Translational Oncology
www.transonc.com

Volume 12 Number 2 February 2019 pp. 292–300 292

A Deep Learning-Based Radiomics Model for Differentiating Benign and Malignant Renal Tumors[1]

Leilei Zhou[*], Zuoheng Zhang[†], Yu-Chen Chen[‡], Zhen-Yu Zhao[§], Xin-Dao Yin[‡] and Hong-Bing Jiang[*,¶]

[*]Department of Medical Equipment, Nanjing First Hospital, Nanjing Medical University, Nanjing 210006, China; [†]State Key Laboratory of Bioelectronics, Jiangsu Key Laboratory for Bio materials and Devices, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China; [‡]Department of Radiology, Nanjing First Hospital, Nanjing Medical University, Nanjing 210006, China; [§]Department of Nuclear Medicine, Nanjing First Hospital, Nanjing Medical University, Nanjing 210006, China; [¶]Nanjing Health Information Center, Nanjing 210003, China

Abstract

OBJECTIVES: To investigate the effect of transfer learning on computed tomography (CT) images for the benign and malignant classification on renal tumors and to attempt to improve the classification accuracy by building patient-level models. *METHODS:* One hundred ninety-two cases of renal tumors were collected and identified by pathologic diagnosis within 15 days after enhanced CT examination (66% male, 70% malignant renal tumors, average age of 62.27 ± 12.26 years). The InceptionV3 model pretrained by the ImageNet dataset was cross-trained to perform this classification. Five image-level models were established for each of the Slice, region of interest (ROI), and rectangular box region (RBR) datasets. Then, two patient-level models were built based on the optimal image-level models. The network's performance was evaluated through analysis of the receiver operating characteristic (ROC) and five-fold cross-validation. *RESULTS:* In the image-level models, the test results of model trained on the Slice dataset [accuracy (ACC) = 0.69 and Matthews correlation coefficient (MCC) = 0.45] were the worst. The corresponding results on the ROI dataset (ACC = 0.97 and MCC = 0.93) were slightly better than those on the RBR dataset (ACC = 0.93 and MCC = 0.85) when freezing the weights before the mixed6 layer. Compared with the image-level models, both patient-level models could discriminate better (ACC increased by 2%-5%) on the RBR and Slice datasets. *CONCLUSIONS:* Deep learning can be used to classify benign and malignant renal tumors from CT images. Our patient-level models could benefit from 3D data to improve the accuracy.

Translational Oncology (2019) 12, 292–300

Introduction

The widespread use of various imaging modalities has increased the incidental detection of renal tumors, particularly computed tomography (CT) [1,2]. Simultaneously, the incidence of benign histology in surgical specimens has also increased [3]. Small renal tumors accounted for 85% of renal cell carcinomas, and 20%-40% of these were benign pathological findings such as cysts and angiomyolipoma(AMLs) [4]. At CT scans, it is not difficult to be diagnosed when macroscopic fat appears, but diagnosis is challenging for AMLs with minimal fat. The differential diagnosis of renal tumors is the most important prognostic factor affecting patient survival and management.

Address all correspondence to: Dr. Yu-Chen Chen, Department of Radiology, Nanjing First Hospital, Nanjing Medical University, Nanjing 210006, China. E-mail: chenyuchen1989@126.com or Prof. Hong-Bing Jiang, Department of Medical Equipment, Nanjing First Hospital, Nanjing Medical University, Nanjing 210006, China. E-mail: cmdjhb@126.com

[1] Funding: This work was supported by National Natural Science Foundation of China (81600638), 14th "Six Talent Peaks" Project of Jiangsu Province (YY-079), Nanjing Health Youth Talent Project, Nanjing Department of Health (QRX11033).
Received 26 July 2018; Revised 28 October 2018; Accepted 29 October 2018

1936-5233/19
https://doi.org/10.1016/j.tranon.2018.10.012

Radiomics has been proposed to extract quantitative features from radiographic images and build models relating image features to pathological results [5]. In the past few years, some radiomics models have been proposed to classify renal tumors. Hodgdon et al. [6] used texture analysis to differentiate AMLs from renal cell carcinoma (RCC). Raman et al. [7] applied a random forest to predict the pathology of renal tumors. Feng et al. [8] used machine-learning–based quantitative texture analysis of CT images to identify different types of small renal tumors. However, the features adopted for these studies were explicitly designed or handcrafted, including their shape, intensity, texture, and wavelet textures [9]. These low-throughput features were selected based on radiologists' expert knowledge, which might limit the potential of the radiomics model.

Recently, the advent of graphics processing units and large training datasets has sparked tremendous advancement in computer vision capabilities with convolutional neural networks (CNNs) [10]. When a sufficiently large training dataset is available, a CNN can automatically extract high-throughput features and avoid the complicated process of artificial feature extraction [11]. CNNs have shown strong performances in medical fields. Esteva et al. [12] trained a CNN model with 2000 dermatological images and the corresponding pathological results and achieved the ability to classify benign and malignant skin cancers. Moreover, Arevalo et al. [13] applied a CNN to classify mammography mass lesions and achieved excellent performance with results of 79.9%-86.0% in terms of area under the receiver operating characteristic (ROC) curve (AUC). However, whether such an approach can assist in accurately differentiating benign from malignant renal tumors based on CT images has not yet been fully explored.

Unfortunately, the labeled medical data are inadequate and not easily available [14]; this is also the case for renal tumor data. Transfer learning is often used to solve small dataset problems. However, most researchers only trained the last fully connected (FC) layers [15,16]. Thus, regulating the trainable layers in transfer learning is worthwhile to investigate which approach is best for renal tumor recognition. Moreover, applications of CNN models in the medical image field mainly utilize 2D data, for example, chest radiograph classification [17] and mitotic detection of histological images [18]. These methods ignore the contexts of image sequences on the *z*-axis. A 3D CNN model was built for a special application [19]. However, there is a dearth of pretrained 3D models. Moreover, such models carry high computational costs, which restrict their application in medical fields.

In this study, we aimed to classify benign and malignant renal tumors from CT images by taking advantage of transfer learning. This paper also discussed the effects of freezing different weight layers during transfer learning and selecting one optimal image-level model. On this basis, two patient-level models were established by merging multislice CT image features. We hypothesized that CNN transfer learning (image-level model) can be used to classify renal tumor CT images and that comprehensive consideration of each patient's full set of images (patient-level model) can improve diagnostic accuracy.

Materials and Methods

General Information

From January 2013 to September 2018, a total of 192 cases of renal tumors (mean maximum diameter, 48.19 mm; range, 5-160 mm) identified by enhanced CT examination were collected in our hospital. The collection of this dataset was approved by the Institutional Review Board, and we obtained waived written informed consent. Of these cases, 127 were males, and 65 were females; their average age was 62.27 ± 12.26 years. All the patients underwent surgery within 15 days after CT examination. The final pathological diagnosis distinguished 134 cases of malignant renal tumors (98, 16, and 20 cases were of American Joint Committee on Cancer stage I, II, and III [20]), including 117 clear-cell RCCs (20, 68, 24, and 5 cases were of Fuhrman grade I, II, III, and IV [21]), 8 papillary RCCs, and 9 other RCC subtypes. The other 58 cases were benign; these included 50 renal cysts and 8 renal AMLs (Table 1).

Acquisition of CT images

The data were collected from a Siemens Somatom Definition Flash dual-source CT and a Philips 128-slice spiral CT. The nonionic iodine contrast agent was bolus, injected into the anterior cubital vein at a dose of 1.5-2 ml/kg and a speed of 2.5-3.0 ml/s by high-pressure syringe with the contrast agent automatic trigger technique. All the studies involved at least one-phase scanning, including corticomedullary phase cases (n = 192), nephrographic phase (n = 188) cases, and excretory phase (n = 118) cases. The scanning parameters were as follows: voltage 120 kV, tube current 260 mAs, thickness of scan layer 5 mm, pitch 1, thickness of reconstruction layer 1 mm, and a matrix of 512 × 512.

Image Preprocessing

Gomes et al. [22] found that the accuracy of RCC subtype differentiation with single-phase corticomedullary contrast-enhanced CT was comparable to that of multiphasic imaging. Yan et al. [23] observed a better tumor classification with corticomedullary phase for clear cell RCC versus papillary RCC. In this study, deep learning was performed on the corticomedullary phase CT images. The CNN model used for transfer learning was Inception V3 [24] pretrained on ImageNet. Bar et al. [25] migrated a CNN model trained on natural images to process tasks using medical images.

Gao et al. [26] also confirmed that images using different CT attenuation channels obtain better classification results than images using a single channel. Therefore, to make full use of the three RGB input channels in Inception V3, we changed the image according to three CT attenuation ranges: normal renal attenuation range (−110 to 190 HU), high attenuation range (20-120 HU), and low attenuation range (−40 to 60 HU). The low attenuation range was used to capture high-intensity patterns such as those from clear-cell RCC and papillary RCC. The normal renal attenuation range was the most commonly used one for the imaging diagnosis of renal regions, as the attenuation value was different between various tissues. The high attenuation range facilitated revealing low-intensity tissue such as cysts and AMLs. The image preprocessing flow was shown in Figure 1*A*.

Table 1. General Information of Dataset

Characteristic	Benign	Malignant	*P* Value
Patients			
Number of samples (*n* %)	58 (30.21)	134 (69.79)	
Tumor location			.002
Exophytic (*n* %)	33 (42.31)	45 (57.69)	
Mesophytic (*n* %)	21 (27.27)	56 (72.73)	
Endophytic (*n* %)	4 (10.81)	33 (89.19)	
Gender			.432
Male (*n* %)	36 (28.35)	91 (71.65)	
Female (*n* %)	22 (33.85)	43 (66.15)	
Age (mean ± SD, years)	62.40 ± 13.83	62.22 ± 11.57	.926

The differences in patient tumor location, gender, and age between the two groups were assessed using χ^2 test, χ^2 test, and one-way ANOVA, respectively.

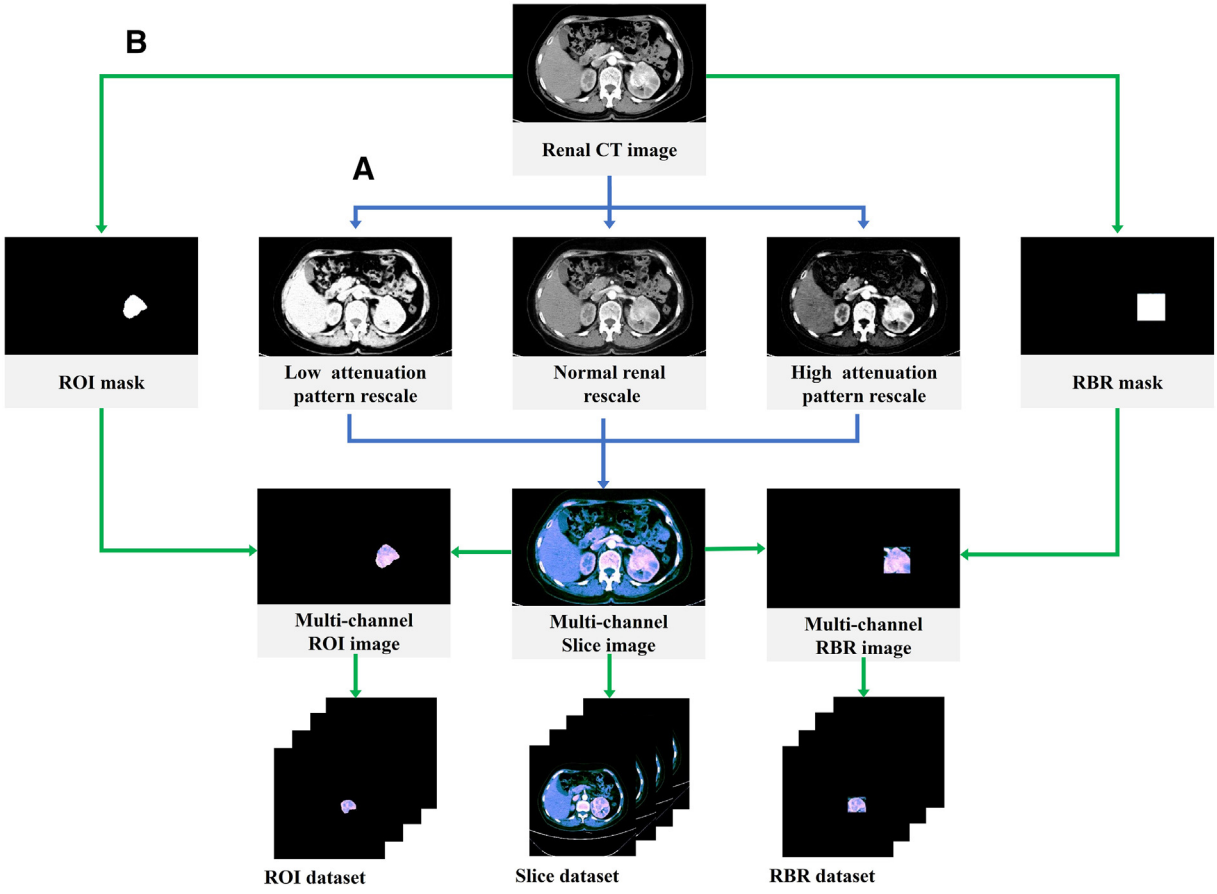

Figure 1. Flowchart of image preprocessing and datasets setting. A (blue lines) was the flow of image preprocessing. An example of renal/high-attenuation/low-attenuation CT windowing for an axis renal CT slice. We encode the renal/high-attenuation/low-attenuation CT windowing into red/green/blue channels. B (green lines) was the flow of datasets setting. The Slice dataset was made up of axial multichannel renal CT slices. The ROI mask was drawn manually by two experienced radiologists on each image of the Slice dataset. The RBR mask was generated from the bounding box of ROI mask's contour. The ROI/RBR dataset consisted of ROI/RBR images which were gotten from slice images and corresponding ROI/RBR masks.

Datasets Setting

In this study, we created three datasets: Slice, region of interest (ROI), and rectangular box region (RBR). The Slice dataset was composed of axial CT images selected based on the maximum lesion diameter and optimal representation of the largest lesion area. The ROI dataset consisted of region-of-interest images manually indicated by two experienced radiologists on each image of the Slice dataset. The RBR dataset consisted of rectangular images generated from the bounding box of tumor's contour in each image of ROI dataset. Some other researchers have also selected the bounding box approach to confirm tumor areas [19,27]. The flow to create these datasets is shown in Figure 1*B*.

Establishment of Model

Image-Level Model. To achieve classification of benign and malignant renal tumors from CT images, we set two nodes (benign/malignant) in the softmax layer of Inception V3. The other structures of the model remained the same and were initialized by the weights trained on ImageNet. To explore the effect of freezing different layers during transfer learning, we chose the mixed0, mixed3, mixed6, mixed9, and mixed10 layers as the dividing points. The layers prior to the dividing point were frozen, which meant that the weights of these layers were not updated but others could be trained during iteration.

Patient-Level Model. We established two patient-level models to make use of the 3D data (Figure 2).

(a) Model one (FC model): For this model, the feature vectors (1024 × 1 × N) were extracted from all the tumor images (N images) of a given patient using the optimal image-level model. Then, we merged them into a one-dimensional vector (1024 × 1) to form the input tensor of the patient-level model with Max pooling layer. This layer was used to uniform the different image sequence lengths. At the end of the model, we added two FC layers (the first one has 1024 nodes, and the second one has two nodes for benign/malignant) and the softmax activation to achieve diagnosis at the patient level.

(b) Model two [gated recurrent unit (GRU) [28] model]: In the GRU model, the feature vectors (1024 × 1 × N) of each patient were extracted in the same way as described above; however, the feature vectors should be concatenated into a 2D array(1024 × N) as the input image sequence. After two GRU layers of this model, softmax activation function was used to achieve the patient-level classification result. This model considered not only all of a given patient's images but also the order of the images.

Training Details

The three datasets were divided into training (80%) and testing dataset (20%), respectively. One hundred fifty-three cases (benign/malignant = 109/44) were assigned to the training dataset. The

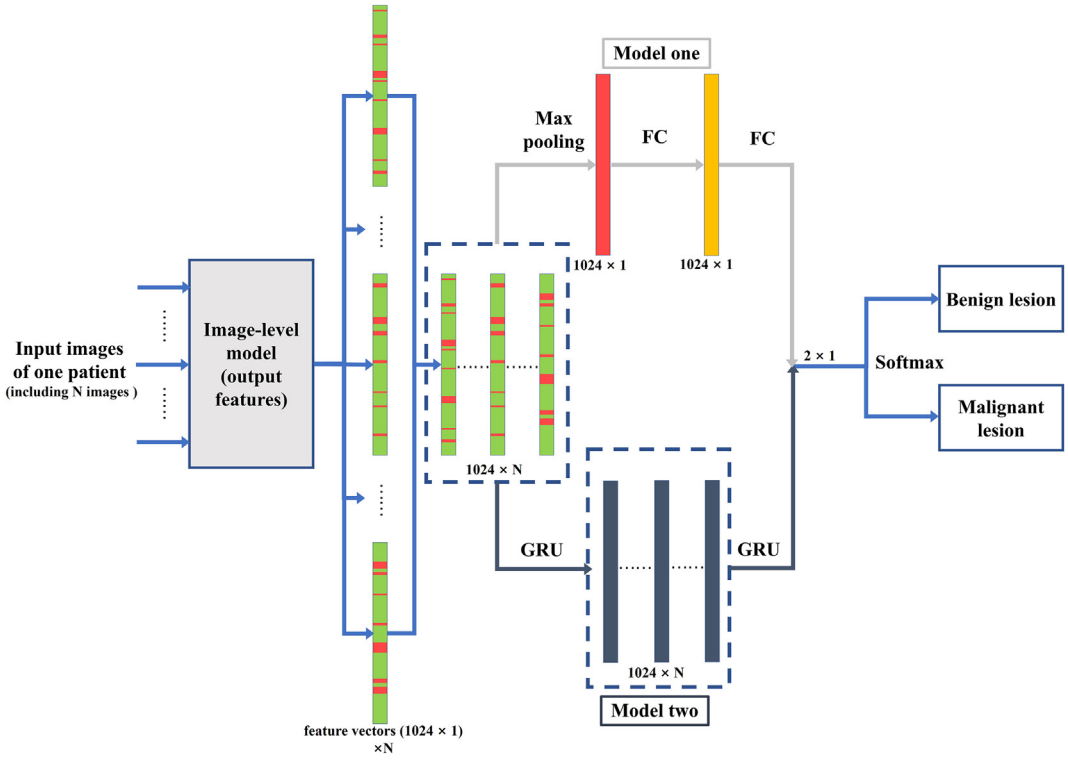

Figure 2. An overview of two patient-level models (model one: FC method, model two: GRU method). Our approach included three steps. The first step was learning intraimage features (1024 × 1) from one patient's N renal tumor image based on optimal image-level model and concatenating features into two-dimensional vectors(1024 × N) as the input tensors of the patient-level model. In the second step, there were differences between the two models. Model 1 added one Max pooling layer to merge the image sequences into a one-dimensional vector and FC layers to learn interimage features, while model 2 only added GRU layers. In the third step, the softmax activation containing two nodes (benign/malignant) was used to realize the diagnosis on the patient-level.

model parameter exploration was performed by five-fold cross-validation on training dataset. The remaining 39 cases (benign/malignant = 25/14) were assigned to the testing dataset.

The training dataset were randomly split into five groups in order to perform a five-fold cross-validation. However, because lesion sizes varied, the number of images differed between patients. If the data were divided according to the number of patients, image numbers between groups would have been prone to serious imbalance. If equal division were performed according to the average number of images instead, the image sequences of some patients might be truncated. It might cause that images of the same patient would appear in both the training and validation set, thus ignoring the individual differences. Therefore, we proposed a new approximate equalization method. Firstly, according to the total number of images, we evaluated the average image number of each group. Then, based on the principle of not truncating a patient's image sequence, when the number of images of the patient who would be divided across two groups exceeded twice the space of the current group, all the images of that patient were assigned to the next group; otherwise, they were added to the end of the current group.

Having fewer pathological samples of rare cases or benign tumors could easily cause data imbalance and cause the recognition result to tend toward the class with more samples [29,30]. This study had the similar risk that the number of the patient with malignant tumors significantly higher than that with benign tumors. We reduced the risk in two ways. On one hand, during the training process, the benign dataset was dynamically oversampled on patient level and image level, respectively, to get a balance between malignant and benign dataset. On the other hand, we selected the optimal threshold by ROC analysis to determine the classification labels.

The server used in this study was equipped with Intel(R) Xeon (R) E5-2650 v4 CPUs @ 2.20 GHz (2 CPUs, 24 cores, 2 threads/core, 128 GB of memory) and an NVIDIA-SMI 384.81. Using the KERAS deep learning framework, based on numerous preliminary experiments, we set the number of iterations to 150 in the image-level model, 100 in the FC model, and 300 in the GRU model. In addition, the learning rate was 0.01, and the momentum was 0.9. At the same time, 2-norm regularization was added to the GRU model.

Evaluation Methods

To eliminate contingencies in the classification results and evaluate the performance of the renal tumor classification model, the results were compared with pathological findings and evaluated by several metrics, including accuracy (ACC), sensitivity (SEN), specificity (SPEC), negative predictive value (NPV), positive predictive value (PPV), Matthews correlation coefficient (MCC), ROC curves, and AUC.

In validation phase, all the metrics were calculated based on the average five-fold cross-validation results. In test phase, the final classification was determined by the majority (≥3) of models with five groups of weights, which were trained from five-fold cross validation. It could make full use of all the weights to obtain higher accurate.

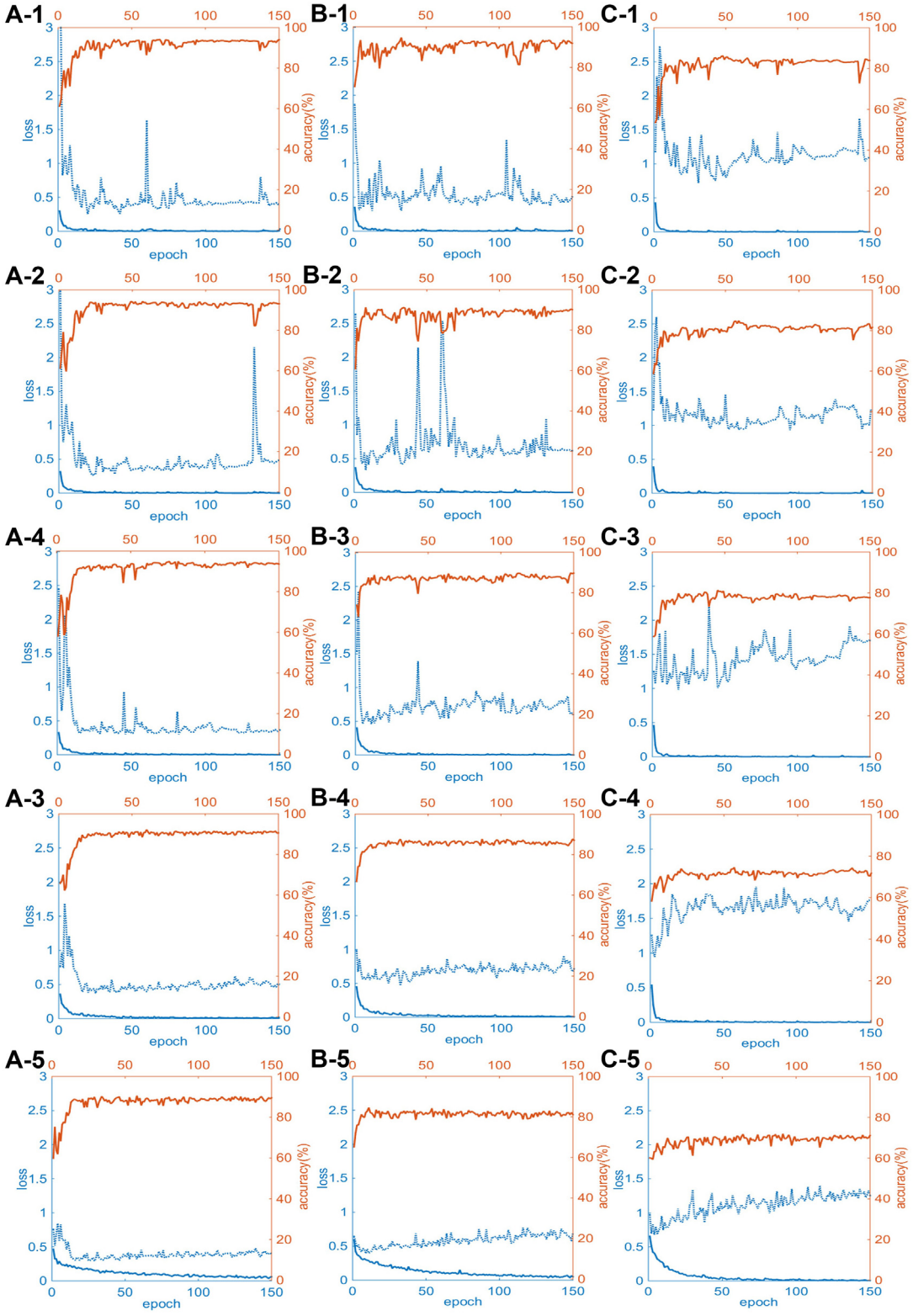

Figure 3. Traces of training loss and validation loss (blue solid and dash lines) and validation accuracy (orange lines). A, B, and C columns were trained on the ROI dataset, RBR dataset, and Slice dataset, respectively. −1, −2, −3, −4, and − 5 denoted freezing the weights of CNN before mixed0, mixed3, mixed6, mixed9, and mixed10 layers, respectively.

Results

Performances with Different Depth Transfer Learnings

As presented in Figure 3, for the models trained on the ROI, RBR, and Slice datasets, the convergence ranges of validation loss were 0.4-0.5, 0.5-0.7, and 1.2-1.7, and the averaged validation accuracies after 100 epochs were 89%-94%, 81%-91%, and 70-83%, respectively. It can also be observed that the ROCs of the ROI and RBR datasets were similar but steeper than that of the Slice dataset in Figure 4, *A-C*. The

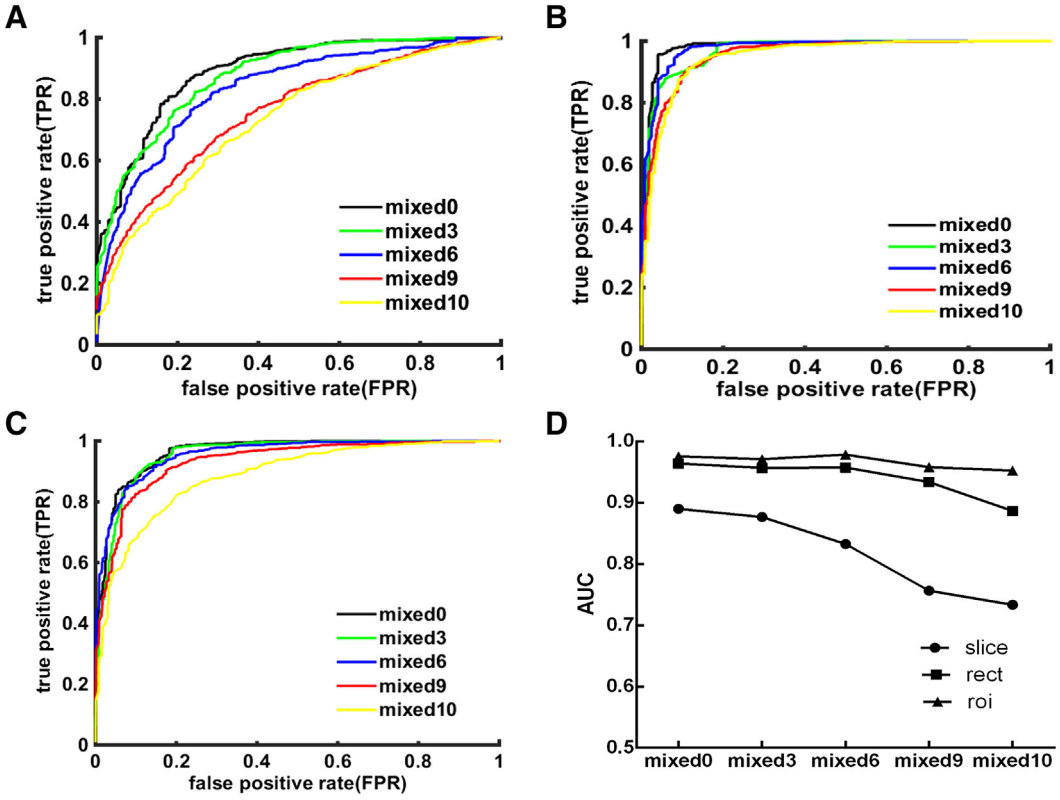

Figure 4. ROC averaged on five-fold cross-validation of the transfer learning with freezing different layers for (A) Slice, (B) ROI, and (C) RBR datasets. (D) The plot of AUC calculated from ROC with freezing different layers for three datasets. ROI and RBR datasets had larger AUCs than Slice dataset with statistical significant (P = .001 and .008, respectively), while the differences between ROI and RBR datasets of AUCs did not reach statistical significant (P = .101).

renal tumor image-level model trained on the ROI dataset was best; its validation loss converged to a smaller value and achieved high validation accuracy. The same model achieved the worst training results on the Slice dataset.

For all the datasets, the validation results of AUCs from freezing the mixed 0, 3, and 6 layers (corresponding to freezing fewer layers) were larger in Figure 4*D*. Other metrics such as SEN, SPEC, PPV, NPV, MCC, and ACC in Figure 5 presented the same trends as the AUCs in Figure 4*D*. As shown in Figure 5, the above indices remained stable and at high levels until the dividing layer exceeded a critical point. Therefore, we selected the image-level model created by freezing the weights before the mixed6, mixed6, and mixed3 layers, respectively, as the optimal transfer learning model for the ROI, RBR, and Slice datasets.

We employed corresponding model parameters and thresholds which yielded better performance in five-fold cross-validation on our testing dataset. The results were listed in Table 3. The selected image-level model obtained good performance on these datasets, particularly on ROI dataset (ACC = 97%).

Comparisons of Patient-Level Models

The two patient-level models based on the selected image-level model and 3D contexts were trained with the iteration stopping criteria (FC model: 100 iterations, GRU model: 300 iterations), which were decided by monitoring the average AUC on validation dataset. The evaluation of the five-fold cross-validation on three datasets was listed in Table 2, which indicated that the training of image-level and two patient-level models on these datasets did not suffer from overfitting.

In the test phrase, we employed the five-fold model parameters and thresholds, and the final results were determined by majority vote of them. The evaluation of test results was depicted in Table 3. The classification performance improved more obviously on the RBR and Slice datasets than on the ROI dataset. On the RBR dataset, the improvements of the two patient-level models were similar (ACC and MCC increased by 2% and 4%) between the two patient-level models compared with the image-level model (Table 3). However, on the Slice dataset, the FC model (ACC increases of 5%) achieved better performance than the GRU model (ACC increases of 3%) (Table 3).

In addition, we did a patient-level classification test with small (maximum diameter <4 cm) and large masses on our testing dataset (large:small = 23:16), respectively. In our training dataset, the number of the larger renal masses was bigger than that of the small renal masses (98 vs. 55). Therefore, we performed the data balancing during training process. All of the patient models had the same accuracy for small (0.94) and large (0.96) renal masses, respectively, regardless of the models consisting of FC or GRU layers, or trained with the ROI or RBR datasets. It meant that our models had almost the same performance to the different size masses generally.

Training Time

As listed in Table 4, the running times of the transfer learning models for renal tumor classification were not significantly different

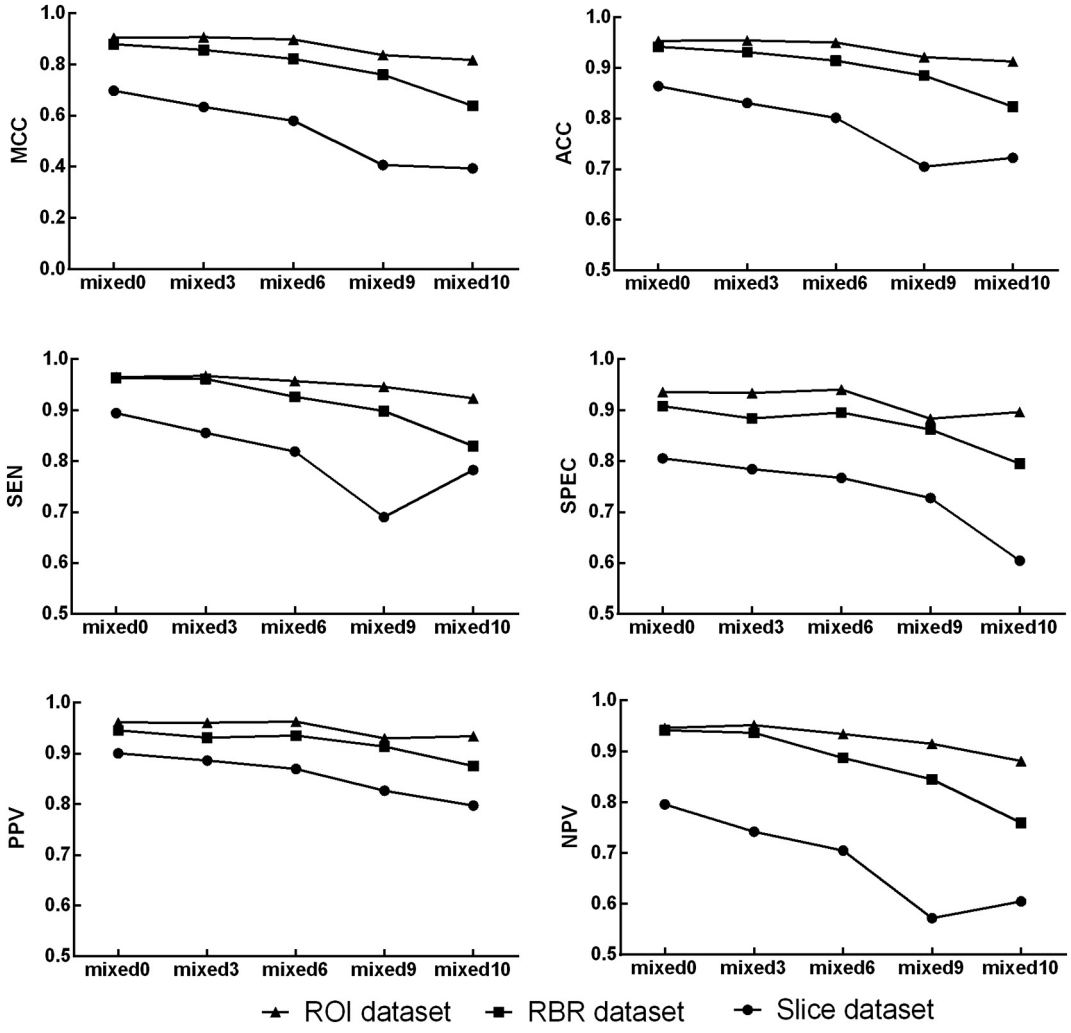

Figure 5. The plot of SEN, SPEC, PPV, NPV, MCC, and ACC with freezing different layers for three datasets.

between the three datasets. However, it took more time to train the tumor classification CNN, which included more trainable layers.

Discussion

In this study, we explored the effects of freezing different numbers of layers during transfer learning and proposed two patient-level models to differentiate benign from malignant renal tumors. Previous studies have demonstrated that renal tumors in CT images can be differentiated using texture analysis. Feng et al. [8] established a support vector machine classifier based on texture features, which were low throughput and predefined by radiologists' expert knowledge, to different small renal tumors, achieving an accuracy

Table 2. Validation Result on These Models Trained by ROI, RBR, and Slice Dataset with Weight of CNN Before Mixed6, Mixed6, and Mixed3 Layers Were Fixed, Respectively

	AUC	ACC	SEN	SPEC	MCC
ROI dataset					
Image-level model	0.98 ± 0.01	0.95 ± 0.01	0.96 ± 0.03	0.94 ± 0.04	0.90 ± 0.02
FC model	0.97 ± 0.02	0.96 ± 0.03	0.96 ± 0.05	0.96 ± 0.06	0.91 ± 0.06
GRU model	0.95 ± 0.04	0.95 ± 0.02	0.96 ± 0.04	0.91 ± 0.09	0.87 ± 0.05
RBR dataset					
Image-level model	0.96 ± 0.04	0.91 ± 0.03	0.93 ± 0.04	0.90 ± 0.09	0.82 ± 0.06
FC model	0.97 ± 0.04	0.95 ± 0.03	0.96 ± 0.04	0.91 ± 0.09	0.88 ± 0.06
GRU model	0.97 ± 0.04	0.97 ± 0.02	0.99 ± 0.02	0.91 ± 0.09	0.92 ± 0.05
Slice dataset					
Image-level model	0.87 ± 0.09	0.83 ± 0.08	0.86 ± 0.07	0.78 ± 0.13	0.63 ± 0.17
FC model	0.89 ± 0.10	0.86 ± 0.10	0.86 ± 0.13	0.86 ± 0.05	0.70 ± 0.18
GRU model	0.81 ± 0.19	0.86 ± 0.09	0.87 ± 0.13	0.81 ± 0.11	0.70 ± 0.14

Image-level model was trained by RBR dataset with weight of CNN before mixed6 layer was frozen. FC model and GRU model were both based on the image-level model. The former was made up of fully connected layers, while the latter consisted of gated recurrent unit layers. Shown in mean ± SD form.

Table 3. Test Result on These Models Trained by ROI, RBR, and Slice Dataset with Weight of CNN Before Mixed6, Mixed6, and Mixed3 Layers Were Fixed, Respectively

	ACC	SEN	SPEC	PPV	NPV	MCC
ROI dataset						
Image-level model	0.97	0.95	0.97	0.95	0.97	0.93
FC model	0.95	0.93	0.86	0.93	1.00	0.89
GRU model	0.95	0.96	0.93	0.96	0.93	0.89
RBR dataset						
Image-level model	0.93	0.87	0.91	0.87	0.97	0.85
FC model	0.95	0.93	0.86	0.93	1.00	0.89
GRU model	0.95	0.93	0.86	0.93	1.00	0.89
Slice dataset						
Image-level model	0.69	0.61	0.42	0.61	0.93	0.45
FC model	0.74	0.76	0.50	0.76	0.70	0.42
GRU model	0.72	0.73	0.43	0.73	0.67	0.35

Image-level model was trained by RBR dataset with weight of CNN before mixed6 layer was frozen. FC model and GRU model were both based on the image-level model. The former was made up of fully connected layers, while the latter consisted of gated recurrent unit layers.

Table 4. Training Time of Different Transfer Learning Models on Three Datasets (min)

	Mixed0	Mixed3	Mixed6	Mixed9	Mixed10
ROI dataset	97.30	83.10	67.80	58.50	56.00
RBR dataset	94.20	80.20	66.30	54.60	51.60
Slice dataset	95.40	80.30	65.00	55.00	54.80

Mixed0, mixed3, mixed6, mixed9, and mixed10 layers were dividing points and represented five different transfer learning models. In these models, the layers before dividing point were frozen, but others could be trained. ROI dataset consisted of region-of-interest images. RBR dataset consisted of rectangular images which were generated from the bounding box of lesion's contour. Slice dataset was made up of CT cross-sectional images containing target area.

of 93.9%. In contrast, our image-level model automatically extracted high-throughput features, avoiding the complicated feature extraction process and obtaining higher accuracy (97% in the model trained on the ROI dataset when the CNN weights prior to mixed6 layer were fixed). CNNs have demonstrated strong performances in medical fields [31], but few applications have focused on renal CT images. Recently, Lee et al. [32] used the deep learning method AlexNet [33] to classify AMLs and renal cell carcinoma, achieving an accuracy of 76.6%. They confirmed that deep features outperformed handcrafted features. However, their method ignored the 3D context, which restricted the recognition performance for renal tumors. Our patient-level models made better use of 3D data, thus improving the accuracy.

Our image-level models for renal tumor classification were separately trained on three different datasets. Both of the average ROC/AUC plots (Figure 4) of validation and metrics on testing dataset (Table 3) demonstrate that the classification ability of the model trained on the Slice dataset was the worst. This was not consistent with a previous study, which reported that chest radiograph or mammogram classification models trained with unclipped images can also achieve good results [17,25]. This discrepancy may have occurred because the abdominal CT Slice images contained multiple organs (e.g., liver, stomach and renal) with weak contrast [34]. The best classification performance was achieved when the models were trained on the ROI dataset because this type of preprocessing eliminates interference from other organs or lesions to the greatest extent. However, drawing the ROI manually involved considerable manual labor, and it was easy to lose details around the boundary regions. Compared with the ROI approach, generating the RBR dataset was much simpler.

Some researchers have demonstrated that classification of small sample size medical image data can be achieved by transfer learning [15,16,29]. To improve the transfer learning performance, we experimented with varying the number of trainable layers. As Figure 4 showed, for the models trained on the ROI and RBR datasets, the classification ability for renal tumors remained at similarly high levels but declined significantly when the weights of more layers were frozen. This phenomenon illustrated that common transfer learning methods (which trained only the last FC layers) [15,16] was not an optimal approach in our study. The cause might be that features extracted directly from the pretrained model were unsuitable and insufficient for renal tumor classification. Considering the validation loss, accuracy (Figure 3), and training time (Table 4), we concluded that the image-level models trained on the ROI and RBR dataset were most appropriate for discriminating between benign and malignant renal tumors when with the weights were frozen before the mixed6 layer. Meanwhile, for the Slice dataset, the optimal classification results were obtained by freezing the weights before the mixed3 layer. In the test, the selected image-level model could obtain state-of-the-art performance on the ROI and RBR dataset (ACC: 97%, 93%; MCC: 93%, 85%).

In clinical diagnoses, an experienced radiologist usually observes and detects tumors based on many slices along the *z*-axis. To make full use of 3D CT images, we established two patient-level models based on the optimal image-level model. By fusing the learning of intraslice and interslice features, the detection performance for renal tumors has obviously been improved except for the model trained on the ROI dataset. These results demonstrate the effectiveness of our patient-level model. Especially, it was inspiring that the performance of our model achieved substantial growth on the RBR and Slice dataset, particularly on RBR dataset approximately caught up with the model trained on the ROI dataset (Tables 2 and 3). The GRU layers could use all the features from the entire series of one patient's images while referring to the order of image sequence [35]. The FC layers also could use all the features after unifying the different image sequence lengths with Max pooling layer. Our patient-level models considered the 3D contexts efficiently and should be improved as greater numbers of 3D images are accumulated.

Our study had several limitations related to the specific application of renal tumor classification as well as to general aspects of deep learning. First, our models could not detect specific subtypes of lesions that radiological experts might diagnose from images, such as clear-cell RCC and papillary RC. In future work, we will collect more data concerning the different subtypes and further investigate fine-grained recognition [36] of renal tumors. Second, the number of renal cases used here was relatively small, which easily caused data distribution imbalance. In our datasets, the number of the low–nuclear grade tumors was significantly bigger than that of the high–nuclear grade tumors. Therefore, although all of them could be classified correctly, it was hard to conclude whether our models had different performance for various nuclear grade tumors because of the small test dataset (low vs. high = 22 vs. 3). Three locations (exophytic, mesophytic, and endophytic) of tumors also had the same problem. We can use data balance and augmentation to reduce these risks, but the amount of data is still the essential reason. In the future, larger-scale and multicenter samples will be acquired to improve the generalization ability of the CNN model for renal tumor classification. Finally, we found that the model trained on the RBR dataset obtained similar on the ROI datasets. However, our RBR dataset was based on the ROI dataset, which was generated by manual drawing. Thus, we plan to perform more work to explore automatic RBR recognition by regions with CNN features (R-CNN) [37].

In conclusion, CNN transfer learning can be used to classify benign and malignant renal tumors from CT images. On our datasets, training layers of approximately half the initially trained model performed better than previous transfer-learning studies in which only the last layer was trained. In addition, our patient-level models notably improved the classification accuracy. Almost all the patients with malignant renal tumors were recognized. Such a recognition ability could eliminate the need for patients identified as benign to undergo invasive procedures. We believe that as the available datasets expand and models are further optimized, CNNs will be able to support clinicians and reduce human errors.

References

[1] Laguna MP, Algaba F, Cadeddu J, Clayman R, Gill I, Gueglio G, Hohenfellner M, Joyce A, Landman J, and Lee B (2014). Current patterns of presentation and treatment of renal masses: a clinical research office of the endourological society prospective study. *J Endourol* **28**, 861–870.

[2] Ljungberg B, Bensalah K, Canfield S, Dabestani S, Hofmann F, Hora M, Kuczyk MA, Lam T, Marconi L, and Merseburger AS (2015). EAU guidelines on renal cell carcinoma: 2014 update. *Eur Urol* **67**, 913–992.

[3] Murphy AM, Buck AM, Benson MC, and Mckiernan JM (2009). Increasing detection rate of benign renal tumors: evaluation of factors predicting for benign tumor histologic features during past two decades. *Urology* **73**, 1297–1298.

[4] Ball MW, Bezerra SM, Gorin MA, Cowan M, Pavlovich CP, Pierorazio PM, Netto GJ, and Allaf ME (2015). Grade heterogeneity in small renal masses: potential implications for renal mass biopsy. *J Urol* **193**, 36–40.

[5] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, Zegers CML, Gillies R, Boellard R, and Dekker A, et al (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* **48**, 441–446.

[6] Hodgdon T, McInnes MDF, Schieda N, Flood TA, Lamb L, and Thornhill RE (2015). Can quantitative CT texture analysis be used to differentiate fat-poor renal angiomyolipoma from renal cell carcinoma on unenhanced CT images? *Radiology* **276**, 787–796.

[7] Raman SP, Chen YF, Schroeder JL, Huang P, and Fishman EK (2014). CT texture analysis of renal masses: pilot study using random forest classification for prediction of pathology. *Acad Radiol* **21**, 1587–1596.

[8] Feng Z, Rong P, Cao P, Zhou Q, Zhu W, Yan Z, Liu Q, and Wang W (2018). Machine learning-based quantitative texture analysis of CT images of small renal masses: differentiation of angiomyolipoma without visible fat from renal cell carcinoma. *Eur Radiol* **28**, 1625–1633.

[9] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, and Rietveld D, et al (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* , 4644. https://doi.org/10.1038/Ncomms5644.

[10] Karpathy A, Toderici G, Shetty S, Leung T, and Sukthankar R (2014). FF Li, Large-scale video classification with convolutional neural networks. CVPR; 2014. p. 1725–1732.

[11] Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, and Darrell T (2013). DeCAF: a deep convolutional activation feature for generic visual recognition. ICML, 32; 2013. p. 647–655.

[12] Esteva A, Kuprel B, Novoa R, Ko J, Swetter S, Blau H, and Thrun S (2017). Corrigendum: dermatologist-level classification of skin cancer with deep neural networks. *Nature* **546**, 686.

[13] Arevalo J, Gonzalez FA, Ramos-Pollan R, Oliveira JL, and Guevara Lopez MA (2015). Convolutional neural networks for mammography mass lesion classification. EMBC, 797. , ; 2015.

[14] Pan SJ and Yang QA (2010). A survey on transfer learning. *IEEE Trans Knowl Data Eng* **22**, 1345–1359.

[15] Kim DH and MacKinnon T (2018). Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol* **73**, 439–445.

[16] Huynh BQ, Hui L, and Giger ML (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J Med Imaging* **3**, 034501.

[17] Cicero M, Bilbily A, Dowdell T, Gray B, Perampaladas K, and Barfett J (2017). Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Invest Radiol* **52**, 281–287.

[18] Ciresan DC, Giusti A, Gambardella LM, and Schmidhuber J (2013). Mitosis detection in breast cancer histology images with deep neural networks. *Med Image Comput Comput Assist Interv* **16**, 411–418.

[19] Dou Q, Chen H, LQ Yu, Zhao L, Qin J, Wang DF, Mok VCT, Shi L, and Heng PA (2016). Automatic detection of cerebral microbleeds from mr images via 3D convolutional neural networks. *IEEE Trans Med Imaging* **35**, 1182–1195.

[20] Frederick LGMD, David LPMD, Irvin DFMD, Fritz C.T.R. RHIT AG, Charles MBMD, Daniel GHMD, and Monica MMD (2010). AJCC Cancer Staging Manual. Springer; 2010 .

[21] Fuhrman SA, Lasky LC, and Limas C (1982). Prognostic significance of morphologic parameters in renal cell carcinoma. *Am J Surg Pathol* **6**, 655–663.

[22] Gomes FV, Matos AP, Palas J, Mascarenhas V, Herédia V, Duarte S, and Ramalho M (2015). Renal cell carcinoma subtype differentiation using single-phase corticomedullary contrast-enhanced CT. *Clin Imaging* **39**, 273–277.

[23] Yan L, Liu Z, Wang G, Huang Y, Liu Y, Yu Y, and Liang C (2015). Angiomyolipoma with minimal fat: differentiation from clear cell renal cell carcinoma and papillary renal cell carcinoma by texture analysis on CT images. *Acad Radiol* **22**, 1115.

[24] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, and Wojna Z (2016). Rethinking the Inception Architecture for Computer Vision. ICVPR; 2016. p. 2818–2826.

[25] Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H, and Ieee (2015). Chest Pathology Detection Using Deep Learning with Non-Medical Training. ISBI; 2015. p. 294–297.

[26] Gao M, Bagci U, L Lu, A Wu, Buty M, Shin HC, Roth H, and Papadakis GZ (2018). A Depeursinge, RM Summers, Holistic Classification of CT Attenuation Patterns for Interstitial Lung Diseases via Deep Convolutional Neural Networks. CMBBE, 6; 2018. p. 1–6.

[27] Liu J, Wang D, L Lu, Wei Z, Kim L, Turkbey EB, Sahiner B, Petrick N, and Summers RM (2017). Detection and diagnosis of colitis on computed tomography using deep convolutional neural networks. *Med Phys* **44**.

[28] Jozefowicz R, Zaremba W, and Sutskever I (2015). An Empirical Exploration of Recurrent Network Architectures. PMLR, 37; 2015. p. 2342–2350.

[29] Ravishankar H, Sudhakar P, Venkataramani R, Thiruvenkadam S, Annangi P, Babu N, and Vaidya V (2016). Understanding the Mechanisms of Deep Transfer Learning for Medical Images. Springer Int Publishing Ag; 2016 188–196.

[30] Raj V, Magg S, and Wermter S (2016). Towards effective classification of imbalanced data with convolutional neural networks. *Lect Notes Artif Intell* **9896**, 150–162.

[31] Esteva A, Kuprel B, RA Novoa J Ko, Swetter SM, Blau HM, and Thrun S (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118.

[32] Lee H, Hong H, Kim J, and Jung DC (2018). Deep feature classification of angiomyolipoma without visible fat and renal cell carcinoma in abdominal contrast-enhanced CT images with texture image patches and hand-crafted feature concatenation. *Med Phys* **45**, 1550–1561.

[33] He KM, Zhang XY, Ren SQ, Sun J, and Ieee (2016). Deep Residual Learning for Image Recognition2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR; 2016. p. 770–778.

[34] Shimizu A, Ohno R, Ikegami T, Kobatake H, Nawano S, and Smutek D (2007). Segmentation of multiple organs in non-contrast 3D abdominal CT images. *Int J Comput Assist Radiol Surg* **2**, 135–142.

[35] Chung J, Gulcehre C, Cho K, and Bengio Y (2015). Gated feedback recurrent neural networks. *Comput Sci* , 2067–2075.

[36] Jia D, Krause J, and Li FF (2013). Fine-Grained Crowdsourcing for Fine-Grained Recognition. CVPR; 2013. p. 580–587.

[37] Girshick R (2015). Fast R-CNN. ICCV; 2015. p. 1440–1448.