Classification of Primary Liver Tumors with Radiomics and Deep Learning based on Multiphasic MRI

MSc Thesis Biomedical Engineering
BM51035 - Medical Physics

by

A. A. Goedhart

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended on Friday April 14, 2023 at 14:30.

Chair and supervisor:	Dr. F.M. Vos	Erasmus MC, TU Delft
Supervisors:	Dr. M.P.A Starmans	Erasmus MC
	Dr. S. Klein	Erasmus MC
Independent committee member:	Dr. D.H.J. Poot	Erasmus MC

Classification of Primary Liver Tumors with Radiomics and Deep Learning based on Multiphasic MRI

A.A. Goedhart[1], M.P.A Starmans[2], F.M. Vos[1,2], and S. Klein[2]

[1]*Delft University of Technology*

[2]*Erasmus Medical Center, Rotterdam*

Abstract

Introduction: Primary liver cancer is a commonly diagnosed cancer and accurate diagnosis is crucial for treatment planning. To differentiate between malignant and benign liver tumors, contrast-enhanced MRI is typically used as it provides information over multiple contrast phases. However, diagnosis based on MRI is challenging. In this study, automatic classification is used to distinguish common primary liver tumors.

Methods: Imaging data from 102 patients with malignant (hepatocellular carcinoma) and benign (focal nodular hyperplasia and hepatocellular adenoma) primary liver tumors was used for binary classification through radiomics and deep learning approaches. The radiomics method was applied with the use of the open-source toolbox WORC. The deep learning model was based on the ResNet-10 architecture. The data input consisted of individual and combined phases of contrast-enhanced T1-weighted and T2-weighted MRI.

Results: The highest performance values were found for the radiomics approach that combined the precontrast, arterial, portal venous, and delayed contrast phases together with T2-weighted MRI, with an AUC of 0.92. The deep learning model scored an AUC of 0.83 with this data input, however substantial overfitting occurred due to the limited sample size.

Conclusion: The radiomics classifiers based on combined contrast-enhanced T1-weighted and T2-weighted MRI can differentiate malignant from benign primary liver tumors with limited data samples. The classification task is too complex with the given data when using a ResNet-10 model and should be applied to an extended dataset.

Keywords: Radiomics, Deep Learning, ResNet, MRI, Post-contrast T1, Liver Cancer

1 Introduction

Primary liver cancer is the sixth most frequently diagnosed cancer and the third most frequent cause of cancer deaths globally, with around 906,000 new cases and 830,000 deaths in 2020 [1]. Whether a primary liver tumor is malignant or benign is crucial for treatment planning. Of the primary malignant lesions, hepatocellular carcinoma (HCC) is the most common phenotype [2] as it accounts for roughly 75% of all liver cancers [3, 4]. The earlier HCC can be detected, the more treatment possibilities there are for the patient, of which examples are liver transplantation, resection, and immunotherapy [5].

Besides primary malignant lesions, a range of primary benign hepatic lesions exists of which frequently occurring examples are focal nodular hyperplasia (FNH) and hepatocellular adenoma (HCA). From these two lesions, FNH is most common with a 0.4–3 % prevalence, followed by HCA with a prevalence between 0.001 and 0.004 % [6–8]. The need for treatment of these benign lesions depends on the presence of symptoms, the risk of future complications, and the risk of malignant transformation. For FNH, patients usually don't experience any symptoms and there is no risk of malignant transformation. In most cases, a follow-up instead of treatment is sufficient [7]. For HCA, there is a higher risk probability, of which bleeding is the highest. HCA can transform into a malignancy and therefore become an HCC lesion over time [7].

To provide the most fitting treatment planning, a fast and accurate diagnosis is needed. A common, early procedure in diagnosis-making is the analysis of magnetic resonance imaging (MRI) by a radiologist. Radiologists use a combination of MRI sequences to analyze lesions in the liver, as they provide complementary information. Sequences used for a standard MRI examination are T2-weighted (T2), pre-contrast and post-contrast T1-weighted (T1), and in-phase and out-of-phase MRI, and diffusion-weighted imaging (DWI) [9]. The diagnosis of HCC with MRI is based on vascular image

characterizations, which appear in typical ways in post-contrast T1 [10]. In contrast-enhanced T1-weighted (CE-T1) imaging, a series of images is acquisitioned over several contrast phases with the use of gadolinium-based contrast agents. The contrast injections emphasize the difference in vascular architecture of the liver parenchyma and most liver lesions [9] and can accentuate phenotype-specific characteristics, which aids radiologists in distinguishing the lesions. For HCC, this typically results in hypervascularity in the arterial contrast phase and washout in the delayed and portal venous phase for instance [10].

Despite the use of contrast enhancement, difficulties in diagnosis can arise when different types of lesions show similar image characterizations. For example, fibrolamellar HCC and FNH are both frequently hypervascular [11]. Another example is a central scar, which is typical for FNH but is also reported in more than a quarter of HCC [11]. HCA can show something that looks like a central scar, which in reality is tissue from fat, necrosis, or old hemorrhage [11]. Difficulties in diagnosis like these may lead to an unnecessary referral from a peripheral to a tertiary care center, which is costly and time-consuming. While diagnosis on imaging remains challenging, the final diagnosis is often based on a biopsy. Besides being time-consuming, taking a biopsy is an invasive method that brings inconveniences to the patient and has the risk of tumor seeding and other complications [12].

Computer-aided diagnosis (CAD) techniques hold a potential solution in the search to reduce the need for biopsies and increase diagnostic accuracy. A CAD tool that can distinguish malignant from benign liver tumors based on quantitative information can decrease the diagnosis process by acting as a quickly available second opinion. In CAD, radiomics and deep learning are the most often applied methods. Both approaches extract high-dimensional, quantitative features from medical images and analyze them for diagnosis or prediction [13]. Whether a radiomics or a deep learning model is the best choice for tumor classification, depends on the complexity of the task and available data. In radiomics, the extracted features are predefined and hand-engineered. They can be divided into histogram, morphologic, and texture features [13]. To extract these features, radiomics methods most often require segmentations of the lesions, which can be time-consuming to make. In deep learning, the extracted features are not predefined and the model learns the best features from the data for the given task [13]. The freedom in feature extraction gives the possibility to have highly accurate classification without the use of segmentations. Therefore, segmentations are often not part of the data input and instead bounding boxes surrounding the tumor are used, which are much easier to define. To perform well, deep learning models need a high number of data samples to generalize well, while datasets in the medical field are often limited in size. In contrast to deep learning, the image information that a radiomics model can learn is limited by the choice of predefined features and important information might therefore be missed. However, this limitation in features can be beneficial for the generalizability of the model when dealing with a small dataset.

For the automatic classification of liver tumors based on CE-T1 MRI, multiple studies have used radiomics methods [10, 14–16] and deep learning methods [17–22] with promising results. In these studies, both radiomics and deep learning models for liver lesion classification have been demonstrated to benefit from the combined usage of various contrast phases and sequences. These studies are described in the literature review in Appendix B, together with more background on primary liver tumors and classification through radiomics and deep learning.

To our best knowledge, no study has been performed on the classification of primary liver tumors based on contrast-enhanced MRI that compares radiomics and deep learning on the same dataset. The primary goal of this research was therefore to develop a method for distinguishing benign and malignant primary liver tumors based on contrast-enhanced MRI imaging with the use of radiomics and deep learning. The secondary aim of this research was to compare the performance of contrast-enhanced and non-enhanced MRI, to analyze the contribution of the use of contrast agents. The third aim was to analyze the deep learning performance without the use of segmentations.

2 Methods

2.1 Data

2.1.1 Dataset description

The dataset used in our study was based on the dataset acquired by Starmans et al. (2021) [23]. This dataset was collected from patients who were diagnosed in or referred to the Erasmus Medical Center (Rotterdam, the Netherlands) between 2002 and 2018. The data consists of the patient's imaging data, age, sex, and liver lesion phenotype. In the study of Starmans et al., only T2 images were used and segmentations were based on these images. For our study, the CE-T1 images of the patients were used. Since the patients originate from different hospitals and had been imaged with different

Figure 1: MRI scans of a patient with a HCA lesion. A: the precontrast phase, no contrast enhancement. B: arterial phase, the contrast is uptaken by the lesion and a heterogeneous pattern can be seen on the spleen. C: the portal venous phase, the spleen is now homogeneous. D: the delayed phase, the image characteristics are almost identical to the portal venous phase, E: the T2-weighted sequence.

protocols, there was much heterogeneity in the contrast-enhanced imaging data. A variety of contrast agents has been used, including Multihance, Gadovist, Dotarem, Magnevist, and Primovist. The number of contrast phases differed per patient and some patients had no contrast-enhanced scans at all and were therefore excluded. For our research, in total four contrast phases were included: the precontrast, arterial, portal venous, and delayed phases. Images made during the hepatobiliary phase were not included in the dataset since only a fraction of the patients were imaged with contrast agents that allow for this contrast phase (Multihance or Primovist). The patients that were diagnosed with HCC, FNH, or HCA were included. A fourth phenotype in the dataset of Starmans et al. [23], i.e. intrahepatic cholangiocarcinoma (iCCA), was not used in our dataset because only four patients could be included. Examples of the included images are shown in Figure 1.

2.1.2 Contrast phase labeling

To compare the classification performance for each contrast phase individually as well as combined, one image per contrast phase for each patient was included in our dataset. To find the CE-T1 images in the dataset, a string search was performed on the following DICOM tags: series description (0008,103E), protocol name (0018,1030), and contrast/bolus agent (0018,0010). Inclusion terms were e.g. 'dynamic', 'multiphase', and the names of contrast agents and the four contrast phases. To select an image per phase, the contrast phases of the available dynamic images had to be identified. The contrast phase labeling steps were based on the protocols described by Donato et al. [24] and were performed under the supervision of an experienced radiologist.

First, the earliest postcontrast image had to be identified, which is the first image to show a hyperintense aorta. The time difference between the first postcontrast images and all the other dynamic images after were calculated.

Second, the image that was made right before the first postcontrast image was selected as the precontrast phase. Although often multiple precontrast images were available and could be found by their series description, this image was selected since it was expected to have the least motion artifacts compared to the following contrast-enhanced images.

Third, the arterial phase can be recognized by heterogeneous enhancement of the kidneys and spleen [25, 26]. Often, multiple images per patient showed this enhancement. Radiologists prefer to evaluate the late arterial phase over the early arterial phase since the contrast agent is more likely to be taken by the tumor. Therefore, the last made arterial image was selected to increase the chance of including a late arterial phase image.

Fourth, the images that no longer showed heterogeneity in the kidneys and spleen were either labeled as the portal venous or delayed phase. Since the portal venous and delayed phases are visually indistinguishable, the DICOM information on acquisition time (0008,0032) was applied in the labeling of these two phases. Any image that was made minimally 2.5 minutes and maximally 7 minutes after the first postcontrast image, was considered to be made during the delayed phase. To have a maximum difference between the portal venous and the delayed phase, the first portal venous and the last delayed images were chosen.

Only patients with all four contrast phases were included in the dataset to compare all phases individually on the same data. Patients with missing acquisition times in the DICOM information were excluded. Since this occurred mostly for patients with HCC, some subjects with benign tumors were excluded to not further imbalance the classes.

2.1.3 Segmentations

The dataset of Starmans et al. [23] included segmentations that were semi-automatically made by radiologists on the T2 images. These segmentations were warped so that they spatially overlapped with the tumors in the dynamic images. To warp these segmentations, a rigid registration of the T2 images to the CE-T1 images was performed with the `elastix` toolbox [27]. The components of the registration consisted of the multi-resolution registration, the mutual information metric, the Euler transform, the adaptive stochastic gradient descent optimizer, and the third-order B-spline interpolator. The masks were warped by using a nearest-neighbor interpolator. The quality of the transformed segmentations was visually analyzed over a sample of the total number of patients.

2.1.4 Data preprocessing

For the radiomics experiments, only z-score normalization was applied on the whole image as the used method allows for different image and voxel sizes.

For the deep learning method, the images were preprocessed to make the image and voxel sizes equal for all samples. First, the images and segmentations were resampled to the same voxel size. This was set to the median voxel size of the images from the portal venous phase: 1.4 x 1.4 x 2.5 mm. Second, the images and segmentations were cropped around the center point of the tumor with a crop size of 192 x 160 x 96 voxels. This bounding box fitted the largest tumor diameters plus a margin for mask shifts due to warping. Padding with the minimal image intensity value was applied if the bounding box fell outside the image. Lastly, z-score normalization was applied to the images.

2.2 Radiomics classification

For the radiomics classification method, the machine learning toolbox WORC (Workflow for Optimal Radiomics Classification) [28] was used. This open-source toolbox optimizes radiomics workflows by performing classification along many combinations and algorithms and hyperparameters and by comparing the performances. WORC uses conventional radiomics pipelines and conventional machine learning algorithms [28]. The workflows in WORC are divided into image acquisition, image preprocessing, segmentation, feature extraction, and data mining [29].

The input data for radiomics with WORC consisted of images, segmentations, and ground truth labels. After pre-processing, 564 quantitative features were extracted from the imaging data within the segmentation boundaries. These features represent information about the intensity, morphology, orientation and positioning, and texture of the segmented lesions [28]. WORC optimizes algorithms and hyperparameters for data mining (e.g. feature selection) and machine learning. In total, 1000 workflows are sampled through a random search of algorithms and hyperparameters. Then, the workflows are ranked based on the F1-score performance of a validation dataset. The 100 workflows highest in rank are ensembled to form the final prediction model.

2.3 Deep learning classification

The deep learning classification was performed with a residual neural network (ResNet) [30], which is a convolutional neural network (CNN) with residual units. The architecture was copied from the Pytorch-based, open-source framework MONAI [31]. The used model is a 3D ResNet-10, which has ten deep layers. This model is the most shallow 3D ResNet within the MONAI framework and is chosen because our dataset was relatively small. A schematic overview of the ResNet-10 architecture is shown in Figure 2. The ten deep layers consist of a convolutional layer, four residual blocks of each of two deep layers, and a final fully connected layer. The first convolutional layer is followed by batch normalization, a ReLU activation function, and max-pooling. A residual block consists of two repetitions of a convolution layer, batch normalization, and a ReLU. Besides the normal residual block outputs, ResNets also have skip connections. In a skip connection, the input is only downsampled before being fed to the next layer and thus does not go through the residual block. Before the input within the residual block goes through the second ReLU layer, the output of the skip connection is added. Together they go through the ReLU and form the input of the next layer. After the last residual block, there is an adaptive average pooling layer that fits its kernel size and stride to a target output of 1 x 1 x 1. Lastly, a fully connected layer is connected with a single output.

2.4 Experimental setup

2.4.1 Fitted models

For both the radiomics and deep learning methods, eight different models were fitted, which included both the images and segmentations as input. To evaluate the predictive value of each MRI sequence, five models of a single sequence were fitted, called Precontrast, Arterial, Portal venous, Delayed, and T2. Next, three

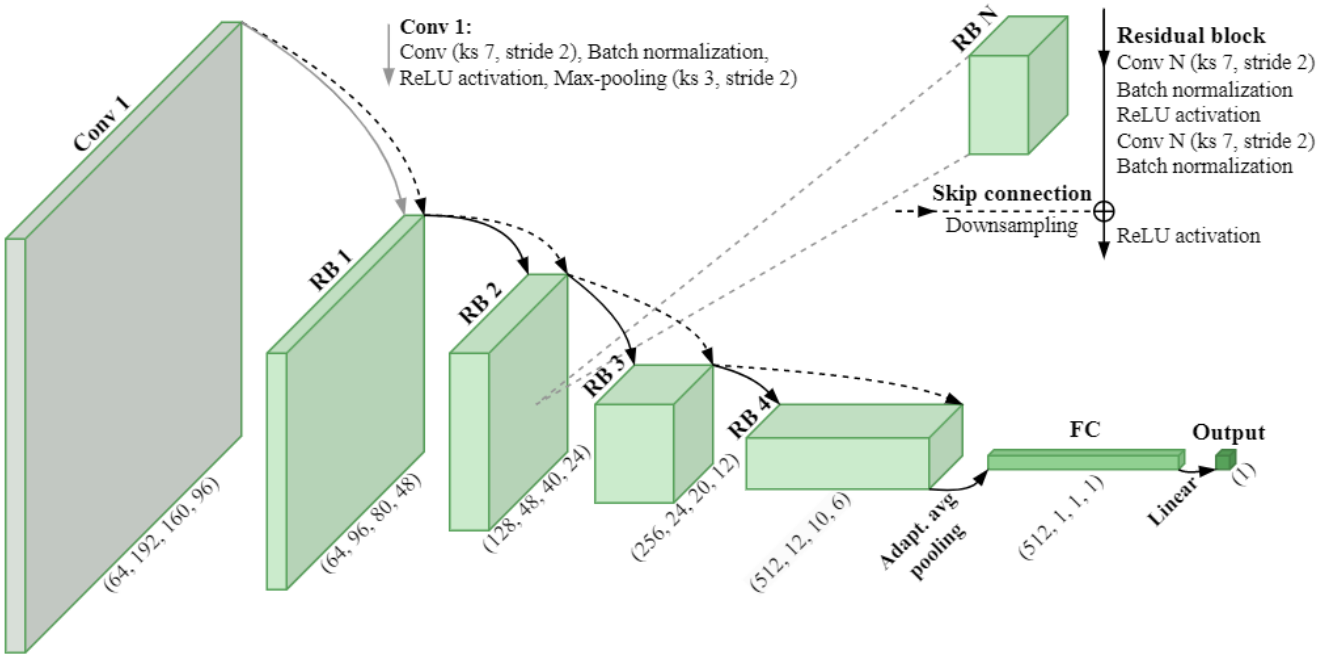

Figure 2: Architecture of the ResNet-10 model. The numbers in brackets are the output dimensions of each layer and represent the number of convolution filters and the width, height, and depth of the image, respectively. The convolution and pooling operations are in 3D and the kernel size and stride apply to all three dimensions. Abbreviations: Conv: convolution, RB: residual block, FC: fully connected, ReLU: rectified linear unit, ks: kernel size, Adapt. avg: adaptive average.

combinations of image inputs were applied: 1) Precontrast + T2, to evaluate the predictive value of all sequences without contrast; 2) All phases, to evaluate the combined predictive value of the four contrast phases; and 3) All phases + T2, to evaluate whether all sequences combined has more value. In the ground truth labels for the binary classification, the positive class represented the malignant tumors, the negative class represented the benign tumors.

2.4.2 Radiomics experiments

The radiomics experiments were performed on AMD Opteron 2378 CPUs. Python version 3.6.8 and WORC version 3.6.0 [29] were used.

The WORC toolbox performed a cross-validation with a 100x stratified random-split for an 80% training and 20% test set. For model optimization, a second internal cross-validation is performed on the training set, with a 5x random-split for an 85 % training and 15 % validation set. With 5 splits for training, 100 splits for testing, and 1000 workflows in the random search, a total of 500,000 workflows were applied per radiomics experiment.

2.4.3 Deep learning experiments

The deep learning experiments were performed on an AMD EPYC 7742 CPU with Nvidia A40 48GB GPUs. Python version 3.7.4, MONAI version 1.1.0 [31], and Pytorch version 1.13.1 [32] were used for fitting the models.

The ResNet-10 model had two input channels per MRI sequence: one for the image and one for the segmentation. After shuffling the data with a fixed random seed, a stratified 5-fold cross-validation was implemented. This was substantially less than the 100x random split of the radiomics method because applying a 100x random split was not feasible for the given time and resources of this research. Data augmentation was applied to the training set to generalize the model. The augmentation was performed with MONAI transforms that were randomly applied with probability p. These transforms consisted of zooming (p = 0.3), flipping (p = 0.5), 20° rotation (p = 0.3), and Gaussian noise with a standard deviation of 0.05 (p = 0.5).

The model had 14,356,929 trainable parameters in total. The batch size was maximized based on the available GPU memory, which allowed for a batch size of 2. An Adam optimizer with a learning rate of 0.0001

was used, after it gave better loss curves compared to a stochastic gradient descent (SGD) optimizer and a range of lower and higher learning rates. Since there were only two classes, the loss function was the binary cross-entropy with logits loss (BCEWithLogitsLoss on PyTorch). This loss functions combines a sigmoid function and binary cross-entropy loss into one layer, turning the output into a probability before the loss is calculated.

The experiments were performed over a total of 200 epochs. The test loss curves showed sharp fluctuations over epochs. Therefore, an exponential moving average filter was applied to smoothen the curves, after which the test output was calculated over an average of the last 5 epochs. A sigmoid layer was applied to the test output to get the prediction probabilities, which were later used for the evaluation metrics.

For the model with the highest AUC value, three more adjusted versions were fitted. In the first of these models, the segmentation input channels were removed. Gradient-weighted class activation mapping (Grad-CAM) [33] was used to make heatmaps for the last convolutional layer of this model and the original one. These heatmaps visualize which regions were important for the final prediction of the CNN. In the other two models, transfer learning was applied with and without the use of segmentations. Transfer learning is often used for limited datasets [34]. The models were initialized with pre-trained weights, from Med3D [35], which are based on 23 medical datasets and are publicly available.

2.5 Statistical analysis of performance

The accuracy, area under the curve (AUC) of the receiver operating characteristics (ROC) curve, F1-score, sensitivity, and specificity were analyzed for both the radiomics and deep learning experiments. The prediction probabilities had values between 0 and 1 and the accuracy, F1-score, sensitivity, and specificity were calculated with a cut-off of 0.5. For the metrics that the radiomics method outputs for the test set evaluation, 95 % confidence intervals were constructed from the 100x random-split cross-validation. These intervals were made with a corrected resampled t-test [36]. For calculating the deep learning evaluation metrics scikit-learn version 1.0.2 [37] was used. The mean value and standard deviation of the 5-fold cross-validation were calculated for all metrics.

The AUC performance of the radiomics models, including 100 values from each split, were compared in a paired corrected t-test [36]. To the pairwise comparison of the eight fitted models, a Bonferonni correction was applied and the calculations were also performed with scikit-learn. This test was not applied to compare the deep learning results because they only consist of five AUC values per model, which makes the test inappropriate.

Boxplots of the AUC values of both methods were made for each fitted model with Seaborn version 0.11.2. AUC values higher than 0.5, were considered to be better than random guessing.

A Mann-Whitney U test was performed on the radiomics features, which tests for significant differences in the distribution between the two classes for each feature [28]. A Bonferonni correction was also applied to this test [29].

2.6 Failed classification analysis

The radiomics method outputs a percentage of correct classifications over the 100 test splits for each subject. To gain insight into why certain lesions could not be classified, failed cases were analyzed with the help of a radiologist. The subjects that had a 0 % correct classification for more than one of the fitted models were selected for the analysis. An experienced radiologist classified the lesions with the use of the T2 and CE-T1 images and stated which sequences were most informative for the decision. The radiologist also labeled the lesions as typical or atypical based on their image characterization. Then, the images were analyzed based on tumor size, image quality, segmentation quality, and the radiologist's input. For the model with the highest AUC value, the lesions that were correctly classified in 50% or less of the test splits, a comparison to the deep learning performance of the corresponding model was made.

3 Results

3.1 Final dataset

The data set consisted of 102 patients in total. Information about age, sex, and phenotype are depicted in Table 1. The malignant class consisted of 40 HCC lesions, and the benign class consisted of 35 HCA and 27 FNH lesions.

Table 1: Clinical characteristics of the dataset.

	Patients	Age*	Male	Female	HCA	FNH	HCC
Benign	62	38 [30, 46]	2	60	35	27	
Malignant	40	68 [60, 73]	22	18			40
Total	102	44 [32, 63]	24	78	35	27	40

*: median [1st quartile, 3rd quartile]

3.2 Warped segmentations

The quality of the warped segmentations was considered sufficient for most scans based on visual analysis. The quality was considered sufficient when the segmentation overlapped well with the tumor, with the focus on the middle axial slices, where the tumor area is typically the largest. Due to warping, few spatial shifts of the segmentation to the lesion occurred both in and out of plane. For smaller tumors, spatial shifts had more effect which led to worse segmentation overlap. In other cases, the segmentations were well aligned. however, with some parts of the tumor missing in the segmentation. These tumor regions were overlooked upon annotation in the T2 images. Examples of warped segmentations for these cases are depicted in Figure 3.

3.3 Radiomics

The performance of the different radiomics experiments is summarized in Table 2. It should be noted that the 95% confidence intervals overlap for all models over all the metrics. From the individual image inputs, the mean AUC values ranged from 0.75 to 0.88. From these models, the phases with contrast enhancement had a higher AUC than the Precontrast and T2 models. From the combined image inputs, the AUC ranged from 0.85 to 0.92. The Precontrast + T2 had a higher AUC than its individual models and performed similarly to the All phases model. In Figure 4, all the AUC values are collated in boxplots. For all performance metrics, the All phases + T2 model had the highest values. According to the pairwise t-test on the AUC, there was no significant difference between any of the models.

The radiomics features associated with a statistically significant difference between the classes were all texture features, except for the three histogram features of the Arterial models. None of the models had significantly different shape features. For the All phases + T2 model, 33 features were statistically significantly different for the two classes. These features had a p-value below 2e-05, which was the Bonferonni corrected p-value of the Mann-Whitney U test. These features were divided into vessel filter (21), local binary pattern (8), Laplacian of Gaussian filter (3), and Gabor filter (1) features.

3.4 Deep learning

The performance of the ResNet-10 model with channels for both images and segmentations is summarized in Table 3. It should be noted that the standard deviations of all metrics are relatively high and extracting from and adding to the mean values gives ranges that overlap for all models. The AUC values of the individually fitted models ranged from 0.77 to 0.80, and of the combined models from 0.79 to 0.83. In Figure 4, all the AUC values are collated in boxplots. In contrast to the radiomics models, the Precontrast model had a higher AUC than the T2 model and the contrast-enhanced models, except for the All phases + T2 model. Like in the radiomics models, the highest AUC can be found for the All phases + T2 model. The mean values of the AUC are lower than for the radiomics models, except for the Precontrast model. Also, the range of the mean AUC values over the models is smaller than for the radiomics models.

Although the performance metrics were within the same range as the radiomics experiments, overfitting on the training sets occurred in all experiments. There was a notable variation in the test loss for the five cross-validation folds and some folds showed overfitting in the test loss straight from the first epoch. The value of the test loss at the first epoch was similar for every fold with a standard deviation range of 0.022-0.053 over all the models. After 200 epochs, the mean value of the test loss remained similar to that of the first epoch, however, the standard deviation range increased to 0.12-0.25. A typical example of the test and train loss curves and the AUC are depicted in Figure 5. This figure shows that the variation of the loss increased over the epochs but the mean AUC and its variation stagnated over time.

Since the highest AUC value was found for the All phases + T2 model, the experiments without segmentation input and with the use of transfer learning were applied to this model. The results are shown in Table 4. The use of pretrained weights decreased all of the performance metrics. Only a slight change was found in the mean values of the evaluation metrics when the segmentation input was removed. The GradCAM visualizations of the models trained with and without segmentations are depicted in Figure 6 for three correctly classified lesions. For two of the three examples, the model did not focus on the area of the tumor for any of the models.

3.5 Failed classification

For the 15 lesions with a 0 % correct classification for more than one radiomics model, no relation was found with tumor size, image quality, segmentation quality, or whether the tumor was typical or atypical on the different sequences. The sequences that the radiologist considered to be uninformative for classification did not directly relate to the sequences of the radiomics models that failed.

For the All phases + T2 radiomics model, 16 lesions had a correct classification score of 50% or less. From these 16 subjects, 11 were incorrectly classified by the All phases + T2 deep learning model.

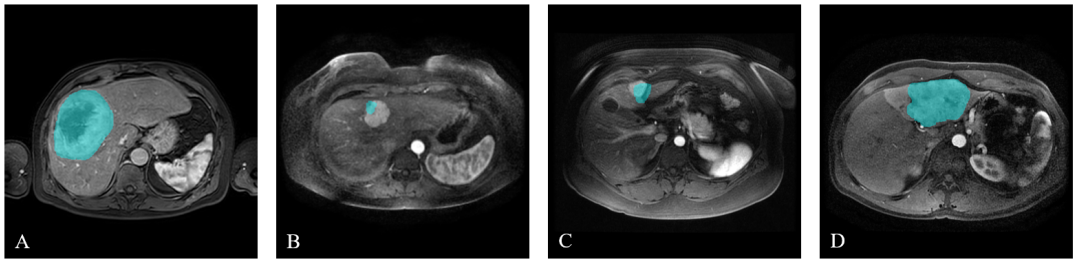

Figure 3: Examples of warped segmentations used for the arterial phase images. A: Good quality segmentation with a 0 % correct classification score in the All phases + T2 radiomics model. B: Lesion with missing tumor regions with a 100 % correct classification score in the All phases + T2 radiomics model. C: Spatial shift in the axial plane with a 100 % correct classification score in the All phases + T2 radiomics model. D: Lesion with missing tumor regions with a 10 % correct classification score in the All phases + T2 radiomics model.

Table 2: Performance of the radiomics experiments. The mean values of the internal cross-validation and the 95 % confidence intervals are reported. Abbreviations: Pre: Precontrast, Art: Arterial, PV: Portal venous, Del: Delayed, AUC: area under the receiver operating characteristic curve. The highest values are written in bold text.

Models	Accuracy	AUC	F1-score	Sensitivity	Specificity
Precontrast	0.70 (0.61, 0.79)	0.75 (0.64, 0.87)	0.69 (0.59, 0.78)	0.48 (0.31, 0.64)	0.84 (0.72, 0.95)
T2	0.74 (0.65, 0.82)	0.80 (0.70, 0.89)	0.73 (0.64, 0.82)	0.58 (0.40, 0.77)	0.83 (0.70, 0.96)
Precontrast + T2	0.80 (0.72, 0.88)	0.85 (0.78, 0.93)	0.79 (0.71, 0.87)	0.69 (0.52, 0.86)	0.86 (0.76, 0.96)
Arterial	0.75 (0.66, 0.85)	0.81 (0.71, 0.92)	0.74 (0.64, 0.84)	0.56 (0.36, 0.76)	0.87 (0.77, 0.97)
Portal venous	0.80 (0.72, 0.88)	0.88 (0.81, 0.95)	0.80 (0.71, 0.88)	0.69 (0.53, 0.85)	0.87 (0.76, 0.98)
Delayed	0.81 (0.73, 0.89)	0.88 (0.82, 0.95)	0.81 (0.72, 0.89)	0.70 (0.54, 0.86)	0.88 (0.78, 0.98)
All phases*	0.80 (0.73, 0.88)	0.87 (0.79, 0.95)	0.80 (0.72, 0.88)	0.68 (0.52, 0.84)	0.88 (0.79, 0.96)
All phases* + T2	**0.85 (0.78, 0.92)**	**0.92 (0.85, 0.98)**	**0.85 (0.78, 0.92)**	**0.78 (0.62, 0.93)**	**0.90 (0.81, 0.99)**

*: All four contrast phases: precontrast, arterial, portal venous, and delayed.

Table 3: Performance of the deep learning experiments. The mean values of the 5-fold cross-validation and the standard deviation are reported. Abbreviations: Pre: Precontrast, Art: Arterial, PV: Portal venous, Del: Delayed, AUC: area under the receiver operating characteristic curve. The highest values are written in bold text.

Models	Accuracy	AUC	F1-score	Sensitivity	Specificity
Precontrast	0.71 ± 0.09	0.80 ± 0.11	0.63 ± 0.12	0.62 ± 0.18	0.77 ± 0.11
T2	0.70 ± 0.07	0.79 ± 0.07	0.62 ± 0.12	0.65 ± 0.19	0.73 ± 0.09
Precontrast + T2	0.74 ± 0.05	0.81 ± 0.06	0.63 ± 0.12	0.60 ± 0.16	0.82 ± 0.03
Arterial	0.72 ± 0.09	0.79 ± 0.08	0.64 ± 0.10	0.62 ± 0.09	0.79 ± 0.13
Portal venous	0.71 ± 0.09	0.77 ± 0.12	0.59 ± 0.11	0.55 ± 0.14	0.81 ± 0.11
Delayed	0.68 ± 0.10	0.77 ± 0.13	0.50 ± 0.24	0.50 ± 0.32	0.79 ± 0.10
All phases*	0.75 ± 0.12	0.79 ± 0.12	**0.68 ± 0.12**	**0.65 ± 0.10**	0.82 ± 0.15
All phases* + T2	**0.75 ± 0.04**	**0.83 ± 0.07**	0.62 ± 0.09	0.55 ± 0.14	**0.87 ± 0.08**

*: All four contrast phases: precontrast, arterial, portal venous, and delayed.

Table 4: Performance of the deep learning All phases* + T2 model with and without the use of segmentations and pretraining. The mean values of the 5-fold cross-validation and the standard deviation are reported. The highest values are written in bold text. AUC: area under the receiver operating characteristic curve.

Segmentations	Pretraining	Accuracy	AUC	F1-score	Sensitivity	Specificity
✓		**0.75 ± 0.04**	**0.83 ± 0.07**	0.62 ± 0.09	0.55 ± 0.14	**0.87 ± 0.08**
		0.75 ± 0.10	0.80 ± 0.08	**0.66 ± 0.14**	**0.60 ± 0.14**	0.85 ± 0.07
✓	✓	0.68 ± 0.09	0.72 ± 0.11	0.57 ± 0.16	0.55 ± 0.19	0.77 ± 0.08
	✓	0.68 ± 0.10	0.71 ± 0.09	0.56 ± 0.14	0.53 ± 0.16	0.79 ± 0.05

*: All four contrast phases: precontrast, arterial, portal venous, and delayed.

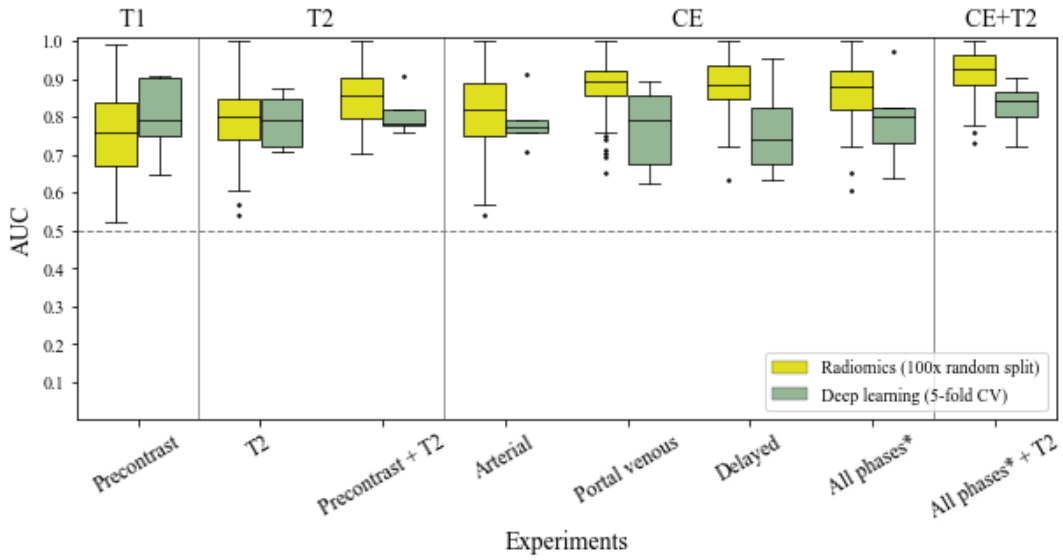

Figure 4: Area under the receiver operating curve (AUC) for the radiomics (100x random split) and deep learning (5-fold cross-validation (CV)) experiments. The middle line in a box represents the median AUC value. The box edges represent the lower and upper quartiles, which form the interquartile range (IQR). The whiskers represent the minimum and maximum value within 1.5*IQR from the box edges, AUC values outside of this range are shown as outliers. The dashed line at 0.5 represents the AUC of random guessing.
*: All four contrast phases: precontrast, arterial, portal venous, and delayed.

4 Discussion

During this research, the classification of malignant and benign primary liver tumors has been studied through the use of different computer-aided diagnosis methods. MRI data of different sequences and contrast phases have been used as input for radiomics and deep learning models. The highest performance metrics were found for the radiomics model with a combined input of all the contrast phases and T2. By extracting information from imaging data in a way that is not possible through visual analysis by humans, radiomics and deep learning can, after further research and validation, improve diagnostic accuracy for primary liver lesions.

4.1 Performance analysis

This study uses the same dataset as Starmans et al. (2021) [23]. The AUC value of 0.74 for the T2 radiomics experiment is comparable to the AUC values of Starmans (0.78 for internal cross-validation and 0.74 and 0.76 for external validation), even though this study only includes 102 of the 187 patients in Starmans's paper. For the radiomics approach, the highest AUC is found for combining all the contrast phases with T2, however, this model does not significantly differ from any of the other models. The performance of the radiomics models that combine sequences fits in the range of results from studies that used radiomics for the classification of HCC and other (benign) lesions with multiphasic MRI [14, 16]. For the deep learning models, the performance is similar to the paper of Jian et al. [17] but lower compared to the other found literature with limited datasets [18, 21, 22]. See Table 1 in Appendix B for a complete overview of similar studies and their performance.

In the evaluation of the radiomics approach, the most relevant features were texture features, in particular vessel filter and local binary pattern features. The vessel filter features appeared most frequently in experiments that included the portal venous and delayed phase. These features might be of importance because the contrast agent specifically enhances the vascularization characterizations of a tumor. The importance of other texture features could be explained by the more heterogeneous image characteristics of the malignant lesions compared to the benign ones. Except for the experiment with arterial phase images, no histogram features were significantly different between the classes. Histogram features are only calculated for regions within the tumor and do therefore not include the intensity differences between normal liver and lesion that are amplified with contrast enhancement. This is different for texture features, which are calculated with filters that blend some information

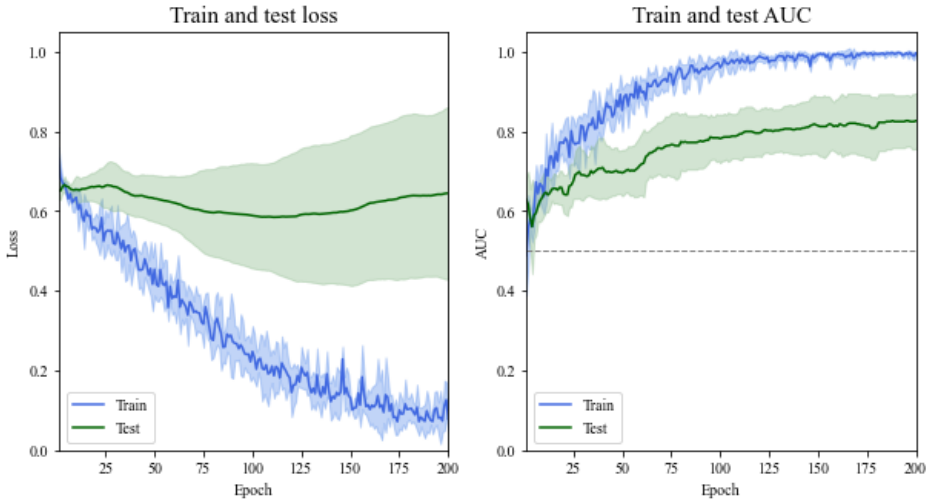

Figure 5: Results of the All phases + T2 deep learning model. Left: Loss curves of the training and test set for the deep learning method. Right: Area under the receiver operating curve (AUC) of the training and test set for the deep learning method. The solid lines represent the mean value and the lighter-colored filled regions represent the mean ± standard deviation. The dashed line at 0.5 represents the AUC of random guessing.

from the tumor edge into the region within the segmentation. Therefore, texture features might take the liver-lesion intensity difference into account, which typically differs for benign and malignant lesions. The shape features were never significantly different over the classes, not even for the original segmentations made on T2, which are considered to be of better quality than the warped segmentations. This is in line with conventional liver lesion diagnosis where the tumor shape plays an unimportant role.

Since the deep learning method does not use predefined features, it potentially could perform well without the input of segmentations. However, the GradCAM visualizations showed that for correctly classified cases, the CNN did not focus on the tumor area for the prediction, also not if the model was trained on the segmentations. Preferably, a CNN extracts information from the lesion since this contains clinically relevant information but this is not the case in these models. Using bounding boxes that include less of the background surrounding the tumor, could aid the model to focus on the tumor areas.

4.2 Limitations of this study

The dataset that has been used to train the machine learning models is relatively small (n=102) for the complexity of the classification task. Much of the total available data had to be discarded as subjects missed acquisition time information due to anonymization methods or missed a contrast phase. Small datasets are likely to cause overfitting. When comparing the radiomics and deep learning methods, overfitting seems to be less of a problem for the radiomics approach. To verify whether overfitting is not a problem in the trained radiomics method, an external test set should be introduced in a future study. For the deep learning models, the applied data augmentation was not enough to prevent overfitting, but it did help to decrease the test loss. That the model did not generalize well, was reflected by the large variance in the test loss. Increasing the epochs caused more overfitting but the AUC stayed stable after 200 epochs and was therefore stopped from that point. After the problem of overfitting was identified, the ResNet-10 model was adjusted to have fewer parameters. A widen factor of 0.5 was applied to halve the number of filters for each layer, which decreased the number of trainable parameters from 14.4 to 3.6 million. However, since this had no effect on the loss and AUC, we chose to continue with the original model to be able to use all the pretrained weights that were available for later experiments. The loss and AUC curves of the smaller ResNet-10 are shown in Figure 9 in Appendix A.

Another limitation of this study was the comparison of the radiomics models to the deep learning models. When comparing the evaluation metrics, one should take into account that the radiomics method outputs 100 values and the deep learning method only outputs five. Statistical tests were not applied to the deep learning results for this reason, which makes the interpretation of these results more subjective. It is

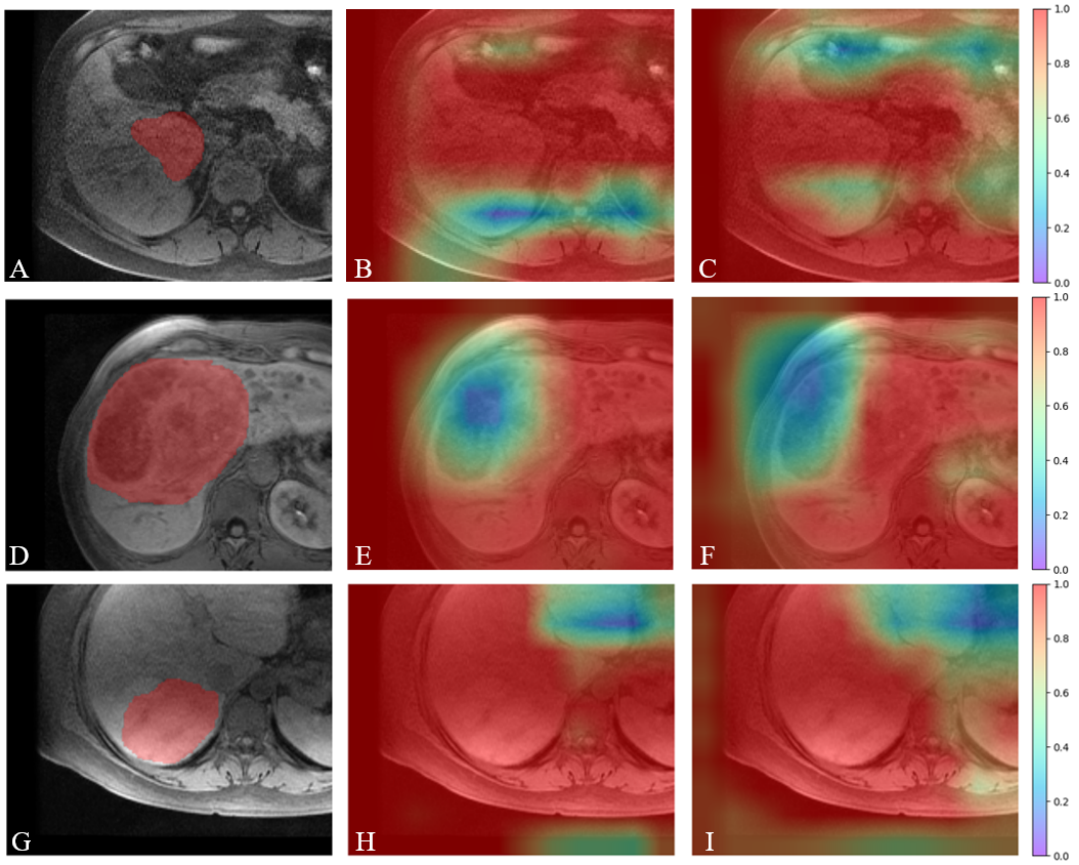

Figure 6: A,D,G: The MRI image with the segmentation in red. B, E, H: The GradCAM visualization of the All phases + T2 model trained on both the images and segmentations. C, F, H: The GradCAM visualizations of the All phases + T2 model trained on the images only. All three examples were correctly classified with the model.

likely that choosing another seed for the train-test splits of the folds will have a big impact on the deep learning results. When comparing the AUC boxplots, the interquartile range (IQR) is much more sensitive to outliers for the deep learning method. On top of that, the IQR changes substantially over epochs, and therefore also the number of outliers. On top of that, the input data of the radiomics models and the deep learning models were differently preprocessed, which makes the comparison of performance less fair. The deep learning input data was resampled and cropped but the input data for radiomics was not. However, radiomics models could benefit from resampled data when combining image inputs of different sequences.

A different limitation of this study is that the deep learning results are only performed for single hyperparameter choices, e.g. the learning rate and optimizer. For this study, the hyperparameters were based on common choices and only compared for a single fold. Variating and comparing hyperparameter selections based on the mean performance of all folds could lead to better performance.

Furthermore, the warped segmentations were considered to be of less quality than the manual segmentations because of the difference in voxel sizes of the T2 and CE-T1 images. When analyzing the subjects for which multiple experiments always failed the classification, no correlation with the segmentation quality was found. Considering the fact that no shape features were significantly different between the classes, this raises the question of how important the quality of segmentation is. In further research, this might be solved by only using segmentations that are of high quality. Since manual segmentations are time-consuming, automatic segmentation with deep learning models like nnU-net [38] might aid speeding up the process.

Lastly, in this research, only binary classification has been performed while more than two phenotypes were included in the dataset. Differentiating between the two benign phenotypes FNH and HCA would be clinically relevant because both phenotypes appear in

similar populations (younger females) and because HCA gives a higher risk of complications and can transform into malignancy. Ideally, more hepatic lesions would be added to the dataset and multi-classification would be performed.

4.3 Future research

As the dataset for this research is relatively small, the first step for future research would be to extend the dataset. Since the highest performance metrics occur for the All phases + T2 experiment, the radiomics model appears to benefit from combining sequences. Except for increasing the sample size, increasing the number of different input images per patient could increase the performance. MRI sequences that could be added in future research are e.g. DWI, in- and out-phase, and the hepatobiliary contrast phase. Another improvement to the radiomics method could be to allow feature extraction on the liver tissue surrounding the tumor, as the intensity difference between tumors and normal lesions is relevant when using contrast enhancement. This could be achieved by dilating the segmentations or using whole liver segmentations. However, since the final radiomics model currently has a good performance, it would be already valuable to compare it to the performance of radiologists on the same sequences and contrast phases.

For the deep learning method, the overfitting is a strong indication that the model needs to be trained on a much larger dataset, preferably in the range of hundreds of scans. Increasing the dataset could be achieved by loosening inclusion and exclusion criteria for both patient and image selection, like tumor size and required sequences. Acquisition times were vital to distinguish between the portal venous and the delayed phase and the most frequent missing phase was the delayed phase. Since the combined image input gave the highest performance, dataset size is apparently more important than correct phase labeling of portal venous and delayed phases. Also, since the portal venous and delayed phases are very similar in image characteristics and performance, for future research, subjects can be included that have at least one image that resembles either of these two phases, even if their exact acquisition time is unknown. To generate more data samples for the deep learning network, 2D models could be used as this provides a sample per slice. However, this would exclude 3D features that might hold important information. For further research, a 2.5D residual neural network might hold a solution for small datasets.

Lastly, patient information like age and sex could give relevant information in diagnostics. They have shown to be strong predictors for binary classification in age-and-sex-only models for the dataset [23]. In typical cases, the phenotypes are correlated with age and sex as HCC appears more in older men and HCA and FNH more in younger women, which is reflected by the used dataset (see Table 1). Since it's desirable to correctly classify rare cases, for this research, the model has only been trained on image information and not on additional features. Automatic classifiers could benefit from sex and age information but future research should consider the risk of missing rare cases.

5 Conclusion

The radiomics classifiers based on combined contrast-enhanced T1-weighted and T2-weighted MRI can differentiate malignant from benign primary liver tumors with limited data samples. We conclude that the classification task is too complex with the given dataset when using a ResNet-10 deep learning model, as overfitting always occurred.

Although adding contrast-enhanced and T2 MRI sequences improved the mean performance in both radiomics and deep learning, a statistically significant improvement was not found.

For future research, the complementary information of multiphasic T1-weighted and T2-weighted MRI should be used for the classification of primary liver tumors. The radiomics performance should be compared to radiologists and validated on an external dataset. Before this deep learning model can be compared to radiologists' performance and tested on an external dataset, it must be trained on a substantially larger dataset for better generalizability.

Acknowledgements

I would like to thank Martijn for supervising me throughout the whole project, for always being supportive and prepared to answer my questions, and for teaching me the essential steps of doing research. I would also like to thank Frans and Stefan for the many insightful ideas and pieces of advice. Furthermore, I wish to thank radiologist Dr. Maarten Thomeer for giving me many insights into the clinical aspects and importance of the project and for guiding me in requiring the final dataset. Lastly, I would like to thank everyone in the BIGR group for sharing their knowledge and experience, which made working on this thesis interesting and enjoyable.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] J. Balogh, D. Victor, E. H. Asham, S. G. Burroughs, M. Boktour, A. Saharia, X. Li, M. Ghobrial, and H. Monsour, "Hepatocellular carcinoma: a review," *Journal of Hepatocellular Carcinoma*, vol. Volume 3, pp. 41–53, Oct. 2016.

[3] N. Razumilava and G. J. Gores, "Classification, diagnosis, and management of cholangiocarcinoma," *Clinical Gastroenterology and Hepatology*, vol. 11, pp. 13–21.e1, Jan. 2013.

[4] K. A. McGlynn, J. L. Petrick, and H. B. El-Serag, "Epidemiology of hepatocellular carcinoma," *Hepatology*, vol. 73, no. S1, pp. 4–13, 2021.

[5] Z. Chen, H. Xie, M. Hu, T. Huang, Y. Hu, N. Sang, and Y. Zhao, "Recent progress in treatment of hepatocellular carcinoma," *Am J Cancer Res*, vol. 10, no. 9, pp. 2993–3036, 2020.

[6] L. M. de Buy Wenniger, V. Terpstra, and U. Beuers, "Focal nodular hyperplasia and hepatic adenoma: Epidemiology and pathology," *Digestive Surgery*, vol. 27, no. 1, pp. 24–31, 2010.

[7] K. J. Oldhafer, V. Habbel, K. Horling, G. Makridis, and K. C. Wagner, "Benign liver tumors," *Visceral Medicine*, vol. 36, no. 4, pp. 292–303, 2020.

[8] "EASL clinical practice guidelines on the management of benign liver tumours," *Journal of Hepatology*, vol. 65, pp. 386–398, Aug. 2016.

[9] L. Vu, J. Morelli, and J. Szklaruk, "Basic MRI for the liver oncologists and surgeons," *Journal of Hepatocellular Carcinoma*, vol. 5, pp. 37–50, Apr. 2018.

[10] R. Hu, H. Li, H. Horng, N. M. Thomasian, Z. Jiao, C. Zhu, B. Zou, and H. X. Bai, "Automated machine learning for differentiation of hepatocellular carcinoma from intrahepatic cholangiocarcinoma on multiphasic MRI," *Scientific Reports*, vol. 12, May 2022.

[11] T. Kim, M. Hori, and H. Onishi, "Liver masses with central or eccentric scar," *Seminars in Ultrasound, CT and MRI*, vol. 30, pp. 418–425, Oct. 2009.

[12] M. A. Silva, B. Hegab, C. Hyde, B. Guo, J. A. C. Buckels, and D. F. Mirza, "Needle track seeding following biopsy of liver lesions in the diagnosis of hepatocellular cancer: a systematic review and meta-analysis," *Gut*, vol. 57, no. 11, pp. 1592–1596, 2008.

[13] B. P. Hyo Jung Park and S. S. Lee, "Radiomics and deep learning: Hepatic applications," *Korean journal of radiology*, vol. 21, no. 4, pp. 387–401, 2020.

[14] M. J. A. Jansen, H. J. Kuijf, W. B. Veldhuis, F. J. Wessels, M. A. Viergever, and J. P. W. Pluim, "Automatic classification of focal liver lesions based on MRI and risk factors," *PLOS ONE*, vol. 14, pp. 1–13, 05 2019.

[15] X. Liu, F. Khalvati, K. Namdar, S. Fischer, S. Lewis, B. Taouli, M. A. Haider, and K. S. Jhaveri, "Can machine learning radiomics provide pre-operative differentiation of combined hepatocellular cholangiocarcinoma from hepatocellular carcinoma and cholangiocarcinoma to inform optimal treatment planning?," *European Radiology*, vol. 31, pp. 244–255, Aug. 2020.

[16] K. Sun, L. Shi, J. Qiu, Y. Pan, X. Wang, and H. Wang, "Multi-phase contrast-enhanced magnetic resonance image-based radiomics-combined machine learning reveals microscopic ultra-early hepatocellular carcinoma lesions," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 49, pp. 2917–2928, Mar. 2022.

[17] W. Jian, H. Ju, X. Cen, M. Cui, H. Zhang, L. Zhang, G. Wang, L. Gu, and W. Zhou, "Improving the malignancy characterization of hepatocellular carcinoma using deeply supervised cross modal transfer learning for non-enhanced MR," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, July 2019.

[18] C. A. Hamm, C. J. Wang, L. J. Savic, M. Ferrante, I. Schobert, T. Schlachter, M. Lin, J. S. Duncan, J. C. Weinreb, J. Chapiro, and B. Letzen, "Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI," *European Radiology*, vol. 29, pp. 3338–3347, Apr. 2019.

[19] C. J. Wang, C. A. Hamm, L. J. Savic, M. Ferrante, I. Schobert, T. Schlachter, M. Lin, J. C. Weinreb, J. S. Duncan, J. Chapiro, and B. Letzen, "Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features," *European Radiology*, vol. 29, pp. 3348–3357, May 2019.

[20] S. Zhen, M. Cheng, Y. Tao, Y. Wang, S. Juengpanich, Z. Jiang, Y. Jiang, Y. Yan, W. Lu, J. Lue, J. Qian, Z. Wu, J. Sun, H. Lin, and X. Cai, "Deep learning for accurate diagnosis of liver tumor based on magnetic resonance imaging and clinical data," *Frontiers in Oncology*, vol. 10, May 2020.

[21] P. M. Oestmann, C. J. Wang, L. J. Savic, C. A. Hamm, S. Stark, I. Schobert, B. Gebauer, T. Schlachter, M. Lin, J. C. Weinreb, R. Batra, D. Mulligan, X. Zhang, J. S. Duncan, and J. Chapiro, "Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver," *European Radiology*, vol. 31, pp. 4981–4990, Jan. 2021.

[22] R. Stollmayer, B. K. Budai, A. Tóth, I. Kalina, E. Hartmann, P. Szoldán, V. Bérczi, P. Maurovich-Horvat, and P. N. Kaposi, "Diagnosis of focal liver lesions with deep learning-based multi-channel analysis of hepatocyte-specific contrast-enhanced magnetic resonance imaging," *World Journal of Gastroenterology*, vol. 27, pp. 5978–5988, Sept. 2021.

[23] M. P. Starmans, R. L. Miclea, V. Vilgrain, M. Ronot, Y. Purcell, J. Verbeek, W. J. Niessen, J. N. Ijzermans, R. A. de Man, M. Doukas, *et al.*, "Automated differentiation of malignant and benign primary solid liver lesions on mri: an externally validated radiomics model," *medRxiv*, Aug 2021.

[24] H. Donato, M. França, I. Candelária, and F. Caseiro-Alves, "Liver MRI: From basic protocol to advanced techniques," *European Journal of Radiology*, vol. 93, pp. 30–39, Aug. 2017.

[25] J. L. Zhang, "Functional magnetic resonance imaging of the kidneys—with and without gadolinium-based contrast," *Advances in Chronic Kidney Disease*, vol. 24, pp. 162–168, May 2017.

[26] T. Vancauwenberghe, A. Snoeckx, D. Vanbeckevoort, S. Dymarkowski, and F. Vanhoenacker, "Imaging of the spleen: what the clinician needs to know," *Singapore Medical Journal*, vol. 56, pp. 133–144, Mar. 2015.

[27] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "elastix: A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[28] M. P. A. Starmans, S. R. van der Voort, T. Phil, M. J. M. Timbergen, M. Vos, G. A. Padmos, W. Kessels, D. Hanff, D. J. Grünhagen, C. Verhoef, S. Sleijfer, M. J. van den Bent, M. Smits, R. S. Dwarkasing, C. J. Els, F. Fiduzi, G. J. L. H. van Leenders, A. Blazevic, J. Hofland, T. Brabander, R. van Gils, G. J. H. Franssen, R. A. Feelders, W. W. de Herder, F. E. Buisman, F. E. J. A. Willemssen, B. Groot Koerkamp, L. Angus, A. A. M. van der Veldt, A. Rajicic, A. E. Odink, M. Deen, J. M. Castillo T, J. F.

Veenland, I. Schoots, M. Renckens, M. Doukas, R. A. de Man, J. N. M. Ijzermans, R. L. Miclea, P. B. Vermeulen, E. E. Bron, M. G. Thomeer, J. J. Visser, W. J. Niessen, and S. Klein, "Reproducible radiomics through automated machine learning validated on twelve clinical applications," *arXiv:2108.08618*, 2021.

[29] M. P. Starmans, S. Van der Voort, T. Phil, and S. Klein, "Workflow for optimal radiomics classification (WORC) documentation," *https://worc.readthedocs.io/en/latest/*.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2015.

[31] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, *et al.*, "MONAI: An open-source framework for deep learning in healthcare," *arXiv:2211.02701*, 2022.

[32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, oct 2019.

[34] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, pp. 500–510, May 2018.

[35] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer Learning for 3D Medical Image Analysis," *arXiv:1904.00625*, 2019.

[36] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[38] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, Dec. 2021.

Appendix A

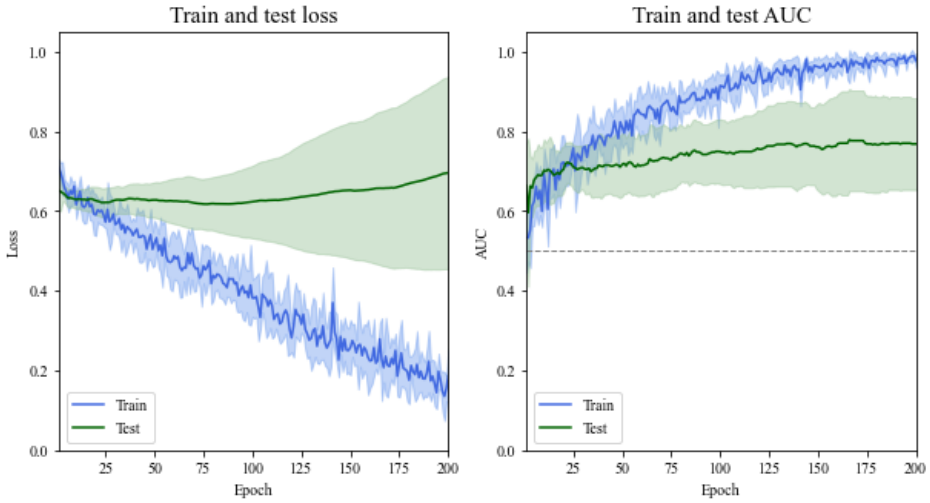

Figure 7: Results of the Portal venous experiment with widen factor = 1.

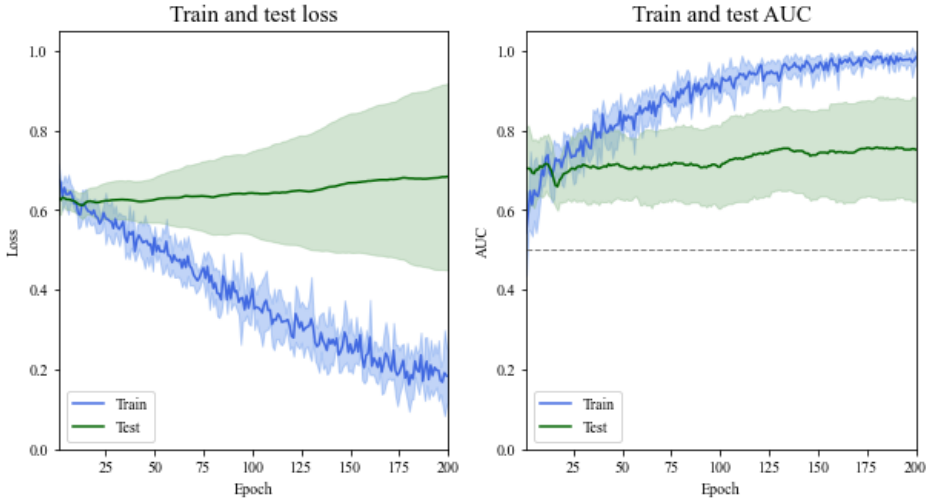

Figure 8: Results of the Portal venous experiment with widen factor = 0.5.

Figure 9: Left: Loss curves of the training and test set for the deep learning method. Right: Area under the receiver operating curve (AUC) of the training and test set for the deep learning method. The solid lines represent the mean value and the lighter-colored filled regions represent the mean ± standard deviation. The dashed line at 0.5 represents the AUC of random guessing.

Appendix B
Classification of primary liver tumors with radiomics and deep learning based on multiphasic MRI: a literature review

A.A. Goedhart

Abstract

Accurate diagnosis of primary liver cancer is crucial for treatment planning. Contrast-enhanced multiphasic MRI is commonly used for the diagnosis of primary liver tumors. The contrast agents emphasize the image characterizations of different liver lesions over multiple contrast phases, which aids radiologists in distinguishing different phenotypes. Despite the use of contrast enhancement, image-based diagnosis can remain challenging, especially when different lesion types show similar image characterizations, often resulting in the need for biopsies. Computer-aided diagnosis techniques like radiomics and deep learning have been used in the literature for the automatic classification of liver lesions with the goal to increase diagnostic accuracy and to reduce the need for biopsies. Radiomics and deep learning models have both been shown to benefit from the combined usage of contrast phases and the literature has presented promising results. However, the generalizability of models is often limited because they have been trained on small and single-center datasets. Also, for both radiomics and deep learning studies on the classification of liver lesions, the standardization of methods is missing. Further studies need to tackle these limitations before the models can be used for clinical practice.

1 Introduction and clinical background

Primary liver cancer is the sixth most frequently diagnosed cancer and the third most frequent cause of cancer deaths globally, with around 906,000 new cases and 830,000 deaths in 2020 [1], and the mortality numbers are still increasing [2]. Whether a primary liver tumor is malignant or benign is crucial for treatment planning. Therefore, accurate diagnosis is essential. A common, early procedure in the diagnosis of primary liver lesions is the analysis of magnetic resonance imaging (MRI) by a radiologist. Radiologists use a combination of MRI sequences to analyze lesions in the liver, as they provide complementary information. Sequences used for a standard MRI examination are T2-weighted (T2), precontrast and postcontrast T1-weighted (T1), in-phase and out-of-phase, and diffusion-weighted imaging (DWI) [3].

1.1 Contrast-enhanced MRI

The diagnosis of primary liver tumors is often based on contrast-enhanced MRI. In contrast-enhanced T1-weighted (CE-T1) images, a series of images is made over several contrast phases with the use of gadolinium-based contrast agents. These contrast phases are defined by the time after intravenous contrast injection and are usually divided into the precontrast phase (before injection), the arterial phase (30 s), the portal venous phase (60-70 s), the delayed phase (3-8 min), and the hepatobiliary phase (20 min - 2 h) [4]. The contrast agent emphasizes the difference in the vascular architecture of the liver parenchyma and most liver lesions [3] and can accentuate phenotype-specific characteristics. Contrast agents can be divided into extracellular and intracellular agents. Extracellular agents include the first four mentioned contrast phases. Examples of intracellular contrast agents are Gadovist, Dotarem, and Magnevist. Intracellular contrast agents are hepato-specific and allow for the fifth contrast phase, the hepatobiliary phase. There are two hepato-specific contrast agents used in the clinic, which are Primovist and Multihance, for which the hepatobiliary phase appears after 20 min and 1-2 h, respectively [4].

1.2 Malignant liver lesions

Of the primary malignant liver lesions, hepatocellular carcinoma (HCC) is the most common phenotype [2] as it accounts for roughly 75% of all liver cancers [5, 6]. HCC occurs 2 to 4 times more in men than in women and the

main risk factors are excessive use of alcohol, hepatitis B and C, diabetes, and obesity [2]. Most patients with HCC also suffer from chronic liver disease and cirrhosis [2]. The earlier that HCC can be detected, the more treatment possibilities there are for the patient, which increases the chance of success for treatment outcomes. Therefore, fast diagnosis is important. Magnetic resonance imaging (MRI) is a preferred imaging technique for HCC diagnosis since it shows higher sensitivity for lesions in patients with chronic liver disease than computed tomography (CT) [2]. The diagnosis of HCC with MRI is based on vascular image characterizations, which appear in typical ways in contrast-enhanced MRI [7]. For HCC, this typically results in hypervascularity in the arterial contrast phase and washout in the delayed and portal venous phase [7]. If imaging alone is not enough for a clear diagnosis, biopsies must be taken. A biopsy has its inconveniences and risks for the patient because it is an invasive method. For example, needle biopsies have the risk of HCC tumor seeding with an incidence of 2.7% [8].

After HCC, cholangiocarcinoma (CCA) is the most common primary malignant lesion found in the liver. CCA can be divided into intrahepatic CCA (iCCA), perihilar CCA (pCCA), and distal CCA (dCCA), of which iCCA is the most common type [5]. Diagnosis of iCCA is more challenging than the diagnosis of HCC [9]. Just like HCC, hepatitis B and C are risk factors for iCCA [5]. Contrast-enhanced MRI contributes to the diagnosis to a large degree because iCCA progressively takes up the contrast agent in the arterial and portal venous phase. In contrast to HCC, iCCA does not show strong washout in the portal venous phase [5]. However, the differentiation between HCC and iCCA can be complicated because differences are often subtle and the hypervascular characterization of iCCA in the arterial phase can lead to it being falsy interpreted as HCC [7]. Further complicating the diagnosis of these malignant lesions is the possibility of a hybrid tumor: combined hepatocellular cholangiocarcinoma (cHCC-CC). In most cases, biopsies have to be taken to confirm the lesion is CCA [5].

1.3 Benign liver lesions

The most common primary benign tumors are hepatic hemangioma (HH), focal nodular hyperplasia (FNH), and hepatocellular adenoma (HCA). HH is the most common lesion of the three as it is present in 0.4–20 % of the population [10]. HH appears 1.2 to 6 times more frequently in women than in men and is most frequently found in women of the age group of 30-50 years. Most patients with HH do not experience any symptoms and the lesion is often found incidentally [10, 11]. For MRI, HH lesions typically show hypointensity on precontrast T1 and hyperintensity on T2 MRI. This does not apply to contrast-enhanced MRI, for which HH can be atypical. During the delayed phase, the lesion can appear hypointense compared to the surrounding liver tissue, which mimics the washout that appears for lesions like HCC. This creates an imaging pitfall and sometimes leads to the need for a biopsy. However, the strong signal on T2 and the enhancement during the arterial phase are usually sufficient for diagnosis [10].

The second most common benign primary liver lesion, FNH, has a prevalence of 0.4–3 % and is usually found incidentally [10, 12]. The lesion is more frequently found in women than in men, with ratios ranging from 2:1 to 26:1. A leading cause of FNH lesions is vascular abnormalities and the lesion is associated with HH. FNH has no transformation into a malignant lesion in further stages, does typically not cause complications, and is mostly asymptomatic [12]. MRI has the overall best performance for FNH diagnosis [10]. On T2, FNH is isotense or slightly hyperintense, while it is isointense or slightly hyper-intense on non-enhanced T1. For FNH, the sensitivity of MRI diagnosis decreases when the lesion does not include a central scar [10]. A central scar appears in 30 % of FNH and shows hypointensity on non-enhanced T1. For contrast-enhanced T1, FNH shows increased signal intensity during the arterial and portal venous phase but the central scar is not enhanced. During the delayed and hepatobiliary phase, FNH (including the central scar) is hyperintense [13]. A pitfall for FNH diagnosis is hypervascularity, as fibrolamellar HCC and FNH are both frequently hypervascular [14]. The central scar can be a pitfall for other lesion: even though it is typical for FNH, it is also reported in more than a quarter of HCC [14]. HCA can show something that looks like a central scar, which in reality is tissue from fat, necrosis, or old hemorrhage [14].

HCA is the third most common benign primary liver with a prevalence between 0.001 and 0.004 % [10]. HCA is mainly found in women and very rarely in men. The lesion is associated with the intake of oral contraceptives because of the steroid sex hormones that those contain. Another factor that is correlated with HCA is glycogen storage disease [10, 12]. Just like the other benign lesions, HCA is often asymptomatic. However, HCA must be treated when it reaches a certain volume. This is because HCA has the risk of complications like hemorrhage and malignant transformation into HCC [10, 12]. Diagnosis of HCA is preferably done on solely imaging because needle biopsies are often not clear enough to confirm the diagnosis and because the tumors are likely to bleed [12]. MRI is the best image modality for HCA [10]. The image characterization on MRI differs per HCA subphenotype. For example, inflammatory HCA can be characterized by hyperintensity on T2 and continuous enhancement during the delayed phase with the use of extracellular contrast agents. For β-catenin HCA, enhancement is shown during the arterial

phase but the intensity can be both continuous or decreased during the delayed phase. In the hepatobiliary phase, typical HCA is not enhanced [13]. The difference in risks of complications of benign tumors emphasizes the importance of distinguishing HCA from FNH and HH. In contrast to FNH and HCA, HH is a nonsolid tumor which makes a diagnosis on imaging relatively easy [15]. When using intracellular contrast agents, both inflammatory and β-catenin HCA can show contrast agent uptake during the hepatobiliary phase, which mimics FNH [16]. However, HCA usually shows more heterogeneous intensities than FNH [13]. In case of doubt, phenotype confirmation can be accomplished after surgical excision [12].

2 Computer-aided image analysis

Despite the use of contrast enhancement, difficulties in diagnosis can arise when different types of lesions show similar image appearances. These difficulties may lead to an unnecessary referral from a peripheral to a tertiary care center and biopsies, which can be costly and time-consuming. Computer-aided diagnosis (CAD) techniques hold a potential solution in the search to reduce the need for biopsies and increase diagnostic accuracy if they overcome the imaging pitfalls. In CAD, radiomics and deep learning are the most researched methods. Both extract high-dimensional, quantitative features from medical images and analyze them for diagnosis or prediction [17].

2.1 Radiomics

Radiomics is an image analysis technique in which a high number of quantitative medical imaging features are extracted for clinical predictions, to gain information that is more complete than from a physician or radiologist alone [18]. A radiomics workflow input requires high-quality images for diagnostics or treatment (for example MRI- or CT-based) and for shape information it also requires lesion segmentations that are either made manually or (semi-)automatically. Image preprocessing steps like normalization and resampling voxel size are usually required to extract radiomics features [17, 19].

The quantitative features are extracted within the segmented region and are based on histogram, morphologic, and texture features [17]. The distribution of intensity values over voxels in the segmentation can be shown in a histogram. Examples of histogram features are magnitude, dispersion, and asymmetry. Features about the size, volume, and shape of the lesion are examples of morphologic features. Information about the spatial relationship between intensities can be found in textural features in which pixel values are compared to surrounding ones. The higher-order texture features are not extracted from the image directly but from filtered versions. Examples of filtering are smoothing Gaussian filters and edge-enhancing Laplacian filters [17].

After the calculation of features, the next step is feature selection. From the hundreds to thousands of calculated features, a large part stays unused [19]. Selection is based on whether the features are independent of the others, reproducible, and prominent [18]. Different selection criteria can be used and ideally features differ for each class with statistical significance. When dealing with a large amount of radiomics features, machine learning can provide help in the feature selection and for the classification based on the selected features [17].

After the features are selected, machine learning classifiers are used for the prediction. Various machine learning methods are used for feature-based classification, like regression, support vector machine (SVM), decision tree, and random forest, which all require hyperparameter optimization [17].

2.2 Deep learning

Deep learning is a subset of machine learning, inspired by the neural networks in the human brain. In deep learning, predictions are made with models of multiple layers of connected neurons that extract features from the input data. For medical image classification, convolutional neural networks (CNN) in combination with supervised learning (training on labeled data) is the most popular method [20]. In a CNN classifier, the middle layers (called hidden layers) consist of repetitions of convolutional layers and pooling layers. Convolutional layers apply filters to the input of which the network learns different image features, like edges and textures. Higher-level features are extracted in deeper layers, which allows the network to learn complex tasks. Pooling layers downsample the convolutional output. The output of the hidden layers is flattened and passed through one or more fully connected layers that apply weights. Lastly, the output layer has a size equal to the classes and returns the class predictions by passing a softmax layer [20, 21].

In contrast to radiomics, features do not need to be predefined and no separate feature extraction step has to be taken for learning input-output relations in deep learning [17, 20, 22]. The freedom in feature extraction gives the possibility to have high-performance classification without the use of segmentations. Therefore, segmentations are often not part of the data input and instead bounding boxes surrounding the tumor are used. However, since deep learning features are not predefined, networks are often seen as a black box, which makes them harder to interpret than radiomics models. Furthermore, deep learning models typically need a large number of training data [17]. If the dataset is too small, the model will

not generalize well and can lead to overfitting, which makes them less suited for limited medical datasets than radiomics.

CNNs for image classification increased in popularity after the introduction of AlexNet when it won the ImageNet classification championship in 2012. Following up on AlexNet, various deeper networks like GoogleNet and VGGNet improved the classification performance further in 2014 [23]. However, neural networks can not be infinitely increased in depth to gain better performance. Deeper networks have the problem of information loss due to many backpropagation steps and the gradients can disappear or explode [24]. The introduction of residual neural networks (ResNets) by He et al. [25] in 2015 contributed to solving these problems. In the architecture, the data does not go through every layer as the network includes so-called skip connections that skip over a block of layers (residual blocks). The output of such a residual block is added together with the output of the skip connection. Therefore, the input for the following layers still contains information from data from earlier layers, as the network leans the differences between the residual block and the skip connection output. This way, ResNet tackles the issue of degradation and allows for deeper architectures with more than 100 deep layers. By adding skip connections between all deep layers, Huang et al. [26] introduced the densely connected convolutional network (DenseNet) in 2016. In a DenseNet, all layers are connected to all other layers within the same dense block. Therefore, the input of a layer consists of the concatenated output of all previous layers. Just like with ResNets, the problem of vanishing gradients is tackled by reusing feature maps from previous layers. Compared to ResNets, DenseNets need fewer parameters for similar test accuracy [26]. In the literature on the two networks, the minimum number of layers for ResNet and DenseNet are 10 and 121, respectively.

3 Automatic classification of liver lesions

For this literature review, studies on the automatic classification of primary liver tumors with the use of contrast-enhanced MRI are discussed. The studies in this review aimed to classify at least one of the following liver lesions: HCC, CCA, HH, FNH, or HCA. The imaging data input of the studies had to include one or more contrast phases. The methods for the automated classification had to be either based on a radiomics model or a CNN deep learning model, or a combination of both. Lastly, only studies from 2019 and on were included. In total, four radiomics studies, five deep learning studies, and one combined study were included in the review. An overview of the studies is shown in Table 1.

3.1 Radiomics studies

Motivated by the lack of automatic classifiers for liver lesions that were fitted on dynamic contrast-enhanced MRI, Jansen et al. [27] developed a classifier that was fitted on T2, precontrast T1, and the late arterial and portal venous contrast phases. From these sequences, 164 features were extracted, which include image features (contrast curve, histogram, and texture) and risk factor features. The added risk factors were the presence of steatosis, cirrhosis, and other known primary tumor in the body. The dataset consisted of 95 patients with 125 benign lesions (HCA, cysts, and HH) and 88 malignant lesions (HCC and metastasis from different sites). The model differentiated the five lesions with a single type of classifier: an extremely randomized trees classifier. This classifier assembles decision trees, of which each tree node applies a random set of thresholds to a random subset of features in order to keep the most informative features. Adding the dynamic images to the T2 increased the performance of the model, and so did adding risk factor features. Further, it was concluded that features were selected from all categories, which indicated the importance of a wide selection of features. The AUC values of the model that included all sequences and all features ranged from 0.88 to 1.00 for the five classes.

A recent study from Sun et al. [28] showed that multiphasic MRI images of small HCC (SHHC) tumors (diameter less than 2 cm) could be differentiated from normal liver tissue and from benign lesions (HH and cysts). In this study, the precontrast, arterial, portal venous, and delayed phase images from 124 subjects were used both individually and combined as model input. The radiomics model extracted 1132 features and they were selected with a least absolute shrinkage and selection operator (LASSO) regression model. A radial basis-function, kernel-based support vector machine classifier was used. First, models were trained on individual contrast phases and the LASSO regression model was used to filter the features. After filtration, new models were fitted on the individual phases and a combination of all four phases. With an AUC of 0.93 and 0.97 for the SHCC-normal tissue and SHCC-benign lesions classification, respectively, the combination of all phases was higher than for the individually fitted models.

In a study by Liu et al. [29] on differentiating HCC from malignant non-HCC liver tumors, models based on different MRI and CT features were compared. The MRI data consisted of non-enhanced sequences (T2, precontrast T1, DWI, and in-phase) and contrast-enhanced T1 images (arterial, portal venous, late venous, delayed, and hepatobiliary phase). Precontrast, arterial, portal venous, and delayed CT phases were also included. The data included 86 lesions of HCC, CCA, and combined hepatocellular cholangiocarcinoma (cHCC-CC). In total,

1419 radiomics features were extracted and an SVM classifier was used. For differentiating HCC, post-contrast phases had a higher AUC (0.79-0.81) than the non-enhanced MRI sequences (0.45-74). The model fitted on the hepatobiliary phase had an AUC of 0.90 but this was considered to be an unreliable result due to the small sample size of 23 lesions. All post-contrast phases individually showed an AUC within the range of 0.79-0.81, compared to the AUC range of 0.49-0.74 for the mentioned non-enhanced sequences. Opposite to the MRI images, for CT the precontrast model had a higher AUC (0.81) than the CT contrast phases (0.52-0.71).

In a study by Hu et al. [7], a radiomics model for the differentiation of HCC and iCCA was researched with manual and automatic optimization. The dataset consisted of 489 subjects and the input data consisted of combined T2 and CE-T1 in the arterial and portal venous phase. From the sequences, 173 image features were extracted, which were repeatedly calculated for different isotropic voxel sizes. On top of that, the sex and age information was added, which result in 6130 features per sequence. The model used variance-based feature selection (VBFS) for different thresholds and the following eight classifiers were used on the training set: SVM, random forest, multilayer perceptron, XGBoost, AdaBoost, extra trees, logistic regression, and gradient boosting. In the manual optimization, for each combination of thresholds and classifiers, a k-fold cross-validation was performed and the mean AUC over the folds was documented. For each classifier, the threshold with the highest AUC was selected for the final model. For the automatic optimization, the Tree-Based Pipeline Optimization Tool (TPOT) was used. TPOT automates the feature extraction, feature selection, and model selection for maximal accuracy, by using a search algorithm. For each run, with a population size of 20, the TPOT classifier randomly generates 10 model pipelines, which form a generation. The best-performing model pipelines of the generation fill 10 % of the population of the next run. The model is run 10 times, forming 10 generations, and in the end 100 % of the population is generated by the selection process. From this population, the contents of two random model pipelines are split and swapped, and mutation operation is applied to the other model pipelines. This is repeated for the 10 generations and then the model pipeline with the highest AUC was selected for the final model. The best-performing model of the manual and automatic optimization methods had similar AUC values on the test set of 0.79-0.80 and 0.76–0.79, respectively. The automated method showed similar sensitivity and specificity to radiologists.

3.2 Deep learning studies

Hamm et al. [30] classified six common liver lesions (FNH, HCC, iCCA, cysts, cavernous hemangioma, and colorectal cancer metastasis) with contrast-enhanced MRI of the late arterial, portal venous and delayed phase. Precontrast images were not included. A CNN of three convolutional layers was trained with cross-validation for 494 subjects. HCC could be distinguished from the rest with an AUC of 0.992 and 92 % accuracy, which was higher than the accuracy of radiologists. For the other lesions, only the sensitivity and specificity across all lesions were given, with an average value of 90 % and 98 %, respectively. In a consecutive study by Wang et al. [31] the CNN was combined with radiomics features. Feature maps were created for interpretability in the decision-making of the model. This was done by labeling a subset of the lesions with radiological features that fitted the lesions, like a central scar, washout, and heterogeneity. For the differentiation of HCC, the sensitivity was 82 %.

In a study by Jian et al. [32], a deeply supervised cross-modal transfer learning CNN model was used for the characterization of HCC. Pretraining was performed on 2D MRI data of 150 subjects, which included the precontrast, arterial, and portal venous phases. During the cross-modal pertaining, the precontrast and contrast-enhanced images were put into the model in separate channels and went through separate convolutional layers before they were combined in a final linear layer, leading to a submodel for the arterial phase and one for the portal venous phase. For the final training and testing, the data consisted of only precontrast images. The motivation for testing without contrast enhancement was to provide a classifier for cases where using a contrast agent is clinically not possible. The best performance was achieved when the submodels were combined and deep supervision was applied, which led to an AUC of 0.82.

Binary classification of HCC versus non-HCC lesions was performed in a study by Oestmann et al. [33]. The dataset consisted of 150 lesions and included lesions with atypical imaging features. The non-HCC class consisted of iCCA, HH, cysts, regenerative nodules, dysplastic nodules, FNH and bile duct adenoma. The dataset included contrast-enhanced MRI of the arterial, portal venous, and delayed phases. Precontrast images were not used. The CNN model was built with three convolutional layers, two maximum pooling layers, and two fully connected layers. The HCC lesions were differentiated with an AUC of 0.912, which demonstrates the model's ability to correctly classify atypical HCC.

The high performance of tumor liver classification with the use of DenseNets was demonstrated by Stollmayer et al. [34]. In this study, a 2D- and a 3D-DenseNet-264 model were compared for the classification of HCC, FNH, and

metastases. The dataset consisted of 216 lesions imaged on T2 and CE-T1 MRI in the arterial, portal venous, and hepatobiliary phases. The 3D images of each sequence were registered, cropped around the lesion, and concatenated into a single image. For the 2D method, the concatenated image consisted of the three axial slices per sequence that were the most representative. For the three lesions, the average AUC values of the 2D and 3D networks were 0.98 and 0.94, respectively. For none of the classes, the difference was statistically significant for the two models. The results of this study indicate that 2D information can be sufficient for the differentiation of liver lesions.

A deep learning system created with a pretrained Google Inception-ResNet V2 CNN by Zhen et al. [35] was trained for binary classification (malignant versus benign), three-way malignancy (HCC, non-HCC primary malignancy like iCCA, and non-hepatic metastasis) and seven-way classification (three-way malignancy, FNH, cyst, HH, and other benign nodules). The image data sequences consisted of T2, DWI, and precontrast, arterial, portal venous, and delayed phases from 1411 subjects. The models were trained on two datasets: one including the contrast phases and one without. For the seven-way classification, the AUC values were in the range of 0.897-0.987 and 0.841-0.965 for the model with and without contrast enhancement, respectively. For all classifiers, the performance was similar to that of radiologists but the use of contrast enhancement did not give statistically better results. With this study, Zhen et al. showed the potential of accurate deep learning diagnosis without the use of contrast agents, in contrast to the other studies mentioned before.

4 Discussion

From the described radiomics and deep learning studies, we can conclude that the automatic classification of primary liver lesions is feasible with the use of contrast-enhanced MRI. Both methods show promising results with the use of combined contrast phases, which indicates that radiomics and deep learning models benefit from the complimentary information of the different phases. The results of the study show that high AUC values can be achieved with limited datasets (150 samples or less) for both radiomics and deep learning. The discussed studies include different phenotypes as classes, which makes direct comparisons of the performances challenging. Additionally, the limitations of the studies have to be considered when interpreting the results.

Limitations of some of the studies are the sample size and the use of single-center data. When training on small datasets, the model is more prone to overfitting and may not generalize well. Using-single center data is also limits the generalizability. As imaging data from different centers vary from each other in imaging protocol and image quality, a model that performs well on multi-center data is more robust. Therefore, the generalizability of single-centered studies should always be verified with an external dataset. The performance of the studies with small datasets, for example of Liu et al. [29] (86 lesions) could probably be improved by only extending the data. However, since cHCC-HCC is less common and they included the hepatobiliary contrast phase that can only be found for Primovist and Multihance, adding more data is difficult. Since not all contrast-enhanced MRI examinations include the hepatobiliary phase, it leads to the discussion of whether including this phase enhances or limits the training performance. To prevent having a too small dataset, it may be preferable to include patients that only miss the hepatobiliary phase, rather than excluding them from the dataset. Another small data study is the study of Sun et al. [28], which achieved a high AUC for a relatively small sample size but the data was single-centered and similar performance is not expected when the model will be tested on external data. The same holds for the study of Stollmayer et al. [34], which had a very high AUC. The used dataset was small and from a single institute, so the generalization of the model is not demonstrated.

Despite the fact that all the studies used contrast-enhanced T1 MRI, several did not make use of the precontrast phase [7, 30, 33, 34]. The precontrast phase is usually always available for patients who have been imaged with contrast agents and its image characterizations can differ for different liver lesions. The study of Jian et al. [32] showed that classification on test sets of precontrast images is possible when pretraining has been performed on precontrast and postcontrast phases. Not including the precontrast in the training data restricts the amount of complimentary information that a model can learn and is therefore a limitation.

The methods of the radiomics studies and of the deep learning studies showed very different measures, which shows a lack of standardization. A clear difference in the radiomics studies is the number of extracted features, varying from 164 to 6130. Also, the methods for feature selection were different and the number of classifiers ranged from one to eight. This shows there is little consensus on what radiomics methods are optimal for liver lesion classification since there are so many options for feature extraction, feature selection, and classification. A framework that optimizes for these steps was introduced by Starmans et al. [19], in order to optimize radiomics classification.

Since the use of deep learning for liver tumor classification is fairly new, standardization is still missing. The depth of the networks and the number of filters per layer of the used CNNs showed substantial variation. The use of well-known, standardized architectures like ResNets

Table 1: Overview of discussed studies. Abbreviations: AUC: area under the receiver operating characteristic curve, DWI: diffusion-weighted imaging, CE: contrast-enhanced, pre: precontrast, art: arterial, pv: portal venous, lv: late venous, del: delayed, hb: hepatobiliary, Sn = sensitivity

Study	Classification	Method	MRI sequences	Sample size	Performance (CE)
Jansen et al. (2019) [27]	5 hepatic lesions	Radiomics	T2 and CE-T1 (pre, art, pv)	213 lesions	AUC = 0.88-1.00
Sun et al. (2022) [28]	Small HCC - non-HCC	Radiomics	CE-T1 (pre, art, pv, del)	124 subjects	AUC = 0.93-0.97
Liu et al. (2021) [29]	HCC - CCA/cHCC-CC	Radiomics	T2, DWI, in-phase, CE-T1 (pre, art, pv, lv, del, hb)	86 lesions	AUC = 0.79-0.81
Hu et al. (2022) [7]	HCC - iCCA	Radiomics	T2 and CE-T1 (art, pv)	489 subjects	AUC = 0.80
Hamm et al. (2019) [30]	6 hepatic lesions	CNN	CE-T1 (art, pv, del)	494 subjects	AUC = 0.992 (for HCC)
Wang et al. (2019) [31]	6 hepatic lesions	CNN/Radiomics	CE-T1 (art, pv, del)	494 subjects	Sn = 82.0 % (for HCC)
Jian et al. (2019) [32]	Detection of HCC	CNN	CE-T1 (pre, art, pv)	150 subjects	AUC = 0.82
Oestmann et al. (2021) [33]	HCC - non-HCC	CNN	CE-T1 (art, pv, del)	150 lesions	AUC = 0.912
Stollmayer et al. (2021) [34]	HCC, FNH, metastases	CNN	T2 and CE-T1 (art, pv, hb)	216 lesions	AUC = 0.98 (2D)
Zhen et al. (2020) [35]	7 hepatic lesions	CNN	T2, DWI, CE-T1 (pre, art, pv, del)	1411 subjects	AUC = 0.897-0.987

and DenseNets could improve the ability to compare studies and to use of transfer learning, as pretrained weights will be easily implemented in the models.

In this literature review, only contrast-enhanced MRI studies have been discussed. However, a substantial number of studies based on contrast-enhanced CT can be found in the literature as this imaging modality is also commonly used for liver lesion patients. As contrast-enhanced CT also contains complimentary information divided over contrast phases, the inclusion of CT-based studies is expected to give a more comprehensive and inclusive overview of the contribution of contrast-enhanced multiphasic imaging for liver tumor diagnosis from the past few years.

5 Conclusion

Radiomics and deep learning models based on multiphasic MRI show promising results for the automatic classification of primary liver lesions. Both radiomics and deep learning models for liver lesion classification have demonstrated to benefit from the combined usage of various contrast phases and sequences. Using automated classification methods in the clinic could aid in the diagnosis of primary liver lesions when differentiation on image characterizations causes difficulties for radiologists. Removing doubts about image-based diagnoses can decrease the need for biopsies and increase the diagnostic accuracy.

For further research, studies should aim for multi-center data of a large sample size as this improves the generalizability of the model. Furthermore, the models could benefit from transfer learning on pretrained multiphasic data as medical imaging data is often limited. To use transfer learning and to compare studies fairly, more standardization is needed. In following studies, the performance always needs to be compared to radiologists and the models always need to be validated on external datasets before they can be implemented in the clinic.

Literature references

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] J. Balogh, D. Victor, E. H. Asham, S. G. Burroughs, M. Boktour, A. Saharia, X. Li, M. Ghobrial, and H. Monsour, "Hepatocellular carcinoma: a review," *Journal of Hepatocellular Carcinoma*, vol. Volume 3, pp. 41–53, Oct. 2016.

[3] L. Vu, J. Morelli, and J. Szklaruk, "Basic MRI for the liver oncologists and surgeons," *Journal of Hepatocellular Carcinoma*, vol. Volume 5, pp. 37–50, Apr. 2018.

[4] H. Donato, M. França, I. Candelária, and F. Caseiro-Alves, "Liver MRI: From basic protocol to advanced techniques," *European Journal of Radiology*, vol. 93, pp. 30–39, Aug. 2017.

[5] N. Razumilava and G. J. Gores, "Classification, diagnosis, and management of cholangiocarcinoma," *Clinical Gastroenterology and Hepatology*, vol. 11, pp. 13–21.e1, Jan. 2013.

[6] K. A. McGlynn, J. L. Petrick, and H. B. El-Serag, "Epidemiology of hepatocellular carcinoma," *Hepatology*, vol. 73, no. S1, pp. 4–13, 2021.

[7] R. Hu, H. Li, H. Horng, N. M. Thomasian, Z. Jiao, C. Zhu, B. Zou, and H. X. Bai, "Automated machine learning for differentiation of hepatocellular carcinoma from intrahepatic cholangiocarcinoma on multiphasic MRI," *Scientific Reports*, vol. 12, May 2022.

[8] M. A. Silva, B. Hegab, C. Hyde, B. Guo, J. A. C. Buckels, and D. F. Mirza, "Needle track seeding following biopsy of liver lesions in the diagnosis of hepatocellular cancer: a systematic review and meta-analysis," *Gut*, vol. 57, no. 11, pp. 1592–1596, 2008.

[9] R. Cannella, T. J. Fraum, D. R. Ludwig, A. A. Borhani, A. Tsung, A. Furlan, and K. J. Fowler, "Targetoid appearance on T2-weighted imaging and signs of tumor vascular involvement: diagnostic value for differentiating HCC from other primary liver carcinomas," *European Radiology*, vol. 31, pp. 6868–6878, Feb. 2021.

[10] "EASL clinical practice guidelines on the management of benign liver tumours," *Journal of Hepatology*, vol. 65, pp. 386–398, Aug. 2016.

[11] N. Bajenaru, V. Balaban, F. Săvulescu, I. Campeanu, and T. Patrascu, "Hepatic hemangioma-review," *Journal of medicine and life*, vol. 8, no. Spec Issue, p. 4, 2015.

[12] L. M. de Buy Wenniger, V. Terpstra, and U. Beuers, "Focal nodular hyperplasia and hepatic adenoma: Epidemiology and pathology," *Digestive Surgery*, vol. 27, no. 1, pp. 24–31, 2010.

[13] J. W. van den Esschert, T. M. van Gulik, and S. S. Phoa, "Imaging modalities for focal nodular hyperplasia and hepatocellular adenoma," *Digestive Surgery*, vol. 27, no. 1, pp. 46–55, 2010.

[14] T. Kim, M. Hori, and H. Onishi, "Liver masses with central or eccentric scar," *Seminars in Ultrasound, CT and MRI*, vol. 30, pp. 418–425, Oct. 2009.

[15] M. P. Starmans, R. L. Miclea, V. Vilgrain, M. Ronot, Y. Purcell, J. Verbeek, W. J. Niessen, J. N. Ijzermans, R. A. de Man, M. Doukas, S. Klein, and M. G. Thomeer, "Automated differentiation of malignant and benign primary solid liver lesions on MRI: an externally validated radiomics model," Aug. 2021.

[16] A. Ba-Ssalamah, C. Antunes, D. Feier, N. Bastati, J. C. Hodge, J. Stift, M. A. Cipriano, F. Wrba, M. Trauner, C. J. Herold, and F. Caseiro-Alves, "Morphologic and molecular features of hepatocellular adenoma with gadoxetic acid–enhanced MR imaging," *Radiology*, vol. 277, pp. 104–113, Oct. 2015.

[17] B. P. Hyo Jung Park and S. S. Lee, "Radiomics and deep learning: Hepatic applications," *Korean journal of radiology*, vol. 21, no. 4, pp. 387–401, 2020.

[18] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts, "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[19] M. P. A. Starmans, S. R. van der Voort, T. Phil, M. J. M. Timbergen, M. Vos, G. A. Padmos, W. Kessels, D. Hanff, D. J. Grünhagen, C. Verhoef, S. Sleijfer, M. J. van den Bent, M. Smits, R. S. Dwarkasing, C. J. Els, F. Fiduzi, G. J. L. H. van Leenders, A. Blazevic, J. Hofland, T. Brabander, R. van Gils, G. J. H. Franssen, R. A. Feelders, W. W. de Herder, F. E. Buisman, F. E. J. A. Willemssen, B. Groot Koerkamp, L. Angus, A. A. M. van der Veldt, A. Rajicic, A. E. Odink, M. Deen, J. M. Castillo T, J. F. Veenland, I. Schoots, M. Renckens, M. Doukas, R. A. de Man, J. N. M. Ijzermans, R. L. Miclea, P. B. Vermeulen, E. E. Bron, M. G. Thomeer,

J. J. Visser, W. J. Niessen, and S. Klein, "Reproducible radiomics through automated machine learning validated on twelve clinical applications." , 2021.

[20] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, pp. 500–510, May 2018.

[21] L. Lu, Y. Zheng, G. Carneiro, and L. Yang, "Deep learning and convolutional neural networks for medical image computing," *Advances in computer vision and pattern recognition*, vol. 10, pp. 978–3, 2017.

[22] V. S. Parekh and M. A. Jacobs, "Deep learning and radiomics in precision medicine," *Expert Review of Precision Medicine and Drug Development*, vol. 4, no. 2, pp. 59–72, 2019. PMID: 31080889.

[23] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of Translational Medicine*, vol. 8, pp. 713–713, June 2020.

[24] C. Chen, C. Chen, M. Ma, X. Ma, X. Lv, X. Dong, Z. Yan, M. Zhu, and J. Chen, "Classification of multi-differentiated liver cancer pathological images based on deep learning attention mechanism," *BMC Medical Informatics and Decision Making*, vol. 22, July 2022.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[26] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016.

[27] M. J. A. Jansen, H. J. Kuijf, W. B. Veldhuis, F. J. Wessels, M. A. Viergever, and J. P. W. Pluim, "Automatic classification of focal liver lesions based on mri and risk factors," *PLOS ONE*, vol. 14, pp. 1–13, 05 2019.

[28] K. Sun, L. Shi, J. Qiu, Y. Pan, X. Wang, and H. Wang, "Multi-phase contrast-enhanced magnetic resonance image-based radiomics-combined machine learning reveals microscopic ultra-early hepatocellular carcinoma lesions," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 49, pp. 2917–2928, Mar. 2022.

[29] X. Liu, F. Khalvati, K. Namdar, S. Fischer, S. Lewis, B. Taouli, M. A. Haider, and K. S. Jhaveri, "Can machine learning radiomics provide pre-operative differentiation of combined hepatocellular cholangiocarcinoma from hepatocellular carcinoma and cholangiocarcinoma to inform optimal treatment planning?," *European Radiology*, vol. 31, pp. 244–255, Aug. 2020.

[30] C. A. Hamm, C. J. Wang, L. J. Savic, M. Ferrante, I. Schobert, T. Schlachter, M. Lin, J. S. Duncan, J. C. Weinreb, J. Chapiro, and B. Letzen, "Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI," *European Radiology*, vol. 29, pp. 3338–3347, Apr. 2019.

[31] C. J. Wang, C. A. Hamm, L. J. Savic, M. Ferrante, I. Schobert, T. Schlachter, M. Lin, J. C. Weinreb, J. S. Duncan, J. Chapiro, and B. Letzen, "Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features," *European Radiology*, vol. 29, pp. 3348–3357, May 2019.

[32] W. Jian, H. Ju, X. Cen, M. Cui, H. Zhang, L. Zhang, G. Wang, L. Gu, and W. Zhou, "Improving the malignancy characterization of hepatocellular carcinoma using deeply supervised cross modal transfer learning for non-enhanced MR," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, July 2019.

[33] P. M. Oestmann, C. J. Wang, L. J. Savic, C. A. Hamm, S. Stark, I. Schobert, B. Gebauer, T. Schlachter, M. Lin, J. C. Weinreb, R. Batra, D. Mulligan, X. Zhang, J. S. Duncan, and J. Chapiro, "Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver," *European Radiology*, vol. 31, pp. 4981–4990, Jan. 2021.

[34] R. Stollmayer, B. K. Budai, A. Tóth, I. Kalina, E. Hartmann, P. Szoldán, V. Bérczi, P. Maurovich-Horvat, and P. N. Kaposi, "Diagnosis of focal liver lesions with deep learning-based multi-channel analysis of hepatocyte-specific contrast-enhanced magnetic resonance imaging," *World Journal of Gastroenterology*, vol. 27, pp. 5978–5988, Sept. 2021.

[35] S. hui Zhen, M. Cheng, Y. bo Tao, Y. fan Wang, S. Juengpanich, Z. yu Jiang, Y. kai Jiang, Y. yu Yan, W. Lu, J. min Lue, J. hong Qian, Z. yu Wu, J. hong Sun, H. Lin, and X. jun Cai, "Deep learning for accurate diagnosis of liver tumor based on magnetic resonance imaging and clinical data," *Frontiers in Oncology*, vol. 10, May 2020.