Learning Cross-Protocol Radiomics and Deep Feature Standardization from CT Images of Texture Phantoms

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]Nuclear Medicine and Molecular Imaging Department, Lausanne University Hospital, Lausanne, Switzerland;
[c]University of Geneva (UNIGE), Geneva, 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 introduces a new method to transform image features in order to improve their stability across scanners. This method 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. In this setting, variations in extracted features will be representative of true physio-pathological tissue changes in the scanned patients. This approach uses a publicly available texture phantom dataset and can be applied to both hand-crafted radiomic and deep features.

Keywords: quantitative imaging, neural network, standardization

1. INTRODUCTION

Radiomics aims at extracting and analyzing large amounts of quantitative features (e.g. volume, shape, intensity and texture) from medical images. The number of related papers has followed an exponential growth since the first publications in 2010.[1–3] Various organs and cancer types have been analyzed with radiomics including lungs,[3–8] liver,[9] breast,[9] head-and-neck[3] and brain gliomas.[10] 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.

Uncovering disease characteristics or predicting a response to treatment relies on the fact that these features describe the patients' biomarkers independently from the image acquisition device or protocol. The same person scanned in different hospitals or with different machines should ideally obtain the same features. Scanning protocols and machines are frequently changed over time and vary across hospitals. Radiomics biomarkers such as texture features can be strongly impacted by these changes.[6] Texture phantom images allow evaluating the variation of features extracted from different scanners and with varying protocols.[6] Several studies have shown a high variability of radiomic features across scanners, limiting their interpretability and comparison.[4,6,7,11,12] 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 definition and implementation on the feature variation is tackled by the Image Biomarker Standardization Initiative (IBSI).[13] Various studies have evaluated the reproducibility and stability of texture features and the influence of scanner variation and reconstruction settings.[4,5,7,11,14] 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. Texture phantom images allow evaluating the variation of features extracted from different scanners and with varying protocols of an unchanged body. It avoids repeatedly

Further author information: Send correspondence to V.A. Email: vincent.andrearczyk@hevs.ch

exposing a patient to radiations and tiring protocols[6] 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.[6] CT images were pre-processed in[15] by resampling and filtering to standardize image pixel sizes, resulting in a reduced variability of radiomic features. Finally, an excellent systematic review of the repeatability and reproducibility of radiomic features with and without phantom studies was recently presented in.[16] We refer to this work for more details on the mentioned analyses and a more exhaustive literature review.

The adequacy of deep learning for texture analysis and medical imaging was extensively demonstrated in various studies.[17–19] This paper is, therefore, not dedicated to yet another illustration of the informativeness and generalization of deep features in a classic recognition or prediction medical task but rather to demonstrate that the performance of quantitative image descriptors can be further improved by using phantom images to obtain stable features across scanners with a robust generalization to unknown texture classes. The obtained features that are independent of the acquisition and reconstruction methods allow clinicians to better evaluate and compare patient biomarkers over time and across scanners and hospitals.

2. METHODS

For a given Region Of Interest (ROI) in an image, a stack of slices or a volume I, we extract a feature vector $\boldsymbol{g}$. The elements of $\boldsymbol{g}$ are generally radiomic features that were shown to vary strongly for a same phantom texture scanned across different scanners.[4, 7, 11] Using a phantom of texture volumes, we train a neural network on top of the radiomic or deep features to classify slices acquired using 17 scanners. In this way, hidden layers yield consistent intra-class values while conserving inter-class variations. In addition, the transformed features $\tau(\boldsymbol{g})$ become standardized (i.e. reduced variations from one scanner to another) 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.

2.1 Pre-processing

The features are extracted from 16cm^2 slices as provided with the dataset.[6] The slices are resized using bilinear interpolation to either (a) in-plane pixel spacing of 1mm^2 for the radiomic features as suggested in Mackin et. al,[6] or (b) to the CNN input size for the deep features (224×224). The Hounsfield Units (HU) range $[-1409, 747]$ is linearly converted into the interval $[0, 255]$ for the input to the CNNs. 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 textures after interpolation. For the CNNs, a three channel input is obtained by duplication in order to use pre-trained networks. As a standard procedure, the images used with the pre-trained CNNs are centered (ImageNet mean subtraction) and scaled (division by the ImageNet standard deviation).

2.2 Feature Extraction

In the first set of experiments, we use radiomic features as a baseline, extracted with the pyRadiomics toolbox.[20] A 97-dimensional feature vector is extracted from each slice, including intensity (i.e. first order statistics) and texture (e.g. co-occurrence and size-zone matrices) features. In the second set of experiments, we use VGG19[21] and ResNet-50[22] to extract deep features of dimension 4096 and 2048 respectively from the texture slices. We remove the prediction layer and extract the penultimate layer output. By averaging the features within each cartridge (i.e. volume of texture, see Section 2.4), we obtain the feature vectors $\boldsymbol{g}_m$, where $m \in \{rad., vgg, res.\}$. Note that this dataset is developed for the analysis of 2D slices, although 3D features could be used with the same standardization method on other datasets.

2.3 Feature Transformation

We design a two-layer Multi-Layer Perceptron (MLP) with 100 hidden neurons (with dropout 0.5 and ReLU activation) that takes the radiomic or deep features as input and is trained to output a class probability with five training classes. This design is motivated by the use of a simple non-parametric yet non-linear transformation where the 100 neurons correspond to the radiomic feature dimensionality (97) for comparison. After training (see Section 2.4), the output of the hidden layer is used as a 100-dimensional feature vector and averaged within cartridges (resulting in $\tau(\boldsymbol{g}_m)$)*. The networks are trained by optimizing the prediction, i.e. the last layer (softmax activated) of the MLP but the feature representation is extracted from the hidden layer.

$\tau(\boldsymbol{g}_m)$ transforms the feature space into a more discriminative and clustered space, in which the features are more stable to scanner variability. This is achieved by learning from the set of training classes a scanner invariance of the learned representation that will generalize to unknown tissue types, as confirmed by the results in Section 3. An overview of the feature extraction and training is illustrated in Fig. 1.

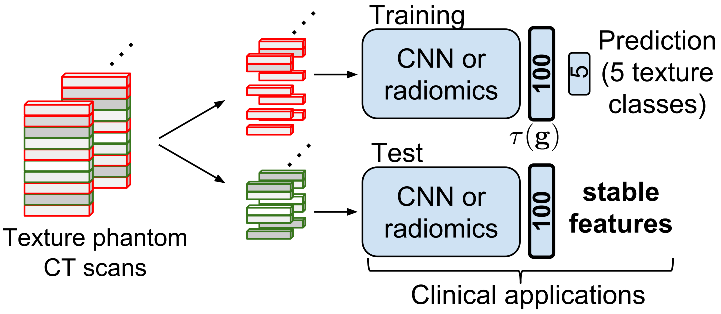

Figure 1: Overview of the feature extraction and training. The CNN is either VGG-19 or ResNet-50 from which the prediction layer is removed.

2.4 Dataset and Training

We use the Credence Cartridge Radiomics (CCR) phantom dataset.[6] The physical phantom contains ten volumes of textures (cartridges) as shown in Figure 2. The cartridges 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 several scanners (from the manufacturers GE, Philips, Siemens and Toshiba), in different centers and with different acquisition protocols and reconstruction algorithms. More information about the scans can be found in.[6] The dataset is publicly available and the experiments are thus fully reproducible. We randomly split the dataset (100 repetitions) to train the networks on half of the texture types and evaluate on the other half (five texture labels) from all the 17 scanners. A number of slices ranging from 6 to 11 depending on the scanners and cartridges are available from each cartridge, as proposed in.[6] From 1360 available slices, we obtain training and testing sets composed of 675 to 685 slices depending on the random splits. The feature vectors are extracted for all the test slices and averaged within the cartridges ($\boldsymbol{g}_{rad.} \in \mathbb{R}^{97}$, $\boldsymbol{g}_{vgg} \in \mathbb{R}^{4096}$, $\boldsymbol{g}_{res.} \in \mathbb{R}^{2048}$ and $\tau(\boldsymbol{g}_m) \in \mathbb{R}^{100}$). The sparsity of the neuron activations results in a few features of $\tau(\boldsymbol{g}_m)$ 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 $\tau(\boldsymbol{g}_m)$ may, therefore, be reduced to $d \leq 100$.

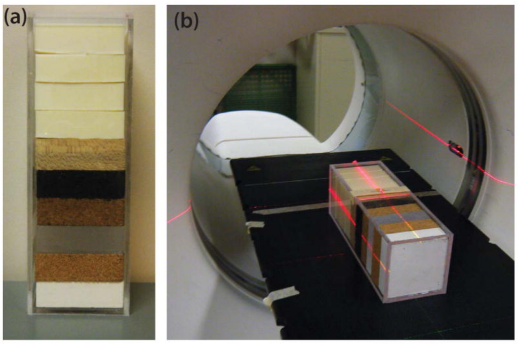

Figure 2: CCR texture phantom volume. Figure reproduced from ([6]).

[*] A transform is learned for each feature extraction method but we keep the same symbol τ for simplicity.

The CNNs are pre-trained on ImageNet[23] to obtain informative deep features despite the limited amount of training data. They are finetuned end-to-end by adding fully-connected layers in place of the MLP described in Section 2.3. The networks are trained with the Adam optimizer with an initial learning rate of 10^{-4}, average decays $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a batch size of 32. The radiomics MLP is trained for more epochs than the CNNs (500 vs. 100) as the former overfits less due to a reduced number of trainable weights and the pre-trained CNNs also enable a faster convergence. The random train/test split is reproduced 100 times with the same splits for all experiments, and the average and standard deviation are reported for each method.

2.5 Dimensionality Reduction

Excellent clustering of unknown textures can be obtained with a simple HU average as this measure separates the cartridges well. However, the informative and discriminative power of a single feature is limited in a real medical image analysis scenario. 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. Applying Principal Component Analysis (PCA) allows evaluating the intra-class variability along the directions of the largest variance in the feature space.

3. EXPERIMENTAL RESULTS

3.1 Evaluation Metrics

Intra-class Correlation Coefficient (ICC): ICC evaluates the clustering of features using their correlation within classes.

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

where n is the number of targets (5 test classes) and k is the number of judges (17 scanners). BMS is the between target mean square, EMS the residual mean square and JMS the between judge mean square. This coefficient ranges from 0 to 1 with values close to 1 indicating high similarities between values of the same class. The ICCs are averaged across all the features. ICC is a standard evaluation method of feature stability, yet we provide 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 $\boldsymbol{g}$ and $\tau(\boldsymbol{g})$ from the test cartridges. 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,[2,6,12] the value of the features is highly correlated with the pixel spacing, limiting their comparison and interpretability. We measure, average and compare the Pearson correlation of the various extracted features with the slice sizes.

3.2 Results

Figure 3 illustrates the improvement of ICC with the proposed standardization method. Considering only the ICC, the radiomic features surprisingly obtain better results than the ResNet ones, although this is contrasted by the supplementary results. As mentioned previously, half of the texture types are used for training (five texture labels), the rest for testing with repeated random splits.

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

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

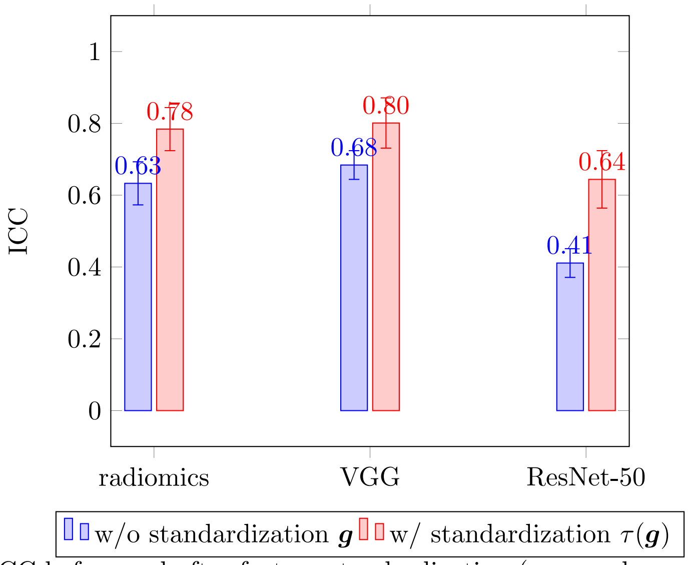

Figure 3: ICC before and after feature standardization (averaged over 100 runs).

	ICC$_{(\uparrow)}$	H$_{(\uparrow)}$	C$_{(\uparrow)}$	V$_{(\uparrow)}$	Cov.$_{(\downarrow)}$	Cor.$_{(\downarrow)}$
Radiomics $\boldsymbol{g}^{rad.}$	0.633$_{\pm0.06}$	0.564$_{\pm0.10}$	0.672$_{\pm0.09}$	0.611$_{\pm0.09}$	0.343$_{\pm0.18}$	0.577$_{\pm0.02}$
MLP radiom. $\tau(\boldsymbol{g}^{rad.})$	0.784$_{\pm0.06}$	0.723$_{\pm0.10}$	0.770$_{\pm0.08}$	0.745$_{\pm0.09}$	0.239$_{\pm0.07}$	0.510$_{\pm0.03}$
VGG $\boldsymbol{g}^{vgg}$	0.684$_{\pm0.04}$	**0.794**$_{\pm0.10}$	0.844$_{\pm0.08}$	0.817$_{\pm0.09}$	0.352$_{\pm0.05}$	0.504$_{\pm0.02}$
MLP VGG $\tau(\boldsymbol{g}^{vgg})$	**0.801**$_{\pm0.07}$	0.790$_{\pm0.11}$	**0.849**$_{\pm0.10}$	**0.817**$_{\pm0.10}$	**0.199**$_{\pm0.08}$	0.503$_{\pm0.04}$
ResNet-50 $\boldsymbol{g}^{res.}$	0.411$_{\pm0.04}$	0.681$_{\pm0.12}$	0.778$_{\pm0.08}$	0.724$_{\pm0.10}$	0.580$_{\pm0.12}$	**0.424**$_{\pm0.01}$
MLP ResNet-50 $\tau(\boldsymbol{g}^{res.})$	0.644$_{\pm0.08}$	0.740$_{\pm0.13}$	0.799$_{\pm0.12}$	0.767$_{\pm0.12}$	0.376$_{\pm0.09}$	0.443$_{\pm0.03}$
Radiomics PCA	0.680$_{\pm0.07}$	0.569$_{\pm0.09}$	0.661$_{\pm0.09}$	0.611$_{\pm0.09}$	3.592$_{\pm1.80}$	0.563$_{\pm0.03}$
MLP radiom. PCA	0.729$_{\pm0.10}$	0.731$_{\pm0.11}$	0.777$_{\pm0.10}$	0.753$_{\pm0.11}$	3.211$_{\pm1.06}$	0.560$_{\pm0.06}$
VGG PCA	**0.814**$_{\pm0.11}$	**0.842**$_{\pm0.10}$	0.876$_{\pm0.08}$	**0.859**$_{\pm0.09}$	42.53$_{\pm16.44}$	0.598$_{\pm0.07}$
MLP VGG PCA	0.775$_{\pm0.10}$	0.831$_{\pm0.12}$	**0.877**$_{\pm0.10}$	0.853$_{\pm0.10}$	**1.38**$_{\pm0.81}$	0.540$_{\pm0.07}$
ResNet-50 PCA	0.730$_{\pm0.10}$	0.748$_{\pm0.10}$	0.829$_{\pm0.08}$	0.785$_{\pm0.09}$	46.02$_{\pm22.57}$	0.563$_{\pm0.07}$
MLP ResNet-50 PCA	0.764$_{\pm0.11}$	0.785$_{\pm0.12}$	0.833$_{\pm0.10}$	0.808$_{\pm0.11}$	2.148$_{\pm0.86}$	**0.528**$_{\pm0.06}$

Table 1: Evaluation of feature stability due 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 pixel spacing (Cor.). The (↑/↓) signs indicate whether higher or lower results are better. Best result without and with PCA are in bold.

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

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

4. DISCUSSION

The transformed features $\tau(\boldsymbol{g}_m)$ largely improve the ICC and other performance measures from the original features $\boldsymbol{g}_m$, illustrating the improved standardization with respect to the scanner type and pixel spacing as well as a generalization to textures that were never seen by the networks. The radiomic features benefit more from the learned transformation than the deep features. Yet, the transformed deep features are globally more robust to scanner variation than the shallow radiomic ones with a significantly better clustering evaluation. The VGG network performs better on this task than ResNet in terms of ICC and clustering, maybe due to ResNet's depth and its difficulty to generalize with the limited amount of training data.

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 discriminatory power in medical applications. The results are provided with four PCA components, yet similar results are observable for other numbers of principal components. It is worth noting that the large covariance of the PCA clusters is a consequence of retaining the components with the largest variance.

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.[2,6,12] The deep features and the standardization method significantly 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.

The pre-training domain (natural color images) is distant from the task domain (CT textures in grey levels). Yet, a good transferability of the pre-trained weights is observed, as well as a quick convergence in finetuning and a good generalization to unknown textures.

A drawback of the proposed feature transformation, 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.

5. CONCLUSIONS

This paper demonstrates an approach to obtain image features $\tau(\boldsymbol{g}_m)$ that are robust to scanner variability by training a neural network on top of radiomic or deep features. 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 results in better performance and generalization for computer-assisted diagnosis, treatment planning and prognosis, in particular when using data from several hospitals or varying acquisition methods.

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.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation (grants PZ00P2_154891 and 205320_179069).

REFERENCES

1. R. Gillies, A. Anderson, R. Gatenby, and D. Morse, "The biology underlying molecular imaging in oncology: from genome to anatome and back again," *Clinical radiology* **65**(7), pp. 517–521, 2010.
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, and R. J. Gillies, "Radiomics: the process and the challenges," *Magnetic Resonance Imaging* **30**(9), pp. 1234–1248, 2012.
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**, 2014.

4. 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), pp. 1391–1397, 2013.
5. 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), pp. 72–87, 2014.
6. D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A. K. Jones, and L. Court, "Measuring CT scanner variability of radiomics features," *Investigative radiology* **50**(11), p. 757, 2015.
7. 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, and 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), pp. 788–795, 2016.
8. R. Paul, S. H. Hawkins, M. B. Schabath, R. J. Gillies, L. O. Hall, and D. B. Goldgof, "Predicting malignant nodules by fusing deep features with classical radiomics features," *Journal of Medical Imaging* **5**(1), p. 011021, 2018.
9. F. Orlhac, S. Boughdad, C. Philippe, H. Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, and I. Buvat, "A post-reconstruction harmonization method for multicenter radiomic studies in PET," *Journal of Nuclear Medicine* , pp. jnumed–117, 2018.
10. Z. Li, Y. Wang, J. Yu, Y. Guo, and W. Cao, "Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma," *Scientific reports* **7**(1), p. 5467, 2017.
11. P. E. Galavis, C. Hollensen, N. Jallow, B. Paliwal, and R. Jeraj, "Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters," *Acta Oncologica* **49**(7), pp. 1012–1016, 2010.
12. M. Robins, J. Solomon, J. Hoye, E. Abadi, D. Marin, and E. Samei, "How reliable are texture measurements?," in *Medical Imaging 2018: Physics of Medical Imaging*, **10573**, p. 105733W, International Society for Optics and Photonics, 2018.
13. A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, "Image biomarker standardisation initiative - feature definitions," *CoRR* **abs/1612.0**, 2016.
14. J. E. van Timmeren, R. T. Leijenaar, W. van Elmpt, J. Wang, Z. Zhang, A. Dekker, and P. Lambin, "Test–retest data for radiomics feature stability analysis: Generalizable or study-specific?," *Tomography* **2**(4), pp. 361–365, 2016.
15. 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), p. e0178524, 2017.
16. A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and reproducibility of radiomic features: A systematic review," *International Journal of Radiation Oncology* Biology* Physics* , 2018.
17. V. Andrearczyk and P. F. Whelan, "Using filter banks in convolutional neural networks for texture classification," *Pattern Recognition Letters* **84**, pp. 63–69, 2016.
18. M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi, "Deep filter banks for texture recognition, description, and segmentation," *International Journal of Computer Vision* **118**(1), pp. 65–94, 2016.
19. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis* **96**(21), 2017.
20. J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research* **77**(21), pp. e104–e107, 2017.
21. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* , 2014.
22. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

23. 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), pp. 211–252, 2015.