POMERANIAN MEDICAL UNIVERSITY IN SZCZECIN

Faculty of Medicine with an English Program

Charlie Hamm M.D.

DISSERTATION

Development of an Interpretable Liver Tumor Diagnosis Tool using Deep Learning

for obtaining the academic degree *Doctor of Philosophy (PhD)* in medical sciences

Supervised by Prof. dr hab. n. med. Wojciech Poncyljusz

Szczecin 2020

The dissertation was prepared on the basis of a thematically coherent set of articles published in scientific journals in accordance with article 13.2 of the Act on Academic Degrees and Academic Title and Degrees and Title in Art, Dz. U. of 27th September 2017. Pos. 1789.

Parts of the work presented here have been published in:

2019

Hamm, Charlie A., Clinton J. Wang*, Lynn J. Savic, Marc Ferrante, Isabel Schobert, Todd Schlachter, MingDe Lin et al. "Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI." European radiology 29, no. 7 (2019): 3338–3347*

**= shared first authorship*
25 pt. Lista "A" MNiSW ; IF: 3.962

2019

Wang, Clinton J., Charlie A. Hamm*, Lynn J. Savic, Marc Ferrante, Isabel Schobert, Todd Schlachter, MingDe Lin et al. "Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features." European radiology 29, no. 7 (2019): 3348-3357.*

**= shared first authorship*
25 pt. Lista "A" MNiSW ; IF: 3.962

TABLE OF CONTENTS

1) ABBREVIATIONS

HCC = Hepatocellular carcinoma

RFA = Radiofrequency ablation

LI-RADS = Liver Imaging Reporting And Data System

CT = Computed tomography

MRI = Magnetic-resonance imaging

US = Ultrasound

BCLC = Barcelona Clinic Liver Cancer

DL = Deep learning

CNN = Convolutional neural network

STARD = Standards for Reporting of Diagnostic Accuracy guidelines

PACS = Picture archiving and communication system

FNH = Focal nodular hyperplasia

ICC = Intrahepatic cholangiocarcinoma

CRC = Colorectal carcinoma

SD = Standard deviation

Sn = Sensitivity

Sp = Specificity

PPV = Positive predictive value

AUC = Area under the curve

OPTN = Organ Procurement and Transplantation Network

NAFLD = Non-alcoholic fatty liver disease

2) INTRODUCTION

Hepatocellular carcinoma (HCC) is a rapidly growing global health problem, representing as it does the most common primary liver cancer and the third most common cause of cancer-related deaths worldwide [1-3]. The stratification and treatment planning of patients with HCC is a challenging task and often requires interdisciplinary co-operation of the clinicians on the tumor board. Radiological information, such as lesion entity, size and vascular involvement, play a pivotal role in the clinical decision-making for such patients [4,5]. Despite many proposed systems for staging and classification, there is currently no globally accepted approach for assessing HCC patients, and prognosis is often poor [6,7]. However, improved prognosis can be achieved when the diagnosis is made at an early stage of the disease, as curative-intent therapies (radiofrequency ablation (RFA), resection) are usually applicable for lesions smaller than 2 cm [8]. This underscores the clinical need for continuous advancements in imaging for early diagnosis of HCC.

A substantive contribution to radiological diagnosis could be made by the introduction of better standardization in the assessment of images and the reporting thereof. This would first of all decrease the potential for variation and for subjective factors in the interpretation of images, and thus in the errors that can ensue from these. Secondly, it would improve – in terms of both accuracy and speed – the communication of results to the clinicians involved. Finally, it would raise the standard of research and the reliability of quality-assurance procedures. The Liver Imaging Reporting and Data System (LI-RADS) was developed to provide a standardized analysis and reporting system for computed tomography (CT) and magnetic-resonance imaging (MRI) of patients at risk of developing HCC [9]. Improved quality and availability of oncological imaging, in combination with standardized reporting systems such as LI-RADS, has decreased the need for invasive biopsy of hepatic lesions larger than 2 cm, propelling imaging-based diagnosis to a more central position in diagnosis of HCC. While LI-RADS has changed the diagnostic workflow for malignant lesions and contributed to a higher quality in diagnosis and reporting [9-11], a majority of studies have shown at best moderate inter-observer agreement for LI-RADS categories [12-18]. In addition, biannual ultrasonography (US), despite its potentially impaired sensitivity in nodular cirrhotic livers, is generally recommended for the surveillance of patients at risk of HCC, facilitating detection at an early stage [19]. However, a recent study showed that MRI is the more cost-effective and sensitive modality in the detection of early-stage HCC in patients at risk [19]. At first glance MRI appears more expensive, but the high detection rate of very-early-stage HCC (Barcelona Clinic Liver Cancer (BCLC) stage 0) has been shown to increase the effectiveness of curative-intent treatment approaches and to engender a lower probability of HCC recurrence and mortality, thereby decreasing overall costs [19].

Given the unmet clinical need for improved HCC diagnosis and the improved soft-tissue contrast resolution of MRI, it is plausible that a deep learning (DL) system could extract hidden information and comprehensively analyze numerous features from MR images. This may lead to higher accuracy in staging and improved treatment planning for cancer patients. The majority of artificial-intelligence techniques in the field of medical imaging rely on training sets with manually defined features, limiting the model to predefined diagnostic patterns. Unlike those techniques, DL systems based on convolutional neural networks (CNNs) do not need any manually defined features to interpret images, and they may even uncover additional differential features not yet identified in current radiological practice [20]. As CNN-based DL systems have shown a potential to improve markedly the process of radiological diagnosis [21-24], there is room for a workflow that brings together the experience of practicing radiologists on the one hand and the computational power of artificial intelligence on the other, with a view to increasing primarily the quality and secondarily the efficiency of patient care. The potential for such a combination of human and computational resources has not yet been fully exploited in the field of HCC.

Although CNNs have demonstrated immense potential to enhance imaging-based diagnosis [23], their “black box” design has so far limited their adoption in clinical routine [25-27]. In their current form, CNNs cannot provide information about the factors used to arrive at predictions, and this in turn can prevent physicians from incorporating computational results into an informed decision-making process. The inability of CNNs to “explain their reasoning” also leads to a dearth of safeguards, and to a lack accountability when they fail. Interpretable DL systems that provide high-quality results in a more transparent manner would help to facilitate the migration of DL systems from the research unit into clinical practice.

3) PURPOSE OF THE STUDY

This study introduces the concept of a comprehensive interpretable DL system for liver tumor diagnosis based on magnetic-resonance images. The purpose of this study was to develop an interpretable deep learning system in which high accuracy was validated by comparison with radiologists’ findings and with a transparency that made it possible to “justify” its decisions to physicians.

4) MATERIALS AND METHODS

A description of the materials and methods used in this work were published in advance [28,29]. Thus, complete details of these can be found in the publications attached to this work.

This was a single-center, retrospective study compliant with the U.S. Health Insurance Portability and Accountability Act. The study design was in agreement with the Standards for Reporting of Diagnostic Accuracy guidelines (STARD). The study was approved by the institutional review board of the unit where the work was performed; informed consent was waived. The two components of the study involved (i) developing and validating a CNN-based liver-tumor classifier, followed by (ii) application of self-engineered algorithms to analyze specific hidden layers of this pre-trained CNN in a model-agonistic approach.

4.1 STUDY COHORT SELECTION

The picture archiving and communication system (PACS) was searched for abdominal MRI examinations between 2010 and 2017 depicting one of the following hepatic lesions: cavernous hemangioma, focal nodular hyperplasia (FNH), simple cyst, intrahepatic cholangiocarcinoma (ICC), colorectal cancer (CRC) metastasis and HCC. Owing to the limited availability of pathological proof, lesions were restricted to those demonstrating typical imaging characteristics. Moreover, additional diagnostic criteria were incorporated, to maximize the certainty of definite diagnosis. Typical imaging features, radiological-histopathological correlation and clinical data were criteria defining the “ground truth” utilized for each lesion type. Diagnosed lesions formally described by the radiology faculty in official reports were validated by another radiological reader according to diagnostic criteria defined for this study, and lesions presenting discrepancies between “ground truth” criteria and inclusion criteria were excluded. A detailed listing of these “ground truth” criteria used can be found in the supplementary material in the publications attached to this work (Tab. S1) [28], which also give further details on the inclusion and exclusion criteria (in the section of “Establishment of ‘ground truth’ cases”) [28].

4.2 MRI ACQUISITION PROTOCOL AND IMAGE PROCESSING

All MRI scans were performed on clinical 1.5 T or 3 T scanners. T1-weighted breath-hold sequences were used, with acquisition times of 12–18 seconds. After a bolus injection of macrocyclic gadolinium-based contrast agent, several post-contrast imaging series were obtained. Images were acquired at three time points after contrast-agent administration: late arterial phase (individually timed, but usually around 20 seconds after contrast injection), portal venous phase (~70 seconds after injection) and delayed venous phase (~3 min after injection). Between 2010 and 2017 several different MRI scanners and imaging protocols were used. However, although scanners and protocols may have differed in specific imaging parameters, the T1-weighted sequences used in this study met the purpose of the study.

Files associated with eligible MRI studies were downloaded from the PACS, and the images from each patient were re-evaluated by a radiological reader to confirm the reported diagnosis. If reference standard and inclusion criteria were fulfilled, then the location and size of a 3D bounding box around the target lesion were recorded manually.

The images were processed using code written in the programming language Python 3.5 (Python Software Foundation, Beaverton, Oregon, USA). Portal-vein and delayed-phase MRI studies were registered to the arterial phase by using affine registration with a mutual information metric. Images were cropped on the basis of the 3D bounding box to the lesion and its surrounding tissue, and cropped regions were then re-sampled to a resolution of 24×24×12 voxels (Fig. 2 in Hamm *et al.* [28]).

The data set comprised 494 lesions. Monte Carlo cross-validation was used for CNN training and testing. In each iteration of training and testing, 10 of the lesions in the data set were chosen at random from each class. Together, the 60 lesions chosen comprised 12% of the dataset. These 60 lesions were assigned to the test set, while the other 434 lesions were assigned to the training set. In order to increase the volume of training samples, images of the training set were augmented by a factor of ca. 100, giving 43,400 images in all. During augmentation, images underwent random scaling, rotation, translation and/or horizontal/vertical flipping. Data augmentation is an established machine-learning technique that allows a model to learn imaging features that are invariant to translation or rotation [30]. Phases were shifted randomly relative to each other to add robustness to imperfectly registered phases. The brightness and contrast of the image were also changed randomly.

4.3 DEEP LEARNING MODEL

For CNN model training a GeForce GTX 1060 (NVIDIA, Santa Clara, California, USA) graphics-processing unit was used. The model was built using Python 3.5 and Keras 2.2 (https://keras.io/) [31] running on a Tensorflow backend (Google, Mountain View, California, USA, https://www.tensorflow.org/). The CNN that was built comprised three convolutional layers, where the first layer had 64 convolutional filters for each of the 3 phases in the original image, and the other two had 128 filters across all phases. The model contained two maximum pooling layers (size 2×2×2 and 2×2×1 respectively), which is a standard deep-learning technique to facilitate learning. The final CNN comprised two fully connected layers, in which the first had 100 neurons while the second utilized a softmax output to six categories, corresponding to the six lesion types (Fig. 3 in Hamm *et al.* [28]). The CNN also used rectified linear units in conjunction with regularization techniques after convolutional and fully connected layers: this facilitates the learning of non-linear features and helps the model to generalize beyond the training set data respectively [30,32].

The selected imaging studies used for training and testing comprised a total of 296 patients, patient and imaging characteristics are displayed below (**Tab. 1** & Fig. 1 in Hamm *et al.* [28]). The training of the CNN was performed with an Adam optimizer [33], utilizing randomly chosen samples from each class from the training dataset. The model was then tested for its ability to classify correctly 60 lesions in the test set (10 from each lesion class). Overall, the model's performance was validated over 20 independent training iterations with different groupings of training and test sets, to yield a more accurate assessment.

Table 1: Patient and image characteristics. The 'total' column does not equal the sum of the rows because some MRI studies had more than one lesion type. (SD = standard deviation; adapted from Hamm *et al.* [28])

Patient characteristics	Cavernous hemangioma	FNH	Cyst	ICC	CRC metastasis	HCC	Total
Number of patients	49	53	37	36	39	88	296
Male Female	17 32	8 45	19 18	18 18	27 12	67 21	155 141
Age at imaging (mean ± SD)	50 ± 11	43 ± 11	62 ± 10	63 ± 14	61 ± 14	63 ± 8	57 ± 14
Image characteristics							
Number of MRI studies	50	57	42	49	44	96	334
Number of lesions	82	84	74	58	87	109	494
Lesion diameter (mm, mean ± SD)	25 ± 11.6	28.4 ± 20.7	21.7 ± 15.5	45 ± 16.8	26.4 ± 12.3	24.4 ± 10	27.5 ± 15.9

4.4 READER STUDY

Classification accuracy was compared between the CNN model and two board-certified radiologists (with respectively 39 and 7 years of experience), who did not take part in selecting the liver lesions used in this study. The reader study was conducted on an OsiriX MD (v.9.0.1, Pixmeo SARL, Switzerland, Geneva) workstation, with several differences as compared with an actual clinical setting. The reader study was performed on an anonymized dataset of 60 lesions (10 randomly selected from each class), and the radiologists were fully blinded to laboratory and clinical data, outcomes, demographics, any prior or follow-up imaging, and to any additional MRI sequences. The randomized test set was generated by using Monte Carlo cross-validation. In order to mimic the radiologists' "first exposure" to the MRI images and to compare their performance to the CNN, results of the reader study were compared after a single iteration. Each radiologist independently classified the same 60 lesions characterized by the model in the test dataset using the original three contrast-enhanced MRI phases. The performance of the radiologists was assessed in terms of (i) their ability to distinguish between the six liver-lesion types and (ii) their performance in respect of the three broader categories in which the application of a DL model to an HCC diagnostic imaging framework is simulated (here, LI-RADS; **Tab. 2**). The radiologists was instructed not to scroll the image beyond the upper and lower edges of the lesion, as this would have risked their noticing any other lesions present within the patient's liver, with the consequent introduction of a possible source of bias. The time taken by the radiologist to perform the assessment was noted; this began with the opening of the MRI phases and ended with the entry of the radiologist's classification of the lesion.

Table 2: Categories used in the reader study. Category 1, six individual lesion types (one out of six); Category 2, three broader categories in accordance to LI-RADS classes (one out of three)

Category 1: Lesion type	**Category 2**: Broader categories (LI-RADS classes)
Cysts	LR-1 (representing benign lesions)
Cavernous hemangiomas	
FNHs	
HCCs	LR-5 (HCC only)
ICCs	LR-M (non-HCC malignancy)
CRC metastases	

4.5 INTERPRETABILITY OF THE DEEP LEARNING MODEL

Full details of the technique of DL interpretability used in this study, with its *post hoc* probabilistic approach for analyzing hidden layers of a CNN, have been published [29]. Therefore, the following section provides only a brief description of this rather technical aspect of the study [29].

A set of fourteen imaging features was identified containing lesion-imaging characteristics that are useful for differentiating between various lesion types in T1-weighted triphasic contrast-enhanced MRI. For each feature, the training set was searched for hepatic lesions that best displayed each feature. Up to 20 example lesions were selected for each feature; this resulted in a total of 224 lesions used across the 14 radiological features. Also, a test set of 60 lesions was labelled with the most clearly dominant imaging features in each image (1-4 features per lesion). In the end, this test set was used for validation of the model's capabilities in feature extraction, and the test set was the same as that used to conduct the reader study described above.

For each radiological feature, ten example lesions were selected randomly from the 224 example lesions and passed through the CNN system, and the pre-activation outputs of the fully connected layer were examined. By comparing these neuronal outputs among the ten examples, each radiological feature was associated with specific patterns in these neurons. The test image was passed through the CNN to obtain its neuronal outputs, which were compared with the patterns of neuronal outputs that were associated with each feature. If the outputs were sufficiently similar to a feature's pattern, the CNN inferred that this feature was present in the test image. The CNN was tested for its ability to identify correctly the radiological features in the test set of 60 lesions. Performance was evaluated in 20 iterations with separately trained models_using different (though overlapping) choices of the ten example lesions. The voxels in the original image that contributed most to the presence of each feature identified were highlighted in feature maps by selecting voxels with the strongest positive correlations with the feature (as determined on the basis of the gradient of neurons in the fully connected layer with respect to the original image's voxels). The relative contribution of each identified feature to the classification of the lesion type was also evaluated (based on the Hessian of the objective function with respect to training examples that contained the feature of interest [34]). Further details of feature identification, mapping and scoring can be found in the supplementary information in the publication by Wang *et al.* [29] and the conference paper of our team [35].

4.6 STATISTICS

For the main analysis, the performance of the model was evaluated by Monte Carlo cross-validation, averaging the sensitivity (Sn), specificity (Sp) and overall accuracy over 20 iterations. With regard to the validation of the CNN by radiological readings, the performances of the model and the radiologists were compared by evaluating their Sn, Sp and overall accuracy on the same single randomly selected test set. In order to compare the model's and radiologists' performance in identifying HCC masses, a receiver operating characteristic curve was plotted. The performance of the model in image-feature extraction and identification was assessed by calculating the positive predictive value (PPV), Sn, precision and recall.

5) RESULTS

The results of this work have been published in advance [28,29], and copies of the publications are attached to this thesis.

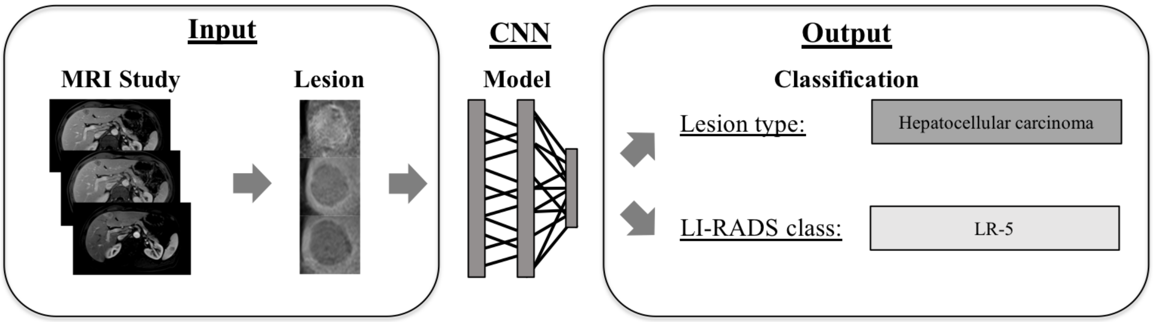
5.1 DEEP LEARNING MODEL

The DL system showed an average test accuracy of 91.9 ± 2.9% (1103/1200) and 94.3% ± 2.9% (1131/1200) among individual lesions and across the three broader categories respectively. The initial training of the CNN took 29 ± 4 minutes. Once the training was completed, the actual run time needed to classify each lesion in the test set was 5.6 ± 4.6 milliseconds. The Sn and Sp achieved by the DL system across the six lesion classes as well as for the three LI-RADS-derived classes is displayed below (**Tab. 3**). The overall accuracy and run times of the model classification are displayed in the Table 3 of Hamm *et al.* [28], which is attached to this work. The workflow of lesion classification by the CNN is illustrated below (**Fig. 1**).

Table 3: Model and radiologist performance metrics for individual lesion types and LI-RADS classes. (Adapted from Hamm *et al.* [28])

	Average of 20 iterations		**Reader study**					
	Model test set		Model		Radiologist 1		Radiologist 2	
	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
Lesion type								
Cavernous hemangioma	91%	99%	100%	100%	100%	96%	100%	94%
FNH	91%	98%	90%	96%	90%	98%	90%	94%
Cyst	99%	100%	100%	100%	90%	96%	100%	98%
ICC	90%	97%	60%	100%	80%	94%	90%	100%
CRC metastasis	89%	98%	100%	94%	50%	92%	70%	96%
HCC	94%	98%	90%	98%	70%	100%	60%	100%
Overall	92%	98%	90%	98%	80%	96%	85%	97%
Derived LI-RADS class								
LR-1 (*n* = 30)	94%	96%	97%	93%	97%	87%	100%	80%
LR-5 (*n* = 10)	94%	98%	90%	98%	70%	100%	60%	100%
LR-M (*n* = 20)	95%	96%	95%	100%	85%	93%	85%	98%
Overall	94%	97%	95%	96%	88%	91%	88%	89%

Figure 1: Workflow of lesion classification by the CNN in the example of HCC classification.



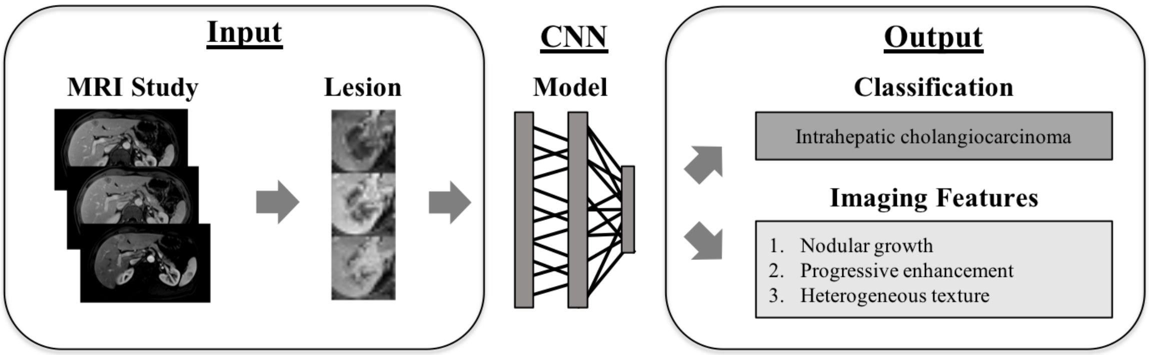
5.2 READER STUDY

In the reader study (described above), the lesions could be classified. The model yielded a mean accuracy of 90% (55/60 lesions), while the two radiologists assessing the same lesions achieved respective accuracies of 80% (48/60) and 85% (51/60). For the three broader categories, the model gave an accuracy of 92% (58/60), against an accuracy of 88% (53/60) for each of the two radiologists. The Sn and Sp across the six lesion types and three broader categories achieved by the CNN and the radiologists in the reader study are given above (**Tab. 3**). The total time required for analyzing each lesion was 0.8 milliseconds for the classification model versus 14 ±10 seconds and 17 ±24 seconds for the radiologists. Additionally, the performance of the model in HCC classification was investigated by plotting a receiver operating characteristic curve. The DL system achieved an area under the curve (AUC) of 0.992 with a high sensitivity at the cost of a few false positives (Sn = 90%, false-positive rate = 2%; Fig. 4 in Hamm *et al.* [28]).

5.3 INTERPRETABILITY OF THE DEEP LEARNING MODEL

A total of 224 annotated images were used across the 14 radiological features, and some images were labelled with up to 4 features. After being presented with a random subset of these examples, the model obtained a PPV of 76.5 ± 2.2% (2553/3339) and an Sn of 82.9 ± 2.6% (2553/3080) in identifying the 1–4 correct radiological features for the 60 manually labelled test lesions over 20 iterations. The workflow of lesion classification and imaging feature identification by the CNN is illustrated below (**Fig. 2**).

Figure 2: Workflow of lesion classification and imaging-feature extraction by the CNN in the example of ICC classification.



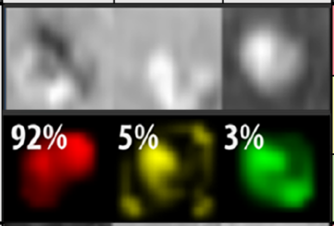
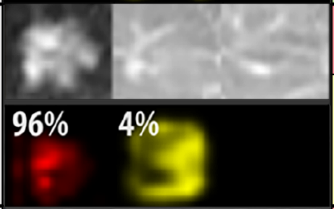
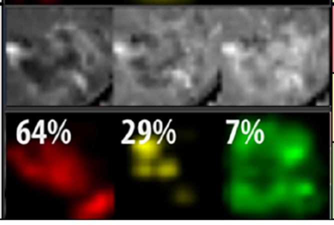
In its assessment of individual features, the CNN performed best for the simpler enhancement patterns. Presented with 2.6 labelled features on average per lesion, its performance was as summarized in **Tab. 4**. For simpler image features (e.g. arterial-phase hyperenhancement, hyperenhancing mass on delayed phase, thin-walled mass), the CNN's performance was good; for more complex ones (e.g. nodularity, infiltrative appearance) it was less so, and the central-scar frequency was grossly overestimated, as there was only one such among the 60 lesions in the test set.

Table 4: Recognition of enhancement pattern by the model over 20 iterations. The PPV and Sn of six example imaging features are shown.

Overall precision	76.5 ± 2.2% (recall = 82.9 ± 2.6%)	
Misclassified lesions	144/1200 (12%)	
	PPV	Sn
Arterial-phase hyperenhancement	91.2% = 343/376	90.3% = 343/380
Hyperenhancing mass on delayed phase	93.0% = 160/172	100% = 160/160
Thin-walled mass	86.5% = 160/185	100% = 160/160
Nodularity	62.9% = 73/116	60.8% = 73/120
Infiltrative appearance	33.0% = 36/109	45.0% = 36/80
Frequency of central scars	32.0% = 16/50	80.0% = 16/20
All features, misclassified lesions only	56.6% = 259/458	63.8% = 259/406

In classifying the lesion type, the CNN model put greater weight on radiological features that appeared more prominent in the image (**Fig. 3**). Hyperenhancing mass in delayed phase was a clearly observed imaging feature in the cavernous hemangioma example, receiving a relevance score of 92%. Arterial-phase hyper-enhancement was likewise clearly seen in the FNH example, and it received a relevance score of 96%. In some of the features with low relevance scores, the feature map was less well defined. For example, heterogeneous lesion of the ICC was assigned a relevance score of 7%, and had a very diffuse feature map. Further details of the mapping of radiological features and their relevance can be found in the supplementary material of the study publication attached to this thesis [29].

Figure 3: 2D slices of the feature maps and relevance scores for the examples of cavernous hemangioma, FNH and ICC with correctly identified features.

Lesion Class	Contrast-enhanced T1w MRI			Feature Relevance	Features identified by the model
	Arterial Phase	Venous Phase	Delayed Phase		
Caverous hemangioma				92%	Hyperenhancing mass in delayed phase
				5%	Nodular peripheral enhancement
				3%	Progressive centripetal filling
FNH				96%	Arterial phase hyperenhancement
				4%	Isointensity in venous/delayed phase
ICC				64%	Progressive hyperenhancement
				29%	Nodularity
				7%	Heterogeneous lesion

6) DISCUSSION

This study demonstrates the development of a proof-of-concept "interpretable" deep learning system for the classification of liver lesions from multiphase contrast-enhanced MRI. In addition to making high-accuracy predictions, this system was found to be capable of justifying its decisions by automatically identifying, mapping and scoring radiological features. The system outperformed radiologists in distinguishing six lesion classes (model accuracy 90%, radiologist accuracies 80% and 85%), as well as in classifying lesions into three broader categories representing the LI-RADS classes for benign, HCC and malignant non-HCC lesions (model accuracy ~92%, radiologist accuracies ~88%), with a classification time of one millisecond per lesion.

Previous studies have demonstrated CNN-based classification of liver lesions on single 2D imaging slices using CT or US imaging [36-38], and this study builds on these approaches by classifying focal liver lesions on the basis of the reference standard of contrast-enhanced MRI. The improved soft-tissue contrast resolution inherent to MRI can enable DL systems to capture a wider variety of imaging features, contributing to superior diagnostic performance. Additionally, the heterogeneity of HCC lesions makes imaging-based diagnosis and staging especially challenging [6,39]. A volumetric approach using 3D data sets may lead to improved detection of enhancement patterns or inhomogeneous growth that may be relevant for lesion classification, while removing the model's dependence upon manual slice selection (and consequent variability) [40]. To take further advantage of available imaging data, the present study introduces a DL system that interprets 3D volumes around each lesion. Moreover, previously published studies have laid the foundation for computational classification of hepatic lesion types by grouping different lesion entities into three to five classes [36-38]. However, when future clinical implementation is considered, it is clear that the challenge of classification becomes increasingly hard to meet when lesions are not grouped. For this, more differential features must be learned, and the chance of achieving the correct classification decreases. The present study included six ungrouped hepatic lesion types, showing high accuracy (~92%). As anticipated, a higher overall accuracy (~94%) was reached with three grouped classes (LR-1, LR-5 and LR-M). In this case, there is no penalty for mistaking slowly filling cavernous hemangiomas for cysts, or for confusing nodular ICCs with CRC metastases. In addition, the heterogeneous imaging protocol (imaging studies from 2010–2017) and the inclusion of previously treated lesions demonstrate the robustness of the DL system toward applications where inhomogeneous data sets and variable lesion appearances are present.

As the strength of DL systems become particularly visible when their performance is compared with that of experienced clinicians, reader studies have become an established tool to investigate their performance and clinical value. The DL system in this study demonstrated a high Sn for CRC metastases and for HCC in comparison with the Sn achieved by radiologists. HCCs with faint enhancement or with unclear washout were prone to be misclassified by radiologists as CRC metastases or FNHs, respectively. In contrast, the improved Sn for identifying HCCs suggests that the CNN could more reliably utilize yet unknown imaging features for classification. It is of note that the diagnostic accuracy of the radiologists might have matched or exceeded the accuracy of the DL system if they had been given access to diagnostically relevant clinical information or other imaging sequences. However, the moderate sensitivity and excellent specificity attained for HCC diagnosis by the radiologists in the reader study match the results of a recent study investigating the performance of LI-RADS for the diagnosis of HCC [41], indicating that radiologists are more likely to miss the diagnosis of HCC if classical imaging features are somewhat ambiguous. As this study only included typical HCCs appearing according to the Organ Procurement and Transplantation Network (OPTN) criteria, it can be hypothesized that radiologists underestimate imaging features if no clinical data are available on the underlying condition. The application of standardized reporting systems, such as LI-RADS, is only targeted for an at-risk population presented with cirrhosis, chronic hepatitis B virus infection without cirrhosis, or current or prior HCC, including liver transplant recipients [42]. However, non-alcoholic fatty liver disease (NAFLD) has emerged as the leading cause of chronic liver disease in most regions of the world, and it is the fastest-growing cause of HCC-related transplants in the United States [43,44]. Furthermore, among new HCC cases without advanced fibrotic liver changes in the United States, NAFLD constitutes the largest etiological proportion of cases [43]. This entity poses an additional challenge to clinical practice paradigms based on HCC risk [43], and it highlights the need for reliable detection and extraction of imaging features within the lesion, despite underlying liver conditions which could bias the predictions of radiologists. The results of the reader study suggest that DL systems may be able to analyze imaging features within a lesion efficiently and possibly even make use of lesion characteristics that are unrecognized by the radiologist.

Good diagnostic performance of the DL system indicates the possibility that CNNs can potentially be utilized as a quick and reliable "second opinion" for a radiologist in the diagnostic workup of focal liver lesions, helping to reduce inter-reader variability and difficulties in interpretation when radiological features are unclear or obscure. However, where patient diagnosis and treatment planning is concerned, it is unlikely that clinicians will accept an automated assessment if they cannot understand the algorithm's reasoning. The method of scoring radiological features, as presented here, allows the algorithm to communicate how it arrives at its conclusion. With this, the referring radiologist can check quickly whether the DL system has detected features of the lesion correctly, by comparing the feature map with the lesion on the actual MRI image. The radiologist is thereby able to verify that detected imaging features correlate with the correct location in the lesion, and to exclude predictions based on incorrectly identified imaging features.

The DL system was able to identify the majority of radiological features consistently, despite being provided with only 10 example lesions per class. Nonetheless, this study has demonstrated that the CNN had growing difficulty in identifying features correctly as the complexity of these features increased. The presence, location, and relevance for classification of simple imaging features – such as hypoenhancing or hyperenhancing masses – were determined reliably and accurately by the CNN, whereas the model performed worse in lesions with imaging features that consisted of patterns over several phases (such as washout or centripetal filling). In particular, the model struggled on more complex features, such as infiltrative appearance, that may appear quite variable across different lesions, suggesting either that more examples of these features are required for training or that these features are not sufficiently well defined by the CNN. Even so, there was a clear correlation between the CNN's misclassifications of the lesion entity and its incorrect identification of radiological features. This could in the future provide clinicians, research workers and other relevant parties with sufficient transparency to make them aware of when – and, importantly, why– the CNN model has failed in individual cases.

As shown in the results on feature relevance **(Fig. 3)**, the model tends to place greater weight on imaging features that have greater uniqueness and differential diagnostic power in the respective lesion class. The method of scoring the relevance of single imaging features enables the interpretable DL system to be utilized as a tool for the validation of imaging guidelines, particularly for entities which are uncommon or have evolving imaging criteria, such as bi-phenotypic tumors and ICCs [11,45,46]. One approach to this might be to present the DL system initially with a large set of candidate imaging features. The features with the highest relevance scores output by the model would then be selected. This would enable one to find out which features have the greatest relevance for members of a given lesion class. This would appear to be especially applicable in HCC diagnosis, as the majority of inter-reader studies have demonstrated an – at best – moderate level of reliability in determining LI-RADS classes [12-17], and the rigidity and complexity of LI-RADS constitutes a major barrier for broad adoption [16,47].

Recent studies have also highlighted issues regarding the application of LI-RADS ancillary features, which are recommended for category adjustment, improved detection, and increased confidence in diagnosis [41,48]. However, these features are based primarily upon a combination of retrospective single-center studies, on biological plausibility and on expert opinion with a somewhat low level of evidence [47,48]. Here again, this problem could be tackled with the help of an interpretable DL system; this would allow an approach to the numerous ancillary imaging features specified in the LI-RADS guidelines, in that it would provide information on the relative importance of the diverse radiological features that go into a differential diagnosis. The CNN might, for example, find application in the validation of additional ancillary features suggested as being of relevance, and in charting their frequency of occurrence by application to a large patient cohort and subsequent analysis of the predictions generated by the CNN. Features found only to have a low frequency, or considered to be of little relevance, could thus be considered for exclusion from the LI-RADS guidelines. An approach of this kind could be a stepping-stone on the path toward the generation of a protocol that could make diagnosis more efficient and more practical in clinical routine [12,16]. Furthermore, the interpretable DL system classified lesions reliably as being 'benign', 'HCC' or 'malignant non-HCC' (roughly corresponding to LR-1, LR-5 and LR-M, respectively) with an accuracy of 94.3%. This DL model could interface with standardized reporting systems by the calculation of an average probability of the finding ‘HCC’ based on the model’s prediction *and* the diagnosis by the radiologist, in order to score lesions that are suspicious for HCC but that lack a definite benign or malignant appearance (i.e. LR-2/3/4). Such shared decision-making would help address the recently indicated need for simplification of LI-RADS in order to integrate it into the radiologist’s normal workflow [47].

The clinical management of patients with liver malignancies depends greatly on radiology reports, which may include vague descriptions and may depend substantially on the experience of the referring radiologist. In a DL system-supported diagnosis, the radiologist could use data on lesion classification and extracted imaging features provided by the DL system, thus supporting his subjective interpretation by adducing quantitative data, as the training of DL systems generally comprises several hundred exemplary lesions. In addition, once the DL system has reached high accuracy levels, it analyses any lesion presented according to a predefined algorithm. Thus DL systems can contribute with quantitative data to more evidence-based radiology reports, leading to higher reproducibility and diagnostic confidence. As opposed to many other malignancies, HCC incidence rates continue to rise [49], which may be expected to result in a continued trend of increasing imaging volumes, requiring more rapid and more reliable techniques for detecting and diagnosing HCC. In addition, emerging risk factors such as NAFLD, diabetes and obesity may challenge the present-day diagnostic frameworks for HCC [43]. Highlighted by the high accuracy in lesion classification and extracted imaging features supporting the prediction, DL systems could support radiologists with reproducible quantitative data and thereby help clinicians to diagnose focal liver lesions earlier and with greater confidence.

As the present study was designed as a proof-of-concept study, there are several limitations that a future multi-center study should address before clinical integration of DL can be considered. As this was a retrospective single-center investigation, only a limited number of imaging studies were available for each class. Thus, only lesions with typical appearance in MRI were used, excluding more complex lesions such as infiltrative HCC subtypes or complicated cysts. Additionally, LI-RADS is only applicable to patients at high risk of HCC, and this study included many lesions in livers without cirrhosis or hepatitis B viral infection, so that the results do not necessarily reflect “real life” performance within an HCC diagnostic framework. Because diagnoses such as FNH or CRC metastasis are much less common in cirrhotic livers, limiting the cohort to cirrhotic patients would have severely reduced the dataset. Yet, as mentioned above, NAFLD is the fastest-growing cause of HCC-related transplantation in the United States and constitutes the largest etiological proportion of cases among new HCC cases without advanced fibrosis or cirrhosis [43], suggesting that the current LI-RADS diagnostic framework will have to be adjusted. Additionally, as this was a retrospective study, with data from a limited number of patients at a single institution, the requisite pathological "ground truth" diagnosis was only available for a restricted number of the study lesions. Thus, this study used only lesions of "typical" appearance, and "ground truth" criteria were carefully selected and defined (**Tab. S1** in Hamm *et al.* [28]). In the case of lesions for which no pathological diagnosis was available, this was replaced by the result of an analysis covering all the accessible image material (T1 pre-contrast, T2 etc.) and all the available clinical data. However, this additional image material was not used in the model training or in the reader study. Therefore, their contribution to the CNN model's performance will have to be assessed in further studies. A further limitation of this study was that the readers had no access to additional information such as clinical data, knowledge of disease progression, or evidence of prior surgery, which a radiologist would utilize in daily practice. Therefore, for such lesions, it is not unreasonable for discrepancies to occur in this study between the "ground truth" and the reader's classification. In the context of these limitations, this approach and selected reference standards were appropriate for the study's purposes of developing a proof-of-concept prototype from available data at a single large academic medical center. Furthermore, there is no established ground truth for describing feature maps or relevance. Therefore, future studies will be designed to demonstrate similar functionality using different choices of radiological features and lesion types, also taking into account the reproducibility of such techniques under different DL models. These limitations should be addressed in the future through progressive refinements with multi-institutional data registries, utilizing larger and more diverse input data and a more complex CNN model capable of analyzing other types of MRI sequences.

7) CONCLUSION

In summary**, this study presents the development of an "interpretable" DL system prototype that exceeds the accuracy of radiologists in classifying hepatic lesions in contrast-enhanced MRI, while allowing insight into the algorithm's decision-making.** As comprehensibility and transparency are key barriers towards the practical integration of DL in clinical practice [50], the interpretable DL system presented here demonstrates its potential as a decision-support tool in liver lesion diagnosis; however, the clinical impact of the decision-support tool needs to be validated in a prospective study before it can be considered for integration into clinical practice.

8) ENGLISH ABSTRACT

Objectives

The purpose of this study was (i) to develop an interpretable deep learning system, of high accuracy, for classifying hepatic lesions in contrast-enhanced MRI, with a transparency that allows justification of its decisions to physicians and (ii) to validate this system by comparison of its diagnostic performance with that of radiologists.

Methods

This study included 296 patients with 494 hepatic lesions in six categories. Lesions were identified by multiphasic MRI and divided into training (n=434) and test (n=60) sets. Established image augmentation techniques were used to increase the number of training samples to 43,400. This training set was input to a custom-made convolutional neural network (CNN), consisting of three convolutional layers with associated rectified linear units, two maximum pooling layers, and two fully connected layers. An Adam optimizer was used for model training. Additionally, up to four key imaging features per lesion were assigned to a subset of each lesion class and a post-hoc algorithm was used to infer the presence of these features in a test set on the basis of activation patterns of the (trained) CNN model. Validation of the CNN was performed by comparing the diagnostic performance of the CNN with that of two board-certified radiologists. This was carried out by Monte Carlo cross-validation, and the CNN's performance on an identical unseen test set was compared with that of the radiologists. Feature maps highlighting regions in the original image that corresponded to particular features were generated. A relevance score was then assigned to each feature identified, denoting the relative importance of the feature for the predicted lesion classification.

Results

The interpretable deep learning (DL) system demonstrated a 92% sensitivity (Sn), a 98% specificity (Sp), and a 92% accuracy. Test set performance in a single run showed an average 90% Sn and 98% Sp across the six lesion types, compared with an average 82.5% Sn and 96.5% Sp for radiologists, respectively. Radiologists achieved an Sn of 60%–70% for classifying hepatocellular carcinoma, while the DL system achieved an Sn of 90%. For the specific case of HCC classification the CNN achieved a receiver operating characteristic area under the curve of 0.992. Computation time per lesion was 5.6 milliseconds.

The positive predictive value and the Sn in identifying the correct radiological features present in each test lesion were 76.5% and 82.9%, respectively, while 12% of the lesions were misclassified; these misclassified lesions led more often to wrongly identified features than the correctly classified ones did (60.4% vs. 85.6%). Original image voxels contributing to each imaging feature were consistent with the feature maps generated, and in each class the most prominent imaging criteria were reflected by their respective feature relevance scores.

Conclusion

This study presents the development of an "interpretable" DL system prototype, the accuracy of which exceeds that of radiologists in classifying hepatic lesions on contrast-enhanced MRI, while illuminating the algorithm's decision-making. The interpretable DL system presented demonstrates potential as a decision-support tool in liver lesion diagnosis; however, the clinical impact of the decision-support tool needs to be validated in a prospective study before the tool can be considered for integration into clinical practice.

9) POLISH ABSTRACT

Cele

Celami tego badania było:

(i) opracowanie wysokiej dokładności systemu głębokiego uczenia się do oceny i klasyfikacji zmian w wątrobie przy pomocy rezonansu magnetycznego z kontrastem, z możliwością oceny podjętej decyzji przez lekarza oraz,

(ii) walidacja tego systemu przez porównanie jego wyników diagnostycznych z wynikami uzyskanymi przez lekarzy radiologów.

Metody

Badanie objęło 296 pacjentów z 494 zmianami chorobowymi wątroby podzielonymi na sześć kategorii. Zmiany zostały zidentyfikowane za pomocą wielofazowego MRI i podzielone na zestawy treningowe (n=434) i testowe (n=60). Dostępne techniki multiplikacji obrazu zostały wykorzystane w celu zwiększenia liczby próbek szkoleniowych do 43 400. Ten zestaw szkoleniowy wprowadzono do stworzonej na zamówienie sieci neuronów konwulsyjnych (CNN), składającej się z trzech warstw konwulsyjnych z powiązanymi prostymi jednostkami liniowymi, dwóch maksymalnych warstw zbiorczych oraz dwóch w pełni połączonych warstw. Optymalizator Adama został użyty w celu szkolenia. Maksymalnie cztery kluczowe cechy na każdą zmianę chorobową zostały przypisane do podzbioru każdej klasy zmiany, dodatkowo

zastosowano algorytm post-hoc do wnioskowania o obecności tych cech w zestawie testowym na podstawie wzorów aktywacji (wytrenowanego) modelu CNN. Walidacja CNN została przeprowadzona poprzez porównanie wyników diagnostycznych CNN z wynikami dwóch specjalistów radiologii. Zostało to przeprowadzone w ramach walidacji krzyżowej Monte Carlo, a wyniki CNN na identycznym, niewidocznym zestawie testowym zostały porównane z wynikami radiologów. Wygenerowane zostały mapy cech wyróżniające regiony na oryginalnym obrazie, które odpowiadały poszczególnym cechom. Następnie do każdej zidentyfikowanej cechy przypisano ocenę istotności, oznaczającą względne znaczenie danej cechy dla przewidywanej klasyfikacji zmiany.

Wyniki

Interpretowalny system głębokiego uczenia się (DL) wykazał 92% czułości (Sn), 98% specyficzności (Sp) i 92% dokładności. Wydajność zestawu testowego w pojedynczym badaniu wykazała średnio 90% Sn i 98% Sp w sześciu typach zmian, w porównaniu do średnio 82,5% Sn i 96,5% Sp dla radiologów. Radiolodzy uzyskali Sn na poziomie 60%-70% do klasyfikacji raka wątrobowokomórkowego, natomiast system DL uzyskał Sn na poziomie 90%. Dla szczególnego przypadku klasyfikacji HCC CNN uzyskał obszar charakterystyki pracy pod krzywą 0,992. Czas obliczeniowy na jedną zmianę wynosił 5,6 milisekundy.

Dodatnia wartość predykcyjna i Sn w identyfikacji prawidłowych cech radiologicznych występujących w każdej badanej zmianie wynosiły odpowiednio 76,5% i 82,9%, podczas gdy 12% zmian było źle sklasyfikowanych; te źle sklasyfikowane zmiany częściej prowadziły do błędnej identyfikacji cech niż prawidłowo sklasyfikowane (60,4% vs 85,6%). Oryginalne woksle obrazowe przyczyniające się do każdej funkcji obrazowania były spójne z wygenerowanymi mapami cech, a w każdej klasie najbardziej znaczące kryteria obrazowania były odzwierciedlone przez ich odpowiednie oceny istotności cech.

Wniosek

W pracy przedstawiono rozwój "interpretowalnego" prototypu systemu głębokiego uczenia się, którego dokładność przewyższa dokładność radiologów w klasyfikowaniu zmian w wątrobie na MRI wzmocnionym kontrastem, przy jednoczesnym oświetleniu procesu podejmowania decyzji przez algorytm. Przedstawiony interpretacyjny system DL wykazuje potencjał jako narzędzie wspomagające podejmowanie decyzji w diagnostyce zmian chorobowych w wątrobie; jednakże wpływ kliniczny narzędzia wspomagającego podejmowanie decyzji musi być zweryfikowany w badaniu prospektywnym, zanim będzie można rozważyć jego włączenie do praktyki klinicznej.

10) REFERENCES

1 El-Serag, H. & Rudolph, K. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology* **132**, 2557 (2007).

2 Wang, H. *et al.* Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* **388**, 1459-1544 (2016).

3 Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **68**, 394-424 (2018).

4 Bruix, J. *et al.* Clinical decision making and research in hepatocellular carcinoma: pivotal role of imaging techniques. *Hepatology* **54**, 2238-2244 (2011).

5 Llovet, J. M., Brú, C. & Bruix, J. in *Seminars in liver disease.* 329-338 (© by Thieme Medical Publishers, Inc.).

6 Huo, T. I., Liu, P. H. & Hsu, C. Y. Staging and restaging for hepatocellular carcinoma: Solution of confusion? *Hepatology* (2018).

7 Serper, M. *et al.* Association of provider specialty and multidisciplinary care with hepatocellular carcinoma treatment and mortality. *Gastroenterology* **152**, 1954-1964 (2017).

8 Llovet, J. M., Bru C Fau - Bruix, J. & Bruix, J. Prognosis of hepatocellular carcinoma: the BCLC staging classification.

9 Mitchell, D. G., Bruix, J., Sherman, M. & Sirlin, C. B. LI-RADS (Liver Imaging Reporting and Data System): summary, discussion, and consensus of the LI-RADS Management Working Group and future directions. *Hepatology* **61**, 1056-1065, doi:10.1002/hep.27304 (2015).

10 Corwin, M. T. *et al.* Nonstandardized Terminology to Describe Focal Liver Lesions in Patients at Risk for Hepatocellular Carcinoma: Implications Regarding Clinical Communication. *AJR. American journal of roentgenology* **210**, 85-90, doi:10.2214/ajr.17.18416 (2018).

11 Mitchell, D. G., Bashir, M. R. & Sirlin, C. B. Management implications and outcomes of LI-RADS-2, -3, -4, and -M category observations. *Abdominal radiology (New York)* **43**, 143-148, doi:10.1007/s00261-017-1251-z (2018).

12 Barth, B. *et al.* Reliability, Validity, and Reader Acceptance of LI-RADS-An In-depth Analysis. *Academic radiology* **23**, 1145 (2016).

13 Bashir, M. *et al.* Concordance of hypervascular liver nodule characterization between the organ procurement and transplant network and liver imaging reporting and data system classifications. *Journal of magnetic resonance imaging: JMRI* **42**, 305 (2015).

14 Davenport, M. S. *et al.* Repeatability of Diagnostic Features and Scoring Systems for Hepatocellular Carcinoma by Using MR Imaging. *Radiology* **272**, 132 (2014).

15 Ehman, E. C. *et al.* Rate of observation and inter-observer agreement for LI-RADS major features at CT and MRI in 184 pathology proven hepatocellular carcinomas. *Abdominal radiology (New York)* **41**, 963-969, doi:10.1007/s00261-015-0623-5 (2016).

16 Fowler, K. J. *et al.* Interreader Reliability of LI-RADS Version 2014 Algorithm and Imaging Features for Diagnosis of Hepatocellular Carcinoma: A Large International Multireader Study. *Radiology* **286**, 173-185, doi:10.1148/radiol.2017170376 (2018).

17 Liu, W. *et al.* Accuracy of the diagnostic evaluation of hepatocellular carcinoma with LI-RADS. *Acta radiologica (Stockholm, Sweden : 1987)*, 284185117716700, doi:10.1177/0284185117716700 (2017).

18 Zhang, Y. D. *et al.* Classifying CT/MR findings in patients with suspicion of hepatocellular carcinoma: Comparison of liver imaging reporting and data system and criteria-free Likert scale reporting models. *Journal of Magnetic Resonance Imaging* **43**, 373-383 (2016).

19 Kim, H. L. *et al.* Magnetic Resonance Imaging Is Cost-Effective for Hepatocellular Carcinoma Surveillance in High Risk Patients with Cirrhosis. *Hepatology*, doi:10.1002/hep.30330 (2018).

20 Greenspan, H., Van Ginneken, B. & Summers, R. M. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging* **35**, 1153-1159 (2016).

21 Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A. & Mougiakakou, S. Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE transactions on medical imaging* **35**, 1207-1216, doi:10.1109/TMI.2016.2535865 (2016).

22 Chartrand, G. *et al.* Deep learning: a primer for radiologists. *Radiographics : a review publication of the Radiological Society of North America, Inc* **37**, 2113-2131 (2017).

23 Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* **1**, e271-e297 (2019).

24 McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89-94 (2020).

25 Dayhoff, J. E. & DeLeo, J. M. Artificial neural networks: opening the black box. *Cancer: Interdisciplinary International Journal of the American Cancer Society* **91**, 1615-1635 (2001).

26 Kiczales, G. Beyond the black box: open implementation. *IEEE Software* **13**, 8, 10-11, doi:10.1109/52.476280 (1996).

27 Olden, J. D. & Jackson, D. A. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* **154**, 135-150, doi:https://doi.org/10.1016/S0304-3800(02)00064-9 (2002).

28 Hamm, C. A. *et al.* Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. *European radiology* **29**, 3338–3347 (2019).

29 Wang, C. J. *et al.* Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. *European radiology* **29**, 3348-3357 (2019).

30 Krizhevsky, A., Sutskever, I. & Hinton, G. E. in *Advances in neural information processing systems.* 1097-1105.

31 Chollet, F. (https://keras.io, 2015).

32 Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

33 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

34 Koh, P. W. & Liang, P. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730* (2017).

35 Wang, C. J., Hamm, C. A., Letzen, B. S. & Duncan, J. S. in *Medical Imaging 2019: Computer-Aided Diagnosis.* 109500U (International Society for Optics and Photonics).

36 Acharya, U. R. *et al.* Automated diagnosis of focal liver lesions using bidirectional empirical mode decomposition features. *Comput Biol Med* **94**, 11-18, doi:10.1016/j.compbiomed.2017.12.024 (2018).

37 Hwang, Y. N., Lee, J. H., Kim, G. Y., Jiang, Y. Y. & Kim, S. M. Classification of focal liver lesions on ultrasound images by extracting hybrid textural features and using an artificial neural network. *Bio-medical materials and engineering* **26**, S1599-S1611 (2015).
38 Yasaka, K., Akai, H., Abe, O. & Kiryu, S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology*, 170706 (2017).
39 Friemel, J. *et al.* Intratumor heterogeneity in hepatocellular carcinoma. *Clinical Cancer Research* **21**, 1951-1961 (2015).
40 Chapiro, J. *et al.* Radiologic-Pathologic Analysis of Contrast-enhanced and Diffusion-weighted MR Imaging in Patients with HCC after TACE: Diagnostic Accuracy of 3D Quantitative Image Analysis. *Radiology* **273**, 746-758, doi:10.1148/radiol.14140033 (2014).
41 Kierans, A. S. *et al.* Validation of Liver Imaging Reporting and Data System 2017 (LI-RADS) Criteria for Imaging Diagnosis of Hepatocellular Carcinoma. *Journal of Magnetic Resonance Imaging* (2018).
42 Cerny, M. *et al.* LI-RADS version 2018 ancillary features at MRI. *Radiographics : a review publication of the Radiological Society of North America, Inc* **38**, 1973-2001 (2018).
43 Kulik, L. & El-Serag, H. B. Epidemiology and Management of Hepatocellular Carcinoma. *Gastroenterology*, doi:10.1053/j.gastro.2018.08.065 (2018).
44 Maurice, J. & Manousou, P. Non-alcoholic fatty liver disease. *Clin Med (Lond)* **18**, 245-250, doi:10.7861/clinmedicine.18-3-245 (2018).
45 Narsinh, K. H., Cui, J., Papadatos, D., Sirlin, C. B. & Santillan, C. S. Hepatocarcinogenesis and LI-RADS. *Abdominal radiology (New York)* **43**, 158-168, doi:10.1007/s00261-017-1409-8 (2018).
46 Tang, A. *et al.* Evidence Supporting LI-RADS Major Features for CT- and MR Imaging-based Diagnosis of Hepatocellular Carcinoma: A Systematic Review. *Radiology* **286**, 29-48, doi:10.1148/radiol.2017170554 (2018).
47 Sirlin, C. B., Kielar, A. Z., Tang, A. & Bashir, M. R. LI-RADS: a glimpse into the future. *Abdominal radiology (New York)* **43**, 231-236, doi:10.1007/s00261-017-1448-1 (2018).
48 Cruite, I. *et al.* in *Seminars in roentgenology.* 301-307 (Elsevier).
49 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA: a cancer journal for clinicians* **66**, 7-30 (2016).
50 Holzinger, A., Biemann, C., Pattichis, C. S. & Kell, D. B. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923* (2017).

11) PUBLICATIONS

ANNEX 1 - Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. European Radiology July 2019, Volume 29, Issue 7, pp 3338–3347

ANNEX 2 - Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. European Radiology July 2019, Volume 29, Issue 7, pp 3348–3357

European Radiology (2019) 29:3338–3347
https://doi.org/10.1007/s00330-019-06205-9

Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI

Charlie A. Hamm[1,2] · Clinton J. Wang[1] · Lynn J. Savic[1,2] · Marc Ferrante[1] · Isabel Schobert[1,2] · Todd Schlachter[1] · MingDe Lin[1] · James S. Duncan[1,3] · Jeffrey C. Weinreb[1] · Julius Chapiro[1] · Brian Letzen[1]

Received: 20 December 2018 / Revised: 6 March 2019 / Accepted: 26 March 2019 / Published online: 23 April 2019

Abstract

Objectives To develop and validate a proof-of-concept convolutional neural network (CNN)–based deep learning system (DLS) that classifies common hepatic lesions on multi-phasic MRI.

Methods A custom CNN was engineered by iteratively optimizing the network architecture and training cases, finally consisting of three convolutional layers with associated rectified linear units, two maximum pooling layers, and two fully connected layers. Four hundred ninety-four hepatic lesions with typical imaging features from six categories were utilized, divided into training ($n = 434$) and test ($n = 60$) sets. Established augmentation techniques were used to generate 43,400 training samples. An Adam optimizer was used for training. Monte Carlo cross-validation was performed. After model engineering was finalized, classification accuracy for the final CNN was compared with two board-certified radiologists on an identical unseen test set.

Results The DLS demonstrated a 92% accuracy, a 92% sensitivity (Sn), and a 98% specificity (Sp). Test set performance in a single run of random unseen cases showed an average 90% Sn and 98% Sp. The average Sn/Sp on these same cases for radiologists was 82.5%/96.5%. Results showed a 90% Sn for classifying hepatocellular carcinoma (HCC) compared to 60%/70% for radiologists. For HCC classification, the true positive and false positive rates were 93.5% and 1.6%, respectively, with a receiver operating characteristic area under the curve of 0.992. Computation time per lesion was 5.6 ms.

Conclusion This preliminary deep learning study demonstrated feasibility for classifying lesions with typical imaging features from six common hepatic lesion types, motivating future studies with larger multi-institutional datasets and more complex imaging appearances.

Key Points

- *Deep learning demonstrates high performance in the classification of liver lesions on volumetric multi-phasic MRI, showing potential as an eventual decision-support tool for radiologists.*
- *Demonstrating a classification runtime of a few milliseconds per lesion, a deep learning system could be incorporated into the clinical workflow in a time-efficient manner.*

Keywords Liver cancer · Deep learning · Artificial intelligence

Charlie A. Hamm and Clinton J. Wang contributed equally to the study.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-019-06205-9) contains supplementary material, which is available to authorized users.

✉ Julius Chapiro
j.chapiro@googlemail.com

[1] Department of Radiology and Biomedical Imaging, Yale School of Medicine, 333 Cedar Street, New Haven, CT 06520, USA

[2] Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Berlin Institute of Health, Institute of Radiology, Humboldt-Universität, 10117 Berlin, Germany

[3] Department of Biomedical Engineering, Yale School of Engineering and Applied Science, New Haven, CT 06520, USA

Abbreviations

CNN	Convolutional neural network
CRC	Colorectal carcinoma
DL	Deep learning
DLS	Deep learning system
FNH	Focal nodular hyperplasia
HCC	Hepatocellular carcinoma
ICC	Intrahepatic cholangiocarcinoma
LI-RADS	Liver Imaging Reporting and Data System
PACS	Picture archiving and communication system
Sn	Sensitivity
Sp	Specificity

Introduction

Liver cancer is the second leading cause of cancer-related deaths worldwide and hepatocellular carcinoma (HCC) represents the most common primary liver cancer [1, 2]. Contrary to many other cancer types, HCC incidence rates continue to rise [3]. Rapid and reliable detection and diagnosis of HCC may allow for earlier treatment onset and better outcomes for these patients. As the availability and quality of cross-sectional imaging have improved, the need for invasive diagnostic biopsies has decreased, propelling imaging-based diagnosis to a more central role, with a unique status especially for primary liver cancer. However, the radiological diagnosis of potentially malignant hepatic lesions remains a challenging task. In this setting, standardized image analysis and reporting frameworks such as the Liver Imaging Reporting and Data System (LI-RADS) can improve radiological diagnosis by reducing imaging interpretation variability, improving communication with referring physicians, and facilitating quality assurance and research [4]. However, the increasing complexity of LI-RADS has made its implementation less feasible in a high-volume practice, leaving an unmet clinical need for computational decision-support tools to improve workflow efficiency.

Machine learning algorithms have achieved excellent performance in the radiological classification of various diseases and may potentially address this gap [5–7]. In particular, a deep learning system (DLS) based on convolutional neural networks (CNNs) can attain such capabilities after being shown imaging examples with and without the disease. Unlike other machine learning methods, CNNs do not require definition of specific radiological features to learn how to interpret images, and they may even discover additional differential features not yet identified in current radiological practice [8]. However, such capabilities have not yet been fully demonstrated in the realm of HCC imaging. Most prior machine learning studies classified liver lesions on 2D CT slices and ultrasound images [9–14]. However, higher performance may be achieved with a model that analyzes 3D volumes of multi-phasic contrast-enhanced MRI, which is the reference standard for image-based diagnosis.

Therefore, this study aimed to develop a preliminary CNN-based DLS that demonstrates proof-of-concept for classifying six common types of hepatic lesions with typical imaging appearances on contrast-enhanced MRI, and to validate performance with comparison to experienced board-certified radiologists.

Materials and methods

This was a single-center engineering development and validation study compliant with the Health Insurance Portability and Accountability Act and the Standards for Reporting of Diagnostic Accuracy guidelines. The study was approved by the institutional review board and informed consent was waived. The two components of the study involved (1) engineering a CNN-based liver tumor classifier, followed by (2) proof-of-concept validation of the final optimized CNN by comparison with board-certified radiologists on an identical unseen dataset. An overview of the model training and validation portions is illustrated in Fig. 1.

Establishment of "ground truth" cases

A medical student (CH) searched the picture archiving and communication system (PACS) for abdominal MRI examinations between 2010 and 2017 depicting one of the following hepatic lesions: simple cyst, cavernous hemangioma, focal nodular hyperplasia (FNH), HCC, intrahepatic cholangiocarcinoma (ICC), and colorectal cancer (CRC) metastasis. Due to the nature of a single-institution investigation with limited availability of pathological proof, lesions were restricted to those displaying typical imaging features, incorporating clinical criteria to maximize the certainty of definite diagnosis. Table S1 contains the selected criteria for the "ground truth" utilized for each lesion type. Diagnosed lesions formally described by radiology faculty on official reports were double-checked post hoc according to these criteria with another radiological reader (BL), and lesions were excluded if they contained discrepancies or displayed poor image quality. Up to three imaging studies per patient were included as long as studies were more than 3 months apart. Up to nine different lesions were used in each study. The majority of included lesions were untreated; treated lesions were only included if the selected lesion showed progression, or the patient underwent loco-regional therapy more than 1 year ago and now presented with residual tumor. Patients younger than 18 years were excluded.

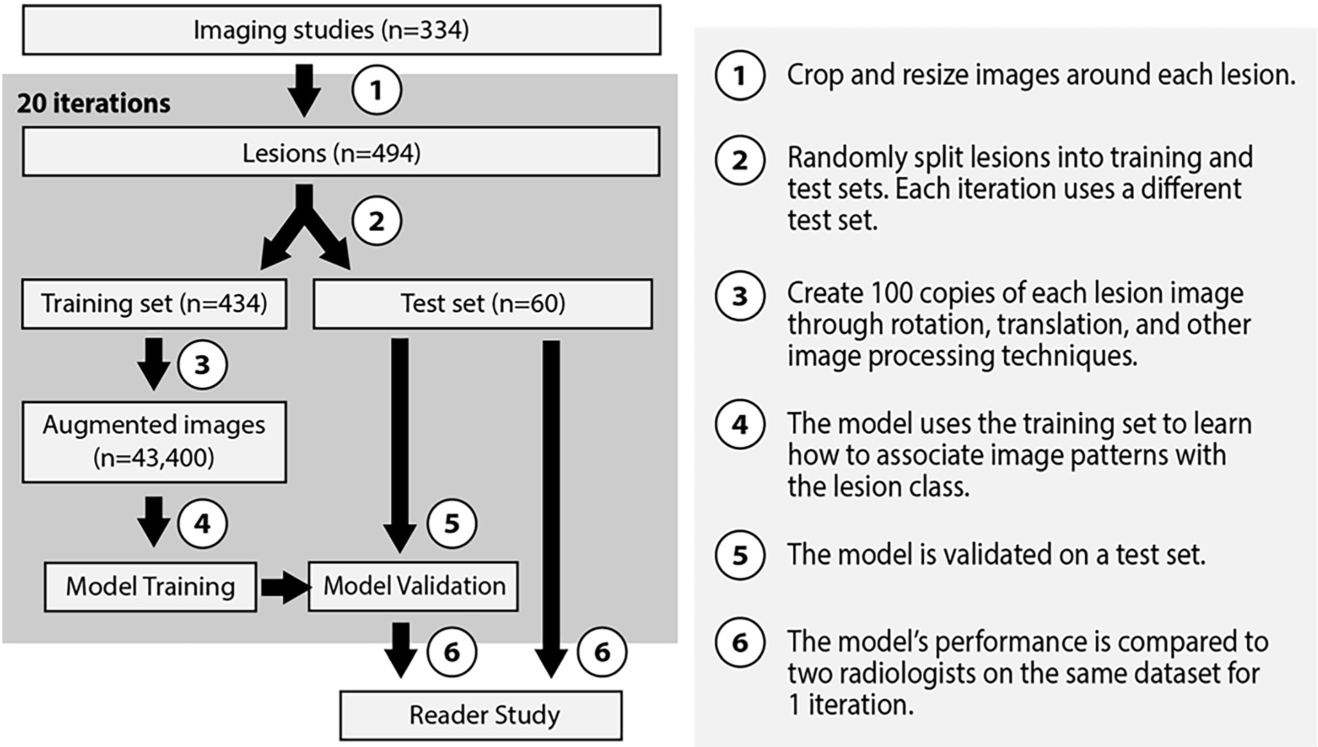

Imaging studies (n=334)
20 iterations
1
Lesions (n=494)
2
Training set (n=434)
Test set (n=60)
3
Augmented images (n=43,400)
4
Model Training
Model Validation
5
6
6
Reader Study
1 Crop and resize images around each lesion.
2 Randomly split lesions into training and test sets. Each iteration uses a different test set.
3 Create 100 copies of each lesion image through rotation, translation, and other image processing techniques.
4 The model uses the training set to learn how to associate image patterns with the lesion class.
5 The model is validated on a test set.
6 The model's performance is compared to two radiologists on the same dataset for 1 iteration.

Fig. 1 Flowchart of the lesion classification approach, including model training, model testing, and reader study

MRI acquisition protocol

This study involved MRI examinations performed from 2010 to 2017 available throughout the institutional PACS, designed to include a heterogeneous collection of MRI scanners and imaging studies. This incorporated both 1.5-T and 3-T MR scanners, including Siemens Aera, Espree, Verio, Avanto, Skyra, and Trio Tim and GE Discovery and Signa Excite scanners. Multi-phasic contrast-enhanced T1-weighted breath-hold sequences from standard institutional liver MR imaging protocols were used with acquisition times of 12–18 s. Several different gadolinium-based contrast agents were used (dosed at 0.1 mmol/kg), including Dotarem (Guerbet), Gadavist (Bayer), Magnevist (Bayer), ProHance (Bracco Diagnostics), and Optimark (Covidien). Post-contrast images were analyzed, including late arterial phase (~ 20 s post-injection), portal venous phase (~ 70 s post-injection), and delayed venous phase (~ 3 min post-injection). Imaging parameters varied across different scanners and time frames; however, the majority were in the range of TR 3–5 ms, TE 1–2 ms, flip angle 9–13°, bandwidth 300–500 Hz, slice thickness 3–4 mm, image matrix 256 × 132 to 320 × 216, and field-of-view 300 × 200 mm to 500 × 400 mm.

Image processing

Eligible MRI studies were downloaded from the PACS and stored as DICOM files. The location and size of a 3D bounding box around the target lesion were manually recorded on the *x*-, *y*-, and *z*-axis. The images were processed and automatically cropped to show only the lesion of interest using code written in the programming language Python 3.5 (Python Software Foundation). The cropped image was then resampled to a resolution of 24 × 24 × 12 voxels (Fig. 2). To minimize bias field effects, cropped images were normalized to intensity levels from − 1 to 1. Affine registration with a mutual information metric was used to register portal venous and delayed phase MRI studies to the arterial phase. Ten lesions from each class were randomly selected to comprise the test set (12% of the entire dataset) using Monte Carlo cross-validation and the remaining lesions comprised the training set. Each image in the training set was augmented by a factor of 100 using established techniques [15] to increase the number of training samples, which allows the model to learn imaging features that are invariant to rotation or translation. During augmentation, images randomly underwent rotation, translation, scaling, flipping, interphase translation, intensity scaling, and intensity shifting.

Deep learning model development

The CNN model was trained on a GeForce GTX 1060 (NVIDIA) graphics processing unit. The model was built using Python 3.5 and Keras 2.2 (https://keras.io/) [16] running on a Tensorflow backend (Google, https://www.tensorflow.org/). Model engineering consisted of iteratively adjusting the network architecture (number of convolutional layers, pooling layers, fully connected layers, and filters for

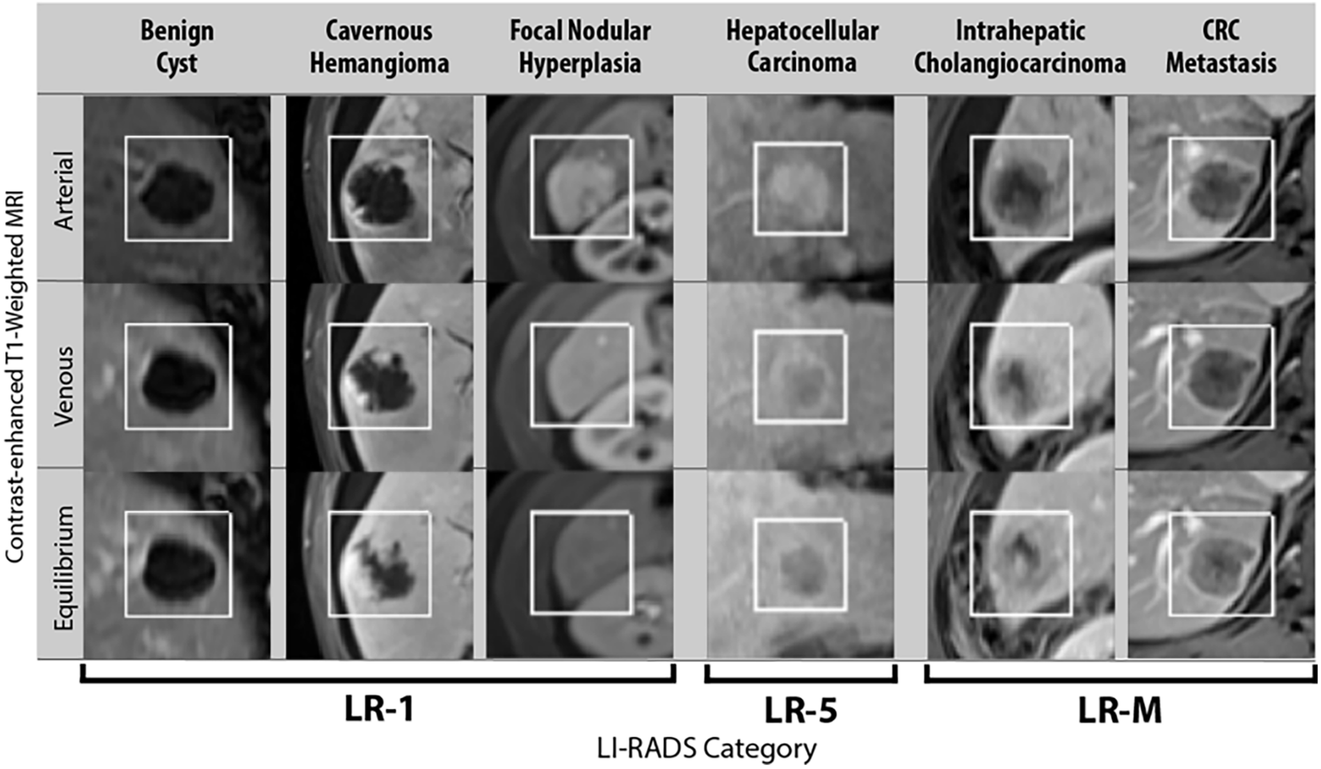


Fig. 2 Sample images of lesion classes and corresponding derived LI-RADS categories. Boxes indicate the cropping of each lesion, which adds padding to the lesion coordinates as determined by a radiologist. The model was able to overcome extrahepatic tissues such as the kidney

each layer, along with parameter optimization) and training cases (removing cases with poor imaging quality or ambiguous imaging features and increasing the number of training samples for lesion classes demonstrating lower performance). The final CNN consisted of three convolutional layers, where the first layer had 64 convolutional filters for each of the three phases in the original image, and the other two had 128 filters across all phases. Each filter generated filtered images by convolving voxels in 3 × 3 × 2 blocks. The model also contained two maximum pooling layers (size 2 × 2 × 2 and 2 × 2 × 1 respectively), which reduce the resolution of filtered images to provide spatial invariance (i.e., a feature that is shifted by a voxel can still be represented by the same neuron, which facilitates learning). The final CNN contained two fully connected layers, one with 100 neurons and the second with a softmax output to six categories that corresponded to the lesion types (Fig. 3). The selected imaging studies spanned 296 patients (155 male/141 female) (Table 1). A total of 334 imaging studies were selected, with a combined total of 494 lesions (74 cysts, 82 cavernous hemangiomas, 84 FNHs, 109 HCCs, 58 ICCs, 87 CRC metastases). The average diameter of all lesions used was 27.5 ± 15.9 mm, ranging from 21.7 ± 15.5 mm for simple cysts to 45 ± 16.8 mm for ICCs (Table 2). The CNN used rectified linear units after each convolutional layer and the first fully connected layer, which helps the model to learn non-linear features [15]. These are used in conjunction with batch normalization and dropout, which are regularization techniques that help the model to generalize beyond the training data [17]. Each CNN was trained with an Adam optimizer using minibatches of five samples from each lesion class. Hyperparameters were chosen via an exhaustive search through a manually specified portion of the search, an approach known in the literature as a grid search [18]. Samples were chosen randomly from the augmented dataset. The model was then tested on its ability to correctly classify 60 lesions in the test dataset (10 from each lesion class) and performance was averaged over 20 independent training iterations with different groupings of training and test datasets to gain a more accurate assessment.

Reader study validation

After development of the CNN model was complete, the classification accuracy of the final CNN was compared with two board-certified radiologists, using an identical set of randomly selected lesions that were unseen by either the CNN model or the radiologists. The two radiologists (39 and 7 years of experience) did not take part in the model training process and were blinded to the lesion selection. The reader study was conducted on an OsiriX MD (v.9.0.1, Pixmeo SARL) workstation. To provide even comparison of input data available to the CNN model, the simulated ready study contained several differences compared to actual clinical practice. The imaging

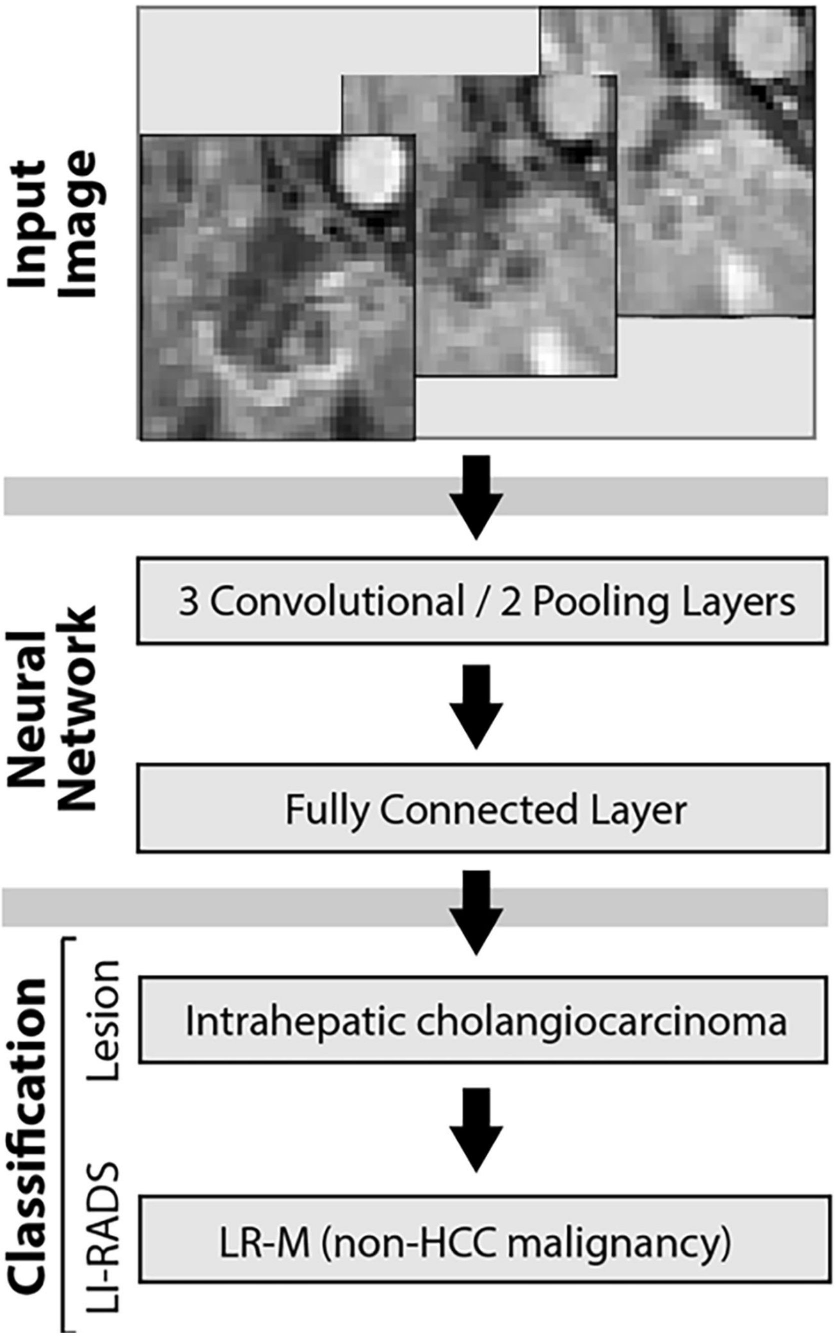

Input Image
Neural Network
3 Convolutional / 2 Pooling Layers
Fully Connected Layer
Classification
Lesion
Intrahepatic cholangiocarcinoma
LI-RADS
LR-M (non-HCC malignancy)

Fig. 3 Neural network model architecture used to infer the lesion entity based on the input image, shown for an example of intrahepatic cholangiocarcinoma. The derived LI-RADS classification follows from the lesion class

studies were anonymized, and the radiologists were fully blinded to clinical data as well as MRI sequences not utilized for the CNN training. The test set for the reader study consisted of 10 randomly selected lesions of each class, 60 lesions in total, while the remaining lesions were assigned to the training set. The randomization was based on Monte Carlo cross-validation and the results of the reader study were compared after a single iteration to mimic their "first exposure" to the images. Each radiologist independently classified the 60 lesions characterized by the model in the test set based on the original three contrast-enhanced MRI phases (late arterial, portal venous, and delayed/equilibrium). Their performance was evaluated in distinguishing the six lesion entities as well as three broader categories that simulate the application of a deep learning model to an HCC diagnostic imaging framework such as LI-RADS. The three broader derived categories were HCCs (corresponding to LR-5), benign lesions (grouping cysts, hemangiomas, and FNHs, corresponding to LR-1), and malignant non-HCC lesions (grouping ICCs and CRC metastases, corresponding to LR-M). The radiologists did not scroll any further than the superior and inferior margins of the lesion in order to avoid revealing

Table 1 Patient characteristics and demographics. Total column does not equal the sum of the rows because some patients had multiple lesion types

Patient characteristics	Cyst	Cavernous hemangioma	FNH	HCC	ICC	CRC metastasis	Total
Number of patients	37	49	53	88	36	39	296
Age at imaging (mean ± SD)	62 ± 10	50 ± 11	43 ± 11	63 ± 8	63 ± 14	61 ± 14	57 ± 14
Gender							
Male	19	17	8	67	18	27	155
Female	18	32	45	21	18	12	141
Ethnicity							
Caucasian	29	39	34	50	25	32	206
African American	2	3	11	12	3	2	32
Asian	3	0	0	3	1	0	5
Other	0	3	2	12	3	4	24
Unknown	3	4	6	11	4	1	29

possible other lesions within the liver and thereby biasing the read. The time from opening the MRI phases until classification of the lesion was recorded.

Statistics

The performance of the model was evaluated by averaging the sensitivity, specificity, and overall accuracy over 20 iterations, as described above. For validation of the CNN with radiological readings, the performances of both the model and the radiologists were computed by evaluating sensitivity, specificity, and overall accuracy on the same single randomly selected test set of unseen cases. Prevalence-based parameters such as positive predictive value and negative predictive value were not applicable for this study. A receiver operating characteristic curve was plotted to compare the model and radiologist performance in identifying HCC masses.

Results

Deep learning model

The final CNN demonstrated a training accuracy of 98.7% ± 1.0 (8567/8680 volumetric samples) across six lesion types and 99.1% ± 0.7 (8602/8680) according to the three general derived LI-RADS categories (Table 3). The average test accuracy was 91.9% ± 2.9 (1103/1200) among individual lesions and 94.3% ± 2.9 (1131/1200) across the three broader categories. The time to initially train the DLS was 29 ± 4 min. Once the model was trained, the actual runtime to classify each lesion in the test dataset was 5.6 ± 4.6 ms.

For the 20 iterations, the average model sensitivity across the six lesion types was 92%, with an average specificity of 98% (Table 4). The model sensitivity for individual lesion types ranged from 89% (177/200) for CRC metastases to 99% (197/200) for simple cysts (Table 4). The corresponding model specificity for individual lesions ranged from 97% (965/1000) for ICC to 100% (1000/1000) for simple cysts. HCC lesions demonstrated a sensitivity of 94% (187/200) and specificity of 98% (984/1000). For the case of the three broader categories, the sensitivity ranged from 94% (187/200 for HCC, 563/600 for benign lesions) to 95% (381/400 for malignant non-HCC lesions). The corresponding specificity ranged from 96% (770/800 for malignant non-HCC lesions, and 577/600 for benign lesions) to 98% (984/1000 for HCC). The study was conducted using the same number of lesions from each class, and thus does not reflect the actual prevalence of each lesion type.

Reader study

Classification of unseen randomly selected lesions included in the reader study demonstrated an average model accuracy of

Table 2 Imaging details for each category of lesion

Image characteristics	Cyst	Cavernous hemangioma	FNH	HCC	ICC	CRC metastasis	Total
Number of patients	37	49	53	88	36	39	296
Number of imaging studies	42	50	57	96	49	44	334
Number of lesions	74	82	84	109	58	87	494
Lesion diameter (mm, mean ± SD)	21.7 ± 15.5	25 ± 11.6	28.4 ± 20.7	24.4 ± 10	45 ± 16.8	26.4 ± 12.3	27.5 ± 15.9

Total column does not equal the sum of the rows because some imaging studies had multiple lesion types

Table 3 Overall accuracy and runtimes for model classification and classification by two radiologists

	Accuracy of lesion classification (mean ± SD %)	Accuracy of derived LI-RADS classification (mean ± SD %)	Runtime (mean ± SD)
Average of 20 iterations			
Model training set	98.7 ± 1.0	99.1 ± 0.7	29 min ± 4
Model test set	91.9 ± 2.9	94.3 ± 2.9	5.6 ms ± 4.6
Reader study (*n* = 60)			
Model	90.0	91.7	1.0 ms ± 0.4
Radiologist 1	80.0	88.3	14 ± 10 s
Radiologist 2	85.0	88.3	17 ± 24 s

90% (55/60 lesions). Radiologist accuracy was 80% (48/60) and 85% (51/60) on these same lesions, respectively (Table 3). The model accuracy for the three broader categories was 92% (58/60), compared with 88% (53/60) for both radiologists. The total elapsed time analyzing each lesion was 0.8 ms for the classification model versus 14 ± 10 s and 17 ± 24 s for the radiologists.

Lesions included in the reader study showed an average CNN model sensitivity of 90% ± 14 (9/10) and specificity of 98% ± 2 (49/50) across the six lesion types. This compared to an average sensitivity of 80% ± 16 (8/10) and 85% ± 15 (8.5/10) and specificity of 96% ± 3 (48/50) 97% ± 3 (48.5/50) for the two radiologists respectively (Table 4). The model sensitivity ranged from 70% (7/10 for FNH) to 100% (10/10 for simple cysts and hemangiomas) with a specificity ranging from 92% (46/50 for HCC) to 100% (50/50 for simple cysts, hemangiomas, and ICC). Radiologist sensitivity ranged from 50% (5/10 for CRC metastases) to 100% (10/10 for simple cysts, hemangiomas), with specificity ranging from 92% (46/50 for CRC metastases) to 100% (50/50 for HCC and ICC). The average model sensitivity for three broader categories was 92% with a specificity of 97%. This compared to the radiologists' sensitivity of 88% and specificity of 89% and 91%, respectively. The model demonstrated highest sensitivity for malignant non-HCC lesions at 95% (19/20) compared to 85% (17/20) for both radiologists, whereas radiologists attained highest sensitivity for benign lesions at 97% (29/30) and 100% (30/30), compared to 90% (27/30) for the CNN.

A receiver operating characteristic curve was constructed by varying the probability threshold at which the CNN would classify a lesion as HCC, with an area under the curve of 0.992 (Fig. 4). This included a true positive rate of 93.5% (187/200) averaged over 20 iterations and a false positive rate of 1.6% (16/1000). When including only lesions within the reader study, the model true positive rate was 90% (9/10), and the false positive rate was 2% (1/50). Radiologists had a true positive rate of 60% and 70% (6/10 and 7/10, respectively) and a false positive rate of 0% (0/50).

Table 4 Model and radiologist performance metrics for individual lesion types and LI-RADS classes

	Average of 20 iterations		Reader study					
	Model test set		Model		Radiologist 1		Radiologist 2	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Lesion type								
Cyst	99%	100%	100%	100%	90%	96%	100%	98%
Hemangioma	91%	99%	100%	100%	100%	96%	100%	94%
FNH	91%	98%	90%	96%	90%	98%	90%	94%
HCC	94%	98%	90%	98%	70%	100%	60%	100%
ICC	90%	97%	60%	100%	80%	94%	90%	100%
CRC metastasis	89%	98%	100%	94%	50%	92%	70%	96%
Overall	92%	98%	90%	98%	80%	96%	85%	97%
Derived LI-RADS class								
LR-1 (*n* = 30)	94%	96%	97%	93%	97%	87%	100%	80%
LR-5 (*n* = 10)	94%	98%	90%	98%	70%	100%	60%	100%
LR-M (*n* = 20)	95%	96%	95%	100%	85%	93%	85%	98%
Overall	94%	97%	95%	96%	88%	91%	88%	89%

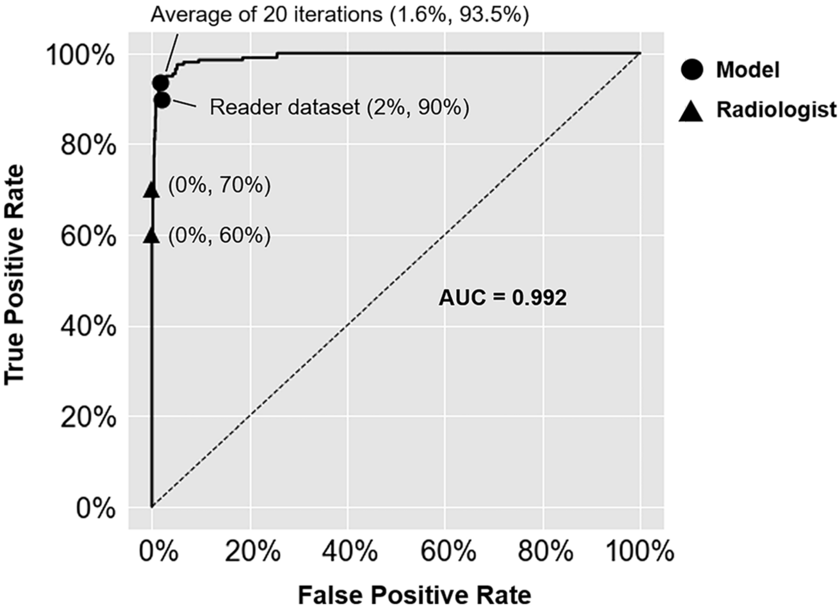


Fig. 4 Model receiver operating characteristic curve for distinguishing HCCs. This model achieves high sensitivity for HCC at the cost of a few false positives. AUC, area under curve

Discussion

This study demonstrates a deep learning–based prototype for classification of liver lesions with typical imaging features from multi-phasic MRI, demonstrating high performance and time efficiency. While the study did not simulate clinical practice conditions, comparison with equivalent data input showed the potential of DL systems to eventually aid in improving radiological diagnosis of six classes of hepatic lesions (model accuracy of 92%, radiologist accuracy of 80% and 85%), as well as three broader categories of benign, HCC, and malignant non-HCC lesions (model accuracy of 94%, radiologist accuracy of 88%), with a classification time of 5.6 ms per lesion.

Building upon prior 2D CT and ultrasound models, the inherent improved soft tissue contrast resolution of MRI can enable this CNN to capture a wider variety of imaging features [14]. Additionally, the 3D volumetric approach may improve detection of inhomogeneous growth or enhancement patterns that may be relevant to lesion classification, while removing the model's variability and dependence on manual slice selection [19, 20]. Furthermore, the use of heterogeneous imaging sources demonstrated the robustness of DLS in the setting of different MRI scanners and acquisition protocols.

Previous studies have paved the way for computational classification of diverse lesion types by grouping hepatic lesion entities into three to five classes [11, 13, 14]. Moving towards clinical implementation, classification becomes increasingly challenging when lesions are ungrouped and single entities are differentiated. In this case, a higher number of differential features must be learned with a lower chance of guessing correctly. The present study included six ungrouped lesion classes, demonstrating a high accuracy level of 91.9%. As expected, the overall accuracy was higher with three grouped classes (94.3%).

Since single-center developmental efforts often suffer from limited datasets, selection of idealized cases is often necessary, making the interpretation of classification results ambiguous. The direct comparison between the DLS and two radiologists allows for better interpretation of performance and potential clinical value. High sensitivity for HCC and CRC metastases was demonstrated relative to radiologists. The radiologists tended to misclassify HCCs with faint enhancement as CRC metastases and HCCs with unclear washout as FNHs, whereas the DLS could more reliably make use of other features to correctly identify the HCCs. Similarly, radiologists misclassified CRC metastases without clear progressive enhancement with cysts, and those with heterogeneous, nodular appearances were misclassified for ICCs, whereas the computational predictions were likely more robust to the absence of these features. Still, the radiologists' diagnostic accuracy may have matched or exceeded the DLS's accuracy if given access to clinical information or additional imaging sequences. As a proof-of-concept study with limited sequences, this simulated environment provided unbiased comparison between the DLS and radiologists with the same available input data.

These performance metrics suggest that a DLS could serve as a quick and reliable "second opinion" for radiologists in the diagnosis of hepatic lesions, helping to reduce interpretation difficulty and inter-reader variability when imaging features are more ambiguous. In HCC diagnosis, most inter-reader studies demonstrated a moderate level of reliability in determining LI-RADS classes [21–26], and the rigor and complexity of LI-RADS constitutes a major barrier for broad adoption [25, 27]. The DLS classified lesions into benign, HCC, and malignant non-HCC lesions (roughly corresponding to LR-1, LR-5, and LR-M respectively) with an accuracy of 94.3%. While this is a preliminary feasibility study with many limitations, it suggests that a DLS could potentially interface with LI-RADS, for example, by averaging the model and radiologist predictions to score lesions that are suspicious for HCC but lack a definite benign/malignant appearance (i.e., LR-2/3/4). Such an implementation could reduce rote manual tasks, helping to simplify LI-RADS for clinical workflow integration [27].

While these results are promising, there are several limitations that make this a preliminary feasibility study. As a single-center investigation, only a limited number of imaging studies were available for each class. Thus, only lesions with typical imaging features on MRI were used, excluding lesions with more ambiguous features or poor image quality as well as more complex lesion types such as infiltrative HCC or complicated cysts. Additionally, LI-RADS is only applicable to patients at high risk for HCC. However, because non-HCC lesions are much less common in cirrhotic livers, this study also included lesions in livers without cirrhotic background or hepatitis-B/C, and thus the input does not identically conform to current consensus. Additionally, due to limited data from a single institution, pathological proof was not available for all lesions. Thus, "ground truth" criteria were carefully selected and defined for each lesion type as thoroughly outlined in Table S1. Notably, for lesions without pathological diagnosis, "ground truth" was established by analyzing all available clinical and imaging data, including T1 pre-contrast, T2, and other sequences. However, these sequences were not used in the model training and subsequent reader study, and thus their potential additive value for the CNN performance needs to be evaluated in further studies. Additionally, the simulated reader comparison did not reflect conditions in clinical practice, as the test set contained equal numbers of each lesion type and participants did not have access to ancillary information such as clinical data. However, this allowed for initial validation of the CNN with radiologists using the same conditions and input data for a more equivalent comparison. Within these limitations, this approach met the study's purpose to demonstrate initial feasibility of a liver MRI lesion classification prototype from available data at one large academic medical center, providing motivation for the establishment of larger multi-institutional databases.

In summary, this preliminary study provides proof of principle for a DLS that classifies six hepatic lesion types on multi-phasic MRI, demonstrating high performance when validated by comparison with board-certified radiologists. As the demands of radiological practice continue to increase, a synergistic workflow that combines the experience and intuition of radiologists with the computational power of DL decision-support tools may offer higher-quality patient care in a time-efficient manner.

Funding BL and CW received funding from the Radiological Society of North America (RSNA Research Resident Grant no. RR1731). JD, JC, ML, and CW received funding from the National Institutes of Health (NIH/NCI R01 CA206180).

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Brian Letzen.

Conflict of interest The authors of this manuscript declare relationships with the following companies: JW: Bracco Diagnostics, Siemens AG; ML: Pro Medicus Limited; JC Koninklijke Philips, Guerbet SA, Eisai Co.

Statistics and biometry One of the authors has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- retrospective
- experimental
- performed at one institution

References

1. El–Serag HB, Rudolph KL (2007) Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology 132: 2557–2576
2. Wang H, Naghavi M, Allen C et al (2016) Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 388:1459–1544
3. Siegel RL, Miller KD, Jemal A (2016) Cancer statistics, 2016. CA Cancer J Clin 66:7–30
4. Mitchell DG, Bruix J, Sherman M, Sirlin CB (2015) LI-RADS (Liver Imaging Reporting and Data System): summary, discussion, and consensus of the LI-RADS Management Working Group and future directions. Hepatology 61:1056–1065
5. Yasaka K, Akai H, Abe O, Kiryu S (2018) Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology 286:887–896. https://doi.org/10.1148/radiol.2017170706
6. Grewal M, Srivastava MM, Kumar P, Varadarajan S (2018) RADnet: radiologist level accuracy using deep learning for hemorrhage detection in CT scans2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp 281–284
7. Klöppel S, Stonnington CM, Barnes J et al (2008) Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. Brain 131:2969–2974
8. Greenspan H, Van Ginneken B, Summers RM (2016) Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging 35:1153–1159
9. Shiraishi J, Sugimoto K, Moriyasu F, Kamiyama N (2008) Computer-aided diagnosis for the classification of focal liver lesions by use of contrast-enhanced ultrasonography. Med Phys 35: 1734–1746

10. Sugimoto K, Shiraishi J, Moriyasu F, Doi K (2010) Computer-aided diagnosis for contrast-enhanced ultrasound in the liver. World J Radiol 2:215
11. Hwang YN, Lee JH, Kim GY, Jiang YY, Kim SM (2015) Classification of focal liver lesions on ultrasound images by extracting hybrid textural features and using an artificial neural network. Biomed Mater Eng 26:S1599–S1611
12. Virmani J, Kumar V, Kalra N, Khandelwa N (2013) PCA-SVM based CAD system for focal liver lesions using B-mode ultrasound images. Def Sci J 63:478
13. Acharya UR, Koh JEW, Hagiwara Y et al (2018) Automated diagnosis of focal liver lesions using bidirectional empirical mode decomposition features. Comput Biol Med 94:11–18
14. Rajpurkar P, Irvin J, Ball RL et al (2018) Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15:e1002686. https://doi.org/10.1371/journal.pmed.1002686
15. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. pp 1097–1105
16. Chollet F (2015) Keras. https://keras.io/. Accessed 15 Oct 2018
17. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167
18. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980
19. Chapiro J, Lin M, Duran R, Schernthaner RE, Geschwind J-F (2015) Assessing tumor response after loco-regional liver cancer therapies: the role of 3D MRI. Expert Rev Anticancer Ther 15:199
20. Chapiro J, Wood LD, Lin M et al (2014) Radiologic-pathologic analysis of contrast-enhanced and diffusion-weighted MR imaging in patients with HCC after TACE: diagnostic accuracy of 3D quantitative image analysis. Radiology 273:746–758
21. Barth B, Donati O, Fischer M et al (2016) Reliability, validity, and reader acceptance of LI-RADS-an in-depth analysis. Acad Radiol 23:1145
22. Bashir M, Huang R, Mayes N et al (2015) Concordance of hypervascular liver nodule characterization between the organ procurement and transplant network and liver imaging reporting and data system classifications. J Magn Reson Imaging 42:305
23. Davenport MS, Khalatbari S, Liu PS et al (2014) Repeatability of diagnostic features and scoring systems for hepatocellular carcinoma by using MR imaging. Radiology 272:132
24. Ehman EC, Behr SC, Umetsu SE et al (2016) Rate of observation and inter-observer agreement for LI-RADS major features at CT and MRI in 184 pathology proven hepatocellular carcinomas. Abdom Radiol (NY) 41:963–969
25. Fowler KJ, Tang A, Santillan C et al (2018) Interreader reliability of LI-RADS version 2014 algorithm and imaging features for diagnosis of hepatocellular carcinoma: a large international multireader study. Radiology 286:173–185
26. Liu W, Qin J, Guo R et al (2017) Accuracy of the diagnostic evaluation of hepatocellular carcinoma with LI-RADS. Acta Radiol. https://doi.org/10.1177/0284185117716700:284185117716700
27. Sirlin CB, Kielar AZ, Tang A, Bashir MR (2018) LI-RADS: a glimpse into the future. Abdom Radiol (NY) 43:231–236

Supplementary table

Lesion Type	Imaging Characteristics	Non-imaging criteria
Cyst	• Sharply defined, thin walled lesion with no septations or signs of hemorrhage or inflammation • Hypointense and no enhancement of content on contrast enhanced phases	Diagnosed solely based on imaging characteristics
Cavernous Hemangioma	• Well-circumscribed, spherical to ovoid mass • Early peripheral, nodular or globular, discontinuous enhancement on arterial phase • Progressive centripetal enhancement with isointensity to blood vessels on portal venous phase • Persistent filling or completely filled hyperintense mass on delayed phase • “Flash-filling” lesions were not included in this study.	Diagnosed solely based on imaging characteristics
Focal Nodular Hyperplasia	• Round shaped focal liver mass with homogenous enhancement and marked hyperintensity in the arterial phase • Lesion blends into the surrounding parenchyma as it becomes isointense on portal venous and delayed phase the • Potential central/stellate scar shows uptake enhancement and is hyperintense on portal venous and delayed phases • Presence of a central scar was not necessary for being classified as classic appearing, since the definition of assuming the presence of a stellate scar as a typical feature is generally discussed in literature	Diagnosed solely based on imaging characteristics
Hepato-cellular carcinoma	**OPTN5A:** • Size: 1-2 cm • Representing all of the following features: o Increased contrast enhancement on arterial phase o Washout during portal venous or delayed phases o Peripheral rim enhancement, illustrating a capsule or pseudocapsule **OPTN5B:** • Size: 2- 5 cm • Arterially hyperenhancing and has at least one of two venous features: o Washout o Peripheral rim enhancement **OPTN5X:** • Size: > 5 cm • Arterially hyperenhancing and has at least one of two venous features: o Washout o Peripheral rim enhancement	• Only lesions which were classified as OPTN 5A, OPTN 5B or OPTN5X HCCs were included. • The classification criteria for HCC in the UNOS/OPTN system were developed in such way HCC can be unequivocally diagnosed by using imaging. The diagnostic imaging criteria driving HCC classification rely on the characteristic appearance of HCC on dynamic multiphasic contrast-enhanced CT scans or MR images. • OPTN class 5 indicates that a nodule meets radiologic criteria for HCC.
Intrahepatic cholangio-	• Either well circumscribed, large with lobulated margins or masses with an infiltrative growth	• Histopathologic report from biopsy or surgery

carcinoma	pattern • Delayed enhancement with a progressive and concentric filling pattern on contrast enhanced phases • Distally adjacent bile ducts may show prominent enlargement	• Clinical information and therapy approach
Colorectal carcinoma metastases	• Well-circumscribed, spherical to ovoid mass • Typical enhancement pattern of hypovascular metastases with a hypointense center and peripheral enhancement; “target” lesion appearance • Potential perilesional enhancement due to tumor vascularity or hepatic edema • over time the central part of the lesion remains hypointense due to necrosis or hypovascularity	• Histopathologic report from biopsy or surgery • Clinical information and therapy approach • Known history of primary malignancy

Table S1: Reference standard for included lesions.

European Radiology (2019) 29:3348–3357
https://doi.org/10.1007/s00330-019-06214-8

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features

Clinton J. Wang[1] · Charlie A. Hamm[1,2] · Lynn J. Savic[1,2] · Marc Ferrante[1] · Isabel Schobert[1,2] · Todd Schlachter[1] · MingDe Lin[1] · Jeffrey C. Weinreb[1] · James S. Duncan[1,3] · Julius Chapiro[1] · Brian Letzen[1]

Received: 27 February 2019 / Accepted: 2 April 2019 / Published online: 15 May 2019

Abstract

Objectives To develop a proof-of-concept “interpretable” deep learning prototype that justifies aspects of its predictions from a pre-trained hepatic lesion classifier.

Methods A convolutional neural network (CNN) was engineered and trained to classify six hepatic tumor entities using 494 lesions on multi-phasic MRI, described in Part 1. A subset of each lesion class was labeled with up to four key imaging features per lesion. A post hoc algorithm inferred the presence of these features in a test set of 60 lesions by analyzing activation patterns of the pre-trained CNN model. Feature maps were generated that highlight regions in the original image that correspond to particular features. Additionally, relevance scores were assigned to each identified feature, denoting the relative contribution of a feature to the predicted lesion classification.

Results The interpretable deep learning system achieved 76.5% positive predictive value and 82.9% sensitivity in identifying the correct radiological features present in each test lesion. The model misclassified 12% of lesions. Incorrect features were found more often in misclassified lesions than correctly identified lesions (60.4% vs. 85.6%). Feature maps were consistent with original image voxels contributing to each imaging feature. Feature relevance scores tended to reflect the most prominent imaging criteria for each class.

Conclusions This interpretable deep learning system demonstrates proof of principle for illuminating portions of a pre-trained deep neural network’s decision-making, by analyzing inner layers and automatically describing features contributing to predictions.

Key Points

- *An interpretable deep learning system prototype can explain aspects of its decision-making by identifying relevant imaging features and showing where these features are found on an image, facilitating clinical translation.*
- *By providing feedback on the importance of various radiological features in performing differential diagnosis, interpretable deep learning systems have the potential to interface with standardized reporting systems such as LI-RADS, validating ancillary features and improving clinical practicality.*
- *An interpretable deep learning system could potentially add quantitative data to radiologic reports and serve radiologists with evidence-based decision support.*

Keywords Liver cancer · Artificial intelligence · Deep learning

Clinton J. Wang and Charlie A. Hamm contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-019-06214-8) contains supplementary material, which is available to authorized users.

✉ Julius Chapiro
j.chapiro@googlemail.com

[1] Department of Radiology and Biomedical Imaging, Yale School of Medicine, 333 Cedar Street, New Haven, CT 06520, USA

[2] Institute of Radiology, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität, and Berlin Institute of Health, 10117 Berlin, Germany

[3] Department of Biomedical Engineering, Yale School of Engineering and Applied Science, New Haven, CT 06520, USA

Abbreviations

CNN	Convolutional neural network
CRC	Colorectal carcinoma
DL	Deep learning
FNH	Focal nodular hyperplasia
HCC	Hepatocellular carcinoma
ICC	Intrahepatic cholangiocarcinoma
LI-RADS	Liver Imaging Reporting and Data System
PPV	Positive predictive value
Sn	Sensitivity

Introduction

Deep learning (DL) systems based on convolutional neural networks (CNNs) have shown potential to revolutionize the process of radiological diagnosis [1–3]. Unlike other artificial intelligence techniques, CNNs do not need to be taught specific radiological features to learn how to interpret images [4]. A synergistic workflow that combines the experience of radiologists and the computational power of artificial intelligence systems may substantially improve the efficiency and quality of clinical care. Part I of this article series demonstrated a proof-of-concept 3D CNN for the classification of liver lesions on multi-phasic MRI [5]. Although CNNs have demonstrated high performance in diagnostic classification tasks, their "black box" design limits their clinical adoption [6–8]. Despite recent advances in interpretable machine learning [9], deep learning models still do not provide information about the factors used in decision-making in a manner that can be understood by radiologists and other physicians, which prevents them from incorporating their results into an informed decision-making process. The inability to explain their reasoning also leads to a lack of safeguards and accountability when they fail. DL systems that demonstrate high accuracy in a more transparent manner are more likely to gain clinical acceptance.

This is especially applicable when incorporating DL into standardized reporting systems such as the Liver Imaging Reporting and Data System (LI-RADS). While LI-RADS has changed the diagnostic workflow of malignant lesions and contributed to higher quality diagnosis and reporting [10–12], most studies have demonstrated moderate inter-observer agreement for LI-RADS categories [13–19]. Recent studies also highlighted issues regarding the application of LI-RADS ancillary features, which are primarily based on a combination of biological plausibility, single-center retrospective studies, and expert opinion with somewhat low level of evidence [20, 21]. For example, the application of such features resulted in an increased number of misclassifications [10, 14, 22] and ancillary features were not seen as a useful tool for assigning definite LR classes [13]. Moreover, the application of a number of ancillary features may be inefficient, as they affected the final diagnosis in at most 10% of cases [13, 19]. The American College of Radiology has called for novel systems to increase the efficiency and accuracy of LI-RADS and to make it more feasible for daily radiology practice [21]. Interpretable DL systems could help to address this gap by automating the validation, detection, and standardized reporting of diagnostic imaging features, providing a way for radiologists to efficiently interact with such tools in a shared decision-making paradigm.

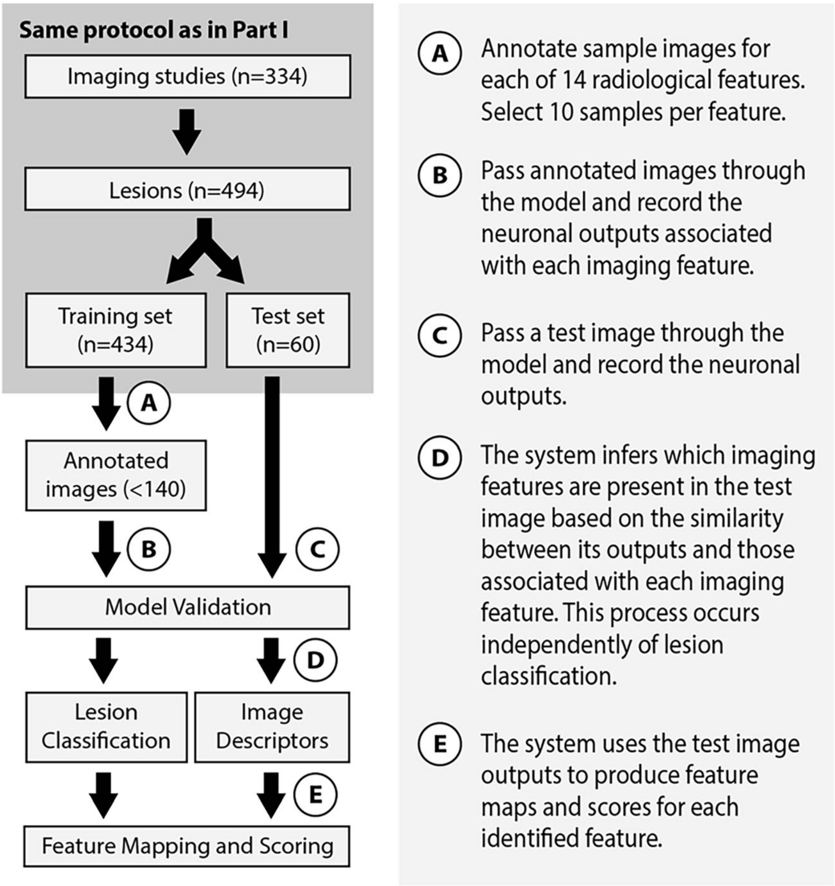
This study investigates an integrative interpretable DL approach for DL systems used in clinical radiology, using techniques for identifying, localizing, and scoring imaging features. In addition to developing a liver lesion classifier for multi-phasic MRI (Part I), the aim of Part II was to develop a proof-of-concept interpretable system that justifies aspects of its decisions through internal analysis of relevant radiologic features.

Materials and methods

Deep learning system and model-agnostic interpretability

This single-center retrospective study is based on an institutional review board–approved protocol, and the requirement for written consent was waived. The specific methods for patient selection, lesion reference standard, MRI technique, image processing techniques, and DL model are described in Part I of this study [5]. Briefly, a CNN was utilized with three convolutional layers and two fully connected layers, which was capable of differentiating benign cysts, cavernous hemangiomas, focal nodular hyperplasias (FNHs), HCCs, intrahepatic cholangiocarcinomas (ICCs), and colorectal carcinoma (CRC) metastases after being trained on 434 hepatic lesions from these classes. This study was integrated into the Part I DL workflow so that the system could be trained to classify lesion types before incorporating techniques to identify, localize, and score their radiological features (Fig. 1). Specifically, the current study utilized the DL model from Part I which has been trained on a large dataset including 494 lesions. Additionally, custom algorithms were applied to analyze specific hidden layers of this pre-trained neural network in a model-agonistic approach. This method is also known as post hoc analysis (not to be confused with the post hoc analysis in statistics) and is generalizable to various pre-trained machine learning neural networks [23, 24]. Under the taxonomy of established interpretability methods, these algorithms fall under the general category of feature summary statistic. In terms of scope, the methods used describe local interpretability where the focus is on individual predictions, as opposed to

Fig. 1 Flowchart of the approach for lesion classification and radiological feature identification, mapping, and scoring. The entire process was repeated over 20 iterations



A Annotate sample images for each of 14 radiological features. Select 10 samples per feature.

B Pass annotated images through the model and record the neuronal outputs associated with each imaging feature.

C Pass a test image through the model and record the neuronal outputs.

D The system infers which imaging features are present in the test image based on the similarity between its outputs and those associated with each imaging feature. This process occurs independently of lesion classification.

E The system uses the test image outputs to produce feature maps and scores for each identified feature.

global scope where the entire model behaviour is analysed. These selected techniques are especially useful for the purposes of communicating feature information to radiologists. These algorithms are described in detail below.

Table 1 Radiological features labeled for each class. A total of 224 example images were used across the 14 radiological features, and some images were labeled with multiple features

Radiological features	Associated lesion types	Number of examples	Frequency in the test set
Arterial phase hyperenhancement	FNH, HCC	20	19/60
Central scar	FNH	10	1/60
Enhancing rim (CRC metastasis), capsule/pseudocapsule (HCC)	CRC metastasis, HCC	20	15/60
Heterogeneous lesion	ICC, HCC (OPTN5B/X)	20	17/60
Hyperenhancing mass on delayed phase	Cavernous hemangioma	17	8/60
Hypoenhancing core (CRC metastasis), hypoenhancing mass (cyst)	Cyst, CRC metastasis	20	20/60
Infiltrative appearance	ICC	15	4/60
Iso-intensity on venous and delayed phase	FNH	20	9/60
Nodularity	ICC	15	6/60
Nodular/discontinuous peripheral hyperenhancement	Cavernous hemangioma	20	10/60
Progressive centripetal filling	Cavernous hemangioma	20	9/60
Progressive hyperenhancement	CRC metastasis, ICC	20	19/60
Thin-walled mass	Cyst	20	8/60
Washout	HCC	20	9/60

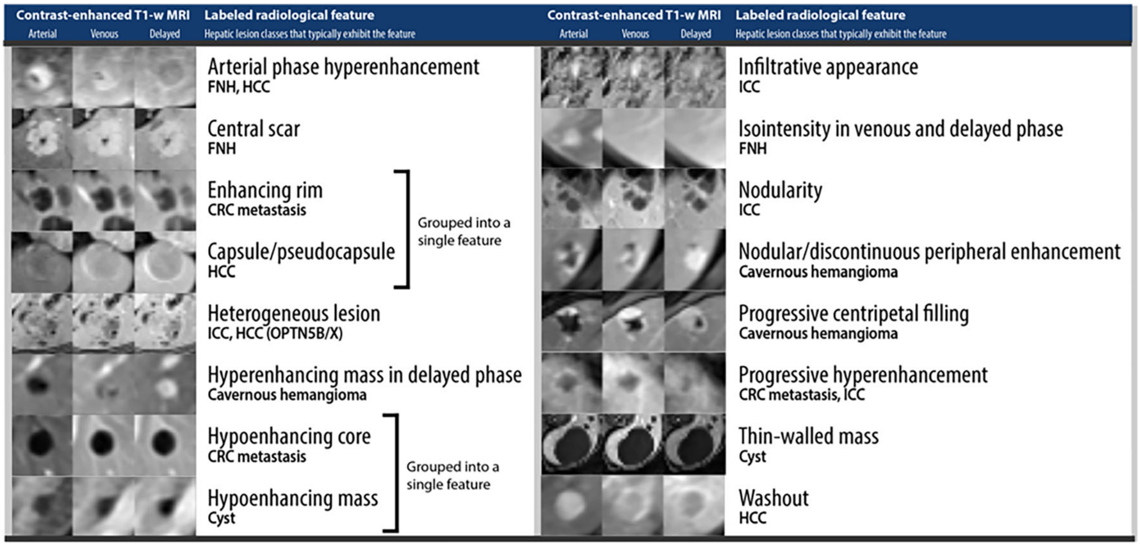

Contrast-enhanced T1-w MRI
Arterial
Venous
Delayed
Labeled radiological feature
Hepatic lesion classes that typically exhibit the feature
Arterial phase hyperenhancement
FNH, HCC
Central scar
FNH
Enhancing rim
CRC metastasis
Capsule/pseudocapsule
HCC
Grouped into a single feature
Heterogeneous lesion
ICC, HCC (OPTN5B/X)
Hyperenhancing mass in delayed phase
Cavernous hemangioma
Hypoenhancing core
CRC metastasis
Hypoenhancing mass
Cyst
Grouped into a single feature
Contrast-enhanced T1-w MRI
Arterial
Venous
Delayed
Labeled radiological feature
Hepatic lesion classes that typically exhibit the feature
Infiltrative appearance
ICC
Isointensity in venous and delayed phase
FNH
Nodularity
ICC
Nodular/discontinuous peripheral enhancement
Cavernous hemangioma
Progressive centripetal filling
Cavernous hemangioma
Progressive hyperenhancement
CRC metastasis, ICC
Thin-walled mass
Cyst
Washout
HCC

Fig. 2 Examples of labeled sample lesions for the 14 radiological features

Radiological feature selection

Fourteen radiological features were selected comprising lesion imaging characteristics that are observable on multi-phasic MRI and are commonly utilized in day-to-day radiological practice for differentiating between various lesion types [25, 26] (Table 1). This includes LI-RADS features for HCC classification, including arterial phase hyperenhancement, washout, and pseudocapsule. Up to 20 hepatic lesions in the training set that best exemplified each feature were selected (Fig. 2). From this sample, ten were randomly selected in each repetition of this study. Imaging features with similar appearances were grouped. A test set of 60 lesions was labeled with the most prominent imaging features in each image (1–4 features per lesion). This test set was the same as that used to conduct the reader study in Part I.

Feature identification with probabilistic inference

For each radiological feature, a subset of ten sample lesions with that feature was passed through the CNN, and the intermediate outputs of the 100 neurons in the fully connected layer were inspected. By analyzing these neuronal outputs among the ten samples, each radiological feature was associated with specific patterns in these neurons. The test image was passed through the CNN to obtain its intermediate outputs, which were compared to the outputs associated with each feature. When the intermediate outputs of a test image are similar to the outputs observed for lesions with a particular feature, then the feature is likely to be present in the test image (see Fig. 3). The intermediate outputs were modeled as a 100-dimensional random variable and the training dataset was used to obtain its empirical distribution (refer to “marginal distributions” and “conditional distributions” in [27]. Using kernel density estimation, the features present in each test

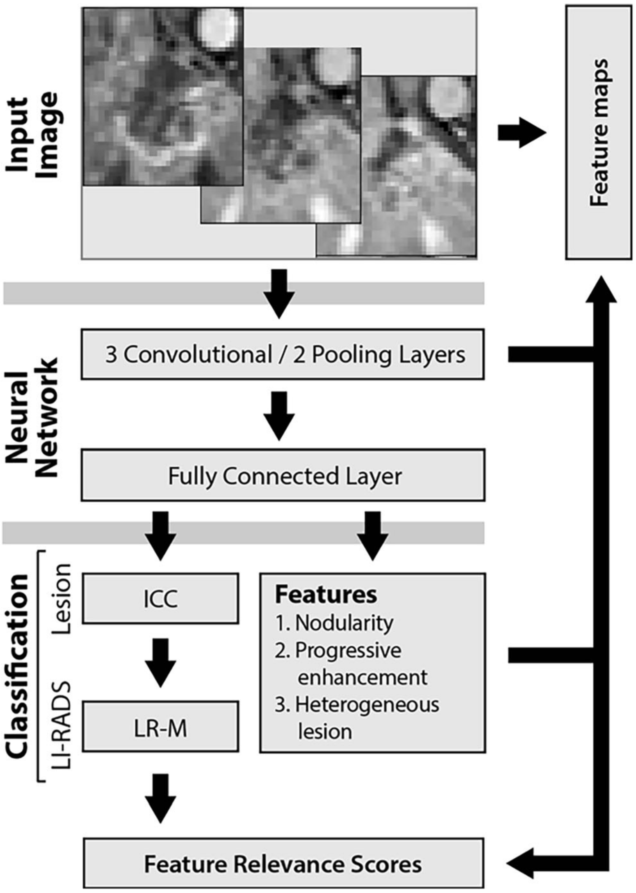


Fig. 3 CNN model architecture used to infer the lesion entity and radiological features based on the input image, shown for an example of intrahepatic cholangiocarcinoma. Patterns in the convolutional layers are mapped back to the input image to establish feature maps for each identified feature. As well, relevance scores are assigned to the features based on the correspondence between patterns in the convolutional layers, the lesion classification, and the identified features

image were probabilistically inferred. The neuronal outputs of augmented versions of all images were used to provide more robust estimates of the probability distributions. As described in Part I, image augmentation creates copies of images with stochastic distortions.

The CNN system's performance was assessed by its ability to correctly identify the radiological features in the test set of 60 labeled lesions. Performance was evaluated in 20 iterations with separately trained models using different choices of the ten sample lesions. Positive predictive value (PPV) and sensitivity (Sn) were measured for the entire population (averaged over the total number of features across all lesions). This was performed for each feature individually and for each lesion class.

Feature mapping with weighted activations

After identifying the radiological features observed in an input lesion image, 3D feature maps were derived from the CNN's layer activations to show where features are observed within each image. For this analysis, the post-activation neuronal outputs of the final convolutional layer were used, which has 128 channels. The original images have 24 × 24 × 12 resolution and pass through padded convolutions and a 2 × 2 × 2 max pooling layer before reaching this layer at 12 × 12 × 6 spatial dimensions. The feature map was constructed for a test image by obtaining this layer's output and applying a weighted average over the 128 channels using different weights for each of the 1–4 radiological features identified within the image. The resulting 12 × 12 × 6 feature maps were upsampled using trilinear interpolation to correspond to the 24 × 24 × 12 resolution of the original image. The mapping to the three MRI phases cannot be readily traced. The channel weights used for each feature were determined by correlating the channel with at most one of the features based on the channel outputs observed in the sample lesions labeled with the feature.

Feature scoring with influence functions

Among the radiological features identified in an image, some features may be more important for classifying the lesion than others. The contribution of each identified feature to the CNN's decision was analyzed by impairing the CNN's ability to learn the specific feature and examining how this impacts the quality of the CNN's classification. If the feature is not important for classifying the lesion, then the CNN should still make the correct decision, even if it can no longer identify the feature. The CNN's ability to learn a particular feature can be hampered by removing examples of that feature from its training set. Although repeatedly removing examples and retraining the model is prohibitively time-consuming, Koh et al. developed an approximation of this process that calculates an "influence function" [28]. The influence function of a feature with respect to a particular image estimates how much the probability of the correct lesion classification deteriorates for that image as examples of the feature are removed from the CNN's training set. Thus, the radiological feature that is most influential for classifying a particular lesion is the feature with the largest influence function for that image. Scores were obtained for each feature by measuring their respective influence functions, then dividing each by the sum of the influences. No ground truth was used for the optimal weighting of radiological features for diagnosing a given image, since a CNN does not "reason" about radiological features in the same way as a radiologist. The definition and further interpretation of the influence function are provided in Supplement 1.

Results

Characteristics of the 296 patients included in this study are described in Part I of this article series. CNN model classification performance is also described in detail in Part I.

Feature identification with probabilistic inference

A total of 224 annotated images were used across the 14 radiological features, and some images were labeled with multiple features. After being presented with a randomly selected subset of 140 out of 224 sample lesions, the model obtained a PPV of 76.5 ± 2.2% and Sn of 82.9 ± 2.6% in identifying the 1–4 correct radiological features for the 60 manually labeled test lesions over 20 iterations (see Table 2).

Among individual features, the model was most successful at identifying relatively simple enhancement patterns. With a mean number of 2.6 labeled features per lesion, the model achieved a precision of 76.5 ± 2.2% with a recall of 82.9 ± 2.6% (see Table 3). It achieved the best performance at identifying arterial phase hyperenhancement (PPV = 91.2%, Sn = 90.3%), hyperenhancing mass on delayed phase (PPV = 93.0%, Sn = 100%), and thin-walled mass (PPV = 86.5%, Sn = 100%). In contrast, the model performed relatively poorly on more complex features, struggling to identify nodularity (PPV = 62.9%, Sn = 60.8%) and infiltrative appearance (PPV = 33.0%, Sn = 45.0%). The CNN also overestimated the frequency of central scars (PPV = 32.0%, Sn = 80.0%), which only appeared once among the 60 test lesions.

The model misclassified lesions with higher frequency when the radiological features were also misclassified. For

Table 2 Precision and recall of the model for determining individual radiological features present in lesion images

Radiological feature	Positive predictive value (mean ± SD)	Sensitivity (mean ± SD)
Arterial phase hyperenhancement	91.2 ± 5.6%	90.3 ± 3.8%
Central scar	32.0 ± 21.7%	80.0 ± 40.0%
Enhancing rim (CRC metastasis), capsule/pseudocapsule (HCC)	74.8 ± 7.5%	75.3 ± 8.7%
Heterogeneous lesion	64.9 ± 4.8%	75.6 ± 5.4%
Hyperenhancing mass on delayed phase	93.0 ± 6.2%	100.0 ± 0.0%
Hypoenhancing core (CRC metastasis), hypoenhancing mass (cyst)	82.4 ± 4.5%	71.3 ± 11.8%
Infiltrative appearance	33.0 ± 11.3%	45.0 ± 10.0%
Iso-intensity on venous and delayed phase	69.5 ± 8.7%	92.2 ± 9.4%
Nodularity	62.9 ± 14.0%	60.8 ± 22.5%
Nodular/discontinuous peripheral hyperenhancement	80.3 ± 10.0%	94.0 ± 7.3%
Progressive centripetal filling	73.7 ± 8.5%	95.0 ± 5.5%
Progressive hyperenhancement	87.1 ± 5.4%	92.6 ± 3.9%
Thin-walled mass	86.5 ± 8.5%	100.0 ± 0.0%
Washout	67.4 ± 10.0%	66.7 ± 9.3%
Overall	76.5 ± 2.2%	82.9 ± 2.6%

the 12% of lesions that the model misclassified over 20 iterations, its PPV and Sn were reduced to 56.6% and 63.8%, respectively. Furthermore, the feature that the model predicted with the highest likelihood was only correct in 60.4% of cases—by comparison, the feature that the model predicts with the greatest likelihood in correctly classified lesions was correct 85.6% of the time.

This effect was also observed when the feature identification metrics are grouped by lesion classes, as the model generally identified features most accurately for classes in which the lesion entity itself was classified with high accuracy. The model obtained the highest PPV for benign cyst features at 100% and lowest for CRC metastasis features at 61.2%. The model attained the highest sensitivity for hemangioma features at 96.1% and lowest for HCC features at 64.2%. The lesion classifier performed better on both cysts (Sn = 99.5%, Sp = 99.9%) and hemangiomas (Sn = 93.5%, Sp = 99.9%) relative to HCCs (Sn = 82.0%, Sp = 96.5%) and CRC metastases (Sn = 94.0%, Sp = 95.9%).

Feature mapping with weighted activations

The feature maps (Fig. 4) were consistent with radiological features related to borders: enhancing rim and capsule/pseudocapsule, and a thin wall yield feature maps that trace these structures. Additionally, the model's feature maps for hypoenhancing and hyperenhancing masses were well localized and consistent with their location in the original image: hypoenhancing core/mass and nodularity had fairly well-defined bounds, as did arterial phase hyperenhancement and hyperenhancing mass in delayed phase. Iso-intensity in venous/delayed phase was also well defined, capable of excluding the hyperenhancing vessels in its map. In contrast, features describing enhancement patterns over time were more diffuse and poorly localized. There was slight misregistration between phases included in the hemangioma example, contributing to artifacts seen in the feature map for nodular peripheral hyperenhancement.

Table 3 Precision and recall of the model for determining the radiological features present in test images grouped by lesion class

Lesion class	Mean number of labeled features per lesion	Precision (mean ± SD)	Recall (mean ± SD)
Benign cyst	1.8	100.0 ± 0.0%	94.7 ± 7.1%
Cavernous hemangioma	2.7	81.9 ± 3.4%	96.1 ± 3.2%
Focal nodular hyperplasia	2.0	77.1 ± 7.7%	95.0 ± 5.7%
Hepatocellular carcinoma	3.2	83.5 ± 5.0%	64.2 ± 6.9%
Intrahepatic cholangiocarcinoma	3.0	69.3 ± 4.0%	83.3 ± 5.2%
Colorectal carcinoma metastasis	2.7	61.2 ± 4.9%	74.4 ± 7.0%
Overall	2.6	76.5 ± 2.2%	82.9 ± 2.6%

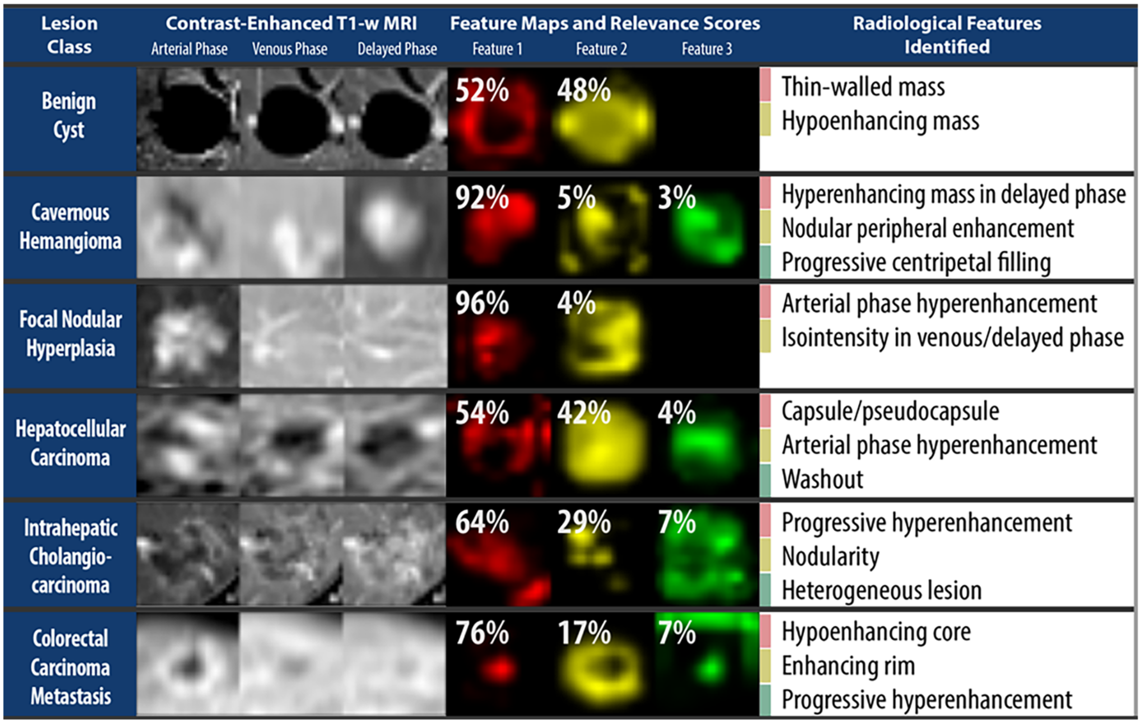

Lesion Class
Contrast-Enhanced T1-w MRI
Arterial Phase
Venous Phase
Delayed Phase
Feature Maps and Relevance Scores
Feature 1
Feature 2
Feature 3
Radiological Features Identified
Benign Cyst
52%
48%
Thin-walled mass
Hypoenhancing mass
Cavernous Hemangioma
92%
5%
3%
Hyperenhancing mass in delayed phase
Nodular peripheral enhancement
Progressive centripetal filling
Focal Nodular Hyperplasia
96%
4%
Arterial phase hyperenhancement
Isointensity in venous/delayed phase
Hepatocellular Carcinoma
54%
42%
4%
Capsule/pseudocapsule
Arterial phase hyperenhancement
Washout
Intrahepatic Cholangio-carcinoma
64%
29%
7%
Progressive hyperenhancement
Nodularity
Heterogeneous lesion
Colorectal Carcinoma Metastasis
76%
17%
7%
Hypoenhancing core
Enhancing rim
Progressive hyperenhancement

Fig. 4 2D slices of the feature maps and relevance scores for examples of lesions from each class with correctly identified features. The color and ordering of the feature maps correspond to the ranking of the feature relevance scores, with the most relevant feature's map in red. The feature maps are created based on the entire MRI sequence, and do not correspond directly to a single phase. These results are taken from a single iteration

Feature scoring with influence functions

The most relevant radiological feature for cavernous hemangiomas was progressive centripetal filling, with a score of 48.6% compared with 34.0% for hyperenhancing mass on delayed phase and 21.6% for nodular/discontinuous peripheral hyperenhancement. Thin-walled mass was a more relevant feature for classifying benign cysts than hypoenhancing mass (67.1% vs. 46.6%; Table 4). The most relevant feature for correctly classifying FNHs was iso-intensity on venous/delayed phase (79.4%), followed by arterial phase hyperenhancement (65.8%) and central scar (37.4%). The relevance scores for HCC imaging features were 49.5% for capsule/pseudo-capsule, 48.5% for heterogeneous lesion, 40.3% for washout, and 38.4% for arterial phase hyperenhancement. The relevance scores for ICC imaging features were 58.2% for progressive hyperenhancement, 47.3% for heterogeneous lesion, 43.8% for infiltrative appearance, and 37.2% for

Table 4 Features ranked by mean relevance score for the features for test lesions in each class. Percentages do not sum to 100% because some lesions only have a subset of the features listed above

Lesion class	Feature 1	Feature 2	Feature 3	Feature 4
Benign cyst	Thin-walled mass (67.1%)	Hypoenhancing mass (46.6%)	N/A	N/A
Cavernous hemangioma	Progressive centripetal filling (48.6%)	Hyperenhancing mass on delayed phase (34.0%)	Nodular/discontinuous peripheral hyperenhancement (21.6%)	N/A
Focal nodular hyperplasia	Isointense on venous/delayed phase (79.4%)	Arterial phase hyperenhancement (65.8%)	Central scar (37.4%)	N/A
Hepatocellular carcinoma	Capsule/pseudo-capsule (49.5%)	Heterogeneous lesion (48.5%)	Washout (40.3%)	Arterial phase hyperenhancement (38.4%)
Intrahepatic cholangiocarcinoma	Progressive hyperenhancement (58.2%)	Heterogeneous lesion (47.3%)	Infiltrative appearance (43.8%)	Nodularity (37.2%)
Colorectal carcinoma metastasis	Progressive hyperenhancement (67.2%)	Hypoenhancing core (52%)	Enhancing rim (46.9%)	N/A

nodularity. The most relevant imaging feature for correctly classifying CRC metastases was progressive hyperenhancement (67.2%), followed by hypoenhancing core (52.0%) and enhancing rim (46.9%).

Discussion

This study demonstrates the development of a proof-of-concept prototype for the automatic identification, mapping, and scoring of radiological features within a DL system, enabling radiologists to interpret elements of decision-making behind classification decisions. While DL algorithms have the opportunity to markedly enhance the clinical workflow of diagnosis, prognosis, and treatment, transparency is a vital component. Indeed, it is unlikely that clinicians would accept automated diagnostic decision support without some measure of "evidence" to justify predictions. The method of identifying and scoring radiological features allows the algorithm to communicate factors used in making predictions. Radiologists can then quickly validate these features by using feature maps or similar interpretability techniques to check whether the system has accurately identified the lesion's features in the correct locations.

The CNN was able to identify most radiological features fairly consistently despite being provided with a small sample of lesions per class, in addition to being trained to perform an entirely different task (classifying the lesion entity in Part I). For many simple imaging features such as hyperenhancing or hypoenhancing masses, the model was able to accurately and reliably determine its presence, location, and contribution to the lesion classification. However, it had greater difficulty identifying or localizing features that consist of patterns over multiple phases than patterns that are visible from a single phase or constant across all phases. It struggled in particular on more complex features that may appear quite variable across different lesions such as infiltrative appearance, suggesting that these features are not well understood by the CNN or that more examples of these features need to be provided. By highlighting which radiological features the CNN fails to recognize, this system may provide engineers with a path to identify possible failure modes and fine-tune the model, for example, by training it on more samples with these features.

A general relationship was observed between the model's misclassification of a lesion entity and its misidentification of radiological features, which could provide researchers and clinicians with the transparency to identify when and how a CNN model fails. If the model predicts non-existent imaging features, clinicians will be aware that the model has likely made a mistake. Moreover, this gives developers an example of a potential failure mode in the model. An interpretable DL system can be utilized as a tool for validation of imaging guidelines, particularly for entities which are uncommon or have evolving imaging criteria, such as bi-phenotypic tumors and ICCs [12, 29, 30]. As shown in the results on feature scoring, the model tends to put greater weight on imaging features that have greater uniqueness and differential diagnostic power in the respective lesion class. An interpretable CNN could be initially presented with a large set of candidate imaging features. Then by selecting the imaging features with the highest relevance score output by the model, one could determine which features are most relevant to members of a given lesion class. This approach also addresses the need for more quantitative evidence-based data in radiology reports.

An interpretable DL system could help to address the large number of ancillary imaging features that are part of the LI-RADS guidelines and similar systems by providing feedback on the importance of various radiological features in performing differential diagnosis. With further refinements, the presented concepts could potentially be used to validate newly proposed ancillary features in terms of frequency of occurrence, by applying it to a large cohort and analyzing the CNN's predictions. Features that are predicted with low frequency or relevance could be considered for exclusion from LI-RADS guidelines. This could be a first step towards providing a more efficient and clinically practical protocol [13, 19]. An interpretable DL model could also enable the automated implementation of such complex reporting systems as LI-RADS, by determining and reporting standardized descriptions of the radiological features present. By enabling such systems to become widely adopted, there is potential for the reporting burden on radiologists to be alleviated, data quality to improve, and the quality and consistency of patient diagnosis to increase.

Since the present study is designed as a proof-of-concept development, there are multiple limitations that future studies will address. As a single-institution study with limited data availability, a relatively small number of sample lesions was included for each lesion type. This will be remedied by eventually utilizing larger multi-institutional datasets. In addition, while feature extraction could be easily validated with ground truth confirmation by radiological readers, there is intrinsically no existing ground truth criteria for validating feature maps and relevance scores. As a result, more formal validation of these elements will require an aggregate of forthcoming studies that demonstrate reproducibility under different DL models and datasets. Such a system would also need to demonstrate similar functionality using different choices of radiological features and lesion types. Future work will demonstrate this technique on LI-RADS ancillary features, which will require incorporating a more complex CNN model capable of analyzing other types of MRI sequences.

In summary, this study demonstrates a proof-of-concept interpretable deep learning system for clinical radiology. This provides a technique for interrogating relevant portions of an existing CNN, offering rationale for classifications through internal analysis of relevant imaging features. With further refinement and validation, such methods have the potential to eventually provide a cooperative approach for radiologists to interact with deep learning systems, facilitating clinical translation into radiology workflows. Transparency and comprehensibility are key barriers towards the practical integration of deep learning into clinical practice [31]. An interpretable approach can serve as a model for addressing these issues as the medical community works to translate useful aspects of deep learning into clinical practice.

Funding BL and CW received funding from the Radiological Society of North America (RSNA Research Resident Grant No. RR1731). JD, JC, ML, and CW received funding from the National Institutes of Health (NIH/NCI R01 CA206180).

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Julius Chapiro.

Conflict of interest The authors of this manuscript declare relationships with the following companies: JW: Bracco Diagnostics, Siemens AG; ML: Pro Medicus Limited; JC: Koninklijke Philips, Guerbet SA, Eisai Co.

Statistics and biometry One of the authors has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- retrospective
- experimental
- performed at one institution

References

1. Rajpurkar P, Irvin J, Zhu K et al (2018) Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15:e1002686. https://doi.org/10.1371/journal.pmed.1002686
2. Chartrand G, Cheng PM, Vorontsov E et al (2017) Deep learning: a primer for radiologists. Radiographics 37:2113–2131
3. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S (2016) Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging 35:1207–1216
4. Greenspan H, Van Ginneken B, Summers RM (2016) Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging 35:1153–1159
5. Hamm CA, Wang CJ, Savic LJ et al (2019) Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol. https://doi.org/10.1007/s00330-019-06205-9
6. Olden JD, Jackson DA (2002) Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecol Model 154:135–150
7. Kiczales G (1996) Beyond the black box: open implementation. IEEE Softw 13(8):10–11
8. Dayhoff JE, DeLeo JM (2001) Artificial neural networks: opening the black box. Cancer 91:1615–1635
9. Olah C, Satyanarayan A, Johnson I et al (2018) The building blocks of interpretability. Distill 3:e10. https://doi.org/10.23915/distill.00010
10. Corwin MT, Lee AY, Fananapazir G, Loehfelm TW, Sarkar S, Sirlin CB (2018) Nonstandardized terminology to describe focal liver lesions in patients at risk for hepatocellular carcinoma: implications regarding clinical communication. AJR Am J Roentgenol 210:85–90
11. Mitchell DG, Bruix J, Sherman M, Sirlin CB (2015) LI-RADS (Liver Imaging Reporting and Data System): summary, discussion, and consensus of the LI-RADS Management Working Group and future directions. Hepatology 61:1056–1065
12. Mitchell DG, Bashir MR, Sirlin CB (2018) Management implications and outcomes of LI-RADS-2, -3, -4, and -M category observations. Abdom Radiol (NY) 43:143–148
13. Barth B, Donati O, Fischer M et al (2016) Reliability, validity, and reader acceptance of LI-RADS-an in-depth analysis. Acad Radiol 23:1145
14. Davenport MS, Khalatbari S, Liu PS et al (2014) Repeatability of diagnostic features and scoring systems for hepatocellular carcinoma by using MR imaging. Radiology 272:132
15. Ehman EC, Behr SC, Umetsu SE et al (2016) Rate of observation and inter-observer agreement for LI-RADS major features at CT and MRI in 184 pathology proven hepatocellular carcinomas. Abdom Radiol (NY) 41:963–969
16. Zhang YD, Zhu FP, Xu X et al (2016) Classifying CT/MR findings in patients with suspicion of hepatocellular carcinoma: comparison of liver imaging reporting and data system and criteria-free Likert scale reporting models. J Magn Reson Imaging 43:373–383
17. Bashir M, Huang R, Mayes N et al (2015) Concordance of hypervascular liver nodule characterization between the organ procurement and transplant network and liver imaging reporting and data system classifications. J Magn Reson Imaging 42:305
18. Liu W, Qin J, Guo R et al (2017) Accuracy of the diagnostic evaluation of hepatocellular carcinoma with LI-RADS. Acta Radiol. https://doi.org/10.1177/0284185117716700:284185117716700
19. Fowler KJ, Tang A, Santillan C et al (2018) Interreader reliability of LI-RADS version 2014 algorithm and imaging features for diagnosis of hepatocellular carcinoma: a large international multireader study. Radiology 286:173–185
20. Cruite I, Santillan C, Mamidipalli A, Shah A, Tang A, Sirlin CB (2016) Liver imaging reporting and data system: review of ancillary imaging features. Semin Roentgenol 51:301–307. https://doi.org/10.1053/j.ro.2016.05.004
21. Sirlin CB, Kielar AZ, Tang A, Bashir MR (2018) LI-RADS: a glimpse into the future. Abdom Radiol (NY) 43:231–236
22. Kim YY, An C, Kim S, Kim MJ (2017) Diagnostic accuracy of prospective application of the Liver Imaging Reporting and Data System (LI-RADS) in gadoxetate-enhanced MRI. Eur Radiol. https://doi.org/10.1007/s00330-017-5188-y
23. Molnar C (2019) Interpretable machine learning. A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/
24. Fisher A, Rudin C, Dominici F (2018) Model class reliance: variable importance measures for any machine learning model class, from the "Rashomon" perspective. arXiv preprint arXiv:180101489
25. Federle MP, Jeffrey RB, Woodward PJ, Borhani A (2009) Diagnostic imaging: abdomen. Published by Amirsys. Lippincott Williams & Wilkins

26. Victoria C, Sirlin CB, Cui J et al (2018) LI-RADS v2018 CT/MRI Manual. Available via https://www.acr.org/-/media/ACR/Files/Clinical-Resources/LIRADS/Chapter-16-Imaging-features.pdf?la=en
27. Everitt BS (2002) The Cambridge dictionary of statistics. Cambridge University Press
28. Koh PW, Liang P (2017) Understanding black-box predictions via influence functions. AarXiv preprint arXiv:170304730
29. Narsinh KH, Cui J, Papadatos D, Sirlin CB, Santillan CS (2018) Hepatocarcinogenesis and LI-RADS. Abdom Radiol (NY) 43:158–168
30. Tang A, Bashir MR, Corwin MT et al (2018) Evidence supporting LI-RADS major features for CT- and MR imaging-based diagnosis of hepatocellular carcinoma: a systematic review. Radiology 286:29–48
31. Holzinger A, Biemann C, Pattichis CS, Kell DB (2017) What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:171209923

Supplement 1

The optimal neural network parameters $\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, \theta)$ minimize the loss function L averaged over training images $\{x_i\}_{i \in \{1,\dots,n\}}$ with corresponding lesion classes y_i. The perturbed parameters θ^*_{pert} were defined as the optimal network parameters when it is trained with a reweighted loss function, in which the loss for a particular training datapoint (x, y) is downweighted by an amount ϵ.

$$\theta^*_{pert}(x, y, \epsilon) = \underset{\theta \in \Theta}{min} \left(\frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, \theta) - \epsilon L(x, y, \theta) \right)$$

Choosing $\epsilon = 0$ yields the unperturbed optimum θ^*, while choosing $\epsilon = \frac{1}{n}$ is equivalent to removing the datapoint from the training set. Thus as ϵ increases within this range, the perturbed network faces a lower penalty for misclassifying the downweighted datapoint, and so it may become harder to correctly classify test cases that are similar to this datapoint. In particular, if the training datapoint (x, y) is helpful for predicting some test datapoint (**x**,**y**), then $L($**x**,**y**$, \theta^*)$ is likely to be less than $L($**x**,**y**$, \theta^*_{pert}(x, y, \epsilon))$ even for small values of ϵ. To quantify this effect, the influence of a training datapoint on a test datapoint can be defined by taking the limit that $\epsilon \to 0$:

$$I(\mathbf{x}, \mathbf{y}, x, y) = \frac{dL(\mathbf{x}, \mathbf{y}, \theta^*_{pert}(x, y, \epsilon))}{d\epsilon} |_{\epsilon=0}$$

The influence function can be extended to represent not just the effect of a single training datapoint, but the influence of a radiological feature as a whole. Specifically, downweighting training datapoints that exemplify a particular feature is likely to hinder the model's ability to recognize that feature. By measuring the deterioration in classification performance as the network loses its ability to detect a particular feature, the feature-level influence function can quantify to what extent different features contribute to a particular prediction. Hence, an influence function was defined with respect to each of the radiological features $\{f_j\}_{j\in\{1,\ldots,14\}}$ by selecting a subset of training datapoints (with indices denoted by $J(f_j) \subseteq \{1,\ldots,n\}$) that corresponded to examples of the feature, and taking the average of their individual influence functions:

$$I_f(\mathbf{x},\mathbf{y},f_j) = \frac{1}{|J(f_j)|}\sum_{k\in J(f_j)} I(\mathbf{x},\mathbf{y},x_k,y_k)$$

When the influence function is large, this suggests that removing even a few examples of a feature from the training set would compromise the model's ability to correctly classify other lesions with the feature, and hence that feature is assigned a higher relevance score. This study computed an approximation to the influence function as derived by Koh et al. [26]. An intuitive description of the implemented approach is described in the body of the paper under “Feature scoring with influence functions”.

12) ACKNOWLEDGEMENT

I would like to extend my sincere thanks to my supervisor ***Prof. dr hab. n. med. Wojciech Poncyljusz****, who consistently supported and motivated me while I was carrying out the work that led to this dissertation. I greatly appreciate the expertise that he has imparted to me, and I particularly wish to thank* ***dr hab. n. med. Elżbieta Petriczko*** *for introducing me to the PhD program at the Pomeranian Medical University.*

I also wish to thank my supervisors at the Interventional Oncology Department in the Yale School of Medicine, ***Dr. Brian Letzen and Dr. Julius Chapiro****, for introducing me to their research groups and for assigning me to such an interesting and challenging scientific topic. Their valuable support in the planning and implementation of this scientific work will always be appreciated.*

The dissertation would not have been possible without the support and advice of ***Mr. Clinton Wang and Dr. Brian Letzen****. I would like to extend my deepest gratitude to them for their never-failing support. Their abundance of ideas and their willingness to discuss gave great impetus to my continuation and completion of this project. Both of these colleagues regularly challenged and enriched my ideas by involving me in constructive and goal-oriented conversations.*

I am also grateful to ***Dr. Lynn Jeanette Savic and Dr. Julius Chapiro*** *for selecting me as a candidate for the "Rising Star" research exchange program and for sharing their scientific expertise with me.*

I gratefully acknowledge the assistance of ***Mrs. Olena Voznyak****, who guided and helped me while I was developing and setting out my individual study plan.*

Last but not least, I cannot begin to express my thanks to my ***family and friends****, who have kept me motivated, focused and self-confident. This dissertation would not have been possible without their changeless support and love. For this, I am deeply indebted to all of them.*