The Stability of Oncologic MRI Radiomic Features and the Potential Role of Deep Learning: a Review

Scalco Elisa[1], Rizzo Giovanna[1] and Mastropietro Alfonso[1]

[1]*Institute for Biomedical Technologies, National Research Council (ITB-CNR), Segrate (MI), Italy*

Abstract

The use of MRI radiomic models for the diagnosis, prognosis and treatment response prediction of tumors has been increasingly reported in literature. However, its widespread adoption in clinics is hampered by issues related to features stability. In the MRI radiomic workflow, the main factors that affect radiomic features computation can be found in the image acquisition and reconstruction phase, in the image pre-processing steps, and in the segmentation of the Region of Interest (ROI) on which radiomic indices are extracted.

Deep Neural Networks (DNNs), having shown their potentiality in the medical image processing and analysis field, can be seen as an attractive strategy to partially overcome the issues related to radiomic stability and mitigate their impact. In fact, DNN approaches can be prospectively integrated in the MRI radiomic workflow to improve image quality, obtain accurate and reproducible segmentations and generate standardized images.

In this review, DNN methods that can be included in the image processing steps of the radiomic workflow are described and discussed, in the light of a detailed analysis of the literature in the context of MRI radiomic reliability.

Keywords

Radiomics, reproducibility, repeatability, deep neural networks, MRI

Corresponding author

Scalco Elisa, PhD

Institute of Biomedical Technologies, National Research Council (ITB-CNR)

Via Fratelli Cervi, 93, 20054 Segrate (MI), Italy

elisa.scalco@itb.cnr.it

Introduction

The inclusion of quantitative information derived from medical imaging in predictive models has led in the last 10 years to an increasing number of studies adopting radiomics, especially in the oncological field (Scalco and Rizzo 2017, Liu *et al* 2019). Radiomic models has been proposed to improve tumor diagnosis and staging, to predict survival, tumor response to treatment and treatment-related toxicities (Liu *et al* 2019). MRI offers the possibility of a multi-parametric characterization of tumoral tissue, which is essential for an accurate tumor diagnosis and prognosis, and radiomics seems to be able to link the quantification of image-based tumor heterogeneity with its biological variability (Stanzione *et al* 2020). For this reason, a large portion of radiomic studies is focused on MRI that, thanks also to the recent technological improvements in several clinical institutes, allows the acquisition of high-quality images.

Radiomics was defined as the high-throughput extraction of large amounts of image features from radiographic images (Lambin *et al* 2012). In the oncologic literature, these features traditionally include shape and morphological features, primary statistics and histogram-based features and texture features. From the very beginning of radiomics definition, it was clear that radiomic features computation suffers from stability issues related to the non-standardization and high variability of image acquisition protocols, ROI definition, and image processing techniques. This is a relevant problem that would have affected the widespread adoption of radiomics in clinical practice (Davnall *et al* 2012). These issues are particularly evident for radiomic features computed from MRI, where image intensities are dependent on scanner manufacturer, magnetic field strength, sequence protocol, image reconstruction and image pre-processing. However, in the first years of radiomics development, the stability of radiomic features was mainly investigated for CT (Zhao 2021) and PET images (Zwanenburg 2019), but not for MRI (Yip and Aerts 2016, Traverso *et al* 2018). Since then, the number of studies focused on it has been growing.

A potential strategy to partially solve the problems related to radiomics stability consisted of attempts to standardize the radiomics workflow to reduce its variability. The most relevant of them is due to the Image Biomarker Standardization Initiative (IBSI) (Zwanenburg *et al* 2020), which tried to define the guidelines for features definition and computation. However, other factors related to image acquisition, Region-Of-interest (ROI) segmentation and image pre-processing are still challenges and require ad-hoc studies. Another way to overcome issues related to handcrafted features extraction may be the adoption of artificial intelligence solutions, and deep learning in particular, that has been recently proposed as a powerful alternative and/or integrative approach to radiomics (Ibrahim *et al* 2021). In fact, deep neural networks (DNNs) can learn features directly from the original images, thus removing many sources of variability. However, deep learning still presents several challenges, and it cannot be already efficiently adopted in place of standard radiomics (Hatt *et al* 2019). Nonetheless, DNNs have shown their potentiality not only in predictive models but also in several other applications related to image generation, image processing and image segmentation. The integration of DNN-based image processing techniques in the radiomic workflow can be seen as a potential strategy to increase radiomics stability, as able to improve image quality, obtain accurate and reproducible segmentations and generate homogeneous images.

This work aims to discuss the potentialities offered by DNNs approaches for MRI image generation, pre-processing and segmentation and their effects on radiomic stability. First, the reference literature about the main sources affecting radiomic features reproducibility and repeatability, when estimated from MRI imaging in oncological applications, is examined. Then, an overview of DNNs in oncologic MRI and their relations with radiomics is reported. Finally, the integration of DNN methods and its effect on MRI radiomics variability is discussed.

Sources of variability for MRI radiomics

Before starting a detailed description and discussion of radiomics stability on MR images, it is worthy to define how it is traditionally quantified in the literature. Stability can be assessed in two different ways, defined as *repeatability* (the variation in repeated measurements made on the same subject under identical conditions) and *reproducibility* (the variation in measurements made on a subject under changing conditions) (Bartlett and Frost 2008). In both cases, agreement (how close two measurements made on the same subject are) and reliability (the magnitude of the measurement error in observed measurements to the inherent variability in the 'error-free' level of the quantity between subjects) can be quantified using different metrics (Bartlett and Frost 2008). The most common agreement metric is the Bland-Altman plot, which estimates the 95% limits of agreement; reliability can be quantified through the Intraclass Correlation Coefficient (ICC), which allows the evaluation of test-retest, intra-rater and inter-rater variability using different formulas (Koo and Li 2016). ICC is the metric mostly used to assess radiomic reliability, it is based on ANOVA and it is applicable for all radiomic features that have continuous values. A systematic review about radiomics reliability quantified by this metric has been recently written by Xue et al. (Xue *et al* 2021a). Other popular metrics commonly adopted in repeatability and reproducibility studies are the Concordance Correlation Coefficient (CCC), the Coefficient of Variation (CoV or CV), Pearson/Spearman correlation and parametric/non-parametric statistical test for mean/median comparison (t-test, ANOVA test, Wilcoxon test, Friedmann test). However, there is no consensus about the best metric to be adopted, nor the cutoff value to separate stable and unstable features (Traverso *et al* 2018).

As already mentioned in the introduction, radiomics stability on MRI is mainly affected by image acquisition, image pre-processing and ROI segmentation, due to the availability of different scanners, protocols, reconstruction methods, to the different contouring performances of experts and to the wide choice of pre-processing algorithms (see Fig.1). Results of stability analyses may differ depending on the MRI contrast (T1, T2, diffusion, perfusion) and on the body district under examination. The generalization of the results is thus a complex task, and this is also the reason of the large number of studies published in the last two-three years. In the following sections, the major sources of variability for MRI radiomics computation will be treated in detail.

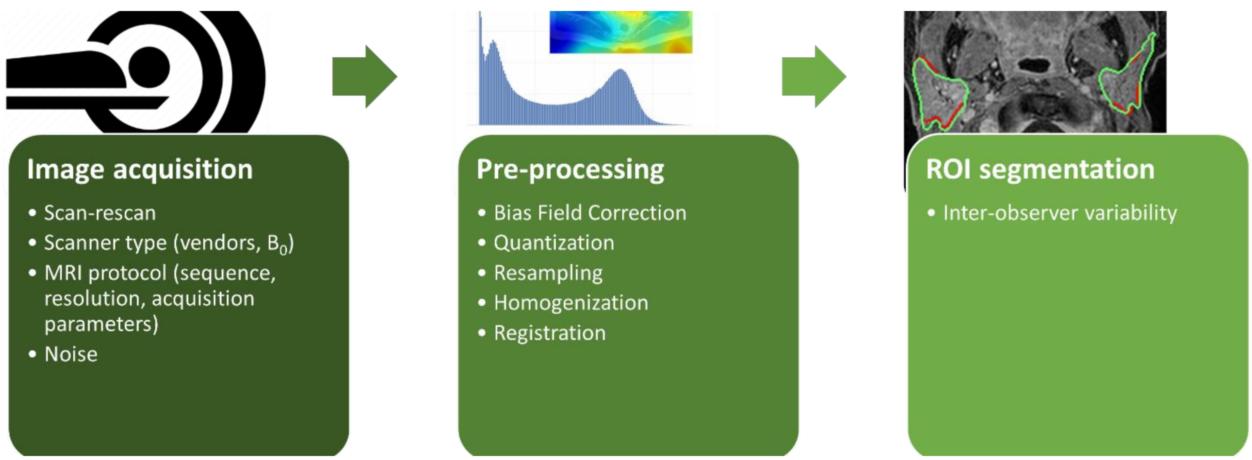

Figure 1. The major sources of variability that affect radiomics computed from MRI images.

I. Image acquisition

Image acquisition, with several different factors that have to be considered, chosen or optimized, is the main source of variability in the MRI radiomic analysis. In fact, a large part of the literature about radiomic feature stability studied the effect of different aspects related to MRI acquisition protocol. In the following points, the main results are summarized (see also Table 1):

- The use of standardized objects or phantoms is typically employed to obtain controlled MRI acquisitions - aimed to prevent unnecessary artifacts and object variability - that are required to assess the stability of radiomic features. This is remarkably true if the impact of several factors related to the MR scanner (e.g., manufacturers, models, magnetic field strength) and to the acquisition protocol (e.g., MR sequence, acquisition parameters, geometrical characteristics, noise) must be explored. As a general rule, the ideal phantom must be stable and mimic the target tissue in both texture characteristics and magnetic properties. However, the use of biological phantoms (e.g., vegetables, fruits, ex-vivo tissue) as target objects, which was proposed in several works (Baeßler *et al* 2018, Bernatz *et al* 2021, Dreher *et al* 2020, Shur *et al* 2021, Mi *et al* 2020, Mayerhoefer *et al* 2009), has limited comparability to in vivo human tissues in terms of relaxation properties and can undergo natural degeneration, which leads to changes in magnetic properties over time. This implies that a direct translation to in-vivo radiomic studies would probably lead to ambiguous results. A more robust approach is represented using synthetic phantoms that have the advantage of being more stable and can be customized to mimic, as much as possible, the heterogeneity and magnetic properties of target tissues (Ammari *et al* 2020, Bianchini *et al* 2021, Lee *et al* 2021). It is interesting to notice that, recently, the approach based on the generation of phantoms based on the use of 3D printing technology was proposed (Rai *et al* 2020) to obtain stable objects with customized shapes and patterns suitable for texture analysis.
- Repeatability – or scan-rescan stability - is one of the most assessed factors to evaluate radiomic features stability. Currently, there is no consensus on the type of features that can be considered repeatable following a scan-rescan test. In some papers, first-order features and histogram descriptors were described as highly repeatable (Eck *et al* 2021, Pandey *et al* 2021, Xue *et al* 2021b), conversely some works found that shape features have better stability (Baeßler *et al* 2018, Fiset *et al* 2019, Hoebel *et al* 2021), whereas in other papers there is not a clear definition of groups for stable features. Furthermore, the percentage of robust features changes remarkably among papers. Interestingly, the repeatability metrics can change depending on the type of target object that is scanned. In fact, they are higher in the case of phantoms if compared to the *in vivo* acquisitions (Lee *et al* 2021), and, additionally, they are lower for tumor tissue if compared to the healthy tissue (Mahon *et al* 2019). As to the repeatability with and without repositioning, Bianchini et al. demonstrated that the scan-rescan test without repositioning has a higher repeatability (>90% of the features was stable) if compared to the scan-rescan with repositioning (80% of stable features) (Bianchini *et al* 2021). In general, it should be emphasized that the

generalizability of repeatability results coming from single center datasets must not be overstated since they can lead to ambiguous conclusions.

- Using MRI systems from different manufactures, even considering the same magnetic field strength, can reduce the number of reproducible features and this issue has to be considered in a multicentric study. Bianchini et al., found that excellent reproducibility between scanners was observed in 4.6% to 15.6% of the analyzed features, at fixed imaging parameters. Furthermore, they observed a higher reproducibility when comparing two scanners of 1.5T and 3T from the same manufacturer than comparing two scanners of 1.5T with same acquisition parameters, but from different vendors (Bianchini *et al* 2021). Poor inter-scanner reproducibility was also observed in other studies (Fiset *et al* 2019, Pandey *et al* 2021) even with consistent parameters across different scanners (Mi *et al* 2020, Bianchini *et al* 2021). This can be probably explained by differences in technical design of the scanners that can induce variability in the image quality. Interestingly, it was found that extracted Apparent Diffusion Coefficient (ADC)-based radiomic features can be stable across multiple centers and scanners (Peerlings *et al* 2019). As to the choice of the more robust features, there is currently no consensus. For example, the majority of the radiomic features, including first-order features, failed to be repeatable across sites in Pandey et al. (Pandey *et al* 2021), whereas first-order statistics were the most robust in Rai et al. (Rai *et al* 2020).
- The static magnetic field strength B_0 (1.5T, 3.0T) affects the image acquisition and the subsequent feature reproducibility. In fact, it was found that radiomic features extracted on 1.5T should not be used interchangeably with 3T when evaluating texture features (Ammari *et al* 2020, Bianchini *et al* 2021). Only one paper didn't find any significant effect of magnetic field strength (Pandey *et al* 2021).
- Image resolution is another relevant determinant for radiomic features reproducibility (Ammari *et al* 2020, Eck *et al* 2021, Baeßler *et al* 2018, Mayerhoefer *et al* 2009, Roy *et al* 2020). It is not clear how changes in resolution can affect the stability since there is no agreement among papers. Mayerohoefer et al. stated that texture features of all categories are increasingly sensitive to acquisition parameter variations with increasing spatial resolution (Mayerhoefer *et al* 2009), whilst Eck et al. reported a reduced stability when the resolution is increased (Eck *et al* 2021). Furthermore, Baeßler et al. described an increase in stability performance associated with an increase of resolution, but this was dependent to the specific MRI sequence (Baeßler *et al* 2018). Finally, Roy et al. argue that the sensitivity of radiomic features to resolution cannot be overstated

since clinical image acquisition pipelines are well standardized within an institution and between institutions (Roy *et al* 2020), thanks to numerous initiatives such as the quantitative imaging network (QIN) and the quantitative imaging biomarker alliance (QIBA). Only one paper reported no differences in radiomic features for different resolutions (Dreher *et al* 2020).

- Regarding the specific MR sequences that are typically used in clinical studies, it was described that even the choice of pulse sequence or contrast weighting can influence the stability of radiomic features. However, as for the other determinants, results in literature were contradictory. In fact, it was reported that the number of stable features was higher for those calculated from FLAIR than from T1-w and T2-w images (Baeßler *et al* 2018), whilst Lee et al. showed that the overall stability for T1-weighted images was slightly lower than that for T2-weighted images (Lee *et al* 2021), whereas Roy et al. stated that features derived from T1w images tended to be more susceptible to noise than T2-w images (Roy *et al* 2020). McHugh et al. suggested that, while quantitative values may be expected to be more repeatable than signal intensities, quantitative maps may be more sensitive to motion and thus affect feature extraction (McHugh *et al* 2021). Regarding diffusion MRI, Dreher et al. found that, under *ex-vivo* conditions, DWI provided robust radiomics features with those from ADC being slightly less robust than from raw DWI (b=500, 1,000 s/mm^2) (Dreher *et al* 2020). Lu et al., in a scan/rescan assessment, found that features computed from ADC images were more repeatable if extracted from a *habitat* region, i.e. the aggressive tumor-like area characterized by restricted diffusion (Lu *et al* 2020). Although we understand that different sequences have intrinsic differences related to the specific pulse scheme, we believe that changes in feature stability are mainly related to other factors, such as sequence parameters that can affect image quality, as discussed in the next point.
- Several studies have tried to explore how acquisition parameters can affect the stability of the radiomic features. Bianchini et al. found a reduction of feature stability with increasing Echo Time (TE) intervals in T2-w images, whereas no changes are attributable to Repetition Time (TR) variations (Bianchini *et al* 2021). These results are just partially in agreement with those described by Eck et al. suggesting that features may be sensitive to changes in TR as well (4% of the descriptors showed excellent robustness). Yuan et al., in a rigorous and detailed study on an anthropomorphic phantom testing different acquisition factors, found that the radiomics feature values changed significantly with the varying TE, TR, Echo Train Length, number of startup RF pulses and flip angles pattern (Yuan *et al* 2021). Finally, Mayerhoefer et al. stated that if the spatial resolution is sufficiently high, variations in Number of Averages (NA), TR, TE, and Spectral

BandWidth (SBW) have little effect on the results of pattern discrimination (Mayerhoefer *et al* 2009). The influence of acceleration factors for image reconstruction (SENSE vs. compressed SENSE) on T2-w and T1-w MRI was also assessed, reporting a higher number of reproducible radiomic features for SENSE reconstruction than for compressed SENSE. Higher acceleration factors led to lower reproducibility and first order and GLCM indices resulted more stable than the other classes (Xue *et al* 2021c, Kim *et al* 2021). Finally, the influence of acquisition parameters was also evaluated in DCE-MRI showing that even the values of texture features extracted from DCE-MRI parametric maps can be influenced by temporal parameters (Crombé *et al* 2019).

All the previous discussed determinants affecting the radiomic reproducibility also have influences on image Signal to Noise Ratio (SNR). Noise represents a crucial aspect to be considered when extracting radiomic features. Differences in SNR could be explained by magnetic field strength, but SNR is also linked to the entire signal acquisition system (coils, electronic device etc.), to the specific sequence parameters, and to the intrinsic magnetic properties of the tissue (that influence T1, T2, T2*). Ammari et al. suggested that SNR should be considered in the interpretation of the texture indices (Ammari *et al* 2020), confirming what reported in a previous work, where it was demonstrated that MRI texture features are influenced by change in SNR (Mayerhoefer *et al* 2009). More recently, Yang et al. found that even under optimal low noise acquisition parameters, measured texture features can differ substantially from the true underlying texture of the tissue, and that details of image acquisition and reconstruction parameters do influence many of the observed quantitative texture features at clinically relevant noise levels (Yang *et al* 2018). SNR has been recognized as the main factor of radiomic instability also in case of increasing acceleration factors for SENSE reconstruction (Xue *et al* 2021c). Finally, Roy et al. suggested that the performance of radiomic features can be improved significantly if SNR > 28, in particular for rich texture images, and that features extracted directly from histogram were less sensitive to noise compared to second order and higher order features, except skewness and kurtosis (Roy *et al* 2020). As a general indication, we recommend to always consider the effect of image SNR and its changes over time in radiomic studies. In fact, alterations of the magnetic properties of the target objects, which are related to the evolution of the tissue under certain physio-pathological conditions (e.g., ageing, tumors, degenerative disorders, etc.), can affect the SNR even if the acquisition parameters are fixed.

Table 1. List of papers which evaluated the effects of MRI acquisition on radiomics stability

Reference	MRI Sequence	Target Application	Type of Analysis	Metrics for Stability
(Ammari *et al* 2020)	T1-w	Phantom; Brain	Reproducibility (magnetic field strength; image resolution)	t-test; Spearman Coefficient
(Baeßler *et al* 2018)	T1-w; T2-w	Phantom	Scan-Rescan with repositioning Reproducibility (image resolution)	CCC, DR, ICC
(Bernatz *et al* 2021)	T2-w; T2 relaxometry ; T1-w	Phantom	Scan-Rescan without repositioning	CCC, DR, ICC, Gini scores
(Bianchini *et al* 2021)	T2-w	Phantom	Scan-Rescan with and without repositioning Reproducibility (Scanner manufacturers; magnetic field strength; TR, TE)	CCC, ICC, CV
(Crombé *et al* 2019)	DCE	Soft-tissue sarcoma	Reproducibility (influence of temporal parameters)	CV
(Dreher *et al* 2020)	DWI	Phantom	Scan-Rescan with repositioning Reproducibility (image resolution)	ICC
(Eck *et al* 2021)	T2-w	Brain	Scan-Rescan Reproducibility (TR, TE, Resolution, Parallel Imaging Acceleration)	CCC
(Fiset *et al* 2019)	T2-w	Cervical cancer	Scan-Rescan with repositioning Reproducibility (Scanner manufacturers; magnetic field strength)	ICC
(Hoebel *et al* 2021)	T2-w; T1-w post contrast	Glioblastoma	Scan-Rescan with repositioning	ICC, Jensen-Shannon divergence
(Kim *et al* 2021)	T1-w	Head	Reproducibility (image reconstruction with acceleration factors)	CCC
(Lee *et al* 2021)	T2-w; T1-w	Phantom; Brain	Scan-Rescan with repositioning Reproducibility (Scanner manufacturers; number of excitation (NEX), slice thickness, phasing steps, and field of view)	CV, ICC
(Lu *et al* 2020)	T2-w; DWI (ADC map)	Prostate cancer	Scan-Rescan with repositioning	CCC
(Mahon *et al* 2019)	T2/T1-w; T1-w	Non-small cell lung cancer	Scan-Rescan without repositioning	CCC
(Mayerhoefer *et al* 2009)	T2-w	Phantom	Reproducibility (NA, TR, TE, SBW, Resolution)	Linear discriminant analysis, k nearest

				neighbor classification
(McHugh *et al* 2021)	T1-w pre/post contrast; T2-w; T1 relaxometry;	Colorectal Cancer Liver Metastases	Reproducibility (different sequences)	ICC, RC
(Merisaari *et al* 2020)	DWI (ADC and Kurtosis maps)	Prostate cancer	Scan-Rescan with repositioning	ICC
(Mi *et al* 2020)	T2-w	Cervix, Phantom	Reproducibility (Scanner manufacturers; TR, TE, Slice Thickness, Resolution)	CV, quartile coefficient of dispersion (QCD)
(Pandey *et al* 2021)	T2-w	Brain	Scan-Rescan with repositioning Reproducibility (Scanner manufacturers; magnetic field strength; spatial variability)	ICC, CV
(Peerlings *et al* 2019)	DWI (ADC maps)	Ovarian, lung, and colorectal liver metastasis cancer	Reproducibility (Scanner manufacturers; magnetic field strength)	CCC
(Rai *et al* 2020)	PD-w	Phantom	Reproducibility (Scanner manufacturers; magnetic field strength)	ICC, CV
(Roy *et al* 2020)	T1-w; T2-w	Breast cancer (clinical and preclinical MRI)	Reproducibility (Resolution, Noise)	CV, RPC, APC
(Saha *et al* 2017)	DCE	Breast cancer	Reproducibility (Scanner manufacturers; magnetic field strength; Slice Thickness)	Two-sample t -test, AUC
(Shur *et al* 2021)	T1-w; T2-w	Phantom	Scan-Rescan with repositioning Reproducibility (resolution; Noise)	CV, CCC
(Xue *et al* 2021b)	T2-w	Prostate cancer	Reproducibility (influence of time) Scan-Rescan with repositioning	ICC, Bland-Altman
(Xue *et al* 2021c)	T2-w	Phantom	Reproducibility (image reconstruction with acceleration factors) Scan-rescan with repositioning	ICC, CV
(Yang *et al* 2018)	T1-w; T2-w	Phantom; Glioma	Reproducibility (Noise)	CV
(Yuan *et al* 2021)	T2-w	Phantom	Reproducibility (TR, TE, Echo Train Length, RF pulse)	% deviation, ICC

II. Image pre-processing

For MR images, pre-processing plays a relevant role within the radiomic workflow since intensity values have to be corrected for different factors to harmonize images before features extraction. The most common pre-processing methods on MRI are pixel size resampling, correction for bias field inhomogeneities, intensity interpolation, discretization and image homogenization (through the intensity normalization, due to a lack of standard and quantifiable interpretation of MRI pixel values), and image registration, as schematically described in Table 2. Several studies have evaluated these factors, separately or in combination. The main results are synthetically reported in the following points:

- In general, the application of a pre-processing pipeline before feature extraction increases radiomics reproducibility, especially in case of multi-centric acquisitions (Chirra *et al* 2019). The combination of bias field correction, voxel interpolation and intensity normalization provided a high number of reproducible and discriminative features computed from prostate T2w-MRI images across four different institutes; however, the adoption of Laws image filters significantly reduced the reproducibility.
- Shape and histogram features were the less sensitive classes to image pre-processing (Moradmand *et al* 2020, Traverso *et al* 2019, Park *et al* 2021). On the contrary, higher order features showed poor reproducibility, since they provide a local measurement by looking at particular patterns inside gray values. For this reason, any image pre-processing may alter the intensity values, and consequently the matrices used for computation of texture features (Traverso *et al* 2019).
- Bias field correction showed a relatively low impact on radiomics stability on brain MRI (Li *et al* 2021, Shiri *et al* 2020, Moradmand *et al* 2020, Um *et al* 2019). However, this phenomenon was explained by Li et al. because their MRI images did not suffer from obvious bias field effects, thus the bias-corrected images were similar to the original images (Li *et al* 2021). Therefore, additional studies with more severe bias field inhomogeneities are further required.
- The choice of the quantization technique may have a relevant impact on features stability, but results differ depending on body site and MRI sequence (Simpson *et al* 2020, Carré *et al* 2020, Traverso *et al* 2020, Schwier *et al* 2019, Traverso *et al* 2019, Duron *et al* 2019). The IBSI community described two different approaches for intensity discretization: the fixed-bin number (FBN), or relative discretization, and the fixed-bin size (FBS), or absolute discretization. As a general guideline, the IBSI recommends the use of FBS for calibrated or quantitative imaging (e.g., CT, PET), whilst FBN was suggested for images with intensity values with arbitrary units (e.g., T1-w

and T2-w MRI). One of the first studies evaluating intensity discretization on multi-parametric MRI of lacrimal gland and breast tumors (Duron *et al* 2019), reported that in an inter-observer setting the highest reproducibility was found for the FBS approach. The authors explained that, when using relative discretization (FBN), the intensity range of the segmented image will impact the bin size, that determines the intensity histogram and the textural matrices used to derive radiomic features. Differences in segmentation in the inter-observer setting lead to slightly different contours providing intensity outliers, that change the intensity range of the ROI. On the contrary, the FBS method is independent from the intensity range of the segmented image and thus it may be less sensitive to inter-observer variability. Another study reported that for low-field strength MRI, the best discretization approach that provided more reproducible features was a fixed-bin number of 64 gray levels combined with histogram equalization. However, when the analysis was carried out on real cases, no optimal combination was found (Simpson *et al* 2020). Therefore, a clear consensus has not yet been reached among the community.

- Pixel resampling and interpolation have slight impact on radiomics stability (Li *et al* 2021, Park *et al* 2021, Traverso *et al* 2019, Molina *et al* 2017), especially for histogram and GLCM features. Results may differ if intensity normalization is performed before pixel resampling; in fact, Park et al. found lower reproducibility with respect to pixel resampling in normalized images than in the original (Park *et al* 2021). This is partly in contrast with the results reported by other studies about intensity normalization and the authors said that a possible explanation was that the normalization might transform the original image and partly loose the original information, even though it could mitigate the influence of various MRI acquisition protocols. This point highlights the needs of choosing a normalization method able to preserve the original content, as also discussed in the next point.
- Among the different intensity normalization methods tested in the literature, it seems that the histogram-matching (HM) algorithm, as proposed by Nyul et al. (Nyul *et al* 2000), was able to obtain higher robustness on T2-w and T1-w MRI (Li *et al* 2021, Crombé *et al* 2021, Carré *et al* 2020, Scalco *et al* 2020, Isaksson *et al* 2020), and to reduce radiomic features variability across different scanners and field strengths (Um *et al* 2019). However, Carré et al. discussed that HM increased stability for histogram features but changes the information for other texture classes; for them, the choice of quantization method impacts and the best choice, in brain MRI, seems to be a quantization with a fixed-bin number of 32 with z-score normalization (Carré *et al* 2020). In the case of prostate T2-w MRI, z-score and HM were both able to preserve the original information

content of the image better than the normalization by the mean intensity value of a homogeneous region (i.e., the urine in the bladder), with the HM slightly increasing the reproducibility for delta-radiomics and in the multi-observer setting (Scalco *et al* 2020). Lower repeatability when using normalization by the mean value of a ROI (muscular tissue in this case) was found also by Schwier et al. (Schwier *et al* 2019). On the other hand, Isaksson et al. reported that T2-w normalization of pelvic images according to pixel intensities within healthy prostate tissue was the most effective way of obtaining the desired properties of normalized images (Isaksson *et al* 2020). Another study also evaluated the predictive power of the features and reported that, although HM was able to provide most repeatable features from T2-w MRI, radiomics computed on quantitative T2 maps was the best compromise between stability and predictive performance (Crombé *et al* 2021). Nonetheless, in a multi-centric study, where radiomics was computed from different scanner, image normalization, even performed with HM, was not able to guarantee features reproducibility and harmonization in the feature domain using the ComBat approach (Johnson *et al* 2007) was essential to remove batch effect (Li *et al* 2021). Regarding ADC images, even if intensity values have a quantitative meaning, an intensity normalization over the population was suggested; in this case it was reported that normalization by the mean value of the urine in cervical cancer patients was the best choice to increase reproducibility (Traverso *et al* 2020).

- Finally, a separated evaluation was performed for the impact of image registration on radiomics. Image registration plays a relevant role when images acquired at different time points have to be realigned to obtain spatial correspondence (e.g., for treatment planning or treatment evaluation). Shiri et al. tried to analyze the effect of different registration algorithms (at increasing degrees of freedom and with different cost functions) on brain MRI images in a test-retest setting. They found high repeatability across all the conditions for a restricted set of features and a high percentage of repeatable features when affine transformation with the optimization of mutual information was considered (Shiri *et al* 2017).

Table 2. List of papers which evaluated the effects of image pre-processing on radiomics stability

Reference	**MRI sequence**	**Target application**	**Type of analysis**	**Metrics for stability**
(Carré *et al* 2020)	CE-T1-w, T2-w	Brain tumor	Effect of intensity normalization and discretization	ICC, CCC
(Chirra *et al* 2019)	T2-w	Prostate cancer	Effect of pre-processing (bias field correction, resampling and normalization) in a multi-site setting	CV

(Crombé *et al* 2021)	T2-w, T2 maps	Abdomen	Effect of intensity normalization and quantitative maps	ICC
(Duron *et al* 2019)	Mp-MRI	Lachrymal gland tumor, breast cancer	Effect of intensity discretization	ICC, CCC
(Isaksson *et al* 2020)	T2-w	Prostate cancer	Effect of intensity normalization	CCC
(Li *et al* 2021)	T1-w	Phantom + Brain tumor	Effect of bias field correction, image harmonization and features harmonization methods	Friedman and Wilcoxon test
(Molina *et al* 2017)	T1-w	Glioblastoma	Effect of pixel resolution and quantization	CV
(Moradmand *et al* 2020)	FLAIR, T1-w, CE-T1-w, T2-w	Glioblastoma	Effect of bias field correction and noise filtering	ICC, CCC, DR
(Park *et al* 2021)	T2-w, CE-T1-w	Cervical cancer	Effect of pixel resampling and interpolation	ICC
(Scalco *et al* 2020)	T2-w	Prostate cancer	Effect of intensity normalization	ICC
(Schwier *et al* 2019)	T2-w, ADC	Prostate cancer	Effect of intensity normalization, image filtering and discretization	ICC
(Shiri *et al* 2020)	T1-w, T2-w	Glioblastoma	Effect of bias field correction algorithm and image registration	ICC
(Simpson *et al* 2020)	Low-field-strength MRI	Phantom + Kidney tumor	Effect of quantization method and number of gray levels	CV, CCC
(Traverso *et al* 2019)	ADC	Rectal cancer	Effect of image filtering, pixel resampling and intensity discretization	CCC
(Traverso *et al* 2020)	ADC	Cervical cancer	Effect of intensity normalization and quantization	ICC
(Um *et al* 2019)	FLAIR, T1-w, CE-T1-w	Glioblastoma	Effect of image pre-processing methods	Wilcoxon test

III. ROI segmentation

The influence of ROI segmentation on radiomics stability is well recognized and extensively reported for every type of imaging modality. On MRI images, the effect of different segmentations mainly depends on tumor types and MRI sequences (see also Table 3); however, some general considerations can be drawn.

- Lower variability in ROI contouring between observers, measured by the Dice coefficient, generally increases radiomics stability. In fact, it was reported in different studies, focused on different type of tumors, that regions that are easy to define, or higher agreement between observers, leaded to higher reproducibility (Granzier *et al* 2020, Saha *et al* 2018).

- Histogram features were generally reported as the most stable against segmentation variability, confirming the results seen for the image pre-processing effects, whereas texture features were the most affected by it (Pati *et al* 2020, Tixier *et al* 2019, Lecler *et al* 2019, Lee *et al* 2017). Histogram features, being calculated from the histogram of the whole region, are less influenced by small contour variations. Shape features showed variable behavior, since they are generally found as the most stable together with histogram features in several studies (Gitto *et al* 2021, Pati *et al* 2020, Lecler *et al* 2019, Kurata *et al* 2021)**,** but in other cases they were found as the most impacted by segmentation (Tixier *et al* 2019). Texture features were the least stable to ROI delineation, as intensity quantization is needed for their computation. The intensity quantization levels are drastically affected by the change in boundaries, even by a few voxels of delineation, hence affecting higher-order features (Dutta *et al* 2021).
- Regarding DW-MRI, contradictory results were found for features extracted from ADC maps, which resulted as more reproducible than T2-w MRI for nasopharyngeal (Yang *et al* 2020) and prostate cancer (Wu *et al* 2019), but as the least reproducible for lacrimal gland tumor in a multi-parametric protocol consisting of axial T1-w MRI, axial DW-MRI, coronal DIXON-T2-w MRI and coronal post-contrast DIXON-T1-w MRI (Lecler *et al* 2019). One study evaluated the reproducibility of features extracted from of IVIM-MRI parametric maps of cervical cancer patients, finding high repeatability and reproducibility for features calculated on ADC, D and D* maps and only moderate for those computed on f maps (Chen *et al* 2020).
- Features that are proven to be stable against tumor segmentation were found to be also more clinically informative (Wu *et al* 2019, Cheng *et al* 2021). On the other hand, Tixier et al. suggested that high robustness should be used only as a criterion for feature exclusion, but not for feature inclusion when a prediction model is built, since highly reproducible features may not be useful predictors of clinical outcome (Tixier *et al* 2019). For this reason, several studies have adopted this approach to perform an initial features selection before model training (Stanzione *et al* 2021, Damascelli *et al* 2021, Suter *et al* 2020). However, it was also reported in a study focused on nasopharyngeal and breast cancer, that features that were not stable to the differences in ROI delineations may not result in poor prediction performance (Zhang *et al* 2019a). An approach to find at the same time stable and discriminative features without the need of multiple observers was proposed by Bologna et al. (Bologna *et al* 2018), by using a set of geometrical transformations

of increasing entity applied on the ROI. They built a distribution of ICC values calculated between the original and transformed contours to find a threshold of ICC able to discriminate between good, unstable, and non-discriminative features.

Table 3. List of papers which evaluated the effects of segmentation on radiomics stability

Reference	**MRI sequence**	**Target application**	**Type of analysis**	**Metrics for stability**
(Bologna *et al* 2018)	DW-MRI	Soft-tissue sarcoma; Oropharyngeal cancer	Method to find stable and discriminative features	ICC
(Chen *et al* 2020)	IVIM-MRI	Cervical cancer	Inter-observer reproducibility	ICC
(Cheng *et al* 2021)	T2-w MRI, CT	Pancreatic tumor	Inter-observer reproducibility	ICC
(Gitto *et al* 2021)	T1-w MRI, T2-w MRI	Chondrosarcoma	Inter-observer reproducibility; 2D vs 3D delineation	ICC
(Granzier *et al* 2020)	T1-w MRI post-contrast	Breast cancer	Inter-observer reproducibility	ICC
(Lecler *et al* 2019)	Multi-parametric MRI	Lacrimal gland tumor	Inter-observer reproducibility; identification of redundant features	ICC, CCC, Spearman Correlation
(Lee *et al* 2017)	T1-w MRI, FLAIR	Glioblastoma	Effect of semi-automatic segmentation methods	ICC
(Lu *et al* 2021)	T2-w MRI	Rectal cancer	Reproducibility between minimum and maximum delineation	ICC
(Pati *et al* 2020)	Multi-parametric MRI	Glioblastoma	Inter-observer reproducibility for tumor sub-compartments segmentation	Spearman correlation
(Saha *et al* 2018)	T1-w MRI	Breast cancer	Inter-observer reproducibility	ICC
(Tixier *et al* 2019)	T1-w MRI, CE-T1-w MRI	Glioblastoma	Inter-observer reproducibility	ICC
(Wu *et al* 2019)	T2-w MRI, DW-MRI	Prostate cancer	Inter-observer reproducibility; prediction of treatment response	ICC
(Yang *et al* 2020)	T2-w MRI, DW-MRI, PET	Nasopharyngeal cancer	Inter-observer reproducibility	ICC
(Zhang *et al* 2019a)	T1-w MRI, T2-w MRI, DW-MRI	Nasopharyngeal cancer; breast cancer	Effect of dilation, erosion and filtering on manual delineation; prediction of metastases	ICC

DNN models in medical imaging

The most popular deep learning models adopted in the context of medical imaging are Convolutional Neural Networks (CNN). The typical architecture of a CNN is composed of convolutional and pooling layers followed by non-linear activation functions; in the last layers, features are flattened to have the final class prediction (see Fig.2a) (Litjens *et al* 2017).

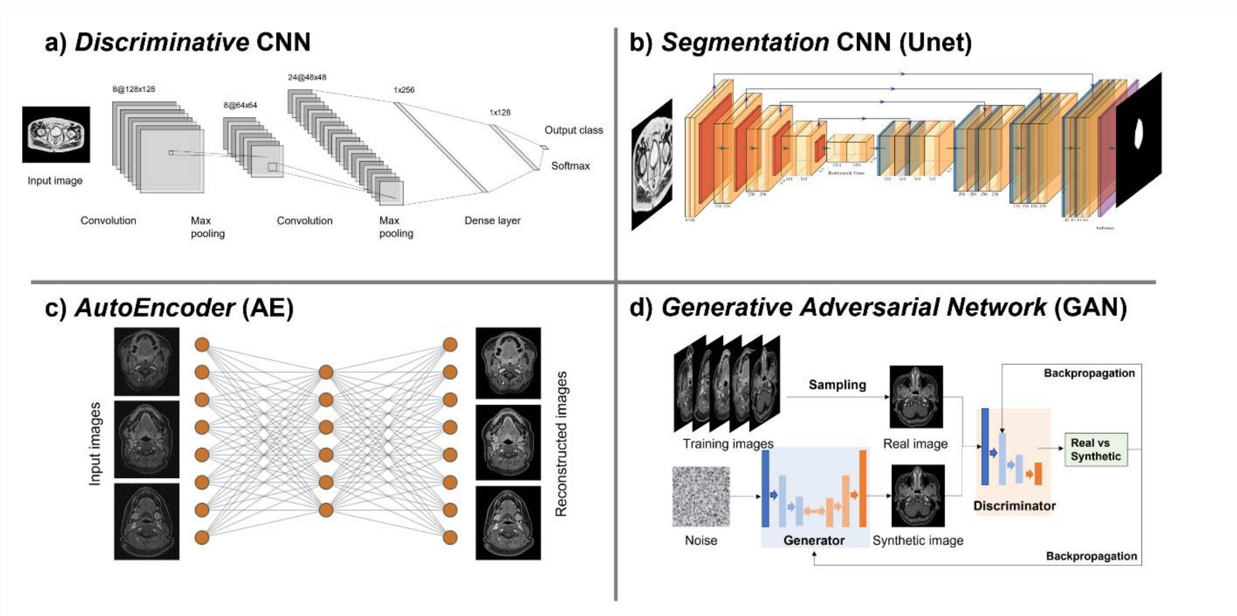

Figure 2. Typical architecture of a CNN. Convolutions are followed by max-pooling layers, and this is repeated several times to increase the depth of the network. Lastly, features are flattened, and a non-linear activation function performs the final class prediction.

Several types of CNN models have been proposed to face the different issues and applications that can be found in medical image processing, from image classification and segmentation to image synthesis and de-noising. Hereafter, we list the DNN models that are mainly adopted in the MRI context, with their properties and limitations. A detailed description of these architectures is beyond the scope of this review and can be found in previous excellent works (Parekh and Jacobs 2019, Afshar *et al* 2019, Litjens *et al* 2017, Yi *et al* 2019).

Discriminative CNNs try to extract features that better distinguish class labels by the minimization of prediction error (Fig.2a). There are several standard architectures that can be adopted, previously developed to solve different classification problems (e.g. *AlexNet*, *VGG*, *Inception*, etc). In this case, it is possible to use weights already optimized on larger datasets and thus a fine tuning approach may be adopted, with a reduced number of parameters to be trained, suitable when a small number of data is available. Alternatively, self-designed CNNs can be trained for a specific task, with the advantage of choosing the best performing architecture in terms of depth of the networks, type of layers, number of input and output channels. However, the training procedure requires a larger number of images compared to the first approach.

Segmentation CNNs are used for semantic segmentation, and the most popular model is the UNet (Ronneberger *et al* 2015) and its variants (Fig.2b). Here the encoder part of the network is followed by a decoder part with upsampling convolutional layers and skip connections to recover spatial information and to obtain segmentation maps at the same resolution as the input (Papadimitroulas *et al* 2021). One of the main advantages of these networks is that they can be effectively trained with a relatively small number of images.

Generative models represent the other major class of deep learning algorithms and they allow the generation of new synthetic images (Parekh and Jacobs 2019). *Autoencoder (AE) networks* are composed by a first encoder part that converts the input images into a latent space, and a second decoder part that takes the latent space and tries to reconstruct the input images (Fig.2c) (Afshar *et al* 2019). In this context, they are mainly adopted for image denoising (Stacked Denosing Autoencoders) and for image generation (Variational Autoencoders). The other family of generative models is represented by *Generative Adversarial Networks (GAN)*: they consist of two neural network architectures, the generator and the discriminator, competing against each other in a zero-sum game framework (Fig.2d) (Goodfellow *et al* 2014). Compared to conventional methods, these deep learning models for image synthesis have shown a superior mapping capability of non-linear relationships and a significant reduction in computing time (Wang *et al* 2021). Furthermore, both AEs and GANs allow the combination of neural networks with other techniques (e.g. wavelet-based transformations) or the integration with a priori-information, rather than adopting an end-to-end deep learning strategy. One of the main challenges in optimizing GANs is to find a good balance between the training of generator and the discriminator, that can lead to the generation of a limited and biased set of synthetic images, compared to the distribution of the target domain (Yi *et al* 2019).

DNNs in radiomics

These DNNs can be integrated in the radiomic workflow at different levels, as schematically depicted in Fig.3. Firstly, they can completely substitute the radiomic process, by adopting a CNN architecture trained end-to-end to directly perform class prediction (Fig.3a): in this case, the original image is used as input in the CNN and the final prediction is given as output, without the need of performing any other additional steps. This approach has the potential advantage of removing all the variability that characterize the typical radiomic workflow, from ROI segmentation to feature computation and model selection, but it requires large and highly curated dataset (Hatt *et al* 2019). In fact, training a network from scratch on a limited size dataset can easily lead to overfitting and to less effective performance; in addition, heterogeneous data are needed to increase the generalizability of the model. A possible solution can be found using data augmentation techniques, that can be carefully chosen with respect to the properties of the original dataset.

An alternative approach is to use pre-trained CNNs to extract learned features that can feed a machine-learning classifier as an alternative to or in combination with traditional radiomic features (Fig.3b); this second option can potentially solve the issues related to the limited datasets, typical for medical imaging, since networks weights have been already optimized and only fine tuning of the last layers may eventually be required.

Deep features have the advantage of retaining complementary information if compared to hand-crafted radiomic features; thus, the combination of these two feature families may improve the predictive performance of a traditional radiomic model (Afshar *et al* 2019). Being generated and extracted directly from the underlying data as the best performing features, they do not generally need feature selection techniques. Regularization techniques to reduce the risk of overfitting are applied directly during the network training phase. On the contrary, a limitation of deep features is the high correlations with the input data since these features are generated from that very data instead of being "a priori" defined (Lohmann *et al* 2020). Another limitations of deep learning approaches, both for end-to-end learning and deep features extraction, is in their low interpretability. This is an open research field that is achieving great attention, and several techniques have been already proposed to improve DNN explainability and reliability, especially for the clinical users. More details can be found in dedicated reviews (Reyes *et al* 2020, Tjoa and Guan 2020).

Considering the high requirements of large, heterogeneous and curated datasets, the low interpretability, and the not yet clear evidence that deep learning features can really outperform traditional radiomics (some works reported only marginal improvements), a complete replacement of radiomics by deep learning cannot be completely sustained yet. Especially in case of limited available data, the radiomic approach can still be preferable and a potential way to obtain improvement in the traditional radiomic models can be found in the integration of DNNs in the image processing steps of the radiomic workflow, before features computation (Fig.3c). In this way, the potentiality of DNNs is exploited to increase image quality, to have more accurate ROI segmentation and to create homogeneous datasets, starting from heterogeneous sources.

In the next section of this review, we will focus on this last point, by exploring the potential use of DNNs on MRI image processing to evaluate if and how these models can improve the stability of radiomic features, with respect to the traditional approach. In the following paragraphs, we will consider DNNs adopted in different tasks related to the MRI image processing steps; the papers presenting novel DNN approaches that have an explicit potential effect on image texture, discussed by the authors, are listed in Table 4.

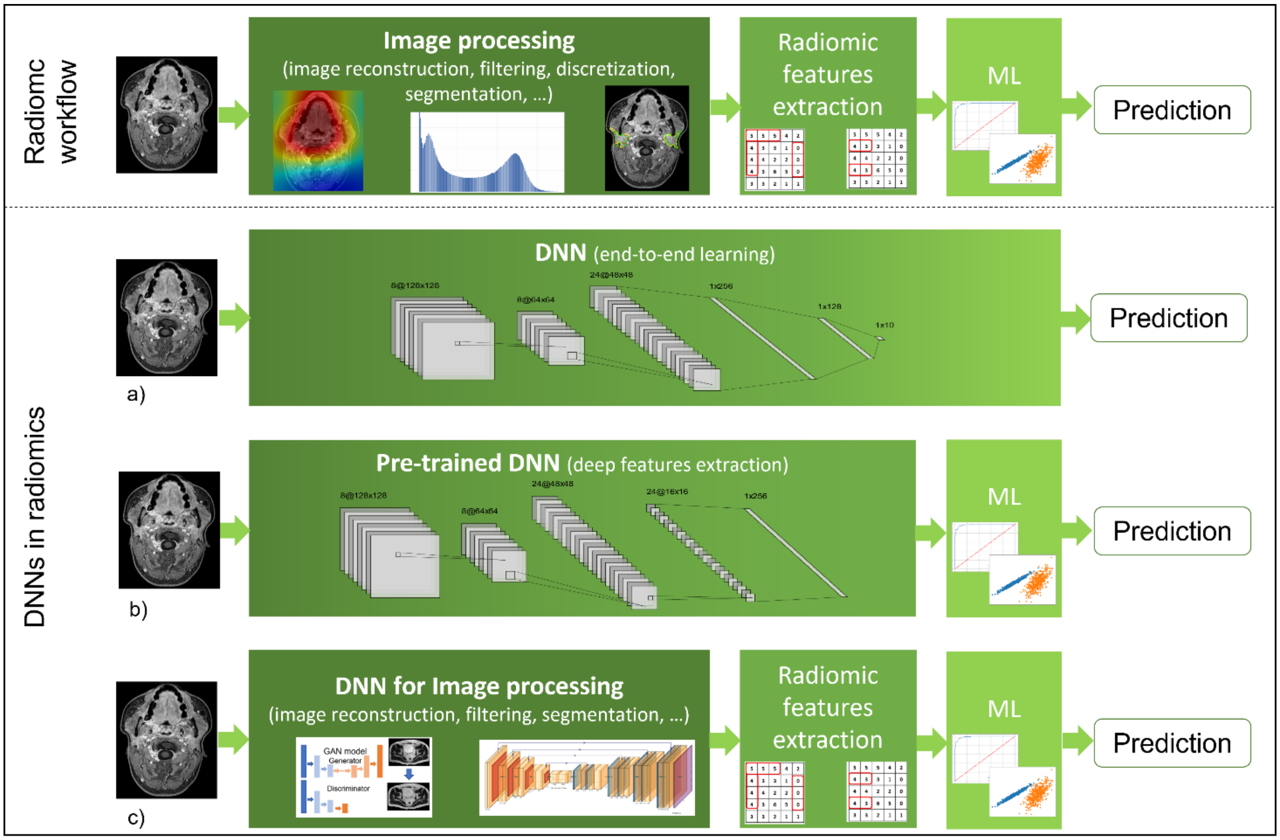

Figure 3. The different levels of DNN integration in the traditional radiomic workflow (top row). A) end-to-end DNN; b) pre-trained DNN for deep features extraction; c) integration of DNN for image processing before radiomic feature extraction. DNN: Deep Neural Networks; ML: Machine Learning.

a. DNNs for image synthesis

To overcome or at least mitigate the effects of MRI acquisition on radiomic features stability, the use of synthetic images generated by DNNs can be an interesting solution to be explored. As a result, one could obtain a more homogeneous dataset from images acquired with different scanners or acquisition protocols, or a set of images with increased quality (higher SNR or resolution), from which extract radiomic features.

Synthetic image generation has been explored for various applications in MRI, including translation between sequence types (e.g., T1-w to T2-w and vice versa), conversion from low-magnetic-field to high-magnetic-field MRI and improvement of image resolution and SNR. A recent review paper described in detail all the relevant technical aspects and the clinical applications related to the generation of synthetic images using deep learning (Wang *et al* 2021).

Deep learning-based methods used for image synthesis usually include AE and GAN models, which are built following a data-driven approach for mapping image intensity. Commonly, the algorithms consist of a training step to learn the actual mapping between the input images and their targets, and a prediction step that generate synthetic target images from inputs. Network training usually needs significant effort in terms of computational times but, once the network is trained, the generation of synthetic data usually takes only a few seconds.

Regarding the generation of images at higher magnetic fields, the most popular applications consider the synthesis of 7.0T MR images from 3.0T MR images, using different architectures, such as Unet (Qu *et al* 2020), auto-encoder (Zhang *et al* 2019b) or GAN (Nie *et al* 2018). All the previous methods can be easily adapted, with few changes, to perform the more suitable conversion from 1.5T MRI to 3.0T MRI. It is worth noticing that recently a DNN architecture reached a significant improvement in SNR and image quality without any loss of information, if compared to traditional techniques, even at very low magnetic field strength <0.3T (Koonjoo et al 2021). The impact on the image texture was not directly assessed in these works. Nonetheless, the use of these types of DNN models has shown superior results in terms of image visual perception (quantified by the Structural Similarity Index) compared to traditional methods (e.g. histogram matching algorithm) (Zhang *et al* 2019b). Furthermore, the AEs and GANs allowed the integration of layers that account for complementary information (spatial and wavelet domain, or spatial and frequency domain), able to enhance the structural information of the target image (Qu *et al* 2020, Zhang *et al* 2019b).

As to the improvement of image resolution, the use of GAN architecture was proposed to recover image details using hundreds of brain images (Galbusera et al 2018, Kim et al 2018)(Galbusera et al 2018, Kim et al 2018). The methods were proven to be effective and could be re-arranged to be used in several district and tissues. Kim et al. proposed the inclusion of cross-contrast high resolution images (e.g., T1w-MRI to generate high resolution T2w-MRI) in the training phase to improve the reconstruction performance. In this way, it was possible to recover both spatial and contrast details. Further studies will be needed to sustain the hypothesis that this approach could also better preserve texture information. However, as discussed by Galbusera et al., models used for the generation of high-resolution images may add details based on information found in similar patients, but not directly visible in the original low-resolution images. This can lead to the introduction of a specific shape, grey level or texture, that may have a strong impact, either positive or negative, on the future clinical applications, and further studies are needed for this purpose.

Considering the translation between sequences types, some papers have been focused on the conversion from T1-w to T2-w using both GAN and U-net architectures (Galbusera *et al* 2018, Chartsias *et al* 2018, Dar *et al* 2019), whereas other papers have explored the possibility to synthesize diffusion *b0* maps from T1-w images (Schilling *et al* 2019) or FLAIR from mpMRI (Wei *et al* 2019). Dar et al. have proposed a new conditional GAN model for multi-contrast MRI, able to preserve intermediate-to-high frequency details via an adversarial loss. They introduced two different loss functions for registered and unregistered multi-contrast images: pixel-wise and perceptual losses for the firsts and a cycle-consistency loss for the seconds (Dar *et al* 2019). The incorporation of this adversarial loss as opposed to typical squared or absolute error loss leads to enhanced capture of detailed texture information about the target contrast, thereby enabling higher synthesis quality (Shan *et al* 2018).

Although DL approaches were not specifically employed to improve radiomics features stability, they could provide a valuable contribution towards the homogenization and standardization of MR images acquired at different sites, with different scanners and protocols, as long as texture details are preserved in the image synthesis process. For this reason, ad-hoc studies will be required to assess how these novel DNN models, specifically designed to improve the visual perception and contrasts of generated images, can affect the real texture of the tissues.

b. DNN for image homogenization

As discussed in a previous dedicated section, image homogenization is a crucial step to increase radiomic reproducibility across different scanners. DNN approaches have been proposed as a promising alternative to the traditional harmonization methods, such as histogram matching and intensity standardization, or in the features domain, such as the ComBat algorithm (Mali *et al* 2021). Examples of MRI image normalization using GAN models can be found in (Modanwal *et al* 2021, 2020, Hognon *et al* 2019, Tixier *et al* 2021, Zhong *et al* 2020). Hognon et al. discussed in their work that the proposed cycleGAN architecture used to homogenized multi-center brain MRI images was able to preserve the within-center radiomic variability, whilst reduce cross-center variability (Hognon *et al* 2019). Dual GANs were proposed to harmonize DTI-MRI of neonatal brain images and they were compared with standard intensity normalization techniques and the ComBat method (Zhong *et al* 2020). The authors reported that the proposed dual GANs map the complex non-linear relationship between different sites better than conventional methods like the scaling and the ComBat.

In a recent work, the histogram matching technique and a cycleGAN model are used to obtain normalized brain MRI images acquired from a multi-center cohort (Tixier *et al* 2021). The authors verified the impact of the two normalization methods on radiomic features and they also evaluated how these methods affected the performance of the radiomic model built for survival prediction. It was shown that normalization using deep learning techniques seemed to outperform the histogram matching technique, as the GAN-based method had the ability to smooth the variability between images and recover the predictive power of some radiomic features, more than the traditional algorithms.

These preliminary results suggest that generative DNNs could be a valid alternative to the feature harmonization when multi-centric datasets are analyzed to infer radiomic models.

c. DNN for image quality improvement

Deep learning approaches have been proposed for the improvement of image quality or the correction of MR artifacts. An application that is currently reaching high interest is the intensity correction for bias field inhomogeneities (Dai *et al* 2020, Venkatesh *et al* 2020). A cycle-GAN architecture was proposed by Dai et al., obtaining positive results, in line with the traditional methods, with the advantage of an easier hyper-parameters tuning and very fast computational times (Dai *et al* 2020). In a different work, the generative DNN was associated with a novel loss function that combined a histogram correlation loss, which tries to bring about pixel consistency in different regions, and a 3D pixel loss, ensuring an accurate spatial localization of the pixels in the generated image. The proposed approach obtained a Structural Similarity Index greater than the other traditional methods and made the proposed approach efficient and accurate for intensity inhomogeneity correction (Venkatesh *et al* 2020). The impact of these methods on radiomics computation was not tested; however, the use of these novel loss functions seems to guarantee a better preservation of texture details.

Quantification of diffusion parameters from DW-MRI is used for the direct estimation of diffusion properties of tumor, such as ADC, or true and pseudo-diffusion and the perfusion fraction from IntraVoxel Incoherent Motion (IVIM) model. If the quantification is performed voxel-by-voxel, diffusion and perfusion maps can be derived and these can be used to compute radiomic features to evaluate the functional heterogeneity of the tissue. In fact, several works have performed radiomic analyses on diffusion maps and features stability has also been studied, as seen in the previous sections (Chen *et al* 2020). Recently, DNN approaches have been proposed to estimate the IVIM coefficients, showing high accuracy, similar or even higher than traditional Bayesian methods, and fast computational times (Kaandorp *et al* 2020, Barbieri *et al* 2020, Vasylechko *et al* 2022). Although the variability of IVIM coefficients estimation using different model fitting methods is widely acknowledged (While 2017, Lanzarone *et al* 2020), their impact on radiomics reproducibility was not studied, but it is likely that they can highly affect the computation of texture features, especially for the least accurate maps (i.e., for specific condition of diffusion and perfusion properties (Scalco *et al* 2021)). DNN, being able to provide robust and accurate estimation of these coefficients, may be considered as a potential strategy to increase radiomics stability, but further investigations are needed.

d. DNN for ROI segmentation

It is generally reported that semi-automatic segmentation approaches lead to higher stability than manual delineation. DNNs offer the possibility to have a completely automatic method able to provide in very short time accurate ROI segmentations learned from data that account for expert knowledge. DNNs can automatically delineate tumor volumes more accurately than other traditional segmentation methods (Havaei *et al* 2017). The effect of DNN MRI segmentation on radiomics stability in oncology was studied only in a few recent works, reporting encouraging results. Different types of U-Net architectures were proposed for endometrial (Kurata *et al* 2021), breast (Dutta *et al* 2021), cervical tumor (Lin *et al* 2020) and glioblastoma (Park *et al* 2020) segmentation, finding comparable or even better reproducibility with respect to manual contours. In particular, Dutta et al. compared the reproducibility of radiomic features extracted from DNN segmentation with the reproducibility obtained considering a consensus map from different manual segmentations (using STAPLE algorithm), in preclinical T1-w and T2-w MRI images of breast cancer. They found that, in test-retest analysis, features computed from DNN segmentation were most reproducible than those extracted from STAPLE delineation for every radiomics class and for both T1-w MRI and T2-w MRI (Dutta *et al* 2021). In another study focused on glioblastoma, the authors evaluated the reproducibility between manual and DNN-based segmentation for radiomic and diffusion/perfusion features in pre- and post-treatment acquisitions. DNN seems to increase the reproducibility in post-treatment glioblastoma when human observers tend to show high variability in segmentation (Park *et al* 2020). In the other two works, radiomic features from DNN segmentation were found to be consistent with the manual delineation, with histogram features that were confirmed as the most reproducible (Kurata *et al* 2021, Lin *et al* 2020).

Considering these preliminary results, the fully automatic ROI segmentation provided by DNNs seems an attractive option in clinics, able to drastically reduce the time required for contouring, by keeping at the same time high segmentation accuracy, high reproducibility and by warrantying also radiomics stability. It is mandatory, thus, to train and test these DNNs on large dataset, using multi-centric and multi-observer delineations, to have high heterogeneity useful to improve the generalizability of the predicted segmentations. Furthermore, U-Net-based segmentations may be even improved by considering novel approaches in deep learning, that can help in increasing reliability and interpretation. One example can be found in Bayesian networks, which allow the quantification of uncertainties related to model errors and inherent randomness of observations (Kwon *et al* 2020). The impact of this type of network on the reproducibility of radiomic features may deserve further investigations.

e. DNN for image registration

Another field of application of deep learning approaches in medical image processing is image registration. In the last years it has been rapidly evolving and it achieved the state-of-the-art performances in many applications (Fu *et al* 2020). In the radiomic workflow, image registration can be used to realign images or to directly deformed contours by applying the estimated transforms. This last case is the most selected choice when non-rigid deformations have to be considered (Yip *et al* 2016). Obviously, a contour propagation method that is fast and requires minimal manual adjustments is desirable and DNNs can offer a valid alternative to traditional methods, such as Free-Form Deformation (FFD) with B-Splines (Rueckert *et al* 1999) or Optical Flow algorithms (Thirion 1998). Eppenhof et al. proposed the use of CNNs to automatically propagate prostate contours at each treatment fraction during MR-guided radiotherapy (Eppenhof *et al* 2020). The 3D-CNN consisted in two parts: in the first one, the reference and target images were given as input and the deformation vector field was estimated as output. In the second part, the estimated deformation field was applied to the prostate segmentation manually delineated on the reference image to obtain the contour on the target image. The accuracy obtained by the CNN was higher than that obtained using the traditional FFD approach.

This kind of DNN approach has not been tested yet in the radiomic context to increase features stability. However, since image registration for contour propagation can be included in the field of automatic segmentation, we can expect that more accurate contour delineation in longitudinal acquisitions may be coupled with more stable radiomic features than traditional image registration methods.

Table 4. DNN models applied on MRI that have effect on image texture

Reference	DNN task	DNN model	Radiomic evaluation	Potential effect on texture
(Dar *et al* 2019)	Image synthesis (multi-contrast MRI generation)	Conditional GAN	No	Introduction of adversarial losses to capture detailed texture information
(Dutta *et al* 2021)	ROI segmentation	UNet	Yes	Features computed from UNet most reproducible than those extracted from STAPLE.
(Galbusera *et al* 2018)	Image synthesis (increase resolution)	GAN	No	Introduction of details not visible in low-resolution images that may alter the real texture.
(Hognon *et al* 2019)	Image homogenization	cycleGAN	Yes	Preservation of the within-center radiomic variability, and reduction of cross-center variability.
(Kim *et al* 2018)	Image synthesis (increase resolution)	GAN	No	Inclusion of high resolution image in another contrast to recover spatial and contrast details
(Kurata *et al* 2021)	ROI segmentation	UNet	Yes	High radiomic stability between automatic and manual segmentations
(Lin *et al* 2020)	ROI segmentation	UNet	Yes	High radiomic stability between automatic and manual segmentations
(Park *et al* 2020)	ROI segmentation	UNet	Yes	Similar or higher radiomic stability for features computed from UNet compared to manual segmentation.
(Qu *et al* 2020)	Image synthesis (from 3T to 7T MRI)	UNet	No	Integration of wavelet domain in DNN to obtain better tissue contrast and greater details
(Tixier *et al* 2021)	Image homogenization	cycleGAN	Yes	GAN-based method had the ability to smooth the variability between images and recover the predictive power of some radiomic features.

Open issues for the integration of DNN in radiomics

In the previous sections, the integration of DNNs in the MRI processing for the extraction of radiomic features was described, by discussing several types of applications. Even if DNNs seem to obtain better results compared to traditional methods in terms of segmentation accuracy, image quality or image homogenization, their effects on radiomic stability are not completely clear, and only some preliminary evidences are present in literature. Image segmentation and image harmonization are the only applications for which ad-hoc studies are available; in these cases, radiomic stability seems to benefit from the introduction of DNN models. More efforts should be performed for the evaluation of the impact of these approaches on radiomics regarding image synthesis and image quality improvement. Furthermore, it should be considered that deep learning algorithms present some open challenges that may hamper their wide adoption, especially within the radiomic workflow.

One of the main limitations in DNNs training is the lack of a true gold standard. For example, networks for image segmentation are trained using labels manually contoured by experts, that are obviously prone to errors: thus, the final estimations are dependent on the operators' variability and expertise. The same problem arose with the generation of new images, e.g., images corrected for bias field inhomogeneities using cycle-GANs (Dai *et al* 2020), where the ground truth was the correction made by the traditional N4ITK method (Tustison *et al* 2010), which requires a specific parameter tuning that necessarily affects the results. Thus, N4ITK corrected images contained unavoidable errors compared to ideal images free from field inhomogeneities, that are, however, difficult to be obtained in practice (Dai *et al* 2020). Also, for the adoption of DNNs in image registration, the lack of known transformation in the training process is a main limitation. In this case, possible solutions can consist in data augmentation, which however could introduce errors and bias, or in transfer learning techniques (Fu *et al* 2020).

The other main limitation of DNNs is the reduced generalization capability. Large, curated and heterogeneous dataset are needed to obtain generalizable models; however, in several cases DNNs were trained and validated on images acquired in the same institute. In addition, it can be also taken into account that synthetic images generated by DNNs can be unpredictable when input images in the inference phase differ significantly from training images. Currently, it is a standard practice the exclusion of unusual cases in the training/validation process; however, unusual cases may be present in the clinical practice, especially when tumor progression/response to treatment is studied. In these cases, the application of DNN-based methods should be approached with caution (Wang *et al* 2021). Further studies are thus needed to obtain robust, reliable and generalizable DNN performances in multi-centric settings.

Another point that should be carefully considered is that even if radiomics stability would be improved by advanced image processing techniques, this could not guarantee that radiomic models would obtain higher predictive performance. In fact, reproducible features are not necessarily clinically informative. Nevertheless, successful radiomics models must be built upon reproducible and robust features (Zhao 2021). In this sense, the potential benefit of integrating DNN models in the MRI image processing before radiomic features extraction should be also evaluated and compared to the end-to-end learning strategy, where DNN would completely replace the radiomic workflow. In this way, it may be possible to understand if increased radiomic stability, as provided by deep learning integration, would also lead to increased radiomic model performance and if this strategy would be preferable to the end-to-end learning. Future studies should be addressed on this specific topic, to sustain the choice of one approach rather than the other.

Conclusion

In this review, we have discussed the main factors that affect radiomic features stability, when computed from MR images, along with the possible contribution of DNN as a potential solution.

The huge variability in the MRI workflow, from image acquisition to feature computation, significantly reduces the reliability of MRI radiomics. By reviewing the recent literature, we recommend an accurate standardization of each step of this workflow, especially for new prospective studies. For retrospective studies, where image acquisition and processing standardization cannot be achieved, possible solutions can be found in the harmonization in the image domain or in the feature domain (or both). Deep learning approaches are an interesting novel strategy that provides high-quality and homogeneous images, potentially able to better preserve textural information, and that can be used as an alternative to or in combination with traditional harmonization approaches. However, we have seen in the previous sections that the most part of these methods has not been evaluated yet in terms of radiomic feature stability. On the contrary, DNNs seem an attractive and ready-to-use solution to increase the reproducibility of ROI segmentation, with some preliminary evaluations on radiomic stability already published.

Future works could deepen this issue, by testing the actual impact of DNNs in the MRI workflow for the computation of more reliable radiomic features, making sure of not altering the original texture of the tissue.

References

Afshar P, Mohammadi A, Plataniotis K N, Oikonomou A and Benali H 2019 From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities *IEEE Signal Processing Magazine* **36** 132–60

Ammari S, Pitre-Champagnat S, Dercle L, Chouzenoux E, Moalla S, Reuze S, Talbot H, Mokoyoko T, Hadchiti J, Diffetocq S, Volk A, Haik M el, Lakiss S, Balleyguier C, Lassau N and Bidault F 2020 Influence of Magnetic Field Strength on Magnetic Resonance Imaging Radiomics Features in Brain Imaging, an In Vitro and In Vivo Study *Frontiers in Oncology* **10** Online: /pmc/articles/PMC7855708/

Baeßler B, Weiss K and Pinto Dos Santos D 2018 Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Investigative radiology* Online: http://www.ncbi.nlm.nih.gov/pubmed/30433891

Barbieri S, Gurney-Champion O J, Klaassen R and Thoeny H C 2020 Deep learning how to fit an intravoxel incoherent motion model to diffusion-weighted MRI *Magnetic Resonance in Medicine* **83** 312–21 Online: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.27910

Bartlett J W and Frost C 2008 Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables *Ultrasound in Obstetrics and Gynecology* **31** 466–75 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/uog.5256

Bernatz S, Zhdanovich Y, Ackermann J, Koch I, Wild P J, dos Santos D P, Vogl T J, Kaltenbach B and Rosbach N 2021 Impact of rescanning and repositioning on radiomic features employing a multi-object phantom in magnetic resonance imaging *Scientific Reports* **11** 1–13 Online: https://www.nature.com/articles/s41598-021-93756-x

Bianchini L, Santinha J, Loução N, Figueiredo M, Botta F, Origgi D, Cremonesi M, Cassano E, Papanikolaou N and Lascialfari A 2021 A multicenter study on radiomic features from T2-weighted images of a customized MR pelvic phantom setting the basis for robust radiomic models in clinics *Magnetic Resonance in Medicine* **85** 1713–26 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/mrm.28521

Bologna M, Corino V D A, Montin E, Messina A, Calareso G, Greco F G, Sdao S and Mainardi L T 2018 Assessment of Stability and Discrimination Capacity of Radiomic Features on Apparent Diffusion Coefficient Images *Journal of Digital Imaging* **31** 879–94

Carré A, Klausner G, Edjlali M, Lerousseau M, Briend-Diop J, Sun R, Ammari S, Reuzé S, Alvarez Andres E, Estienne T, Niyoteka S, Battistella E, Vakalopoulou M, Dhermain F, Paragios N, Deutsch E, Oppenheim C, Pallud J and Robert C 2020 Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics *Scientific Reports* **10** 1–15 Online: https://www.nature.com/articles/s41598-020-69298-z

Chartsias A, Joyce T, Giuffrida M V and Tsaftaris S A 2018 Multimodal MR Synthesis via Modality-Invariant Latent Representation *IEEE Transactions on Medical Imaging* **37** 803–14

Chen H, He Y, Zhao C, Zheng L, Pan N, Qiu J, Zhang Z, Niu X and Yuan Z 2020 Reproducibility of radiomics features derived from intravoxel incoherent motion diffusion-weighted MRI of cervical cancer:

Acta Radiologica **62** 679–86 Online: https://journals.sagepub.com/doi/abs/10.1177/0284185120934471

Cheng S, Shi H, Lu M, Wang C, Duan S, Xu Q and Shi H 2021 Radiomics Analysis for Predicting Malignant Potential of Intraductal Papillary Mucinous Neoplasms of the Pancreas: Comparison of CT and MRI *Academic Radiology* **In Press**

Chirra P, Leo P, Yim M, Bloch B N, Rastinehad A R, Purysko A, Rosen M, Madabhushi A and Viswanath S E 2019 Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI *Journal of Medical Imaging* **6** 024502 Online: https://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-6/issue-02/024502/Multisite-evaluation-of-radiomic-feature-reproducibility-and-discriminability-for-identifying/10.1117/1.JMI.6.2.024502.full

Crombé A, Buy X, Han F, Toupin S and Kind M 2021 Assessment of Repeatability, Reproducibility, and Performances of T2 Mapping-Based Radiomics Features: A Comparative Study *Journal of Magnetic Resonance Imaging* **54** 537–48 Online: www.lifexsoft.org

Crombé A, Saut O, Guigui J, Italiano A, Buy X and Kind M 2019 Influence of temporal parameters of DCE-MRI on the quantification of heterogeneity in tumor vascularization *Journal of Magnetic Resonance Imaging* **50** 1773–88 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/jmri.26753

Dai X, Lei Y, Liu Y, Wang T, Ren L, Curran W J, Patel P, Liu T and Yang X 2020 Intensity non-uniformity correction in MR imaging using residual cycle generative adversarial network *Physics in Medicine & Biology* **65** 215025 Online: https://iopscience.iop.org/article/10.1088/1361-6560/abb31f

Damascelli A, Gallivanone F, Cristel G, Cava C, Interlenghi M, Esposito A, Brembilla G, Briganti A, Montorsi F, Castiglioni I and de Cobelli F 2021 Advanced imaging analysis in prostate mri: Building a radiomic signature to predict tumor aggressiveness *Diagnostics* **11** 594 Online: https://www.mdpi.com/2075-4418/11/4/594/htm

Dar S U H, Yurt M, Karacan L, Erdem A, Erdem E and Cukur T 2019 Image Synthesis in Multi-Contrast MRI with Conditional Generative Adversarial Networks *IEEE Transactions on Medical Imaging* **38** 2375–88

Davnall F, Yip C S P, Ljungqvist G, Selmi M, Ng F, Sanghera B, Ganeshan B, Miles K A, Cook G J and Goh V 2012 Assessment of tumor heterogeneity: An emerging imaging tool for clinical practice? *Insights into Imaging* **3** 573–89

Dreher C, Kuder T A, König F, Mlynarska-Bujny A, Tenconi C, Paech D, Schlemmer H P, Ladd M E and Bickelhaupt S 2020 Radiomics in diffusion data: a test–retest, inter- and intra-reader DWI phantom study *Clinical Radiology* **75** 798.e13-798.e22

Duron L, Balvay D, vande Perre S, Bouchouicha A, Savatovsky J, Sadik J-C, Thomassin-Naggara I, Fournier L and Lecler A 2019 Gray-level discretization impacts reproducible MRI radiomics texture features ed Y Fan *PLOS ONE* **14** e0213459 Online: http://dx.plos.org/10.1371/journal.pone.0213459

Dutta K, Roy S, Whitehead T D, Luo J, Jha A K, Li S, Quirk J D and Shoghi K I 2021 Deep Learning Segmentation of Triple-Negative Breast Cancer (TNBC) Patient Derived Tumor Xenograft (PDX) and

Sensitivity of Radiomic Pipeline to Tumor Probability Boundary *Cancers* **13** 3795 Online: https://www.mdpi.com/2072-6694/13/15/3795/htm

Eck B, Chirra P v., Muchhala A, Hall S, Bera K, Tiwari P, Madabhushi A, Seiberlich N and Viswanath S E 2021 Prospective Evaluation of Repeatability and Robustness of Radiomic Descriptors in Healthy Brain Tissue Regions In Vivo Across Systematic Variations in T2-Weighted Magnetic Resonance Imaging Acquisition Parameters *Journal of Magnetic Resonance Imaging* **54** 1009–21 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/jmri.27635

Eppenhof K A J, Maspero M, Savenije M H F, de Boer J C J, van der Voort van Zyp J R N, Raaymakers B W, Raaijmakers A J E, Veta M, van den Berg C A T and Pluim J P W 2020 Fast contour propagation for MR-guided prostate radiotherapy using convolutional neural networks *Medical Physics* **47** 1238–48 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/mp.13994

Fiset S, Welch M L, Weiss J, Pintilie M, Conway J L, Milosevic M, Fyles A, Traverso A, Jaffray D, Metser U, Xie J and Han K 2019 Repeatability and reproducibility of MRI-based radiomic features in cervical cancer *Radiotherapy and Oncology* **135** 107–14 Online: https://www.sciencedirect.com/science/article/pii/S0167814019301057

Fu Y, Lei Y, Wang T, Curran W J, Liu T and Yang X 2020 Deep learning in medical image registration: a review *Physics in Medicine & Biology* **65** 20TR01 Online: https://iopscience.iop.org/article/10.1088/1361-6560/ab843e

Galbusera F, Bassani T, Casaroli G, Gitto S, Zanchetta E, Costa F and Sconfienza L M 2018 Generative models: an upcoming innovation in musculoskeletal radiology? A preliminary test in spine imaging *European Radiology Experimental* **2** 1–13 Online: https://eurradiolexp.springeropen.com/articles/10.1186/s41747-018-0060-7

Gitto S, Cuocolo R, Emili I, Tofanelli L, Chianca V, Albano D, Messina C, Imbriaco M and Sconfienza L M 2021 Effects of Interobserver Variability on 2D and 3D CT- and MRI-Based Texture Feature Reproducibility of Cartilaginous Bone Tumors *Journal of Digital Imaging* 1–13 Online: https://link.springer.com/article/10.1007/s10278-021-00498-3

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative Adversarial Nets *Advances in neural information processing systems* **27**

Granzier R W Y, Verbakel N M H, Ibrahim A, van Timmeren J E, van Nijnatten T J A, Leijenaar R T H, Lobbes M B I, Smidt M L and Woodruff H C 2020 MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability *Scientific Reports* **10** 1–11 Online: https://www.nature.com/articles/s41598-020-70940-z

Hatt M, Parmar C, Qi J and el Naqa I 2019 Machine (Deep) Learning Methods for Image Processing and Radiomics *IEEE Transactions on Radiation and Plasma Medical Sciences* **3** 104–8

Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, Pal C, Jodoin P M and Larochelle H 2017 Brain tumor segmentation with Deep Neural Networks *Medical Image Analysis* **35** 18–31

Hoebel K v., Patel J B, Beers A L, Chang K, Singh P, Brown J M, Pinho M C, Batchelor T T, Gerstner E R, Rosen B R and Kalpathy-Cramer J 2021 Radiomics Repeatability Pitfalls in a Scan-Rescan MRI Study of Glioblastoma *Radiology: Artificial Intelligence* **3** e190199

Hognon C, Tixier F, Gallinato O, Colin T, Visvikis D, Jaouen V and Stan A 2019 Standardization of Multicentric Image Datasets with Generative Adversarial Networks *IEEE MIC2019* 9–10 Online: https://hal.archives-ouvertes.fr/hal-02447807

Ibrahim A, Primakov S, Beuque M, Woodru H C, Halilaj I, Wu G, Refaee T, Granzier R, Widaatalla Y, Hustinx R, Mottaghy F M and Lambin P 2021 Radiomics for precision medicine : Current challenges , future prospects , and the proposal of a new framework *Methods* **188** 20–9

Isaksson L J, Raimondi S, Botta F, Pepa M, Gugliandolo S G, de Angelis S P, Marvaso G, Petralia G, de Cobelli O, Gandini S, Cremonesi M, Cattani F, Summers P and Jereczek-Fossa B A 2020 Effects of MRI image normalization techniques in prostate cancer radiomics *Physica Medica* **71** 7–13 Online: https://doi.org/10.1016/j.ejmp.2020.02.007

Johnson W E, Li C and Rabinovic A 2007 Adjusting batch effects in microarray expression data using empirical Bayes methods *Biostatistics* **8** 118–27 Online: https://academic.oup.com/biostatistics/article/8/1/118/252073

Kaandorp M P T, Barbieri S, Klaassen R, van Laarhoven H W M, Crezee H, While P T, Nederveen A J and Gurney-Champion O J 2020 Improved unsupervised physics-informed deep learning for intravoxel-incoherent motion modeling and evaluation in pancreatic cancer patients *Magnetic Resonance Imaging* **86** 2250–65 Online: http://arxiv.org/abs/2011.01689

Kim K H, Do W J and Park S H 2018 Improving resolution of MR images with an adversarial network incorporating images with different contrast *Medical Physics* **45** 3120–31 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/mp.12945

Kim M, Jung S C, Park J E, Park S Y, Lee H and Choi K M 2021 Reproducibility of radiomic features in SENSE and compressed SENSE: impact of acceleration factors *European radiology* **31** 6457–70 Online: https://pubmed.ncbi.nlm.nih.gov/33733690/

Koo T K and Li M Y 2016 A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research *Journal of Chiropractic Medicine* **15** 155–63 Online: http://dx.doi.org/10.1016/j.jcm.2016.02.012

Kurata Y, Nishio M, Moribata Y, Kido A, Himoto Y, Otani S, Fujimoto K, Yakami M, Minamiguchi S, Mandai M and Nakamoto Y 2021 Automatic segmentation of uterine endometrial cancer on multi-sequence MRI using a convolutional neural network *Scientific Reports* **11** 1–10 Online: https://www.nature.com/articles/s41598-021-93792-7

Kwon Y, Won J H, Kim B J and Paik M C 2020 Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation *Computational Statistics and Data Analysis* **142** 106816

Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout R G P M, Granton P, Zegers C M L, Gillies R, Boellard R, Dekker A and Aerts H J W L 2012 Radiomics: Extracting more information from medical images using advanced feature analysis *European Journal of Cancer* **48** 441–6

Lanzarone E, Rizzo G, Mastropietro A, Scalco E and Vidiri A 2020 A novel bayesian approach with conditional autoregressive specification for intravoxel incoherent motion diffusion-weighted MRI *NMR in Biomedicine* **33** e4201

Lecler A, Duron L, Balvay D, Savatovsky J, Bergès O, Zmuda M, Farah E, Galatoire O, Bouchouicha A and Fournier L S 2019 Combining Multiple Magnetic Resonance Imaging Sequences Provides Independent Reproducible Radiomics Features *Scientific Reports* **9** 2068 Online: http://www.nature.com/articles/s41598-018-37984-8

Lee J, Steinmann A, Ding Y, Lee H, Owens C, Wang J, Yang J, Followill D, Ger R, MacKin D and Court L E 2021 Radiomics feature robustness as measured using an MRI phantom *Scientific Reports* **11** 1–14 Online: https://www.nature.com/articles/s41598-021-83593-3

Lee M, Woo B, Kuo M D, Jamshidi N and Kim J H 2017 Quality of Radiomic Features in Glioblastoma Multiforme: Impact of Semi-Automated Tumor Segmentation Software *Korean Journal of Radiology* **18** 498–509 Online: https://synapse.koreamed.org/articles/1027345

Li Y, Ammari S, Balleyguier C, Lassau N and Chouzenoux E 2021 Impact of Preprocessing and Harmonization Methods on the Removal of Scanner Effects in Brain MRI Radiomic Features *Cancers* **13** 3000 Online: https://www.mdpi.com/2072-6694/13/12/3000/htm

Lin Y C, Lin C H, Lu H Y, Chiang H J, Wang H K, Huang Y T, Ng S H, Hong J H, Yen T C, Lai C H and Lin G 2020 Deep learning for fully automated tumor segmentation and extraction of magnetic resonance radiomics features in cervical cancer *European Radiology* **30** 1297–305

Litjens G, Kooi T, Bejnordi B E, Setio A A A, Ciompi F, Ghafoorian M, van der Laak J A W M, van Ginneken B and Sánchez C I 2017 A survey on deep learning in medical image analysis *Medical Image Analysis* **42** 60–88

Liu Z, Wang S, Dong D, Wei J, Fang C, Zhou X, Sun K, Li L, Li B, Wang M and Tian J 2019 The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges *Theranostics* **9** 1303 Online: /pmc/articles/PMC6401507/

Lohmann P, Bousabarah K, Hoevels M and Treuer H 2020 Radiomics in radiation oncology—basics, methods, and limitations *Strahlentherapie und Onkologie 2020 196:10* **196** 848–55 Online: https://link.springer.com/article/10.1007/s00066-020-01663-3

Lu H, Parra N A, Qi J, Gage K, Li Q, Fan S, Feuerlein S, Pow-Sang J, Gillies R, Choi J W and Balagurunathan Y 2020 Repeatability of Quantitative Imaging Features in Prostate Magnetic Resonance Imaging *Frontiers in Oncology* **0** 551

Lu H, Yuan Y, Zhou Z, Ma X, Shen F, Xia Y and Lu J 2021 Assessment of MRI-Based Radiomics in Preoperative T Staging of Rectal Cancer: Comparison between Minimum and Maximum Delineation Methods *BioMed Research International* **2021**

Mahon R N, Hugo G D and Weiss E 2019 Repeatability of texture features derived from magnetic resonance and computed tomography imaging and use in predictive models for non-small cell lung cancer outcome *Physics in Medicine and Biology* **64**

Mali S A, Ibrahim A, Woodruff H C, Andrearczyk V, Müller H, Primakov S, Salahuddin Z, Chatterjee A and Lambin P 2021 Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods *Journal of Personalized Medicine* **11** 842 Online: https://www.mdpi.com/2075-4426/11/9/842/htm

Mayerhoefer M E, Szomolanyi P, Jirak D, Materka A and Trattnig S 2009 Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: An application-oriented study *Medical Physics* **36** 1236 Online: http://scitation.aip.org/content/aapm/journal/medphys/36/4/10.1118/1.3081408

McHugh D J, Porta N, Little R A, Cheung S, Watson Y, Parker G J M, Jayson G C and O'Connor J P B 2021 Image Contrast, Image Pre-Processing, and T1 Mapping Affect MRI Radiomic Feature Repeatability in Patients with Colorectal Cancer Liver Metastases *Cancers 2021, Vol. 13, Page 240* **13** 240 Online: https://www.mdpi.com/2072-6694/13/2/240/htm

Merisaari H, Taimen P, Shiradkar R, Ettala O, Pesola M, Saunavaara J, Boström P J, Madabhushi A, Aronen H J and Jambor I 2020 Repeatability of radiomics and machine learning for DWI: Short-term repeatability study of 112 patients with prostate cancer *Magnetic Resonance in Medicine* **83** 2293–309 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/mrm.28058

Mi H, Yuan M, Suo S, Cheng J, Li S, Duan S and Lu Q 2020 Impact of different scanners and acquisition parameters on robustness of MR radiomics features based on women's cervix *Scientific Reports* **10** 1–12 Online: https://www.nature.com/articles/s41598-020-76989-0

Modanwal G, Vellal A, Buda M and Mazurowski M A 2020 MRI image harmonization using cycle-consistent generative adversarial network *SPIE Medical Imaging* 36

Modanwal G, Vellal A and Mazurowski M A 2021 Normalization of breast MRIs using cycle-consistent generative adversarial networks *Computer Methods and Programs in Biomedicine* **208** 106225

Molina D, Pérez-Beteta J, Martínez-González A, Martino J, Velasquez C, Arana E and Pérez-García V M 2017 Lack of robustness of textural measures obtained from 3D brain tumor MRIs impose a need for standardization *PLOS ONE* **12** e0178843 Online: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0178843

Moradmand H, Aghamiri S M R and Ghaderi R 2020 Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma *Journal of Applied Clinical Medical Physics* **21** 179–90 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/acm2.12795

Nie D, Trullo R, Lian J, Wang L, Petitjean C, Ruan S, Wang Q and Shen D 2018 Medical Image Synthesis with Deep Convolutional Adversarial Networks *IEEE Transactions on Biomedical Engineering* **65** 2720–30

Nyul L G, Udupa J K and Zhang X 2000 New variants of a method of MRI scale standardization *IEEE Transactions on Medical Imaging* **19** 143–50

Pandey U, Saini J, Kumar M, Gupta R and Ingalhalikar M 2021 Normative Baseline for Radiomics in Brain MRI: Evaluating the Robustness, Regional Variations, and Reproducibility on FLAIR Images *Journal of Magnetic Resonance Imaging* **53** 394–407 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/jmri.27349

Papadimitroulas P, Brocki L, Christopher Chung N, Marchadour W, Vermet F, Gaubert L, Eleftheriadis V, Plachouris D, Visvikis D, Kagadis G C and Hatt M 2021 Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization *Physica Medica* **83** 108–21

Parekh V S and Jacobs M A 2019 Deep learning and radiomics in precision medicine *Expert Review of Precision Medicine and Drug Development* **4** 59–72 Online: https://www.tandfonline.com/doi/abs/10.1080/23808993.2019.1585805

Park J E, Ham S, Kim H S, Park S Y, Yun J, Lee H, Choi S H and Kim N 2020 Diffusion and perfusion MRI radiomics obtained from deep learning segmentation provides reproducible and comparable diagnostic model to human in post-treatment glioblastoma *European Radiology* **31** 3127–37 Online: https://link.springer.com/article/10.1007/s00330-020-07414-3

Park S-H, Lim H, Bae B K, Hahm M H, Chong G O, Jeong S Y and Kim J-C 2021 Robustness of magnetic resonance radiomic features to pixel size resampling and interpolation in patients with cervical cancer *Cancer Imaging* **21** 1–11 Online: https://cancerimagingjournal.biomedcentral.com/articles/10.1186/s40644-021-00388-5

Pati S, Verma R, Akbari H, Bilello M, Hill V B, Sako C, Correa R, Beig N, Venet L, Thakur S, Serai P, Ha S M, Blake G D, Shinohara R T, Tiwari P and Bakas S 2020 Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the Ivy Glioblastoma Atlas Project (Ivy GAP) dataset *Medical Physics* **47** 6039–52 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/mp.14556

Peerlings J, Woodruff H C, Winfield J M, Ibrahim A, van Beers B E, Heerschap A, Jackson A, Wildberger J E, Mottaghy F M, DeSouza N M and Lambin P 2019 Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial *Scientific Reports 2019 9:1* **9** 1–10 Online: https://www.nature.com/articles/s41598-019-41344-5

Qu L, Zhang Y, Wang S, Yap P T and Shen D 2020 Synthesized 7T MRI from 3T MRI via deep learning in spatial and wavelet domains *Medical Image Analysis* **62** 101663

Rai R, Holloway L C, Brink C, Field M, Christiansen R L, Sun Y, Barton M B and Liney G P 2020 Multicenter evaluation of MRI-based radiomic features: A phantom study *Medical Physics* **47** 3054–63 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/mp.14173

Reyes M, Meier R, Pereira S, Silva C A, Dahlweid F-M, Tengg-Kobligk H von, Summers R M and Wiest R 2020 On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities *Radiology: Artificial Intelligence* **2** e190043 Online: https://pubs.rsna.org/doi/abs/10.1148/ryai.2020190043

Ronneberger O, Fischer P and Brox T 2015 U-net: Convolutional networks for biomedical image segmentation *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol 9351 (Springer Verlag) pp 234–41

Roy S, Whitehead T D, Quirk J D, Salter A, Ademuyiwa F O, Li S, An H and Shoghi K I 2020 Optimal co-clinical radiomics: Sensitivity of radiomic features to tumour volume, image noise and resolution in co-clinical T1-weighted and T2-weighted magnetic resonance imaging *EBioMedicine* **59** 102963

Rueckert D, Sonoda L I, Hayes C, Hill D L, Leach M O and Hawkes D J 1999 Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging* **18** 712–21 Online: http://www.ncbi.nlm.nih.gov/pubmed/10534053

Saha A, Harowicz M R and Mazurowski M A 2018 Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors *Medical Physics* **45** 3076–85 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/mp.12925

Saha A, Yu X, Sahoo D and Mazurowski M A 2017 Effects of MRI scanner parameters on breast cancer radiomics *Expert Systems with Applications* **87** 384–91 Online: https://www.sciencedirect.com/science/article/pii/S0957417417304463

Scalco E, Belfatto A, Mastropietro A, Rancati T, Avuzzi B, Messina A, Valdagni R and Rizzo G 2020 T2w-MRI signal normalization affects radiomics features reproducibility *Medical Physics* **47** 1680–91

Scalco E, Mastropietro A, Bodini A, Marzi S and Rizzo G 2021 A Multi-Variate framework to assess reliability and discrimination power of Bayesian estimation of Intravoxel Incoherent Motion parameters *Physica Medica* **89** 11–9

Scalco E and Rizzo G 2017 Texture analysis of medical images for radiotherapy applications *The British journal of radiology* **90** 20160642

Schilling K G, Blaber J, Huo Y, Newton A, Hansen C, Nath V, Shafer A T, Williams O, Resnick S M, Rogers B, Anderson A W and Landman B A 2019 Synthesized b0 for diffusion distortion correction (Synb0-DisCo) *Magnetic Resonance Imaging* **64** 62–70

Schwier M, van Griethuysen J, Vangel M G, Pieper S, Peled S, Tempany C, Aerts H J W L, Kikinis R, Fennessy F M and Fedorov A 2019 Repeatability of Multiparametric Prostate MRI Radiomics Features *Scientific Reports* **9** 9441 Online: http://www.nature.com/articles/s41598-019-45766-z

Shan H, Zhang Y, Yang Q, Kruger U, Kalra M K, Sun L, Cong W and Wang G 2018 3-D Convolutional Encoder-Decoder Network for Low-Dose CT via Transfer Learning From a 2-D Trained Network *IEEE Transactions on Medical Imaging* **37** 1522–34

Shiri I, Abdollahi H, Shaysteh S and Mahdavi S R 2017 Test-Retest Reproducibility and Robustness Analysis of Recurrent Glioblastoma MRI Radiomics Texture Features *Iranian Journal of Radiology* 48035

Shiri I, Hajianfar G, Sohrabi A, Abdollahi H, Shayesteh S P, Geramifar P, Zaidi H, Oveisi M and Rahmim A 2020 Repeatability of radiomic features in magnetic resonance imaging of glioblastoma: Test–retest and image registration analyses *Medical Physics* **47** 4265–80 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/mp.14368

Shur J, Blackledge M, D'Arcy J, Collins D J, Bali M, O'Leach M and Koh D-M 2021 MRI texture feature repeatability and image acquisition factor robustness, a phantom study and in silico study *European Radiology Experimental* **5** 1–11 Online: https://eurradiolexp.springeropen.com/articles/10.1186/s41747-020-00199-6

Simpson G, Ford J C, Llorente R, Portelance L, Yang F, Mellon E A and Dogan N 2020 Impact of quantization algorithm and number of gray level intensities on variability and repeatability of low field strength magnetic resonance image-based radiomics texture features *Physica Medica* **80** 209–20

Stanzione A, Cuocolo R, Verde F, Galatola R, Romeo V, Mainenti P P, Aprea G, Guadagno E, del Basso De Caro M and Maurea S 2021 Handcrafted MRI radiomics and machine learning: Classification of indeterminate solid adrenal lesions *Magnetic Resonance Imaging* **79** 52–8

Stanzione A, Gambardella M, Cuocolo R, Ponsiglione A, Romeo V and Imbriaco M 2020 Prostate MRI radiomics: A systematic review and radiomic quality score assessment *European Journal of Radiology* **129** 109095

Suter Y, Knecht U, Alão M, Valenzuela W, Hewer E, Schucht P, Wiest R and Reyes M 2020 Radiomics for glioblastoma survival analysis in pre-operative MRI: Exploring feature robustness, class boundaries, and machine learning techniques *Cancer Imaging* **20** 1–13 Online: https://link.springer.com/articles/10.1186/s40644-020-00329-8

Thirion J-P 1998 Image matching as a diffusion process: an analogy with Maxwell's demons *Medical Image Analysis* **2** 243–60 Online: http://www.sciencedirect.com/science/article/pii/S1361841598800224

Tixier F, Jaouen V, Hognon C, Gallinato O, Colin T and Visvikis D 2021 Evaluation of conventional and deep learning based image harmonization methods in radiomics studies *Physics in Medicine and Biology* **66**

Tixier F, Um H, Young R J and Veeraraghavan H 2019 Reliability of tumor segmentation in glioblastoma: Impact on the robustness of MRI-radiomic features *Medical Physics* **46** 3582–91 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/mp.13624

Tjoa E and Guan C 2020 A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI *IEEE Transactions on Neural Networks and Learning Systems* **32** 4793–813

Traverso A, Kazmierski M, Shi Z, Kalendralis P, Welch M, Nissen H D, Jaffray D, Dekker A and Wee L 2019 Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing *Physica Medica* **61** 44–51

Traverso A, Kazmierski M, Welch M L, Weiss J, Fiset S, Foltz W D, Gladwish A, Dekker A, Jaffray D, Wee L and Han K 2020 Sensitivity of radiomic features to inter-observer variability and image pre-processing in Apparent Diffusion Coefficient (ADC) maps of cervix cancer patients *Radiotherapy and Oncology* **143** 88–94

Traverso A, Wee L, Dekker A and Gillies R 2018 Repeatability and reproducibility of radiomic features: A systematic review *International Journal of Radiation Oncology*Biology*Physics* **102** 1143–58 Online: http://linkinghub.elsevier.com/retrieve/pii/S0360301618309052

Tustison N J, Avants B B, Cook P A, Yuanjie Zheng, Egan A, Yushkevich P A and Gee J C 2010 N4ITK: Improved N3 Bias Correction *IEEE Transactions on Medical Imaging* **29** 1310–20 Online: http://ieeexplore.ieee.org/document/5445030/

Um H, Tixier F, Bermudez D, Deasy J O, Young R J and Veeraraghavan H 2019 Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets *Physics in Medicine & Biology* **64** 165011 Online: https://iopscience.iop.org/article/10.1088/1361-6560/ab2f44

Vasylechko S D, Warfield S K, Afacan O and Kurugol S 2022 Self-supervised IVIM DWI parameter estimation with a physics based forward model *Magnetic Resonance in Medicine* **87** 904–14

Venkatesh V, Sharma N and Singh M 2020 Intensity inhomogeneity correction of MRI images using InhomoNet *Computerized Medical Imaging and Graphics* **84** 101748

Wang T, Lei Y, Fu Y, Wynne J F, Curran W J, Liu T and Yang X 2021 A review on medical imaging synthesis using deep learning and its clinical applications *Journal of Applied Clinical Medical Physics* **22** 11–36 Online: https://onlinelibrary.wiley.com/doi/full/10.1002/acm2.13121

Wei W, Poirion E, Bodini B, Durrleman S, Colliot O, Stankoff B and Ayache N 2019 Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis *https://doi.org/10.1117/1.JMI.6.1.014005* **6** 014005 Online: https://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-6/issue-1/014005/Fluid-attenuated-inversion-recovery-MRI-synthesis-from-multisequence-MRI-using/10.1117/1.JMI.6.1.014005.full

While P T 2017 A comparative simulation study of bayesian fitting approaches to intravoxel incoherent motion modeling in diffusion-weighted MRI *Magnetic Resonance in Medicine* **78** 2373–87

Wu S, Jiao Y, Zhang Y, Ren X, Li P, Yu Q, Zhang Q, Wang Q and Fu S 2019 Imaging-Based Individualized Response Prediction Of Carbon Ion Radiotherapy For Prostate Cancer Patients *Cancer Management and Research* **11** 9121 Online: /pmc/articles/PMC6817347/

Xue C, Yuan J, Lo G G, Chang A T Y, Poon D M C, Wong O L, Zhou Y and Chu W C W 2021a Radiomics feature reliability assessed by intraclass correlation coefficient: a systematic review *Quantitative Imaging in Medicine and Surgery* **11** 4431 Online: /pmc/articles/PMC8408801/

Xue C, Yuan J, Poon D M C, Zhou Y, Yang B, Yu S K and Cheung Y K 2021b Reliability of MRI radiomics features in MR-guided radiotherapy for prostate cancer: Repeatability, reproducibility, and within-subject agreement *Medical physics* **48** 6976–86 Online: https://pubmed.ncbi.nlm.nih.gov/34562286/

Xue C, Zhou Y, Lo G G, Wong O L, Yu S K, Cheung K Y and Yuan J 2021c Reliability of radiomics features due to image reconstruction using a standardized T 2 -weighted pulse sequence for MR-guided radiotherapy: An anthropomorphic phantom study *Magnetic resonance in medicine* **85** 3434–46 Online: https://pubmed.ncbi.nlm.nih.gov/33404129/

Yang F, Dogan N, Stoyanova R and Ford J C 2018 Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: A simulation study utilizing ground truth *Physica Medica* **50** 26–36 Online: https://www.sciencedirect.com/science/article/pii/S1120179718304721?via%3Dihub

Yang P, Xu L, Cao Z, Wan Y, Xue Y, Jiang Y, Yen E, Luo C, Wang J, Rong Y and Niu T 2020 Extracting and Selecting Robust Radiomic Features from PET/MR Images in Nasopharyngeal Carcinoma *Molecular Imaging and Biology 2020 22:6* **22** 1581–91 Online: https://link.springer.com/article/10.1007/s11307-020-01507-7

Yi X, Walia E and Babyn P 2019 Generative adversarial network in medical imaging: A review *Medical Image Analysis* **58** 101552

Yip S S F and Aerts H J W L 2016 Applications and limitations of radiomics *Physics in Medicine and Biology* **61** R150–66

Yip S S F, Coroller T P, Sanford N N, Huynh E, Mamon H, Aerts H J W L and Berbeco R I 2016 Use of registration-based contour propagation in texture analysis for esophageal cancer pathologic response prediction *Physics in Medicine and Biology* **61** 906–22

Yuan J, Xue C, Lo G, Wong O L, Zhou Y, Yu S K and Cheung K Y 2021 Quantitative assessment of acquisition imaging parameters on MRI radiomics features: a prospective anthropomorphic phantom study using a 3D-T2W-TSE sequence for MR-guided-radiotherapy *Quantitative imaging in medicine and surgery* **11** 1870–87 Online: https://pubmed.ncbi.nlm.nih.gov/33936971/

Zhang X, Zhong L, Zhang B, Zhang L, Du H, Lu L, Zhang S, Yang W and Feng Q 2019a The effects of volume of interest delineation on MRI-based radiomics analysis: evaluation with two disease groups *Cancer Imaging* **19** Online: /pmc/articles/PMC6925418/

Zhang Y, Yap P T, Qu L, Cheng J Z and Shen D 2019b Dual-domain convolutional neural networks for improving structural information in 3 T MRI *Magnetic Resonance Imaging* **64** 90–100

Zhao B 2021 Understanding Sources of Variation to Improve the Reproducibility of Radiomics *Frontiers in Oncology* **11** 633176

Zhong J, Wang Y, Li J, Xue X, Liu S, Wang M, Gao X, Wang Q, Yang J and Li X 2020 Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: Application to neonatal white matter development *BioMedical Engineering Online* **19** 1–18 Online: https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-020-0748-9

Zwanenburg A 2019 Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis *European Journal of Nuclear Medicine and Molecular Imaging* **46** 2638–55 Online: https://link.springer.com/article/10.1007/s00259-019-04391-8

Zwanenburg A, Vallières M, Abdalah M A, Aerts H J W L, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga R J, Boellaard R, Bogowicz M, Boldrini L, Buvat I, Cook G J R, Davatzikos C, Depeursinge A, Desseroit M-C, Dinapoli N, Dinh C V, Echegaray S, Naqa I el, Fedorov A Y, Gatta R, Gillies R J, Goh V, Götz M, Guckenberger M, Ha S M, Hatt M, Isensee F, Lambin P, Leger S, Leijenaar R T H, Lenkowicz J, Lippert F, Losnegård A, Maier-Hein K H, Morin O, Müller H, Napel S, Nioche C, Orlhac F, Pati S, Pfaehler E A G, Rahmim A, Rao A U K, Scherer J, Siddique M M, Sijtsema N M, Fernandez J S, Spezi E, Steenbakkers R J H M, Tanadini-Lang S, Thorwarth D, Troost E G C, Upadhaya T, Valentini V, Dijk L V van, Griethuysen J van, Velden F H P van, Whybra P, Richter C and Löck S 2020 The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping *Radiology* **295** 328–38