Diagnostic and Interventional Imaging (2019) **100**, 227—233

ORIGINAL ARTICLE / *Computer developments*

Diagnosis of focal liver lesions from ultrasound using deep learning

B. Schmauch[a], **P. Herent**[a], **P. Jehanno**[a,*], **O. Dehaene**[b], **C. Saillard**[a], **C. Aubé**[c], **A. Luciani**[d], **N. Lassau**[e,f], **S. Jégou**[a]

[a] *Owkin Inc, Research and Development Laboratory, 75, rue de Turbigo, 75003 Paris, France*
[b] *École Centrale d'Electronique (ECE), 75015 Paris, France*
[c] *Radiology Department, CHU Angers, 49933 Angers, France*
[d] *Radiology Department, AP—HP, Hôpitaux Universitaires Henri-Mondor, 94010 Creteil, France*
[e] *Radiology Department, Institut Gustave Roussy, 94805 Villejuif, France*
[f] *IR4M, UMR8081CNRS, Université Paris-Sud, Université Paris-Saclay, 94805 Villejuif, France*

Received 21 January 2019; accepted 22 February 2019

KEYWORDS
Artificial intelligence;
Deep learning;
Focal liver lesions;
Ultrasound;
Radiology

Abstract
Purpose: The purpose of this study was to create an algorithm that simultaneously detects and characterizes (benign vs. malignant) focal liver lesion (FLL) using deep learning.
Materials and methods: We trained our algorithm on a dataset proposed during a data challenge organized at the *2018 Journées Francophones de Radiologie*. The dataset was composed of 367 two-dimensional ultrasound images from 367 individual livers, captured at various institutions. The algorithm was guided using an attention mechanism with annotations made by a radiologist. The algorithm was then tested on a new data set from 177 patients.
Results: The models reached mean ROC-AUC scores of 0.935 for FLL detection and 0.916 for FLL characterization over three shuffled three-fold cross-validations performed with the training data. On the new dataset of 177 patients, our models reached a weighted mean ROC-AUC scores of 0.891 for seven different tasks.
Conclusion: This study that uses a supervised-attention mechanism focused on FLL detection and characterization from liver ultrasound images. This method could prove to be highly relevant for medical imaging once validated on a larger independent cohort.

* Corresponding author.
E-mail address: paul.jehanno@owkin.com (P. Jehanno).

https://doi.org/10.1016/j.diii.2019.02.009

The number of imaging examinations in modern medicine is steadily increasing, along with the complexity of interpretation and demands on providers for access to healthcare data, collaborative decision making, and quality accountability measures [1]. Radiologists are exposed to decision fatigue, due to known susceptibility factors including prolonged shifts, sleep deprivation, and performance of high-volume and high-complexity tasks [2,3]. Decision fatigue can lead to medical errors, with missed, incorrect, or delayed diagnoses estimated up to 10—15% in radiology [4]. The two most frequent causes of errors are missed findings (42% of recounted errors) and satisfaction of search [4]. Moreover, interpretations by radiologists are prone to high intra- and inter-individual variability [5,6].

Artificial intelligence shows promise in many applications in radiology. Deep learning is a subtype of machine learning, called ''deep'' because of digital architecture that uses a large number of layers of artificial neurons, called neural networks [7]. In computer vision, most deep learning papers use convolutional neural networks (CNN), following its success in data challenges in outperforming standard computer vision algorithms [8]. The application of deep learning in radiology is potentially vast and could revolutionize each step of the medical-imaging pipeline [9—11].

Few studies have focused on liver ultrasound and deep learning. Park et al. focused on the classification of fibrosis [12], Choi et al. reported work on the classification of LI-RADS based on radiological reports of ultrasound monitoring for hepatocellular carcinoma [13], and Liu et al. showed promising results on the diagnosis of cirrhosis based on hepatic capsule morphology [14]. Human performance for characterizing focal liver lesion (FLL) is limited with an area under the ROC curve (AUC) between 0.72 and 0.74 [12].

The purpose of this study was to create an algorithm that simultaneously detects and characterizes (benign vs. malignant) FLL using deep learning.

Materials and methods

Preprocessing

The dataset was provided during a public challenge during the *2018 Journées Francophones de Radiologie* in Paris, France. Despite the standardization already performed by the challenge organizers, the data were highly heterogeneous in terms of size and luminosity. First, the images were cropped to maximally remove the black borders and standardize the aspect ratio. Furthermore, we performed a normalization based on the observation that the upper part of the considered images consisted mostly of abdominal tissues, which share a common pattern. Thus, the intensity peak of this region should vary little between images. However, the image acquisition conditions varied, with the intensity peak values ranging from 30 to 110 units. Consequently, we rescaled every image to shift this peak to a common value of 70 units (corresponding to the average value of the dataset).

We selected the top 20% of pixels from the image, excluding those that were black, and estimated the peak intensity by taking the median value (m) of this selection. Then, we applied Eq. (1) to every pixel x_{ij} (Fig. 1).

$$x_{ij} \leftarrow x_{ij} \times \frac{70}{m} \tag{1}$$

Finally, we resized every image to 240×345 or 480×690. This choice had little impact on the performance of the models. In the description of the model below, we considered the first case.

Feature extraction

To extract features from images, we used a 50-layer residual neural network (ResNet50), pretrained on the ImageNet dataset, from which we removed the last two layers [15]. This network was designed for color images. Thus, we had to copy each grayscale image three times in order to simulate red, green and blue channels. For an input image of size $3 \times 240 \times 345$, the network produced a feature map with a dimension of $2048 \times 8 \times 11$. A first simple approach consisted of averaging this representation over the spatial dimensions using Eq. (2).

$$x_k = \frac{1}{8 \times 11} \sum_{i,j} x_{kij} \tag{2}$$

This gave a feature vector of 2048 for each image, which was fed to a unique densely connected layer with seven neurons, one for each classification task (detection and characterization). The main drawback of this approach was that it gave the same importance to each pixel of the image, including regions of little interest such as the tissues surrounding liver.

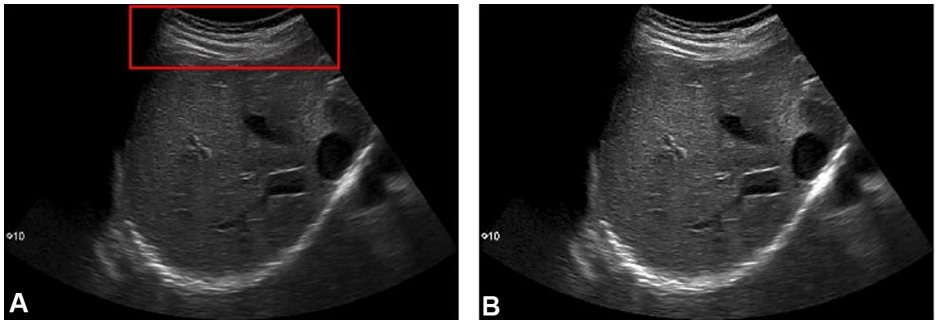

Figure 1. Ultrasound images of the liver from the training set before (A) and after (B) intensity rescaling based on the observation that all abdominal tissue should share a common pixel intensity profile in all liver ultrasound images. The region used for computing the scaling factor is outlined in red on A.

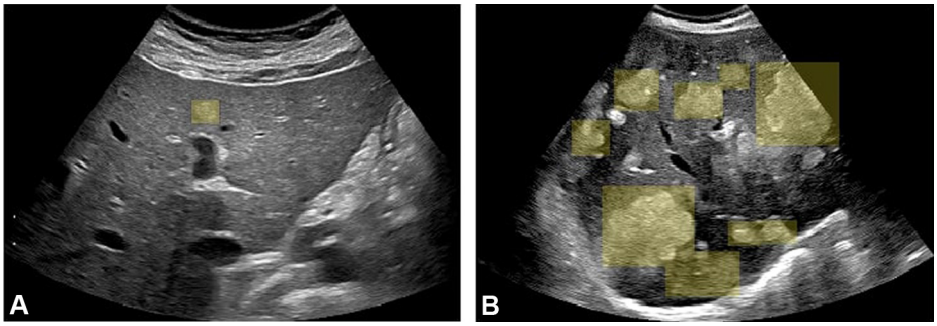

Figure 2. Ultrasound images of focal liver lesions from the training set. Bounding boxes annotations are superimposed on the original image for an angioma (a) and multiple metastasis (b). Those annotations were generated a radiologist using dedicated tool.

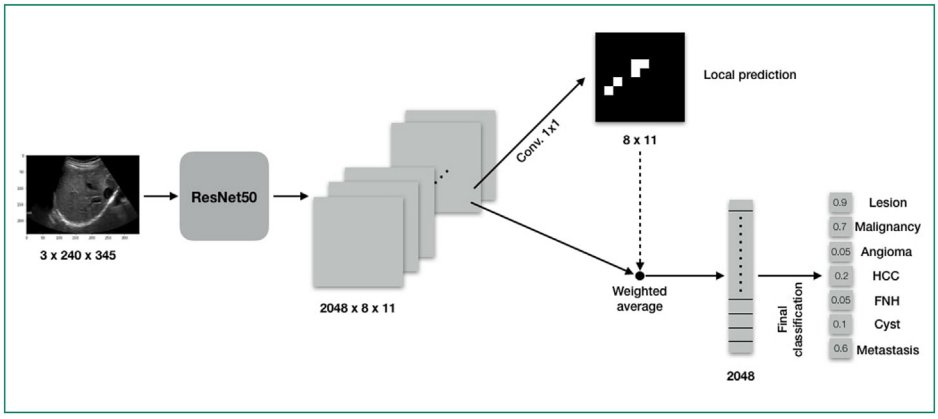

Figure 3. Architecture of the neural network. Each image of size 240×345 is fed into the ResNet50 Neural Network which produces 2048 images of size 8×11. They are then fed to the upper branch also called the ''attention block'' of the algorithm that learns to detect anomalies in the image. The latter are also fed to a second branch that averages features maps over the selected areas. Finally, the 2048 features are fitted to a logistic regression that outputs a score ranging from 0 to 1 for each category of breast lesion. This score can be interpreted as the probability of presence of such lesion in the image.

Supervised attention mechanism

One of the main difficulties of this task is that liver lesions vary widely in appearances and size. We facilitated learning by decomposing classification into two steps:

- detection of abnormalities on the ultrasound image;
- classification of these lesions.

These two steps were simultaneously performed by two branches of the same model. For the first, we created and used additional labels for localization. These labels consisted of bounding boxes surrounding the lesions. These annotations did not require precise characterization. Thus, they were rapidly performed by a 5th-year resident in radiology (P.H.) (Fig. 2).

For each image, we generated a binary mask of the same size, indicating the presence or absence of lesions. We reduced the size of this mask to match the output dimensions of the ResNet (i.e., 8×11 pixels for an input image with spatial dimensions of 240×345). The localization module was a single 1×1 convolution, applied to the output of the ResNet. This transformed the $2048 \times 8 \times 11$ representation into a single image with the dimensions 8×11, to which we applied a sigmoid function to generate a prediction between 0 and 1. This module was trained to reproduce the binary mask generated from the annotations.

Then, this local prediction was used to guide the main module, responsible for determining the presence of FLL in the image level and their characterization. More precisely, we used the local prediction to compute a weighted average of the final feature map, shown below in Eq. (3), in which $p_{i,j}$ is the local prediction for pixel (i, j). If the localization module predicted a uniform probability of a FLL over the entire image, the formula was equivalent to the spatial average of the simple model, whereas, when the module predicted the presence of a lesion in a single pixel with high confidence, only the feature vector extracted from this pixel was used for the final prediction.

$$x_k = \frac{\sum_{i,j} p_{i,j} x_{kij}}{\sum_{i,j} p_{i,j}} \tag{3}$$

The final prediction was performed by a densely connected layer with seven neurons, one for each prediction: FLL detection, malignancy, angioma, hepatocellular carcinoma, focal nodular hyperplasia, cyst, or metastasis. The architecture of the model is shown in Fig. 3. Furthermore, this attention mechanism allowed interpretation of the model's predictions. To do so, we resized the 8×11 attention map $\frac{p_{ij}}{\sum_{i,j} p_{i,j}}$, and to the original image dimensions (i.e., 240×345). We superposed this map over the image to see the areas considered by the model to make its decision

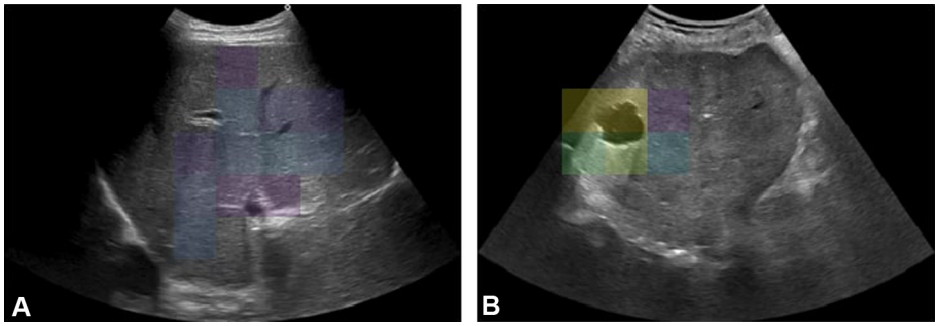

Figure 4. Figures show two examples of attention maps generated by the model for a homogeneous liver without lesions (a) and one with a biliary cyst (b). In (a), the model indiscriminately covers a large part of the liver, whereas, in (b), the model focuses its attention on a restricted area containing the lesion.

(Fig. 4). In particular, this allowed us to identify reasons why the model may have failed.

Implementation

Given the relatively small number of images available for training, we performed data augmentation to avoid overfitting. We applied random transformations, such as cropping, modifications of the aspect ratio, small translations and rotations. Counter-intuitively, applying random horizontal flips to images during training slightly improved performance, even though images generated in this manner are not realistic, as the liver is not a symmetric organ. This may be due to the fact that our model focuses on parts of the image, and the mirror image of a lesion is still a plausible lesion. We also applied mixup with a coefficient of 0.2 [16].

Our models were simultaneously trained on the three tasks evaluated in this challenge (lesion detection, malignancy and lesion classification). This multitask setting contributed to limiting overfitting. However, these tasks were not necessarily learned at the same pace. Thus, we saved three copies of the weights, chosen depending on the performance of the model on a validation set during cross-validation. For example, when the model reached its best AUC for lesion detection, we saved the first copy of its weights which was used only for this task. We used stochastic gradient descent with Nesterov momentum to train the models. The results were highly variable due to the small amount of data. We thus performed three-fold cross-validation, which we repeated for three different splits of the data, to account for this variability. We repeated nine experiences during which we selected randomly 245 images (out of the 367 images of the training set) to train our neural network, and estimated its performance by computing an AUC over the 122 images left. We then computed the mean scores over those nine different experiences to evaluate our model before executing it on the test set provided by the organizers of the challenge.

Results

The details of the dataset are provided in Table 1. We performed experiments with several variations of the model, making the following changes:

Table 1 Number of samples provided by the organizers of the data challenge for each category.

Lesion Type	Training set	Test set*
Homogeneous Liver	258 (70%)	Unknown
Angioma	17 (4.6%)	Unknown
Metastasis	48 (13%)	Unknown
HCC	6 (1.6%)	Unknown
Cyst	30 (8.2%)	Unknown
HNF	8 (2.2%)	Unknown
Total	367 (100%)	177

The test set was used by the data challenge organizers to evaluate the algorithms. The data challenge organizers did not indicate the lesion types present in the test set of liver ultrasound images. HCC indicates hepatocellular carcinoma. FNH indicates focal nodular hyperplasia

- replacement of the ResNet used for feature extraction with a 121-layer DenseNet [17];
- modification of the attention focus: this was achieved by applying a factor λ to the local prediction before applying the sigmoid function: the smaller this factor, the more diffuse the attention, whereas larger values focused the attention more on a single pixel. In particular, making $\lambda = 0$ amounts to suppressing attention, whereas making $\lambda = 1$ provides no modification to the original formula. It is also possible to simultaneously use different scales;
- use of higher resolution images: 480×690 instead of 240×345.

These models demonstrated similar performance, although calculating the average of their predictions improved ROC-AUC scores during cross-validation. The average ROC-AUC scores achieved by our three best architectures for the repeated cross-validation (*i.e.*, three times three-fold) as well as the result of the ensemble (obtained by averaging predictions of the three models) are shown in Table 2. The AUC for FLL detection was computed over the entire dataset, whereas the AUCs for diagnosing malignancy and lesion characterization were evaluated on the subset of images in which a lesion was actually present (109/367 [29.7%] images in the training data). The same ensemble achieved a weighted AUC of 0.891 on the challenge test set, using the metric defined in Eq. (1). Detailed scores,

Table 2. Results of the different models for the three classification tasks.

Dimensions	Extractor	λ	AUC lesion	AUC mal.	AUC type
240 × 345	ResNet 50	1	0.913 (0.027)	0.930 (0.040)	0.895 (0.060)
480 × 690	ResNet 50	1	0.906 (0.029)	0.910 (0.044)	0.910 (0.060)
240 × 345	DenseNet 121	0.25, 1	0.912 (0.021)	0.917 (0.044)	0.900 (0.062)
—	Ensemble	—	**0.935** ((0.022)	**0.942** ((0.044)	**0.916** ((0.058)

Scores were computed as the mean of three shuffled three-fold cross-validation on the training set. Nine experiments in which 245/367 images (66.8%) of the training set were randomly used to train our algorithm and the scores were computed from the results obtained on the remaining 122/367 images (33.2%). AUC indicates area under the receiver operator characteristic curve. Data are presented as mean. Numbersin parentheses are standard deviation. In bold we highlighted the best scores for each task (detection of lesion/malignancy and lesion type). We see that the best score is achieved by the Ensemble model.

Table 3 Area under the ROC curve (AUC) scores by lesion type.

Class	AUC
Angioma	0.898 (0.067)
Metastasis	0.886 (0.052)
HCC	0.931 (0.072)
Cyst	0.954 (0.017)
HNF	0.909 (0.084)
Average	0.916 (0.058)

AUC indicates area under the receiver operator characteristic curve. HCC indicates hepatocellular carcinoma. FNH indicates focal nodular hyperplasia. Data are presented as means. Numbers in parentheses are standard deviations. Scores were computed as the mean of three shuffled three-fold cross-validations on the training set. Nine experiments in which 245/367 images (66.8%) of the training set were randomly used to train our algorithm and the scores were computed from the results obtained on the remaining 122/367 images (33.2%).

calculated using Eq. (4) below, are shown in Table 3 and corresponding ROC curves are provided in Fig. 5.

$$Score = 0.5 \times AUC_{lesion} + 0.3 \times AUC_{\frac{benign}{malignant}} + 0.2 \times \sum_{lesion\ types} AUC_{lesion\ type} \quad (4)$$

The attention mechanism of our model allowed interpretation of the model's predictions using heat maps (Fig. 4), which helped us identifying false-positive findings. For example, on a subset of ultrasound images, the algorithm misclassified blood vessels as cysts.

Discussion

This study is the first to assess the performance of automatic detection of FLL from ultrasound images, and yielded promising results given the relatively small amount of data. The training dataset consisted of only 367 images, with some lesion types poorly represented such as HCC. Despite these limitations, cross-validation yielded good performance. It is likely that the use of larger databases would further increase the accuracy of the model.

The supervised attention model was beneficial for two reasons: first, providing better labels in the image, such as where the model focused its attention for prediction, improved interpretability of the results. Indeed, the heatmaps made it possible to better understand how the model performed and why misclassifications occurred. For example, a common mistake was confusion between a blood vessel (portal vein or hepatic vein) and a cyst, which suggests areas for improvement in this area by teaching the model to detect normal physiological structures to avoid confusion with potential lesions. Second, annotations by the radiologist increased the performance of the models with a 0.05 increase in AUC compared to the simple approach.

Annotating datasets is a time-consuming task using classical research tools not amenable to a radiological workflow. One of the challenges in deep-learning is to build powerful tools for radiologists, data scientists and patients. Abajian et al. have already identified this key point [18]. In their study, they underlined that annotations made by radiologists such as measurements are present in the DCOM files but are currently insufficiently used [18].

Here, we developed a tool that enabled rapid labeling for better performance, creating bounding box instead of well-defined segmentation, which permitted annotation of the entire dataset in approximately half an hour. The use of heatmaps was beneficial in assessing clinical relevance of the algorithm prediction. In other studies, this method has helped to provide more confident predictions for detecting pneumonia on chest X-rays by revealing the source of infection in images, and predicting bone age, showing growth-plate cartilage to be relevant for prediction [19,20]. More generally, heatmaps are among the methods used in the research field of algorithm interpretability and are particularly relevant for computer vision. In other words, this method allows one to ''see what the algorithm sees'' and could be implemented in the radiological workflow [21].

A limitation of the dataset used in this study was that the labels ''hepatocellular carcinoma'' or ''metastasis'' do not make much sense in clinical practice, as it is difficult to characterize such FLL with ultrasound only [22]. In general, liver ultrasound is a first line examination that needs further examinations for further characterization. However, characterization of some benign FLL, such as cavernous

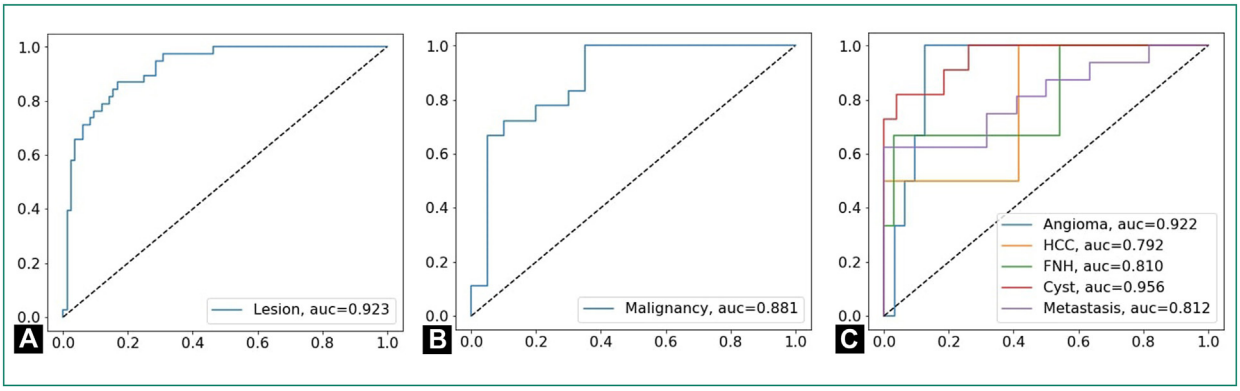

Figure 5. Diagrams show receiver operating characteristic (ROC) curves for focal liver lesion (FLL) detection (a), diagnosis of malignant FLL (b), and correct classification into one out five categories (c) obtained with one model on 122 images.

hemangioma or biliary cyst is possible with ultrasound when these FLL display typical features.

In conclusion, the validation of our algorithm on an independent dataset is an essential step in demonstrating the generalizability of the model, but the promising results appear to corroborate the relevance of applying deep-learning models to ultrasound images.

Human and animal rights

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans as well as in accordance with the EU Directive 2010/63/EU for animal experiments.

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s).

The authors declare that they obtained a written informed consent from the patients and/or volunteers included in the article. The authors also confirm that the personal details of the patients and/or volunteers have been removed.

Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Disclosure of interest

Paul Herent is a part-time consultant at Owkin. Benoit Schmauch, Simon Jégou, Paul Jehanno and Charlie Saillard are employees at Owkin. Olivier Dehaene was a data scientist intern at Owkin at the time the Data Challenge of JFR 2018 occurred and is now a full-time employee at Owkin.

C. Aubé, A. Luciani, N. Lassau declare that they have no competing interest

References

[1] Reiner BI, Krupinski E. The insidious problem of fatigue in medical imaging practice. J Digit Imaging 2012;25:3—6.
[2] Gaba DM, Howard SK. Fatigue among clinicians and the safety of patients. N Engl J Med 2002;347:1249—55.
[3] MacDonald W. The impact of job demands and workload on stress and fatigue. Aust Psychol 2003;38:102—17.
[4] Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015;35:1668—76.
[5] Muenzel D, Engels HP, Bruegel M, Kehl V, Rummeny EJ, Metz S. Intra- and inter-observer variability in measurement of target lesions: implication on response evaluation according to RECIST 1.1. Radiol Oncol 2012;46:8—18.
[6] Suzuki C, Torkzad MR, Jacobsson H, Astrom G, Sundin A, Hatschek T, et al. Interobserver and intraobserver variability in the response evaluation of cancer therapy according to RECIST and WHO-criteria. Acta Oncol 2010;49:509—14.
[7] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436—44.
[8] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual Recognition Challenge. Int J Comput Sci 2015;115:211—52.
[9] Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. Nature 2018;555:487—92.
[10] Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. Jpn J Radiol 2018;36:257—72.
[11] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500—10.
[12] Park H, Park JY, Kim DY, Ahn SH, Chon CY, Han KH, et al. Characterization of focal liver masses using acoustic radiation force impulse elastography. World J Gastroenterol 2013;19:219—26.
[13] Choi H, Banerjee I, Sagreiya H, Kamaya A, Rubin D, Desser T.Machine learning for rapid assessment of outcomes of an ultrasound screening and surveillance program in patients at risk for hepatocellular carcinoma. 2018.
[14] Liu X, Song JL, Wang SH, Zhao JW, Chen YQ. Learning to diagnose cirrhosis with liver capsule guided ultrasound image classification. Sensors 2017;17:149.
[15] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit abs/1512.03385, 2016. http://arxiv.org/abs/1512.03385.

[16] Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. Mixup: beyond empirical risk minimization. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit abs/1710.09412, 2017. http://arxiv.org/abs/1710.09412.

[17] Huang G, Liu Z, van der Maaten L, Weinberger K. Densely Connected Convolutional Networks. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit abs/1608.06993, 2016. http://arxiv.org/abs/1608.06993.

[18] Abajian A, Levy M, Rubin D. Informatics in radiology improving clinical work flow through an AIM Database: a sample web-based lesion tracking application. Radiographics 2012;32:1543—52.

[19] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. ArXiv Preprint, 2017, ArXiv :1711.05225.

[20] Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018;287:313—22.

[21] Holzinger A, Biemann C, Pattichis C, Kell BD. What do we need to build explainable AI systems for the medical domain?; 2017 [ArXiv Preprint ArXiv :1712.09923].

[22] Aubé C, Bazeries P, Lebigot J, Cartier V, Boursier J. Liver fibrosis, cirrhosis, and cirrhosis-related nodules: imaging diagnosis and surveillance. Diagn Interv Imaging 2017;98:455—68.