Radiomics

Introduction

As explained previously, there are globally two ways for establishing a diagnosis of diseases such as liver cancers: the biopsy, or the diagnostic imaging.

The biopsy, as detailed earlier, suffers from a lot of drawbacks. It remains an invasive procedure, with a high cost in terms of resources, and does not consider the tumor heterogeneity.

The diagnostic imaging on the other hand, is not invasive, provides information about the tumor shape, the growth over the time, and is less prone to bias due to tissue heterogeneity.

The recent improvements in the medical imaging field allow acquisition of data being more and more relevant, thus enabling a better estimation of the phenotypical characteristics of the patients.

In the case of brain imaging, the augmentation of contrast, on MR images, thanks to the injection of contrast agents such as gadolinium-based agents (method mentioned before) is an important technique for the evaluation of brain and liver tumors [1–3]. This tool allows a delineation of large tumors and an early detection of small metastatic lesions. The different MRI sequences (e.g. the T1 weighted sequences) also allow an internal separation of the tissue within the same tumor (active vs necrotic part of the tumor) [1].

Support brought by the innovations in the medical imaging field have been demonstrated on other organs such as the liver [4], the breast [5] or the colon [6] with a consequent benefit in terms of diagnosis.

However, even though the advancements in the medical imaging fields allowed those performances, the interpretation of the medical images remains subjective and not quantitative. In order to correctly provide a diagnosis that will not depend on the observer, one can extract and use the characteristics previously difficult, even impossible to distinguish with the naked eye.

Introduced in the 80s, CAD (*Computer Assisted Diagnosis*) tools were the first to implement this method, to establish a link between the imaging features and the biological characteristics of the patients [7].

In order to go along with those new systems, standard were introduced such as the one created by the WHO (*World Health Organization*) or the RECIST (*Response Evaluation Criteria in Solid Tumors*) [8], were the objective was to assess the evolution of the disease following the progression of the tumor size, but here again, those criteria suffer from a too high dependence with the observers.

The term *radiomics* was introduced in the early 2010s, allowing the computation of more features than

the traditional CADs (more than a thousand vs only a dozen previously) and bringing a more complete diagnosis, since CADs were often limited to distinguish benign vs malignant lesions [9].

This new technique allows some breakthroughs in various applications such as the cancer diagnosis, the detection of the tumors (with the identification of malignant lesions), their classification, the estimation of the patient survival, the prediction of the aggressivity of the tumors, their recurrence, or the advancement of the disease.

In the clinical practice, this new method also allows an improvement in the way biopsies are performed, with the identification of the areas where the extraction should be performed [10] or even by prediction when a biopsy is helpful or not [11].

Compared to above-mentioned criteria based on a naked-eye examination, we are now able to rely on a computer to analyse the gray-levels at a finer scale. Therefore, 2 approaches exist, the HCR (*Hand-Crafted Radiomics*) based on mathematical engineered features, relying on the textural and intensity based properties of the volume of interest, and the DLR (*Deep-Learning radiomics*), where the retained features will directly be computed from the input data without any prior knowledge.

Handcrafted Radiomics

In this section we will describe the HCR pipeline, by first exposing the different steps of the classical workflow, before analysing the different studies that used HCR on patients suffering from HCCs. We will conclude with the different improvements that should be brought to enhance the power of radiomics.

A conventional radiomics workflow (based on HCR features) starts with the acquisition and the reconstruction of medical images, followed by the segmentation of those images, which is a critical step since HCR features are extracted from the segmented sections, and many tissues do not have distinct boundaries [10]. Once the different areas segmented, the features are extracted and quantified, and a statistical analysis is performed to select only the most relevant one. The final step consists in building a model that will use the selected features to perform the wanted task, which is often either the tumor characterization or its prognosis. The pipeline is illustrated in the figure below 1.

We will now describe those different steps, before analyzing recent state-of-the-art HCR studies, before establishing a list of measures needed to be taken in order to improve the quality and the reproducibility of future radiomics works.

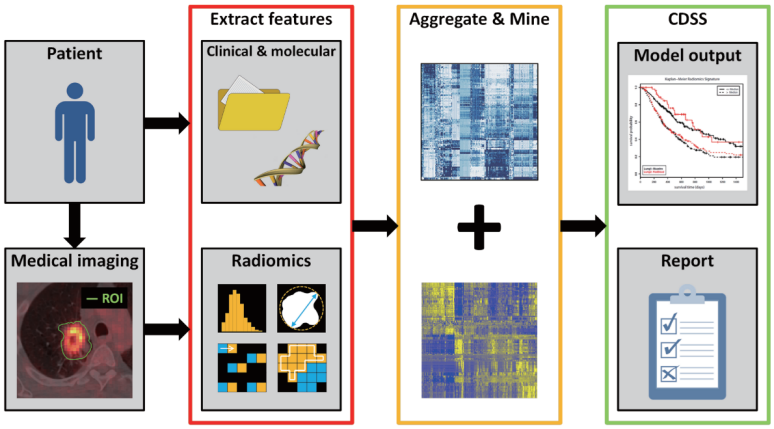

Figure 1 Overview of methodological processes in radiomics: data discovery, collection and preparation, model(s) development/validation and implementation, assessment of clinical utility.

Figure 1: Radiomics classical workflow as depicted by **©Scrivener et al.** [12]

HCR workflow

As explained previously, ultrasonography (US) is the recommended modality as primary imaging test for surveillance. If the surveillance is positive, CT or MR examinations are performed for the diagnosis and the staging of the disease. For the reasons exposed previously, namely the availability, and its robustness when compared to MRI, we will focus on HCR studies based on CT imaging data.

Without entering into the details of how a CT scan works, we can assume that performances of the CT imaging depend mainly on some settings such as the slice thickness, the capability for projecting the density variations into image intensities and the reconstruction algorithm which aims at converting tomographic measurements into cross-sectional images.

It has been demonstrated that radiomics features can differ between different scanners with the same settings [13]. It is also common to differentiate CT images into two categories, the screening where low dose images are used and the diagnosis with higher quality of contrast obtained with higher doses [14]. Worth mentioning that the majority of the studies that are being investigated belong to the second category.

Images are typically combined with other clinical sources when computing the radiomics features. Among them, gene expressions, clinical data such as the age, the gender or the past medical history, blood biomarkers or other prognostic markers such as the tumor size, the stage or the recurrence are the main non-imaging sources of data that are used in the radiomics workflow. However, they can be difficult to acquire, normalize and integrate in a radiomics pipeline, therefore, features are most commonly extracted from images only.

Segmentation

Historically, the segmentation was performed manually, hence, sensitive to the inter-observer variability, as depicted below in the figure 2 & 3.

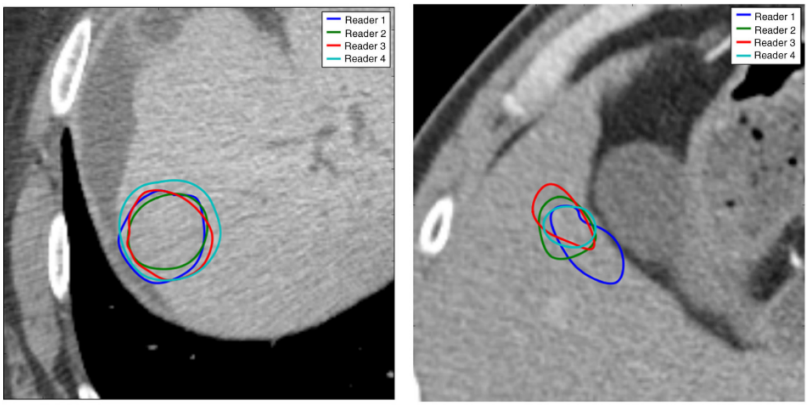

Fig. 2 Two samples of 14 slices selected for segmentation by all four readers. Each closed boundary represents a different reader's segmentation. (a) Relatively high overlap (54%) and (b) relatively low overlap (23%).

Figure 2: Inter-observer variability in the tumor segmentation, as reported by © **Echegaray et al.** [15]

To reduce this bias, semi-automatic segmentation techniques were developed. Those based on the intensity of the pixels suffer from the fact that the intensity of the pixels, in the case of abdominal organs, is often close from one organ to another. On the other hand, models based on statistical models often require the computation of an energy function, that can involve a large number of parameters, thus being difficult to compute and optimize.

(More details concerning the liver semantic segmentation techniques can be found in the **Semantic Segmentation chapter**).

Once the volume of interest delineated, different features can be extracted. On one hand, features can be chosen a priori for their capacity to translate the physiological behavior expected by the experts. For example, studies performed on the lungs showed a correlation between the textural homogeneity and the survival of the patients [17, 18], or the grade [19]. This knowledge can also be used in the case of brain tumors to assess the response to a treatment, by observing for example the vascular or cellular density [2, 20]. However, a prior knowledge is not always for the wanted task, therefore, the alternative is to extract a huge quantity of features and to determine the most relevant one by using for example some machine learning algorithms.

Features

Features can be regrouped into different categories depending on their statistical order (first, second or higher order). For the features belonging to the first statistical order, the volume of interest is

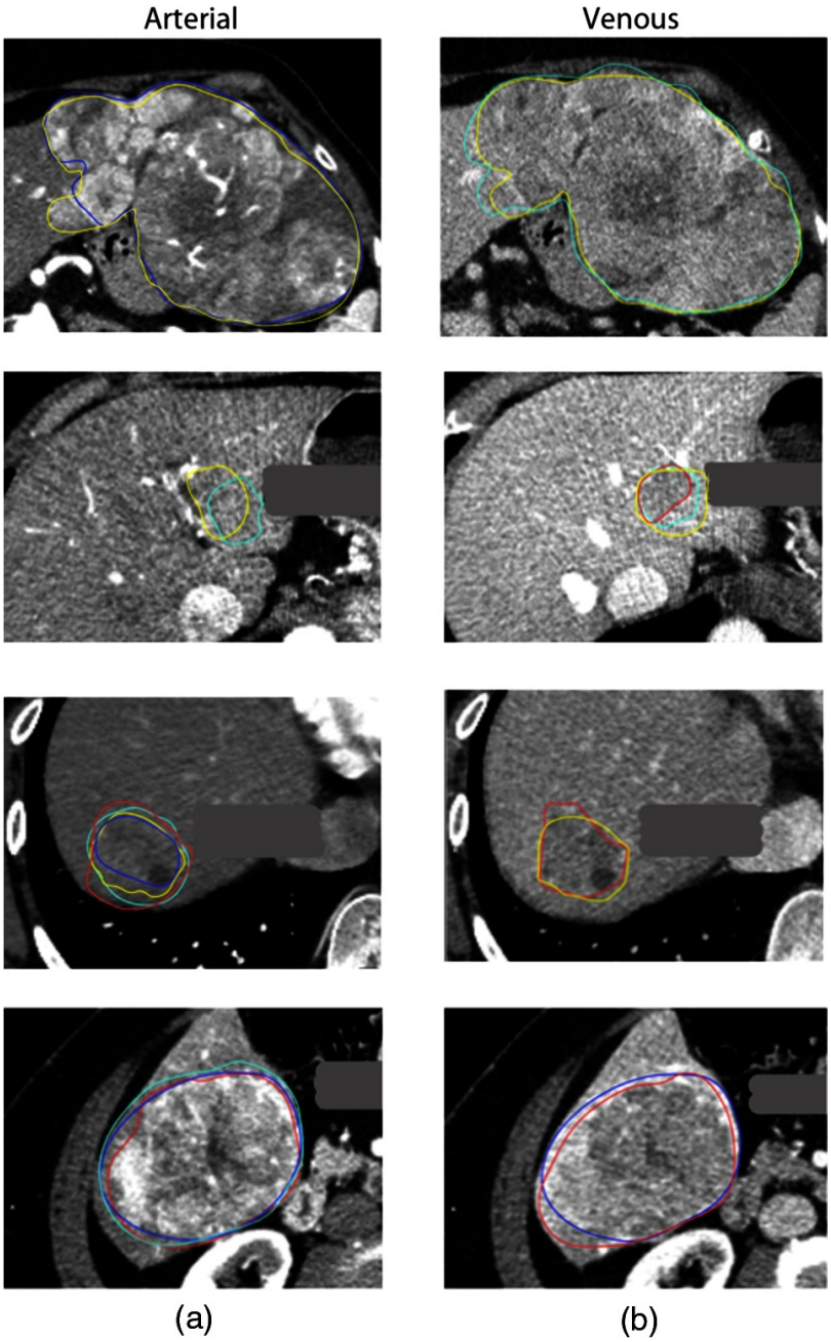

Figure 3: Inter-observer variability between 4 different observers on multiphasic images (a corresponding to arterial phase images and b to portal venous phase images), with a dice similarity coefficient from 0.19 to 0.93, **©Bakr et al.** [16]

transformed into an histogram, and different values are computed such as the uniformity or the entropy. Even though those features are often sensitive to the acquisition settings such as the slice thickness or even the way the histogram is computed, they permitted the prediction of the malignancy of breast lesions [21].

Shape features are often extracted directly from the VOI in order to analyze its geometrical properties

(such as the overall volume occupied by the tumor, its sphericity, its roughness or its fractal dimension). Those features also permitted the prediction of the response to treatment in previous studies [14].

Second order statistical features are meant to extract the textural properties of the volume, by considering the neighboring relationship between pixels. This will play an important role especially for the characterization of the heterogeneity of the tissues. This relation is captured by several descriptive matrices (GLCM, GLRLM, ...) [21].
Finally, higher order features allow the extraction of imaging features in various frequency domains (the Wavelet features are the most commonly used higher order features [14]).

Features Selection

As listed here, a large quantity of features can be extracted and they tend sometimes to be highly correlated, which in some cases can cause overfitting during the creation of a predictive model. A diminution phase of the number of features is often required, either in a supervised manner (features are selected for their discriminant power in regards with the wanted task) or in non-supervised manner (the main objective is to suppress the redundant features without considering the different labels) [21].

Among the supervised methods, one can distinguish the univariate ones, where the features are tested one by one depending on their contribution to the wanted task (Wilcoxon or Fisher test), from the multivariate where the features are regrouped into subsets before being tested against the output class.

Unsupervised methods (such as the PCA: *principal component analysis*) are less prone to overfitting since they do not consider the label of the data, but their main goal is to reduce the dimensionality of the features space.

Once the number of features is reduced, the next step is to construct a predictive model, by using either clustering methods (patients are regrouped based on a metric depending on the retained features) or classification ones (where models such as RF: *random forests* or SVM: *support vector machines* are trained from the selected features in order to predict the wanted clinical criteria). Concerning the prediction of the survival, it is common to implement slightly different models such as the *Kaplan-Meier* or the *Cox Proportional Hazard* [22].

In the radiomics studies, one of the main goals is the stability of the features against the pre-treatment steps described above. In order to reach this objective, it is possible for the patients to undergo the medical imaging examinations several times (*test-retest*), and the segmentations can be performed by several experts or even by the same expert several times [23].

In summary, when dealing with classical radiomics pipelines, reaching the best results will often depend

on the best combinations between the extraction of the features, the technique used to reduce the number of features and the method implemented to create the model.

Every modification on the cited steps can have a huge impact on the predictable performances of the created model.

In the next section, we will analyze the different HCR studies performed on patients suffering from HCC and who underwent CT examination. We will first describe the different choices made in each step of the classical pipeline, before presenting ways to improve the quality of future HCR work.

HCR applied to the liver

In order to analyze the different methods implemented in the HCR field, 15 studies performed on patients suffering from HCC and who underwent CT scan examination were reviewed. Initially, 23 primary liver cancer-related studies have been scanned in our review [24], we then selected the 15 HCC-related ones.

We will first describe the different targets of the studies and the details of the cohorts through the number of patients and the clinical criteria that preceded their selection.

We will then compare the different imaging acquisition protocols, and the way the regions/volumes of interest are delineated. Finally, we will analyze the different features that appeared to be relevant in the studies, before proposing some tracks to improve the reproducibility and the performances of future radiomics work.

Details concerning the experimental settings of the studies, the endpoints and the different endpoints can be found in the table 1.

Author	Modality & slice thickness	Mean tumor size	Treatment	#Patients & Inclusion Criteria	Segmentation	Computed features	Retained features	Retained features category	Study endpoints	Results	%RQS (total points)
Cozzi et al. [25]	NECT 3mm	-	Radiotherapy (volumetric modulated arc therapy)	**138** Patients BCLC stages from A to C, Child-Pugh stages A-B	Segmentation done using the CTV (clinical target volume) which is manually contoured (whole tumor analysis)	**35** extracted features 6 geometry and histogram 29 GLM	Compacity Energy GLNU	Quantitative	OS & local control of the tumor	AUC of the model is 0.80 Survival could be predicted with a radiomics signature	14 (5)
Zhou et al. [33]	Contrast CT (30 and 60s) 1.25mm	-	Hepatectomy	**215** Patients who underwent partial hepatectomy	Largest cross-sectional area of the tumor, manual delineation Exclusion of necrosis 2 experts	**300** features (Mean, SD, Kurtosis, Skewness, GLM)	Histogram features (skewness, energy, means, ...)	Quantitative	Recurrence	First-order statistical features combined with clinical factors can predict early recurrence	25 (9)
Akai et al. [26]	Contrast CT (27-28, 40 and 90s) 5mm	3.7cm (2.4-7.0cm)	Hepatectomy	**127** patients	Manually setting the ROI to include the tumor within the slice at its max diameter. Single radiologist	**96** features (mean, sd, positive calue pixels, entropy, kurtosis, skewness)	Entropy, skewness and kurtosis	Quantitative	OS & DFS	First-order statistical features were sufficient to predict postoperative survival	25 (9)
Chen et al. [27]	Contrast CT (25 and 70s) 1.25mm	-	Hepatectomy	**61** patients with only one lesion and survival above 3 months	ROI was delineated around the tumor outline at the longest dimension 2 experts	**84** features 12 Gabor 9 Wavelet 7 GLCM	Textural features, Gabor and Wavelet as key features	Quantitative	OS & DFS	Tumor prognosis could be predicted using Gabor and Wavelet responses	17 (6)
Li et al. [28]	Contrast CT (70s) 1.25mm	8.0cm (5.1-18.7cm)	Hepatectomy or TACE	**130** patients 86 treated by LR and 22 by TACE	Irregular ROI manually drawn around the largest-cross sectional tumor outline 2 radiologists	**27** features (Wavelet)	2 Wavelet features correlated with survival	Quantitative	OS and Treatment sensitivity	Wavelet features correlated with survival suggesting a suitable treatment choice	19 (7)

Author	Modality & slice thickness	Mean tumor size	Treatment	#Patients & Inclusion Criteria	Segmentation	Computed features	Retained features	Retained features category	Study endpoints	Results	%RQS (total points)
Raman et al. [38]	Contrast CT (25s) 3mm	Adenoma 7 ± 3cm FNH 6 ± 3cm HCC 8 ± 3cm	-	**80** patients 17 FNH 19 Adenomas 25 HCCs 19 normal livers	ROIs were selected from multiple axial slices (from 5 to 10 slices) 2 experts	**32** features (mean, SD, entropy, skewness, kurtosis)	SD and Mean of histogram	Quantitative	Diagnosis	First-order statistical features able to differenciate 3 types of hypervascular lesions with a 15% of error rate)	3 (1)
Kuo et al. [34]	Contrast CT (30-35 and 60-70s) 2.5mm	-	-	**30** Patients no patients received chemo before resection	No segmentation Images analyzed visually by 2 experts.	**6** imaging traits (internal arteries, textural heterogeneity, wash-in-wash-out, necrosis, tumor margin)	Tumor margin Internal arteries	Semantic	MVI status	The tumor margin showed strong correlation with MVI, TNM. Internal arteries showed correlation with MVI	19 (7)
Banerjee et al. [29]	Contrast CT (30-35, 60-70, 180-300s) 2.5-3mm	2.8cm (1.8-4.5cm)	Hepatectomy or LT[1]	**157** patients 72 resection 85 LT[1] MVI diagnosed in 45 patients	Only imaging features were evaluated by 5 radiologists	**3** imaging traits (internal arteries, hypodense halo, tumor-liver difference)	The 3 imaging traits were retained	Semantic	OS and RFS	Combination of the three different imaging traits was correlated with MVI	53 (19)
Renzulli et al. [35]	Contrast CT (25-30, 45-60, 180-300s) 2.5mm	3.3cm (1.8-5.2cm)	Hepatectomy	**125** patients where hepatic resection was indicated	Only imaging features were evaluated by 2 radiologists	**5** imaging traits Dimensions Lesions number Non-smooth margins TTPVI [2] Peritumoral enhancement	All except the lesion dimensions	Semantic	MVI status	The 4 retained traits were correlated with the presence of MVI in HCC	8 (3)

[1]Liver transplantation

[2]Two-Trait Predictor of Venous Invasion: Internal arteries and Hypoattenuating halo

Author	Modality & slice thickness	Mean tumor size	Treatment	#Patients & Inclusion Criteria	Segmentation	Computed features	Retained features	Retained features category	Study endpoints	Results	%RQS (total points)
Segal et al. [30]	Contrast CT (three-phasic)	-	Hepatectomy	Train: **30** Test: **32**	No segmentation, visual examination 2 radiologists.	**32** imaging traits (Capsule, Wash-in-Wash-out, Tumor-Liver difference, . . .)	Internal arteries and hypodense halo	Semantic	OS & MVI	Internal arteries combined with hypodense halo can predict OS, MVI Both are correlated with HCC gene expression	42 (15)
Zheng et al. [31]	Contrast CT (22 and 60s) 5mm	-	Hepatectomy	Train: **212** Test: **107** patients without anticancer therapy	ROI delineated around the tumor outline of the largest cross-sectinal area. 2 radiologists	**110** GLM features	6 GLCM features	Quantitative	Recurrence & OS	Textural features sufficient to predict postoperative recurrence and survival	47 (17)
Peng et al. [36]	Contrast CT (30, 60 and 120s) 5mm	4.9-6.4cm	-	Train: **184** Test: **120** Partial hepatectomy with pathologically confirmed HCC	ROI semi-automatically segmented in the largest cross-sectional area 2 experts	**5** imaging traits (tumor margin, peritumoral enhancement, hypoattenuating halo, internal arteries, tumor-liver difference) & **980** quantitative features	Nonsmooth tumor margins, hypoattenuating halos and internal arteries + 8 radiomics features (Entropy, shape, GLRLM, GLCM)	Semantic & Quantitative	MVI status	Radiological features and a radiomics signature computed with first-order statistical features showed correlation with MVI	47 (17)
Bakr et al. [16]	Contrast CT (AR with bolus tracking, PV, delay) Thickness $\leq$ 3mm	7.4cm	-	**28** patients with surgical resection of a previously untreated HCC	3 ROIs were placed on different cross sections of the tumors (one central, one superior and one inferior) 4 radiologists	**464** features (intensity, texture, shape)	Textural features	Quantitative	MVI status	Textural features computed using single- or combined-phased images were correlated with MVI	3 (1)

Author	Modality & slice thickness	Mean tumor size	Treatment	#Patients & Inclusion Criteria	Segmentation	Computed features	Retained features	Retained features category	Study endpoints	Results	%RQS (total points)
Taouli et al. [37]	Contrast CT (AR with bolus tracking, PV at 70s, Delay at 180s)	5.7±3.2cm	-	**38** patients **26** CT/ 12 MRI 36 Liver resection 2 LT[3]	Global inspection of the imaging traits and "slice-wise" evaluation for the enhancement ratio and the wash-out ratio 2 radiologists	**11** imaging traits (wash-in-washout, hypovascularity) **3** Quantitative features (contrast ratios)	Infiltrative pattern, mosaic appearance, presence of MVI, large size	Semantic & Quantitative	Signature of MVI & aggressive phenotype	Correlation was found between some imaging traits and the aggressive profile of the tumors	19 (7)
Xia et al. [32]	Contrast CT (30, 55 70, 300s) 2.5-5mm	12 tumors smaller than 5cm, 26 larger	Hepatectomy or LT	**38** patients	Tumor was firstly delineated then divided into 3 spatially distinct subregions (using a multiparametric clustering) 1 radiologist	**37** features (1st order, geometry, textural) And **4** features for the whole tumor	Volume of transition region & cluster prominence	Quantitative	OS	The volume of transition between tumor and liver, and the heterogeneity of the lesion were correlated with survival.	22 (8)

Table 1: HCR reviewed studies details

[3]Liver Transplantation

Experimental setup

The vast majority of the studies were designed to predict the survival of the patients after surgery, or any other type of treatment [25–32]. In clinical trials, the traditional way to evaluate the survival is through the OS (Overall Survival), which corresponds to the duration from either the date of diagnosis of the disease or the start of its treatment, to either the end of the trial or the death of the patient. Being often assimilated to the survival rate, new finer metrics tend to be preferred such as the DFS (Disease Free Survival), which corresponds to the duration from the beginning of the treatment to the date of the recurrence of the disease.

Other ways to evaluate the response to a given treatment were also evaluated in some of the reviewed studies, such as the presence or absence of recurrence [31, 33], the local control which assess the end of the growth of a tumor [25], and other ways to compute the sensitivity to a treatment [28, 34].

Another important aspect that is often assessed by the reviewed studies is the physiological changes brought by the disease, such as the aggressive profile of the tumor, usually translated by the presence of MVI (*MicroVascular Invasion*), and its association to genes expression [16, 29, 30, 34–37].

The entire 15 studies had a retrospective design, and the number of selected patients varied from 28 [16] to 319 [31], with a median of 125 patients per study, as illustrated in the figure 4.

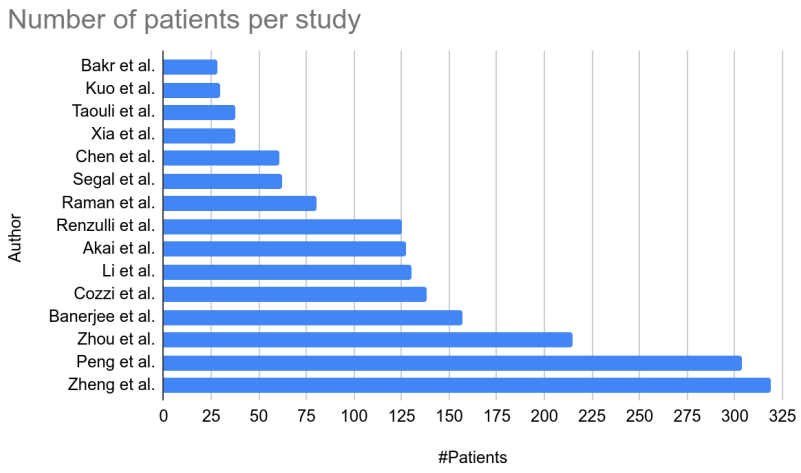

Figure 4: Number of patients included in the reviewed HCR studies

The relatively low number of patients can be explained by the strong inclusion criteria set within the studies (size and number of lesions, baseline imaging examination within a given period of time before initial treatment, . . .)

They were all built with data acquired on patients who underwent CT examination, except one, that decided to mix data obtained from both CT and MR examinations [37].

Regarding the CT scan protocol, only one study decided to use images acquired before the injection of contrast medium [25], whereas two other studies used images acquired at only one phase (early arterial phase for [38], and portal venous phase for [28]), and the remaining studies analyzed multiphase images. Among them, four studies utilized images acquired at both early arterial and portal venous phases [27, 31, 33, 34], while the rest were based on traditional triphasic images [16, 26, 29, 30, 32, 35–37]. Concerning the acquisition protocols, early arterial phase images are most of the time acquired around 30s after the injection of contrast agent (between 22 and 35s), and some studies used bolus tracking method to estimate the best acquisition moment instead of using the same timing for all the patients. Portal venous phase images are acquired between 45 and 70s after the injection, and the delayed images obtained during an even larger interval (between 90 and 300s after the injection).
Knowing that images are the key elements in the computation process of the radiomics features, this high variation within the acquisition protocols is the first reason why the standard HCR pipeline should be standardized.

Once the images acquired, the following step consists in selecting the region of interest to compute the features, or evaluate the physiological properties of the lesions to the naked eye.

ROI selection

The selection of the region/volume of interest and/or the assessment of the physiological characteristics of the tumor is often performed by one or multiple experienced radiologists.
This step of the pipeline was performed by only a single expert in rare cases in the reviewed studies [26,32], whereas it was usually performed by two experts [27, 28, 30, 31, 33–38].
When more than 1 expert is involved, the authors decided to implement ways for quantifying the inter and intra-observer variability, such as the ICC (*Intraclass Correlations Coefficient*) [28, 31], or the Cohen k statistics [29, 35].

Regarding now the selection of the ROI, we can separate the reviewed studies by the type of features being extracted.

On one hand, when using imaging traits, it is common to use the whole tumor to perform the evaluation. For example when evaluating the vascular invasion, some features like the peritumoral enhancement need to be estimated globally [29, 30, 34, 35].
On the other hand, when using computational features, the analysis can be performed on the whole tumor or on a single slice. Some studies decided to compute the features on the entire 3D ROI, as an example Cozzi et al. used the entire tumor independently on the tissues present within it, Xia et al. decided to separate areas based on their textural properties as depicted in the figure 5, but the analysis

remains global. Other studies placed everal ROIs all throughout the tumor (3 ROIs for Bakr et al. and from 5 to 10 for Raman et al.). However, the majority of the studies decided to place the ROI at the largest cross-sectional area [26–28, 31, 33].

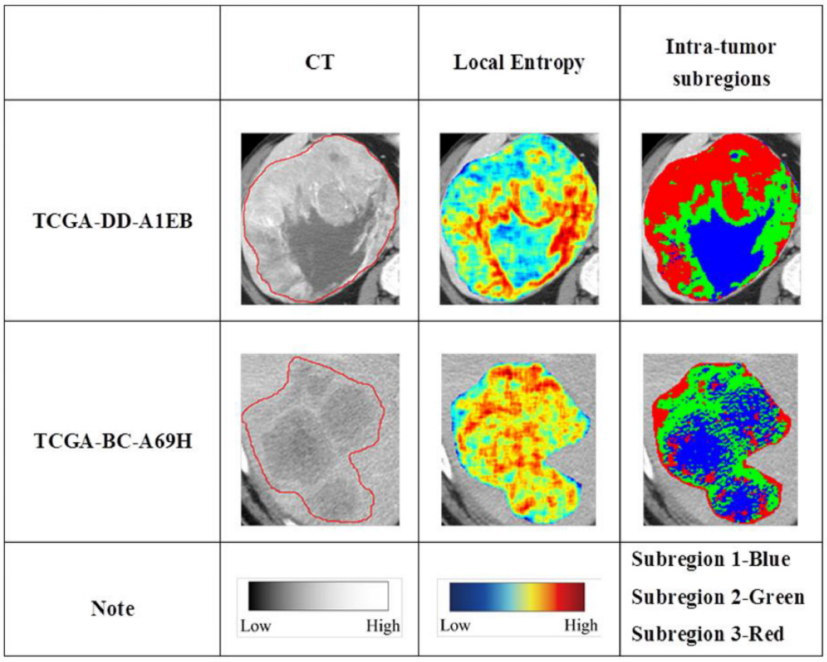

Figure 5: Illustration of intratumor partition for two representative patients (TCGA-DD-A1EB and TCGA-BC-A69H). The first column shows the tumor outlines on the original CT image. The second column shows the heatmaps of calculated local entropy on tumor images. The third column shows the three subregions marked with different colors after intratumor partitioning. **©Xia et al.** [32]

Note that some studies decided to compute the features using both the entire volume and the slice-wise approach (e.g. Taouli et al. evaluated imaging traits globally and computed the ratio using a slice-wise fashion, Peng et al. did the same in their study by computing features using an ROI placed at the largest-cross sectional area, and evaluating imaging traits globally).

Worth mentioning that before the computation of features, it is common to filter the images using different kernel sizes, in order to enhance different elements of the volumes such as the blood vessels for example. All reviewed studies that computed quantitative features filtered their images with a Laplacian of Gaussian algorithm with various sizes, as depicted in the figure 6.

Features selection

Once the ROI delineated and pre-processed, the following step consists in extracting the features, and as explained previously, the choice on which features to extract depends on the type of features the study is going to rely on, quantitative or semantic.

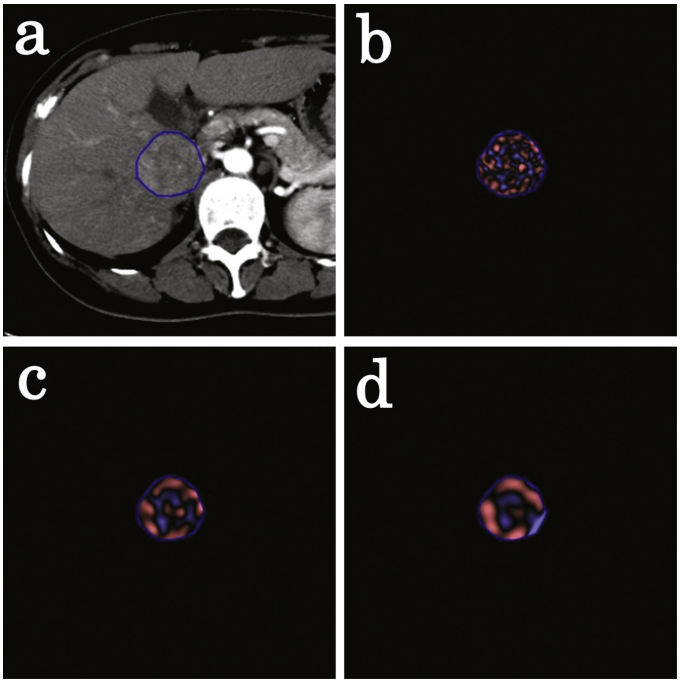

Figure 6: Screenshot of the CT texture analysis software. A polygonal ROI was drawn on the tumor (a). Processed images using Laplacian of Gaussian filters with SSF of 2 mm (b), 4 mm (c), and 6 mm (d) were automatically generated. The images were displayed using a red or blue scale showing negative or positive pixels, respectively. **©Akai et al.** [26]

In the case of quantitative features, the reviewed studies often decided to focus on a single category of features (first-order, textural features, higher-order...), thus obtaining a relative small number of features (27 for Li et al., 32 for Raman et al.). However, despite choosing a specific category of features, this number can increase, when combining the native features with the spatial filter, and the different contrast enhanced phases, as in Akai et al. where a total 96 features are extracted from the initial 6 histogram-based features [26].

Some other studies decided to extract the maximum possible features by combining the previously mentioned group of features, and thus obtained 300+ features [16, 33, 36]. The problem in this case, often called "*curse of dimensionality*", corresponds to a high number of features relative to the number of individuals, and that can cause some troubles when further training the predictable model.

When imaging traits are preferred to classical radiomics features, the number of extracted characteristics is generally below 10, with a high predominance of changes brought by the hepatocarcinogenesis, such as the presence of internal arteries, or the wash-in wash-out effect (example of imaging traits can be observed in the figure 7 & 8). Even though the number of extracted features is often small, they often correspond to absence or presence of physiological properties that are sometimes difficult to quantify and that will highly depend on the observer's experience.

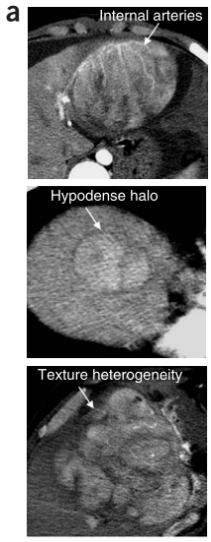

Figure 7: From top to bottom: Internal arteries, Hypodense halo, Textural heterogeneity, as illustrated by **©Segal et al.** [30]

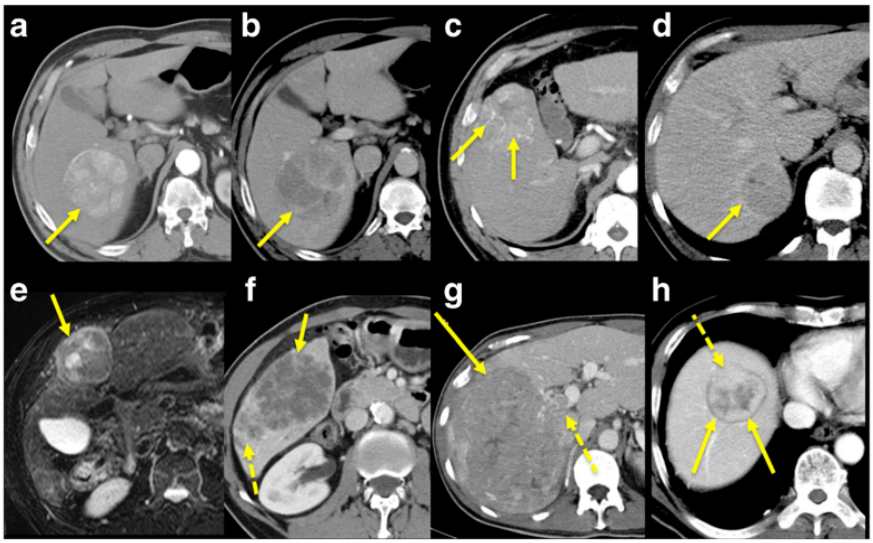

Figure 8: **©Taouli et al.** [37] (a,b) wash-in/wash-out pattern and mosaic appearance, without capsule/ pseudocapsule. (c) internal arteries (arrows) (d) pseudo-capsule (arrow) (e) [MR image] hyperintense encapsulated with mosaic appearance (arrow) (f) internal necrosis and satellite lesions posteriorly (dashed arrow) (g) right portal vein invasion (dashed arrow) (h) extra-nodular growth anteriorly (dashed arrow)

After the different features are obtained, the next step in the pipeline consists in the selection of the features and the building of the predictive model.

The vast majority of the reviewed studies decided to implement a logistical regression model in order to assess the correlation between features and the study endpoint. Among the existing methods,

time-related approaches such as the Cox regression model [25, 28, 29, 31, 32] and the Kaplan-Meier survival analysis [26, 27, 30, 32] are the most commonly used, especially because those are historical ways to predict the survival of the patients.

Different statistical approaches were used by the studies that tried to determine the link between selected features and study endpoints such as the LASSO (*Least Absolute Shrinkage and Selection Operator*) algorithm [16, 33, 36]. In the other studies, other approaches were implemented, for example, Raman et al. performed a PCA followed by a MANOVA (*Multivariate Analysis Of VAriances*) to create clusters among patients for the classification of hypervascular lesions, whereas Renzulli et al. evaluated the positive and negative predictive values of the selected features against the microvascular invasion status of the patients.

The list of discriminant features per study is given in the tables 2 & 3 , with a separation between quantitative and semantic features.

Table 2: List of quantitative features used in the reviewed studies

First Order Statistics

- Shape [25]
- Skewness [26, 28, 33]
- Kurtosis [26]
- Mean [25, 33, 38]
- Energy [25, 33]
- Entropy [26, 36]
- Peak [16]
- Standard deviation [32]
- Enhancement ratio [37]
- Tumor-Liver difference [37]

Second Order Statistics

- Gray Level matrices [25, 31, 36]
- Cluster prominence [32]

Higher Order Statistics

- Wavelets [16, 27, 28]
- Gabor [16, 27]

Morphological features

- Tumor margin volume [32]
- Tumor size[1] [35, 37]

[1] Tumor size in the two studies is not given as the exact volume, but rather a classification of tumors in different categories (for example the categories were smaller than 2cm, between 2 and 5 cm, and larger than 5 cm in Renzulli et al. [35])

Table 3: List of semantic features used in the reviewed studies

Two Traits Predictor of Venous Invasion

- Internal arteries [29, 30, 34–37]
- Hypoattenuating halos [29, 30, 35, 36]

Intensity-related features

- Peritumoral enhancement [35]
- Presence of a Tumor-Liver difference [29]

Textural-related features

- Non-smooth Tumor Margins [34–36]
- Infiltrative patterns [37]
- Mosaic appearance [37]

Concerning the studies based on quantitative features we can notice that first-order statistical features

is the most common discriminant type, which is normal because this group of histogram-based characteristics is often implemented in the existing radiomics tools, whereas higher order statistical features require more advanced knowledge to be implemented, and often lack of interpretability.

Even though Li et al. decided to extract only Wavelet features because they consider that the current way of computing textural features is too dependent on the imaging acquisition settings, first and second-order statistical features remain a good indicator of the textural heterogeneity which is often correlated with the physiological advances of the disease [28].

Worth also noting that imaging traits, often analyzed either alone, or in combination with quantitative features, remain discriminant enough in a lot of studies, because they are directly linked to the physiological changes produced by the disease. Despite needing expertise to be extracted, such as in Segal et al. where 32 different imaging traits were analyzed, their predictable power drives us to consider them in future radiomics studies and focus on a way to quantify them [30].

Study reproducibility

Although the majority of the reviewed studies obtained good predictable results in regards with the wanted prediction task, their stability to the experimental settings and their reproducibility remain questionable.

In 2017, Lambin et al. who remains one of the founders of the radiomics fields [39] published a study proposing a way to rethink the HCR pipeline and assess the robustness of future radiomics studies [40]. This assessment is performed thanks to the RQS (*Radiomics Quality Score*), which evaluates a total of 16 components with various weights. The evaluation of the different components allows the computation of a score ranging from 0 to 36 points where highest weights are given to criteria allowing a better reproducibility of the study, such as the prospective aspect of the study (7 points over 36) or the presence of a validation step in the proposed workflow (5 points over 36, with a penalty of 5 points when no validation at all is present).

Our reviewed studies were evaluated in regards with the RQS, in consensus with a medical research fellow (Wakabayashi Taïga) [24].

The details of the different scores per study are reported in the table 4.

Table 4: RQS score details per criteria for the reviewed studies

Criteria	Bakr et al. [16]	Kuo et al. [34]	Taouli et al. [37]	Xia et al. [32]	Chen et al. [27]	Segal et al. [30]	Raman et al. [38]	Renzulli et al. [35]	Akai et al. [26]	Li et al. [28]	Cozzi et al. [25]	Banerjee et al. [29]	Zhou et al. [33]	Peng et al. [36]	Zheng et al. [31]
Image Protocol quality	1	2	2	2	2	1	2	2	2	2	1	2	2	2	2
Multiple segmentations	1	0	0	0	1	1	1	0	0	1	0	1	1	1	1
Phantom Studies	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Multiple time points	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Features reduction	-3	3	3	3	3	3	-3	3	3	3	3	3	3	3	3
Mutlivariate analysis	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1
Biological correlates	0	1	1	1	0	1	0	0	0	0	0	1	0	0	0
Cut-off analysis	0	0	0	0	1	1	1	0	1	1	1	0	1	1	1
Discrimination statistics	2	1	0	0	1	1	1	1	2	1	1	1	1	1	1
Calibration Statistics	0	0	1	2	0	1	0	0	1	1	1	1	0	2	1
Prospective study	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Validation	-5	-5	-5	-5	-5	3	-5	-5	-5	-5	-5	5	-5	2	2
Gold standard comparison	2	2	2	2	0	0	2	0	2	0	2	2	2	2	2
Clinical utility	2	2	2	2	2	2	2	2	2	2	0	2	2	2	2
Cost-effectiveness analysis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Open science Data	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Total	**1**	**7**	**7**	**8**	**6**	**15**	**1**	**3**	**9**	**7**	**5**	**19**	**9**	**17**	**17**

The mean obtained RQS by the reviewed studies was 8.73 +/- 5.57 points, which corresponds to less than 23% of the maximal possible score, and only one study obtained a more than 50% of the maximum points, which translates the lack of robustness of the current HCR state-of-the-art studies applied to the HCC.

The figure 9 allows us to see the most respected criteria in regards with the RQS guidelines.

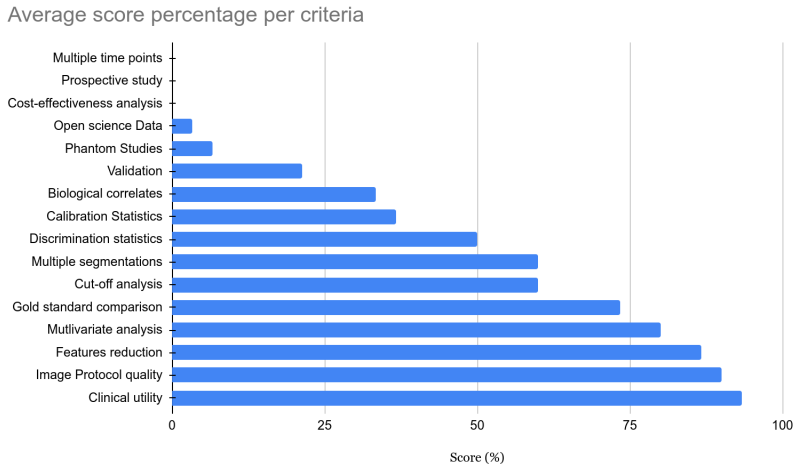

Figure 9: Average score percentage per criteria in the RQS for the reviewed studies

The criteria the less respected by the different reviewed studies were all those related with the reproducibility and the robustness. Among them, the analysis of images acquired at multiple times, the implementation of phantom studies to detect inter-scanner differences and features sensitive to those settings, or even the prospective design of the studies that will ensure inclusion of patients undergoing the same protocols were almost always neglected. Even though the different reviewed studies obtained successful results in regards with the wanted task, their robustness relative to the experimental settings and their reproducibility can be questioned, especially when analyzing their results obtained on the radiomics quality scoring system [40].

One way to allow a better reproducibility for the future radiomics studies is to fulfil the maximum possible criteria introduced by the RQS standard, and to use standard ways to compute the quantitative features, thanks to open-source libraries such as pyradiomics [41]. Another way is to reduce as much as possible the impact of human-based interpretation in the HCR workflow. This can be done by replacing the hand-crafted annotations, and the engineered process of features computation and selection by a standardized, automated and data-driven pipeline, similar to what is performed in more recent DLR studies.

Deep Learning Radiomics

In this section we will first present the differences between the HCR and the DLR strategies, before describing in detail the DLR concept. We will then present the different reviewed studies tackling liver-related problems using a DLR approach. We will outline the different steps of their pipelines such as the use of multiphasic images, the way they incorporated experts' annotations and their choice regarding the deep network architectures.

Difference between HCR and DLR

Being a young field, radiomics starts to mature when applied to several organs such as the lungs or the breast, but it still struggles when applied to the liver, especially because of the scarcity of available data and the complexity of its anatomy.

As exposed previously, the first studies targeting liver cancer almost always rely on hand crafted features. The main limitation observed is the lack of reproducibility originating from engineered features, that sometimes result from complex processes which can be difficult to imitate, and often fail to work on different databases. HCR rely on manual or semi-automatic expert annotations, which often require a complex and time-consuming process, that also provide several constraints, such as the poor inter-expert reproducibility. The extracted features are not necessarily relevant to encode the observed structure, it is needed to increase their number, hence requiring a complex dimensionality reduction step. Another limitation is the difficulty to find the perfect association between features extraction, selection and the statistical analysis used to reach the wanted target.

Those limitations, and the rise of new computational resources associated with the emergence of deep learning techniques allowed the development of a new branch in the radiomics field, called DLR (Deep-Learning Radiomics) [9].

DLR workflow

In this branch of radiomics, the features are extracted through a deep-learning process, avoiding a manual extraction.

A neural network can be trained to generate the most relevant features. These features can either be kept in the network for the final pathological target prediction, or used as input in a different model (such as a SVM or a RF). Compared to the HCR, no prior knowledge is required, and the features can be extracted in an end-to-end manner, using only the raw images as input, and without

necessarily providing any segmentation. It has also been demonstrated that performances of those networks increase with the size of the training dataset [42]. Eliminating the segmentation phase when evaluating the diagnosis, allows to reduce the workload of experts, and provides a solution to the observer-dependency. When training DLR networks, the original image can be combined with the segmentation or any other pre-processed step result such as the gradient image for example [43] to improve the relevance of the extracted features.

Generally, DLR studies can be classified depending on the type of input used, the training strategy or the type of architecture chosen to extract the features.

As input, deep radiomics networks can consider 2D slices independently, however this technique does not bring sufficient information since the decision mainly depends on the entire volume of interest. The different outputs obtained in a slice-wise fashion can be fused to get a volume-wise decision. The volume by itself can also directly be used as input, however it can raise several issues such as the size of the voxels or the slice thickness. Finally, the classification can be performed by considering a series of volumes corresponding to the entire examination of the patient [44], but here again the question regarding the normalization of the input can be raised.

Once the type of input is chosen, the studies differ depending on the training strategy. The networks can be trained *from scratch* using only the available data or a pre-existing architecture can be utilized. In the first case, the obtained network will be specific to the wanted task, but this specialization can also lead to troubles such as overfitting or the sensitivity to imbalance data. The impact of those problematics can be limited with the help of data augmentation (use of existing data to generate new artificial samples) [45], multi-task training (where the number of parameters is limited by the training of several task using the same network) or the incorporation of the proportion of each class present in the data when building the cost function [46]. The other strategy consists in using an architecture pre-trained most of the time on natural images, and then re-train only a specific part on the wanted task [47–49]. It is worth noting that this type of training constrains the pipeline to be slice-wise since existing architectures are often pre-trained on 2D images.

Finally, the features can be extracted using either a supervised or an unsupervised approach. In the supervised case, the most commonly used networks are based on convolutional layers (CNN), followed by one or multiple dense layers to predict the output class. While the network is trained to perform the classification, the features are extracted either after a fully connected layer [49], or after one of the convolutional layers [50].

Other variants that are also built with convolutional layers as key components can also be implemented

(RNN: *Recurrent Neural Networks*, LSTM: *Long Short Term Memory* or *Capsule Net*). Their goal is to get rid of the limitations caused by the input format that need to be fixed, and by the difficulty to consider an entire 3D volume during the training [51]. In the unsupervised case, the objective is to let the extracted features be responsible for the data distribution, so that new cases can be created following this distribution. The most commonly used architecture in this case is the *auto-encoder*, made up with a part that contracts the information (encoder), in order that the most useful one is conserved to regenerate the original data (decoder). Auto-encoder can be built on top of convolutional layers [47], or trained with the aim of being insensitive to the noise added to the input data [43, 52]. Following the same principles which are to reconstruct the original input data using only the most relevant features, some studies implemented DBN (*Deep Belief Networks*) [43] or Deep Boltzmann Machines [53].

Some studies are referred to as hybrid, when features are combined with other sources of data (combination of different modalities [22] or association with clinical data such as genomic data [54], or when only a part of the pipeline implements deep learning methods, either for the extraction of the features [49] or when the decision is taken with a fusion between HCR and DLR features [48].

DLR applied to the liver

Even though the number of studies targeting the liver is increasing, the vast majority of them can be categorized as HCR, and only a few are currently based on deep learning. The main reason behind that is the late emergence of deep learning and the recent outbreak of new architectures and concepts that are often first developed and evaluated in other fields than the medical imaging one (e.g. Residual Network, DenseNet, Capsule Net).

However, we have decided to review the first DLR-liver related studies, in order to understand how DL architectures can successfully be incorporated in a radiomics pipeline. We describe the study endpoints, the data preprocessing, the implementation and training strategies, before analyzing their performances. A more detailed analyze of the reviewed studies is available here.

Study Endpoint

Within the reviewed DLR studies, the majority of them are targeting a diagnosis, either the classification of FLLs (*Focal Liver Lesions*) [55–58] or the estimation of the fibrosis stage [59], when two of them focused on the response to treatments, either for recurrence after resection [60] or the response after TACE (*TransArterial ChemoEmbolization*) [61].

We selected the studies using CT images to perform their analysis, and realized that the vast majority of them used multiphasic images to perform their research, knowing that the evolution of contrast medium

is often correlated with the pathological features of the liver as mentioned in the **Medical Context**. The only one that used single phase images, was the one targeting an estimation of the fibrosis stage, and their method was based on portal phase images only [59]. Regarding the multiphasic studies, there is no consensus about the delay between the injection of the contrast agent and the acquisition of the different phases. They tend to prefer triphasic acquisition, with images acquired before the injection of contrast agent, at early arterial phase and a third phase, either portal venous [56, 58, 60] or a delayed one [55, 57]. Peng et al. decided to get rid of the NECT (*Non-Enhanced CT*) phase, but still chose a triphasic acquisition (AR, PV, DELAY).

Image processing pipeline

Concerning the image processing pipeline, the data used to train the deep networks are most commonly selected via placement of a bounding box by the experts around the hepatic lesion, before a registration between the different phases in case of multiphasic acquisition to counter the effects of body motion and/or breathing.

The manual placement of the *ROIs* is usually done by one or more experts on the raw image [55–57,59,60], and only one study decided to perform an automatic segmentation of both the parenchyma and the lesion with the application of a random-walker algorithm, before being checked by experts [58]. When the method is based on a 2D approach, selected images are often those presenting the maximal cross-sectional proportion of the lesions, except one study targeting the estimation of the fibrosis stage, that centered the ROI so it displayed the ventral aspect of the liver [59]. Only one study incorporated 3D information in their pipeline, but they also started the placement of the VOI with the slice presenting the maximal proportion of tumor, and extended it to adjacent slices [62].

After placing a bounding box, the images are registered either manually [55,56,60] or via the application a non rigid registration with anatomical constraints [58]. No real registration was mentioned for two studies [59,61] but Yamada et al. evaluated the effects of the registration in the prediction performances of their networks (as depicted in the figure 10), and after training several networks with misregistered data (shifted, rotated, skewed) they concluded that in the vast majority of the cases, no statistical significance can be found between the performances of the networks trained with manually registered data and those of the networks trained with misregistered data [55].

Worth noting that some studies performed their analysis on cropped or resized JPEG images [57,59,60]. This choice might cause loss of data, since JPEG images are encoded with 256 values in each one of the RGB channel whereas the raw data is used elsewhere with the precise HU intensities.

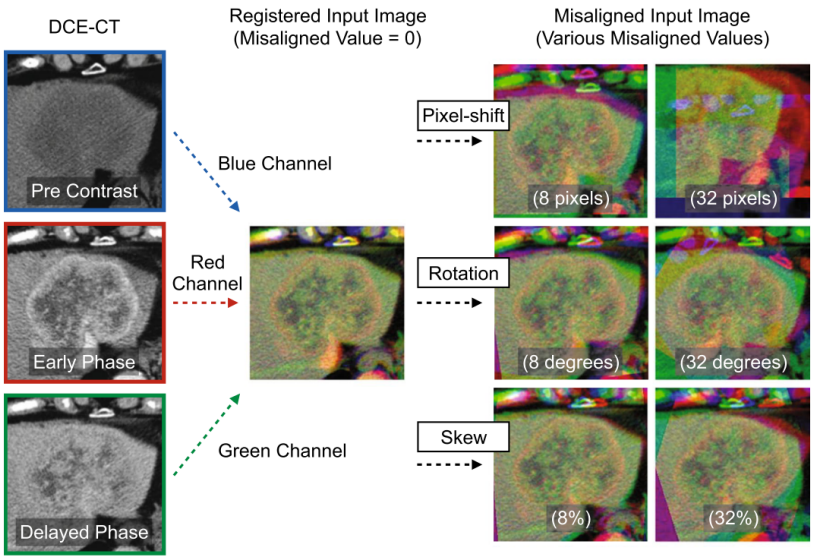

Figure 10: Illustration by **©Yamada et el.** of the manually registered images and the effect of transformations (shift, rotation, skew) in the three phases [55]

Training strategies

Finally, in the reviewed studies, only two of them combined images with clinical data [59, 60], whereas the other only used image data, which is understandable because clinical data are often difficult to retrieve, and can also be difficult to integrate in a deep network, as illustrated in the figure 11.

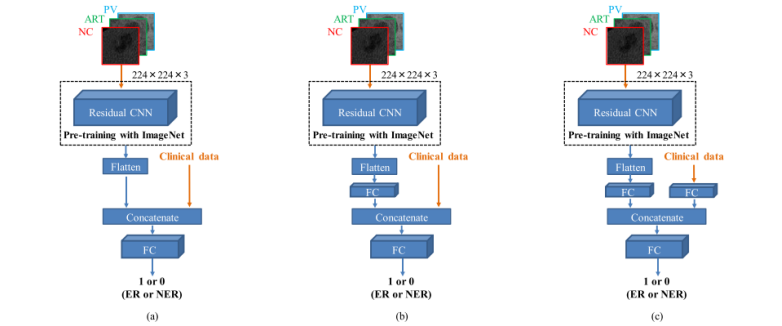

Figure 2. Four deep learning-based combined models with clinical data and CT image data. (a) Model A, (b) model B and (c) model C

Figure 11: Illustration given by **©Wang et al.** of the different tested deep networks where raw images are combined with the clinical data [60]

The reviewed studies differ mainly in the way they built their deep architecture.

In most of the cases, convolutional layers are used for the extraction of the most discriminant features. The main question being whether to use a pre-trained network or to train a network from scratch. In the case of fine-tuning, the most commonly used architectures are the AlexNet and the ResNet [55, 56, 60, 61], but it is also usual to compare the results obtain by different pre-trained architectures [55, 60]. The

general method is to recycle an architecture trained on a huge dataset such as ImageNet, freeze the weights of the early layers (responsible for the high levels features), adjust and train the last layers on the current database to be more specific. An illustration of this process can be found in the figure 12.

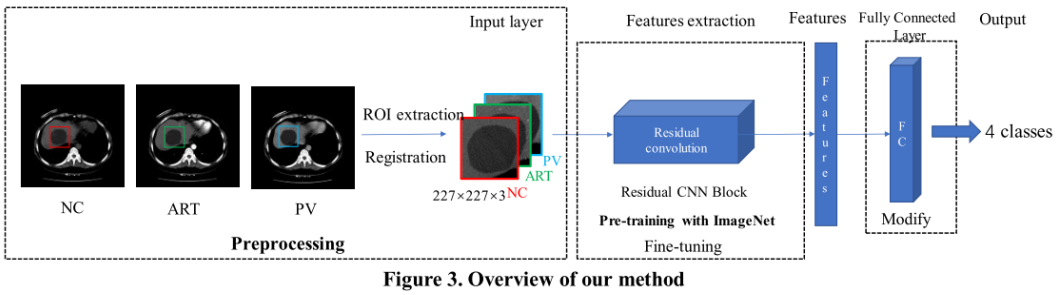

Figure 12: Illustration of the pre-training strategy adopted by **©Yamada et al.** [55]

The rest of the reviewed studies created a custom architecture and trained it from scratch [57–59].

Two of them used classical convolutional layers followed by max pooling layers, early in the network to extract relevant features, as depicted in the figure 13.

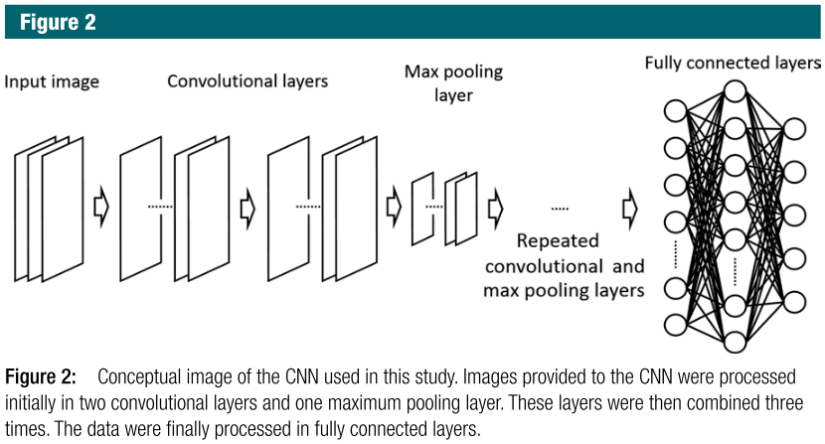

Figure 13: Architecture used by **©Yasaka et al.** where the features are first extracted through convolutional layers before the prediction is performed using fully connected layers. [57]

Since multiphasic studies often stacked the different phases as channels to feed the deep network, one study decided to extract the temporal information through a different paradigm by using LSTM layers [58]. As depicted in the figure 14, their architecture first extracted the features using what they called a "ResGLBlock" per phase, with two scaled data as input (a large one with the entire lesion, and a smaller one with finer details), and conserved the temporal information via bidirectional LSTM layers.

When using a pre-trained architecture, the input size is often dictated by the native architecture (224x224x3 for example for the ResNet architecture [60, 61]), and the different studies need to resample

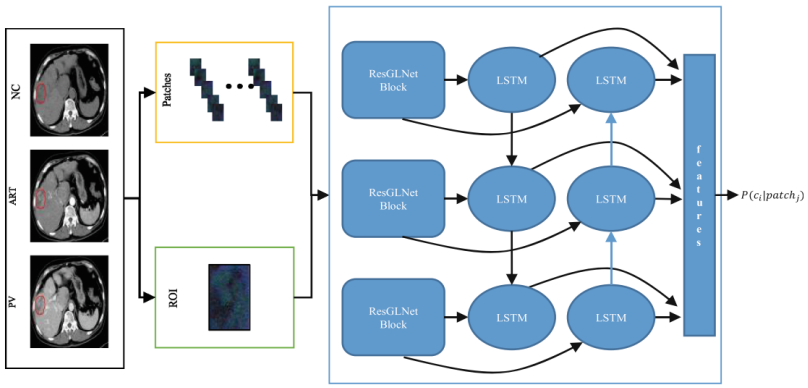

Fig. 1. The flowchart of our framework

Figure 14: Architecture used by **©Liang et al.** where patches of different scales are extracted from the original images, before being used to train residual networks connected with bidirectional LSTM layers [58]

their inputs to fulfill those requirements, which can sometimes affect the performances of the network, whereas custom architectures allow a usage of custom sizes [57–59]. However, the size of the lesions or other extracted ROI is often different from one patient to the other, therefore, this problem is still open.

One way to render the deep networks robust to those changes is through data augmentation. Yasaka et al. for example trained the network with patches cropped at different resolutions from the initial lesion ROI after application of standard geometrical transformations such as rotation or shift [57]. The same process of extracting different patches from an initial ROI is performed by two other studies, where the goal is also to balance the different classes [59, 61].

The networks are then usually trained in a cross validation fashion [55, 57, 59, 60], or validated on external dataset [61] to be less affected by the effect of randomness, and to be less prone to overfitting.

Performances

Regarding their performances on their testing sets, the different studies concluded first that fine tuning allows an improvement in the accuracy of the DLR network, when compared to training from scratch (e.g. Wang et al. reported an improvement from 83.7 to 91.2% regarding the classification accuracy of their model when using a pre-trained network) [55, 56]. Several studies also demonstrated that multiphase images increase the performances or the DLR networks, when compared to single phase input only [57]. Instead of training the DLR networks only with images, it is possible to combine them with clinical data which can be difficult to collect, and challenging to integrate in a deep architecture, but are proven to improve the accuracy of the networks in some cases [60].

Reported results showed moderate to good accuracy for the wanted tasks. For example, the reported

mean accuracy is above 0.90 for the studies targeting a classification between Focal Nodular Hyperplasias, Cysts, HCC and Hemangioma (0.91 in both [56, 58]), and it slightly drops to 0.84 when more complex categories are integrated (iCC, combined HCC and difference between HCC and early HCCs) [59]. Another study performed the classification between HCCs and non HCCs, by additionally incorporating the differentiation stages for the HCC group, but still had comparable performances than experienced radiologists in the diagnostic performances [55].

The study targeting the estimation of the fibrosis stage reported results less accurate than those obtained using elastography data (MRE: *Magnetic resonance elastography* or TE: *Transient elastography*), but they were the first to perform this analysis on CT images, and their results could be improved with the inclusion of volumetric information, and other sources of data. [59].

Finally, the studies predicting a response to a treatment reported a high accuracy with an AUC of 0.82 when predicting the recurrence after TACE [60], and accuracy above 0.83 in the two external validation sets when estimating the response of TACE in HCC [61].

Those results still can be improved, especially by increasing the size of the cohort, or by replacing the manual placement of the bounding boxes with an automatic segmentation method in order to reduce the dependency to single or multiple experts [59, 61].

As a conclusion, the different reviewed studies tend to agree on the fact that a multiphase analysis is necessary to precisely describe and encode the pathological features of the disease. For the rest of the pipeline, no real consensus exists but several strategies are implemented especially to compensate for the small size of the databases. Worth also noting that the reviewed studies correspond to the first DLR liver-related applications, therefore, they tend to tackle the less complex problems such as the FLLs classification. With future improvements regarding DL applied to the medical imaging field, and with more publicly available data, the next DLR liver-related studies will be ready to tackle more complex challenges.

References

[1] Antonios Drevelegas. Imaging of brain tumors with histological correlations. *Imaging of Brain Tumors with Histological Correlations*, pages 1–432, 2011.

[2] Mu Zhou, Lawrence Hall, Dmitry Goldgof, Robin Russo, Yoganand Balagurunathan, Robert Gillies, and Robert Gatenby. Radiologically Defined Ecological Dynamics and Clinical Outcomes in Glioblastoma Multiforme: Preliminary Results. *Translational Oncology*, 7(1):5–13, 2014.

[3] Yee Liang Thian, Angela M. Riddell, and Dow Mu Koh. Liver-specific agents for contrast-enhanced MRI: Role in oncological imaging. *Cancer Imaging*, 13(4):567–579, 2013.

[4] Fergus Davnall, Connie S P Yip, Gunnar Ljungqvist, Mariyah Selmi, Francesca Ng, Bal Sanghera, Balaji Ganeshan, Kenneth A. Miles, Gary J. Cook, and Vicky Goh. Assessment of tumor heterogeneity: An emerging imaging tool for clinical practice? *Insights into Imaging*, 3(6):573–589, 2012.

[5] Bas B Koolen, Marie-jeanne T F D Vrancken Peeters, and Tjeerd S Aukema. 18F-FDG PET / CT as a staging procedure in primary stage II and III breast cancer : comparison with conventional imaging techniques. pages 117–126, 2012.

[6] Dushyant V. Sahani, Mohammad Ali Bajwa, Yasir Andrabi, Surabhi Bajpai, and James C. Cusack. Current status of imaging and emerging techniques to evaluate liver metastases from colorectal carcinoma. *Annals of Surgery*, 259(5):861–872, 2014.

[7] Kunio Doi. Computer-aided diagnosis in medical imaging : Historical review , current status and future potential. 31:198–211, 2007.

[8] C. Carl Jaffe. Measures of response: RECIST, WHO, and new alternatives. *Journal of Clinical Oncology*, 24(20):3245–3251, 2006.

[9] Parnian Afshar, Arash Mohammadi, Konstantinos N. Plataniotis, Anastasia Oikonomou, and Habib Benali. From Hand-Crafted to Deep Learning-based Cancer Radiomics: Challenges and Opportunities. 2018.

[10] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2):563–577, 2016.

[11] Ying Liu, Jongphil Kim, Yoganand Balagurunathan, Qian Li, Alberto L. Garcia, Olya Stringfield, Zhaoxiang Ye, and Robert J. Gillies. Radiomic Features Are Associated With EGFR Mutation Status in Lung Adenocarcinomas. *Clinical Lung Cancer*, 17(5):441–448.e6, sep 2016.

[12] Madeleine Scrivener, Evelyn E. C. de Jong, Janna E. van Timmeren, Thierry Pieters, Benoît Ghaye, and Xavier Geets. Radiomics applied to lung cancer: a review. *Translational Cancer Research*, 5(4):398–409, 2016.

[13] Roberto Berenguer, María Del Rosario Pastor-Juan, Jesús Canales-Vázquez, Miguel Castro-García, María Victoria Villas, Francisco Mansilla Legorburo, and Sebastià Sabater. Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters. *Radiology*, 288(2):407–415, 2018.

[14] Rajat Thawani, Michael McLane, Niha Beig, Soumya Ghose, Prateek Prasanna, Vamsidhar Velcheti, and Anant Madabhushi. Radiomics and radiogenomics in lung cancer: A review for the clinician. *Lung Cancer*, 115(June 2017):34–41, 2018.

[15] Sebastian Echegaray, Olivier Gevaert, Rajesh Shah, Aya Kamaya, John Louie, Nishita Kothary, and Sandy Napel. Core samples for radiomics features that are insensitive to tumor segmentation: method and pilot study using CT images of hepatocellular carcinoma. *Journal of Medical Imaging*, 2(4):041011, 2015.

[16] Shaimaa Bakr, Sebastian Echegaray, Rajesh Shah, Aya Kamaya, and John Louie. Noninvasive radiomics signature based on quantitative analysis of computed tomography images as a surrogate for microvascular invasion in hepatocellular carcinoma: a pilot study. *Journal of Medical Imaging*, 4(04):1, 2017.

[17] Balaji Ganeshan, Elleny Panayiotou, Kate Burnand, Sabina Dizdarevic, and Ken Miles. Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: A potential marker of survival. *European Radiology*, 22(4):796–802, 2012.

[18] Yoganand Balagurunathan, Yuhua Gu, Hua Wang, Virendra Kumar, Olya Grove, Sam Hawkins, Jongphil Kim, Dmitry B. Goldgof, Lawrence O. Hall, Robert A. Gatenby, and Robert J. Gillies. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. *Translational Oncology*, 7(1):72–87, 2014.

[19] Omar S. Al-Kadi and D. Watson. Texture analysis of aggressive and nonaggressive lung tumor CE CT images. *IEEE Transactions on Biomedical Engineering*, 55(7):1822–1830, 2008.

[20] Yue Cao, Christina I. Tsien, Vijaya Nagesh, Larry Junck, Randall Ten Haken, Brian D. Ross, Thomas L. Chenevert, and Theodore S. Lawrence. Clinical investigation survival prediction in high-grade gliomas by MRI perfusion before and during early stage of RT. *International Journal of Radiation Oncology Biology Physics*, 64(3):876–885, 2006.

[21] Vishwa Parekh and Michael A. Jacobs. Radiomics: a new application from established techniques. *Expert Review of Precision Medicine and Drug Development*, 1(2):207–226, 2016.

[22] Anastasia Oikonomou, Farzad Khalvati, Pascal N. Tyrrell, Masoom A. Haider, Usman Tarique, Laura Jimenez-Juan, Michael C. Tjong, Ian Poon, Armin Eilaghi, Lisa Ehrlich, and Patrick Cheung. Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy. *Scientific Reports*, 8(1):1–11, 2018.

[23] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, nov 2017.

[24] Taiga Wakabayashi, Farid Ouhmich, Cristians Gonzalez-Cabrera, Emanuele Felli, Antonio Saviano, Vincent Agnus, Peter Savadjiev, Thomas F. Baumert, Patrick Pessaux, Jacques Marescaux, and Benoit Gallix. Radiomics in hepatocellular carcinoma: a quantitative review, sep 2019.

[25] Luca Cozzi, Nicola Dinapoli, Antonella Fogliata, Wei-Chung Hsu, Giacomo Reggiori, Francesca Lobefalo, Margarita Kirienko, Martina Sollini, Davide Franceschini, Tiziana Comito, Ciro Franzese, Marta Scorsetti, and Po-Ming Wang. Radiomics based analysis to predict local control and survival in hepatocellular carcinoma patients treated with volumetric modulated arc therapy. *BMC Cancer*, 17(1):829, 2017.

[26] H. Akai, K. Yasaka, A. Kunimatsu, M. Nojima, T. Kokudo, N. Kokudo, K. Hasegawa, O. Abe, K. Ohtomo, and S. Kiryu. Predicting prognosis of resected hepatocellular carcinoma by radiomics analysis with random survival forest. *Diagnostic and Interventional Imaging*, 99(10):643–651, 2018.

[27] Shuting Chen, Yanjie Zhu, Zaiyi Liu, and Changhong Liang. Texture analysis of baseline multiphasic hepatic computed tomography images for the prognosis of single hepatocellular carcinoma after hepatectomy: A retrospective pilot study. *European Journal of Radiology*, 90:198–204, 2017.

[28] Meng Li, Sirui Fu, Yanjie Zhu, Zaiyi Liu, Shuting Chen, Ligong Lu, and Changhong Liang. Computed tomography texture analysis to facilitate therapeutic decision making in hepatocellular carcinoma. *Oncotarget*, 7(11):13248–13259, 2016.

[29] Sudeep Banerjee, David S. Wang, Hyun J. Kim, Claude B. Sirlin, Michael G. Chan, Ronald L. Korn, Aaron M. Rutman, Surachate Siripongsakun, David Lu, Galym Imanbayev, and Michael D. Kuo. A computed tomography radiogenomic biomarker predicts microvascular invasion and clinical outcomes in hepatocellular carcinoma. *Hepatology*, 62(3):792–800, 2015.

[30] Eran Segal, Claude B. Sirlin, Clara Ooi, Adam S. Adler, Jeremy Gollub, Xin Chen, Bryan K. Chan, George R. Matcuk, Christopher T. Barry, Howard Y. Chang, and Michael D. Kuo. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nature Biotechnology*, 25(6):675–680, 2007.

[31] Bo Hao Zheng, Long Zi Liu, Zhi Zhi Zhang, Jie Yi Shi, Liang Qing Dong, Ling Yu Tian, Zhen Bin Ding, Yuan Ji, Sheng Xiang Rao, Jian Zhou, Jia Fan, Xiao Ying Wang, and Qiang Gao. Radiomics score: a potential prognostic imaging feature for postoperative survival of solitary HCC patients. *BMC cancer*, 18(1):1148, 2018.

[32] Wei Xia, Ying Chen, Rui Zhang, Zhuangzhi Yan, Xiaobo Zhou, Bo Zhang, and Xin Gao. Radiogenomics of hepatocellular carcinoma: multiregion analysis-based identification of prognostic imaging biomarkers by integrating gene data—a preliminary study. *Physics in Medicine & Biology*, 63(3):035044, feb 2018.

[33] Ying Zhou, Lan He, Yanqi Huang, Shuting Chen, Penqi Wu, Weitao Ye, Zaiyi Liu, and Changhong Liang. CT-based radiomics signature: a potential biomarker for preoperative prediction of early recurrence in hepatocellular carcinoma. *Abdominal Radiology*, 42(6):1695–1704, jun 2017.

[34] Michael D. Kuo, Jeremy Gollub, Claude B. Sirlin, Clara Ooi, and Xin Chen. Radiogenomic Analysis to Identify Imaging Phenotypes Associated with Drug Response Gene Expression Programs in Hepatocellular Carcinoma. *Journal of Vascular and Interventional Radiology*, 18(7):821–830, 2007.

[35] Matteo Renzulli, Stefano Brocchi, Alessandro Cucchetti, Federico Mazzotti, Cristina Mosconi, Camilla Sportoletti, Giovanni Brandi, Antonio Daniele Pinna, and Rita Golfieri. Can Current Preoperative Imaging Be Used to Detect Microvascular Invasion of Hepatocellular Carcinoma? *Radiology*, 279(2):432–442, may 2016.

[36] Jie Peng, Jing Zhang, Qifan Zhang, Yikai Xu, Jie Zhou, and Li Liu. A radiomics nomogram for preoperative prediction of microvascular invasion risk in hepatitis b virus-related hepatocellular carcinoma. *Diagnostic and Interventional Radiology*, 24(3):121–127, 2018.

[37] Bachir Taouli, Yujin Hoshida, Suguru Kakite, Xintong Chen, Poh Seng Tan, Xiaochen Sun, Shingo Kihira, Kensuke Kojima, Sara Toffanin, M. Isabel Fiel, Hadassa Hirschfield, Mathilde Wagner, and Josep M. Llovet. Imaging-based surrogate markers of transcriptome subclasses and signatures in hepatocellular carcinoma: preliminary results. *European Radiology*, 27(11):4472–4481, 2017.

[38] Siva P. Raman, James L. Schroeder, Peng Huang, Yifei Chen, Stephanie F. Coquia, Satomi Kawamoto, and Elliot K. Fishman. Preliminary Data Using Computed Tomography Texture

Analysis for the Classification of Hypervascular Liver Lesions. *Journal of Computer Assisted Tomography*, 39(3):1, 2015.

[39] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G P M Van Stiphout, Patrick Granton, Catharina M L Zegers, Robert Gillies, Ronald Boellard, André Dekker, and Hugo J W L Aerts. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48(4):441–446, 2012.

[40] Philippe Lambin, Ralph T.H. Leijenaar, Timo M. Deist, Jurgen Peerlings, Evelyn E.C. De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben T.H.M. Larue, Aniek J.G. Even, Arthur Jochems, Yvonka Van Wijk, Henry Woodruff, Johan Van Soest, Tim Lustberg, Erik Roelofs, Wouter Van Elmpt, Andre Dekker, Felix M. Mottaghy, Joachim E. Wildberger, and Sean Walsh. Radiomics: The bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, 14(12):749–762, 2017.

[41] Joost J.M. Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, 2017.

[42] Jie Zhi Cheng, Dong Ni, Yi Hong Chou, Jing Qin, Chui Mei Tiu, Yeun Chung Chang, Chiun Sheng Huang, Dinggang Shen, and Chung Ming Chen. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Scientific Reports*, 6(March):1–13, 2016.

[43] Wenqing Sun, Bin Zheng, and Wei Qian. Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis. *Computers in Biology and Medicine*, 89:530–539, 2017.

[44] Wei Shen, Mu Zhou, Feng Yang, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. Learning from Experts: Developing Transferable Deep Features for Patient-Level Lung Cancer Prediction. volume 8150, pages 124–131. 2016.

[45] Devinder Kumar, Audrey G. Chung, Mohammad J. Shaifee, Farzad Khalvati, Masoom A. Haider, and Alexander Wong. Discovery Radiomics for Pathologically-Proven Computed Tomography Lung Cancer Prediction. volume 1, pages 54–62. 2017.

[46] Amir Jamaludin, Timor Kadir, and Andrew Zisserman. SpineNet: Automatically pinpointing classification evidence in spinal MRIs. *Lecture Notes in Computer Science (including subseries*

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9901 LNCS:166–175, 2016.

[47] Oier Echaniz and Manuel Graña. Ongoing Work on Deep Learning for Lung Cancer Prediction. volume 10338 of *Lecture Notes in Computer Science*, pages 42–48. Springer International Publishing, Cham, 2017.

[48] Benjamin Q. Huynh, Hui Li, and Maryellen L. Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, 2016.

[49] Rahul Paul, Samuel H Hawkins, Yoganand Balagurunathan, Matthew B Schabath, Robert J Gillies, Lawrence O Hall, and Dmitry B Goldgof. Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma. *Tomography*, 2(4), 2016.

[50] Zeju Li, Yuanyuan Wang, Jinhua Yu, Yi Guo, and Wei Cao. Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Scientific Reports*, 7(1):1–11, 2017.

[51] Shekoofeh Azizi, Sharareh Bayat, Pingkun Yan, Amir Tahmasebi, Jin Tae Kwak, Sheng Xu, Baris Turkbey, Peter Choyke, Peter Pinto, Bradford Wood, Parvin Mousavi, and Purang Abolmaesumi. Deep recurrent neural networks for prostate cancer detection: Analysis of temporal enhanced ultrasound. *IEEE Transactions on Medical Imaging*, 37(12):2695–2703, 2018.

[52] Bum Chae Kim, Yu Sub Sung, and Heung Il Suk. Deep feature learning for pulmonary nodule classification in a lung CT. *4th International Winter Conference on Brain-Computer Interface, BCI 2016*, pages 2–4, 2016.

[53] Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, 101(1):569–582, nov 2014.

[54] Nastaran Emaminejad, Wei Qian, Yubao Guan, Maxine Tan, Yuchen Qiu, Hong Liu, and Bin Zheng. Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients. *IEEE Transactions on Biomedical Engineering*, 63(5):1034–1043, 2016.

[55] Akira Yamada, Kazuki Oyama, Sachie Fujita, Eriko Yoshizawa, Fumihito Ichinohe, Daisuke Komatsu, and Yasunari Fujinaga. Dynamic contrast-enhanced computed tomography diagnosis

of primary liver cancers using transfer learning of pretrained convolutional neural networks: Is registration of multiphasic images necessary? *International Journal of Computer Assisted Radiology and Surgery*, 14(8):1295–1301, 2019.

[56] Weibin Wang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, Qingqing Chen, Dong Liang, Lanfen Lin, Hongjie Hu, and Qiaowei Zhang. Classification of Focal Liver Lesions Using Deep Learning with Fine-Tuning. In *Proceedings of the 2018 International Conference on Digital Medicine and Image Processing - DMIP '18*, pages 56–60, New York, New York, USA, 2018. ACM Press.

[57] Koichiro Yasaka, Hiroyuki Akai, Osamu Abe, and Shigeru Kiryu. Deep learning with CNN showed high diagnostic performance in differentiation of liver masses at dynamic CT. *Radiology*, 286(3—March):887–896, 2018.

[58] Dong Liang, Lanfen Lin, Hongjie Hu, Qiaowei Zhang, Qingqing Chen, Yutaro Lwamoto, Xianhua Han, and Yen-Wei Chen. Combining Convolutional and Recurrent Neural Networks for Classification of Focal Liver Lesions in Multi-phase CT Images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11071 LNCS, pages 666–675. 2018.

[59] Koichiro Yasaka, Hiroyuki Akai, Akira Kunimatsu, Osamu Abe, and Shigeru Kiryu. Deep learning for staging liver fibrosis on CT: a pilot study. *European Radiology*, 28(11):4578–4585, 2018.

[60] Weibin WANG, Qingqing CHEN, Yutaro IWAMOTO, Xianhua HAN, Qiaowei ZHANG, Hongjie HU, Lanfen LIN, and Yen-Wei CHEN. Deep Learning-Based Radiomics Models for Early Recurrence Prediction of Hepatocellular Carcinoma with Multi-phase CT Images and Clinical Data. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4881–4884. IEEE, jul 2019.

[61] Jie Peng, Shuai Kang, Zhengyuan Ning, Hangxia Deng, Jingxian Shen, Yikai Xu, Jing Zhang, Wei Zhao, Xinling Li, Wuxing Gong, Jinhua Huang, and Li Liu. Residual convolutional neural network for predicting response of transarterial chemoembolization in hepatocellular carcinoma from CT imaging. *European Radiology*, 30(1):413–424, 2020.

[62] Da-wei Yang, Xi-bin Jia, Yu-jie Xiao, Xiao-pei Wang, Zhen-chang Wang, and Zheng-han Yang. Noninvasive Evaluation of the Pathologic Grade of Hepatocellular Carcinoma Using MCF-3DCNN: A Pilot Study. *BioMed Research International*, 2019:1–12, apr 2019.