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.

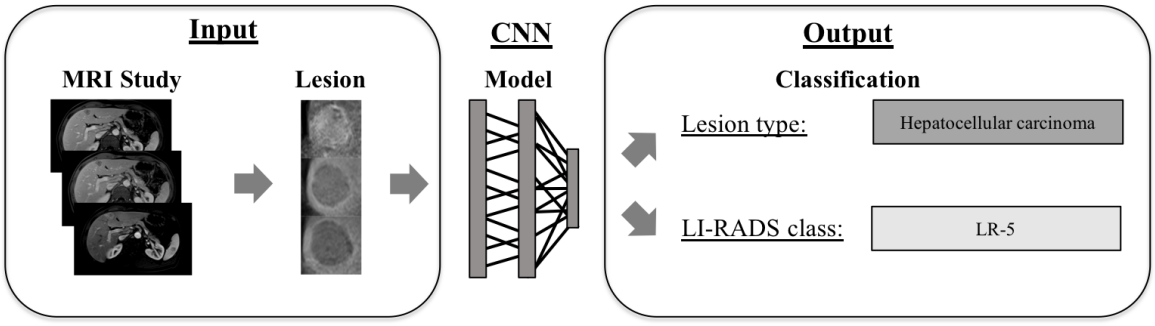
5.1 DEEP LEARNING MODEL

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

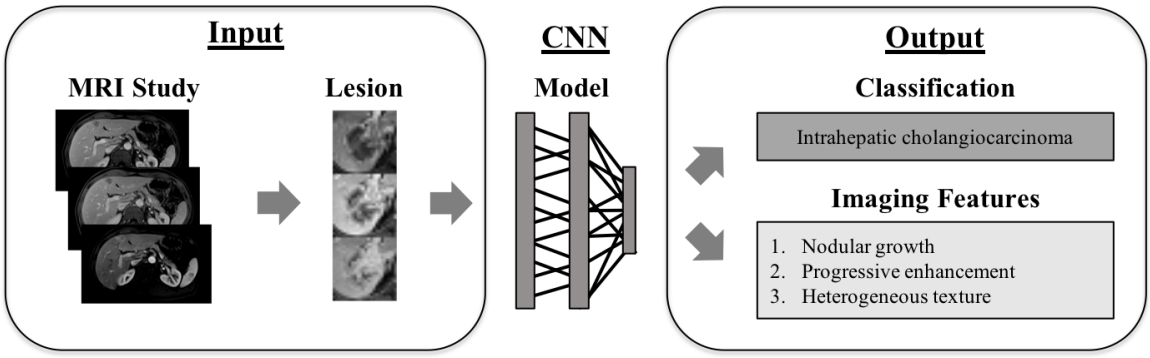
In the reader study (described above), the lesions could be classified. The model yielded a mean accuracy of 90% (55/60 lesions), while the two radiologists assessing the same lesions achieved respective accuracies of 80% (48/60) and 85% (51/60). For the three broader categories, the model gave an accuracy 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.



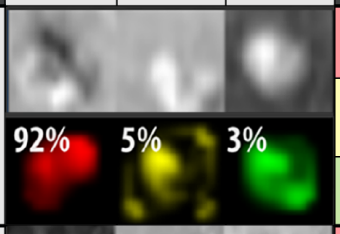
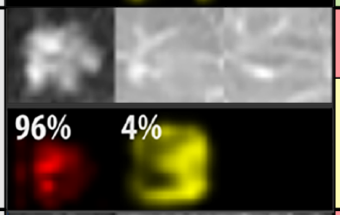
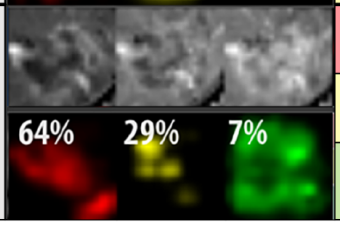
In its assessment of individual features, the CNN performed best for the simpler enhancement patterns. Presented with 2.6 labelled features on average per lesion, its performance was as summarized in **Tab. 4**. For simpler image features (e.g. arterial-phase hyperenhancement, hyperenhancing mass on 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% 5% 3%	Hyperenhancing mass in delayed phase Nodular peripheral enhancement Progressive centripetal filling
FNH				96% 4%	Arterial phase hyperenhancement Isointensity in venous/delayed phase
ICC				64% 29% 7%	Progressive hyperenhancement Nodularity 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

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Hamm, C.A., Wang, C.J., Savic, L.J. et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29, 3338–3347 (2019). https://doi.org/10.1007/s00330-019-06205-9

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

Wang, C.J., Hamm, C.A., Savic, L.J. et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29, 3348–3357 (2019). https://doi.org/10.1007/s00330-019-06214-8

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.*