Multi-parametric MRIs based assessment of Hepatocellular Carcinoma Differentiation with Multi-scale ResNet

Xibin Jia[1], Yujie Xiao[1], Dawei Yang[2], Zhenghan Yang[2,*] and Chen Lu[1]
[1] Faculty of Information Technology, Beijing University of Technology
Beijing, China
[e-mail: jiaxibin@bjut.edu.cn, xyj@bjut.emails.edu.cn]
[2] Department of Radiology, Beijing Friendship Hospital, Capital Medical University
Beijing, China
[e-mail: zhenghanyang@263.net, Dawei-yang@vip.163.com]
*Corresponding author: Zhenghan Yang

Received February 28, 2019; revised March 28, 2019; accepted April 15, 2019; published October 31, 2019

Abstract

To explore an effective non-invasion medical imaging diagnostics approach for hepatocellular carcinoma (HCC), we propose a method based on adopting the multiple technologies with the multi-parametric data fusion, transfer learning, and multi-scale deep feature extraction. Firstly, to make full use of complementary and enhancing the contribution of different modalities viz. multi-parametric MRI images in the lesion diagnosis, we propose a data-level fusion strategy. Secondly, based on the fusion data as the input, the multi-scale residual neural network with SPP (Spatial Pyramid Pooling) is utilized for the discriminative feature representation learning. Thirdly, to mitigate the impact of the lack of training samples, we do the pre-training of the proposed multi-scale residual neural network model on the natural image dataset and the fine-tuning with the chosen multi-parametric MRI images as complementary data. The comparative experiment results on the dataset from the clinical cases show that our proposed approach by employing the multiple strategies achieves the highest accuracy of 0.847±0.023 in the classification problem on the HCC differentiation. In the problem of discriminating the HCC lesion from the non-tumor area, we achieve a good performance with accuracy, sensitivity, specificity and AUC (area under the ROC curve) being 0.981±0.002, 0.981±0.002, 0.991±0.007 and 0.999±0.0008, respectively.

Keywords: Multi-parametric MRI, data fusion, transfer learning, deep learning, hepatocellular carcinoma differentiation

The authors sincerely thank radiologists and pathologists in Beijing Friendship Hospital of Capital Medical University for their efforts to provide the clinical imaging data and carefully verify the ground truth of the datasets. This work is supported by Beijing Natural Science Foundation (No. 7184199), National Natural Science Foundation of China (No. 61871276), Capital's Funds for Health Improvement and Research (No. 2018-2-2023), Research Foundation of Beijing Friendship Hospital, Capital Medical University (No. yyqdkt2017-25) and WBE Liver Fibrosis Foundation (No. CFHPC2019006).

1. Introduction

Hepatocellular cancer (HCC) is an epithelial tumor that originates in the liver. It is the most common primary malignant tumor of the liver and the third most deadly cancer in the world [1], [2]. The differentiation of HCC is one of the most important factors among the multiple factors predicting the recurrence of HCC [3]. Liver biopsy is a gold standard to obtain the differentiation of HCC before surgery, but it cannot be widely used in clinical practice, due to its limitations of invasiveness, sampling error and bleeding. Therefore, there is an urgent need for a non-invasive method to accurately assess the differentiation of HCC.

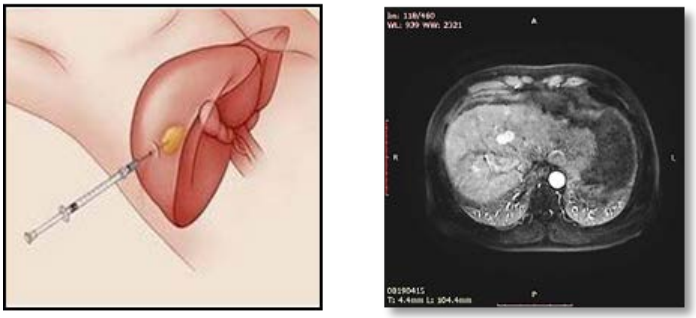


Fig. 1. Liver biopsy (left) and radiographic examination (right)

Although radiology departments of some hospitals have completed millions of radiology imaging examinations, most of them do not have labeled information. The professionalism of medical imaging leads to its particularity. The labeling of medical image data can only rely on professional and experienced doctors in the medical field, so constructing a labeled medical image dataset requires a lot of labor and material resources. Therefore, the size of such medical image dataset is typically small compared to those in the field of computer vision.

On the other hand, the complexity of medical images and the diversity of image capturing methods increase the difficulty of computer-aided diagnosis. First of all, most medical images have three-dimensional, in other words, we can obtain 3D volumes of data through only a radiology imaging examination. In addition, different imaging modes and parameters can be used to obtain different modalities. Each modality of data can reflect different views of human tissue or organ, such as the multi-parametric MRI image used in this paper, including multiple modes, such as DCE-MRI, T2WI, T1 parametric WI and reverse phase, and all of them can provide important information for the doctor to diagnose diseases.

As a consensus effective method for the natural image understanding, deep learning has attracted more researches in the intelligent medical image diagnosis in recent years. It has promising potential in the discriminate representation learning to reveal the essential features of the complex objects. However, deep learning based medical image classification for auto-diagnosis has two main challenges, 1) the lack of well-labeled training data due to the limited amount of the clinical cases and tough requirement of annotation by the professional radiologists and 2) the weakness of cluster characteristics due to the small between-class distance of medical images between lesions and surroundings and the large within-class distance caused by the diversity of cases and imaging technologies.

In this study, we explore the effects of multi-parametric data fusion, transfer learning, and multi-scale features on medical imaging diagnostics, when we are facing the above two

challenges, the lack of sufficient labeled training data to train deep learning models and the complexity and diversity of medical images. At first, we retrospectively collected the clinical multi-parametric magnetic resonance imaging (mp-MRI) data of patients who were diagnosed as hepatocellular carcinoma (HCC) by pathology examination. Secondly, we train the proposed multi-scale deep residual neural network on natural image dataset as the initial state of the network. Then, we fine-tune the network on the collected medical image dataset. Finally, we combine the doctor's experience to comprehensively analyze and summarize the experimental results.

The rest of the paper is organized as follows. In Section 2, we review the related work on the computer-aided diagnosis and deep learning. Section 3 mainly introduces the collected mp-MRI data and our inspiration for data fusion idea based on the analysis of collected data and doctor's clinical diagnosis experience. In Section 4, we elaborate the specific details of the proposed model based on the multi-scale deep residual neural network with the pre-training processing on the natural images. Section 5 presents comparative experiments and performance evaluation with multiple metrics of classification accuracy, sensitivity, specificity and AUC on the clinical datasets from the collected T2WI, T1WI and DCE-MRI images with the standard pathology annotations. The conclusion and future work are given in Section 6.

2. Related Work

In recent years, the computer-aided diagnosis (CAD) based on deep learning methods has been applied to detect, segment, and hierarchical diagnose brain, retina, breast, lung, and abdominal lesions [4]–[7]. Wu Zhou et al. proposed that texture features indexed by mean and GLN based on the arterial phase of DCE-MR images reflect biologic aggressiveness, and have potential applications to predict the histological grading of HCC preoperatively [8]. However, there is no related work to apply deep learning methods in the assessment of the HCC differentiation due to the lack of sufficient calibration training data, the complexity of medical images and the diversity of image capturing methods.

2D Convolutional Neural Networks (CNNs), which have powerful feature expression capability, have shown their better performance on classification, target detection, and segmentation of natural images than traditional feature extraction methods [9]–[11]. Deep learning achieves such rapid development, due to not only the improvement and innovation of algorithms itself but also the massive data and powerful computing resources. For example, many excellent models, such as VGG net [12] and ResNet [13] have been proposed in the aspect of natural image recognition. These methods have achieved good recognition performance on common image datasets, such as Imagenet, COCO, CIFAR and MNIST.

However, having plenty of well-labeled training data is not always feasible in many practical applications especially in the medical image processing areas. Referring to the human perception experience to do determination of a new object based on transferring the knowledge from the known domain to the unknown domain, transfer learning methods have been explored, which pre-train models from other domains of images like natural image datasets, and then fine-tune models with medical image data to improve the effect of diagnosis under the situation without massive well-annotated training samples [14]–[17]. In the paper [14], the data in the source domain and target domain are Human Epithelial-2 (HEp-2) cell images, and cross-modal transfer learning is performed between similar data. We want to further pre-train on the general image, so that the network has a better initialization, and then train on the multi-parametric MRI image. Paper [15] pre-trained the model on five texture

databases so that the model can distinguish the texture differences well, and then apply the trained model to the pattern analysis of the lung image. We pre-train the model on the dataset of natural images used for image classification and expect to have a better initialization of the model, to ensure and accelerate the convergence of the model. Then we fine-tune the model on the medical image, expecting to learn the fusion information of the multi-modality medical image. What we try to extract is more than texture. In [17], by using a large amount of data to pre-train the model, then fixing a part of the weight of the model, and then using the medical image to train the weights of other parts of the model, their work proves the feasibility of the transfer learning method in the medical image classification. Based on the Resnet architecture [13], we use the transfer learning method and propose a multi-modality data fusion and multi-scale feature fusion method for multi-parametric MRI images. We verify the feasibility of transfer learning, multi-modality data fusion, and multiscale feature fusion on medical image classification through experiments on clinical data.

Objects in nature have different states of expression, which depend on the scale of observation. The images of scales in the scale space will become more and more blurred, which can simulate the formation of the target on the retina of eyes when the target is from near to far. When using a computer model to analyze an unknown scene, the model does not know the object dimensions in the image in advance. The pyramid is a multi-scale representation of the image. The pyramid multi-resolution generation is faster and takes up less storage space. Representative methods for image multi-resolution and multi-scale applications are Laplacian gold tower [18] and scale-invariant feature transform [19], which combine scale space expression and pyramid multi-resolution expression. In addition to image object detection, Kaiming He et al. [20] found that the space pyramid pooling layer cannot convert the size of the candidate region into a fixed-length feature for the prediction of the target category. Tsung-Yi Lin et al. [21] proposed an FPN (Feature Pyramid Network) algorithm which simultaneously utilizes the high-level semantics of low-level features and high-level features, and achieves the prediction by merging the features of these different layers. In terms of image segmentation, Kamnitsas et al. [22] used multi-resolution medical images as the input of multi-branch networks to enhance the tumor segmentation effect. Since the use of feature pyramid networks or multi-branch networks will increase the number of parameters of the network, this paper uses spatial pyramid pooling to obtain features of different scales. Similarly, the size of HCC tumors also has different scales. In order to adapt to different sizes of tumors, we first normalize the tumor data of all patients to the same size and then use the spatial pyramid pool to obtain the characteristics of the data at different scales.

3. Motivation

3.1 Multi-parametric MRI

We constructe a multi-parametric magnetic resonance imaging (mp-MRI) dataset to evaluate the differentiation of HCC. One of the most valuable MRI sequence images for the MCC diagnosis is DCE-MRI. The DCE-MRI data contains six phases, which are plain scan, early arterial phase, late arterial phase, portal phase, equilibrium phase, and delay phase, which are represented by S0-S5 in **Fig. 2**. Each phase data contains three dimensions (coronal plane, sagittal plane and cross section), we extract HCC tumor (ROI) data from each three-dimensional MRI image according to doctors' labeled information. The extracted diagrams of DCE-MRI data are shown in **Fig. 2**.

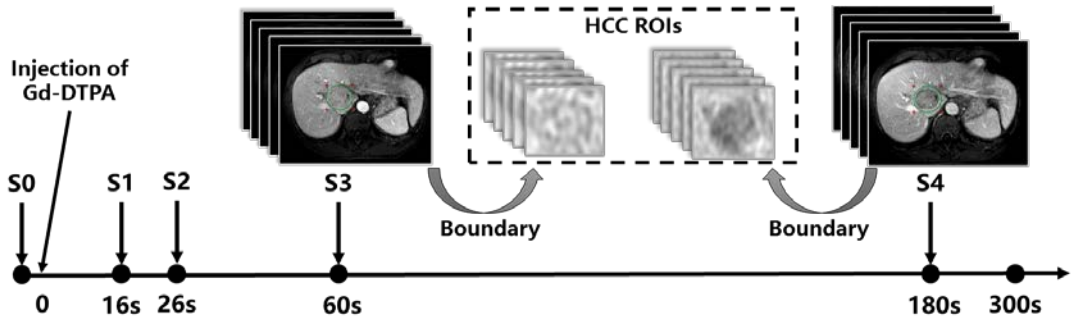


Fig. 2. Six phases of DCE-MRI image

Fig. 2 shows six phases of DCE-MRI image. S0 represents the MR images acquired before injection of Gd-DTPA contrast agent, and S1, S2, S3, and S4 represent the MR images acquired 16, 26, 60, and 180 seconds after the injection of Gd-DTPA contrast agent, respectively. The collected images are all three-dimensional stereoscopic images.

In addition to the DCE-MRI with time series information, the same method is used to obtain the labeled images of T2 weighted imaging (T2WI), chemical shift imaging (in-phase T1WI/IN and out of phase T1WI/OUT). Since the scanning thickness (6.0mm) of the T2WI, T1WI/IN and T1WI/OUT is greater than that of DCE-MRI (2.2mm), two-dimensional regions of each tumor data are re-trained in three modes.

We extract 166 effective lesion samples based on the doctors' labeled information, and each sample data contain nine two-dimensional regions and nine sequences that are illustrated in **Fig. 3**.

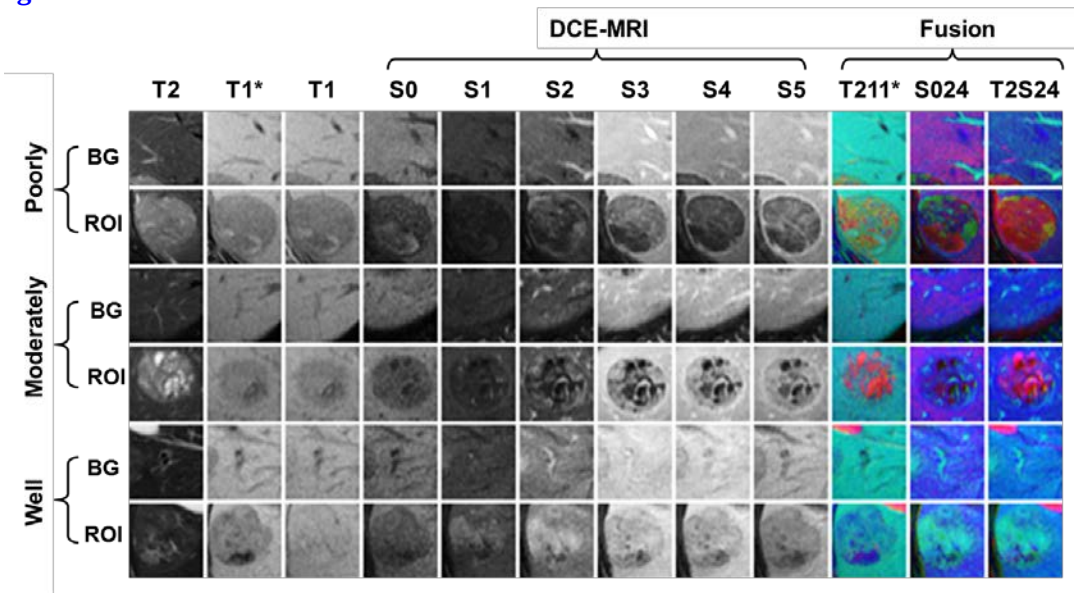


Fig. 3. T2WI, T1WI and reverse phase, DCE-MRI images and their fused images

In **Fig. 3**, T2, T1* and T1 are the abbreviation of T2WI, T1WI/IN and T1WI/OUT, respectively. S0, S1, S2, S3, S4 and S5 represent the different sequence series of DCE-MRI in **Fig. 2**, respectively. In the last three columns, we use three of the previous nine columns gray image of MRI sequences as the three channels of a colorful image. T211* represents the fusion of T2WI, T1WI/IN and T1WI/OUT, S024 represents the fusion of S0, S2 and S4. T2S24

represents the fusion of T2WI, S2 and S4. BG and ROI represent the non-tumor region of liver and HCC tumor region, respectively. The poorly, moderately and well are three grades: poorly differentiated, moderately differentiated and well differentiated of HCC, respectively.

As we can see from **Fig. 3**, the tumor details in the color image are more prominently displayed than that in the gray image, and the fused images obtained with different modality combination are also different. Through experiments, we choose modalities with complementary characteristics for the fusion, which facilitate to achieve better performance for classification. To use two-dimensional tumor image for diagnosis, single modality image and multi-modality fused image can correspond to grayscale and color images in the natural images respectively, which allows us to pre-train the model on natural image dataset and then fine-tune the pre-trained model on the acquired medical image dataset for disease diagnosis.

We normalize all data samples in **Fig. 3** into the same size. However, in fact, the size of the tumor is different, and the size information of the tumor plays an important role in the clinical diagnosis. In order to use the tumor scale information, it is necessary to learn from the method of processing multi-scale targets in natural images. After the images pass through several convolutional layers, the spatial pyramids are used to extract the multi-scale features of the medical images.

3.2 Time-signal intensity curve

In the paper, the data-fusion performance analysis is done from the perspective of the qualitative analysis with the data visualization and the quantitative analysis with the time-signal intensity curvy (TSIC). We calculate the time-signal intensity curves (TSIC) of DCE-MRI by quantifying the change of MRI signal intensity with time of DCE-MR sequences. Fabijańska et al. [23] found that TSIC has three main typical patterns, which respectively correspond to the degrees of malignant characterization of the prostate cancer, can be used to diagnose the malignant degree of the prostate cancer. In this paper, we also calculate the average signal strength value of the region as the current sequence strength by extracting the 16×16 voxel center region of the ROI of S0~S5 images. The TSIC curves of the three grades of HCC differentiation and background regions are shown in **Fig. 4**.

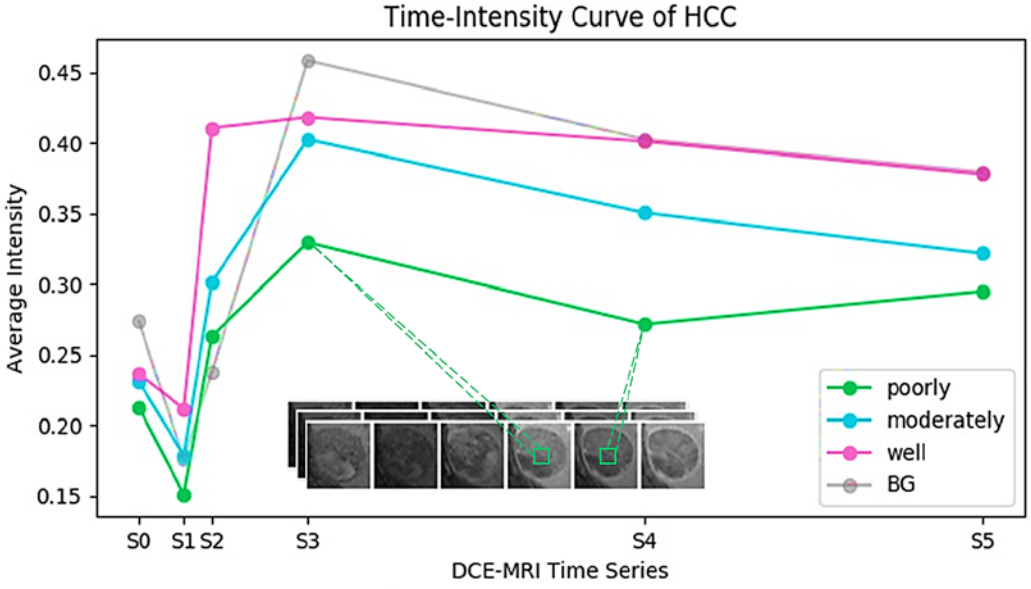


Fig. 4. Average TSIC curve

In **Fig. 4**, the horizontal axis of the graph is the sequence series of the DCE-MRI, and the vertical axis is the average signal intensity of the one DCE-MRI sequence, the TSIC curves of HCC area, including well differentiated, moderately differentiated, poorly differentiated, and non-tumor area (background, BG) are separately calculated.

According to the DCE-MRI TSIC curve of lesions, it can be seen that the overall trend of the three HCC differentiation data is similar, but there is a progressive relationship between three curves of lesions. The well differentiated with the lowest degree of malignancy is closer to the non-tumor area. The curve trend reveals the practical diagnosis experience that the blood perfusion characteristics of tumors reflected by DCE-MRI images play an important role in the diagnosis of HCC differentiation. Therefore, making full use of multi-parametric MRI images with enhancing their contribution is beneficial to achieve a more accurate diagnosis of HCC differentiation than the single modality of MRI.

4. Methodology

4.1 Multi-scale residual neural network

Based on the above data analysis, we propose our methodology from three perspectives. One is to select a deep network with good performance in feature representation learning for the complex objects: HCC lesion with multi-modality MRI. One is the data-fusion scheme in making full use of data complementary and enhancing contribution in classification. Another one is transfer learning to do the pr-training with mitigating the impact of lack of training samples. The details are illustrated as follows.

The basic network used in this paper is ResNet20. To have better ability in revealing the feature of the lesion, the approach is adopted referring to [13] which replaces the GAP layer (Global Average Pooling) with the SPP (Spatial Pyramid Pooling) layer to extract features from different scales as shown in **Fig. 5**. Therefore, together with the Resnet and SPP framework. it is helpful to increase the depth of the network with avoiding the gradient disappearance and facilitates to learn the discriminative representation with multiple scales.

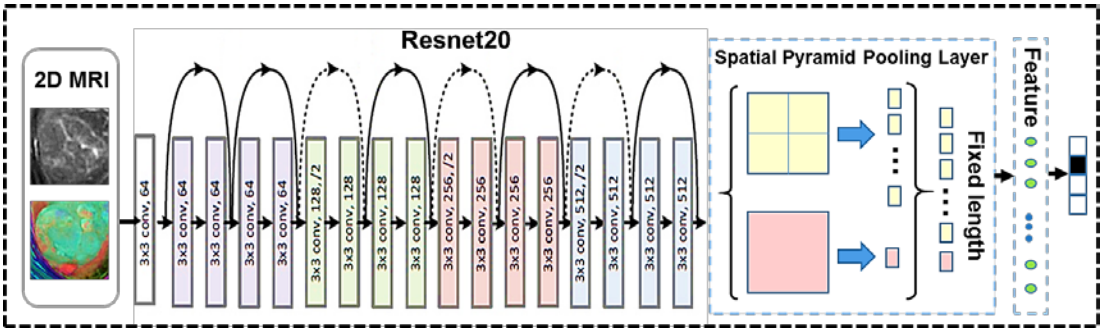


Fig. 5. ResNet 18+SPP

As shown in the left side of **Fig. 5**, the inputs of the network can be the single mode gray image or the fusion mode colorful image. The input image is fed into ResNet20 to extract the image features, and then the features are sent to the Spatial Pyramid Pooling Layer to obtain a 320-dimensional multi-scale feature.

The 320-dimensional multi-scale feature extracted from the above network is fed into the single-layer fully connected neural network, and the number of neurons of the fully connected neural network is 2 (for visualization), 16 or 32, respectively. The network is pre-trained on

MNIST [24] and CIFAR10 [25] to extract features of 2D single mode and mixed mode patches. The model in this study is established by using PyTorch [26] and trained on a computer with a GeForce GTX 1080 (NVIDIA, Santa Clara, Calif) graphics processing unit, a Core i7-6700K 4.00-GHz (Intel, Santa Clara, Calif) central processing unit, and 16 GB of random access memory. The performance of the network on MNIST and CIFAR10 is shown in **Table 1**.

Table 1. Benchmark results of the ResNet20+SPP model

Model		**MNIST**		**CIFAR10**	
Type	Feature number	Train	Test	Train	Test
ResNet	2	99.58	99.18	94.08	86.76
ResNet+SPP	2	99.98	99.21	94.39	87.44
ResNet+SPP	16	99.97	99.54	99.17	90.50
ResNet+SPP	32	100	99.61	99.10	90.48

The experiment results on the public datasets: MNIST and CIFAR10 show that the improved discriminative characteristic with SPP on the multi-scale residual neural network than without SPP based on the visualization of feature distribution.

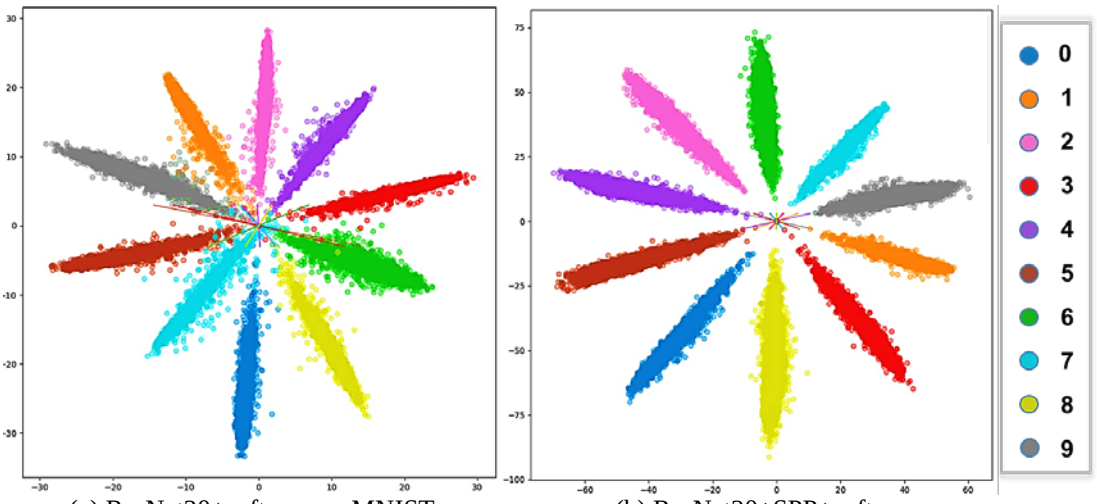
4.2 Loss function

The usual classification task uses softmax to obtain probabilities and trains the network by minimizing the cross-entropy loss.

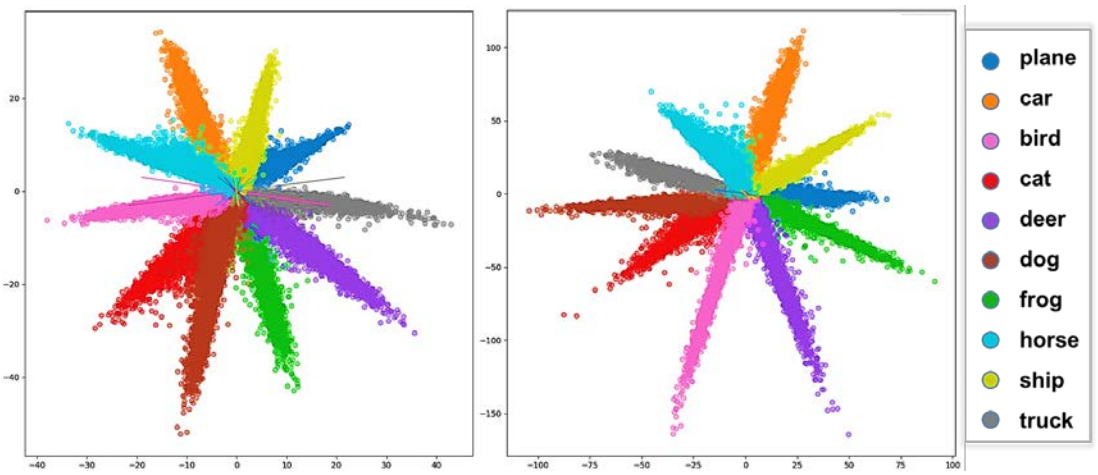
$$L(x,y) = -\sum_{j=1}^{k} 1\{y = j\} \log_2(f_j(x)), \tag{1}$$

where $1\{\bullet\}$ is an indicator function, it's value is 1 only when the function's argument is true, $f_j(x)$ represents the probability that the model predicts that sample x belongs to the category j, k represents the number of categories.

The above loss function is used to train the network which is embedded as a 2-dimensional feature on the MNIST dataset. The visualization of the extracted features is shown in **Fig. 6**.



(a) ResNet20+softmax on MNIST (b) ResNet20+SPP+softmax



(c) ResNet20+softmax on CIFAR10 (d) ResNet20+SPP+softmax on CIFAR10

Fig. 6. Visualization of network feature mapping

Fig. 6 (a) and **Fig. 6** (c) shows the 2D features extracted after training the original ResNet20 on MNIST and CIFAR10, respectively. **Fig. 6** (b) and **Fig. 6** (d) show the 2D features extracted after replacing the GAP Layer in ResNet20 with SPP Layer. The extracted 2D features are trained by using the cross-entropy loss function and softmax classifier.

It can be seen from **Table 1** that replacing GAP layer with SPP layer can improve the performance of handwriting and natural image classification. It can be seen from **Fig. 6** (a) and **Fig. 6** (b) that comparing with the features obtained by original ResNet20, the network using the multi-scale features spatially diverge from the origin points. The multi-scale features are not concentrated near the origin points and conducive to classification, which is consistent with the classification results in **Table 1**. As is known that CIFAR10 is more complex than MNIST, feature extraction on CIFAR10 is relatively more difficult. Compare with **Fig. 6** (c), the overlap between the characteristics of different categories of CIFAR10 data in **Fig. 6** (d) is significantly reduced, which also means that such features are more easily distinguishable.

5. Experiment Results and Analysis

5.1 Medical image data preprocessing

In this study, we obtain 51 effective HCC lesions from 48 patients by using the GE 3.0T 750 magnetic resonance imaging system in Beijing Friendship Hospital of Capital Medical University. Since the pathology is the gold standard, an experienced pathologist judges the degree of HCC differentiation based on the microscopic findings of surgically resected HCC specimens. The labels of pathological HCC differentiation are confirmed by liver biopsy examination. Multi-parametric nuclear magnetic images used in this study include T2 weighted imaging (T2WI), chemical shift imaging (in-phase and out of phase) and dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI). A radiologist who has been engaged in medical imaging diagnosis for 10 years labeled the location of HCC tumors, and then a chief radiologist who has been engaged in abdominal imaging diagnosis for 30 years examined these labels.

We normalize the volume intensities to the range of $[0,1]$ by using Equation (2) for the DCE-MRI, T2WI, and T1WI/IN, T1WI/OUT images have higher intensity values than natural

images, and there are some sparse noises with high-intensity values.

$$I' = (I - I_{\min}) / (I_{\max} - I_{\min}) \quad , \tag{2}$$

where I' is the normalized image gray value, I is the original gray value of the image, $I_{\min}$ is the small gray value of the image, and $I_{\max}$ is maximum intensity value after trimming the top 1% grayscale values. Such preprocessing has been used in [6], [27].

Due to the limitation of quality and data completeness of medical image data, the size medical image datasets are typically small compared to those in computer vision. In order to prevent the over-fitting, the rotating 90° and horizontal and vertical flip can increase the size of the original dataset by 3 and 2 times, respectively. Therefore, after using the data augmentation method, the number of samples in the dataset can be 6 times of its original size.

5.2 Dataset partitioning

For the task to distinguish HCC and non-tumor area (BG), there are 166 samples of HCC and 166 samples of background, and the dataset is randomly divided into a training dataset and testing dataset, according to the ratio of 3:2, and then augmented the training data and testing data, separately.

Table 2. Specification of the numbers of samples in the three categories of for BG and HCC for collection/augmentation

Datasets	Samples | Augmentation	
	HCC	BG
Training	99|594	99|594
Testing	67|402	67|402
Total	166|996	166|996

To distinguish the differentiation of HCC, the specific numbers of samples in the three categories of the dataset for collection/augmentation are shown in **Table 3**.

Table 3. Specification of the numbers of samples in the three categories of the dataset for collection/augmentation

Datasets	Samples | Augmentation		
	Poorly	Moderately	Well
Training	10|60	68|408	20|120
Testing	8|48	46|276	14| 84
Total	18|108	114|684	34|204

5.3 Performances of single-parametric MR images and fusion data

In order to ensure the reliability of the results, we repeat the experiment 10 times, and then calculate the average and standard deviation of all the experiments. The 7 separate modes, T2WI, S0, S1, S2, S3, S4, and S5 in **Fig. 3** and the fusion mode, T211* (fusion of T2WI, T1WI/IN, T1WI/OUT), T2S24 (fusion of T2WI, S2, S4), S024 (fusion of S0, S2, S4) and T2S02 (fusion of T2WI, S0, S2) are used to train the original ResNet20 to distinguish HCC and non-tumor area (BG). The experimental results of ResNet20 are shown in **Table 4**.

Table 4. Results of ResNet20 for distinguishing HCC and non-tumor area (BG)

	Accuracy	Sensitivity	Specificity	AUC
T2WI	0.891±0.012	0.905±0.031	0.874±0.021	0.890±0.010
S0	**0.941±0.007**	**0.941±0.005**	**0.940±0.016**	**0.941±0.008**
S1	0.847±0.014	0.805±0.030	0.899±0.026	0.852±0.013
S2	0.917±0.000	0.929±0.011	0.902±0.021	0.915±0.010
S3	0.898±0.012	0.915±0.025	0.877±0.012	0.896±0.010
S4	0.857±0.014	0.848±0.016	0.867±0.032	0.858±0.015
S5	0.869±0.012	0.842±0.030	0.904±0.024	0.873±0.011
T211*	0.857±0.018	0.849±0.010	0.867±0.018	0.858±0.010
T2S24	0.897±0.023	0.856±0.036	0.947±0.023	0.902±0.023
S024	0.961±0.017	0.959±0.012	0.961±0.011	0.960±0.014
T2S02	**0.974±0.004**	**0.960±0.008**	**0.991±0.007**	**0.976±0.004**

*Data expressed as the mean±SD

The analysis of the individual modality indicators is shown in **Table 4**, The modalities of T2WI (T2 weighted imaging), S0 (plain scan) and S2 (late arterial phase) achieve better performance for the distinguishing of the HCC and BG than other modalities. It can be seen from the comparison results of fusion mode and single mode that S024 and T2S02 achieve better effects than the single mode, while T211* and T2S24 are worse than the single mode of T2WI. Therefore, multiple modality fusions need to be selected. It is necessary to find complementary data for the fusion to get better results. The fusion of T2WI (T2 weighted imaging), S0 (plain scan), and S2 (late arterial phase) of DCE-MRI images achieves a better performance, which is consistent with the doctors' diagnostic experience.

5.4 Performance analysis with Transfer learning

For the task of distinguishing between HCC and non-tumor areas, radiologists are able to make judgments based on medical images, but they usually have to carefully observe the appearance of tissue in multiple images, like DCE-MRI, T2WI and DWI (Diffusion weighted imaging), to make comprehensive judgments. The accuracy of the results is related to the radiologist's experience, and the judgment process usually takes a long time. In this section, we use the clinical MR images to train the model for distinguishing between HCC and non-tumor areas.

In order to verify the feasibility of transfer learning in medical imaging diagnosis, ResNet20 [13] and ResNet20+SPP in **Fig. 5** are trained respectively on CIFAR10 dataset. The output features of ResNet20+SPP are encoded into a 32-dimensional feature vector, which achieves an accuracy rate of 90.19% in the verification set of CIFAR10.

In this part, The fusion of T2WI (T2 weighted imaging), S0 (plain scan) of DCE-MRI, and S2 (late arterial phase) of DCE-MRI are used as the three channels of input images to fine-tune the ResNet20 (fine-tuned ResNet in **Fig. 7** and **Fig. 8**) and ResNet20+SPP (fine-tuned ResNet+SPP in **Fig. 7** and **Fig. 8**) models that have been trained on CIFAR10 dataset. In addition, a comparison experiment, training ResNet20 (ResNet in **Fig. 7** and **Fig. 8**) and ResNet20+SPP (ResNet+SPP in **Fig. 7** and **Fig. 8**) from random initialization, is set up. The learning rate of the fine-tuning is set to 0.0005. The average accuracy curves and the average ROC curves on the testset during the training of the four models in the 10 repeated tests are illustrated in **Fig. 7** and **Fig. 8**.

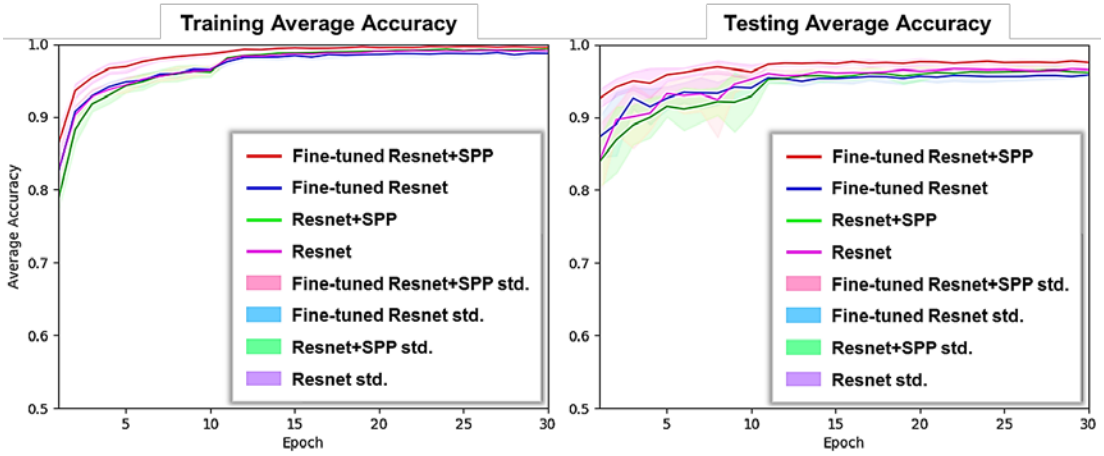


Fig. 7. Training and Testing process

In **Fig. 7**, the four lines represent the mean values of accuracy over 30 epochs for the four models mentioned above. The shadow parts represent the standard deviations of the accuracy.

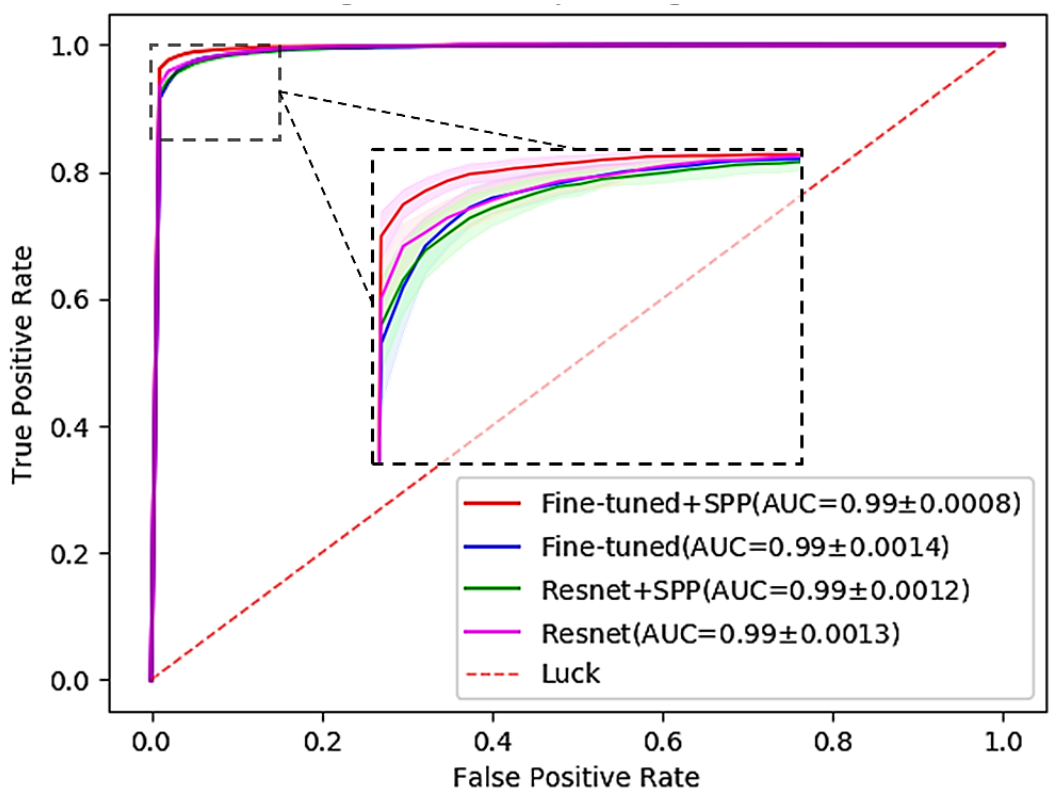


Fig. 8. Average Receiver operating characteristic

In **Fig. 8**, the four lines represent the mean values of AUC of 10-time experiments for the four models mentioned above. The shadow parts represent the standard deviations of the AUC. From the training process, it can be seen that the pre-trained model can converge more quickly. The evaluation results of the above four model for distinguishing between HCC and non-tumor areas are shown in **Table 5**.

Table 5. Results of distinguishing HCC and non-tumor area (BG)

	Accuracy	Sensitivity	Specificity	AUC
ResNet	0.974±0.004	0.960±0.008	0.991±0.007	0.991±0.0013
ResNet+SPP	0.968±0.005	0.957±0.012	0.987±0.009	0.990±0.0012
Pre-trained ResNet	0.975±0.010	0.979±0.015	**0.992±0.016**	0.993±0.0014
Pre-trained ResNet+SPP	**0.981±0.002**	**0.983±0.008**	0.991±0.007	**0.999±0.0008**

*Data expressed as the mean±SD

As is shown in **Table 5**, the Fine-tuned ResNet+SPP can achieve the highest accuracy. The final test accuracy, sensitivity, specificity and AUC are 0.981±0.002, 0.983±0.008, 0.991±0.007 and 0.999±0.0008, respectively. It can also be seen from the average ROC curves that Fine-tuned ResNet+SPP achieves better performance other methods.

5.5 Performance of assessment of HCC differentiation

Even experienced radiologists cannot accurately distinguish HCC differentiation based on MR images. The gold standard to evaluate the differentiation of HCC is a pathological examination. In this section, we use the clinical MR images to train the model and evaluate the HCC differentiation. The results of HCC differentiation using the eight separate modalities of T2WI, T1*, S0, S1, S2, S3, S4, and S5 in **Fig. 3** are shown in **Fig. 9**.

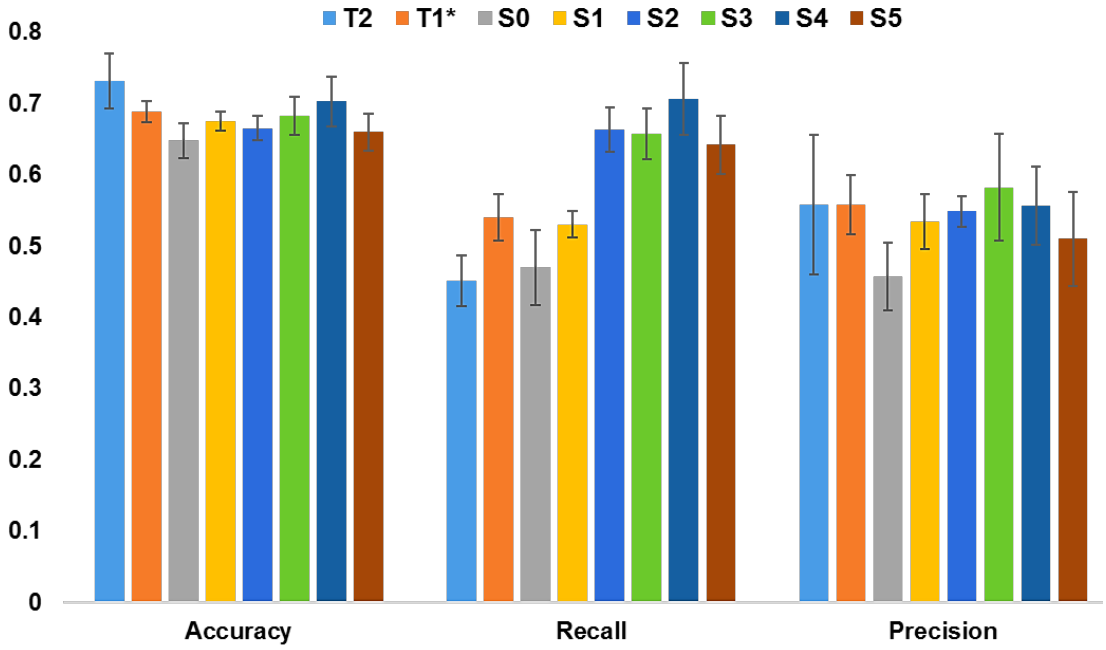


Fig. 9. The contribution of each mode

In **Fig. 9**, the height of the histograms represents the mean values of evaluation indexes, and the standard deviations of these values are also illustrated in the figure. The modalities of T2WI, T1WI/OUT and S4 (equilibrium phase) achieve better performance for the HCC differentiation than other modalities. However, later phases of DCE-MRI, like S2 (late arterial phase), S3 (portal phase), S4 (equilibrium phase) and S5 (delay phase), have a higher recall for

distinguishing HCC differentiation. In order to combine the advantages of multiple MRI data and find the best data fusion scheme to distinguish the degree of HCC differentiation, multiple groups of experiments are compared. There are 56 combinations without repetition to choose three modes from eight modes for combination, we can't experiment in all cases. According to the results in **Fig. 9** and the doctor's experience, we list the following reasonable combinations, T211* (fusion of T2WI, T1WI/IN, T1WI/OUT), T2S24 (fusion of T2WI, S2, S4), S024 (fusion of S0, S2, S4) and T21*S4 (fusion of T2WI, T1WI/OUT and S4). The results for distinguishing HCC differentiation are shown in **Table 6**.

Table 6. Data fusion results of the ResNet20+SPP

	Pre-train	Accuracy	Sensitivity	Precision	F1
T211*	False	0.699±0.053	0.586±0.096	0.554±0.081	0.479±0.093
	True	0.681±0.025	0.578±0.073	0.539±0.045	0.463±0.046
T2S24	False	0.697±0.044	0.683±0.063	0.569±0.072	0.549±0.064
	True	0.706±0.041	0.569±0.695	0.531±0.035	0.500±0.037
S024	False	0.716±0.014	**0.769±0.012**	0.594±0.022	0.605±0.021
	True	0.724±0.055	0.763±0.066	0.590±0.052	0.597±0.063
T21*S4	False	0.766±0.038	0.734±0.069	0.632±0.037	0.614±0.051
	True	**0.777±0.033**	0.768±0.038	**0.655±0.016**	**0.636±0.033**

*Data expressed as the mean±SD

As can be seen from the results of several groups of experiments in **Table 6**, the combination of images using only several phases of DCE-MRI has the highest recall. T21*S4 (fusion of T2WI, T1WI/OUT and equilibrium phase of DCE-MRI) possessing the advantages of T2WI, T1WI and DCE-MRI achieve better comprehensive classification performance. In most cases, transfer learning can improve classification accuracy. All the results in **Table 6** show the results after embedding features into two-dimensional features. In the following experiments, we increase the dimension of features to improve the representation ability of the model.

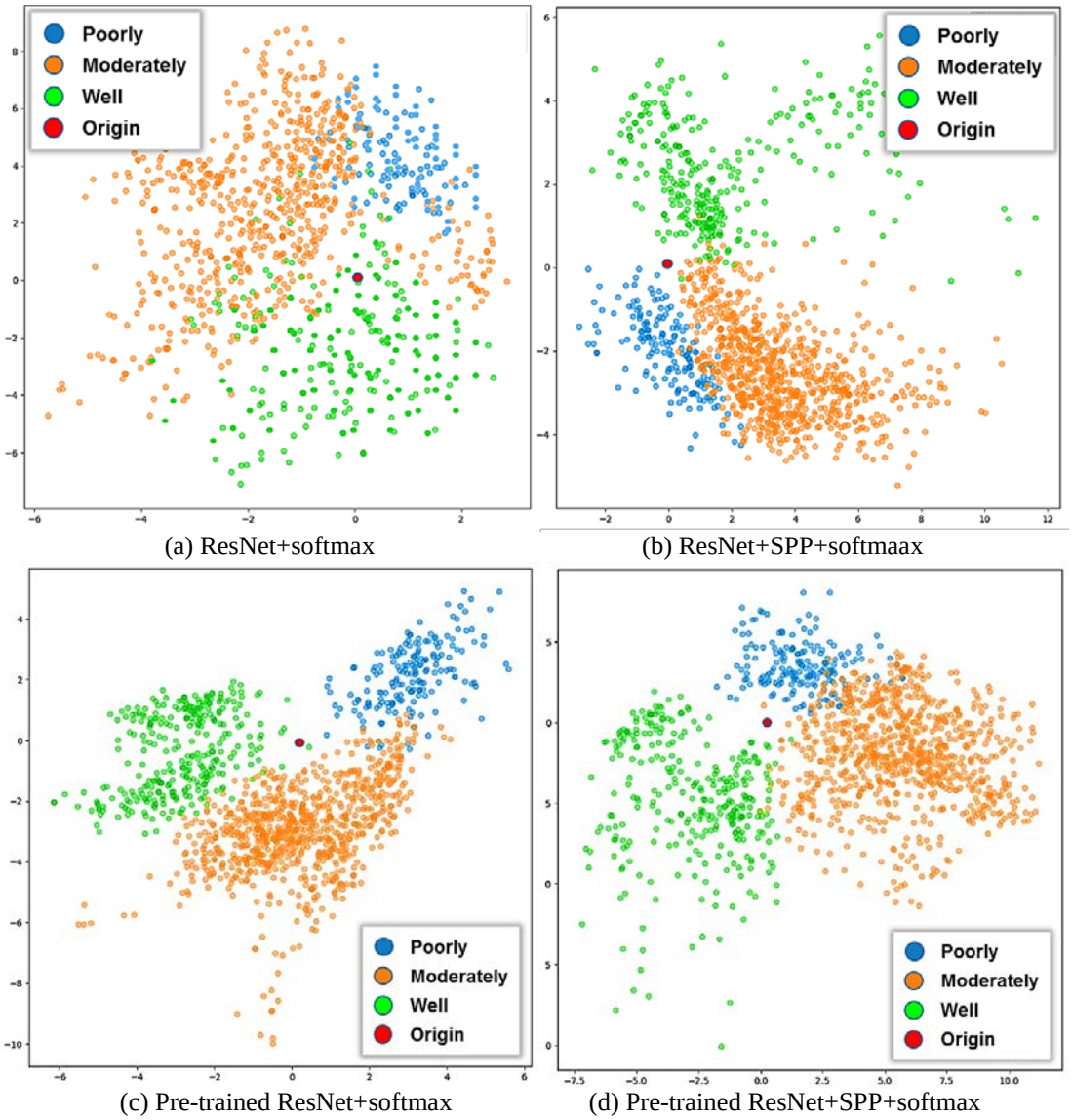
The fusion of T2WI, T1WI/OUT and S4 (equilibrium phase of DCE-MRI) is used to fine-tune the network. After fine-tuning the pre-trained ResNet+SPP on the training dataset, the performance on the testing dataset is shown in **Table 7**. Features are embedded as 2, 16, 32 and 64-dimensional feature vectors, respectively.

Table 7. Results of pre-trained ResNet+SPP discriminate HCC differentiation

Feature	Accuracy	Recall	Precision	F1
2	0.777±0.033	0.768±0.038	0.655±0.016	0.636±0.033
16	0.828±0.056	0.849±0.052	0.688±0.046	0.697±0.061
32	**0.847±0.023**	**0.839±0.028**	**0.724±0.0212**	**0.737±0.032**
64	0.836±0.043	0.834±0.042	0.716±0.034	0.704±0.051

*Data expressed as the mean±SD

The visualization of the extracted features is shown in **Fig. 10**. The red dots in **Fig. 10** represent the origin point, and the other three colors represent the positions in the feature space after the data of three categories are embedded into two-dimensional features.



(a) ResNet+softmax (b) ResNet+SPP+softmaax

(c) Pre-trained ResNet+softmax (d) Pre-trained ResNet+SPP+softmax

Fig. 10. The visualization of network feature mapping on the discrimination of HCC differentiation

The weights of the model in **Fig. 10** (a) and **Fig. 10** (b) are randomly initialized, and then the model is trained directly on medical images. The model in **Fig. 10** (c) and **Fig. 10** (d) is pre-trained on CIFAR10, and the weights of the model except the output layer are retained. Then, the learning rate is set to half of the original to fine-tune the model on medical images. The extracted two-dimensional features are visualized.

Compared with the features in **Fig. 10** (a) and **Fig. 10** (b) extracted by randomly initialized model, the medical image features of the same category in **Fig. 10** (c) and **Fig. 10** (d), which are extracted by pre-trained model, are more concentrated in spatial distribution. **Fig. 10** (c) is the 2-dimensional features of testing dataset samples extracted after training the original ResNet20 on the training dataset. **Fig. 10** (d) is the 2-dimensional features of testing dataset samples extracted after training the ResNet+SPP on the training dataset. The features in **Fig. 10** (c) and **Fig. 10** (d) diverge outward from the origin in the feature space. Compared with **Fig. 10** (c),

the distribution of features in **Fig. 10** (d) spatially diverges faraway from the origin. Using multi-scale features is able to avoid the feature mapping near the origin and improve the classification performance.

6. Conclusion

In this study, we retrospectively collected the clinical multi-parametric magnetic resonance imaging (mp-MRI) data of patients who were diagnosed as hepatocellular carcinoma (HCC) by the pathology examination. We calculate and analyze the characteristics of DCE-MR time-signal strength curves.

Firstly, we use single mode data (including T2WI, T1WI/IN, T1WI/OUT and five sequence series of DCE-MRI) to train basic network separately. From the performance of single mode for the HCC differentiation and distinguishing HCC and non-tumor area (BG) on the basic network, we can draw a conclusion that 1) T2WI (T2 weighted imaging), S0 (plain scan) and S2 (late arterial phase) achieve better performance for the distinguishing of HCC and BG than other modalities. 2) The modalities of T2WI, T1WI/OUT and S4 (equilibrium phase) achieve better performance for the HCC differentiation than other modalities.

In addition, we explore the effects of multi-parametric MRI fusion, transfer learning, and multi-scale features extraction on medical imaging diagnostics. From the performance of fusion data for the HCC differentiation and the distinguishing of HCC and non-tumor area (BG), we can draw the following three conclusions. 1) It is necessary to find complementary data for the fusion, which can achieve better results when using multi-parametric magnetic resonance imaging (mp-MRI) data. 2) Although there is a big difference between natural images and medical images, the model pre-trained on the natural image dataset can ensure and accelerate the convergence of the training on medical image datasets, and improve the performance; 3) In order to train the convolutional neural network with batch, we have to normalize the size of samples. However, the original size of the tumor has a significant role in the clinical diagnosis. The proposed multi-scale deep residual neural network, which can extract multi-scale features of image samples, can help to improve the performance of medical image classification. In a word, our current pilot tests show that the deep learning based methods provide the primary acceptable diagnosis with quantity assessment, whilst the human radiologists can only do the qualitative evaluation.

References

[1] J. D. Yang and L. R. Roberts, “Hepatocellular carcinoma: A global view,” *Nature Reviews Gastroenterology & Hepatology*, vol. 7, no. 8, p. 448, 2010. Article (CrossRef Link)

[2] M. Sherman, “Hepatocellular Carcinoma: Screening and Staging,” *Clinics in Liver Disease*, vol. 15, no. 2, pp. 323–334, 2011. Article (CrossRef Link)

[3] J. M. Regimbeau et al., “Risk factors for early death due to recurrence after liver resection for hepatocellular carcinoma: Results of a multicenter study †,” *Journal of Surgical Oncology*, vol. 85, no. 1, pp. 36–41, 2004. Article (CrossRef Link)

[4] A. Esteva et al., “A guide to deep learning in healthcare,” *Nature Medicine*, vol. 25, pp. 24-29. 2019. Article (CrossRef Link)

[5] G. Litjens et al., “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, no. 9, pp. 60-88, 2017. Article (CrossRef Link)

[6] Q. Dou et al., “Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1182–1195, 2016. Article (CrossRef Link)

[7] K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61-78, 2017. Article (CrossRef Link)
[8] W. Zhou et al., "Malignancy characterization of hepatocellular carcinomas based on texture analysis of contrast-enhanced MR images," *Journal of Magnetic Resonance Imaging*, vol. 45, no. 5, pp. 1476–1484, 2017. Article (CrossRef Link)
[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. of International Conference on Neural Information Processing Systems*, pp. 1097–1105, 2012. Article (CrossRef Link)
[10] W. Liu et al., "SSD: Single Shot MultiBox Detector," in *Proc. of European Conference on Computer Vision*, pp. 21–37, 2015. Article (CrossRef Link)
[11] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. Article (CrossRef Link)
[12] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computer Science*, 2014. Article (CrossRef Link)
[13] K. He, X. Zhang, S. Ren, and S. Jian, "Deep Residual Learning for Image Recognition," in *Proc. of International Conference on Computer Vision and Pattern Recognition*, 2016. Article (CrossRef Link)
[14] H. Lei et al., "A Deeply Supervised Residual Network for HEp-2 Cell Classification via Cross-Modal Transfer Learning," *Pattern Recognition*, vol. 79, pp. 290-302, 2018. Article (CrossRef Link)
[15] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. Mougiakakou, "Multisource Transfer Learning With Convolutional Neural Networks for Lung Pattern Analysis," *IEEE Journal of Biomedical & Health Informatics*, vol. 21, no. 1, pp, 76-84, 2016. Article (CrossRef Link)
[16] H. Chang, J. Han, C. Zhong, A. M. Snijders, and J. Mao, "Unsupervised Transfer Learning via Multi-Scale Convolutional Sparse Coding for Biomedical Applications," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 40, no. 5, pp. 1182–1194, 2018. Article (CrossRef Link)
[17] D. S. Kermany et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018. Article (CrossRef Link)
[18] P. Burt and E. Adelson, "The Laplacian Pyramid as a Compact Image Code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, Apr. 1983. Article (CrossRef Link)
[19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. of IEEE International Conference on Computer Vision*, 1999. Article (CrossRef Link)
[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2014. Article (CrossRef Link)
[21] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. of International Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017. Article (CrossRef Link)
[22] K. Kamnitsas et al., "Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks," in *Proc. of International Conference on Information Processing in Medical Imaging*, pp. 597-609, 2017. Article (CrossRef Link)
[23] A. Fabijańska, "A novel approach for quantification of time-intensity curves in a DCE-MRI image series with an application to prostate cancer," *Computers in Biology & Medicine*, vol. 73, pp. 119–130, 2016. Article (CrossRef Link)
[24] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, and others, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. Article (CrossRef Link)
[25] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Computer Science Department, University of Toronto, Tech. Rep*, 2009. Article (CrossRef Link)

[26] A. Paszke et al., "Automatic differentiation in pytorch," in *Proc. of 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, 2017. Article (CrossRef Link)

[27] H. Chen, L. Yu, Q. Dou, L. Shi, V. C. T. Mok, and P. A. Heng, "Automatic detection of cerebral microbleeds via deep learning based 3D feature representation," in *Proc. of IEEE International Symposium on Biomedical Imaging*, pp. 764–767, 2015. Article (CrossRef Link)

Xibin Jia: She is a Professor in the Information faculty at the Beijing University of Technology. She received Ph.D. degree in computer science and technology from Beijing University of Technology in 2007. Her current main research interests include visual information cognition and computing, sentimental analysis, affection computing, medical image analysis and diagnosis.

Yujie Xiao: He is currently pursuing a master degree at Beijing University of Technology, Beijing, China. His current research interests include computer vision and medical image analysis and diagnosis.

Dawei Yang: He received the B.S.degree from China Medical university, Shenyang and M.S.degree from Beijing Hospital, Ministry of Health. His current research interets include the deep learning in liver disease diagnosis.

Zhenghan Yang: He is a chief physician and professor of radiology at Beijing Friendship Hospital Affiliated to the Capital University of Medical Sciences. He received Ph.D. degree in Beijing Medical University in 1999. His current main research interests include imaging diagnosis of abdominal diseases, early imaging diagnosis of hepatocellular carcinoma and precancerous lesions, development and application of new MRI technology.

Chen Lu: He is currently pursuing a master degree at Beijing University of Technology, Beijing, China. His current research interests include person re-identification and image processing.