Physica Medica 83 (2021) 108–121

Contents lists available at ScienceDirect

Physica Medica

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

Original paper

Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization

Panagiotis Papadimitroulas[a,b], Lennart Brocki[c], Neo Christopher Chung[c,d], Wistan Marchadour[e], Franck Vermet[f], Laurent Gaubert[e,g], Vasilis Eleftheriadis[a], Dimitris Plachouris[b], Dimitris Visvikis[e], George C. Kagadis[b,*], Mathieu Hatt[e]

[a] *Bioemission Technology Solutions - BIOEMTECH, Athens, Greece*
[b] *3DMI Research Group, Department of Medical Physics, University of Patras, Rion GR 265 04, Greece*
[c] *University of Warsaw - Institute of Informatics, Warsaw, Poland*
[d] *University of California Los Angeles (UCLA) School of Medicine – Departments of Physiology and Medicine (Cardiology), USA*
[e] *LaTIM, INSERM, UMR 1101, Univ Brest, Brest, France*
[f] *LBMA, CNRS, UMR 6205, Univ Brest, Brest, France*
[g] *ENIB, Brest, France*

ARTICLE INFO

Keywords:
Deep learning
Machine learning
Convolutional neural network
Radiomics
Data curation
Explainability
Interpretability

ABSTRACT

Over the last decade there has been an extensive evolution in the Artificial Intelligence (AI) field. Modern radiation oncology is based on the exploitation of advanced computational methods aiming to personalization and high diagnostic and therapeutic precision. The quantity of the available imaging data and the increased developments of Machine Learning (ML), particularly Deep Learning (DL), triggered the research on uncovering "hidden" biomarkers and quantitative features from anatomical and functional medical images. Deep Neural Networks (DNN) have achieved outstanding performance and broad implementation in image processing tasks. Lately, DNNs have been considered for radiomics and their potentials for explainable AI (XAI) may help classification and prediction in clinical practice. However, most of them are using limited datasets and lack generalized applicability. In this study we review the basics of radiomics feature extraction, DNNs in image analysis, and major interpretability methods that help enable explainable AI. Furthermore, we discuss the crucial requirement of multicenter recruitment of large datasets, increasing the biomarkers variability, so as to establish the potential clinical value of radiomics and the development of robust explainable AI models.

1. Introduction

1.1. AI in oncology

Healthcare is expected to be highly impacted by machine learning (ML)-based artificial intelligence (AI). As deep learning (DL) relying on neural networks trained with large datasets has demonstrated state-of-the-art performances in numerous applications, massive structural changes in information and data processing in this sector are expected. Oncology is especially targeted by these developments, cancer being a major worldwide issue (18.1 million cases and 9.6 million deaths in 2018, respectively 22 and 13 million projected for 2030) [1]. Regarding predictive modeling based on multimodal medical imaging such as CT (computed tomography), PET/CT (positron emission tomography / CT) or MRI (magnetic resonance imaging), both academic and private research rely on ML/DL methods, however their clinical implementation and acceptability are currently lacking.

For decades, in medical oncology, patients suffering from cancer underwent diagnosis imaging acquisitions including PET/CT/MRI, where anatomical and functional information were combined to provide prognosis of the disease and an effective treatment plan. The extensive use of advanced hybrid imaging scanners increased the amount of diagnostic data in daily routine, enhancing the need of computational support for fast and accurate diagnosis [2]. Daily clinical applications seem to take more and more advantage of the rapid developments of AI alongside the evolution of computer science. Applications of medical imaging in oncology and image-guided radiotherapy include early diagnosis, staging, treatment decision and planning, monitoring, and

[*] Corresponding author at: 3DMI Research Group, Department of Medical Physics, Schoolf of Medicine, University of Patras, Rion GR 265 04, Greece.
E-mail address: gkagad@gmail.com (G.C. Kagadis).

https://doi.org/10.1016/j.ejmp.2021.03.009
Received 2 December 2020; Received in revised form 1 March 2021; Accepted 3 March 2021
Available online 22 March 2021

patient follow-up [3]. A patient's management may be optimized based on predictive models which are able to identify patients at risk of future treatment failures and recurrence. As some patients do not respond fully to the standard of care, different therapeutic strategies could be established based on these predictions. In addition, it is crucial to integrate data from several sources (clinical, imaging, dosimetry, genetics, toxicity, etc.) to improve predictive ability [4].

Besides automation in different stages of image processing, ML/DL opened a new era in clinical oncology, providing a more exhaustive and fast extraction of features from the diagnostic data, including some that may not be directly captured by the naked eye, including the expertly trained one. Quantification analysis of such features alongside with the combination of conventional anatomical and functional characteristics could further characterize tumors' profiles such as aggressiveness or potential of response to therapy, thereby informing clinical decision [5,6]. Radiomics and biomarkers selection and quantification are strongly interdependent with advanced ML/DL algorithms, which should be carefully used and extensively evaluated before being deployed in clinical practice. There are still several limitations and challenges to be addressed in the clinical application of AI in oncology, including the explainability and interpretability of the models, the sensitivity of the features' extraction, the reproducibility of the quantitative feature selection and the harmonization of the data.

1.2. *AI approaches using oncological biomarkers and radiomics*

On the one hand, radiomics has been introduced as the high-throughput extraction of "engineered" (or "handcrafted") features from medical images [6]. It has the potential to provide a quantitative signature of tumors' characteristics that cannot be appreciated visually [7] and has shown promising results in identifying tumor subtypes and in predicting outcome [8] by relying on ML methods to exploit radiomic features in combination with clinical or other variables to build predictive models. The majority of radiomics studies have been focused on oncology applications. On the other hand, the use of DL and specifically convolutional neural networks (CNNs) in computer vision have led to state-of-the-art results in filtering, segmentation, classification, and synthesis (image-to-image translation) including for medical images [9]. For these applications, the amount of available data (i.e. labeled pixels/voxels in the case of segmentation, filtering, synthesis) is usually sufficient for training deep networks. On the other hand, the attempts for predictive modeling in radiomics [10–12] where labels are on a patient basis (i.e., one label per 3D image volume, instead of per pixel/voxel) did not lead to very large improvements compared to the standard radiomics approach, showing in some instances similar but complementary predictive power (i.e., combining both approaches leads to better results), given the comparatively smaller amount of available training samples (for instance, several hundreds of patients in radiomics studies versus millions of images in ImageNet [13]). Nonetheless, the current research trends are clearly to rely more on DL-based techniques, as they may allow for a higher level of automation compared to the traditional workflow and may therefore facilitate its clinical translation.

1.3. *Interpretability of radiomics*

There is a plethora of research and review studies investigating the extraction of radiomics and the optimal combination with other diagnosis biomarkers, to be used in clinical applications. However, there is a clear limitation on the translation of such procedures in oncology practice and their explainability in terms of clinical routine. The majority of the available studies lack concrete, reproducible results, applicable to a larger set of applications and differentiated data [14]. There is a big challenge in the scientific community to translate and effectively use the multi-parametric models combining advanced mathematical models with numerous variables of clinically derived biomarkers [15,16]. In several studies there is the attempt to apply ML/DL methods in clinical routine applications. Such applications are described in Section 3.3.

Particularly in the context of DL the decision-making process of models is not transparent to humans and therefore interpretability is a crucial issue, especially in a potentially high-risk field such as radiomics. Advantages of an interpretable model are a raised confidence that the model will behave in the expected way when presented with unseen data and also a higher trust and acceptance of models by end users, e.g. physicians. Interpretability is therefore an important challenge that needs to be tackled in order to facilitate clinical implementation of DL models.

In this review we present an overview of the state-of-the-art of Deep Neural Networks (DNNs) on oncological applications, using radiomics. There is a focus on the latest developments and the future perspectives regarding interpretability and harmonization of imaging biomarkers. In Section 2 we present the main radiomics classification with their definition and also take a brief look at the different approaches of extracting radiomic features from medical images. In Section 3 the focus shifts on DNNs, first explaining in general the architecture of neural networks, multilayer perceptrons (MLPs) and CNNs, which are widely used in medical imaging. In Section 3.3 we present several clinical applications of DNNs in oncology, highlighting their advantages as well as possible drawbacks. The two final sections are concerned with two of the major challenges to the clinical application of DNNs, namely model interpretability and multicenter harmonization. In Section 4 the subject of Explainable Artificial Intelligence (XAI) is introduced through three major interpretability methods. Finally, in Section 5 are presented methods of processing imaging data for use in AI models that tackle issues associated with data curation, medical confidentiality, multicenter harmonization, expanding datasets and model generalization.

2. Radiomics classification

2.1. *Feature based radiomics*

Conventional radiomic approaches are usually known as feature-based radiomics, which are automatically or semi-automatically derived from medical images. Some of these features aim to the maximum exploitation of available diagnostic clinical data, by uncovering "hidden", difficult or impossible to appreciate with the naked eye, features for clinical use.

The standard approach to extract radiomics features requires the definition of the Volume or Region Of Interest (VOIs/ROIs) in the applied images. There are recent studies, showing the enhanced quality of information derived when hybrid imaging data (PET/MRI, PET/CT) is used in contrast to the use of each one modality alone [17,18]. In order to enable high reproducibility and interpretability of radiomics, a well-defined processing procedure of the data is required prior to the calculation of the handcrafted features themselves. Such processes are analytically described in Section 4. There are a large number of features (even more than 1000), based on mathematical models, usually considered in radiomic studies and they can be categorized into 4 main groups [2]:

1. Shape features [19]: provide quantitative description of geometric properties of the ROIs/VOIs, such as surface area, total volume, diameter, sphericity or surface-to-volume ratio.
2. First order statistics (histogram-based features): describe the fractional volume for the selected region of voxels and the distribution of the voxels' intensity, for example minimum, maximum, mean, variance, skewness, or kurtosis.
3. Second order statistics (textural features): These features are extracted based on matrices derived from intensity relationships of neighboring voxels in a 3D image [20], such as:

a. Gray Level Co-occurrence Matrix (GLCM): describes the spatial distribution of gray level intensities within a 3D image [21].

b. Gray Level Run Length Matrix (GLRLM): is defined as the number of contiguous voxels that have the same gray level value and it characterizes the gray level run lengths of different gray level intensities in any direction [22].
c. Gray Level Size Zone Matrix (GLSZM): quantifies gray level zones, the number of connected voxels that share the same gray level intensity, in a 3D image [23].
d. Neighbouring Gray Tone Difference Matrix (NGTMD): quantifies the difference between a gray value and the average gray value of its neighbours within a distance δ [24].
e. Gray Level Dependence Matrix (GLDM): quantifies the number of connected voxels within a distance δ that are dependent on the center voxel [25].

Second order features include entropy, uniformity, contrast, homogeneity, dissimilarity and correlation.

4. Higher order statistics features: These features are obtained by statistical methods after applying filters or mathematical transformations to the image, in order to highlight repeating patterns, edges, histogram-oriented gradients, or local binary patterns of the segmentation. These include fractal analysis, Minkowski functionals, wavelet and Fourier transformations, as well as Laplacian transformations of Gaussian-filtered images, which can extract areas with increasingly coarse texture patterns [26].

2.2. Deep learning radiomics (DLR) features

Deep learning based radiomic (DLR) features are obtained by normalizing the information from deep neural networks, especially CNNs, designed for image segmentation. The main hypothesis here is that once the image has been segmented accurately by a DNN, DNND the information about the segmented region is already stored within the network [27].

The first layer of an image processing DNN, whose architecture is described in detail in Sections 3.1 and 3.2, generally implements nonlinear template-matching at a relatively fine spatial resolution, extracting basic features of the data, thus detecting primitive patterns such as lines and edges. Subsequent layers learn to recognize particular spatial combinations of previous features, generating patterns of patterns in a hierarchical manner [28]. The higher layers of a deep neural network can often produce higher level features, which when the deep neural network's input is a medical image can be similar to the handcrafted radiomics features. These deep learning based radiomics features can be extracted from the last layers of the network. In this way a DNN can be used to convert 3D images into 1D vectors to allow medical image processing through deep learning, i.e. in an end-to-end fashion, or conventional machine learning methods.

The effectiveness of the deep learning radiomics features is highly related to the quality of the segmentation and the volume of the training dataset [29]. Therefore, in contrast to feature-based radiomics, large datasets are necessary to identify a relevant and robust feature subset. One other limitation of deep learning-based radiomics is the high correlation between the features and the input data, as the DLR features are generated from that very data without the application of prior knowledge [2].

3. Deep learning

Conventional machine learning had limited success in translating radiomic features into improving classification and prediction of cancer in clinical settings. Recently, deep learning has shown great potentials to improve feature engineering, classification, and prediction in medical imaging [9]. In this section, we review fundamentals of DNNs and CNNs.

3.1. Neural networks and multilayer perceptron (MLP)

To classify and predict clinical outcomes, supervised learning algorithms are trained on explanatory variables (e.g., input features) and response variables (e.g., output labels). In radiomics, classification tasks include diagnosis or prediction of response to therapy (e.g., benign vs. malignant lesions, responders vs. non responders to chemoradiotherapy), whereas regression tasks include time-to-event prediction (e.g., disease-free survival).

Generally, deep learning models consist of layers of connected neurons (Fig. 1), where the single neurons are defined through simple activation functions. By combining a large number of nodes and layers, deep learning can learn complex and nonlinear functions between input features and output labels, achieving high performance in a variety of computer vision problems [30,31]. In a MLP, input features (such as medical images) are trained against output labels, while adjusting parameters to maximize prediction accuracy (Fig. 1).

The network transforms an input into an output by a process called forward pass which consists in taking, in each layer, a weighted sum of inputs (resulting in z) and applying an activation function (f), usually the logistic function $f(x) = 1/(1+exp(-x))$ or the rectified linear unit (ReLU) $f(x) = max(0,x)$. The purpose of such an architecture is to find a (non-linear) combination of the input features such that the classes in consideration become linearly separable [32].

A hidden layer(s) is essentially performing automated feature engineering, which finds informative combinations of input features. In conventional radiomics, the process of finding suitable combinations of input features has to be performed manually which is referred to as feature engineering. Handcrafted features are derived using expert knowledge and some of them could be highly informative of cancer, whereas others could be irrelevant. An arduous process for feature evaluation and selection is therefore needed to obtain accurate models. The introduction of a hidden layer, or several hidden layers in case of a deep architecture, automates this process, by using an iterative process of feeding labeled data into the network and updating parameters (weights and biases) in a process called backpropagation [33]. Hence the network learns directly from the data which features are relevant for the task at hand.

In practice, one commonly updates the weights using stochastic gradient descent [34], which uses an estimate obtained from a randomly chosen subset of the training dataset. Updating of the weights is repeated for many subsets until the loss function is not decreasing anymore and the model has converged.

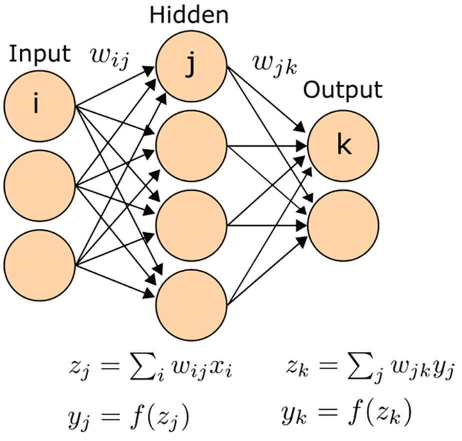

Fig. 1. Architecture of an MLP. Each layer is fully connected to the previous and following layer whereas within each layer the neurons are not connected. The index $i$is labeling the input features x and indices j,k are labeling the neurons in each layer, where $x$are the input features, w are the weights and $z$are the weighted sums of inputs and y activations.

The performance of the trained model is evaluated on a test dataset, which has been held out from training. If the performance of the model is significantly worse in a test dataset than on a training dataset, it may suggest overfitting, where the model has adjusted to inconsequential peculiarities during training and does not generalize well beyond this particular training dataset. While we may attempt to reduce the model complexity in a conventional overfitting context, deep learning takes advantage of over-parameterized regimes [35]. In over-parameterized deep learning models, one may combat overfitting with data augmentation [36], and weight regularization [37 38]. Additionally, cross validation could be used to select the best performing model out of multiple models under consideration [39].

3.2. Convolutional neural networks (CNN)

Multilayer perceptrons are not well suited to classify image data. First, the array representing the image has to be flattened into a one-dimensional input vector, removing spatial structures. Second, the MLP is not shift invariant such that a displacement of an input image fails the trained classification task. CNNs [40,41] overcome these challenges, accepting and being robust against shift of images or objects (Fig. 2).

A CNN typically consists of (a) convolutional layers that perform feature extraction, that are connected to (b) a MLP whose labels are response variables. The convolutional layers are organized in feature maps whose units are related to local patches of the previous layer through a small array of weights called a kernel or a filter. The value of each unit is obtained by calculating the weighted sum of activations of the previous layer using the kernel and applying an activation function. The process of obtaining the feature map is referred to as a discrete convolution of the kernel and the previous layer, hence the name. In a simplified form, it can be written as $z^l{}_{i,j} = f(x^l{}_{i,j}), x^l{}_{i,j} = \Sigma_{m,n} w^l{}_{m,n} z^{l-1}{}_{i+m,j+n} + b^l{}_{ij}$,

where f is the activation function, w is the kernel, z^{l-1} is the feature map in the previous layer and b is the bias. Intuitively, the convolution can be understood as scanning the image with a kernel and storing, for each position of the kernel, the result in the feature map. The weights are shared within each feature map, resulting in shift invariance and reduction of parameters.

DL models with CNNs typically learn a hierarchy of features, where higher-level features are composed of lower-level ones. As an illustration, the first layers may learn edges, which are then combined to shapes and parts which comprise the objects to be classified. This composition of features explains why it is crucial to downsample the image or feature maps via pooling [42] or larger strides [43], because in this way kernels in the deeper layers “see” a larger portion of the original image. The training of the network can be performed just in the same way as MLP, namely by using backpropagation of the loss to update the weights.

CNNs have been hugely successful in computer vision and excel at classification [31,44], object detection [45,46], and segmentation [47,48]. They have also been used in other fields such as speech recognition [49] and natural language processing [50].

3.3. Applications of deep learning in medical imaging

In recent years there has been a surge of applications of deep learning techniques in medical image analysis (see in-depth reviews in [9,51]). In many cases the proposed models perform as well or even outperform health-care professionals, for example in the classification of diseases [52]. Here we review selected applications structured by the tasks which they perform, namely classification, detection, segmentation and registration. Table 1 summarizes the several applications incorporating the performed techniques.

3.3.1. Classification

The problem of classification of medical images can be divided into two subproblems [9]: image/exam classification and object/lesion classification. Image classification considers an image as a whole to predict a diagnostic output, e.g. presence of a certain disease. Object classification on the other hand is concerned with the classification of predefined patches of an image, e.g. whether a nodule is benign or cancerous.

In image classification, especially in medical imaging, transfer learning is a very popular approach due to the comparatively small number of available images for a given task. Transfer learning uses the convolutional layers of a classifier previously trained on a different dataset as a feature extractor which especially for small datasets leads to improved accuracy. This approach has been successfully applied, for example, in the classification of skin cancer [53] and diabetic retinopathy [54] with accuracy comparable to human experts.

Object classification is more involved in the sense that it requires global information about the location of the object as well as local information about the object itself. For this reason, pretrained networks can not so easily be utilized and a so-called multi stream architecture is a popular approach. In [55] several CNNs are trained on different scales of nodule patches and the extracted features are combined and fed into the MLP and [56] uses a similar approach but considers multiple resolutions instead of scales.

3.3.2. Detection

In computer vision object detection seeks to locate and identify instances from a predefined number of classes in images, where usually the location of the objects is indicated by rectangular bounding boxes. Specifically, in medical image analysis one commonly differentiates the tasks of localization of anatomical structures and detection of objects and lesions.

Most approaches to identify anatomical structures in 3D images translate the problem into a 2D classification problem. The basic idea is to first train a CNN on orthogonal slices of the 3D volume to classify the presence of a certain structure and to subsequently obtain the

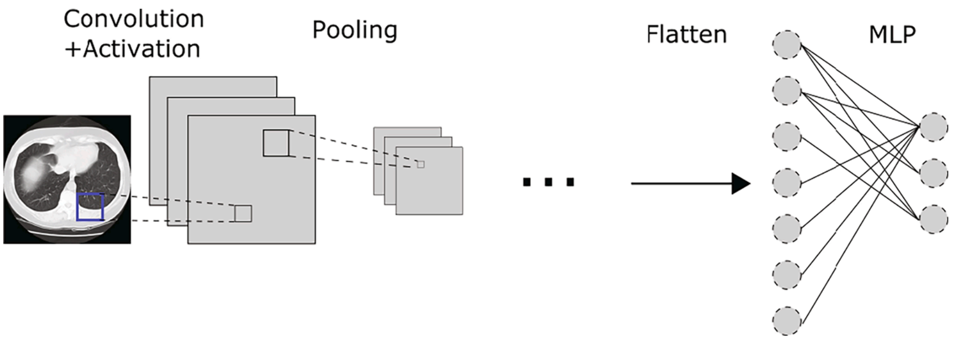

Fig. 2. Typical architecture of a CNN. In the first stage feature extraction is performed using convolutional and pooling layers, typically there are several such layers connected to each other which makes the network “deep”. The second stage consists of a MLP which is using the extracted features to perform class predictions.

Table 1
Summary of indicative techniques and examples of corresponding clinical applications on classification, segmentation, detection and registration using DNN/CNN.

Method	Technique	Application	Study
Classification	Image classification	Skin Cancer	A Esteva *et al.* 2017 [53]
	Image classification	Diabetic retinopathy	V Gulshan *et al.* 2016 [54]
	Object classification	Lung Nodules	W Shen *et al.* 2015 [55]
	Object classification	Skin Lesions	J Kawahara *et al.* 2016 [56]
Detection	3D translation to 2D classification	Bone localization	D Yang *et al.* 2015 [57]
	3D translation to 2D classification	Heart/Aorta localization	B de Vos *et al.* 2016 [58]
	Pixel wise classification	Histopathology	G Litjens *et al.* 2016 [60]
	Pixel wise classification	Coronary calcium scoring	JM Wolterink *et al.* 2016 [61]
	DL using 3D CNN	MRI Brain metastasis	O Charron *et al.* 2018 [62]
	DL using 3D CNN	MRI Brain Metastasis	E Grovik *et al.* 2020 [63]
Segmentation	U-Net Convolutional Network	Neuronal structures in electron microscopy	O Ronneberger *et al.* 2015 [64]
	3D U-Net Convolutional Network	Volumetric segmentation Xenopus kidney	O Cicek *et al.* 2016 [65]
	V-Net: Fully Convolutional Network	MRI prostate volumetric Segmentation	F Milletari *et al.* 2016 [66]
	Multi-scale 3D CNN connected with Conditional Random Field	MRI Brain lesions (injuries, tumors, ischemic stroke)	K Kamnitsas *et al.* 2017 [67]
	Supervised 3D supervoxel learning	Multimodal MRI Brain tumor	M Soltaninejad *et al.* [68]
	Fully CNN combined with Non-quantifiable Local Texture Feature	MRI Brain tumor	W Deng *et al.* 2019 [69]
	Adaptive Neuro Fuzzy Inference System with Textural Features	Glioma Brain tumor	A Selvapandian *et al.* 2018 [70]
Registration	CNN to derive transformation parameters	Neonatal brain tumor	M Simonovsky 2016 [71]
	CNN regression: Pose Estimation via Hierarchical Learning	Total Knee Arthroplasty Kinematics & X-ray transeophageal echocardiography	S Miao *et al.* 2016 [72]
	CNNs using artificial examples to adjust the transformation parameters	Lung radiotherapy	M Foote *et al.* 2019 [73]
	3D CNN	Proton Therapy prostate cancer	M Elmahdy *et al.* 2019 [74]

localization of it by calculating the intersection of slices which have been predicted to obtain the structure. This approach has been successfully applied to automatically localize landmarks on the distal femur [57] and the heart, aortic arch and descending aorta [58].

In order to perform object or lesion detection many authors perform pixel wise classification, which is usually obtained through a sliding window technique [59]. Intuitively, the idea is to train a classifier on small patches of images and to obtain pixel-wise predictions by classifying the patch around the pixel. Since a convolution also consists of sliding windows (kernels) this approach can be performed very efficiently for CNNs by turning a classifier into a fully convolutional network [59]. Two selected applications of this technique are in histopathological diagnosis [60] and coronary calcium scoring [61]. In a recent study, 3D convolutional neural network (DeepMedic), was applied and evaluated to both detect and segment brain metastasis on MRI data [62]. Accordingly E. Grovik et al. [63], used a DL approach based on a fully-CNN, to demonstrate automated detection and segmentation of brain metastases on multisequence MRI data.

3.3.3. *Segmentation*

The purpose of medical image segmentation is to find structures of interest, such as tumors and lesions, and marking the constituting pixels with the same label. Deep learning techniques have proven to be very effective in this task and segmentation is in fact the problem which is most commonly tackled using CNNs [9].

The most well-known CNN architecture used for segmentation for medical images is U-net [64], which uses upsampling convolutional layers to obtain segmentation maps with the same resolution as the input. This architecture allows training the model using entire images end-to-end, which allows the model to utilize the whole context of the image. There exist several variants of U-net, most notably ones that allow processing of 3D images [65,66]. The segmentation of lesions requires to combine models for object detection and segmentation and has been successfully implemented in [67].

M. Soltaninejad *et al.* [68] investigated a supervised learning based multimodal MRI brain tumour segmentation technique using textural features from supervoxels in a limited number of clinical datasets, concluding that increased number of data could provide higher accuracy in the segmentation process. In addition, W. Deng et al. [69] developed a brain tumor segmentation method integrating fully convolutional neural networks and dense micro-block difference features and compared their results with traditional MRI brain tumor segmentation techniques. The study used BRATS 2015 (Brain tumor image segmentation benchmark) and the training of the algorithms was based on 100 patients with MRI brain tumor data. Another recent study was evaluated using BRATS for the performance of the detection of tumor regions in Glioma brain data. Features extraction applied and were used for training applying an Adaptive Neuro Fuzzy Inference system (ANFIS) approach for the classification of a brain image into a healthy or an abnormal - Glioma - brain image [70]. Recently, DNNs were applied in automatic segmentation of brain metastases. A dataset of ~500 imaging data were used for the evaluation of the method, resulting in sensitivity and specificity which varied according to the size of the lesions.

3.3.4. *Registration*

The registration of medical images seeks to align images by finding appropriate coordinate transformations that maximize a certain similarity measure.

Simonowsky et al. [71] use CNNs to construct such a similarity measure for two patches from different modalities. Using this measure, they are also able to derive optimized transformation parameters to spatially align the patches. In order to perform a 3D model to 2D X-ray registration Miao *et al.* [72] use CNNs to directly learn the transformation by training the network using artificial examples which have been obtained by manually adjusting the transformation parameters. DL approaches are extensively under investigation on lung radiotherapy applications. M. Foote *et al.* [73], designed a patient-specific motion subspace and a DNN to recover anatomical positions to define the 2D-3D deformation of the lungs. In addition, a recent study investigated the development and validated a robust and accurate registration pipeline for automatic contouring for online adaptive Intensity-Modulated Proton therapy (IMPT) for prostate cancer applications [74]. There are a plethora of registration applications in medical imaging utilizing DNNs [75].

3.3.5. *Radiomics*

The radiomics community has started relying on DL techniques, to address some of the remaining challenges and limitations of the usual radiomics workflow [76,77]. This includes automation of the detection and segmentation steps, as well as harmonizing images through synthesis generative methods (see Section 5.3). Some studies have also explored relying on one or several deep networks to achieve predictions either by extracting features (that are subsequently combined through standard machine learning techniques) or as an end-to-end tool up to the prediction task [12,27,78–82]. Indeed, training a deep network from scratch on a limited size dataset can often be less efficient. One can thus extract "deep features'' from images using pre-trained networks. These "rough" to "fine" features at different scales through different layers can be exploited directly as well as combined with other handcrafted radiomic features to build even more accurate models [10,83–85] as shown in some studies listed in the non-exhaustive Table 2 below.

However, relying on DL methods in radiomics also requires addressing new challenges and facing several issues. These include the need for appropriate training with data augmentation techniques, constraints and prior knowledge due to the limited size of available datasets and their high level of heterogeneity, especially when training networks from scratch, as shown by some studies that achieved some success without having very large datasets to train their networks, as listed in the Table 3 below

Another issue that has not yet been fully addressed even in recent studies is the lack of interpretability of the models built through the use of deep networks (see Section 4).

4. Explainable artificial intelligence (XAI)

The high performance of end-to-end deep neural networks comes at the cost of high complexity and vast number of parameters. We may not be able to understand and explain why a deep learning model has made certain classifications in image analysis. This type of algorithm is often referred to as a "black box", in which we cannot comprehend internal decision processes. The final outputs (e.g., classifications or statistics) are accepted without justifications.

There are several benefits to expect from improved explainability of radiomics models, especially if they relied on deep learning methods. First, specialists can better understand how the models they develop learn from data, which can allow them to improve the models, especially in understanding how they potentially fail in new data. Second, non-specialists and especially end-users such as physicians, could better grasp the inner workings of the tools they rely on to make decisions for patients' management, which will increase their confidence in relying on them. In turn, confidence of patients in the tools will be increased if the physician can explain to them why he trusts the tool.

Table 2
Some examples of studies comparing and combining a standard radiomics approach with a deep learning one (mostly extraction of "deep" features using pre-trained networks.

Study	**Application/ Endpoint**	**Image modality (ies)**	**Methods**	**Conclusions**
Paul *et al.* [85]	Lung nodules classification (malignant vs. benign)	Low dose CT	Three strategies were compared and combined: standard radiomics, pre-trained CNN and CNN trained from scratch with data augmentation.	Combining all three strategies led to the best performance
Ning *et al.* [84]	gastrointestinal stromal tumors classification (malignant vs. benign)	CT	Standard radiomics vs. Pre-trained CNN based features, and combination into random forest	Combining both outperforms each approach separately
Antropova *et al.* [10]	Breast lesions classification (malignant vs. Benign)	FFDM, US, DCE-MRI	Standard radiomics vs. Pre-trained CNN based features, and combination into support vector machine	Combination always led to best results in all the three image modalities

Table 3
Examples of studies showing only marginal improvement using CNN compared to standard radiomics, implementing different strategies (e.g., data augmentation) to compensate for usual DNNs drawbacks (e.g., limited data size available for training, lack of interpretability).

Study	Application/ Endpoint	Image modality (ies)	Methods	Conclusions
Diamant *et al.* [12]	Head and neck cancer outcome prediction	CT	CNN trained from scratch on a 2D pre-segmented slice of the tumor (use of data augmentation by a factor of 20)	Slightly better performance using CNN compared to standard radiomics but not for all endpoints
Ypsilantis *et al.* [80]	Esophageal cancer response to therapy prediction	PET	CNN trained from scratch on a set of fused 2D pre-segmented slices of the tumor (use of data augmentation by a factor of greater than 55)	Slightly better performance using CNN compared to standard radiomics
Hosny *et al.* [79]	Lung cancer survival prediction	CT	3D CNN trained on pre-segmented volumes (use of data augmentation by a factor of 32000)	Slightly better performance of CNN over engineered features but not for all datasets.

Even though in principle one can follow every processing step, a huge number of parameters – e.g., the popular VGG-16 model has 138 million [86] – is making it infeasible to infer meaningful explanations of behaviors of the model in this way. Research in explainability and interpretability seeks to develop methods to reveal the behavior of a given model or to build models that are inherently more comprehensible for humans.

The concept of XAI is highly diverse, ranging from human computer interactions to visualization, and to interpretability metrics [87]. What it means for an algorithm to explain or how to evaluate interpretability are an active area of research and beyond the scope of this review. Instead, we focus on visual and statistical approaches that help us understand the application-based rationale behind deep learning models in the context of medical imaging. Understanding how exactly a model arrives at its predictions is important to ensure algorithmic fairness, identify a potential bias in a training dataset and to build trust that it performs on new data in an expected way [88]. Especially in sensitive fields such as radiomics, explainability is therefore a crucial criterion for widespread adoption. We summarize them in three categories:

4.1. Proxy models and model compression

Simpler and smaller models are more comprehensible as well as more efficient. Therefore, one may use more conventional statistical models to explain the operating characteristics of deep learning. A major challenge of interpreting deep neural networks is often raised due to the non-linearity of how input features are processed and incorporated into successive layers. Therefore, once deep neural networks are trained and demonstrate high performance, we can distill them into more conventional models [89]. Local Interpretable Model-agnostic Explanations (LIME) aims to explain a complex non-linear model by fitting a locally linear model in the vicinity of a certain prediction [90].

Beyond using a simpler proxy model merely for explanations, model compression seeks to capture the full spectrum (e.g., local and global) of accuracy while drastically reducing the number of parameters and complexities [91]. Particularly, Ba and Caruana [92] demonstrate that a shallow feed-forward net can learn the complex function previously learned by a deep model while maintaining accuracy. Hinton and Frost [93] devised a method to distill a deep learning model into a soft decision tree. In particular, they proposed to use predicted labels from a trained deep learning model, instead of a limited number of true labels, and to introduce adaptive penalties for regularization. They were able to build relatively compact decision trees with a slight reduction in prediction accuracy. Such soft decision trees can better represent a hierarchy of decisions that a human can interpret.

4.2. Visualization of intermediate features

Convolutional neural networks enabled high-performance deep learning in computer vision. For radiomics, convolutional layers can be seen as automated feature engineering that maximizes the prediction accuracy during training. Therefore, it is of great interest to identify which features have actually been learned by the convolutional layers. To this end, Olah et al. [94] proposed to perform a gradient ascent in the input space with respect to the activation of either a single unit in a feature map or for a whole feature map. Concretely, one starts with a pure noise as the input and iteratively changes its value in direction of the gradients in an optimization procedure. This leads to input images which maximally activate certain units or whole feature maps and therefore visualize what patterns the network is sensitive to.

Deconvolution [95,96] which is an inverse function of convolution, takes a different route to visualize learned features in convolutional layers. Essentially, once the model is trained, we set one of the output classes to one and other classes to zero and propagate through the network back to the input space. This backwards query maps the activation of the given output class back to the input and the resulting image can be understood as the network's internal representation of the output class [95,96]. Instead of starting from an output class, one may also arbitrarily start from an activation in any intermediate layer. The resulting image visualizes what shape or pattern this layer represents and is sensitive too.

4.3. Importance estimators and relevance scores

Input features, such as pixel or voxel values in medical images, are ultimately dictating classification. It is therefore of great interest to estimate the relative importance of input pixels for the classifications made by a model, i.e. to estimate which input pixels are the most relevant for a specific prediction. Because importance estimators can be visualized in the same dimensions as inputs, they are often referred to as saliency maps. There are two major approaches in which such a saliency map can be obtained. First, the perturbation methods measure the degradation of prediction accuracy, when small parts of the image are permuted, blurred, or generally perturbed [97–99].

Second, the gradient methods calculate the gradients of the class score with respect to the input pixels, where the class score is the activation of the neuron in the output vector corresponding to the class of interest [100,101]. There exist modifications of the standard method of gradient calculation via the chain rule. SmoothGrad [102] introduces imperceptible noises to the input image, which may result in more robust importance estimators. In Guided Backpropagation [43], negative gradients are set to 0, effectively discarding suppression of neuron activation. Rectified Gradient [103] generalizes this by allowing layer-wise thresholding with an extra hyper parameter. Grad-CAM [104] calculates the gradient of the class score with respect to channels, i.e. feature maps, in the convolutional layers instead of the input pixels. Thus, instead of the importance of input pixels, rather the importance of high-level features learned by the intermediate layers is quantified. The resulting coarse saliency map can be upscaled to the input dimension and combined with aforementioned pixel-level fine grained saliency maps to obtain a high-resolution and class-discriminative importance estimator. Fig. 3 depicts a clinical example of tumors and gradient class activation maps (Grad-CAM) [12].

Due to a lack of ground truth and several related methods, one must be careful in using importance estimators. Particularly, many of the proposed importance estimators are motivated mainly by visual appeal, such as high contrast and reduced noise. Many of these de-noised saliency maps may result in strong biases that do not correspond to true interpretability of underlying deep learning models [105,106]. The degradation of prediction accuracy while masking important pixels has been used to evaluate saliency maps [107]. It has, however, been pointed out that the observed degradation is potentially not only due to removing important pixels but intertwined with a deviation from the distribution of natural input images [108].

Overall, these three major categories of interpretability methods are widely used in application of deep learning models, although they have not yet been extensively exploited to help explain DL based radiomics models. From simplifying complex models to visualizing features that are important for predictions, one should inspect and scrutinize models to better understand operating characteristics. Further development of XAI will likely contribute to facilitating clinical translation of deep learning based radiomics.

5. Imaging data processing

5.1. Data curation

A typical patient medical record today might have an abundance of information sourcing from a standard blood test, to more advanced imaging studies, i.e. Computed Tomography, etc., as well as various omics tests. From the advent of Computerized Tomography in the 1970 s, the amount of medical image data has been steadily increasing in the healthcare enterprise. A typical CT in the 1970 s contained ~ 40 5-mm slices, while today it can contain more than ~ 2 k 512 × 512 slices. Likewise, the various exams that a patient is prescribed have increased in information, complexity, come from various healthcare reference centers, and need to conform to the various guidelines and directives of the hospital and the National or International Healthcare System.

Although small datasets may suffice for the training of AI algorithms, large, well-curated datasets with associated annotations are deemed necessary for AI algorithms in the clinical setting [109]. To this direction the preparation and organization of data from various sources through their lifecycle, rendering them available for processing for research and/or educational purposes is fundamental and is called data curation [110]. Data curation includes several steps like Ethical Approval, De-identification, appropriate labeling and pre-processing, as well as specific dataset types.

The U.S. Health Insurance Portability and Accountability Act (HIPAA), and the E.U. General Data Protection Regulation (GDPR) require data de-identification of patient data. De-identification is the procedure of removing patient specific sensitive information, like name, address, contact information, to name just a few [111]. This type of identification information is present in various data, such as DICOM medical images. There are several toolkits that remove this sensitive information like Conquest DICOM software [112], RSNA Clinical Trial Processor (CTP) [113], K-Pacs [114], DICOM library [115], DICOM-works [116], PixelMed DICOMCleaner [117], DVTK DICOM anonymizer [118], YAKAMI DICOM tools [119], *etc.* Furthermore, they can opt to convert data to a different file format such as NIfTI (Neuroimaging Informatics Technology Initiative) [120] so as DICOM metadata

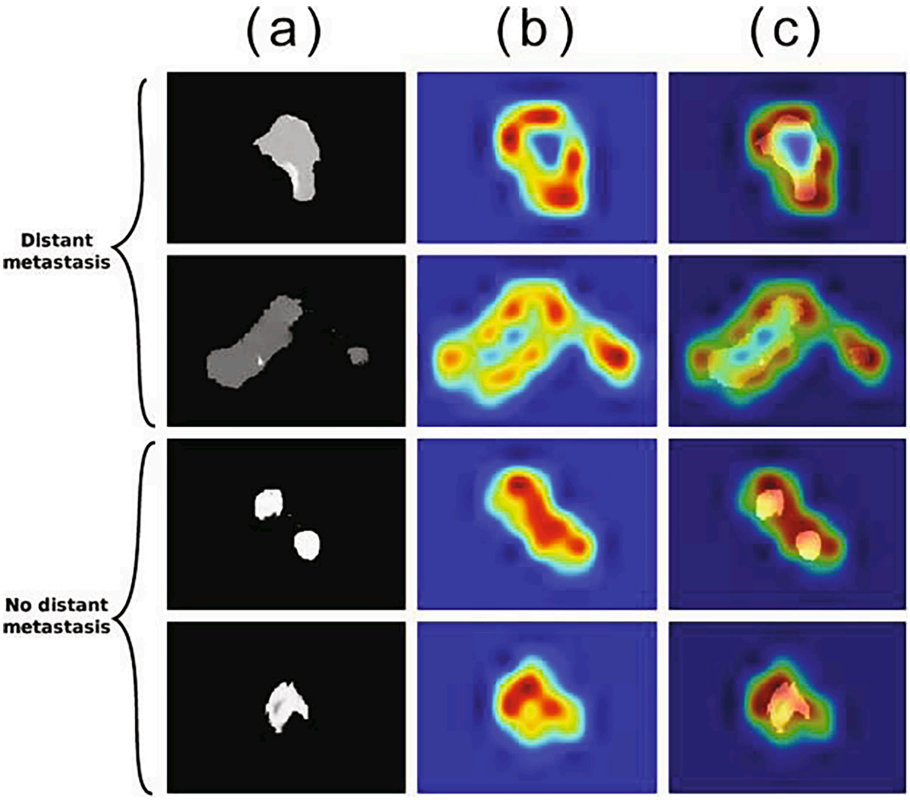

Fig. 3. Each row represents patients who did and did not developed distant metastasis in head and Neck Cancer. A) Raw images imported to the model, b) Gradient class activation map (Grad-CAM) of the penultimate convolution block, c) Merged image of columns a, and b. Red represents a region more significant to the designated classification (reproduced from the study of Diamant *et al.* [12]). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sensitive information is removed, leaving only the image voxel size and patient position for the AI algorithm.

Currently developed AI instances are generally based on supervised learning approaches [121]. To this end, the (surrogate of) ground truth, which is usually an established diagnosis (e.g., based on biopsy) or a known outcome (response to therapy established after treatment, follow up registration of events such as recurrence or death), needs be linked to the image of the patient. After this procedure, which is called labeling, the AI algorithm can be trained and tested on datasets. Although supervised learning dominates the AI field, unsupervised and semi-supervised learning can be used, especially when all or most of the data to be clustered/classified cannot be labelled using ground truth.

Another issue is that the dataset types may be coming from different manufacturers, vendors, institutions, countries (and thus different populations). If an AI algorithm is trained with data from a specific institution, on a specific vendor machine, and on a specific population, then the algorithm's performance might be overfitted to these and may not generalize well to other types of data. The AI algorithm should be thoroughly evaluated for generalization in other situations, where it might not work as efficiently or even completely fail. So, it is necessary that information used to train the AI algorithm come from various sources, or from a specific source if the algorithm is going to be deployed on a specific target population [122].

5.2. Multi-center-harmonization

Because most of previously published radiomics studies have been carried out using small, retrospective and monocenter cohorts of patients, the level of evidence regarding the potential added or complementary value of radiomics compared to clinical standard variables or simple metrics (such as in PET, metabolic volume and basic SUV max, peak or mean measurements) is considered to be rather weak [123,124]. In addition, the developed models are rarely tested on external datasets, even less often on several ones [125,126]. There is therefore a crucial need for the field to move away from the analysis of such small datasets, towards much larger ones in order to establish the potential clinical value of radiomics. With this need, comes the requirement of multicenter recruitment to achieve larger numbers. Another advantage of multicenter studies is the inherent variability in the data, which can make the cohort more representative and thus lead to more robust inference of models [127,128].

Collecting data from several centers is however complex for legal, ethical, administrative and technical reasons, although approaches such as distributed learning (a.k.a. federated learning in the machine learning community) consisting of the data not leaving the centers (only the models' parameters/weights are exchanged), can alleviate some of these issues [129]. Nonetheless, whether radiomic features are extracted from images stored locally in each center, or from images collected and stored in a centralized database, another major issue has to be considered. Indeed, most radiomic features have been shown to be highly sensitive to variations in several factors, including variability in scanner manufacturer, generation and models, acquisition protocols and reconstruction settings [130–132].

In some cases, factors involving relatively modest effects on the image characteristics from a visual point of view, can still have very important impact on some handcrafted radiomic features values and distributions (some being less robust than others to these effects). Thus, pooling these features together to perform any statistical analysis and build models can therefore lead to unreliable results, either hiding existing correlations or on the opposite, creating false discovery of correlative relationships [131,133]. Although it has not been extensively studied yet, using as input to CNNs PET/CT images with different characteristics and properties may also make the training of the network more complex or require more data than homogeneous datasets. On the other hand, a DL model could also benefit from heterogeneous data since it potentially leads to a model that is able to generalize better to unseen data. To understand the impact of harmonization in the context of DL therefore requires further investigation.

This variability in scanner manufacturers and models generations, acquisition protocols and reconstruction algorithms and settings are currently a clinical practice reality, and will likely remain as such in the near future. In addition, one has to emphasize on the fact that this variability may also exist within a single center. For instance, when a PET/CT scanner is replaced by a newer model from the same manufacturer, or by another model by a different manufacturer, images before and after the replacement will likely have different characteristics and extracted features will exhibit some changes in response to those. Similarly, if the center has several scanners, differences in manufacturer / model / acquisition / reconstruction may also exist amongst patients of the cohort. Finally, a given scanner may be used differently by different radiologists/nuclear medicine physicians (favoring different reconstruction algorithms or settings for example). As such, there may be larger differences between images acquired within a given center using different scanners than between two centers relying on the exact same model and associated acquisition protocol and reconstruction settings. Therefore, the lack of harmonization procedures of images and / or radiomic features is also a potential limitation within a single center context.

This has two important implications: i) first, when sharing image data for radiomic studies purposes, anonymization should be performed with caution, making sure that information relevant for the purpose of harmonization is kept in the DICOM files, such as for instance metadata about the scanner manufacturer and model, acquisition protocol (injected dose, etc.) as well as technical settings for the reconstruction (i.e., algorithm, implemented corrections, parameters, etc.); ii) when carrying out a radiomic analysis relying either on the extraction of handcrafted features or on the use of deep neural networks for feature extraction, metadata in DICOM files should be carefully checked to ensure proper data curation and extracting all the *a priori* knowledge about the acquisition and reconstruction of images in order to identify potential sources of bias and variability.

Taking these sources of variability into account is thus primordial for consistent and robust findings in any radiomics studies, even more so when multicenter data is considered. There exist a number of different approaches to address this issue, that can be classified in two groups: methods that address the issue in the image domain (i.e., before extracting features, either handcrafted ones or learning them directly from the images via a convolutional neural network) or in the features' domain (i.e., within or after the feature extraction step). On the one hand, addressing the issue in the image domain consists in harmonizing images directly so they have the same (or closer) properties (resolution, noise, texture, etc.). On the other hand, addressing it in the feature domain consists in harmonizing the features values, by either modifying how they are calculated (so they are less dependent on the varying factors in images) or directly modifying their distributions *a posteriori* so they can be pooled in the statistical analysis. Both approaches may also be combined, although this has not been extensively investigated yet. Most of the studies discussed below focus on one aspect or the other.

5.3. Harmonization in the image domain

5.3.1. Standardization of imaging procedures

One way to reduce the variability of the image properties is to standardize the acquisition and reconstruction protocols to achieve more similar images, according to specific criteria. This is the case in PET/CT where guidelines have been specifically developed to achieve images with closer recovery coefficients and SUV measurements across scanners [134–136]. Indeed, these existing standardization guidelines are mainly focused on qualitative and basic quantitative measurements and do not specifically encompass radiomic features values and distributions as a criterion to aim for in achieving standardization. Although these long-lasting standardization efforts need to be consolidated and maybe expanded to better take into account radiomics, their ability to help in decreasing the variations in radiomic features distributions across different sites, may nonetheless remain insufficient to compensate for the existing (and here to stay) diversity of scanner models and manufacturers proprietary reconstruction algorithms and post-processing tools across the various clinical centers. One recent study evaluated the performance of existing standardization guidelines for PET/CT imaging (i.e., EARL (European Association of Nuclear Medicine (EANM) Research Ltd. [135]) to reduce the variability of radiomic features across different scanner models and reconstruction settings [137]. They relied on a 3D printed phantom scanned on different systems across several centers. The differences between features extracted from PET images reconstructed with each clinical preferred setting and those extracted from the EARL-compliant reconstructions were important. A large percentage of radiomic features exhibited significant differences, even after standardizing the imaging procedures (acquisition protocols, reconstruction settings). This approach is feasible only for prospectively collected images, where it is allowed to modify the acquisition parameters. However, the majority of radiomics studies are retrospective [138]. Therefore, they are carried out by collecting images that have already been acquired and reconstructed. To evaluate the impact of different reconstructions, requires the storage of the raw data, which is rarely done in daily practice [139]. For retrospectively collected images, an approach that can work on already reconstructed images is therefore necessary.

5.3.2. Processing images

One approach is to apply image processing techniques before handcrafted feature extraction or analysis through a CNN. A common and popular example of such pre-processing is interpolation of all considered images to a common voxel size and applying filtering techniques so they would have similar resolution and noise characteristics. It is not trivial to implement, as there exist dozens of algorithms for image interpolation and filtering, so figuring out the most effective combination could be quite challenging and time-consuming. However, as isotropic voxels are recommended in the specific context of handcrafted textural features calculation by the IBSI (image biomarker standardization initiative) guidelines [16], interpolation to a common isotropic voxel size is often performed as a default pre-processing step in recent radiomic studies, so if images with variable reconstruction matrix sizes are considered, it can be beneficial also, although the choice of the common size parameter might be tricky. CNN also usually require images of identical size to be input in the network, so they are also interpolated before being fed to networks [140].

It has been suggested that interpolating images to a common voxel size for the purpose of harmonizing images and obtaining comparable (i.e. poolable in the statistical analysis) handcrafted radiomic features may be insufficient to fully remove the center effect [141]. Filtering images to achieve a similar spatial resolution may be quite detrimental in terms of textural analysis, if the common lowest denominator is chosen [142], which means higher resolution images are smoothed, hence removing details.

Another promising recently developed approach consists in relying on image synthesis through deep networks, such as GANs. The idea is to synthetize images with more similar properties for the specific goal of harmonization, so that handcrafted radiomic features extracted from harmonized images are comparable, or to facilitate training of deep neural network modeling. A recent work investigated the effect of different reconstruction kernels on radiomic features and evaluated the benefit on handcrafted features reproducibility to train a CNN to convert images from one reconstruction kernel to another, in a database of 104 lung cancer patients [143]. It demonstrated that different reconstruction kernels led to most of the features having significantly different distributions (595 out of 702), whereas after the proposed CNN-based image conversion, a larger percentage of features did not exhibit significant differences anymore (57%, 403 out of 702). Almost half of the features continued to exhibit differences, however. Another recent work relied on a two-step framework for multicenter image-based standardization using conditional generative adversarial networks (cGANs) to harmonize multicenter MRI brain images [144], while another relied on bidirectional translation between unpaired MRI images through a cycle-consistent GAN that uses 2 generator-discriminator pairs to achieve harmonization of DCE-MR images of breast [145]. A third study implemented a dual-GAN framework to harmonize diffusion tensor imaging (DTI) derived metrics on neonatal brains, demonstrating improved harmonization performance compared to standard approaches including voxel-wise scaling, and ComBat [146]. However, these studies did not extensively evaluate the resulting impact on multicenter radiomic studies. One recent work did so in the context of multicenter CT images, by relying on a GAN trained on different datasets to learn how to harmonize from one domain to another. Then a lasso classifier to stratify patients according to survival was trained using 77 radiomic features and evaluated in a cross-validation framework across the different domains [147]. Results show that relying on harmonized images to extract radiomic features improved the performance of the lasso classifier by an average of 11% in area under the receiver operating characteristic curve (from 3 to 32%).

5.4. Harmonization in the feature domain

5.4.1. Selection of features based on their reliability

One strategy consists in eliminating radiomic features because they have been identified to be unreliable (i.e., exhibit unreasonable variations in response to small variability of acquisition and reconstruction settings [130] or in a test–retest framework [148] before even considering them in any statistical analysis [149]. This can help build models with higher validation performance when tested on new data, as the features included in the model are expected to be robust to potential differences in image properties. Another advantage of this approach is that it reduces substantially the amount of variables to deal with in the modeling step, which can facilitate selection of features and building of multiparametric models. However, a drawback to consider is the potential loss of clinically-relevant information carried by the discarded features. One can only hope that the predictive power can still be found in the remaining features. In addition, the identification of the features that are both sufficiently reliable and carry enough clinically relevant information needs to be performed for each combination of clinical application and type of imaging data to be most appropriate for each case.

5.4.2. Modifying the feature's definitions

Because a number of radiomic features have been shown to be dependent on the number of voxels included in the calculation [150], it has been proposed to revise the feature definitions themselves to remove or reduce this dependency by including the number of voxels in the mathematical formulation [151]. Coincidentally, this can contribute to reduce the differences between radiomic features due to being extracted from images with different voxel sizes (and therefore different number of voxels for similar volumes of interest), as it has been shown in texture phantom data acquired in 8 different CT scanners from 3 different manufacturers [151], further validated in images of lung cancer [152].

5.4.3. Normalization

A large number of statistical methods have been proposed for statistical normalization [153].

A number of studies specifically evaluated the benefit of normalization techniques for the purpose of correcting biases and differences in radiomic features due to variations in imaging devices, acquisition protocols or reconstruction. A method for feature correction and bias reduction due to difference in exposure in the CT acquisition was proposed, by learning from phantom and clinical data how to model the differences, and then applying that learned correction to features values, thereby demonstrating at least 2 times standard deviation reduction for 47 out of 62 features [154]. Another recent work trained a deep neural network to standardize radiomic and "deep" features across scanners models and acquisition and reconstruction settings, relying on a publicly available texture phantom dataset [155]. It also showed the ability to transfer the learned standardization to new data coming from unknown scanners. The use of normalization to obtain more robust predictive radiomic models for validation in external data was demonstrated by normalizing features separately for each dataset rather than performing the normalization for all datasets combined [156]. Another study relied on z-score normalization to harmonize radiomic features extracted from pretreatment MRI for building a model predicting response in a multicenter study of 275 cervical cancer patients from 8 different centers [157]. High performance was obtained for the predictive models, although the study did not report performance without the normalization.

5.4.4. Batch effect removal

ComBat is designed to estimate a batch-specific transformation to express all data in a common space devoid of center effects [158] and has been shown to provide satisfactory results even for small datasets [159]. An extensive comparison of the previously described normalization techniques in the specific context of radiomics has not yet been carried out, although previous comparisons between ComBat and similar techniques for batch effect correction in different fields (including genomics) indicated superiority of ComBat. Recently, a study compared ComBat with SVD decomposition and voxel size resampling in the context of CT imaging using phantom data and a clinical cohort of patients with colorectal/renal cancer liver metastases [160]. The results indicated that the best harmonization was achieved with ComBat.

ComBat was first evaluated for harmonization of radiomic features in the context of PET [133] imaging and was later evaluated for CT [161] and MRI [162]. It has been exploited in a number of radiomic clinical studies to improve results of predictive models: in FDG PET and MRI radiomics for locally advanced cervical cancers (accuracy improved from 76 to 81% before harmonization to 81–97% after ComBat applied to the three centers) [163] and in FDG PET/CT for early-stage lung cancer where features had lower predictive power without harmonization, and ComBat allowed validating the model trained in 3 centers when applied to the fourth one [164]. Its benefit was also evaluated in the context of DCE MRI images of breast cancer to differentiate 3150 malignant and benign lesions, where classification performance using harmonized features was significantly higher ($p < 0.001$) [165]. The method was recently used in a multicenter CT study to harmonize radiomic features extracted from the different CT scanners in order to build reliable models predictive of outcome of COVID-19 patients [166]. Unfortunately, this study did not report performance of radiomic features without ComBat. Finally, in a recent study, radiomic models were trained to identify malignant nodules in early diagnosis of lung cancer with low-dose CT and externally validated [167]. All models had a high performance in the external validation set (AUC above 0.82), and this was not significantly altered when relying on ComBat-harmonized features.

Combat therefore seems a promising operational and simple method to perform harmonization of radiomic features, providing the number of labels is reasonable and the sources of variations can be identified and labeled. In case of very high heterogeneity where the number of labels to use with ComBat would be too high with respect to the number of patients, unsupervised clustering can be relied upon to identify potential labels to use for harmonization [168]. To avoid features to lose their physical meaning after harmonization, a variant of ComBat allowing to select a reference to align other labels to (instead of averaging all distribution to an arbitrary grand mean), name M−ComBat, can be used with no loss of performance. Finally, the robustness of the estimation can also be improved through bootstrapping and Monte Carlo (B-Combat) [168].

6. Discussion and conclusions

In modern radiation oncology, AI techniques have found several applications in many research domains, ranging from image processing for diagnosis to the optimization of precise therapeutic protocols. The exploitation of imaging biomarkers and radiomics features provided a new metric for quantitative image analysis, aiming to support clinical decisions, in detection, characterization and treatment planning on several pathologies. DNNs seem to provide the potential for a revolution in the field of medical imaging and radiotherapy, opening a new era on the personalization of diagnostic and therapeutic radiation protocols [169]. Although the rapid and increasing developments of DL approaches in imaging biomarkers, the state-of-the-art methodologies in radiomics provide several limitations and need to address great challenges in terms of explainability, interpretability and homogenized approaches (multi-center data harmonization).

Despite the developments of DL models in radiomics, it is still an issue the concept of understanding and explaining the way that the classifications and the predictions are done. The concept of XAI, the metrics used and the evaluation of interpretability is highly debated and under investigation in the scientific community. Furthermore, the repeatability, transferability and reproducibility of radiomics are of high interest as they are dependent on each specific imaging acquisition. In a recent study, the authors investigated imaging biomarker radiomics that seemed to be repeatable and reproducible within the reviewed studies [170]. In this framework, IBSI provided guidelines and radiomics nomenclatures and definitions, to support the verification of feature extraction in the field of radiomics [16].

DLR features may provide advantages in DNNs showing higher generalization and transferability compared to feature based radiomics. However, even the models developed for DLR, they still lack reliability and explainability for their application in clinical practice.

It is of crucial importance to put effort on the standardization of the models and generalization, applying harmonization methodologies to enable the understanding of several published results. Small datasets, dependency on image acquisition protocols (data analysis, imaging modality, quality of image, processing methods), however, are still difficult problems which complicate reaching this understanding. There are already several open-source software, available in the scientific community for radiomics research, such as, Keras [171], TensorFlow [172], LifeX [173], MaZda [174], PyTorch and PyRadiomics [175]. Even though the procedures applied and the workflow in such packages are not simple and lack generalization, with result not to allow researchers to fully comprehend the results and even more to reproduce them.

Last but not least, the major issue of radiomics, is their interpretability in clinical routine. Till now, most of radiomics extraction and imaging biomarkers analysis are used as "black box", making it impossible to clinically translate their usage [29].

It is highly encouraging that the aforementioned limitations are now well-known and are recently widely discussed in literature, moving the research on more focused studies, on addressing the challenges, and finding ways for the clinical exploitation of AI developments in the field of radiomics.

Declaration of Competing Interest

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

Acknowledgements

This study was co-financed by:

a) the CHIST-ERA grant [CHIST-ERA-19-XAI-007] with project acronym INFORM, by General Secretariat for Research and Innovation (GSRI) of Greece, National Science Centre (NCN) of Poland [2020/02/Y/ST6/00071] and Agence Nationale de la Recherche (ANR) of France.

b) the National Science Centre of Poland [2016/23/D/ST6/03613].

c) Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Reinforcement of Postdoctoral Researchers - 2nd Cycle" (MIS-5033021), implemented by the State Scholarships Foundation (IKY).

d) the European Commission under the Marie Skłodowska-Curie Actions - Research and Innovation Staff Exchange project AERAS (contract n. 872735).

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424.

[2] Lohmann P, Bousabarah K, Hoevels M, Treuer H. Radiomics in radiation oncology-basics, methods, and limitations. Strahlenther Onkol 2020;196:848–55.

[3] Zafra M, Ayala F, Gonzalez-Billalabeitia E, Vicente E, Gonzalez-Cabezas P, Garcia T, et al. Impact of whole-body 18F-FDG PET on diagnostic and therapeutic management of Medical Oncology patients. Eur J Cancer 2008;44:1678–83.

[4] Jaffray DA, Das S, Jacobs PM, Jeraj R, Lambin P. How Advances in Imaging Will Affect Precision Radiation Oncology. Int J Radiat Oncol Biol Phys 2018;101: 292–8.

[5] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14:749–62.

[6] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441–6.

[7] Radiomics AH. there is more than meets the eye in medical imaging. Proceedings of the SPIE. 2016;9785.

[8] Hatt M, Cheze Le Rest C, Tixier F, Badic B, Schick U, Visvikis D. Radiomics: data are also images. J Nucl Med. 2019;60:38S–44S.

[9] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.

[10] Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med Phys. 2017;44:5162–71.

[11] Bibault JE, Giraud P, Housset M, Durdux C, Taieb J, Berger A, et al. Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. Sci Rep. 2018;8:12611.

[12] Diamant A, Chatterjee A, Vallieres M, Shenouda G, Seuntjens J. Deep learning in head & neck cancer outcome prediction. Sci Rep. 2019;9:2764.

[13] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115:211–52.

[14] Song J, Yin Y, Wang H, Chang Z, Liu Z, Cui L. A review of original articles published in the emerging field of radiomics. Eur J Radiol. 2020;127:108991.

[15] Schick U, Lucia F, Bourbonne V, Dissaux G, Pradier O, Jaouen V, et al. Use of radiomics in the radiation oncology setting: Where do we stand and what do we need? Cancer/Radiotherapie. 2020;24:755–61.

[16] Zwanenburg A, Vallieres M, Abdalah MA, Aerts H, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 2020;295: 328–38.

[17] Lohmann P, Kocher M, Ceccon G, Bauer EK, Stoffels G, Viswanathan S, et al. Combined FET PET/MRI radiomics differentiates radiation injury from recurrent brain metastasis. Neuroimage Clin 2018;20:537–42.

[18] Lohmann P, Kocher M, Ruge MI, Visser-Vandewalle V, Shah NJ, Fink GR, et al. PET/MRI Radiomics inpatients with brain metastases. Front Neurol. 2020;11:1.

[19] Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging. 2012;30:1234–48.

[20] Parekh V, Jacobs MA. Radiomics: a new application from established techniques. Expert Rev Precis Med Drug Dev. 2016;1:207–26.

[21] Nanni L, Brahnam S, Ghidoni S, Menegatti E, Barrier T. Different approaches for extracting information from the co-occurrence matrix. PLoS ONE 2013;8:e83554.

[22] Xiaoou T. Texture information in run-length matrices. IEEE Trans Im Proc. 1998; 7:1602–9.

[23] Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, et al. Texture indexes and gray level size zone matrix. Application to cell nuclei classification. Pattern Recognition and Information Processing. 2009:140–5.

[24] Amadasun M, King R. textural features corresponding to textural properties. IEEE Systems, Man and Cybernetics. 1989;19:1264–74.

[25] Sun C, Wee W. Neighboring gray level dependence matrix for texture classification. Comput Graph Image Process. 1982;20:297.

[26] Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp. 2018;2: 36.

[27] Li Z, Wang Y, Yu J, Guo Y, Cao W. Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. Sci Rep. 2017;7: 5467.

[28] Vial A, Stirling D, Field M, Ros M, Ritz C, Carolan M, et al. The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. Transl Cancer Res. 2018;7:803–16.

[29] Avanzo M, Wei L, Stancanello J, Vallieres M, Rao A, Morin O, et al. Machine and deep learning methods for radiomics. Med Phys. 2020;47:e185–202.

[30] Kaiming H, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv:1512.03385 [cs.CV].

[31] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In 'Advances in Neural Information Processing Systems', edited by F. Pereira and C. J. C. Burges and L. Bottou and K. Q. Weinberger 2012;25.

[32] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.

[33] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323:533–6.

[34] Robbins H, Monro S. A Stochastic Approximation Method. Ann Math Statist. 1951;22:400–7.

[35] Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. Proc Natl Acad Sci U S A. 2019;116: 15849–54.

[36] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6:60.

[37] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Machine Learning Res 2014;15:1929–58.

[38] Krogh A, Hertz JA. A Simple Weight Decay Can Improve Generalization. 4th International Conference on Neural Information Processing Systems: Morgan-Kaufmann. p. 950-7.

[39] Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. J Royal Statistical Society: Series B. 1974;36:111–33.

[40] Le Cun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc of the IEEE. 1998.

[41] Le Cun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Handwritten digit recognition with a back-propagation network. NeurIPS Proceedings. p. 396-404.

[42] Scherer D, Muller A, Behnke S. Evaluation of pooling operations in convolutional architectures for object recognition. 20th International Conference on Artificial Neural Networks. Thessaloniki, Greece2010.

[43] Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. 2014.

[44] Rawar W, Wang Z. Deep convolutional neural networks for image classification: A comprehensive review. Neural Comput 2017;29:2352–449.

[45] Girshick R, Donahue J, Darrell T, J. M. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition: IEEE; 2014. p. 580-7.

[46] Redmon J, Divvala S, Girshick R, Fargadi A. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition: IEEE; 2016. p. 779–88.

[47] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. arXiv:1505.04366 [cs.CV]. 2015.

[48] Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 [cs.CV]. 2017.

[49] Hinton G, Deng L, Yu D, Dahl GE, Mohamed A, Jaitly N, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Process Mag 2012;29:82–97.

[50] Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning. ICML '08: Proceedings of the 25th international conference on Machine learning2008. p. 160-7.

[51] Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng. 2017;19:221–48.

[52] Liu X, faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health. 2019;1:E271-E97.

[53] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542: 115–8.

[54] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016;316:2402–10.

[55] Shen W, Zhou M, Yang F, Yang C, Tian J. Multi-scale Convolutional Neural Networks for Lung Nodule Classification. Inf Process Med Imaging. 2015;24: 588–99.

[56] Kawahara J, Hamarneh G. Multi-resolution-Tract CNN with Hybrid Pretrained and Skin-Lesion Trained Layers. International Workshop on Machine Learning in Medical Imaging2016. p. 164-71.

[57] Yang D, Zhang S, yan Z, Tan C, Li K, Metaxas D. Automated anatomical landmark detection ondistal femur surface using convolutional neural network. 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI); 2015. p. 17-21.

[58] de Vos BD, Wolterink JM, Viergever MA, Isgum I. 2D image classification for 3D anatomy localization: employing deep convolutional neural networks. SPIE Medical Imaging; 2016.

[59] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. arXiv:14114038 [csCV]; 2015.

[60] Litjens G, Sanchez CI, Timofeeva N. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep. 2016;6:26286.

[61] Wolterink JM, Leiner T, de Vos BD, van Hamersvelt RW, Viergever MA, Isgum I. Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. Med Image Anal. 2016;34:123–36.

[62] Charron O, Lallement A, Jarnet D, Noblet V, Clavier JB, Meyer P. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. Comput Biol Med. 2018;95:43–54.

[63] Grovik E, Yi D, Iv M, Tong E, Rubin D, Zaharchuk G. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. J Magn Reson Imaging. 2020;51:175–82.

[64] Ronneberger O, DFischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI 2015: Medical Image Computing and Computer-Assisted Intervention; 2015. p. 234-41.

[65] Cicek O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. MICCAI 2016 : Medical Image Computing and Computer-Assisted Intervention 2016.

[66] Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. arXiv:160604797v1 [csCV]; 2016.

[67] Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal. 2017;36:61–78.

[68] Soltaninejad M, Yang G, Lambrou T, Allinson N, Jones TL, Barrick TR, et al. Supervised learning based multimodal MRI brain tumour segmentation using texture features from supervoxels. Comput Methods Programs Biomed. 2018;157: 69–84.

[69] Deng W, Shi Q, Luo K, Yang Y, Ning N. Brain Tumor Segmentation Based on Improved Convolutional Neural Network in Combination with Non-quantifiable Local Texture Feature. J Med Syst. 2019;43:152.

[70] Selvapandian A, Manivannan K. Fusion based Glioma brain tumor detection and segmentation using ANFIS classification. Comput Methods Programs Biomed. 2018;166:33–8.

[71] Simonovsky M, Gutierrez-Becker B, Mateus D, Navab N, Komodakis N. A Deep Metric for Multimodal Registration. MICCAI 2016: Medical Image Computing and Computer-Assisted Intervention2016. p. 10-8.

[72] Miao S, Wang ZJ, Liao R. A CNN Regression Approach for Real-Time 2D/3D Registration. IEEE Trans Med Imaging 2016;35:1352–63.

[73] Foote MD, Zimmerman BE, Sawant A, Joshi SC. Real-Time 2D-3D Deformable Registration with Deep Learning and Application to Lung Radiotherapy Targeting. Information Processing in Medical Imaging IPMI 20192019. p. 265-76.

[74] Elmahdy MS, Jagt T, Zinkstok RT, Qiao Y, Shahzad R, Sokooti H, et al. Robust contour propagation using deep learning and image registration for online adaptive proton therapy of prostate cancer. Med Phys. 2019;46:3329–43.

[75] Beoveiri HR, Khayami R, Javidan R, Mehidizadeh A. Medical image registration using deep neural networks: A comprehensive review. Comput Electr Eng 2020; 87:106767.

[76] Hatt M, Le Rest CC, Tixier F, Badic C, Schick U, Visvikis D. Data are also images. J Nucl Med. 2019;60:38S–44S.

[77] Hatt M, Parmar C, Qi J, El Naqa I. Machine (Deep) Learning Methods for Image Processing and Radiomics. IEEE Trans Rad Plasma Med Sci 2019;3:104–8.

[78] Amyar A, Ruan S, Gardin I, Chatelain C, Decazes P, Modzelewski R. 3-D RPET-NET: Development of a 3-D PET Imaging Convolutional Neural Network for Radiomics Analysis and Outcome Prediction. IEEE Trans Radiation Plasma Med Sci 2019;3:225–31.

[79] Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med. 2018;15:e1002711.

[80] Ypsilantis PP, Siddique M, Sohn HM, Davies A, Cook G, Goh V, et al. Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. PLoS ONE 2015;10:e0137036.

[81] Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, et al. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. Sci Rep. 2017;7: 10353.

[82] Zheng X, Yao Z, Huang Y, Yu Y, Wang Y, Liu Y, et al. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nat Commun. 2020;11:1236.

[83] Choi YS, Bae S, Chang JH, Kang SG, Kim SH, Kim J, et al. Fully Automated Hybrid Approach to Predict the IDH Mutation Status of Gliomas via Deep Learning and Radiomics. Neuro Oncol 2020.

[84] Ning Z, Luo J, Li Y, Han S, Feng Q, Xu Y, et al. Pattern Classification for Gastrointestinal Stromal Tumors by Integration of Radiomics and Deep Convolutional Features. IEEE J Biomed Health Inform. 2019;23:1181–91.

[85] Paul R, Hawkins SH, Schabath MB, Gillies RJ, Hall LO, Goldgof DB. Predicting malignant nodules by fusing deep features with classical radiomics features. J Med Imaging (Bellingham). 2018;5:011021.

[86] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:14091556v6 [csCV]2015.

[87] Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:170208608v2 [statML]2017.

[88] Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining Explanations: An Overview of Interpretability of Machine Learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)2018. p. 80-9.

[89] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. 2015.

[90] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD '16. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 1135–44.

[91] Bucila C, Caruana R, Niculescu-Mizil A. Model ompression. KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006: 535–541.

[92] Ba J, Caruana R. Do Deep Nets Really eed to be Deep? in Advances in Neural Information Processing Systems, edited by Z. Ghahramani and M. Welling and C. Cortes and N. Lawrence and K. Q. Weinberger. 2014:27.

[93] Frosst N, Hinton G. Distilling a Neural Network Into a Soft Decision Tree. arXiv: 1711.09784 [cs.LG]. 2018.

[94] Distill. Feature Visualization: How neural networks build up their understanding of images. https://distill.pub/2017/feature-visualization/. Accessed on November 28, 2020.

[95] Zeiler MD, Taylor GW, Fergus R. Adaptive Deconvolutional Networks for Mid and High Level Feature Learning. ICCV 20112011.

[96] Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. ECCV 2014: Computer Vision 2014:818–33.

[97] Samek W, Binder A, Montavon G, Lapuschkin S, Muller K. Evaluating the Visualization of What a Deep Neural Network Has Learned. EEE Trans Neural Networks Learning Syst 2017:2660–73.

[98] Fong RC, Vedaldi A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. 2017 IEEE International Conference on Computer Vision (ICCV) 2017. p. 3449-57.

[99] Zintgraf LM, Cohen TS, Adel T, Welling M. Visualizing deep neural network decisions: prediction difference analysis. ICLR 20172017.

[100] Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, Muller KR. How to Explain Individual Classification Decisions. J Machine Learning Res 2010;11: 1803–31.

[101] Erhan D, Bengio Y, Courville A, Vincent P. Technical Report 1341: Visualizing Higher-Layer Features of a Deep Network. Universite de Montreal; 2009.

[102] Smilkov D, Thorat N, Kim B, Viegas F, Wattenberg M. SmoothGrad: removing noise by adding noise. arXiv:170603825v1 [csLG]2017.

[103] Kim B, Seo J, Jeon S, Koo J, Choe J, Jeon T. Why are Saliency Maps Noisy? Cause of and Solution to Noisy Saliency Maps. arXiv:190204893 [csLG]2019.

[104] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV)2017. p. 618-26.

[105] Brocki L, Chung NC. Input Bias in Rectified Gradients and Modified Saliency Maps. 2021 IEEE International Conference on Big Data and Smart Computing (BigComp). https://doi.org/10.1109/BigComp51126.2021.00036.

[106] Abebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity Checks for Saliency Maps. arXiv: 1810.03292 [cs.CV]. 2018.

[107] Samek W, Binder A, Montavon G, Lapuschkin S, Muller KR. Evaluating the Visualization of What a Deep Neural Network Has Learned. IEEE Trans Neural Netw Learn Syst. 2017;28:2660–73.

[108] Hooker S, Erhan D, Kindermans PJ, Kim B. A Benchmark for Interpretability Methods in Deep Neural Networks. arXiv:180610758v3 [csLG].

[109] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing Medical Imaging Data for Machine Learning. Radiology 2020;295: 4–15.

[110] Yamoah GG, Cao L, Wu CW, Beekman FJ, Vandeghinste B, Mannheim JG, et al. Data Curation for Preclinical and Clinical Multimodal Imaging Studies. Mol Imaging Biol. 2019;21:1034–43.

[111] Aryanto KY, Oudkerk M, van Ooijen PM. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. Eur Radiol. 2015;25: 3685–95.

[112] van Herk M. Conquest DICOM software. http://ingenium.home.xs4all.nl/dicom.html. Accessed on January 26, 2021.

[113] RSNA. CTP-The RSNA Clinical Trial Processor. http://mircwiki.rsna.org/index.php?title=CTP-The_RSNA_Clinical_Trial_Processor. Accessed on January 26, 2021.

[114] Knopke A. K-Pacs. http://k-pacs.net/. Accessed on January 26, 2021.

[115] Library D. DICOM Library - Anonymize, Share, View DICOM files ONLINE. http://www.dicomlibrary.com/. Accessed on January 26, 2021.

[116] DicomWorks. DicomWorks - Free DICOM software. http://www.dicomworks.com/. Accessed on January 26, 2021.

[117] Publishing P. PixelMed Java DICOM Toolkit. http://www.pixelmed.com/. Accessed on January 26, 2021.

[118] DVTk. DVTk Project. http://www.dvtk.org/. Accessed on January 26, 2021.

[119] Yakami M. YAKAMI DICOM Tools. http://www.kuhp.kyoto-u.ac.jp/%7Ediag_rad/intro/tech/dicom_tools.html. Accessed on January 26, 2021.

[120] NIfTI. Neuroimaging Informatics Technology Initiative. https://nifti.nimh.nih.gov/. Accessed on January 26, 2021.

[121] Gatos I, Tsantis S, Spiliopoulos S, Karnabatidis D, Theotokas I, Zoumpoulis P, et al. Temporal stability assessment in shear wave elasticity images validated by deep learning neural network for chronic liver disease fibrosis stage assessment. Med Phys. 2019;46:2298–309.

[122] Kagadis GC, Drazinos P, Gatos I, Tsantis S, Papadimitroulas P, Spiliopoulos S, et al. Deep learning networks on chronic liver disease assessment with fine-tuning of shear wave elastography image sequences. Phys Med Biol. 2020;65:215027.

[123] Pinto Dos Santos D, Dietzel M, Baessler B. A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol. 2020.

[124] Yip SS, Aerts HJ. Applications and limitations of radiomics. Phys Med Biol. 2016; 61:R150–66.

[125] Zwanenburg A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. Eur J Nucl Med Mol Imaging. 2019;46:2638–55.

[126] Zwanenburg A, Lock S. Why validation of prognostic models matters? Radiother Oncol. 2018;127:370–3.

[127] Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present... any future? Eur J Nucl Med Mol Imaging. 2017;44:151–65.

[128] O'Connor JP, Aboagye EO, Adams JE, Aerts HJ, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14: 169–86.

[129] Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept. Radiother Oncol. 2016;121:459–67.

[130] Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol. 2010;49:1012–6.

[131] Reuze S, Orlhac F, Chargari C, Nioche C, Limkin E, Riet F, et al. Prediction of cervical cancer recurrence using textural features extracted from 18F-FDG PET images acquired with different scanners. Oncotarget. 2017;8:43169–79.

[132] Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. J Nucl Med. 2015; 56:1667–73.

[133] Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. J Nucl Med. 2018;59:1321–8.

[134] Aide N, Lasnon C, Veit-Haibach P, Sera T, Sattler B, Boellaard R. EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. Eur J Nucl Med Mol Imaging. 2017;44:17–31.

[135] Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging. 2015;42:328–54.

[136] Kaalep A, Sera T, Rijnsdorp S, Yaqub M, Talsma A, Lodge MA, et al. Feasibility of state of the art PET/CT systems performance harmonisation. Eur J Nucl Med Mol Imaging. 2018;45:1344–61.

[137] Pfaehler E, van Sluis J, Merema BBJ, van Ooijen P, Berendsen RCM, van Velden FHP, et al. Experimental Multicenter and Multivendor Evaluation of the Performance of PET Radiomic Features Using 3-Dimensionally Printed Phantom Inserts. J Nucl Med. 2020;61:469–76.

[138] Pinto Dos Santos D, Dietzel M, Baessler B. A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol. 2021;31:1–4.

[139] Tankyevych O, Tixier F, Antonorsi N, Filali Razzouki A, Mondon R, Pinto-Leite T, et al. Can alternative PET reconstruction schemes improve the prognostic value of radiomic features in non-small cell lung cancer? Methods 2020.

[140] Roboflow. You might be resizing your images incorrectly. https://blog.roboflow.com/you-might-be-resizing-your-images-incorrectly/. Accessed on November 28, 2020.

[141] Orlhac F, Humbert O, Boughdad S, Lasserre M, Soussan M, Nioche C, et al. Validation of a harmonization method to correct for SUV and radiomic features variability in multi-center studies. J Nucl Med. 2018;59:288.

[142] ADNI. Pet Acquisition. http://adni.loni.usc.edu/methods/pet-analysis-method/pet-analysis/. Accessed on November 28, 2020.

[143] Choe J, Lee SM, Do KH, Lee G, Lee JG, Lee SM, et al. Deep Learning-based Image Conversion of CT Reconstruction Kernels Improves Radiomics Reproducibility for Pulmonary Nodules or Masses. Radiology 2019;292:365–73.

[144] Hognon C, Tixier F, Gallinato O, Colin T, Visvikis D, Jaouen V. Standardization of Multicentric Image Datasets with Generative Adversarial Network. IEEE MIC2019.

[145] Modanwal G, Vellal A, Buda M, Mazurowski MA. MRI image harmonization using cycle-consistent generative adversarial network. SPIE Medical Imaging 20202020.

[146] Zhong J, Wang Y, Li J, Xue X, Liu S, Wang M, et al. Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development. Biomed Eng Online. 2020;19:4.

[147] Li Y, Han G, Wu X, Li Z, Zhao K, Zhang Z, et al. Normalization of multicenter CT radiomics by a generative adversarial network method. Phys Med Biol 2020.

[148] Desseroit MC, Tixier F, Weber WA, Siegel BA, Cheze Le Rest C, Visvikis D, et al. Reliability of PET/CT Shape and Heterogeneity Features in Functional and Morphologic Components of Non-Small Cell Lung Cancer Tumors: A Repeatability Analysis in a Prospective Multicenter Cohort. J Nucl Med. 2017;58:406–11.

[149] Desseroit MC, Visvikis D, Tixier F, Majdoub M, Perdrisot R, Guillevin R, et al. Development of a nomogram combining clinical staging with (18)F-FDG PET/CT image features in non-small-cell lung cancer stage I-III. Eur J Nucl Med Mol Imaging. 2016;43:1477–85.

[150] Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, et al. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. Transl Cancer Res 2016;5:349–63.

[151] Shafiq-Ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys. 2017;44:1050–62.

[152] Shafiq-Ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. Sci Rep. 2018;8: 10545.

[153] Da-Ano R, Visvikis D, Hatt M. Harmonization strategies for multicenter radiomics investigations. Phys Med Biol 2020.

[154] Zhovannik I, Bussink J, Traverso A, Shi Z, Kalendralis P, Wee L, et al. Learning from scanners: Bias reduction and feature correction in radiomics. Clin Transl Radiat Oncol. 2019;19:33–8.

[155] Andrearczyk V, Depeursinge A, Muller H. Neural network training for cross-protocol radiomic feature standardization in computed tomography. J Med Imaging (Bellingham). 2019;6:024008.

[156] Chatterjee A, Vallieres M, Dohan A. Creating robust predictive radiomic models for data from independent institutions using normalization. IEEE Trans Radiat Plasma. Med Sci. 2019;1.

[157] Sun C, Tian X, Liu Z, Li W, Li P, Chen J, et al. Radiomic analysis for pretreatment prediction of response to neoadjuvant chemotherapy in locally advanced cervical cancer: A multicentre study. EBioMedicine. 2019;46:160–9.

[158] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.

[159] Goh WWB, Wang W, Wong L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. Trends Biotechnol. 2017;35:498–507.

[160] Ligero M, Jordi-Ollero O, Bernatowicz K, Garcia-Ruiz A, Delgado-Munoz E, Leiva D, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur Radiol 2020.

[161] Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. Radiology 2019;291:53–9.

[162] Orlhac F, Lecler A, Savatovski J, Goya-Outi J, Nioche C, Charbonneau F, et al. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. Eur Radiol. 2020.

[163] Lucia F, Visvikis D, Vallieres M, Desseroit MC, Miranda O, Robin P, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2019;46:864–77.

[164] Dissaux G, Visvikis D, Da-Ano R, Pradier O, Chajon E, Barillot I, et al. Pretreatment (18)F-FDG PET/CT Radiomics Predict Local Recurrence in Patients Treated with Stereotactic Body Radiotherapy for Early-Stage Non-Small Cell Lung Cancer: A Multicentric Study. J Nucl Med. 2020;61:814–20.

[165] Whitney HM, Li H, Ji Y, Liu P, Giger ML. Harmonization of radiomic features of breast lesions across international DCE-MRI datasets. J Med Imaging (Bellingham). 2020;7:012707.

[166] Wu Q, Wang S, Li L, Wu Q, Qian W, Hu Y, et al. Radiomics Analysis of Computed Tomography helps predict poor prognostic outcome in COVID-19. Theranostics. 2020;10:7231–44.

[167] Garau N, Paganelli C, Summers P, Choi W, Alam S, Lu W, et al. External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis. Med Phys 2020.

[168] Da-Ano R, Masson I, Lucia F, Dore M, Robin P, Alfieri J, et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. Sci Rep. 2020;10:10248.

[169] Parekh VS, Jacobs MA. Radiomic synthesis using deep convolutional neural networks. arXiv:1810.11090 [cs.CV]. 2018.

[170] Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys. 2018;102:1143–58.

[171] KERAS. Deep Learning with KERAS Radiomics. https://github.com/decordoba/deep-learning-with-Keras-Radiomics) Accessed on Jan. 29, 2021.

[172] Abadi M, Barham P, Chen J, Chen U, Davis A, Dean J, et al. Tensorflow: a system for large-scale machine learning. 12th symposium on operating systems design and implementation2016. p. 265-83.

[173] Nioche C, Orlhac F, Boughdad S, Reuze S, Goya-Outi J, Robert C, et al. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. Cancer Res. 2018;78:4786–9.

[174] Szczypinski PM, Strzelecki M, Materka A, Klepaczko A. MaZda–a software package for image texture analysis. Comput Methods Programs Biomed. 2009;94:66–76.

[175] van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77:e104–7.