ORIGINAL ARTICLE

Retrospective Study

Multiphase convolutional dense network for the classification of focal liver lesions on dynamic contrast-enhanced computed tomography

Su-E Cao, Lin-Qi Zhang, Si-Chi Kuang, Wen-Qi Shi, Bing Hu, Si-Dong Xie, Yi-Nan Chen, Hui Liu, Si-Min Chen, Ting Jiang, Meng Ye, Han-Xi Zhang, Jin Wang

ORCID number: Su-E Cao 0000-0002-0756-1957; Lin-Qi Zhang 0000-0002-0607-6300; Si-Chi Kuang 0000-0003-3674-651X; Wen-Qi Shi 0000-0003-2497-3299; Bing Hu 0000-0002-8270-433X; Si-Dong Xie 0000-0003-1280-5706; Yi-Nan Chen 0000-0003-0858-2087; Hui Liu 0000-0001-6218-8123; Si-Min Chen 0000-0001-7073-1472; Ting Jiang 0000-0002-6630-3392; Meng Ye 0000-0003-2210-3396; Han-Xi Zhang 0000-0001-9489-3062; Jin Wang 0000-0002-7956-9579.

Su-E Cao, Lin-Qi Zhang, Si-Chi Kuang, Wen-Qi Shi, Bing Hu, Si-Dong Xie, Si-Min Chen, Ting Jiang, Han-Xi Zhang, Jin Wang, Department of Radiology, The Third Affiliated Hospital, Sun Yat-Sen University, Guangzhou 510630, Guangdong Province, China

Yi-Nan Chen, Hui Liu, Meng Ye, Department of Scientific and Technological Research, 12 Sigma Technologies, Beijing 100102, China

Corresponding author: Jin Wang, MD, Doctor, Professor, Department of Radiology, The Third Affiliated Hospital, Sun Yat-Sen University, No. 600, Tianhe Road, Tianhe District, Guangzhou 510630, Guangdong Province, China. wangjin3@mail.sysu.edu.cn

Abstract

BACKGROUND
The accurate classification of focal liver lesions (FLLs) is essential to properly guide treatment options and predict prognosis. Dynamic contrast-enhanced computed tomography (DCE-CT) is still the cornerstone in the exact classification of FLLs due to its noninvasive nature, high scanning speed, and high-density resolution. Since their recent development, convolutional neural network-based deep learning techniques has been recognized to have high potential for image recognition tasks.

AIM
To develop and evaluate an automated multiphase convolutional dense network (MP-CDN) to classify FLLs on multiphase CT.

METHODS
A total of 517 FLLs scanned on a 320-detector CT scanner using a four-phase DCE-CT imaging protocol (including precontrast phase, arterial phase, portal venous phase, and delayed phase) from 2012 to 2017 were retrospectively enrolled. FLLs were classified into four categories: Category A, hepatocellular carcinoma (HCC); category B, liver metastases; category C, benign non-inflammatory FLLs including hemangiomas, focal nodular hyperplasias and adenomas; and category D, hepatic abscesses. Each category was split into a training set and test set in an

approximate 8:2 ratio. An MP-CDN classifier with a sequential input of the four-phase CT images was developed to automatically classify FLLs. The classification performance of the model was evaluated on the test set; the accuracy and specificity were calculated from the confusion matrix, and the area under the receiver operating characteristic curve (AUC) was calculated from the SoftMax probability outputted from the last layer of the MP-CDN.

RESULTS
A total of 410 FLLs were used for training and 107 FLLs were used for testing. The mean classification accuracy of the test set was 81.3% (87/107). The accuracy/specificity of distinguishing each category from the others were 0.916/0.964, 0.925/0.905, 0.860/0.918, and 0.925/0.963 for HCC, metastases, benign non-inflammatory FLLs, and abscesses on the test set, respectively. The AUC (95% confidence interval) for differentiating each category from the others was 0.92 (0.837-0.992), 0.99 (0.967-1.00), 0.88 (0.795-0.955) and 0.96 (0.914-0.996) for HCC, metastases, benign non-inflammatory FLLs, and abscesses on the test set, respectively.

CONCLUSION
MP-CDN accurately classified FLLs detected on four-phase CT as HCC, metastases, benign non-inflammatory FLLs and hepatic abscesses and may assist radiologists in identifying the different types of FLLs.

Key words: Deep learning; Convolutional neural networks; Focal liver lesions; Classification; Multiphase computed tomography; Dynamic enhancement pattern

Core tip: We developed and evaluated a deep learning-based convolutional neural network (CNN) to classify focal liver lesions (FLLs) on multiphase computed tomography. The most important highlight of the current study is that, to the best of our knowledge, this study is the first to employ four-channel input data to preserve the dynamic enhancement properties. The combination of the lesion's dynamic enhancement pattern with a CNN can imitate the image diagnosis of radiologists and is expected to improve diagnostic accuracy. It was interesting to note that the accuracy and specificity of differentiating each category from others were high. This model may become an efficient tool to assist radiologists in the classification of FLLs.

INTRODUCTION

The frequency of detection of focal liver lesions (FLLs) has increased due to the widespread application of imaging techniques[1,2]. Because the treatment of FLLs depends on the nature of the lesion, the ability to accurately distinguish the types of FLLs is an important step in the management of these patients. Currently, dynamic contrast-enhanced computed tomography (DCE-CT) is commonly used for the noninvasive detection and characterization of FLLs due to its high scanning speed and high-density resolution[3,4]. The appearances, especially the dynamic enhancement patterns of FLLs on CT imaging, are essential for categorizing lesions. With the careful evaluation of CT images, diagnosis with a relatively high accuracy can be achieved for most liver lesions. However, in current clinical practice, the evaluation of CT images is mainly performed by radiologists. The results are influenced by the radiologist's experience and are generally subjective. Radiologists have began investigating the potential of computer-aided diagnostic systems to overcome these limitations. Rather

than using qualitative reasoning, artificial intelligence (AI) conducts quantitative assessments by automatically identifying imaging information[5]. Therefore, AI can assist radiologists in making more accurate imaging diagnoses and substantially reduces the radiologists' workload.

Traditional machine learning algorithms need features to be predefined and require the placement of complexly shaped regions of interest (ROIs) on images[6-8]. The predefined features are applied in various combinations to effectively determine the diagnosis using traditional machine learning algorithms, but the combinations are usually incomprehensive and result in low accuracy. Today, deep learning-based algorithms are widely used due to their automatic feature generation and image classification abilities[9,10]. A convolutional neural network (CNN) is considered the first truly successful deep-learning method based on a multilayer hierarchical network, and shows high performance in the image analysis field[9-11]. CNN has been successfully applied to analyze the medical images of patients with many diseases such as pulmonary tuberculosis, breast cancer, brain tumors, and some hepatic diseases[12-19]. However, few studies have attempted to apply CNN in the differential diagnosis of FLLs, and these studies have limited value. The dynamic enhancement pattern of FLLs is essential for making differential diagnoses and may have a complementary role to CNN in the diagnostic workup of FLLs.

Hence, we developed and evaluated an automated multiphase convolutional dense network (MP-CDN) that uses four channels of input data to classify FLLs on four-phase CT.

MATERIALS AND METHODS

Patients

The retrospective study was reviewed and approved by our institutional review board, and written informed consent was obtained from the patients whose data were analyzed. Two radiologists (Cao SE and Shi WQ, both with 5 years of experience in imaging diagnosis) searched for patients with FLLs in the picture archiving and communication system (PACS). The images of patients who underwent a four-phase DCE-CT examination and for whom FLLs were confirmed by histopathological evaluation or were diagnosed based on a combination of clinical and radiological findings with follow-up were collected for further screening. The exclusion criteria were as follows: Lesions larger than 10 cm; images with prominent artifacts; and prior local-regional therapy prior to the CT examination.

Standard of classification

The lesions were classified into four categories according to different pathological types and treatment decisions. (1) Category A was hepatocellular carcinoma (HCC), which was confirmed by histopathologic evaluation after surgery or biopsy. (2) Category B represents liver metastases derived from different primary sites such as colorectal cancer, gastric carcinoma, breast cancer, lung cancer, thyroid cancer, malignant jejunal stromal tumor, duodenal papillary carcinoma, and laryngocarcinoma. The primary lesions were confirmed by a pathological examination, but the metastatic lesions were diagnosed based on the clinical data, patient history, other follow-up CT, magnetic resonance imaging, and positron emission tomography/CT scans. For liver metastases, the follow-up time was 60 d to 1230 d, and the median was 300 d. (3) Category C was defined as benign non-inflammatory FLLs, including hemangiomas, focal nodular hyperplasias (FNHs), and adenomas. A total of 27 lesions, including all adenomas, were confirmed by a histopathological evaluation after surgery, while the remaining 135 lesions were diagnosed based on imaging diagnostic criteria from the CT scan in combination with the clinical information and follow-up MRI; the follow-up time was 90 d to 1800 d, and the median was 330 d. And (4) Category D was hepatic abscesses. The diagnosis of hepatic abscess was based on typical imaging findings, clinical aspects, laboratory findings, and microbiology on blood or aspirate culture results. While all patients received early empirical antibiotic treatment, 37% patients underwent percutaneous or surgical drainage. A longer follow-up with a median time of 100 d (range, 60-365 d) confirmed the remission or absence of signs and symptoms together with imaging studies without findings compatible with hepatic abscess after treatment.

Finally, a total of 375 patients with 517 lesions were enrolled in this study from 2012 to 2017. Each category was split into a training set and test set. Patients who underwent CT scan before June 2016 were used for training, while those after June

2016 were used for testing. The ratio between training set and test set was approximately 8:2.

Basic information about the patients was obtained from the hospital information system, including gender, age, surgical and pathological reports, lesion size, and follow-up time.

Input data: CT imaging protocol

A 320-detector CT scanner (Aquilion ONE; Toshiba Medical Systems, Otawara, Japan) was used to acquire four-phase DCE-CT imaging protocols including precontrast phase (PP), arterial phase (AP), portal venous phase (PVP), and delayed phase (DP). The following scan parameters were used: A peak tube voltage of 120 kV, a tube rotation time of 0.5 s per rotation, a pitch factor of 0.828, a field of view of 35 cm × 35 cm, a matrix of 512 × 512, and automatic tube current modulation.

The first phase was PP to cover the whole liver. The next three phases were contrast-enhanced phases with the same scanning range after the intravenous injection of low osmolar nonionic contrast medium (Ioversol-350; Tyco Healthcare, Montreal, Quebec, Canada and Isovue-370, Bracco Diagnostics, Guangzhou, China) into the right antecubital vein at an injection rate of 3 mL/s and a dose of 1.5 mL/kg body weight, followed by a 20-mL saline chaser.

The AP was acquired by performing a bolus tracking technique. The AP was scanned 15 s after CT attenuation of the aorta at the level of the diaphragm had reached 200 Hounsfield Units. For the PVP, images were acquired 30 s after the AP. The DP was scanned 45 s after the PVP. All images were reconstructed in the axial plane with a slice thickness of 5 mm and interval of 5 mm using a kernel for the evaluation of soft tissues (FC19) and then sent to the PACS.

Input data: CT imaging annotation

The CT imaging annotation was manually and independently performed by four radiologists (all had at least 4 years of imaging experience), and the results were reviewed by a radiologist with 20 years of imaging experience. For each patient, the four-phase CT images were manually loaded into 3D Slicer (https://www.slicer.org). The boundary of each lesion was manually drawn slice-by-slice along the visible borders of the lesion using the annotation module available in 3D Slicer. The classification of the type of each lesion was manually annotated using a home-developed lesion annotation module in 3D Slicer.

Input data: CT imaging processing pipeline

The four phases were organized in a sequence according to the acquisition time and fed into the image processing pipeline, as shown in Figure 1. The inner-phase registration and normalization were used to achieve volume-wise processing. The inner-phase registration was performed by using a nonrigid registration module implemented in Elastix (http://elastix.isi.uu.nl) with PVP as the reference phase, and then each phase was linearly normalized to (-1, 1) with a corresponding HU of (0, 300). Cropping and resizing were performed for lesion-wise processing using the Python library scikit-image 0.15.0 (https://scikit-image.org/scikit-image 0.15.0). For each lesion, a three-dimensional bounding box was generated to cover the lesion boundary and extended with a spare boundary of 10 mm along each direction. After extracting the bounding box of the lesion, ROIs were cropped from the PVP. The ROI was a square on each axial plane, the length of the side was 1.5 times the value of the longest side of the bounding box on the axial plane, and the center point was the projection of the center point of the bounding box on each axial plane. Then the bounding boxes were propagated on other phases to crop the lesion. Following lesion cropping, each cropped ROI was resized into an identical shape in the size of 128 × 128. ROIs from five slices centered at the lesion were extracted and stacked together to form a (128, 128, 5) tensor as the input data for each phase.

Deep convolutional network architecture

The deep convolutional network was designed following the concept of the automatic extraction of useful features from each phase and then the sequential combination of each phase's features to achieve classification, as detailed in Figure 2. Each phase's automatic feature extraction was implemented using a densely connected stack of two-dimensional convolutional, center-cropping and max-pooling layers, where the convolutional kernel size was 3 × 3; the cropping and pooling size was 2 × 2; and the activation layer used the "ReLU" activation function. Then, the four-phase convolutional layers were flattened and sequentially connected to the last dense layer
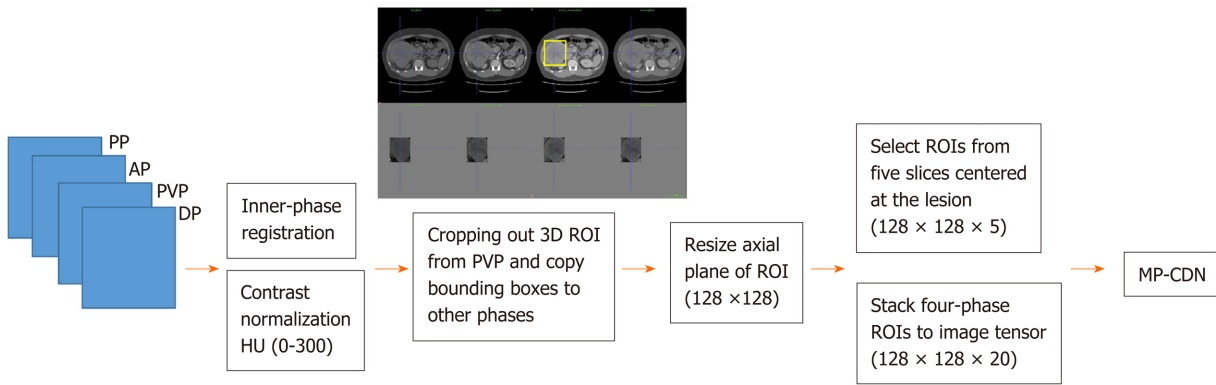
Figure 1 Four-phase images processing pipeline for multiphase convolutional dense network. AP: Arterial phase; DP: Delayed phase; HU: Hounsfield unit; MD-CDN: Multiphase convolutional dense network; PP: Precontrast phase; PVP: Portal venous phase; ROI: Region of interest.
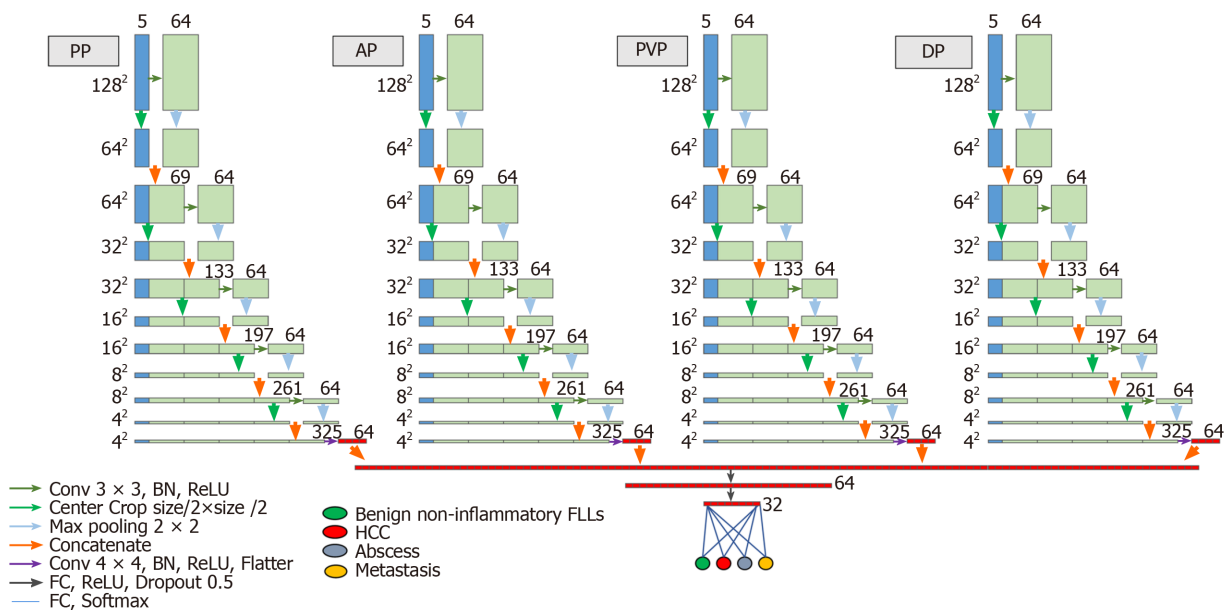
Figure 2 Architecture of the proposed multiphase convolutional dense network. AP: Arterial phase; DP: Delayed phase; FLLs: Focal liver lesions; HCC: Hepatocellular carcinoma; PP: Precontrast phase; PVP: Portal venous phase.

with SoftMax activation for classification purposes. The sequential connection of each phase's CNN network block was designed to preserve the dynamic enhancement properties.

The deep convolutional network was a 2.5 D MP-CDN with the four phases of resized multichannel images as the input (the slice was used as the channel dimension in this network). The classification tasks consisted of training and testing, in which the training task was performed with a batch size of 100 and the test task was performed once for each lesion.

Training and evaluation

For the training set, data augmentation options, which include scaling and rotation, were applied to each ROI. An augmented training dataset with a size 21 times greater than the raw dataset was used to train the model. The test set without augmentation was directly used to assess the model.

During the training phase, the category label was converted to 0.0 or 1.0 as the SoftMax probability to train the model. During the testing phase, the category label included the binary label and probability label, where the binary label was 1.0 or 0.0 corresponding to the class with the largest or non-largest probability from the SoftMax layer. In terms of probability label, the result was derived from the SoftMax probability outputted from the last layer of the MP-CDN.

Model implementation

The model was programmed using Python3.7 (https://www.python.org/) under the deep learning model development framework of Keras (https://keras.io) with the TensorFlow (https://www.tensorflow.org) backend. The network weights were optimized using the Adam optimizer, the learning rate was 0.00001 and the loss function was categorical cross-entropy. A graphics processing unit (GPU) (NVIDIA Titian 1080Ti) was used to accelerate the model training and testing phases.

Statistics

The distributions of age, sex, and lesion size in each of the sets (training and test sets) were compared using SPSS 17.0 software (SPSS Inc., Chicago, IL, United States). Quantitative variables were compared using the Wilcoxon rank sum test or *t*-test, and qualitative variables were compared using the chi-squared test.

The classification performance of the model was assessed on the test set: The accuracy, specificity, and sensitivity for differentiating each category from the others were calculated from the confusion matrix from the confusion matrix, and the area under the receiver operating characteristic (ROC) curve (AUC) was calculated from the SoftMax probability outputted from the last layer of the MP-CDN using SPSS 17.0 Software.

The model was further evaluated by applying a "phase cheating" experiment on the test set. The "phase cheating" experiment was implemented by eliminating one or more phases from the four phases and replacing it with the wrong phase(s) before feeding it into the model. The design idea of this experiment was based on the following concepts: (1) The liver lesion's dynamic enhancement pattern is vital in differential diagnosis; (2) Our model was designed to accommodate the correct sequence of four phases, which preserved the dynamic enhancement properties; and (3) The "phase cheating" experiment was used to test whether our model had learned this important dynamic enhancement pattern. If the phases were replaced by a certain phase (the so-called "phase cheating" experiment), its dynamic enhancement pattern might be different and may result in an incorrect category prediction. We re-evaluated the classification performance by comparing the AUCs between the model in the normal set and that in the "phase cheating" sets by using MedCalc Software (version 11.4.2 for Windows, MedCalc Software bvba).

Statistical significance was defined as $P < 0.05$.

RESULTS

Of the 15680 patients with FLLs treated at our hospital from 2012 to 2017, 375 patients with 517 lesions met the inclusion criteria. Of the 517 FLLs, 410 FLLs (88 HCCs, 89 metastases, 128 benign non-inflammatory FLLs, and 105 abscesses) were used for training, and 107 FLLs (23 HCCs, 23 metastases, 34 benign non-inflammatory FLLs, and 27 abscesses) were used for testing. Table 1 presents the basic and detailed information of each dataset.

The confusion matrix analysis on the test set is shown in Table 2. Of the 23 HCCs, 17 lesions were correctly classified, 4 lesions were misclassified as benign non-inflammatory FLLs, and the remaining 2 lesions were misclassified as metastases. It was interesting to note that all metastases (23 lesions) were correctly classified. Of the 34 benign non-inflammatory FLLs, 25 lesions were correctly classified, 3 lesions were misclassified as HCC, 3 lesions were misclassified as metastases, and the remaining 3 lesions were misclassified as hepatic abscesses. Of the 27 hepatic abscesses, 22 lesions were correctly classified, 3 lesions were misclassified as metastases, and the remaining 2 lesions were misclassified as benign non-inflammatory FLLs. The representative correctly classified and misclassified examples of each category are shown in Figure 3. The accuracy/specificity/sensitivity of differentiating each category from others were 0.916/0.964/0.739, 0.925/0.905/1.0, 0.860/0.918/0.735 and 0.925/0.963/0.815 for HCC, metastases, benign non-inflammatory FLLs, and abscesses, respectively.

ROC analysis was performed on the test set. The AUC (95% confidence interval [CI]) for differentiating each category from the others was 0.92 (0.837-0.992), 0.99 (0.967-1.00), 0.88 (0.795-0.955) and 0.96 (0.914-0.996) for HCC, metastases, benign non-inflammatory FLLs, and abscesses, respectively (Figure 4A). The model's classification probability was calibrated for each category, as shown in Figure 4B, and the Brier scores were 0.104, 0.080, 0.124, and 0.074 for HCC, metastases, benign non-inflammatory FLLs, and hepatic abscesses, respectively.

Table 3 shows the AUC and *P* value when using the "phase cheating" sets

Table 1 The basic information and detail distribution of each dataset

		Training set	Test set	*P* value
Category A: HCC	No. of lesions/No. of patients	88/79	23/22	
	Age (median [range]) in yr	49 (24-81)	49.5 (33-70)	0.726
	Sex (percentage of women)	6/79 (7.6%)	5/22 (22.7%)	0.044
	Size of lesion (mean ± SD) in mm	60.6 ± 36.3	63.0 ± 45.4	0.789
	Histopathologic diagnosis (No. of lesions/No. of patients)			
	Surgery	79/70	20/19	
	Biopsy	9/9	3/3	
Category B: Metastases	No. of lesions/No. of patients	89/34	23/14	
	Age (Median [range]) (yr)	58.5 (23-79)	58 (23-79)	0.937
	Sex (Percentage of women)	8/34 (23.5%)	6/14 (42.9%)	0.181
	Size of lesion (mean ± SD) in mm	23.0 ± 13.9	22.7 ± 11.5	0.937
	Primary tumors (No. of lesions/No. of patients)			
	Colorectal cancer	40/20	10/6	
	Gastric carcinoma	13/3	3/2	
	Breast cancer	2/1	0/0	
	Lung cancer	14/4	4/2	
	Thyroid cancer	16/4	4/2	
	Malignant jejunal stromal tumor	2/1	1/1	
	Duodenal papillary carcinoma	0/0	1/1	
	Laryngocarcinoma	2/1	0/0	
Category C: Benign non-inflammatory FLLs	No. of lesions/No. of patients	128/97	34/32	
	Age (median [range]) in yr	34 (17-82)	34 (10-74)	0.729
	Sex (percentage of women)	52/97 (53.6%)	16/32 (50.0%)	0.723
	Size of lesion (mean ± SD) in mm	41.9 ± 30.5	52.9 ± 28.4	0.060
	Histological type (No. of lesions/No. of patients)			
	Hemangioma	55/35	15/15	
	FNH	67/58	17/15	
	Adenoma	6/4	2/2	
Category D: Hepatic abscesses	No. of lesions/No. of patients	105/77	27/20	
	Age (median [range]) in yr	54 (4-82)	55 (25-82)	0.936
	Sex (percentage of women)	24/77 (31.2%)	7/20 (35.0%)	0.743
	Size of lesion (mean ± SD) in mm	64.5 ± 34.9	63.8 ± 24.2	0.916

FLLs: Focal liver lesions; FNH: Focal nodular hyperplasias.

compared to the normal set. The AUCs were lower for the "phase cheating" set with eliminating AP and/or PVP than for the normal set in differentiating HCC from the others ($P < 0.05$). When we replaced PP with AP, there was no significant difference between the AUCs of the normal set and "phase cheating" sets in differentiating HCC from the others ($P > 0.05$). Figure 5 shows the heatmaps of the predicted category when using the "phase cheating" sets compared to the normal set.

Table 2 The confusion matrix analysis on test set

		Ground truth				Positive predictive value
		Benign non-inflammatory FLLs	Metastases	HCCs	Hepatic abscesses	
Prediction	Benign non-inflammatory FLLs	25	0	4	2	0.806
	Metastases	3	23	2	3	0.742
	HCCs	3	0	17	0	0.85
	Hepatic abscesses	3	0	0	22	0.88
	Sensitivity	0.735	1	0.739	0.815	
	Specificity	0.918	0.905	0.964	0.963	
	Accuracy	0.86	0.925	0.916	0.925	
	Mean accuracy	0.813				

HCCs: Hepatocellular carcinomas; FLLs: Focal liver lesions.

Table 3 The model's performance comparison between the normal set and "phase cheating" sets

Policy	HCCs (AUC [95%CI]/*P* value)	Metastases (AUC [95%CI]/*P* value)	Benign non-inflammatory (AUC [95%CI]/*P* value)	Hepatic Abscesses (AUC [95%CI]/*P* value)
PP + AP + PVP + DP	0.92 (0.837-0.992)	0.99 (0.967-1.00)	0.88 (0.795-0.955)	0.96 (0.914-0.996)
AP + AP + PVP + DP	0.820 (0.705-0.905)/0.0699	0.901 (0.805- 0.960)/0.0289	0.893 (0.809-0.949)/0.2502	0.924 (0.823-0.977)/0.3387
PP + PVP + PVP + DP	0.704 (0.565-0.821)/0.0017	0.930 (0.832- 0.981)/0.2573	0.799 (0.701- 0.877)/0.0924	0.938 (0.846-0.984)/0.4317
PP + AP + AP + DP	0.768 (0.643-0.866)/0.0013	0.833 (0.714 -0.916)/0.0120	0.864 (0.774-0.929)/0.9720	0.935 (0.846-0.981)/0.4047
PP + AP+ PVP + PVP	0.911 (0.815-0.967)/0.6404	0.959 (0.882- 0.992)/0.4066	0.913 (0.832-0.963)/0.7877	0.831 (00.716- 0.914)/0.0184
PP + AP + AP + AP	0.672 (0.542-0.785)/< 0.0001	0.758 (0.692- 0.909)/0.0079	0.863 (0.773-0.927)/0.3188	0.806 (0.690- 0.893)/0.0475
PP + PVP + PVP + PVP	0.721 (0.584-0.834)/0.0019	0.913 (0.807-0.972)/0.1165	0.775 (0.675-0.857)/ 0.0247	0.900 (0.796- 0.962)/0.7491
PP + DP+ DP+ DP	0.652 (0.513-0.774)/0.0002	0.818 (0.692-0.909)/0.0079	0.790 (0.688-0.870)/0.0356	0.904 (0.802-0.964)/0.7911
AP + AP + AP + AP	0.573 (0.443- 0.696)/< 0.0001	0.674 (0.548-0.785)/< 0.0001	0.833 (0.739- 0.904)/0.3375	0.697 (0.567- 0.807)/0.0019
PVP + PVP + PVP+ PVP	0.697 (0.554-0.817)/0.0029	0.859 (0.748- 0.934)/0.0101	0.794 (0.693- 0.874)/0.1144	0.782 (0.650-0.882)/0.0278
DP + DP + DP + DP	0.697 (0.562-0.811)/0.0007	0.787 (0.666- 0.880)/0.0008	0.751 (0.646-0.838)/0.0387	0.873 (0.760-0.946)/0.1805

AP: Arterial phase; AUC: Area under the receiver operating characteristic curve; CI: Confidence interval; DP: Delayed phase; FLLs: Focal liver lesions; HCC: Hepatocellular carcinoma; PP: Precontrast phase; PVP: Portal venous phase.

DISCUSSION

The correct diagnosis of liver lesions before treatment is of great significance. In our study, a classification system was proposed based on the features derived from the four-phase DCE-CT images. The AUC (95%CI) for differentiating each category from the others was 0.92 (0.837-0.992), 0.99 (0.967-1.00), 0.88 (0.795-0.955), and 0.96 (0.914-0.996) for HCC, metastases, benign non-inflammatory FLLs, and hepatic abscesses, respectively, indicating that the classification system is highly capable of distinguishing one lesion type from the others.
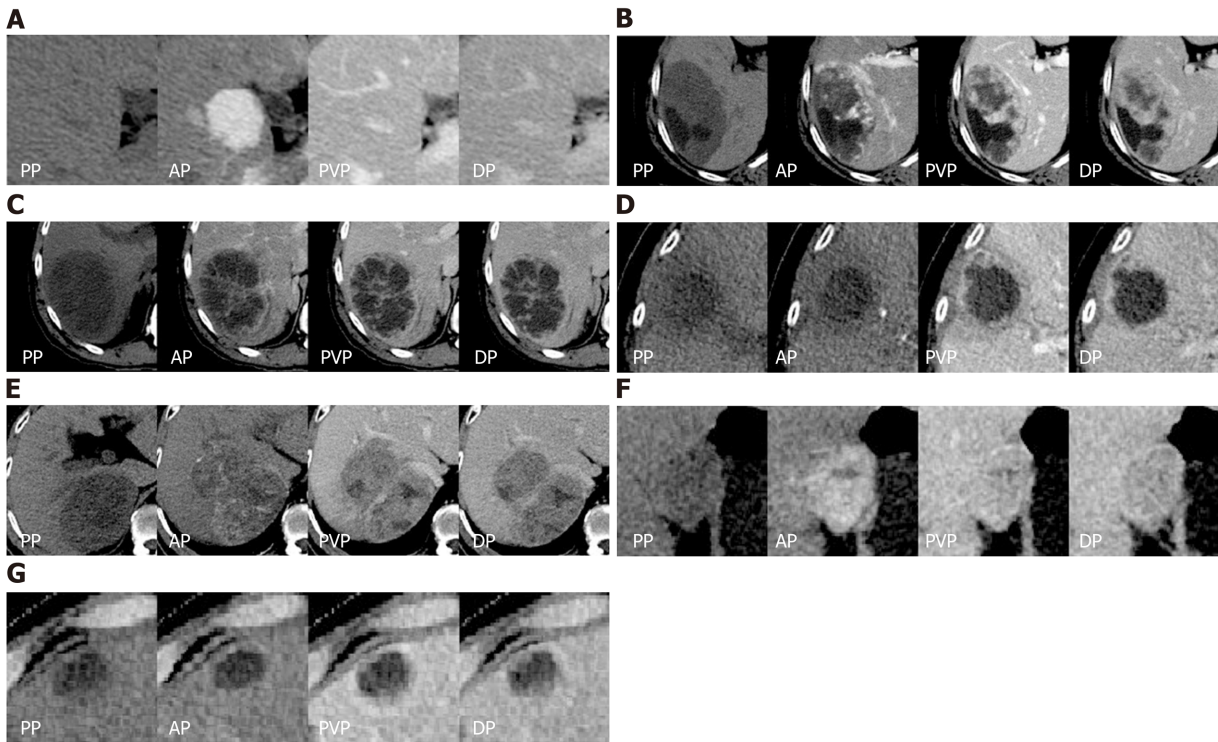
Figure 3 The representative correctly classified and misclassified categories. For each patient, axial four-phase (PP, AP, PVP, DP) computed tomography images were obtained and focal liver lesions were diagnosed by histopathologic evaluation after biopsy or surgery. A: A 33-year-old man with focal nodular hyperplasia was correctly classified as category C; B: A 54-year-old woman with hemangioma was misclassified as category D; C: A 52-year-old man with hepatic abscess was correctly classified as category D; D: An 82-year-old woman with hepatic abscess was misclassified as category B; E: A 55-year-old man with HCC was correctly classified as category A; F: A 38-year-old woman with HCC was misclassified as category C; G: A 75-year-old man with liver metastases derived from colorectal cancer was correctly classified as category B. And there was no misclassification for the metastasis group. AP: Arterial phase; DP: Delayed phase; PP: Precontrast phase; PVP: Portal venous phase.
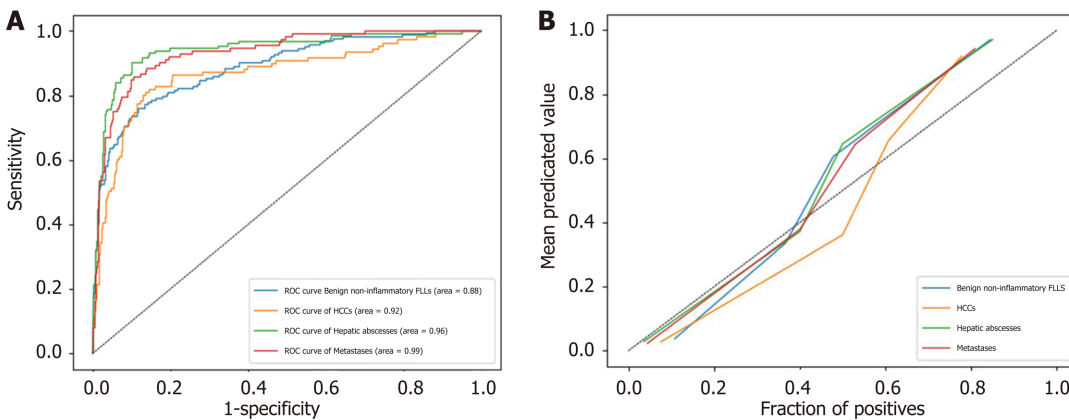

Figure 4 The receiver operating characteristic analysis of model's classification performance on test set and calibration curve of model's classification probability for each category. A: The receiver operating characteristic analysis of model's classification performance on test set; B: Calibration curve of model's classification probability for each category. FLLs: Focal liver lesions; HCC: Hepatocellular carcinoma; ROC: Receiver operating characteristic.

Since the different types of FLLs have different outcomes and require different clinical interventions, the current challenge in determining an accurate diagnosis involves not only effectively differentiating between benign and malignant FLLs according to the medical image but also accurately recognizing the different types of FLLs. A previous study[20] proposed a novel two-stage multiview learning framework for the ultrasound-based computer-aided diagnosis of benign and malignant liver tumors. Although both HCC and metastases are malignant liver tumors, their treatment strategies are completely different; thus, more accurate classification is

Figure 5 Predicted probability heatmaps. The top color bar represents the classification probability of the model from 0 to 1, which corresponds to dark blue to bright yellow. A: Shows the results from normal four-phase input; B: Shows the results from different "phase cheating" sets as indicated in the policy of input data; C: Shows the representative examples. AP: Arterial phase; DP: Delayed phase; PP: Precontrast phase; PVP: Portal venous phase.

needed. Yasaka *et al*[15] investigated the feasibility of applying deep learning models for liver lesion classification using CT images and showed good model performance. However, their standard of classification was based on the radiologic features. HCC is treated differently from metastases, as are abscesses and FNHs. In our study, the category label obtained from the combination of contemporaneous histology and treatment decisions should have more practically applicable value.

Notably, the sensitivity for distinguishing HCC was not high (0.739) in our study, similar to that of previous studies. The range of sensitivities reported in the literature for the detection of HCC on DCE-CT is 50%-75%[21-24]. However, the diagnosis of the lesions may vary depending on the imaging modality. Hamm *et al*[18] developed a CNN model based on MRI images for liver lesion classification, demonstrating high sensitivity. Previous studies[24,25] also reported the superiority of MRI over CT. However, in clinical practice, CT is more accessible and more inexpensive than MRI. Those patients who have a contraindication for MRI due to a comprehensive past history and clinical evaluation are candidates for the CT examination. Our model should be made available to these patients.

The interpretation of how neural networks, particularly deep neural networks, obtain the conclusion is difficult, and these networks are criticized as black boxes[26]. To evaluate whether our model correctly learned useful features from the four-phase CT images, we applied a "phase cheating" experiment on the test set. Compared to the normal set, the performance of the deep-learning network in differentiating HCC from others was dramatically degraded once the placeholder on AP and/or PVP was occluded ($P < 0.05$). This finding probably indicates that the networks make decisions by using accurate distinguishing features, AP hypervascularity and washout in the PVP, which is consistent with the clinical diagnostic criteria for HCC[26]. However, there was no significant difference in the AUCs for differentiating HCC from others between the normal set and the "phase cheating" set when PP was replaced by AP. This result was likely because most lesions are hypodense in the PP[27,28] and the normal hepatic parenchyma shows only minimal enhancement during the AP. The degree of enhancement of lesions in the AP was obtained by comparing the normal hepatic parenchyma around the lesions. In addition, the enhanced scans and the PP have the same value in the diagnosis of calcium, necrosis and gas in the lesion.

One issue for supervised learning is overfitting[29], which normally shows good fit on training data but performs poorly on unseen test data. When the size of training set

is small, this phenomenon becomes more apparent. To avoid overfitting, we applied various regulation techniques in the model during training, such as adding normalization layers to generalize the model, applying L2 regulation to the filters, adding a dropout layer, and augmenting the data to accommodate data variation. The Brier scores for HCCs, metastases, benign non-inflammatory FLLs and hepatic abscesses also suggest that our model is accurate and reasonable.

Our study had several limitations. First, we only evaluated the four-phase CT images and did not consider the clinical information, such as an increased alpha-fetoprotein level and a history of hepatitis B, C infection or liver cirrhosis, which might suggest HCC[29]. Second, we only trained and evaluated the model in a single center setting using a single CT scanner, where there might be a data bias that may lead to model bias. The model should display better generality if more variable data are analyzed. Third, the sample size of the test set was relatively small. Therefore, a larger sample is needed for further studies. Finally, we did not include lesions larger than 10 cm due to the balance among network depth, input matrix size, receptive field size, and memory load. For larger lesions, a higher matrix input size and a deeper network depth are needed, causing a rapid increase in memory requirement, which exceeds the capacity of the current GPUs.

In conclusion, the MP-CDN showed a high differential diagnostic performance for classifying FLLs as HCC, metastases, benign non-inflammatory FLLs and hepatic abscesses in four-phase CT images. If trained on a larger sample or a diverse cohort imaged with a variety of CT scanners, the MP-CDN could become an efficient tool to assist radiologists in accurate identification of the different types of FLLs. However, further evaluation of this model in a multicenter setting is necessary to evaluate its clinical utility.

ARTICLE HIGHLIGHTS

Research background
The accurate classification of focal liver lesions (FLLs) is essential to properly guide treatment options and predict prognosis. Dynamic contrast-enhanced computed tomography (DCE-CT) is commonly used for the noninvasive detection and exact classification of FLLs due to its high scanning speed and high-density resolution. Since their recent development, convolutional neural network (CNN)-based deep learning techniques have been recognized to have high potential for image recognition tasks.

Research motivation
Since the different types of FLLs have different outcomes and require different clinical interventions, the current challenge in determining an accurate diagnosis involves not only effectively differentiating between benign and malignant FLLs according to the medical image but also accurately recognizing the different types of FLLs. Our purpose was to develop and evaluate a deep learning-based CNN to classify FLLs on multiphase CT. Our CNN model is expected to become an efficient tool to assist radiologists in accurately identifying the different types of FLLs.

Research objectives
The appearances, especially the dynamic enhancement patterns of FLLs on CT imaging, are essential for categorizing lesions. We employed a four-channel input data to preserve the dynamic enhancement properties. The combination of the lesion's dynamic enhancement pattern with a CNN can imitate the image diagnosis of radiologists and is expected to improve diagnostic accuracy.

Research methods
A total of 517 FLLs scanned on a 320-detector CT scanner using a four-phase DCE-CT imaging protocol (including precontrast phase, arterial phase, portal venous phase, and delayed phase) from 2012 to 2017 were retrospectively enrolled. FLLs were classified into four categories: Category A, hepatocellular carcinoma (HCC); category B, liver metastases; category C, benign non-inflammatory FLLs including hemangiomas, focal nodular hyperplasias and adenomas; and category D, hepatic abscesses. Each category was split into a training set and test set in an approximately 8:2 ratio. The CNN model with a sequential input of the four-phase CT images was developed to automatically classify FLLs. The classification performance of CNN model was evaluated on the test set: The accuracy, specificity and sensitivity were

calculated from the confusion matrix, and the area under the receiver operating characteristic curve (AUC) was calculated from the SoftMax probability outputted from the last layer of the CNN model.

Research results

A total of 410 FLLs were used for training and 107 FLLs were used for testing. The accuracy/specificity/sensitivity of differentiating each category from others were 0.916/0.964/0.739, 0.925/0.905/1.0, 0.860/0.918/0.735 and 0.925/0.963/0.815 for HCC, metastases, benign non-inflammatory FLLs, and abscesses on the test set, respectively. The AUC (95% confidence interval) for differentiating each category from others was 0.92 (0.837-0.992), 0.99 (0.967-1.00), 0.88 (0.795-0.955) and 0.96 (0.914-0.996) for HCC, metastases, benign non-inflammatory FLLs, and abscesses on the test set, respectively. Also, for this study, we only trained and evaluated the CNN model in a single center setting using a single CT scanner, where there might be a data bias that may lead to model bias. Further evaluation of this model in a multicenter setting is needed to evaluate its clinical utility.

Research conclusions

Overall, our CNN model showed a high differential diagnostic performance for classification FLLs as HCC, metastases, benign non-inflammatory FLLs and hepatic abscesses in four-phase CT image and could become an efficient tool to assist radiologists in accurate identification of the different types of FLLs.

Research perspectives

Further multicenter studies are necessary to evaluate the clinical utility of our CNN model. In addition, it's worth to evaluate the clinical information whether can further improve the perform of CNN model.

REFERENCES

1   **Horta G**, López M, Dotte A, Cordero J, Chesta C, Castro A, Palavecino P, Poniachik J. [Benign focal liver lesions detected by computed tomography: Review of 1,184 examinations]. *Rev Med Chil* 2015; **143**: 197-202 [PMID: 25860362 DOI: 10.4067/S0034-98872015000200007]

2   **Kaltenbach TE**, Engler P, Kratzer W, Oeztuerk S, Seufferlein T, Haenle MM, Graeter T. Prevalence of benign focal liver lesions: ultrasound investigation of 45,319 hospital patients. *Abdom Radiol (NY)* 2016; **41**: 25-32 [PMID: 26830608 DOI: 10.1007/s00261-015-0605-7]

3   **Heimbach JK**, Kulik LM, Finn RS, Sirlin CB, Abecassis MM, Roberts LR, Zhu AX, Murad MH, Marrero JA. AASLD guidelines for the treatment of hepatocellular carcinoma. *Hepatology* 2018; **67**: 358-380 [PMID: 28130846 DOI: 10.1002/hep.29086]

4   **The American College of Radiology**.   CT/MRI LI-RADS® v2018 CORE. Available from: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/LI-RADS/CT-MRI-LI-RADS-v2018

5   **Ambinder EP**. A history of the shift toward full computerization of medicine. *J Oncol Pract* 2005; **1**: 54-56 [PMID: 20871680 DOI: 10.1200/JOP.2005.1.2.54]

6   **Gletsos M**, Mougiakakou SG, Matsopoulos GK, Nikita KS, Nikita AS, Kelekis D. A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier. *IEEE Trans Inf Technol Biomed* 2003; **7**: 153-162 [PMID: 14518728 DOI: 10.1109/titb.2003.813793]

7   **Huang YL**, Chen JH, Shen WC. Diagnosis of hepatic tumors with texture analysis in nonenhanced computed tomography images. *Acad Radiol* 2006; **13**: 713-720 [PMID: 16679273 DOI: 10.1016/j.acra.2005.07.014]

8   **Mougiakakou SG**, Valavanis IK, Nikita A, Nikita KS. Differential diagnosis of CT focal liver lesions using texture features, feature selection and ensemble driven classifiers. *Artif Intell Med* 2007; **41**: 25-37 [PMID: 17624744 DOI: 10.1016/j.artmed.2007.05.002]

9   **Lakhani P**, Gray DL, Pett CR, Nagy P, Shih G. Hello World Deep Learning in Medical Imaging. *J Digit Imaging* 2018; **31**: 283-289 [PMID: 29725961 DOI: 10.1007/s10278-018-0079-6]

10   **Biswas M**, Kuppili V, Saba L, Edla DR, Suri HS, Cuadrado-Godia E, Laird JR, Marinhoe RT, Sanches JM, Nicolaides A, Suri JS. State-of-the-art review on deep learning in medical imaging. *Front Biosci (Landmark Ed)* 2019; **24**: 392-426 [PMID: 30468663]

11   **Litjens G**, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. *Med Image Anal* 2017; **42**: 60-88 [PMID: 28778026 DOI: 10.1016/j.media.2017.07.005]

12   **Shen W**, Zhou M, Yang F, Yang C, Tian J. Multi-scale Convolutional Neural Networks for Lung Nodule Classification. *Inf Process Med Imaging* 2015; **24**: 588-599 [PMID: 26221705 DOI: 10.1007/978-3-319-19992-4_46]

13   **Yasaka K**, Akai H, Kunimatsu A, Abe O, Kiryu S. Liver Fibrosis: Deep Convolutional Neural Network for Staging by Using Gadoxetic Acid-enhanced Hepatobiliary Phase MR Images. *Radiology* 2018; **287**: 146-155 [PMID: 29239710 DOI: 10.1148/radiol.2017171928]

14   **Kermany DS**, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X,

Yan F, Dong J, Prasadha MK, Pei J, Ting MYL, Zhu J, Li C, Hewett S, Dong J, Ziyar I, Shi A, Zhang R, Zheng L, Hou R, Shi W, Fu X, Duan Y, Huu VAN, Wen C, Zhang ED, Zhang CL, Li O, Wang X, Singer MA, Sun X, Xu J, Tafreshi A, Lewis MA, Xia H, Zhang K. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* 2018; **172**: 1122-1131.e9 [PMID: 29474911 DOI: 10.1016/j.cell.2018.02.010]

15   **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* 2018; **286**: 887-896 [PMID: 29059036 DOI: 10.1148/radiol.2017170706]

16   **Lakhani P**, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* 2017; **284**: 574-582 [PMID: 28436741 DOI: 10.1148/radiol.2017162326]

17   **Albarqouni S**, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N. AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE Trans Med Imaging* 2016; **35**: 1313-1321 [PMID: 26891484 DOI: 10.1109/TMI.2016.2528120]

18   **Hamm CA**, Wang CJ, Savic LJ, Ferrante M, Schobert I, Schlachter T, Lin M, Duncan JS, Weinreb JC, Chapiro J, Letzen B. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. *Eur Radiol* 2019; **29**: 3338-3347 [PMID: 31016442 DOI: 10.1007/s00330-019-06205-9]

19   **Kutlu H**, Avcı E. A Novel Method for Classifying Liver and Brain Tumors Using Convolutional Neural Networks, Discrete Wavelet Transform and Long Short-Term Memory Networks. *Sensors (Basel)* 2019; 19 [PMID: 31035406 DOI: 10.3390/s19091992]

20   **Guo LH**, Wang D, Qian YY, Zheng X, Zhao CK, Li XL, Bo XW, Yue WW, Zhang Q, Shi J, Xu HX. A two-stage multi-view learning framework based computer-aided diagnosis of liver tumors with contrast enhanced ultrasound images. *Clin Hemorheol Microcirc* 2018; **69**: 343-354 [PMID: 29630528 DOI: 10.3233/CH-170275]

21   **Addley HC**, Griffin N, Shaw AS, Mannelli L, Parker RA, Aitken S, Wood H, Davies S, Alexander GJ, Lomas DJ. Accuracy of hepatocellular carcinoma detection on multidetector CT in a transplant liver population with explant liver correlation. *Clin Radiol* 2011; **66**: 349-356 [PMID: 21295772 DOI: 10.1016/j.crad.2010.11.012]

22   **Libbrecht L**, Bielen D, Verslype C, Vanbeckevoort D, Pirenne J, Nevens F, Desmet V, Roskams T. Focal lesions in cirrhotic explant livers: pathological evaluation and accuracy of pretransplantation imaging examinations. *Liver Transpl* 2002; **8**: 749-761 [PMID: 12200773 DOI: 10.1053/jlts.2002.34922]

23   **Ladd LM**, Tirkes T, Tann M, Agarwal DM, Johnson MS, Tahir B, Sandrasegaran K. Comparison of hepatic MDCT, MRI, and DSA to explant pathology for the detection and treatment planning of hepatocellular carcinoma. *Clin Mol Hepatol* 2016; **22**: 450-457 [PMID: 27987537 DOI: 10.3350/cmh.2016.0036]

24   **Burrel M**, Llovet JM, Ayuso C, Iglesias C, Sala M, Miquel R, Caralt T, Ayuso JR, Solé M, Sanchez M, Brú C, Bruix J; Barcelona Clínic Liver Cancer Group. MRI angiography is superior to helical CT for detection of HCC prior to liver transplantation: an explant correlation. *Hepatology* 2003; **38**: 1034-1042 [PMID: 14512891 DOI: 10.1053/jhep.2003.50409]

25   **Kim BR**, Lee JM, Lee DH, Yoon JH, Hur BY, Suh KS, Yi NJ, Lee KB, Han JK. Diagnostic Performance of Gadoxetic Acid-enhanced Liver MR Imaging versus Multidetector CT in the Detection of Dysplastic Nodules and Early Hepatocellular Carcinoma. *Radiology* 2017; **285**: 134-146 [PMID: 28609205 DOI: 10.1148/radiol.2017162080]

26   **Zhang Z**, Beck MW, Winkler DA, Huang B, Sibanda W, Goyal H; written on behalf of AME Big-Data Clinical Trial Collaborative Group. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann Transl Med* 2018; **6**: 216 [PMID: 30023379 DOI: 10.21037/atm.2018.05.32]

27   **Ma Y**, Zhang XL, Li XY, Zhang L, Su HH, Zhan CY. [Value of computed tomography and magnetic resonance imaging in diagnosis and differential diagnosis of small hepatocellular carcinoma]. *Nan Fang Yi Ke Da Xue Xue Bao* 2008; **28**: 2235-2238 [PMID: 19114368]

28   **Li CS**, Chen RC, Tu HY, Shih LS, Zhang TA, Lii JM, Chen WT, Duh SJ, Chiang LC. Imaging well-differentiated hepatocellular carcinoma with dynamic triple-phase helical computed tomography. *Br J Radiol* 2006; **79**: 659-665 [PMID: 16641423 DOI: 10.1259/bjr/12699987]

29   **Cook JA**, Ranstam J. Overfitting. *Br J Surg* 2016; **103**: 1814 [PMID: 27901280 DOI: 10.1002/bjs.10244]

30   **Bruix J**, Sherman M; Practice Guidelines Committee, American Association for the Study of Liver Diseases. Management of hepatocellular carcinoma. *Hepatology* 2005; **42**: 1208-1236 [PMID: 16250051 DOI: 10.1002/hep.20933]

31   **Yeh MM**, Daniel HD, Torbenson M. Hepatitis C-associated hepatocellular carcinomas in non-cirrhotic livers. *Mod Pathol* 2010; **23**: 276-283 [PMID: 19935643 DOI: 10.1038/modpathol.2009.174]