used to maximize inter-class feature variation while minimizing intra-class feature variation.

We extract observed features $\boldsymbol{g}^m_{k,l}$, where $k = 1, ..., K$ represents the class, $l = 1, ..., L$ is the scan of origin (including variation of scanner type, acquisition protocol and reconstruction algorithm) and m is the feature extraction method. In our setup, we have $L = 10$, $K = 17$ and $m \in \{rad., vgg, res.\}$ standing for radiomics, VGG and ResNet. For each feature extraction method m, we want to find a feature transformation $\tau^m(\boldsymbol{g}^m_{k,l})$ that both (a) minimizes the variation due to scanner, i.e. $\tau^m(\boldsymbol{g}^m_{k,l})$ is independent from l and (b) maintains the discriminability of the features, i.e. high-throughput quantitative features with optimal class separability and informativity of texture variations. We will drop the m index for readability when it is not essential, or use it only for the features, implying it for the transform function, e.g. $\tau(\boldsymbol{g}^m)$. This problem statement is summarized in Figure 2. The extraction and transformation are in practice performed at a slice level, whereas $\boldsymbol{g}_{k,l}$ and $\tau(\boldsymbol{g}_{k,l})$ are averages of the features and of their transformed counterparts within texture volumes. We add an index s (e.g. $\boldsymbol{g}_{k,l,s}$) to refer to features extracted from a single slice. Note that averaging the features within volumes (2.5D) is a common practice in radiomic studies and the phantom volumes can be considered homogeneous (stationary) [32] as each of them is composed of a unique material.

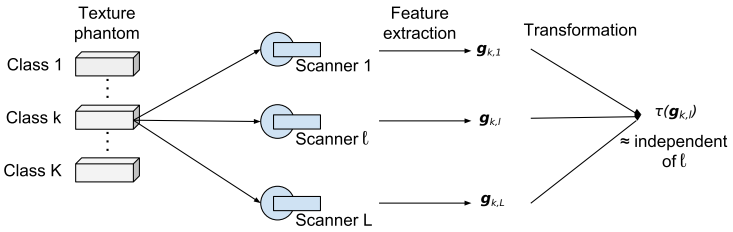


Figure 2: Overview of the problem statement in which the function τ is sought to obtain informative features ideally independent of the scanner types.

2.2. Dataset

We use the Credence Cartridge Radiomics (CCR) phantom dataset developed in [7]. The physical phantom contains ten volumes of textures (cartridges) as shown in Figure 3. The cartridge materials were selected to span the range of radiomic features found in scanned lung tissue and tumors (non small cell lung cancer), for example in terms of density and texture. The developed methods are therefore strongly expected to generalize to clinical images. The dataset consists of 17 CT scans of this volume produced by different scanners (from the manufacturers GE, Philips, Siemens and Toshiba), in different centers and with different acquisition protocols and reconstruction algorithms. Although it is a 3D volume, it is designed for the analysis of 2D slices (the method is termed 2.5D in [7] and commonly used in medical imaging). Contour positions of individual slices inside the cartridges (6 to 11 slices per cartridge) are provided to extract the features. More information about the scanners and the scanning protocols can be found in [7]. The dataset is publicly available and the experiments are thus fully reproducible.

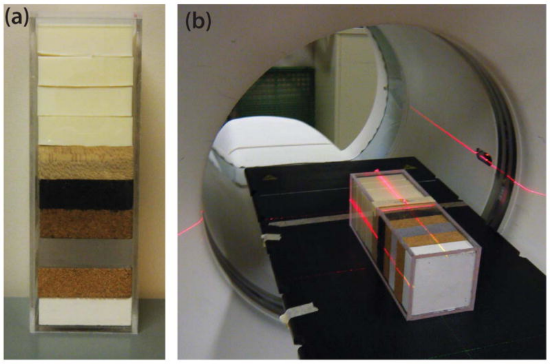


Figure 3: Texture phantom volume used to acquire the CCR dataset (Figure reproduced from [7]).

2.3. Pre-processing

The features are extracted from $16cm^2$ slices as provided in [7]. The slices are resized using bilinear interpolation to either (a) in-plane pixel spacing of $1mm^2$ for the radiomic features as suggested in [7], or (b) to the CNN input size for computing the deep features (224×224). The Hounsfield Units (HU) range $[-1,409;747]$ is linearly converted into the interval $[0;255]$ for the input of CNNs as in [30]. This reduces the dynamic range in the images but we expect these effects to be limited. The effect of interpolation is limited as the textures are relatively homogeneous in the phantom and in addition we learn a stable representation of the texture after interpolation. For the CNNs, a three channel input is obtained by duplication in order to use networks pre-trained on color images from ImageNet (more information on transfer learning in Section 4.3). As a standard procedure to keep the same pixel value range as the pre-trained domain, the image intensity histograms used with the pre-trained CNNs are centered and scaled similarly to the images used for pre-training (i.e. ImageNet mean subtraction and division by the ImageNet standard deviation).

2.4. Feature Extraction

As a baseline, we use radiomic features extracted with the pyRadiomics toolbox [9] with a fixed bin width of 25. A 97-dimensional feature vector is extracted from each slice. The extracted features include intensity features, i.e. first order statistics and texture features including grey-level co-occurrence, run length, size zone, and dependence matrices as well as a neighboring grey tone difference matrix.

As a second set of features, we use VGG19 [33] and ResNet-50 [34] to extract deep features from the slices. We remove the prediction layer and extract the penultimate layer output. The VGG and ResNet-50 features are of dimension $d = 4,096$ and $d = 2,048$ respectively.

By averaging the features $\boldsymbol{g}_{k,l,s}$ over all slices within each cartridge, we obtain the feature vectors $\boldsymbol{g}_{k,l}$.

2.5. Feature Transformation

We design a two-layer Multi-Layer Perceptron (MLP) with 100 hidden neurons (with standard dropout 0.5 and ReLU activation). The design of this network is motivated by a simple non-parametric yet non-linear transformation where the 100 neurons are used to correspond to the radiomic feature dimensionality (97) for comparison. For a given feature extraction method m, the MLP takes the observed features $\boldsymbol{g}^m_{k,l,s}$ as input, and is trained to output a class probability with five training texture classes (i.e. five output neurons). After training (see Section 2.6), the output of the hidden layer is used as a 100-dimensional feature vector that performs the transformation $\tau(\boldsymbol{g}_{k,l,s})$. Once again, by averaging the features $\tau(\boldsymbol{g}_{k,l,s})$ within each cartridge, we obtain the feature vectors $\tau(\boldsymbol{g}_{k,l})$.

The MLP performs a transformation of the feature space into a discriminative and clustered space, in which the features $\tau(\boldsymbol{g}_{k,l})$ are more stable to scanner variability. This is achieved by learning from the set of training slices and the ground truth of the texture types, and assuming that the scanner invariance of the learned representation will generalize to unknown tissue types if the changes are relatively systematic. We then evaluate the stability of the original features $\boldsymbol{g}^m_{k,l}$ and their transformed counterparts $\tau(\boldsymbol{g}^m_{k,l})$ for all the methods m.

When finetuning the CNNs, the MLP is connected to the penultimate layer that outputs $\boldsymbol{g}_{k,l,s}$ and we freeze all but the last two trainable layers (MLP layers) as in a standard finetuning. This is equivalent to training the MLP on the extracted features $\boldsymbol{g}_{k,l,s}$. An overview of the deep feature training and extraction is illustrated in Fig. 4.

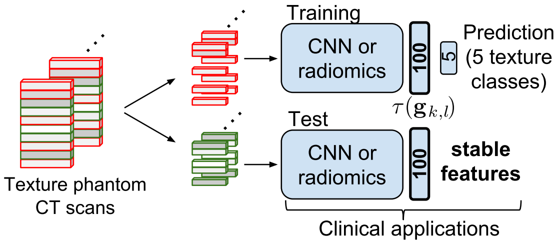


Figure 4: Overview of the feature extraction and training. The CNN is either VGG-19 or ResNet-50 from which the prediction layer is removed.

2.6. Training

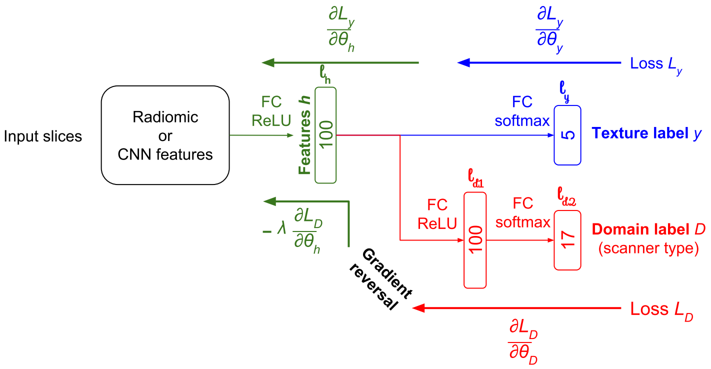
We randomly split the dataset (100 repetitions) to train the networks to classify half of the texture types. For each run, five texture volumes from all the 17 scans are used for training, the remaining five are kept for testing. A number of slices ranging from 6 to 11 depending on the scans and cartridges are available from each cartridge, as proposed in [7]. From 1,360 available slices in total, we obtain training and test sets composed of a number of slices between 675 and 685 depending on the random splits.

The networks are trained by optimizing the class prediction, i.e. the last layer (softmax activated) of the MLP, but the feature representation is extracted from a hidden layer with output $\tau(\boldsymbol{g}_{k,l,s})$. More details on the training setup are provided in Section 3.2.

The methods presented until here including feature extraction and transformation are first evaluated using all scanners for training and testing (but different texture classes) in Sections 4.1 to 4.3. The standardization with unknown test scanners is then evaluated in Section 4.4.

2.7. Domain Adversarial Training

Domain adversarial training of neural networks [35] is a method inspired by the domain adaptation theory to optimize a main learning task while minimizing the discrimination between domains. In this feature standardization task, the domain is the scan. This was used in medical imaging to increase generalization to new data with different imaging protocols in brain lesion segmentation in [36] and for dealing with appearance variability of histopathology images due to acquisition variation between pathology labs in [37]. The idea is that removing the domain information, i.e. the scanner type and protocol, from the trained features enhances their stability. For this, we assume that all the slices from a given scan come from the same data distribution constituting a domain. Domain adversarial training is employed to learn the texture cartridge classification as explained in Section 2.6, while limiting the possibility to recover the domain of origin from the learned features (i.e. lower overall correct classification of the domain classifier). Figure 5 illustrates our domain adversarial training. The domain classifier with trainable parameters θ_D contains two fully connected layers of 100 and 17 neurons respectively. The prediction layer contains 17 neurons for classifying the 17 scans. The label classifier in blue predicts the texture labels with parameters θ_y. Finally, the parameters θ_h of the hidden layer with features $h = \tau(\boldsymbol{g}^{m}_{k,l,s})$ are trained with both the label loss gradients $\frac{\partial L_y}{\partial \theta_h}$ and the reversed domain loss gradients $-\lambda \frac{\partial L_D}{\partial \theta_h}$, where λ weighs the importance of the domain adversarial training of the features h. Results using domain adversarial training are presented in Section 4.5.



$\frac{\partial L_y}{\partial \theta_h}$
$\frac{\partial L_y}{\partial \theta_y}$
Loss L_y
ℓ_h
ℓ_y
Input slices
Radiomic or CNN features
FC ReLU
Features h
100
FC softmax
5
Texture label y
ℓ_{d1}
ℓ_{d2}
$-\lambda \frac{\partial L_D}{\partial \theta_h}$
FC ReLU
100
FC softmax
17
Domain label D (scanner type)
Gradient reversal
Loss L_D
$\frac{\partial L_D}{\partial \theta_D}$

Figure 5: Our domain adversarial network architecture. The domain D is the scanner type and the classification label y is the texture volume. The standardized features $h = \tau(\boldsymbol{g}^{m}_{k,l,s})$ are extracted from layer l_h. Best viewed in color.

3. Experimental Setup and Evaluation

The feature vectors are extracted for all the test slices and averaged within the cartridges ($\boldsymbol{g}^{rad.}_{k,l} \in \mathbb{R}^{97}$, $\boldsymbol{g}^{vgg}_{k,l} \in \mathbb{R}^{4096}$, $\boldsymbol{g}^{res.}_{k,l} \in \mathbb{R}^{2048}$ and $\tau(\boldsymbol{g}^{m}_{k,l}) \in \mathbb{R}^{100}$). The sparsity of the neuron activations results in a few features of $\tau(\boldsymbol{g}_{k,l})$ being zero for all the slices of a test set. These features are removed from the sets in each of the 100 runs. The dimensionality of $\tau(\boldsymbol{g}_{k,l})$ may, therefore, be reduced to $d \leq 100$.

3.1. Evaluation Metrics

Several metrics are employed to evaluate the stability of the features and their dependence on the scanners, acquisition protocols and reconstruction algorithms as listed in the following.

Intraclass Correlation Coefficient (ICC). Intra-class Correlation Coefficient (ICC) evaluates the clustering of features from several classes using the correlation of features within classes as

$$ICC = \frac{BMS - EMS}{BMS + (k-1)EMS + \frac{k}{n}(JMS - EMS)}, \tag{1}$$

where n is the number of targets (5 test classes) and k is the number of judges (17 scans, or 8 in Section 4.4). BMS is the Between targets Mean Square, EMS the residual mean square and JMS the between Judges Mean Square. The ICC ranges from 0 to 1 with values close to 1 indicating high similarity between values of the same class. The coefficients are averaged across all d features.

In Section 4.5, we also evaluate the scan domain ICC, i.e. the correlation within scans that we want to minimize. ICC is a standard evaluation method of feature stability, yet we provided other measures for a more exhaustive evaluation.

Clustering. For further analysis of class separability, clustering based measures are also standard, where cluster dispersion measured under Gaussianity assumption is reasonable. We apply a Gaussian Mixture Model (GMM) with five components corresponding to the five test classes to cluster the features $\boldsymbol{g}_{k,l}^{radiomics}$ or $\tau(\boldsymbol{g}_{k,l}^{m})$ from the test volumes. We evaluate the clustering results using the ground truth test labels. We measure and report the homogeneity, completeness, V-measure (harmonic mean of the latter) and the average covariance of the mixture components. The homogeneity and completeness are in the range $[0, 1]$. The former is highest if the clusters contain only cartridges of a single class, the latter if all cartridges of a given class are elements of the same cluster.

Correlation with pixel spacing. As pointed out in other studies [2, 7, 15], we noted that the value of the features is highly correlated with the pixel spacing, limiting their comparison and interpretability. We measure, average and compare the absolute Pearson correlation coefficients of the various extracted features with the resolution of the slices. It is worth noting that this metric only reflects linear relations between the features and the pixel spacing. Nonlinear dependencies are evaluated by recovering the scan classification from the features with a non-linear MLP classifier in the domain adversarial experiments in Section 4.5.

Dimensionality Reduction. An excellent clustering and ICC of unknown textures can be obtained with a simple HU averaging as this measure separates the cartridges well in the CCR dataset. However, the informative and discriminative power of such a simple feature is limited in a real medical image analysis scenario, where texture rather than only density are important to separate tissue types. On the other hand, a higher dimensional feature vector with highly correlated features can also result in an excellent clustering and ICC. Yet, such non-informative redundancy offers little interest in the description of biomarkers for more complex medical imaging tasks. Principal Component Analysis (PCA) allows evaluating the intra- and inter-class variability along the directions of the largest variance in the feature space. PCA and t-distributed Stochastic Neighbor Embedding (t-SNE) are also used to illustrate the stability of the features across scanners in a 2-dimensional plot (see Figures 7 and 8). t-SNE is a non-linear dimensionality reduction technique used for visualization of high-dimensional data points. Feature vectors are modeled by 2D points so that similar vectors result in nearby points and dissimilar vectors in distant points.

3.2. Training Setup

The CNNs are pre-trained on ImageNet [38] to obtain informative deep features despite the limited amount of training data. They are then finetuned end-to-end by adding fully-connected layers in place of the MLP. The CNNs and MLPs are trained with the Adam optimizer [39] with standard hyperparameters, namely an initial learning rate of 10^{-4}, average decays β_1 and β_2 of 0.9 and 0.999 respectively, and a batch size of 32. The deep CNNs are trained for 100 epochs and the shallow network (MLP on top of radiomic features) for

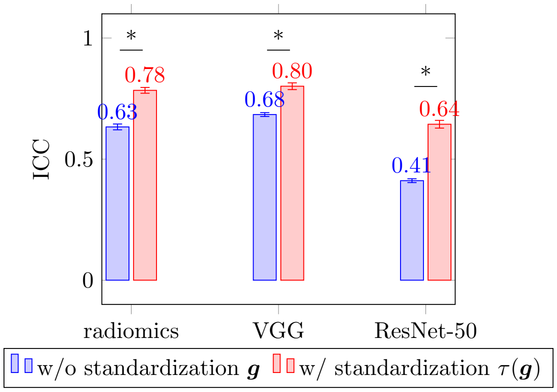


Figure 6: ICC before and after feature standardization when averaged over 100 runs with 95% confidence interval. Asterisks represent statistically significant differences ($p{-}value < 0.0001$).

500 epochs. The radiomics MLP is trained for more epochs than the CNNs as the former overfits less due to a reduced number of trainable weights (the CNNs' second last dense layer is substantially wider than the radiomic feature dimensionality). The pre-trained CNNs also enable a faster convergence. The random train/test split is reproduced 100 times, with the same splits kept unchanged for all experiments. The average and standard deviation are reported for each method.

4. Results

4.1. Results with Known Scanners

As a first set of experiments, the training and test slices originate from the same 17 scans. As mentioned in Section 2.6, half of the texture types are used for training (five texture labels), the rest for testing with repeated random splits. Figure 6 illustrates the statistically significant improvement of ICC (with $p - value < 0.0001$ for the three methods) with the proposed standardization method. Considering only the ICC, the radiomic features surprisingly obtain better results than the ResNet ones (with $p - value < 0.0001$), although this is contrasted by the supplementary results.

More results are provided in Table 1, supporting our hypothesis that robust features are obtained using the proposed training scheme.

	ICC$_{(\uparrow)}$	H$_{(\uparrow)}$	C$_{(\uparrow)}$	V$_{(\uparrow)}$	Cov.$_{(\downarrow)}$	Cor.$_{(\downarrow)}$
Radiomics $\boldsymbol{g}^{rad.}$	0.633$_{\pm 0.06}$	0.564$_{\pm 0.10}$	0.672$_{\pm 0.09}$	0.611$_{\pm 0.09}$	0.343$_{\pm 0.18}$	0.577$_{\pm 0.02}$
MLP radiom. $\tau(\boldsymbol{g}^{rad.})$	0.784$_{\pm 0.06}$	0.723$_{\pm 0.10}$	0.770$_{\pm 0.08}$	0.745$_{\pm 0.09}$	0.239$_{\pm 0.07}$	0.510$_{\pm 0.03}$
VGG $\boldsymbol{g}^{vgg}$	0.684$_{\pm 0.04}$	**0.794**$_{\pm 0.10}$	0.844$_{\pm 0.08}$	**0.817**$_{\pm 0.09}$	0.352$_{\pm 0.05}$	0.504$_{\pm 0.02}$
MLP VGG $\tau(\boldsymbol{g}^{vgg})$	**0.801**$_{\pm 0.07}$	0.790$_{\pm 0.11}$	**0.849**$_{\pm 0.10}$	**0.817**$_{\pm 0.10}$	**0.199**$_{\pm 0.08}$	0.503$_{\pm 0.04}$
ResNet-50 $\boldsymbol{g}^{res.}$	0.411$_{\pm 0.04}$	0.681$_{\pm 0.12}$	0.778$_{\pm 0.08}$	0.724$_{\pm 0.10}$	0.580$_{\pm 0.12}$	**0.424**$_{\pm 0.01}$
MLP ResNet-50 $\tau(\boldsymbol{g}^{res.})$	0.644$_{\pm 0.08}$	0.740$_{\pm 0.13}$	0.799$_{\pm 0.12}$	0.767$_{\pm 0.12}$	0.376$_{\pm 0.09}$	0.443$_{\pm 0.03}$
Radiomics PCA	0.680$_{\pm 0.07}$	0.569$_{\pm 0.09}$	0.661$_{\pm 0.09}$	0.611$_{\pm 0.09}$	3.592$_{\pm 1.80}$	0.563$_{\pm 0.03}$
MLP radiom. PCA	0.729$_{\pm 0.10}$	0.731$_{\pm 0.11}$	0.777$_{\pm 0.10}$	0.753$_{\pm 0.11}$	3.211$_{\pm 1.06}$	0.560$_{\pm 0.06}$
VGG PCA	**0.814**$_{\pm 0.11}$	**0.842**$_{\pm 0.10}$	0.876$_{\pm 0.08}$	**0.859**$_{\pm 0.09}$	42.53$_{\pm 16.44}$	0.598$_{\pm 0.07}$
MLP VGG PCA	0.775$_{\pm 0.10}$	0.831$_{\pm 0.12}$	0.877$_{\pm 0.10}$	0.853$_{\pm 0.10}$	**1.38**$_{\pm 0.81}$	0.540$_{\pm 0.07}$
ResNet-50 PCA	0.730$_{\pm 0.10}$	0.748$_{\pm 0.10}$	0.829$_{\pm 0.08}$	0.785$_{\pm 0.09}$	46.02$_{\pm 22.57}$	0.563$_{\pm 0.07}$
MLP ResNet-50 PCA	0.764$_{\pm 0.11}$	0.785$_{\pm 0.12}$	0.833$_{\pm 0.10}$	0.808$_{\pm 0.11}$	2.148$_{\pm 0.86}$	**0.528**$_{\pm 0.06}$

Table 1: Evaluation of feature stability linked to scan variation (average and standard deviation for 100 runs). From left to right: ICC, GMM cluster homogeneity (H), GMM cluster completeness (C), GMM cluster V-measure (V), average GMM cluster covariance (Cov.) and correlation with resolution (Cor.). The (↑/↓) signs indicate whether higher or lower results are better. Best results are marked in bold ($p - value < 0.001$).

Figures 7 and 8 illustrate PCA and t-SNE representations, respectively, to investigate the influence of standardization on class clusters for several training runs. It includes the following features: radiomics, MLP radiomics, VGG and MLP VGG.

4.2. Computational Time

The networks are implemented in Keras [40] with a TensorFlow backend. The computation time is reported in Table 2 using a Titan Xp GPU.

Table 2: Training and inference time (675 test slices) of the networks.

Method	Training time	Test time
MLP radiomics	42.5 s	25 ms
MLP VGG	337.3 s	3.7 s
MLP ResNet-50	252.6 s	3.2 s

4.3. Transfer Learning

The pre-training domain (natural color images from ImageNet) is distant from the task domain (CT textures with grey levels in HU). Yet, a good transferability of the pre-trained features is observed despite the limited amount of training data, as well as a quick convergence in finetuning and a good generalization to unknown textures. These results confirm previous studies showing

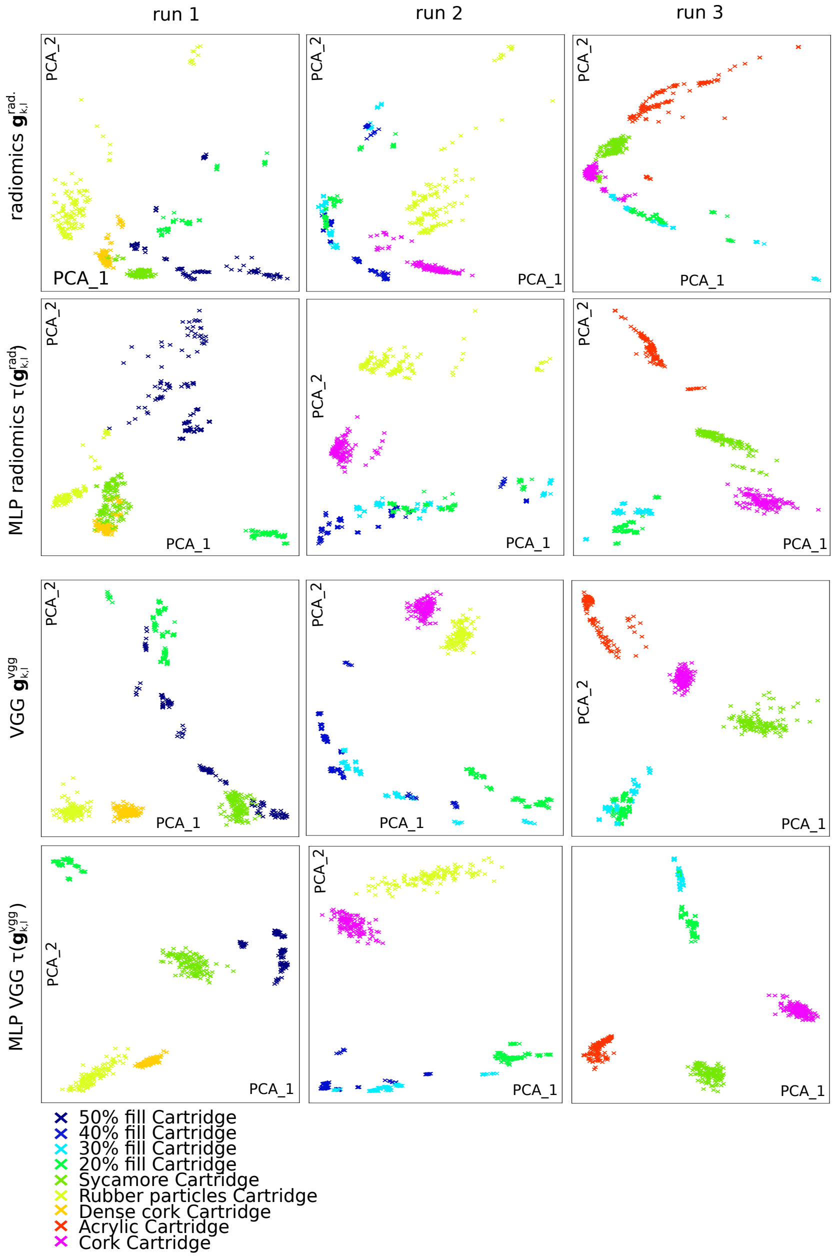


Figure 7: Two component PCA representation of the features on the test set for three distinct runs. The runs with different training/testing splits are shown from left to right. The correspondence between colors and texture types is shown at the bottom. The trained MLP reduces the intra-class variation and increases the inter-class variation of the radiomic and VGG features. Best viewed in color.

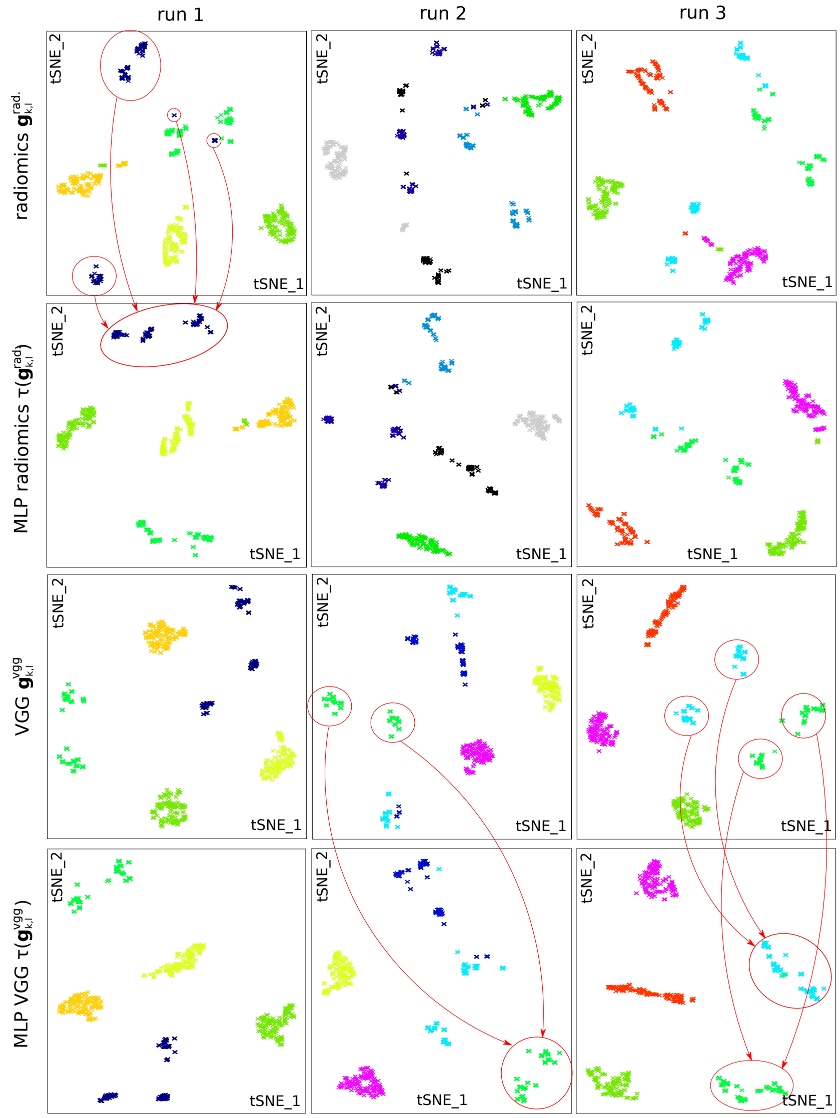
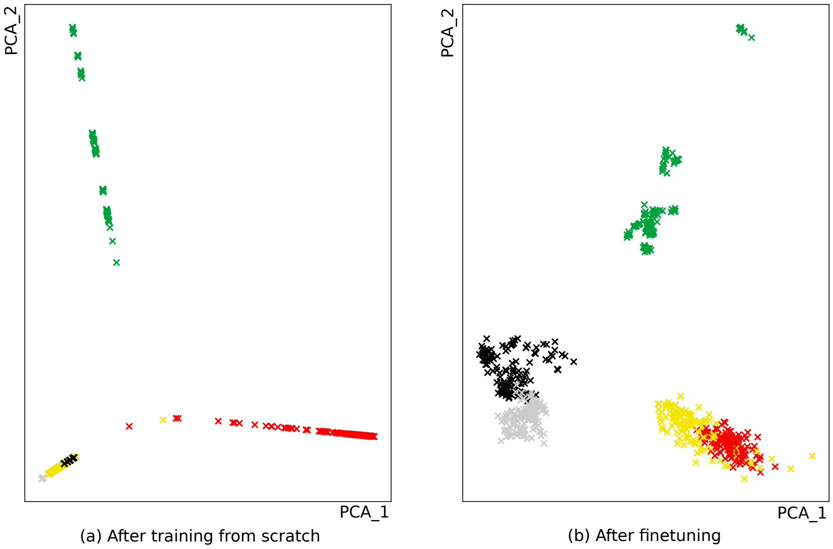


Figure 8: t-SNE representation of the features on the test set for three distinct runs (same runs as Figure 7). The trained MLP reduces the intra-class variation and increases the inter-class variation of the radiomic and VGG features (some improvements are shown by red arrows). Best viewed in color.



(a) After training from scratch

(b) After finetuning

Figure 9: Example of the first two PCA components of the features extracted on the test set from MLP ResNet-50 (a) trained from scratch and (b) finetuned from ImageNet pre-training. The features from scratch are highly correlated and non discriminative. Best viewed in color.

the need for transfer learning in computer vision [33–35] and medical imaging tasks [4, 24, 26, 27, 29–31] with limited training data. When training the CNNs from scratch, the dataset was too limited to obtain informative features that could characterize a ROI. Despite a texture classification accuracy reaching more than 90% (overall correct classification on the balanced test set), the networks were strongly relying on basic HU characteristics of the slices, without learning meaningful (texture) features. Figure 9 illustrates the first two PCA components of the features extracted from ResNet-50 trained from scratch. The features are mostly aligned on a single axis in the hidden space, which shows that they are neither informative nor complementary and that the network probably learned only the average intensity. Similar results were obtained with VGG trained from scratch.

4.4. Results with Unknown Scanners

In a second set of experiments, we evaluate the generalization of the proposed method to new scanners and protocols not used for training. In this setup, we still repeatedly split the training and test texture classes (5 training and 5 test texture classes), and also split the scans (8 training, 9 test scans). These results with the test scanners unknown to the models are summarized in Table 3. The hypothesis is that standardization over a few scanners also extends to unknown scanners, as at least some of the changes in the produced images are systematic, even if they are not the same.

	ICC$_{(\uparrow)}$	H$_{(\uparrow)}$	C$_{(\uparrow)}$	V$_{(\uparrow)}$	Cov.$_{(\downarrow)}$	Cor.$_{(\downarrow)}$
Radiomics $\boldsymbol{g}^{rad.}$	$0.615_{\pm 0.04}$	$0.581_{\pm 0.06}$	$0.672_{\pm 0.08}$	$0.623_{\pm 0.07}$	$0.277_{\pm 0.16}$	$0.584_{\pm 0.02}$
MLP radiomics $\tau(\boldsymbol{g}^{rad.})$	$0.766_{\pm 0.06}$	$0.678_{\pm 0.09}$	$0.732_{\pm 0.08}$	$0.704_{\pm 0.09}$	$0.270_{\pm 0.08}$	$0.519_{\pm 0.02}$
VGG $\boldsymbol{g}^{vgg}$	$0.697_{\pm 0.05}$	$\mathbf{0.829}_{\pm 0.09}$	$\mathbf{0.880}_{\pm 0.08}$	$\mathbf{0.854}_{\pm 0.08}$	$0.306_{\pm 0.05}$	$\mathbf{0.284}_{\pm 0.07}$
MLP VGG $\tau(\boldsymbol{g}^{vgg})$	$\mathbf{0.809}_{\pm 0.07}$	$0.821_{\pm 0.11}$	$\mathbf{0.880}_{\pm 0.08}$	$0.848_{\pm 0.10}$	$\mathbf{0.169}_{\pm 0.06}$	$0.291_{\pm 0.07}$
ResNet-50 $\boldsymbol{g}^{res.}$	$0.427_{\pm 0.05}$	$0.737_{\pm 0.13}$	$0.838_{\pm 0.09}$	$0.782_{\pm 0.11}$	$0.498_{\pm 0.07}$	$0.288_{\pm 0.05}$
MLP ResNet-50 $\tau(\boldsymbol{g}^{res.})$	$0.609_{\pm 0.11}$	$0.760_{\pm 0.14}$	$0.824_{\pm 0.11}$	$0.790_{\pm 0.13}$	$0.350_{\pm 0.09}$	$0.290_{\pm 0.05}$
Radiomics PCA	$0.660_{\pm 0.06}$	$0.566_{\pm 0.08}$	$0.644_{\pm 0.07}$	$0.602_{\pm 0.07}$	$3.496_{\pm 2.04}$	$0.572_{\pm 0.03}$
MLP radiomics PCA	$0.730_{\pm 0.10}$	$0.697_{\pm 0.11}$	$0.757_{\pm 0.08}$	$0.725_{\pm 0.09}$	$3.704_{\pm 1.02}$	$0.636_{\pm 0.09}$
VGG PCA	$\mathbf{0.812}_{\pm 0.11}$	$\mathbf{0.854}_{\pm 0.10}$	$\mathbf{0.892}_{\pm 0.08}$	$\mathbf{0.872}_{\pm 0.09}$	$41.54_{\pm 18.92}$	$0.295_{\pm 0.10}$
MLP VGG PCA	$0.772_{\pm 0.10}$	$0.831_{\pm 0.10}$	$0.884_{\pm 0.09}$	$0.856_{\pm 0.09}$	$\mathbf{1.126}_{\pm 0.51}$	$0.291_{\pm 0.09}$
ResNet-50 PCA	$0.729_{\pm 0.11}$	$0.759_{\pm 0.13}$	$0.849_{\pm 0.09}$	$0.800_{\pm 0.11}$	$33.92_{\pm 11.87}$	$0.291_{\pm 0.11}$
MLP ResNet-50 PCA	$0.746_{\pm 0.12}$	$0.774_{\pm 0.14}$	$0.827_{\pm 0.11}$	$0.799_{\pm 0.13}$	$1.98_{\pm 0.86}$	$\mathbf{0.285}_{\pm 0.09}$

Table 3: Evaluation of feature stability with test scanners different from the training scanners. See Table 1 for metric descriptions. Best results are marked in bold $(p - value < 0.001)$.

4.5. Standardization with Domain Adversarial Training

For the third set of experiments, we evaluate the domain adversarial training with known test scanners. As a first result of training with the domain adversarial network, we investigate the capacity of discarding the domain (scanner) information. For this, we train with different values of $\lambda \in \{-1, 0, 1\}$. When $\lambda = 1$, the adversarial part of the network behaves in a normal domain adversarial scheme. Layers l_{d1} and l_{d2} (see Figure 5) try to recover the domain information from the features $h = \tau(\boldsymbol{g}^{m}_{k,l,s})$ in a categorical domain classification, while layer l_h is finetuned to limit this recovery, together with learning the texture classification. When $\lambda = 0$, l_{d1} and l_{d2} still learn to recover the domain information but l_h is now only trained to classify the texture. In this case, the quantitative features are trained in a similar manner as in the previous experiments. Finally, when $\lambda = -1$, l_{d1}, l_{d2} and l_h are trained together to recover the domain information and no adversarial training takes place. Table 4 summarizes the domain classification accuracy reached after training using these setups. It is worth noting that the test sets are balanced and that a random domain classification accuracy is $\frac{1}{17} = 5.9\%$ whereas a random texture label classification accuracy is $\frac{1}{5} = 20\%$. The CNNs and the radiomics MLP are trained for 100 and 500 epochs respectively, similarly to the previous experiments. As the domain classification is more challenging than the texture classification, we balance the neural network update by giving more weight to the domain adversarial update than the texture one (1 and 0.5 respectively). We observe that for all the setups, the texture classification still reaches nearly 100% accuracy as discriminating the texture volumes is a simpler task than discriminating the scanner of origin. We also report the domain ICC, which corresponds to an intra-scan correlation coefficient.

λ	MLP radiomics		MLP VGG		MLP ResNet-50	
	Accuracy	ICC	Accuracy	ICC	Accuracy	ICC
-1	$34.6_{\pm 4.4}$	$0.14_{\pm 0.05}$, $0.20_{\pm 0.07}$	$23.7_{\pm 2.4}$	$0.09_{\pm 0.02}$, $0.19_{\pm 0.08}$	$38.6_{\pm 9.9}$	$0.11_{\pm 0.03}$, $0.29_{\pm 0.11}$
0	$16.6_{\pm 2.0}$	$0.13_{\pm 0.07}$, $0.17_{\pm 0.08}$	$6.5_{\pm 1.0}$	$\mathbf{0.06}_{\pm 0.05}$, $\mathbf{0.09}_{\pm 0.08}$	$8.0_{\pm 1.2}$	$\mathbf{0.05}_{\pm 0.03}$, $\mathbf{0.10}_{\pm 0.09}$
1	$\mathbf{8.9}_{\pm 1.2}$	$\mathbf{0.12}_{\pm 0.08}$, $\mathbf{0.15}_{\pm 0.10}$	$\mathbf{5.9}_{\pm 1.0}$	$0.08_{\pm 0.05}$, $0.10_{\pm 0.07}$	$\mathbf{5.6}_{\pm 0.9}$	$0.08_{\pm 0.04}$, $0.14_{\pm 0.10}$

Table 4: Domain classification accuracy (% overall correct classification with standard deviation) and domain ICC of the extracted features with different domain adversarial setups. The domain ICC represents the averaged correlation of the features with the scanners of origin. Lower accuracy and lower ICC is better as it shows that the adverarial part is not able to recover the domain information (i.e. scanner typex) and that the features withing domains are less correlated. We report the ICC with all the features followed by the ICC with the first four PCA components. Best results in bold for each network ($p - value < 0.001$).

5. Discussions

The results with known scanners (reported in table 1) show that training the MLP on top of the radiomic features ($\tau(\boldsymbol{g}^{rad.})$, we drop the k and l indexes for simplicity) improves the generalization and standardization with respect to the scanner type, acquisition protocols and reconstruction algorithms. The radiomic features benefit more from the MLP stabilization method than the deep features. Recall, that the dimensionality of the non-standardized deep features $\boldsymbol{g}^{vgg}$ and $\boldsymbol{g}^{res.}$ is substantially larger than the other features (4,096 and 2,048 vs. approximately 100). The standardized deep features, in particular $\tau(\boldsymbol{g}^{vgg})$, are more robust to scanner variation than the ones trained on radiomic features with a better ICC and clustering evaluation.

The results (ICC, homogeneity, completeness and V-measure in Table 1) obtained after applying PCA to the features confirm the superiority of the transformed deep features over the radiomic ones. The low ICC and clustering measures of radiomic feature PCA components and their transformed counterparts reflect the feature correlation, their limited informativeness and discriminative power in medical applications. The results are provided with four PCA components, yet similar results are observable for other numbers of principal components as well. It is worth noting that the large covariance of the PCA clusters is a consequence of retaining the components with the largest variance.

Consistently, we notice from the PCA and t-SNE visualizations (Figures 7 and 8) that the trained MLP features clearly improve the stability of the radiomic features, while the VGG and transformed VGG features are even more stable with a better intra-class clustering and inter-class separability. The standardization improvement from the learned transform is less evident for the VGG features, as also shown in Table 1. Some improved clusters in the t-SNE representations are shown by red circles in Figure 8.

The correlation of the features with the pixel spacing of the scanners (see last column of Table 1) is lower with the trained features. In particular, the radiomic features $\boldsymbol{g}^{rad.}$ present the largest correlation, in line with other studies [2, 7, 15]. The deep features and the standardization method reduce this correlation, illustrating the improved robustness and generalization of the features. The VGG network performs globally better on this task than ResNet. This is potentially due to the latter's depth, leading to a difficulty to generalize with the limited amount of training data and a larger amount of information extracted on the scanner of origin.

From another visualization of the features with t-SNE and PCA (see Figure 10 where the colors represent the scanners), we notice that two scans (S2 and T2, see [7] for details on the scanners and protocols) lead to correlated features that are well separated from the other scans. Other such correlations include slices from scanners GE4, GE5 and GE6 as well as P1, P3, P4 and P5. These scans were mostly acquired with scanners from the same manufacturer, with similar acquisition protocols and reconstruction algorithms. This observation could be further investigated by a correlation analysis of features from pairs of scans, yet this is out of the scope of this paper. As shown in the results, our standardization approach limits these correlation effects. The results with unknown scanners (table 3) show that a network trained with images from a set of training scanners generalizes the learned standardization transform $\tau(\boldsymbol{g})$ to images from new scanners. It extends the potential of this standardization method as it is not necessary to scan the phantom with all the scanners to obtain stable features from them. For instance, our method can be used with images acquired by scanners no longer in operation for retrospective studies, which is extremely important for the secondary use of image data that is regularly needed for research studies.

Comparing Tables 1 and 3, one can see that the results using a subset of scans for training are almost as good as those using all the scans. This observation underlines the excellent generalization to unknown scanners. It is worth noting, however, that the number of test samples is different in these two experiments. For a valid comparison, we evaluated the stability of features with a training set including all scans similarly to Section 4.1 but the same number of test scans as in Section 4.4, i.e. nine scans. Similar results were obtained with a minor deterioration of the results when using this subset of training scans.

Noticeably, as reported in the domain adversarial Section 4.5 (Table 4), the MLP with radiomic features benefits more from the domain adversarial training ($\lambda = 1$) than the deep features. The domain accuracy and domain ICC are significantly reduced from the training without adversarial and with non-adversarial ($\lambda = 0$ and $\lambda = -1$ respectively). The deep CNN features seem to already discard the domain information when training to classify only textures ($\lambda = 0$) by extracting meaningful information about the texture that is not correlated with the scanner type. This observation is even more striking with the VGG features from which the domain can almost not be recovered at

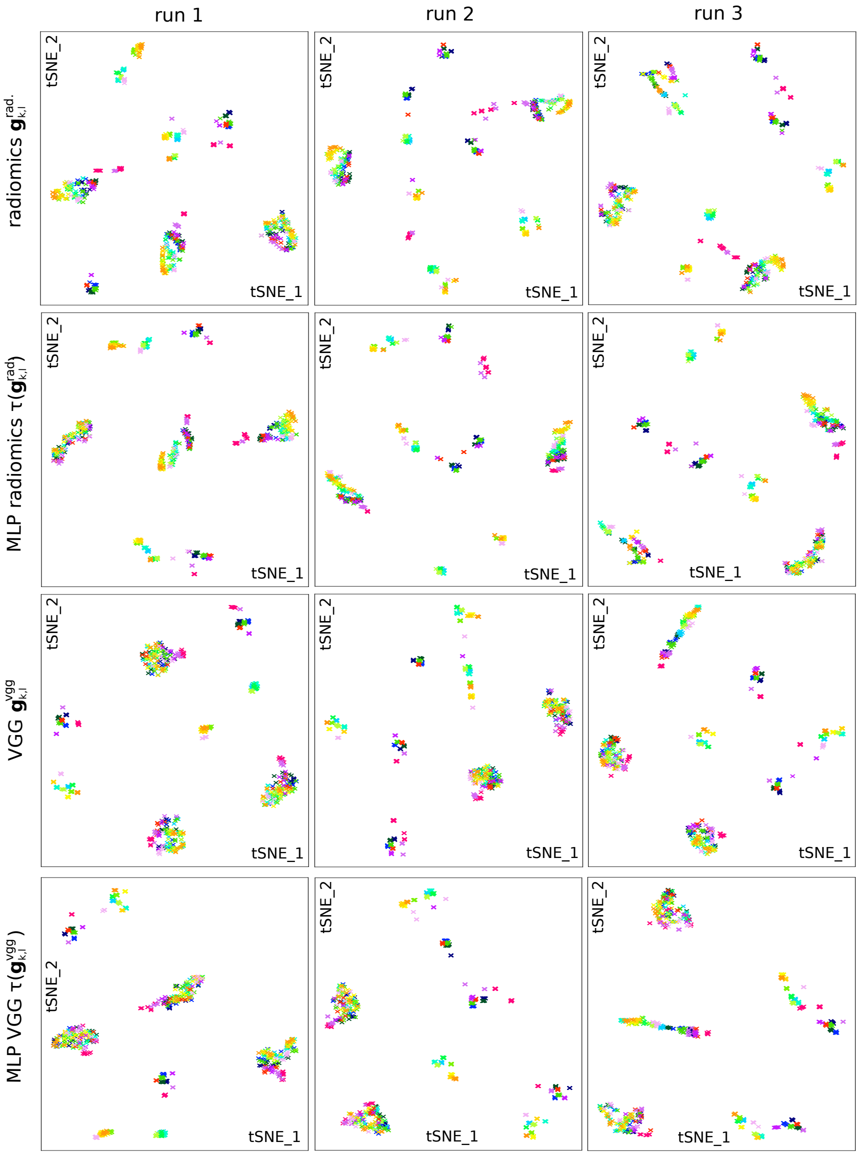


Figure 10: t-SNE visualization of the features on the test set with colours corresponding to the 17 scans of origin

. The features are the same and may be viewed with those in Figure 8 (in which colours correspond to the texture class). Various clusters are formed by features extracted from the same scans, although this is reduced by the proposed standardization method. Best viewed in color.

all (6.5%). The scanner ICC of the CNN features is, therefore, not reduced by using domain adversarial training despite the scanner (domain) accuracy reduction, which is the optimization goal of the adversarial training. If more normalization by domain adversarial training is required, more weight can be given to the domain adversarial contribution in the network update. Finally, the texture ICC and clustering, as reported in Table 1, are not improved by domain adversarial training since discarding the domain information does not entail a better texture intra-class clustering.

A drawback of the proposed feature transformation with and without domain adversarial is the limited direct interpretability of the generated features as compared to some classical radiomic features. However, although interesting studies have investigated the interpretation of radiomic features and their link with biological characteristics, standard radiomic features are rarely interpreted directly and individually. Prediction performance of a set of descriptors is usually analyzed and validated, which is also possible with the proposed learned features.

6. Conclusions

This paper proposes an approach to standardize image features to make them robust to scanner variability by training a neural network on top of radiomic or deep features extracted from CT images. The network learns a function $\tau(\boldsymbol{g}_{k,l})$ that outputs features independent of the scanner type l. The standardization is based on the idea that the same features should be extracted from different scans of the same phantom volume and that standardizing for a set of characteristic textures should generalize to other types of textures and tissue types. The standardized discriminative and quantitative features can be extracted from patient scans to characterize ROIs (e.g. texture in a tumor region) independently from the acquisition and reconstruction protocols. This robustness is expected to improve performance and generalization for retrieval, computer-assisted diagnosis, predictive treatment planning and prognosis, in particular when using data from several hospitals or varying acquisition methods. The presented results particularly motivate the use of deep CNN features in radiomics studies with data from more than a single scanner type, as more stable features can be obtained than with classic hand-crafted texture features.

We showed that the learned standardization can be generalized to new images from unknown scanners, which is important as it is common to use old data for which such a standardization can no longer be done. We also evaluated domain adversarial training to remove information about the scanner and protocol from the extracted features. In this setting, the network representation is trained to enable an accurate texture classification while avoiding the recovery of a scan of origin classification. This method should be used to avoid intra-scan clustering of texture features that does not underline true physio-pathological tissue changes. We are confident that this approach will play an important role in the standardization of features with larger datasets and/or other architectures and training schemes in future applications. Besides, as shown in [37], adversarial training can be used in combination with augmentation and normalization techniques with complementary benefits.

Finally, although this study did not evaluate real patient data, the texture phantom was designed to mimic actual biomedical tissue types (particularly non small cell lung cancer commonly analyzed in radiomics) and it allowed a controlled analysis to isolate the variation due to scanner variation. Future work is foreseen on the evaluation of the approach on prognosis, prediction and diagnosis of real patient data, which requires the extraction of visual features as image biomarkers.

7. Acknowledgments

This work was partly supported by the Swiss National Science Foundation under grants PZ00P2_154891 and 205320_179069. The Titan Xp used for this research was donated by the NVIDIA Corporation. The authors state no conflict of interest and have nothing to disclose.

References

[1] R. Gillies, A. Anderson, R. Gatenby, D. Morse, The biology underlying molecular imaging in oncology: from genome to anatome and back again, Clinical radiology 65 (7) (2010) 517–521.

[2] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. W. L. Aerts, A. Dekker, D. Fenstermacher, D. B. Goldgof, L. O. Hall, P. Lambin, Y. Balagurunathan, R. A. Gatenby, R. J. Gillies, Radiomics: The process and the challenges, Magnetic Resonance Imaging 30 (9) (2012) 1234–1248.

[3] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, et al., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nature communications 5.

[4] R. Paul, S. H. Hawkins, M. B. Schabath, R. J. Gillies, L. O. Hall, D. B. Goldgof, Predicting malignant nodules by fusing deep features with classical radiomics features, Journal of Medical Imaging 5 (1) (2018) 011021.

[5] R. T. Leijenaar, S. Carvalho, E. R. Velazquez, W. J. Van Elmpt, C. Parmar, O. S. Hoekstra, C. J. Hoekstra, R. Boellaard, A. L. Dekker, R. J. Gillies, et al., Stability of FDG-PET radiomics features: An integrated analysis of test-retest and inter-observer variability, Acta oncologica 52 (7) (2013) 1391–1397.

[6] Y. Balagurunathan, Y. Gu, H. Wang, V. Kumar, O. Grove, S. Hawkins, J. Kim, D. B. Goldgof, L. O. Hall, R. A. Gatenby, et al., Reproducibility and prognosis of quantitative features extracted from CT images, Translational oncology 7 (1) (2014) 72–87.

[7] D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A. K. Jones, L. Court, Measuring CT scanner variability of radiomics features, Investigative radiology 50 (11) (2015) 757.

[8] F. H. van Velden, G. M. Kramer, V. Frings, I. A. Nissen, E. R. Mulder, A. J. de Langen, O. S. Hoekstra, E. F. Smit, R. Boellaard, Repeatability of radiomic features in non-small-cell lung cancer [18F] FDG-PET/CT studies: impact of reconstruction and delineation, Molecular imaging and biology 18 (5) (2016) 788–795.

[9] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H. J. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer Research 77 (21) (2017) e104–e107.

[10] D. Mackin, X. Fave, L. Zhang, J. Yang, A. K. Jones, C. S. Ng, et al., Harmonizing the pixel size in retrospective computed tomography radiomics studies, PloS One 12 (9) (2017) e0178524.

[11] F. Orlhac, S. Boughdad, C. Philippe, H. Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, I. Buvat, A post-reconstruction harmonization method for multicenter radiomic studies in PET, Journal of Nuclear Medicine (2018) jnumed–117.

[12] N. Antropova, B. Q. Huynh, M. L. Giger, A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets, Medical physics.

[13] Z. Li, Y. Wang, J. Yu, Y. Guo, W. Cao, Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma, Scientific reports 7 (1) (2017) 5467.

[14] P. E. Galavis, C. Hollensen, N. Jallow, B. Paliwal, R. Jeraj, Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters, Acta Oncologica 49 (7) (2010) 1012–1016.

[15] M. Robins, J. Solomon, J. Hoye, E. Abadi, D. Marin, E. Samei, How reliable are texture measurements?, in: Medical Imaging 2018: Physics of Medical

Imaging, Vol. 10573, International Society for Optics and Photonics, 2018, p. 105733W.

[16] A. Zwanenburg, S. Leger, M. Vallières, S. Löck, Image biomarker standardisation initiative - feature definitions, CoRR abs/1612.0.

[17] X. Fave, D. Mackin, J. Yang, J. Zhang, D. Fried, P. Balter, D. Followill, D. Gomez, A. Kyle Jones, F. Stingo, et al., Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer?, Medical physics 42 (12) (2015) 6784–6797.

[18] J. Yan, J. L. Chu-Shern, H. Y. Loi, L. K. Khor, A. K. Sinha, S. T. Quek, I. W. Tham, D. Townsend, Impact of image reconstruction settings on texture features in 18F-FDG PET, Journal of Nuclear Medicine 56 (11) (2015) 1667–1673.

[19] M. Bogowicz, O. Riesterer, R. Bundschuh, P. Veit-Haibach, M. Hüllner, G. Studer, S. Stieb, S. Glatz, M. Pruschy, M. Guckenberger, et al., Stability of radiomic features in CT perfusion maps, Physics in medicine and biology 61 (24) (2016) 8736.

[20] J. E. van Timmeren, R. T. Leijenaar, W. van Elmpt, J. Wang, Z. Zhang, A. Dekker, P. Lambin, Test–retest data for radiomics feature stability analysis: Generalizable or study-specific?, Tomography 2 (4) (2016) 361–365.

[21] K. Yasaka, H. Akai, D. Mackin, L. Court, E. Moros, K. Ohtomo, S. Kiryu, Precision of quantitative computed tomography texture analysis using image filtering: A phantom study for scanner variability, Medicine 42 (2017) 60–88.

[22] C. Caramella, A. Allorant, F. Orlhac, F. Bidault, B. Asselain, S. Ammari, P. Jaranowski, A. Moussier, C. Balleyguier, N. Lassau, et al., Can we trust the calculation of texture indices of CT images? A phantom study, Medical physics.

[23] A. Traverso, L. Wee, A. Dekker, R. Gillies, Repeatability and reproducibility of radiomic features: A systematic review, International Journal of Radiation Oncology* Biology* Physics.

[24] R. Paul, M. Shafiq-ul Hassan, E. G. Moros, R. J. Gillies, L. O. Hall, D. B. Goldgof, Stability of deep features across CT scanners and field of view using a physical phantom, in: Medical Imaging 2018: Computer-Aided Diagnosis, Vol. 10575, International Society for Optics and Photonics, 2018, p. 105753P.

[25] S. Dutta, J. Fan, D. Chevalier, Study of CT image texture using deep learning techniques, in: Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment, Vol. 10577, International Society for Optics and Photonics, 2018, p. 1057712.

[26] V. Andrearczyk, P. F. Whelan, Using filter banks in convolutional neural networks for texture classification, Pattern Recognition Letters 84 (2016) 63–69.

[27] M. Cimpoi, S. Maji, I. Kokkinos, A. Vedaldi, Deep filter banks for texture recognition, description, and segmentation, International Journal of Computer Vision 118 (1) (2016) 65–94.

[28] H. Zhang, J. Xue, K. Dana, Deep TEN: Texture encoding network, arXiv preprint arXiv:1612.02844.

[29] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 96 (21).

[30] R. Paul, S. H. Hawkins, Y. Balagurunathan, M. B. Schabath, R. J. Gillies, L. O. Hall, D. B. Goldgof, Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma, Tomography: a journal for imaging research 2 (4) (2016) 388.

[31] V. Andrearczyk, A. Depeursinge, H. Müller, Learning cross-protocol radiomics and deep feature standardization from CT images of texture phantoms, SPIE Medical Imaging 2019 (in press).

[32] A. Depeursinge, Multiscale and multidirectional biomedical texture analysis: Finding the needle in the haystack, in: Biomedical Texture Analysis, Elsevier, 2017, pp. 29–53.

[33] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[35] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, Journal of Machine Learning Research 17 (59) (2016) 1–35.

[36] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al., Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, in: International Conference on Information Processing in Medical Imaging, Springer, 2017, pp. 597–609.

[37] M. W. Lafarge, J. P. Pluim, K. A. Eppenhof, P. Moeskops, M. Veta, Domain-adversarial neural networks to address the appearance variability of histopathology images, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2017, pp. 83–91.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[39] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proc. of the 3rd International Conference on Learning Representations (ICLR), 2015.

[40] F. Chollet, et al., Keras, `https://keras.io` (2015).

Vincent Andrearczyk received a double Masters degree in electronics and signal processing from ENSEEIHT, France and Dublin City University, in 2012 and 2013 respectively. He completed his PhD degree on deep learning for texture and dynamic texture analysis at Dublin City University in 2017. He is currently a post-doctoral researcher at the University of Applied Sciences and Arts Western Switzerland with a research focus on deep learning for medical image analysis, texture feature extraction and standardization and multi-modal data analysis.

Adrien Depeursinge received the B.Sc. and M.Sc. degrees in electrical engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland with a specialization in signal processing. From 2006 to 2010, he performed his Ph.D. thesis on medical image analysis at the University Hospitals of Geneva (HUG). He then spent two years as a Postdoctoral Fellow at the Department of Radiology of the School of Medicine at Stanford University. He has currently a joint position as an Associate Professor at the Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), and as a Senior Research Scientist at EPFL.

Henning Müller studied medical informatics at the University of Heidelberg, Germany and did his PhD degree at the University of Geneva, Switzerland in 2002. Since 2002 Henning has been working in medical informatics at the University Hospitals of Geneva where he habilitated in 2008 and was named titular professor in 2014. Since 2007 he has been a professor in business informatics at the HES-SO Valais in Sierre, Switzerland. In 2015-2016, Henning was a visiting professor at the Martinos Center in Boston, MA, USA part of Harvard Medical School and the Massachusetts General Hospital (MGH) working on projects in medical imaging and system evaluation in the context of the Quantitative Imaging Network (QIN) of the National Cancer Institutes (NCI).

List of Figures

9 Example of the first two PCA components of the features extracted on the test set from MLP ResNet-50 (a) trained from scratch and (b) finetuned from ImageNet pre-training. The features from scratch are highly correlated and non discriminative. Best viewed in color.

10 t-SNE visualization of the features on the test set with colours corresponding to the 17 scans of origin