Japanese Journal of Radiology (2018) 36:257–272
https://doi.org/10.1007/s11604-018-0726-3

REVIEW

Deep learning with convolutional neural network in radiology

Koichiro Yasaka[1] · Hiroyuki Akai[1] · Akira Kunimatsu[1] · Shigeru Kiryu[2] · Osamu Abe[3]

Received: 28 December 2017 / Accepted: 26 February 2018 / Published online: 1 March 2018

Abstract
Deep learning with a convolutional neural network (CNN) is gaining attention recently for its high performance in image recognition. Images themselves can be utilized in a learning process with this technique, and feature extraction in advance of the learning process is not required. Important features can be automatically learned. Thanks to the development of hardware and software in addition to techniques regarding deep learning, application of this technique to radiological images for predicting clinically useful information, such as the detection and the evaluation of lesions, etc., are beginning to be investigated. This article illustrates basic technical knowledge regarding deep learning with CNNs along the actual course (collecting data, implementing CNNs, and training and testing phases). Pitfalls regarding this technique and how to manage them are also illustrated. We also described some advanced topics of deep learning, results of recent clinical studies, and the future directions of clinical application of deep learning techniques.

Keywords Deep learning · Convolutional neural network · CT · MRI · PET

Introduction

Recently, deep learning is gaining attention as a technique for realizing artificial intelligence. This technique is known to be useful for recognition or categorization of images, speech recognition, and natural language processing [1]. Deep learning is one of the machine learning methods. An artificial neural network is utilized as one of the methods of conventional machine learning techniques, and a deeply stacked artificial neural network is utilized in deep learning. Several types of deep neural networks exist, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). RNNs deal with sequential input data and are commonly utilized for speech and language tasks. For image tasks, CNNs are known to be useful. The CNN has its roots in a neocognitron proposed by Fukushima et al. in [2]. The idea of neocognitron was based on the biological mechanisms of visual recognition at primary visual cortex (V1) of the vertebrate investigated by Hubel and Wiesel [3]. In 2012, the CNN proposed by Krizhevsky et al. highly gathered attention for its high performance in image recognition tasks compared to conventional methods at ImageNet Large Scale Visual Recognition Competition (ILSVRC) [4]. Since then, many models for image recognition tasks have been developed based on deep learning techniques. And recently, application of deep learning methods with CNNs to radiological images is beginning to gain wide attention [5, 6].

This article illustrates pitfalls and some advanced techniques of deep learning with CNNs as well as basic technical knowledge along the actual course. Results of recent clinical studies and the future direction of clinical application of this technique are also described.

From visual assessment to radiomics and deep learning

Radiological images have been interpreted by radiologists. Most evaluations have been performed in subjective and/or qualitative ways (Fig. 1a). With the increase in data size of radiological images, radiomics strategies with which

✉ Koichiro Yasaka
koyasaka@gmail.com

[1] Department of Radiology, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

[2] Department of Radiology, Graduate School of Medical Sciences, International University of Health and Welfare, 4-3 Kozunomori, Narita, Chiba, Japan

[3] Department of Radiology, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

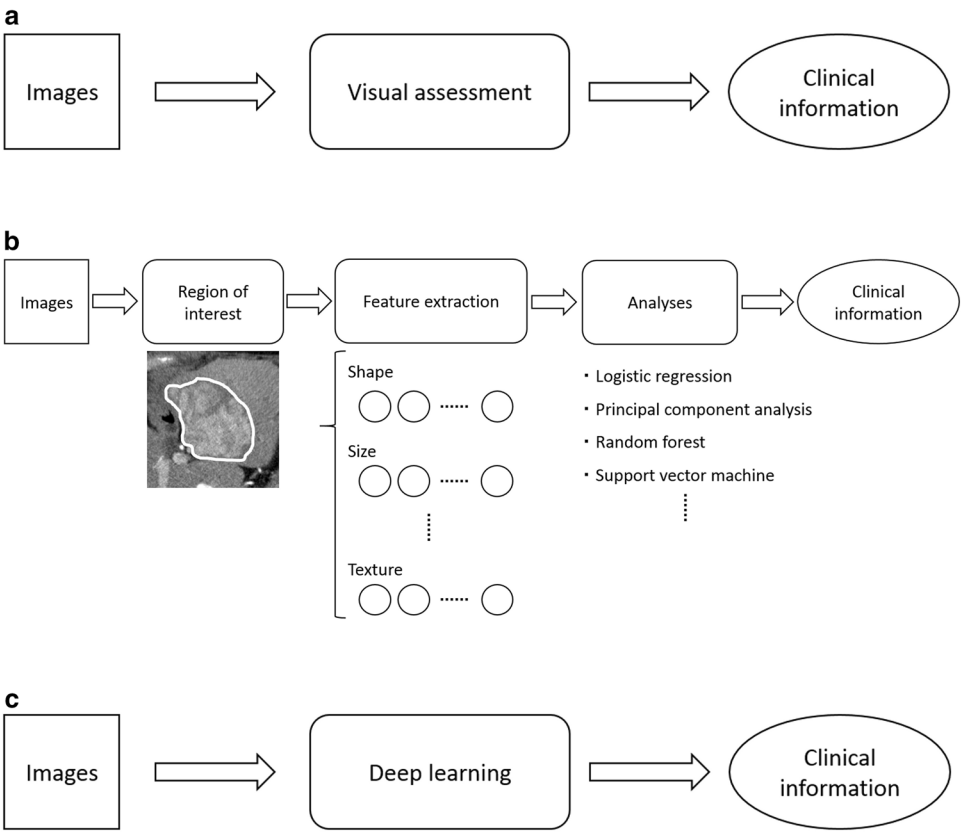

Fig. 1 **a** Radiological images have been visually assessed by radiologists. **b** With radiomics studies, the following are performed: region of interest placement (white freehand line), feature extraction from images, and analyses (selection of parameters and building models) with or without a conventional machine learning process. **c** For deep learning, which is sometimes referred to as an end-to-end strategy, images themselves can be fed to convolutional neural networks

high-throughput extraction of many quantitative parameters from images is performed (Fig. 1b) have been gaining attention [7]. With this technique, regions of interest are placed on images and many features of lesions are extracted as quantitative parameters. Because features of lesions are expressed with continuous variables, this strategy has a potential to capture more detailed features of lesions compared with the visual assessment. Radiomics studies have successfully correlated information extracted from radiological images with clinically important information in estimating subtypes of tumors [8–10], estimating prognosis of patients with tumors [11–15], evaluating the risk of recurrence of tumors [16], and evaluating treatment response of tumors [17].

Features of deep learning with convolutional neural network

Deep learning is one of the machine learning methods, such as support vector machine, random forest, and artificial neural network. These methods allow relating the input data to teaching data without being explicitly programmed. A deeply stacked artificial neural network is utilized in deep learning.

With the deep learning technique, pattern recognition of images can be performed without the human hand once the program was established. Therefore, this technique would also have a potential to perform high-throughput procession of radiological images in correlating the images to the clinical information.

With deep learning, images can be fed to CNNs as is, and important features can be automatically learned (Fig. 1c). Therefore, information extraction from images in advance of the learning process is not necessary with this technique. At shallow layers, simple features such as edges within images are learned. And at deep layers near the output layer, more complex high-order features are known to be learned [1, 18].

Models based on deep learning techniques do not necessarily require placing complex shaped regions of interest on images. Therefore, this technique would be easy to apply to daily clinical practice.

Overfitting

Models obtained from deep learning are known to face an overfitting problem. These complex models try to associate input data with teaching data through a learning process; however, the models also try to fit to the noise included in the training data. Therefore, the trained model does not necessarily show high performance in unexperienced cases as achieved in cases used for training. Therefore, in building deep learning models, several considerations should be

taken into account, which are summarized in Table 1 and described in several parts of this article.

Preparation of input and teaching data

The overview of deep learning process is shown as a scheme in Fig. 2. In advance of deep learning, input data and teaching data are needed to be prepared. As for the sample size of training data, collecting as many data as possible would be preferable in general; because training with small sample size would be related to high risk of overfitting (Table 1).

Input data

Generally, radiological images are grayscale images. Grayscale values for each pixel within images can be described as matrices. For RGB color images, three matrices are used to represent information regarding red, green, and blue. For deep learning, the number of matrices is termed as the channel. For grayscale images, the channel of input data is one and for RGB color-scale images, the channel of input data is three. If combinations of grayscale images are used, the channel of input data can be multiple. Or several types of images can be concatenated as one image. Some researchers used the concatenated images as input data for learning (Fig. 3) [19, 20].

Care should be paid to the data volume of input images. Use of input data with large data volume would be associated with some problems. To deal with such data, computers with large memories would be necessary. Input data with large data volume would also be associated with large CNNs which contain large numbers of parameters to adjust. Training of such large CNNs requires more calculations and longer times. To solve these problems, cropped images from the original image and/or resized images are utilized for many researches.

Table 1 Countermeasures for overfitting problem

Situations	Countermeasures
Preparation of input data	Collecting as many cases as possible
	Augmentation
Convolutional neural network	Dropout
	Batch normalization
Training and testing	Performance evaluation at separate testing cases
	Early stopping
Advanced technique	Transfer learning

Input data: augmentation

For training data, augmentation is sometimes performed (Fig. 4). Image augmentation would be preferable for some reasons. Augmentation of images results in increased numbers of data sets for training; thus it would be beneficial to reduce the risk of the overfitting problem (Table 1). Even if the same patient is examined, images would not be identical between examinations due to slight differences in positioning. Therefore, use of rotated or parallel-shifted images would get the model robust for slight differences in patient positions. Mirrored images are also sometimes utilized. Use of noise-added images would be useful for getting the model to be robust for different degrees of image noise. For

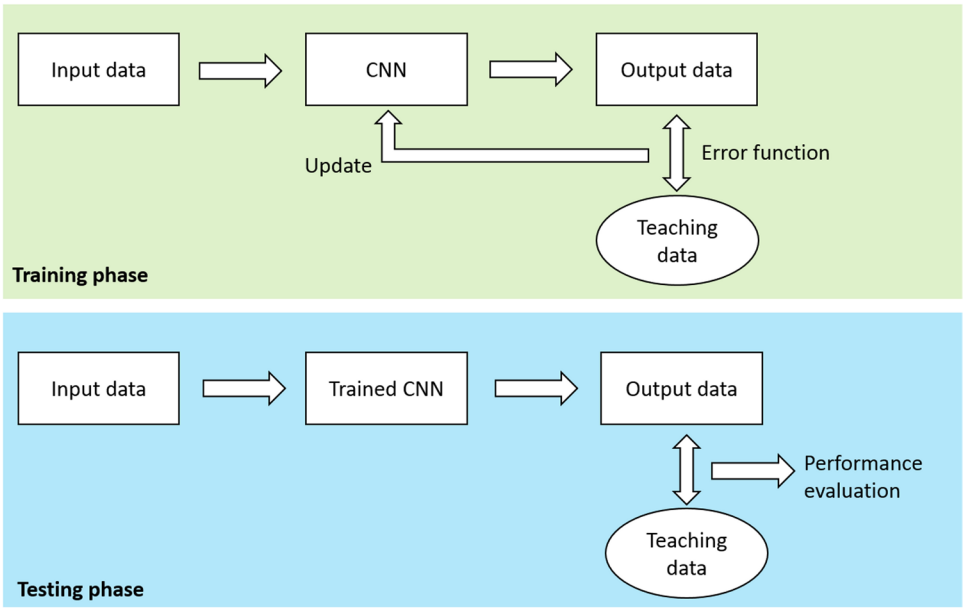

Fig. 2 Overview of typical deep learning process. For a training phase, output data from the convolutional neural network (CNN) are fed to an error function together with teaching data. Parameters within the CNN are adjusted so that the errors would become smaller. For a testing phase, input data, which are separately prepared from training data, are fed to the trained CNN. Performance of the trained CNN is evaluated by using outputted data from the trained CNN and corresponding teaching data

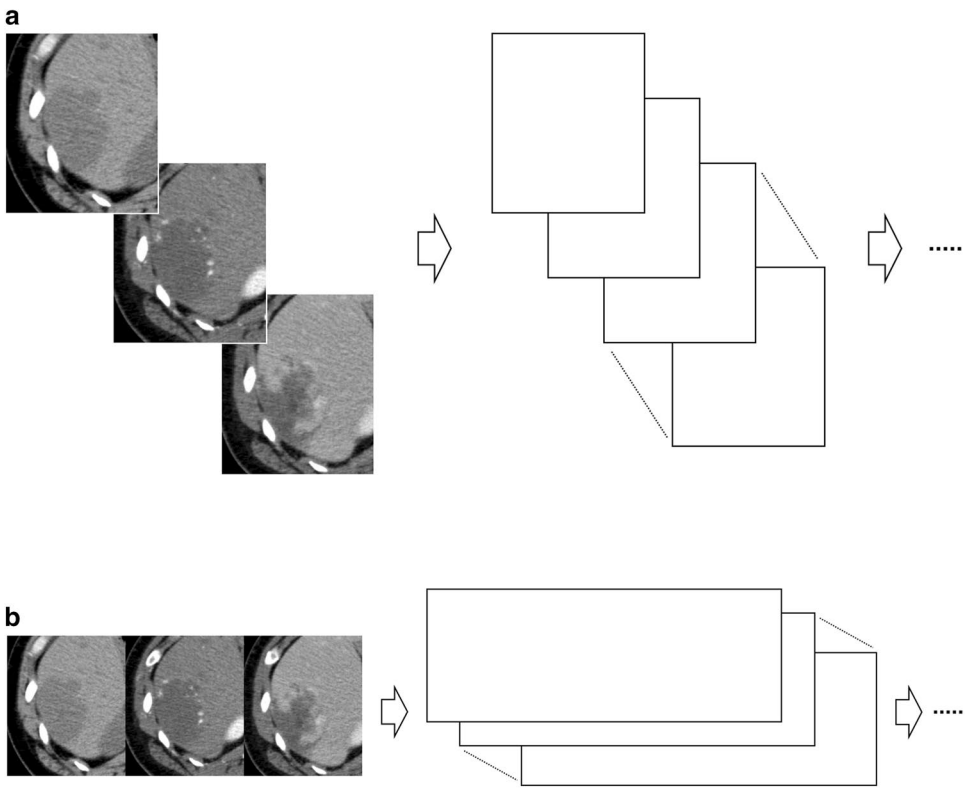

Fig. 3 A combination of input images can be dealt with CNNs in several ways. For example, a combination of unenhanced computed tomography (CT) images, arterial phase CT images, and delayed phase CT images can be fed to CNNs as (**a**) multiple channel images or as (**b**) a concatenated image

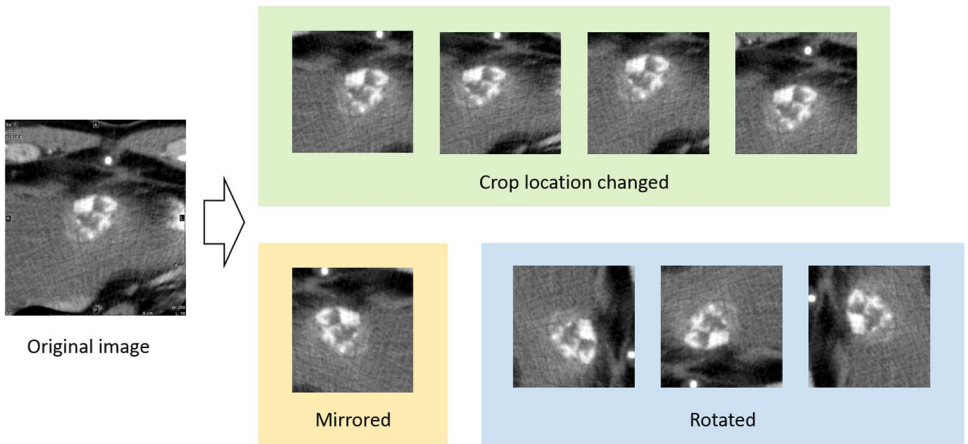

Fig. 4 Image augmentation. For example, from original images, cropped images with different locations, mirrored images, and rotated images are generated

magnetic resonance (MR) images, the pixel values are not absolute values; therefore, use of contrast-changed or brightness-changed images might mitigate this problem.

Teaching data

Most studies regarding clinical application of CNN models based on radiological images are supervised learning. And for supervised learning, teaching data need to be prepared. Blood test data, clinical diagnosis data, pathological diagnosis data, etc., which researchers wish to predict from input data, are used as teaching data. The type of data can be either (a) nominal variables (for example, a lesion is benign or malignant), (b) ordinal variables (for example, breast density is fatty, scattered, heterogeneously dense, or dense), (c) continuous variables (for example, estimated glomerular filtration rate values), or even (d) images (for example, generating noise-reduced images from CT scanned with reduced-dose levels).

Implementation of convolutional neural network

Hardware and software

CNNs include many parameters, and through a training phase, huge amount of calculations are required to adjust them. Development of graphical processing unit (GPU) and general-purpose computing on GPUs (GPGPU) have contributed to recent advancement of deep learning [1]. GPU is suitable for parallel processing and accelerates calculations in a training phase. Compute Unified Device Architecture (CUDA) supported by NVIDIA (Calif, US) is a major platform for GPGPU. Open CL (Khronos; Oregon, US) is also available for this purpose. And to deal with huge amount of data, computers with large random access memories would be preferable.

For deep learning, the programming language of Python (https://www.python.org/) is mainly utilized. Several frameworks for deep learning, such as Caffe (http://caffe.berkeleyvision.org/), Chainer (https://chainer.org/), TensorFlow (https://www.tensorflow.org/), and so on, allow us to write programs with fewer lines.

Overview of convolutional neural network

There are several types of convolutional neural networks. Important components of CNNs are convolutional layers, max pooling layers, and fully connected layers (Fig. 5). The form of input layer should be identical to the input data. The form of output layer should be identical to the teaching data. For example, for multiple classification tasks, the numbers of units in the output layer should be equal to the numbers of categories included in the teaching data. Layers between the input layer and the output layers are called as hidden layers.

Fully connected layer

A fully connected layer is comprised of several units. Output signals from units in a previous layer are weighted and summed for all units (Fig. 6). The weighted and summed signal from a previous layer is also added with a constant referred to as bias, and then is fed to an activation function. In recent articles, rectified linear unit (Relu) function is commonly used (except for the output layer) as the activation function [21], which returns the input value itself when the input values are more than zero or returns zero when the input values are equal to or less than zero (i.e., $y = \max(x, 0)$, where x and y denote the input value and the output value, respectively). For the output layer, an identity function or a softmax function is commonly used for regression tasks and multiple classification tasks, respectively (Fig. 6).

Convolutional layer

While artificial neural networks, which are comprised of only fully connected layers, cannot take spatial information into account (they deal with information of distant pixels and close pixels equally in principle), convolutional layers can treat the input image as is (values of close pixels and distant pixels are processed in different ways). At convolutional layers (Fig. 7), several filters are convoluted to image data from a previous layer. The convolution function is performed repeatedly displacing the location of filters with S pixels. The number of S is referred to as the stride. The values within filters are adjusted through learning processes.

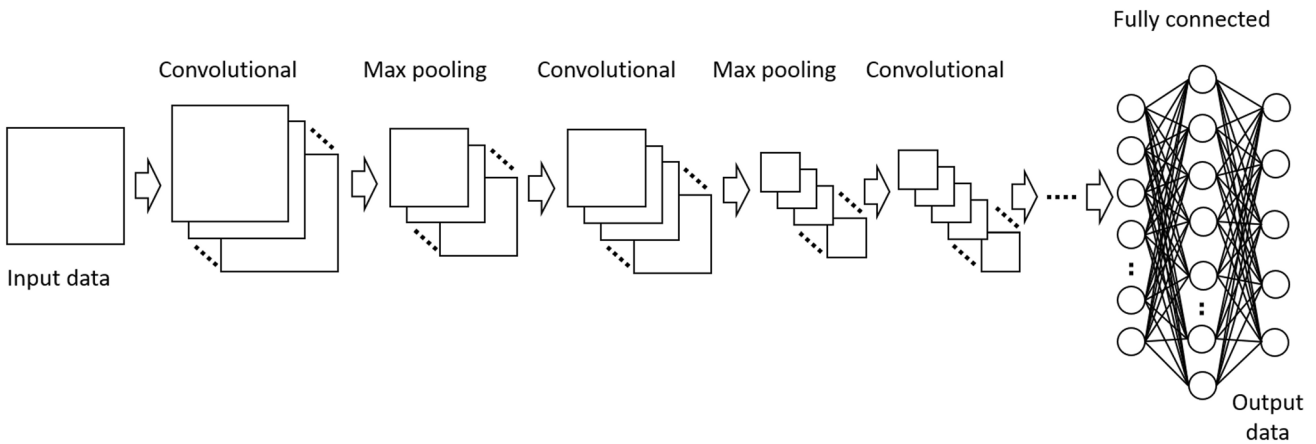

Fig. 5 Overview of a convolutional neural network. Input data are first processed with convolutional layers and max pooling layers. These data are further processed in fully connected layers

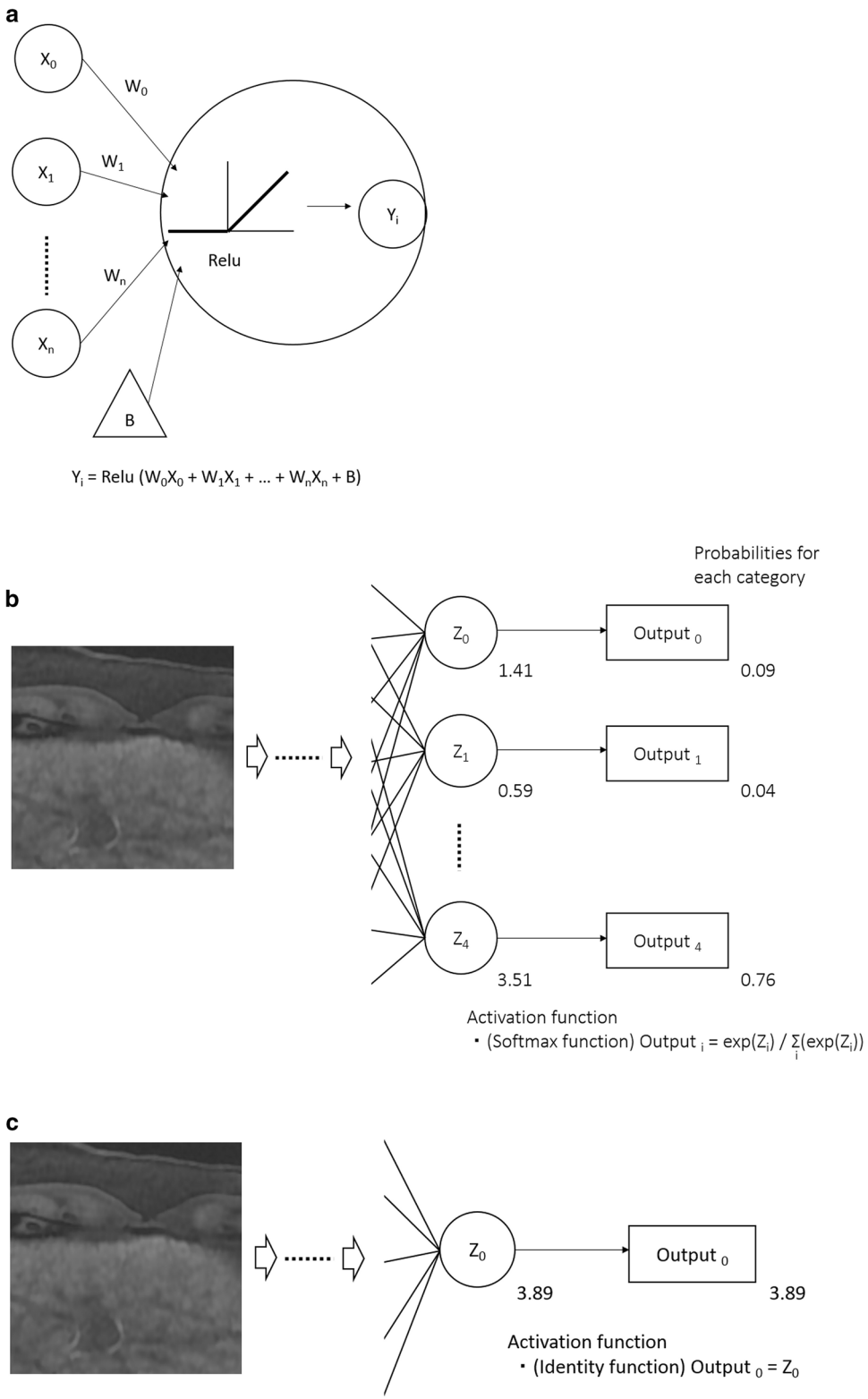

Fig. 6 **a** Output signals from units in a previous layer (X_0, X_1, …, X_n) are weighted (W_0, W_1, …, W_n) and summed, and a constant value referred to bias (B) is added before processed with rectified linear unit (Relu) function. Relu function returns the input value itself when the input values are more than zero or returns zero when the input values are equal to or less than zero [$y = \max(x, 0)$]. **b**, **c** For example, magnetic resonance image of liver surface is fed to CNNs which is trained to predict liver fibrosis stage. **b** This model was developed to return probabilities for each fibrotic stage. The values for each unit in the output layer are 1.41 for Z_0 (correspond to score for $F0$ [normal liver]), 0.59 for Z_1 (correspond to $F1$),…, and 3.51 for Z_4 (correspond to $F4$ [cirrhosis]). By using softmax function, which is commonly used for multiple classification tasks, these values are processed so that they are within the range of 0 and 1 and the sum of them equals to 1. Therefore, output values can be regarded as probabilities for each category. In this case, the model judged the liver is most likely to be at the stage of $F4$ with probability of 0.76. **c** This model was developed to return a single value which would be close to fibrotic stage. For regression tasks as in this case, the identity function, which returns a value equals to an input value, is commonly utilized. With appropriately determined cutoff values, this score can be used to predict fibrotic stage

Pooling layer

At the max pooling layer or average pooling layer, maximum or average values within $H \times H$ pixels are returned, respectively (Fig. 7). Max pooling layers are commonly used in recent studies. In general, for max pooling layers, the stride is also set as H. Therefore, longitudinal and

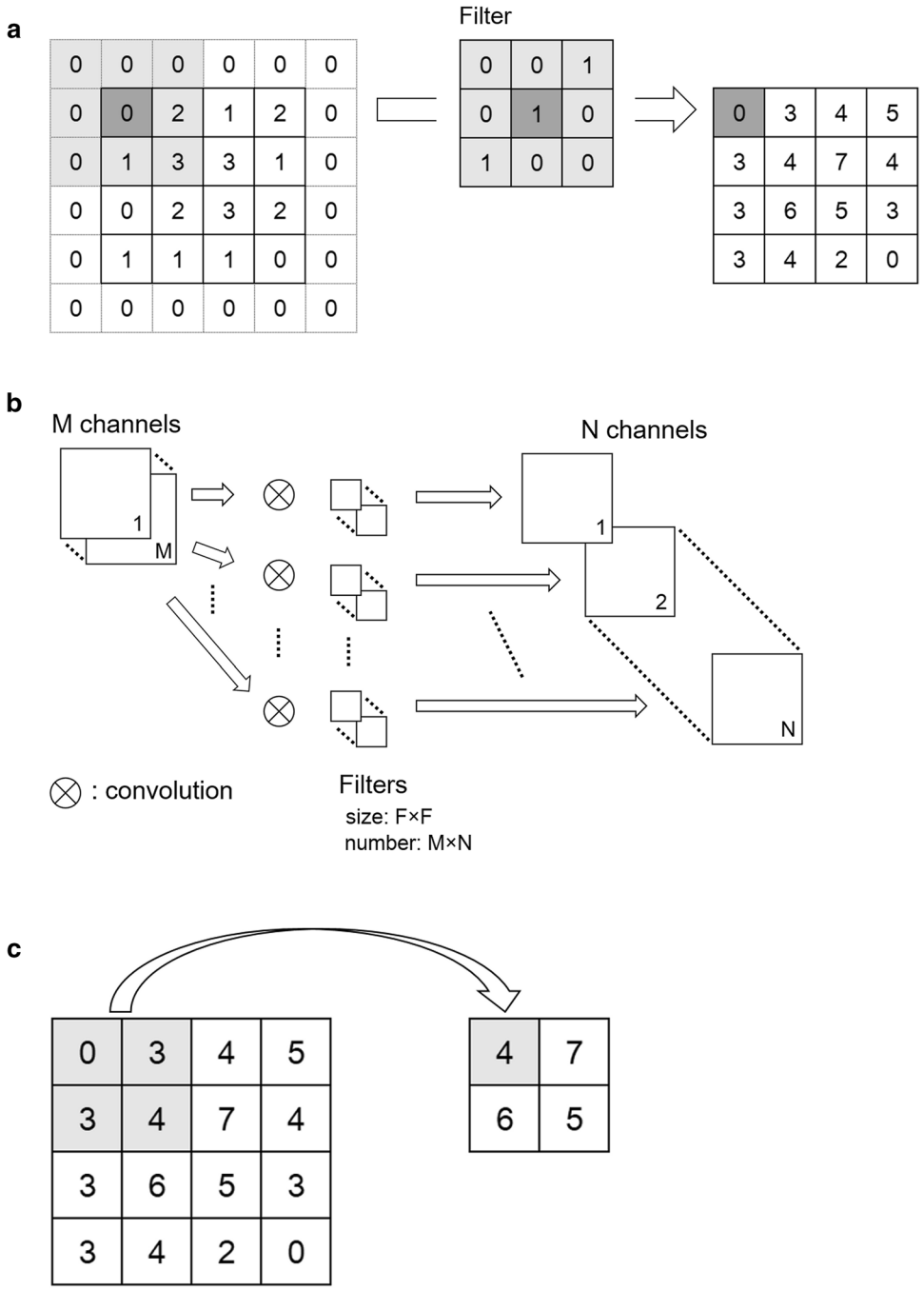

Fig. 7 Schemes of **a**, **b** convolutional layer and **c** max pooling layer. **a** Filters with the size of $F \times F$ (F is set as 3 in this figure) are convoluted to input images. To perform the convolution function for peripheral pixels of input image, values of zero are padded one pixel outside of the image. **b** A combination of M filters is convoluted to M channels of input images from a previous layer to produce an output image. This process is repeated N times to produce N channels of output images. **c** Maximum values within $H \times H$ pixels (H is set as 2 in this figure) are extracted at the max pooling layer

transverse sizes of output images from this layer become $1/H$ of those of input images. Channels of images do not change through this layer. Max pooling function simply returns maximum values; therefore, this function is not adjusted through the training process. As the complex cells within the primary visual cortex are less sensitive for the difference in positions, this layer allows the model to be robust for slight differences in locations of objects within images.

Dropout

With a dropout technique, nodes within CNNs are eliminated (or dropped out) with a certain probability at the training phase (Fig. 8). At the testing phase, all the nodes are utilized. Eliminating units at the training phase prevents the units from adapting too much. This technique is known to be useful for reducing the risk of the overfitting problem (Table 1) [22].

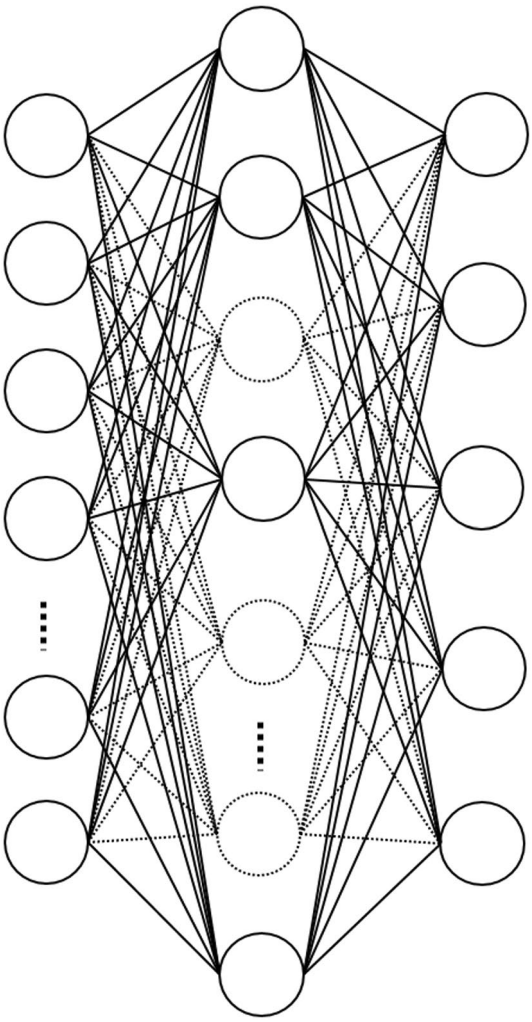

Fig. 8 A scheme of dropout function. Units within the CNNs are eliminated at a certain probability for the training phase with the dropout technique. In this figure, some units at the second layer are eliminated

Batch normalization

With batch normalization, data are normalized (distribution of data are changed so that the mean and variance would become 0 and 1, respectively) in units of the minibatch (a group of data, which is described later). This technique is known to be useful for reducing the risk of the overfitting problem (Table 1) [23]. This would also accelerate the learning process, and according to Ioffe et al. the speed of the learning process with this technique became 17 times faster than that without this technique [23].

Training and testing with convolutional neural networks

Because models obtained through deep learning and conventional machine learning are known to face the overfitting problem, its performance should be evaluated with cases which are not included in the training phase (Table 1). For conventional machine learning, the performance of models is commonly evaluated with cross-validation techniques (leave-one-out or K-fold cross validation). With this method, patient data are split to training data and testing data several times; and in each split, different patients are assigned to test data. For deep learning, because the training process requires huge amount of calculations, this technique is not commonly used for evaluations of model's performance [24]. Instead, a hold-out method, in which testing data are prepared separately from training data, is commonly utilized to evaluate the performance of models obtained with deep learning techniques.

Training

For a training phase, output data from CNNs and teaching data are fed to a function, referred to as an error function. And the errors are backpropagated to CNNs. Parameters within CNNs are adjusted so that errors would become smaller. For regression tasks and multiple classification tasks, mean squared error and softmax cross entropy, respectively, are commonly used as the error function. Optimizers work to adjust parameters within CNNs. There are several types of optimizers, such as stochastic gradient descent, AdaGrad [25], and Adam [26]. In using stochastic gradient descent, learning rate, which regulates how much the parameters within the CNN are changed by one learning process, is set in advance of the learning process. For AdaGrad, the learning rate is adaptively changed throughout the learning process. The learning rate is gradually reduced as the learning process progresses with this optimizer. Adam was proposed by Kingma et al. in 2015. This method was designed to combine the advantages of some optimizers including AdaGrad. In many recent studies, Adam is popularly used; however, which optimizer is suitable for learning differs from tasks to tasks. To understand how these optimizers adjust parameters, explanation with mathematical formula is required, which is beyond our scope. Thanks to several deep learning frameworks, implementation of optimizers can be performed with fewer lines of programs.

The learning processes, in which input data are fed to CNNs and parameters within CNNs are adjusted, are iterated with units of (a) single input datum (online learning),

(b) groups of input data (minibatch learning), or (c) all the input data (batch learning). Because the amount of calculations for one learning process becomes very large, batch learning is unpopular. In recent studies, minibatch learning is popularly performed. When all the training data are used once, it is counted as one epoch has ended. For example, for minibatch learning with the number of data included in each group is *X* and the number of all the training data is *N*, the learning process is iterated with *N*/*X* times for one epoch. With minibatch learning, data are usually shuffled and assigned to different groups for each epoch. Several epochs are performed so that the error would be sufficiently reduced (Fig. 9a). Sometimes, repeating the epochs further does not necessarily result in reduced errors. In such situations, models are trying to fit only to training data (i.e., repeating epochs results in overfitting). The early stopping technique would be one of the strategies to manage this situation (Table 1). With this technique, training of the model is stopped when the error of the model in test data shows a local minimum value (Fig. 9b).

Testing

For testing, output values from the trained CNN are compared with teaching data. And the performance of the model is evaluated with appropriate methods, such as sensitivity, specificity, area under the receiver operating characteristic curve (AUC), Dice similarity coefficient score, etc. For testing, the amount of calculations is not as large as for training.

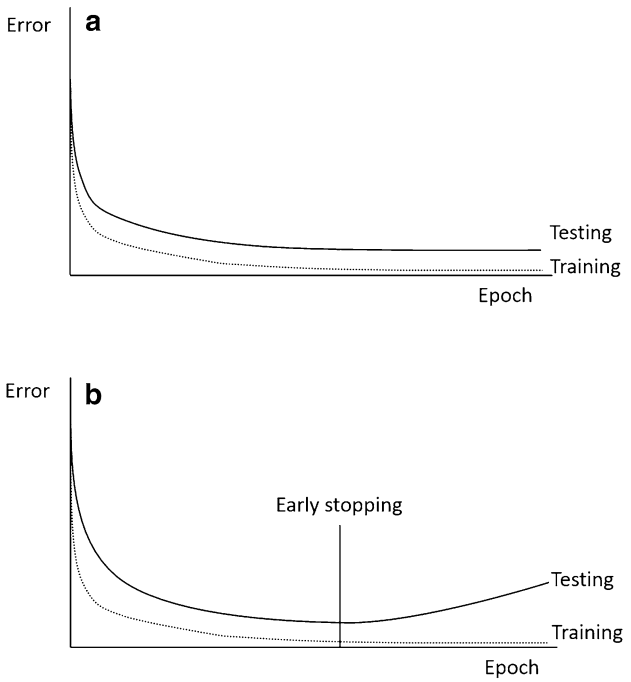

Fig. 9 Learning curves. **a** In this situation, repeating epochs result in decreased errors for the training phase and the testing phase. **b** Sometimes, repeat of epochs would not necessarily result in decrease of errors, due to the overfitting problem. In such a situation, early stopping technique might be useful to mitigate this problem

Advanced topics of deep learning

In this section, we illustrate some advanced topics of deep learning.

Several types of convolutional neural network

Since the high-performance CNN was introduced by Alex Krizhevsky et al. (which is comprised of 5 convolutional layers, 3 max pooling layers, and 3 fully connected layers and is sometimes called as AlexNet) [4], the structure of CNNs has been evolving. GoogLeNet [27] and ResNet [28] are famous CNNs, which gathered attention for their high performance in ILSVRC 2014 and ILSVRC 2015, respectively. GoogLeNet includes inception structures (Fig. 10a), in which input data are processed with different functions and the results are concatenated. ResNet uses skip structures (Fig. 10b) with which input data are added to processed data. Conventionally, use of deeper CNNs was not necessarily associated with high performance. However, skip structures allow deeper CNNs with improved performance. Though it might not necessarily hold true for all tasks, Korfiatis et al. reported that ResNet with 50 layers showed better performance compared with ResNet with 18 layers and ResNet with 34 layers in a certain task [29]. A CNN referred to as Texture CNN, which is useful for capturing texture of images, was also reported [30]. With Texture CNN, energy measures are pooled at the last convolution layer and are connected to a fully connected layer (Fig. 10c).

Use of images as teaching data

For CNN models which predict clinically useful information, the size of data decreases at deeper layers. In other words, data are downsampled within CNNs. On the contrary, for CNN models which use images as teaching data, (a) upsampling of the data is performed at layers near the output layer or (b) downsampling is not even performed. Fully convolutional networks (FCNs) were proposed by Long et al. for semantic segmentation tasks [31]. In this network, fully connected layers are eliminated. This network includes upsampling layers near the output layer. Generally, at deep layers in CNNs, high order features can be captured; however, through processing with pooling layers, spatial information is lost. FCN employs skip connections from shallow downsampling layers to upsampling layers near the output layer in order to overcome this problem. U-net [32] and SegNet [33], which have similar structures to FCN,

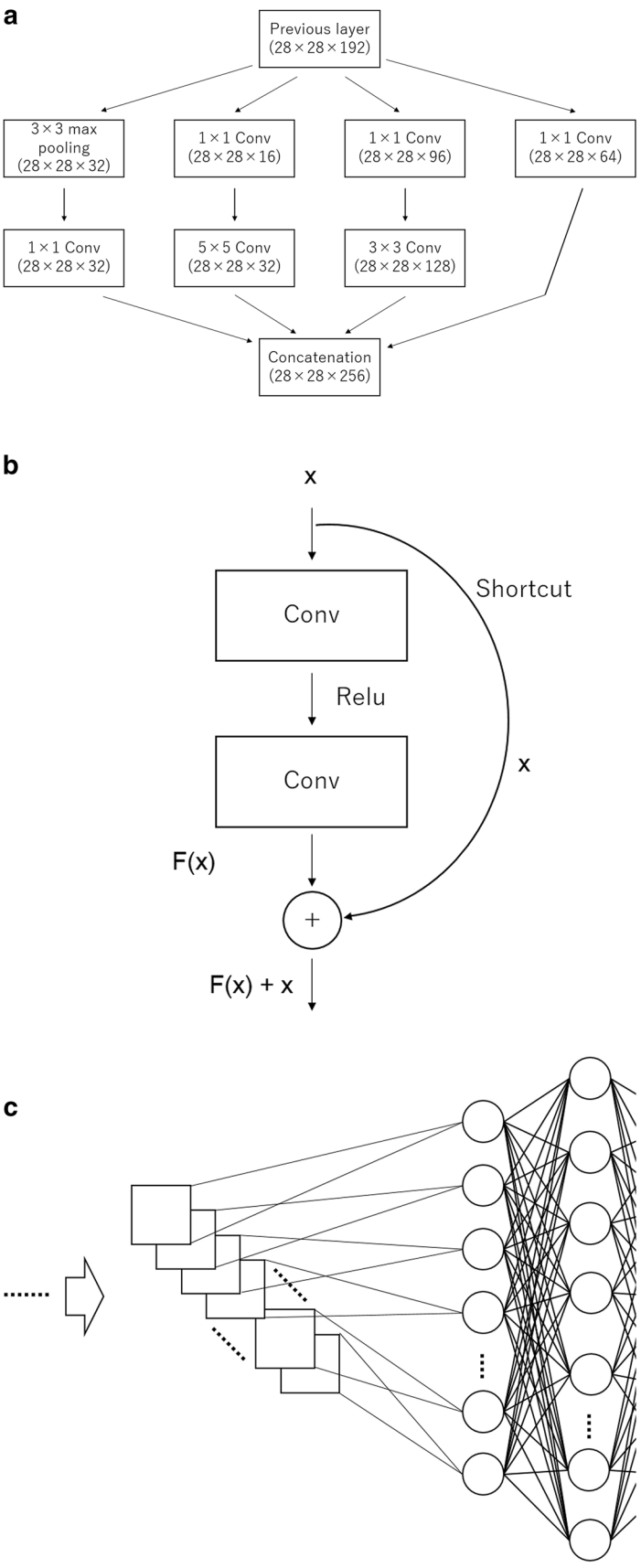

Fig. 10 **a** With the inception structure, image data from a previous layer are processed in four different ways, and these processed data are concatenated. In second row of each box within parentheses, example (image data height) × (image data width) × (image data channel) data are provided. With this structure, 1 × 1 convolutional layers are added in advance of 3 × 3 convolutional and 5 × 5 convolutional layers which are computationally heavy burden. With 1 × 1 convolutional layers, channels can be reduced; therefore, the computational burden would be reduced (for example, channels of input data are reduced from 192 to 96 for 3 × 3 convolutional layers and from 192 to 16 for 5 × 5 convolutional layers). **b** With the skip structure, processed data and input data are concatenated. This structure allows efficient backpropagation of errors. **c** With the Texture CNN, the energy measure of each processed image datum at the last convolution layer is pooled and is connected to each unit in a fully connected layer

are also known to be useful for semantic segmentation of images. U-net [32] was developed for semantic segmentation of biomedical images. This network employed a large number of feature channels so that the context information could be propagated to higher resolution layers. In SegNet [33], information regarding which pixel was used at pooling layer was taken into account at upsampling layers. Very deep Residual Encoder-Decoder Network (RED-Net) [34], which has skip connections, was also introduced for image restoration (for denoising and super-resolution). Pooling layers as well as fully connected layers are not included in the RED-Net.

Ensemble

Combining several different CNNs is useful to improve the performance of models. This technique is called as ensemble learning. With this technique, same input data are fed to several CNNs and the output data from these CNNs are integrated to derive the final result of models. Lakhani et al. utilized this technique (ensemble of AlexNet and GoogLeNet) to build better model in predicting lung tuberculosis from chest radiography [35].

Transfer learning

Transfer learning [36, 37] is a technique gathering wide attention. With this technique, a CNN trained with large sample size data in task A is transferred to task B. And the CNN is fine-tuned with data of task B. This technique is useful when we want to get a model of the task B for which labeled data is hard to collect for a large sample size (Fig. 11). According to Lakhani et al. models pre-trained by using ImageNet data (http://www.image-net.org/) showed higher performance compared with models trained from scratch for detection of tuberculosis with chest X-ray [35]. How many sample size is required to fine-tune pre-trained CNNs may differ from task to task. Mohamed et al. reported that the performance of the model was not so different when different sizes of training data were used for fine-tuning [38]. On the contrary, Larson et al. reported that the performance of the model improved when increased numbers of training data for fine-tuning were included [39].

Clinical application of convolutional neural network

Some researchers have developed CNN models by using radiological images. In this section, we introduce recent studies investigating clinical application of CNNs in terms of (a) lesion detection, (b) lesion evaluation, (c) estimation of patient survival, (d) segmentation of images, and (e) others.

Lesion detection

CNN models for detection of lesions based on chest radiography and mammography have been developed. According to Lakhani et al. by using deep learning models based on chest radiography, tuberculosis could be detected with the AUC of 0.99 [35]. Other researchers reported that pleural effusion, pulmonary edema, consolidation, cardiomegaly, and pneumothorax could be detected with AUCs of 0.96, 0.87, 0.85, 0.88, and 0.86, respectively, by using deep learning models based on chest radiography [40]. It was reported

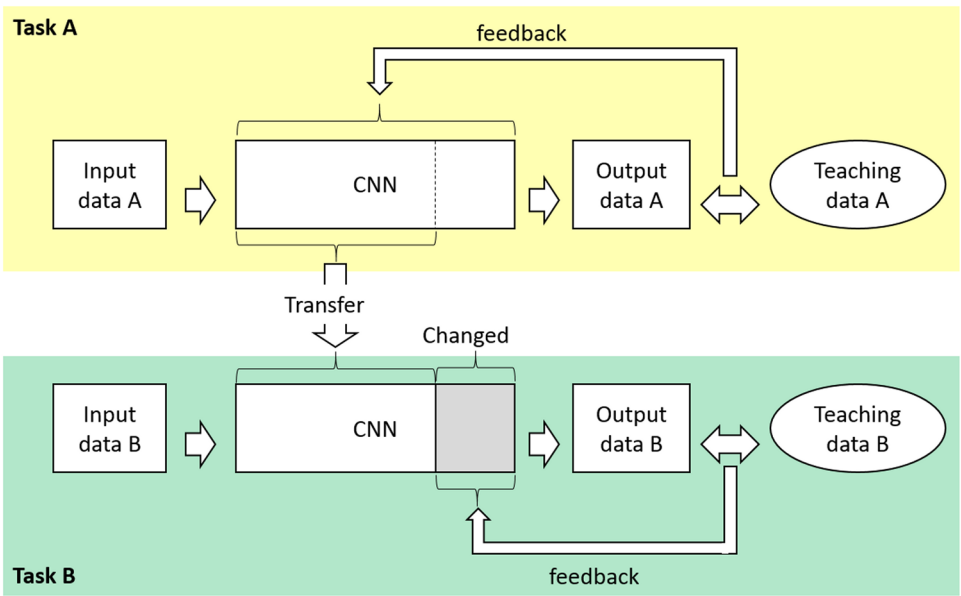

Fig. 11 A scheme of transfer learning. A CNN trained with task A, for which data with large sample size can be prepared, is transferred to task B. And this transferred CNN is further fine-tuned (only the changed layers are further adjusted in this figure) with some data for task B

that breast cancers could be detected by using models based on mammography with the AUC of 0.82, which was comparable to experienced radiologists (AUCs of 0.79–0.87) [41].

CNN models which use computed tomography (CT) and MR images were also developed. By using deep learning models based on brain CT, abnormal findings of hemorrhage, mass effect, and hydrocephalus were reportedly to be detected with sensitivity/specificity of 0.90/0.85. And also, suspected acute infarct could be detected with sensitivity/specificity of 0.62/0.96 [42]. According to Wang et al. when T2 weighted MR images were fed to trained CNNs, prostate carcinomas could be detected with the AUC of 0.84, which was significantly higher than the conventional machine learning method [support vector machine based on scale-invariant feature transform (SIFT) features] with the AUC of 0.70 [43].

CNN models which can detect multiple lesions in a patient have also been developed. By using CNN models based on MR imaging (T1 weighted images and fluid-attenuated inversion recovery images), lacunes were reportedly to be detected with sensitivity of 0.97 and false positive findings of 0.13 per slice [44]. Nakao et al. have developed methods which automatically detect cerebral aneurysms with magnetic resonance angiography [19]. They reported that by using CNN models, cerebral aneurysms could be detected with sensitivity of 0.94 and false positive of 2.9 per case or sensitivity of 0.70 and false positive of 0.26 per case. Deep learning techniques were also useful in detection of lung nodules. According to Jiang et al. lung nodules could be found with sensitivity of 0.80 and 4.7 false positive findings per scan or with sensitivity of 0.94 and 15.1 false positive findings per scan with chest CT [45].

Evaluation of lesion

The CNN is also useful in evaluations of lesions. For gliomas, variation of isocitrate dehydrogenase 1 (IDH1) was reported to be associated with better overall survival [46, 47]. And patients with glioblastoma containing a methylated O-6-methylguanine methyltransferase (MGMT) promotor is known to be benefited from temozolomide [48]. According to Chang et al. IDH1 status in grades II–IV gliomas could be predicted with accuracy of 0.86 by using CNN models based on MR images, and the performance was improved to accuracy of 0.89 when age was also incorporated [49]. Korfiatis et al. have developed CNN models which could classify T2 weighted images to glioblastoma with methylated MGMT, glioblastoma without methylated MGMT, or no tumor, with accuracy of 0.95 at slice base [29].

For head and neck imaging, Chi et al. reported that the malignancy of thyroid nodules could be predicted with accuracy of 0.96–0.98 by using CNN models based on ultrasound (US) images [50]. Rhabdomyosarcomas are mainly divided to embryonal subtype and alveolar subtype, and the latter is known to be associated with poor prognosis [51]. Banerjee et al. have reported that these two entities, which originated in the head and neck region, could be differentiated with accuracy of 0.85 by using CNN models based on MR images [24].

For lung imaging, Anthimopoulos et al. reported that the pattern of interstitial lung diseases (healthy, ground glass opacity, micronodules, consolidation, reticulation, and honeycoming) could be differentiated with accuracy of 0.86 by using CNN models based on chest CT images [52]. By using chest CT images, malignancy of lung nodules were reportedly to be predicted with accuracy of 0.84–0.90 [53, 54].

Deep learning techniques were also applied to liver imaging. A CNN model predicting fibrosis stage as fibrosis score obtained through deep learning (F_{DL} score) was developed by using gadoxetic acid-enhanced hepatobiliary phase MR images [55]. F_{DL} score showed strong correlation with pathologically evaluated fibrosis stage. And the prediction of cirrhosis, advanced fibrosis, and significant fibrosis could be possible by using the model with AUC of 0.84, 0.84, and 0.85, respectively. By using CNN models based on dynamic CT images (unenhanced, arterial phase, and delayed phase), liver masses could be classified into five categories [category A, classical hepatocellular carcinomas (HCCs); category B, malignant liver tumors other than classic and early HCCs; category C, indeterminate masses or mass-like lesions and rare benign liver masses; category D, hemangiomas; and category E, cysts] with sensitivity of 0.71, 0.33, 0.94, 0.90, and 1.00, respectively [56]. Ben-Cohen et al. developed CNN models which predict the primary origin of liver metastasis among four sites (melanoma, colorectal cancer, pancreatic cancer, breast cancer) by using CT images. They reported that primary site could be predicted with accuracy of 0.56. If the prediction was made as top-2 and top-3 classification tasks, the accuracy was 0.83 and 0.99, respectively [57].

Patient survival

Lao et al. have developed a deep learning-based radiomics model based on MR images to estimate the survival of patients with glioblastomas. This model utilized quantitative parameters outputted by CNNs as deep learning-based radiomics features. After feature selection, the established model enabled prediction of overall survival with c-index of 0.71 [58]. A CNN model based on chest CT was reported to be useful for chronic obstructive pulmonary disease patients or smokers in predicting mortality with c-indices of 0.60–0.72 [20]. Future researches are expected to develop models which predict patient survival more accurately from data including radiological images of various body regions.

Segmentation

Segmentation of lesions with CT images is an important process for radiotherapy planning, and application of deep learning models to this topic was studied by some investigators. For treatment of head and neck cancers, organs at risk could be segmented with Dice similarity coefficient of 0.90 for mandible and 0.87 for spinal cord [59]. For patients scheduled for liver stereotactic body radiation therapy, portal vein was reportedly to be segmented with Dice similarity coefficient of 0.83 by using models based on CNN techniques [60]. For treatment of rectal cancer, clinical target volume was reportedly to be segmented with Dice similarity coefficient of 0.88 [61].

Other topics

With the CNN model trained by using hand radiographs, skeletal maturity could be estimated with accuracy similar to that of an expert radiologist [39]. In predicting the density of background breast (scattered density vs. heterogeneously dense), CNN models demonstrated high classification performance with the AUC of 0.99 [38].

The deep learning technique was also used to generate pseudo-CT images from MR images for attenuation correction in positron emission tomography (PET)/MR. This technique achieved successful MR based attenuation correction of PET results in brain and pelvic regions. Liu et al. reported that significantly lower PET reconstruction errors were achieved with deep learning based MR attenuation correction compared with Dixon-based soft-tissue and air segmentation and anatomic CT-based template registration [62]. According to Leynes et al. models which use Dixon MR images and proton-density-weighted zero echo-time MR images to synthesize pseudo-CT images enabled PET quantification closer to CT based attenuation correction when compared to Dixon MR images based conventional PET quantification in pelvic PET/MR [63].

Improving CT image quality was reported to be possible by using the deep learning technique. CT images scanned with reduced radiation dose levels are known to suffer from increased image noise. Though iterative reconstruction techniques are known to be useful for improving image quality for reduced-dose CT [64–67] or routine-dose CT [68–70], these techniques are known to be associated with long image reconstruction time. Chen et al. reported that quality of CT images could be improved in terms of qualitative and quantitative analyses with less time by using models based on deep learning techniques [71, 72]. Considering the potentials of CNN models reported by these studies, further studies are expected to examine whether these techniques are useful for evaluations of diseases.

Future directions of convolutional neural network

Several studies have successfully built models to obtain clinically useful information in certain tasks. For some tasks, CNNs achieved higher performance compared to conventional machine learning methods [43]. And for some tasks, CNNs achieved the performance almost equal to radiologists [39, 41]. Models based on deep learning techniques could generate clinically useful images from original images (such as pseudo-CT images obtained from MR images) [62, 63].

In the future, clinical application of CNN models for diagnoses of other diseases will be investigated. Transfer learning might be useful in developing CNN models for relatively rare diseases. Because the deep learning techniques are evolving rapidly, performance of some CNN models might be further improved by incorporating such state-of-the-art deep learning techniques in the future. Models will also evolve so that they require fewer preprocessing. In addition to these topics, the effect of reconstruction kernel or thickness of images on the performance of deep learning models also needs to be investigated [5]. Investigation regarding how the difference of scanners and image reconstruction techniques affect the performance would also be important. While radiomics parameters are known to be affected by the difference of scanners and image reconstruction techniques with various degrees [73, 74], these effects on performance of deep learning models are not well known.

Some limitations associated with deep learning techniques should be acknowledged. First, though CNNs are useful in radiological diagnoses for some tasks, how they make judgement is a black box. Second, the optimal structure and hyper-parameters of CNNs and the numbers of cases needed to train models differ from task to task. Hyper-parameters which achieved high performance in one task do not necessarily result in good performance in other tasks. Third, many cases are required for training process with deep learning techniques. This problem might be mitigated by using transfer learning techniques. For some rare diseases, though it would be associated with privacy concerns, a public database which allows collecting enough numbers of labeled datasets might be necessary to develop CNN models. Finally, though, the deep learning technique has been successfully applied to some tasks, developed models are valid in only a dedicated task and they are not applicable to other tasks.

Because of the potential of deep learning technique, wide attention by many researchers, and the freely available deep learning frameworks, many CNN models useful for daily clinical practice will be further developed in the future. We hope that this article will help radiologists to understand the recent and future studies regarding deep learning techniques.

Funding No specific funding was disclosed.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical statement This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
2. Fukushima K, Miyake S. Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recognit. 1982;15:455–69.
3. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol. 1962;160:106–54.
4. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing System 25 (NIPS 2012). 2012. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks:Published. Accessed 14 Dec 2017.
5. Kahn CE Jr. From images to actions: opportunities for artificial intelligence in radiology. Radiology. 2017;285:719–20.
6. Dreyer KJ, Geis JR. When machines think: radiology's next frontier. Radiology. 2017;285:713–8.
7. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016;278:563–77.
8. Skogen K, Ganeshan B, Good C, Critchley G, Miles K. Measurements of heterogeneity in gliomas on computed tomography relationship to tumour grade. J Neurooncol. 2013;111:213–9.
9. Lubner MG, Stabo N, Abel EJ, Del Rio AM, Pickhardt PJ. CT textural analysis of large primary renal cell carcinomas: pretreatment tumor heterogeneity correlates with histologic findings and clinical outcomes. AJR Am J Roentgenol. 2016;207:96–105.
10. Yasaka K, Akai H, Nojima H, Shinozaki-Ushiku A, Fukayama M, Nakajima J, et al. Quantitative computed tomography texture analysis for estimating histological subtypes of thymic epithelial tumors. Eur J Radiol. 2017;92:84–92.
11. Kickingereder P, Burth S, Wick A, Gotz M, Eidel O, Schlemmer HP, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. Radiology. 2016;280:880–9.
12. Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, et al. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non-small cell lung cancer. Radiology. 2016;281:947–57.
13. Yip C, Landau D, Kozarski R, Ganeshan B, Thomas R, Michaelidou A, et al. Primary esophageal cancer: heterogeneity as potential prognostic biomarker in patients treated with definitive chemotherapy and radiation therapy. Radiology. 2014;270:141–8.
14. Ng F, Ganeshan B, Kozarski R, Miles KA, Goh V. Assessment of primary colorectal cancer heterogeneity by using whole-tumor texture analysis: contrast-enhanced CT texture as a biomarker of 5-year survival. Radiology. 2013;266:177–84.
15. Kiryu S, Akai H, Nojima M, Hasegawa K, Shinkawa H, Kokudo N, et al. Impact of hepatocellular carcinoma heterogeneity on computed tomography as a prognostic indicator. Sci Rep. 2017;7:12689.
16. Li H, Zhu Y, Burnside ES, Drukker K, Hoadley KA, Fan C, et al. MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of MammaPrint, Oncotype DX, and PAM50 gene assays. Radiology. 2016;281:382–91.
17. Goh V, Ganeshan B, Nathan P, Juttla JK, Vinayan A, Miles KA. Assessment of response to tyrosine kinase inhibitors in metastatic renal cell cancer: CT texture as a predictive biomarker. Radiology. 2011;261:165–71.
18. Le QV, Ranzato M, Monga R, Devin M, Chen K, Corrado GS, et al. Building high-level features using large scale unsupervised learning. International Conference on Machine Learning. 2012. http://icml.cc/2012/papers. Accessed 14 Dec 2017.
19. Nakao T, Hanaoka S, Nomura Y, Sato I, Nemoto M, Miki S, et al. Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. J Magn Reson Imaging. 2017. https://doi.org/10.1002/jmri.25842.
20. Gonzalez G, Ash SY, Vegas Sanchez-Ferrero G, Onieva Onieva J, Rahaghi FN, Ross JC, et al. Disease staging and prognosis in smokers using deep learning in chest computed tomography. Am J Respir Crit Care Med. 2018;197:193–203.
21. Nair V, Hinton G. Rectified linear units improve restricted Boltzmann machines. International Conference on Machine Learning. 2010. Accessed 14 Dec 2017.
22. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1929–58.
23. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Cornell University Library. 2015. http://arxiv.org/abs/1502.03167. Accessed 30 April 2017.
24. Banerjee I, Crawley A, Bhethanabotla M, Daldrup-Link HE, Rubin DL. Transfer learning on fused multiparametric MR images for classifying histopathological subtypes of rhabdomyosarcoma. Comput Med Imaging Graph. 2017. https://doi.org/10.1016/j.compmedimag.2017.05.002.
25. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res. 2011;12:2121–59.
26. Kingma DP, Ba JL. Adam: a method for stochastic optimization. Cornell University Library. 2014. http://arxiv.org/abs/1412.6980. Accessed 30 April 2017.
27. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Cornell University Library. 2014. https://arxiv.org/abs/1409.4842. Accessed 14 Dec 2017.
28. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Cornell University Library. 2015. https://arxiv.org/abs/1512.03385. Accessed 14 Dec 2017.
29. Korfiatis P, Kline TL, Lachance DH, Parney IF, Buckner JC, Erickson BJ. Residual deep convolutional neural network predicts MGMT methylation status. J Digit Imaging. 2017;30:622–8.
30. Andrearczyk V, Whelan PF. Using filter banks in convolutional neural networks for texture classification. Cornell University Library. 2016. https://arxiv.org/abs/1601.02919. Accessed 30 April 2017.
31. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 2015. http://ieeexplore.ieee.org/document/7298965/?reload=true. Accessed 14 Dec 2017.
32. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Cornell University Library. 2015. https://arxiv.org/abs/1505.04597. Accessed 14 Dec 2017.
33. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. Cornell University Library. 2015. https://arxiv.org/pdf/1511.00561. Accessed 14 Dec 2017.

34. Mao X, Shen C, Yang Y. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. Cornell University Library. 2016. https://arxiv.org/abs/1603.09056. Accessed 14 Dec 2017.
35. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284:574–82.
36. Caruana R. Multitask learning. Mach Learn. 1997;28:41–75.
37. Bengio Y. Deep learning of representations for unsupervised and transfer learning. In: JMLR: Workshop and Conference Proceedings. 2012;17–37.
38. Mohamed AA, Berg WA, Peng H, Luo Y, Jankowitz RC, Wu S. A deep learning method for classifying mammographic breast density categories. Med Phys. 2018;45:314–21.
39. 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. 2017. https://doi.org/10.1148/radiol.2017170236.
40. Cicero M, Bilbily A, Colak E, Dowdell T, Gray B, Perampaladas K, et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. Invest Radiol. 2017;52:281–7.
41. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol. 2017;52:434–40.
42. Prevedello LM, Erdal BS, Ryu JL, Little KJ, Demirer M, Qian S, et al. Automated critical test findings identification and online notification system using artificial intelligence in imaging. Radiology. 2017;285:923–31.
43. Wang X, Yang W, Weinreb J, Han J, Li Q, Kong X, et al. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. Sci Rep. 2017;7:15415.
44. Ghafoorian M, Karssemeijer N, Heskes T, Bergkamp M, Wissink J, Obels J, et al. Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. Neuroimage Clin. 2017;14:391–9.
45. Jiang H, Ma H, Qian W, Gao M, Li Y. An automatic detection system of lung nodule based on multi-group patch-based deep learning network. IEEE J Biomed Health Inform. 2017. https://doi.org/10.1109/JBHI.2017.2725903.
46. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008;321:1807–12.
47. Yan H, Parsons DW, Jin G, McLendon R, Rasheed BA, Yuan W, et al. IDH1 and IDH2 mutations in gliomas. N Engl J Med. 2009;360:765–73.
48. Hegi ME, Diserens AC, Gorlia T, Hamou MF, de Tribolet N, Weller M, et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med. 2005;352:997–1003.
49. Chang K, Bai HX, Zhou H, Su C, Bi WL, Agbodza E, et al. Residual convolutional neural network for determination of IDH status in low- and high-grade gliomas from MR imaging. Clin Cancer Res. 2017. https://doi.org/10.1158/1078-0432.CCR-17-2236.
50. Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. J Digit Imaging. 2017;30:477–86.
51. Malempati S, Hawkins DS. Rhabdomyosarcoma: review of the Children's Oncology Group (COG) Soft-Tissue Sarcoma Committee experience and rationale for current COG studies. Pediatr Blood Cancer. 2012;59:5–10.
52. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging. 2016;35:1207–16.
53. Song Q, Zhao L, Luo X, Dou X. Using deep learning for classification of lung nodules on computed tomography images. J Healthc Eng. 2017;2017:8314740.
54. Nibali A, He Z, Wollersheim D. Pulmonary nodule classification with deep residual networks. Int J Comput Assist Radiol Surg. 2017;12:1799–808.
55. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Liver fibrosis: deep convolutional neural network for staging by using gadoxetic acid-enhanced hepatobiliary phase MR images. Radiology. 2017. https://doi.org/10.1148/radiol.2017171928.
56. Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology. 2018;286:887–96.
57. Ben-Cohen A, Klang E, Diamant I, Rozendorn N, Raskin SP, Konen E, et al. CT image-based decision support system for categorization of Liver metastases into primary cancer sites: initial results. Acad Radiol. 2017;24:1501–9.
58. Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, et al. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. Sci Rep. 2017;7:10353.
59. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys. 2017;44:547–57.
60. Ibragimov B, Toesca D, Chang D, Koong A, Xing L. Combining deep learning with anatomy analysis for segmentation of portal vein for liver SBRT planning. Phys Med Biol. 2017;62:8943–58.
61. Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. Med Phys. 2017;44:6377–89.
62. Liu F, Jang H, Kijowski R, Bradshaw T, McMillan AB. Deep learning MR imaging-based attenuation correction for PET/MR imaging. Radiology. 2018;286:676–84.
63. Leynes AP, Yang J, Wiesinger F, Kaushik SS, Shanbhag DD, Seo Y, et al. Direct PseudoCT generation for pelvis PET/MRI attenuation correction using deep convolutional neural networks with multi-parametric MRI: zero echo-time and dixon deep pseudoCT (ZeDD-CT). J Nucl Med. 2017. https://doi.org/10.2967/jnumed.117.198051.
64. Yasaka K, Katsura M, Akahane M, Sato J, Matsuda I, Ohtomo K. Model-based iterative reconstruction for reduction of radiation dose in abdominopelvic CT: comparison to adaptive statistical iterative reconstruction. Springerplus. 2013;2:209.
65. Katsura M, Matsuda I, Akahane M, Sato J, Akai H, Yasaka K, et al. Model-based iterative reconstruction technique for radiation dose reduction in chest CT: comparison with the adaptive statistical iterative reconstruction technique. Eur Radiol. 2012;22:1613–23.
66. Pickhardt PJ, Lubner MG, Kim DH, Tang J, Ruma JA, del Rio AM, et al. Abdominal CT with model-based iterative reconstruction (MBIR): initial results of a prospective trial comparing ultralow-dose with standard-dose imaging. AJR Am J Roentgenol. 2012;199:1266–74.
67. Yamada Y, Jinzaki M, Tanami Y, Shiomi E, Sugiura H, Abe T, et al. Model-based iterative reconstruction technique for ultralow-dose computed tomography of the lung: a pilot study. Invest Radiol. 2012;47:482–9.
68. Deak Z, Grimm JM, Treitl M, Geyer LL, Linsenmaier U, Korner M, et al. Filtered back projection, adaptive statistical iterative reconstruction, and a model-based iterative reconstruction in abdominal CT: an experimental clinical study. Radiology. 2013;266:197–206.
69. Yasaka K, Katsura M, Hanaoka S, Sato J, Ohtomo K. High-resolution CT with new model-based iterative reconstruction with resolution preference algorithm in evaluations of lung nodules:

comparison with conventional model-based iterative reconstruction and adaptive statistical iterative reconstruction. Eur J Radiol. 2016;85:599–606.
70. Yasaka K, Furuta T, Kubo T, Maeda E, Katsura M, Sato J, et al. Full and hybrid iterative reconstruction to reduce artifacts in abdominal CT for patients scanned without arm elevation. Acta Radiol. 2017;58:1085–93.
71. Chen H, Zhang Y, Zhang W, Liao P, Li K, Zhou J, et al. Low-dose CT via convolutional neural network. Biomed Opt Express. 2017;8:679–94.
72. Chen H, Zhang Y, Kalra MK, Lin F, Chen Y, Liao P, et al. Low-dose CT with a residual encoder-decoder convolutional neural network. IEEE Trans Med Imaging. 2017;36:2524–35.
73. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring computed tomography scanner variability of radiomics features. Invest Radiol. 2015;50:757–65.
74. Yasaka K, Akai H, Mackin D, Court L, Moros E, Ohtomo K, et al. Precision of quantitative computed tomography texture analysis using image filtering: a phantom study for scanner variability. Medicine (Baltimore). 2017;96:e6993.