Japanese Journal of Radiology (2019) 37:15–33
https://doi.org/10.1007/s11604-018-0795-3

INVITED REVIEW

Technical and clinical overview of deep learning in radiology

Daiju Ueda[1] · **Akitoshi Shimazaki**[1] · **Yukio Miki**[1]

Received: 15 October 2018 / Accepted: 17 November 2018 / Published online: 1 December 2018

Abstract

Deep learning has been applied to clinical applications in not only radiology, but also all other areas of medicine. This review provides a technical and clinical overview of deep learning in radiology. To gain a more practical understanding of deep learning, deep learning techniques are divided into five categories: classification, object detection, semantic segmentation, image processing, and natural language processing. After a brief overview of technical network evolutions, clinical applications based on deep learning are introduced. The clinical applications are then summarized to reveal the features of deep learning, which are highly dependent on training and test datasets. The core technology in deep learning is developed by image classification tasks. In the medical field, radiologists are specialists in such tasks. Using clinical applications based on deep learning would, therefore, be expected to contribute to substantial improvements in radiology. By gaining a better understanding of the features of deep learning, radiologists could be expected to lead medical development.

Keywords Deep learning · Artificial intelligence · AI · Neural network · Radiology · Review

Abbreviations

NLP	Natural language processing
ANN	Artificial neural network
AUC	Area under the curve
ROC	Receiver operating characteristic
CNN	Convolutional neural network
SR	Super resolution
LR	Low resolution
HR	High resolution
GAN	Generative adversarial networks
NAS	Neural architecture search
ILSVRC	ImageNet large-scale visual recognition challenge
FCN	Fully convolutional network
CRF	Conditional random field
R-CNN	Regions with convolutional neural network features
YOLO	You only look once
SSD	Single shot MultiBox detector
PSP	Pyramid scene parsing
FSRCNN	Fast super resolution convolutional neural network
ESPCN	Efficient sub-pixel convolutional neural network
VDSR	Very deep super resolution
DRCN	Deeply-recursive convolutional network
EDSR	Enhanced deep super resolution network
RDN	Residual dense network
DBPN	Deep back-projection networks
ZSSR	Zero shot super resolution
CBOW	Continuous bag-of-words
GloVe	Global vectors for word representation
DCGAN	Deep convolutional generative adversarial network
XOGAN	Generative adversarial network with XO-structure
ENAS	Efficient neural architecture search
DARTS	Differentiable architecture search
NAO	Neural architecture optimization
HMH	Hemorrhage, mass effect, or hydrocephalus
CT	Computed tomography
SAI	Suspected acute infarct
HCC	Hepato-cellular carcinoma
MR	Magnetic resonance
MCI	Mild cognitive impairment
ICH	Intracranial hemorrhage
EDH/SDH	Epidural/subdural hemorrhage

✉ Daiju Ueda
ueda.daiju@gmail.com

[1] Department of Diagnostic and Interventional Radiology, Osaka City University Graduate School of Medicine, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan

SAH	Subarachnoid hemorrhage
ASL	Arterial spin labeling
VN	Variational network
PICS	Parallel imaging and compressed sensing
DnCNN	Denoising convolutional neural network
PE	Pulmonary embolism
AI	Artificial intelligence

Introduction

Deep learning has led to dramatic, state-of-the-art improvements in speech recognition, visual object recognition, object detection, and natural language processing (NLP) [1]. Two key aspects are common among the various high-level descriptions of deep learning [2]: models consisting of multiple layers or stages of nonlinear information processing; and methods for supervised or unsupervised learning of feature representation at successively higher and more abstract layers.

Deep learning is a subfield of artificial neural networks (ANNs) in machine learning. Broadly speaking, there have been three waves of development [3]: the deep learning known as cybernetics in the 1940s–1960s; the deep learning known as connectionism in the 1980s–1990s; and the current resurgence under the name of deep learning, which began in 2006. The first wave started with the development of theories of biological learning [4, 5] and implementations of the first models, such as the perceptron [6], which enabled the training of a single neuron. The second wave started with back-propagation [7] to train an ANN with one or two hidden layers. The third and current wave started with the development of a kind of ANN, called a deep belief network, which could be efficiently trained using a strategy called greedy layer-wise pretraining [8–10]. The ANN in the third wave is called deep learning. In contrast to deep learning, the networks in the previous two waves were called shallow learning. Compared with the first two waves, the third has been achieving substantial development.

In the scope of radiology, the first research on ANNs was reported by Asada et al. [11] in 1990. Shallow learning was applied to the differential diagnosis of interstitial lung disease. There were only three layers (Fig. 1). On the other hand, the first research about deep learning was reported by Cicero et al. [12], who classified chest X-ray abnormalities using an algorithm developed by GoogLeNet [13] (Fig. 1). The former achieved an area under the receiver operating characteristic curve (AUC-ROC) of 0.97 for intestinal lung diseases, while the latter achieved an AUC-ROC of 0.85–0.96 for differentiating a normality and five abnormalities. At a glance, both algorithms achieved high performance; however, there was a great difference. The primary difference between the past ANN (shallow learning) and present ANN (deep learning) is the learning process. In the former research, the shallow ANN did not need images, but did need the radiologists' interpretations of those images. The ANN could not learn from the images alone. For example, when a chest X-ray showed a nodule in the upper left area, the radiologist needed to input that information into the ANN. In other words, radiologists were needed to extract features before the ANN could do so. On the other hand, deep learning can develop algorithms using images only and then immediately begin extracting features from those images. The learning process in deep learning is, therefore, an end-to-end approach.

The remainder of this review is divided into four sections: a technical overview, clinical applications, a quality proof, and the future of deep learning in medicine. In the technical and clinical overview sections, techniques and applications are explained in chronological order. In the quality proof section, the reliability and accuracy of clinical applications are discussed. In the last section on future perspectives, some possibilities about the future of deep learning and associated changes in radiology are discussed.

Technical overview

In this section, deep learning techniques are summarized. This aim of this overview was to provide a historical and practical introduction of deep learning. The deep learning principles concerning convolutional neural networks (CNNs) were summarized in a previous review [14]. From the viewpoint of radiology, deep learning techniques are divided into the following five categories: classification, object detection, semantic segmentation, image processing, and NLP. Classification is a process in which objects are recognized, differentiated, and understood. Object detection is the extraction of particular areas from images. Semantic segmentation is the partition of an image into several coherent parts; it describes the process of associating each pixel with a class label. For example, finding a nodule from mammography is "object detection", evaluating a nodule to determine whether it is benign or malignant is "classification", and extracting contours from a nodule is "segmentation". Image processing is the technique of converting images. Examples of classification, detection, and segmentation are shown in Figs. 2 and 3. In the scope of image processing techniques, super resolution (SR) is the most well-known and important. The SR technique enhances images from low resolution (LR) to high resolution (HR) by extracting parameters. Examples of SR are shown in Fig. 4. Finally, NLP is defined as an automatic manipulation of natural language. In radiology, NLP is a fundamental method used to extract data from radiological reports or clinical records for analysis. A

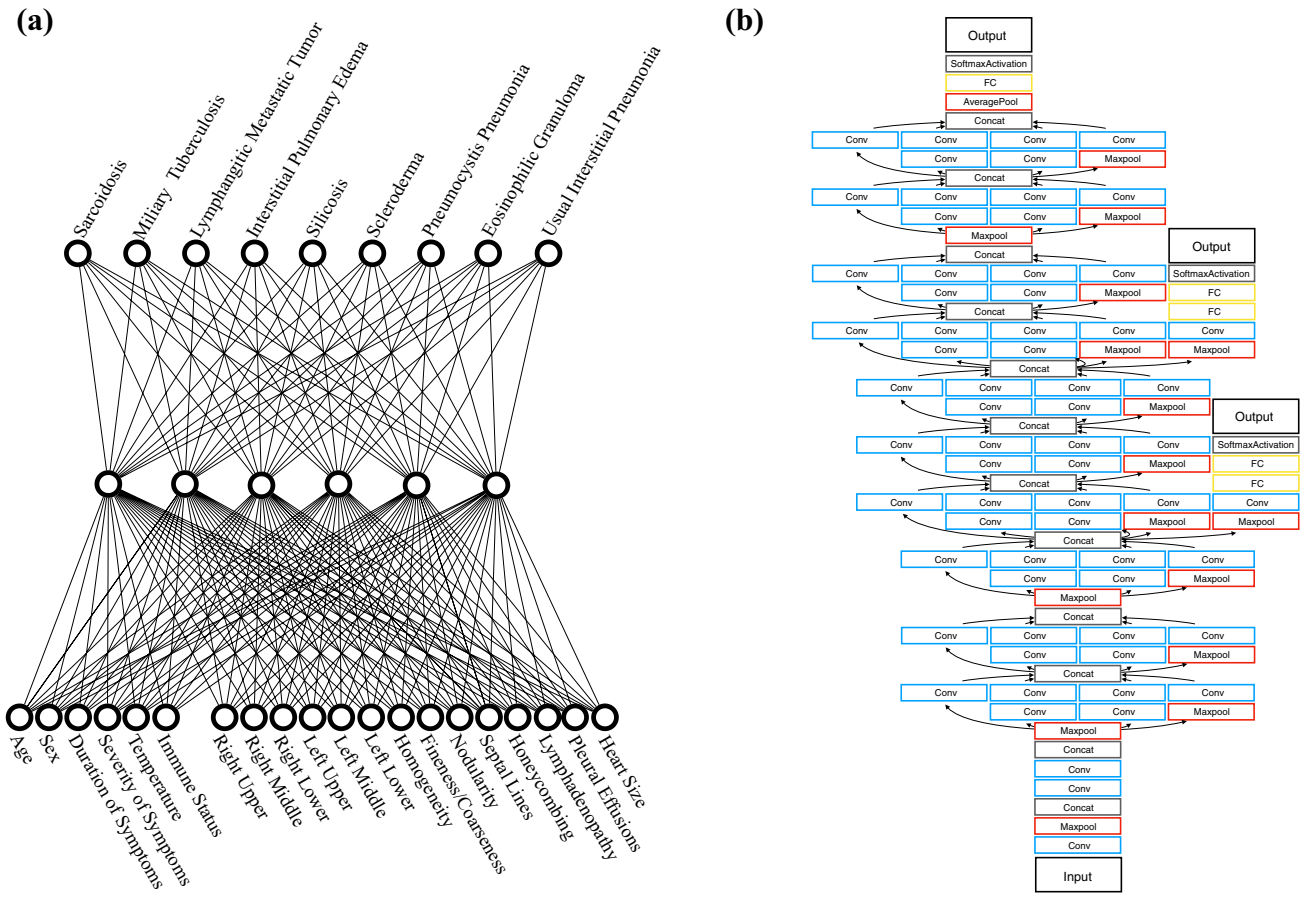

Fig. 1 The first models of shallow and deep learning in radiology. **a** Three layers, including one hidden layer, in the research of Asada et al. [11]. **b** GoogLeNet in the research of Cicero et al. [12]

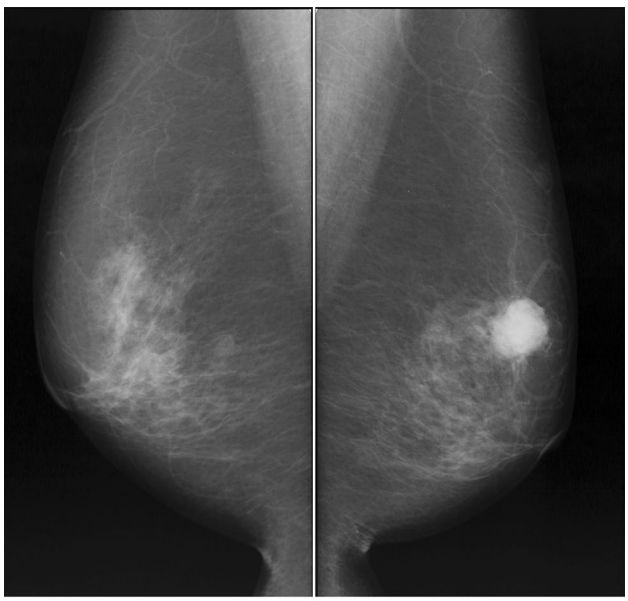

Fig. 2 Examples of classification

brief technical overview of deep learning models is shown in Fig. 5.

The remainder of this section is composed of eight subsections; the first five describe deep learning architectures by category, and the final three describe important techniques in deep learning—the fine-tuning technique, generative adversarial networks (GANs), and neural architecture search (NAS).

Classification

Classification is the most well-known category in deep learning and triggered the deep learning boom. The core technology in deep learning is classification and the key for classification is a CNN. The CNN has its roots in a neocognitron proposed by Fukushima et al. [15]. The idea of neocognitron was based on the biological mechanisms of visual recognition in the primary visual cortex of a vertebrate [16]. The earliest CNNs were used to recognize individual objects in extremely small, tightly cropped images [7]. Since then, there has been a gradual increase in the size of images. The

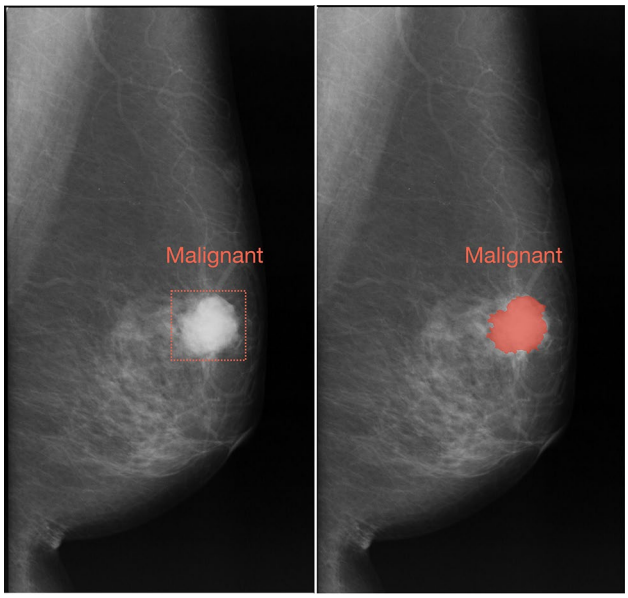

Fig. 3 Examples of object detection and semantic segmentation. Finding a nodule from mammography is a type of object detection. Extracting contours from a nodule is a type of segmentation

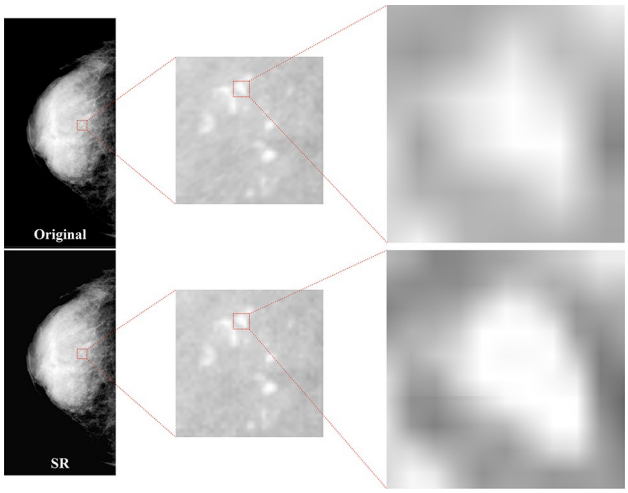

Fig. 4 Examples of super resolution. Upper images are original mammography images, and lower images are mammography images applying super resolution

largest competition in object recognition is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which has been held annually since 2010. A dramatic moment in the meteoric rise of deep learning came in 2012, when, for the first time, a CNN—AlexNet—won top prize in the ILSVRC [17]. Subsequently, VGGNet [18] won second prize in 2013, GoogLeNet [13] won first prize in 2014, and ResNet [19] won first prize in 2015, when it finally achieved a top-5 error rate, down to 3.6%; compare this with the human rate,

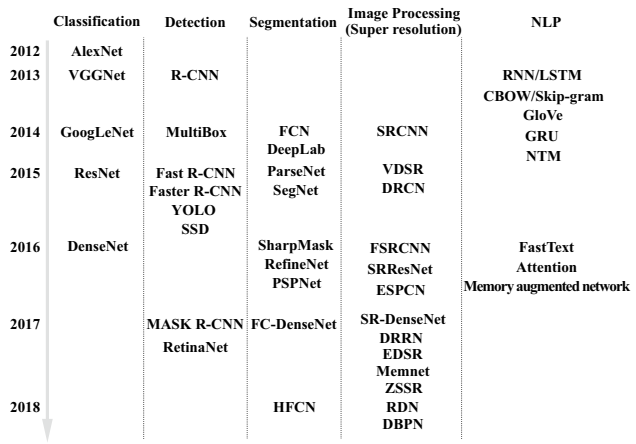

Fig. 5 Technical overview of the deep learning models

which is said to be about 5% [3]. Generally speaking, the deeper the network, the better the accuracy. This is because deeper networks can express more complicated shapes, as the number of convolutional layers deepens. There are 5, 16, 59, and 150 layers in AlexNet, VGGNet-19, GoogLeNet (Inception v1), and ResNet-152, respectively. Unfortunately, however, the deeper the network, the more difficult the back-propagation. To achieve good balance between network depth and back-propagation, GoogLeNet introduced the "Inception module" to the network while ResNet introduced a "skip connection". The goal of the Inception module is to act as a multi-level feature extractor by computing several convolutions within the same module. One of the most beneficial aspects of this architecture is that it allows for the number of units at each stage to be significantly increased without substantial increase in computational complexity. The main feature of ResNet is the skip connection, in which a few convolution layers at a time are bypassed. Each bypass gives rise to a residual block in which the convolution layers predict a residual that is added to the block's input tensor, which promotes effective learning. DenseNet [20] is constructed from "dense blocks", which connect each layer to every other in a feed-forward fashion. For each layer, the feature maps of all preceding layers are used as inputs, and its own feature maps are used as inputs in all subsequent layers.

Object detection

Based on the techniques developed in classification, object detection methods, which have deeper architectures with the capacity to learn more complex features, have been greatly improving. In addition, their expressivity and robust training algorithms allow informative object representations to be learned without the need to design features manually. The frameworks of generic object detection methods can be categorized into two main types [21]. The first type is a region

proposal-based framework, which follows the traditional object detection pipeline, that is, it generates region proposals and then classifies each one into object categories. The second type is a regression-based framework, which regards object detection as a classification problem and thereby adopts a unified framework to achieve the final results (categories and locations) directly. Deep learning techniques developed in classification are used as the core technology for both types. The region proposal-based methods mainly include regions with CNN features (R-CNN) [22], Fast R-CNN [23], Faster R-CNN [24], and Mask R-CNN [25]. The regression-based frameworks mainly include MultiBox [26], You only look once (YOLO) [27], the single shot MultiBox detector (SSD) [28], and RetinaNet [29]. The correlations between these two pipelines are bridged by the anchors introduced in Faster R-CNN.

The first region proposal-based detection model is R-CNN [22]. R-CNN adopts a selective search to generate about 2 k region proposals for each image. The selective search method relies on simple bottom-up grouping and saliency cues to provide more accurate candidate boxes of arbitrary sizes quickly and reduce the searching space in object detection. Fast R-CNN [23] jointly optimizes classification and bounding box regression tasks. Faster R-CNN [24] uses an additional subnetwork to generate region proposals and enables nearly cost-free region proposals to share full-image convolutional features with the detection network. Mask R-CNN [25] extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition, thereby providing instance segmentations for objects. The first regression-based detection model is MultiBox [26]. MultiBox produces scored class-agnostic region proposals. A unified loss was introduced to bias both the localization and confidence of multiple components to predict the coordinates of class-agnostic bounding boxes. YOLO [27] accomplishes object detection via a fixed-grid regression, and as the whole detection pipeline is a single network, it can be optimized end-to-end directly for detection performance. SSD [28] was inspired by the anchors adopted in MultiBox and multi-scale representation. SSD takes advantage of a set of default anchor boxes with different aspect ratios and scales to discretize the output space of bounding boxes. To handle objects with various sizes, the network fuses predictions from multiple feature maps with different resolutions. RetinaNet [29] improves class imbalances that lead to relatively low accuracy by reshaping the standard cross entropy loss such that it down-weights the loss assigned to well-classified examples.

Semantic segmentation

The most successful state-of-the-art deep learning techniques for semantic segmentation stem from fully convolutional networks (FCNs) [30], in which existing and well-known classification models are transformed into fully convolutional ones by replacing the fully connected layers. In this way, semantic segmentation is dependent on classification techniques. ResNet has been extended to work as an FCN (e.g., DeepLab [31], RefineNet [32], and the pyramid scene parsing network (PSPNet) [33]), yielding good results based on different segmentation benchmarks. FC-DenseNet [34] introduced DenseNet, which is suitable for semantic segmentation as it naturally induces skip connections and multi-scale supervision. Although FCN-based decoder structures are considered the most popular and successful in terms of segmentation, some other remarkable structures exist. SegNet [35], which is classified as autoencoder network, is a typical example [36]. The core trainable engine consists of an encoder network, a corresponding decoder network, and a pixel-wise classification layer. The architecture of the encoder network is topologically identical to the 13 convolutional layers in VGGNet.

Semantic segmentation is a problem that requires the integration of information from various spatial scales [37] and also implies balancing local and global information. On the one hand, fine-grained or local information is crucial to achieving good pixel-level accuracy. On the other hand, it is also important to integrate information from the global context of the image to be able to resolve local ambiguities. Many approaches can be taken to make CNNs aware of global information, including refinement as a post-processing step with conditional random fields (CRFs) (DeepLab [38]), dilated convolutions (DeepLab [31]), multi-scale prediction [39], and feature fusion (ParseNet [40], SharpMask [41]). CRFs enable the combination of low-level image information, such as interactions between pixels, with the output of multi-class inference systems that produce per-pixel class scores. This combination is especially important to capture both long-range dependencies, which CNNs fail to consider, and fine local details. DeepLab [38] makes use of the fully connected pairwise CRF [42, 43]. Dilated convolutions [44], also referred to as à trous convolutions, support exponentially expanding receptive fields without losing resolution. The dilated convolutions are regular ones that make use of upsampled filters, which were introduced to an improved version of DeepLab [31]. Multi-scale networks [39] are another possible way to deal with context knowledge integration. Multi-scale networks generally make use of multiple networks that target different scales and then merge the predictions to produce a single output. As a result, the network becomes more robust to scale variations. Future fusion is a technique that consists of merging a global feature (extracted from a previous layer in a network) with a more local feature map extracted from a subsequent layer. ParseNet [40] uses the average feature for a layer to augment the features at each location to add global context.

SharpMask [41] introduced a progressive refinement module to incorporate features from the previous layer to the next in a top-down architecture. RefineNet [32] combined encoder–decoder architecture into dilated convolutions. PSPNet [33] is composed of a pyramid pooling module and dilated convolutions. Highly fused convolutional networks (HFCNs) [45] adopt a strategy involving the upsampling of multiple steps and the combining of feature maps in pooling layers with corresponding unpooling layers. As a result, HFCNs make use of the feature and reduce loss reduction when the loss is back-propagated.

Image processing

Image processing is the technique used to convert images. Among the image processing techniques, SR [46] is the most well known and important. SR refers to the task of restoring HR images from one or more LR observations of the same scene. The most basic model is SRCNN [47, 48], which is a three-layer CNN. The functions of these three nonlinear transformations are patch extraction, nonlinear mapping, and reconstruction. Fast SRCNN (FSRCNN) [49] is the first network to utilize this normal deconvolution layer; it uses nearest-neighbor interpolation, in which the points in the upsampled features are repeated several times in each direction; this configuration of upsampled pixels is redundant. Efficient sub-pixel CNN (ESPCN) [50] is composed of a CNN architecture, where the feature maps are extracted in the LR space, and an efficient sub-pixel convolution layer, which learns an array of upscaling filters to upscale the final LR feature maps into the HR output. Very deep super resolution (VDSR) [51] is the first very deep model based on VGGNet. Deeply recursive convolutional network (DRCN) [52], which is based on VDSR, is composed of the same convolution kernel in the nonlinear mapping part, and presents a deep recursive layer up to 16 recursions. SR-ResNet [53], which is based on ResNet, is made up of 16 residual units. DRRN [54] is a proposed method in which basic residual units are rearranged in a recursive topology to form a recursive block. Enhanced deep super resolution network (EDSR) [55] removes the usage of batch normalization, resulting in high performance. SR-DenseNet [56], Memnet [57], and residual dense network (RDN) [58] have been proposed based on DenseNet. SR-DenseNet further concatenates all the features from different blocks before the deconvolution layer, which has been shown to improve performance effectively. Memnet uses residual units recursively to replace the normal convolutions in the block of the basic DenseNet and adds dense connections among different blocks. RDN uses residual units recursively to replace the normal convolutions in the block of the basic DenseNet. Deep back-projection networks (DBPNs) [59] use deep architectures to simulate iterative back-projection and further improves performance with dense connections. Zero shot super resolution (ZSSR) [60] is the first work to combine deep architectures with internal-example learning.

Natural language processing (NLP)

Natural language processing is a theory-motivated range of computational techniques for the automatic analysis and representation of human language [61]. The main use of NLP in radiology is the task of semantic classification in radiological reports. In general, NLP consists of two steps. The first step, called distributed representation, is representing each word, character, and sentence into vectors, and the second is treating the representations with a CNN.

The main advantage of the first step, distributional vectors, is that they capture similarity between words. Distributed representation is an internal representation of the observed data done in such a way that they are modeled as being explained by the interactions of many hidden factors. One particular factor learned from the configurations of others can often generalize well to new configurations. In introducing distributed representation into NLP, a character or word is represented by distributional vectors; for example, every color is represented by red, blue, and green vectors. This kind of vector composition also allows it to address the question "King − Man + Woman = ?" and arrive at the result "Queen". Distributed representation has some important frameworks, including Word2Vec [62], global vectors for word representation(GloVe) [63], and fastText [64, 65]. In Word2Vec, two concepts, continuous bag-of-words (CBOW) and skip-gram models, have been proposed to construct high-quality distributed vector representations efficiently. GloVe is another well-known word embedding method that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods. In addition, fastText improves the representation of words by using character-level information in morphologically-rich languages; it approaches the skip-gram method by representing words as bag-of-characters n-grams [66].

The second step is the use of deep learning. Some trials have applied CNNs to NLP. In the first reported application of deep learning to NLP [67], a simple CNN framework outperformed most state-of-the-art approaches in NLP. However, deep learning is naturally suitable for the use of recurrent architecture to deal with sentences, because words and sentences are dependent on context. Some networks have connoted recurrent neural networks (RNNs) [68], long short-term memory (LSTM) [69, 70], gated recurrent units (GRUs) [71], recursive neural networks [72], and neural Turing machines (NTMs) [73]. RNNs [68] are the most well-known networks; RNNs memorize previous computations and use this information in current processing. In practice,

however, these simple RNNs suffer from the infamous vanishing gradient problem, which makes it difficult to learn and tune the parameters of the earlier layers in the network. LSTM allows errors to back-propagate through an unlimited number of time steps. GRU involves lesser complexity and has empirically similar performance to LSTM. Attention [74] mechanisms encode a variable-length input sentence into a fixed-length vector. With an attention mechanism, it is no longer necessary to try to encode the full-source sentence into a fixed-length vector. Attention mechanisms enable the model to learn what to attend to based on the input sentence and what it has produced so far. Similar to RNNs, recursive neural networks are natural mechanisms that model sequential data. This is because language can be seen as a recursive structure where words and sub-phrases comprise other higher-level phrases in a hierarchy. NTMs are models that connect neural networks to external memory resources, which they can interact with by attentional processes. In addition, a memory-augmented network [75] was proposed to assimilate and leverage new data rapidly to make accurate predictions after only a few samples.

Transfer learning

Transfer learning is a machine learning technique where a model trained on one task is re-purposed on a second related task. It is a popular approach in deep learning, where pre-trained models are used as the starting point for computer vision and NLP tasks owing to the vast compute and time resources required for these problems by ANNs. In image recognition tasks, early layers extract more generic features such as edges, shapes, and textures [76]. Only the last one or two layers of a CNN extract more complex features and perform the most complex tasks of summarizing the vectorized image data into the classification data. Transfer learning uses these earlier layers of the CNN as a feature extractor in other networks.

Generative adversarial networks (GANs)

Generative adversarial networks [77] are a generative modeling approach based on differentiable generator networks. GANs provide a way to learn deep representations without extensively annotated training data. They are composed of two parts that update alternatively—a generator and a discriminator—and when the discriminator can no longer provide useful information to the generator, that is, when the outputs of the generator totally confuse the discriminator, the optimization procedure is completed.

Deep convolutional GANs (DCGANs) [78] allow the training of a pair of deep convolutional generator and discriminator networks. In image processing, pix2pix [79], CycleGAN [80], and GAN with XO-structure (XOGAN) [81] are used for image-to-image translation. The pix2pix networks learn not only the mapping from input image to output image, but also a loss function to train the mapping. CycleGAN represents an approach for learning to translate an image from a source domain to a target domain in the absence of paired examples. XOGAN provides one-to-many unsupervised image translation problems in which an input sample from one domain can correspond to multiple samples in the other. SRGAN, in which the generator of GAN is the SR-ResNet [53], as mentioned before, uses SR. It extends earlier efforts by adding an adversarial loss component that constrains images to reside on manifolds. In NLP, GAN is used to generate text. TextGAN [82] is a generic framework employing LSTM and CNN for adversarial training to generate realistic text. SeqGAN [83] trains a language model using policy gradients to train the generator to fool a CNN-based discriminator that discriminates between real and synthetic text. MaskGAN [84] employs an actor–critic training procedure on a task designed to provide rewards at every time step.

Neural architecture search (NAS)

Neural architecture search uses machine learning to automate the design of ANNs. The basic search algorithm is to propose a candidate model, evaluate it against a dataset, and then use the results as feedback to teach the NAS network. Designing effective neural network architectures is crucial for the performance of deep learning. While many impressive results have been achieved through significant manual architecture engineering, this process typically requires years of extensive investigation by human experts. Therefore, automatic architecture design has recently attracted much attention. Recently, NAS has achieved higher performance than manual architecture engineering. It is currently being applied to and improving the performance of every deep learning area, including a gradual introduction to medical imaging processing [85].

GeNet [86] is the first network to apply a genetic algorithm for exploring efficient neural network architectures. An encoding method to represent each network structure with a fixed-length binary string is proposed, and then some popular genetic operations, such as mutation and crossover, are applied to explore the search space efficiently. NASNet [87] uses a RNN to generate the model descriptions of neural networks and train the RNN with reinforcement learning to maximize the expected accuracy of the generated architectures. Efficient NAS (ENAS) [88] shares parameters among child models, which allows it to deliver strong empirical performances at much lower computing costs. Differentiable architecture search (DARTS) [89] is based on the continuous relaxation of architecture representation, allowing efficient searches of the architecture using gradient

descent. Neural architecture optimization (NAO) [90] optimizes network architectures by mapping them into a continuous vector space, where optimization is conducted via a gradient-based method.

Clinical overview

Numerous other studies on ANNs have been reported since the first by Asada et al. [11]. A PubMed search with regard to "neural network" and "deep learning" before October 23, 2018 found 94 papers in journals such as *Radiology*, *Investigative Radiology*, the *American Journal of Roentgenology*, the *American Journal of Neuroradiology*, *European Radiology*, the *Korean Journal of Radiology*, and the *Japanese Journal of Radiology*. Of these, 62 involved shallow learning and 32 deep learning. This review summarizes the 32 deep learning studies, which consist of 13 on classification, one on detection, six on semantic segmentation, five on image processing, and two on NLP. Five studies involve the combination of two or three deep learning techniques.

In this article, datasets can be roughly divided into three types. Training and validation datasets are used for developing algorithms. Developing datasets include both training and validation datasets. Training datasets comprise examples used for learning, namely to fit the parameters of a classifier, whereas validation datasets are used to tune the hyperparameters while providing an unbiased evaluation of the model fit to the training dataset. Finally, test datasets are used for evaluating the algorithm. These should be independent from developing datasets.

Classification

In total, 13 articles about classification in radiology were found in the PubMed search. Cicero et al. [12] used GoogLeNet to classify abnormalities (i.e., cardiomegaly, effusion, consolidation, edema, and pneumothorax) in frontal chest radiographs and develop an algorithm; the number of developing and testing data were 32,586 and 2443 images, respectively. The datasets were collected from a single institution. For the abnormal categories, the AUCs were 0.96 for pleural effusion, 0.87 for pulmonary edema, 0.85 for consolidation, 0.88 for cardiomegaly, and 0.86 for pneumothorax.

Lakhani et al. [91] evaluated the efficacy of CNNs for classifying tuberculosis in chest radiographs using both GoogLeNet and AlexNet. The number of developing and testing data were 857 (417 containing tuberculosis) and 150 (75 containing tuberculosis) images, respectively. The datasets were collected from four different datasets. The best-performing classifier, which was an ensemble [92] of the two networks, had an AUC of 0.99.

Prevedello et al. [93] developed two algorithms. First, they evaluated the performance of a deep learning algorithm for the classification of hemorrhage, mass effect, or hydrocephalus (HMH) in head computed tomography (CT) examinations. Second, they evaluated the performance of an algorithm for the detection of suspected acute infarct (SAI). GoogLeNet was used for the both algorithms. The HMH algorithm was developed by 246 developing examinations (containing 100 positive cases: 76, 13, and 11 cases of hemorrhage, mass effect, and hydrocephalus, respectively; and 146 negative cases: 22, 24, and 100 cases of SAI, encephalomalacia, and no critical findings, respectively) and 130 testing examinations (containing 50 positive cases: 38, 5, and 7 cases of hemorrhage, mass effect, and hydrocephalus; and 71 negative cases: 21, 9, and 50 cases of SAI, encephalomalacia, and no critical findings, respectively). The algorithm performance for HMH showed an AUC of 0.91. The SAI algorithm was developed by 71 developing examinations (containing 22 positive cases of SAI and 49 negative cases: 24 and 25 cases of encephalomalacia and no critical findings, respectively) and 49 testing examinations (containing 21 positive cases of SAI and 28 negative cases: 9 and 19 cases of encephalomalacia and no critical findings, respectively). The algorithm performance for SAI showed an AUC of 0.81. In that study, the dataset were collected from a single institution.

Kim et al. [94] evaluated the accuracy and efficiency of a new automatic software system for bone age assessment and validated its feasibility in clinical practice. However, the developing algorithm is unclear because the electronic supplement did not work. The algorithm was built using 18,940 images of left-hand radiographs from a single institution. The test dataset was collected from a different institution than the training one and was composed of 200 left-hand radiographs. They determined that the concordance rate was 70% between the automatic software system and the reference bone age. When the automatic software system was implemented as the second opinion in daily clinical practice, the radiologists' reading times were reduced by 29% on average, without compromising the accuracy of bone age estimations.

Yasaka et al. [95] investigated diagnostic performance using a deep learning method with a CNN for the differentiation of liver masses in dynamic contrast agent-enhanced CT images. Masses were diagnosed according to the following five categories: category A, classic hepato-cellular carcinomas (HCCs); category B, malignant liver tumors other than classic and early HCCs; category C, indeterminate masses or mass-like lesions, including early HCCs and dysplastic nodules, and rare benign liver masses other than hemangiomas and cysts; category D, hemangiomas; and category E, cysts. An architecture with six convolutional layers was applied. The numbers of developing and testing data were

1068 examinations (240, 121, 320, 207, and 180 cases for categories A, B, C, D, and E, respectively) and 100 examinations (21, 9, 35, 20, and 15 cases for categories A, B, C, D, and E, respectively). All datasets were collected by a single institution. The median accuracy for the differential diagnosis of liver masses for the test dataset was 0.84, and the median AUC-ROC for differentiating categories A–B from C–E was 0.92.

Larson et al. [96] compared the performance of a deep-learning bone age assessment model based on hand radiographs with that of expert radiologists. ResNet was used for the algorithm. The algorithm was built using 14,036 images of left-hand radiographs from two institutions. The test datasets were collected from the same institution as the developing data and a different institution than the developing dataset. They were composed of 200 and 913 left-hand radiographs, respectively. The mean difference between the bone age estimates of the model and reviewers was 0 years, with root mean square and median absolute deviations of 0.63 and 0.50 years, respectively.

Yasaka et al. [97, 98] reported two additional studies about deep learning. One was the application of deep learning to liver fibrosis in magnetic resonance (MR) imaging [98] and the other was in CT imaging [97]. First, they investigated the performance of a CNN for the staging of liver fibrosis using gadoxetic acid-enhanced hepatobiliary phase MR imaging. An architecture with three convolutional layers was applied. The developing and test datasets comprised 534 examinations (54, 53, 81, 113, and 233 patients with fibrosis stages F0, F1, F2, F3, and F4, respectively) and 100 examinations (10, 10, 15, 20, and 45 patients with fibrosis stages F0, F1, F2, F3, and F4, respectively), respectively. All datasets were collected by a single institution. Fibrosis stages F4, F3, and F2 were diagnosed with AUCs of 0.84, 0.84, and 0.85, respectively. The Spearman rank correlation coefficient was 0.63. Second, they investigated whether liver fibrosis could be staged by deep learning techniques based on CT images. An architecture with four convolutional layers was applied. The developing and test datasets comprised 396 examinations (113, 36, 56, 66, and 125 patients with fibrosis stages F0, F1, F2, F3, and F4, respectively) and 100 examinations (29, 9, 14, 16, and 32 patients with fibrosis stages F0, F1, F2, F3, and F4, respectively), respectively. All datasets were collected by a single institution. The AUCs for diagnosing significant fibrosis (≥F2), advanced fibrosis (≥F3), and cirrhosis (F4) using F-scores were 0.74, 0.76, and 0.73, respectively. The Spearman rank correlation coefficient was 0.48.

Noguchi et al. [99] developed algorithms to classify head MR sequences and applied pretrained AlexNet and GoogLeNet. They enrolled 78 patients with mild cognitive impairment (MCI) having apparently normal head MR images and 78 patients with intracranial hemorrhage (ICH) having morphologically deformed head MR images. They used 1872 images of the first proximal and middle slices of six MR sequences; 468 of these images were then randomly split into an 82%:9%:9% ratio for training (384 images), validation (42 images) and testing (42 images). The pretrained AlexNet had accuracies of 73, 74, 73, and 61% in the middle slices of the MCI group, middle slices of the ICH group, first slices of the MCI group, and first slices of the ICH group, respectively, while the pretrained GoogLeNet had accuracies of 100, 98, 93, and 95%, respectively.

England et al. [100] developed a deep learning algorithm to diagnose traumatic pediatric elbow effusion on lateral radiographs using DenseNet. The algorithm had an AUC of 0.94 on the test set. The developing dataset comprised 772 images (582 and 190 patients with no-effusion and effusion, respectively) and the test dataset comprised 129 examinations (96 and 33 patients with no-effusion and effusion, respectively). These datasets were collected from a single institution.

Kim et al. [101] compared the diagnostic performance of a deep learning algorithm developed using ResNet with that of radiologists in diagnosing maxillary sinusitis on Waters' view radiographs. The developing dataset contained 9000 radiographs. Two test datasets were prepared, one containing 140 radiographs from the same institution as the developing dataset, and the other containing 200 radiographs from a different institution. The AUCs of the deep learning algorithm were 0.83–0.89 and 0.75–0.84 for the test sets from the same and different institutions, respectively, while the AUCs of the radiologists were 0.83–0.89.

Lehman et al. [102] developed a deep learning algorithm using a ResNet model to assess mammographic breast density. In their retrospective study, a deep CNN was trained to assess breast density based on the original interpretation of 41,479 digital screening mammograms by an experienced radiologist. The resulting algorithm was tested on a held-out test set of 8677 cases. In addition, five radiologists performed a reader study on 500 mammograms randomly selected from the test dataset. Finally, the algorithm was implemented in routine clinical practice, where eight radiologists reviewed 10,763 consecutive mammograms assessed using the model. All datasets came from a single institution. The model showed good agreement with radiologists in the test dataset ($\kappa = 0.67$) and with radiologists in consensus in the reader study dataset ($\kappa = 0.78$). There was also very good agreement ($\kappa = 0.85$) with radiologists in the clinical implementation dataset.

Ueda et al. [103] developed and evaluated a supportive algorithm using deep learning for detecting cerebral aneurysms using time-of-flight MR angiography to provide a second assessment of images already interpreted by radiologists. ResNet was used to distinguish true aneurysms. The algorithm detected cerebral aneurysms in radiological

reports with a sensitivity of 91–93% and improved aneurysm detection compared with initial reports by 4.8–13%. They prepared one dataset from two institutions for training and two datasets from two institutions for the tests. The numbers of aneurysms were 683 for the training dataset, 592 for one test dataset, and 74 for another test dataset.

Object detection

Only one article about detection in radiology was found in the PubMed search. Chang et al. [104] evaluated a CNN optimized for the detection and quantification of intraparenchymal, epidural/subdural (EDH/SDH), and subarachnoid hemorrhages (SAH) on non-contrast CT images. That CNN was based on mask R-CNN [25]. Their model made it easier to understand and deal with object detection tasks, but it included not only detection, but also segmentation. The developing and test datasets comprised 10,159 examinations, including 901 hemorrhage cases, and 682 examinations, including 92 hemorrhage cases, respectively. The datasets were collected from a single institution. The AUC of all ICH cases for the prospective test dataset was 0.95. The mean validation Dice similarity coefficients for intraparenchymal hemorrhage, EDH/SDH, and SAH were 0.93, 0.86, and 0.77, respectively.

Semantic segmentation

Six studies were found on semantic segmentation in radiology. Becker et al. [105] applied deep learning to mammography to distinguish malignant from benign and normal lesions. The name of the network was "ViDi Red", but its detailed architecture remains unclear. They developed two algorithms. In study 1, the numbers of developing and test data were 286 examinations, including 143 malignant cases, and 70 examinations, including 35 malignant cases, respectively. The datasets were collected from different institutions. In study 2, the numbers of developing and test datasets comprised 895 examinations, including 125 malignant cases, and 251 examinations, including 18 malignant cases, respectively. The datasets were collected from a single institution. That study showed that deep learning algorithms designed for generic image analysis can be trained to assess breast cancers from mammography data with AUCs of 0.79 in study 1 and 0.82 in study 2.

Norman et al. [106] analyzed the use of CNNs for automated cartilage and meniscus segmentation in knee MR imaging. The network used was U-net [107]. U-net was proposed for the medical imaging community in the same period as SegNet [35]. U-net transfers the entire feature map to corresponding decoders and then concatenates them to upsampled decoder feature maps; unlike SegNet, it does not reuse pooling indices. Two algorithms were developed: one for three-dimensional double-echo steady-state (3D-DESS) image sequences and the other for T1rho-weighted image sequences. The 3D-DESS and T1rho-weighted sequence examinations were collected by different institutions. For its development, 415 examinations of a 3D-DESS sequence from a single institution and 158 examinations of T1rho-weighted sequence from a single institution were used. Next, algorithms were developed for the 3D-DESS and T1rho-weighted sequences. For testing, 16 examinations of 3D-DESS sequences from a single institution and 158 examinations of T1rho-weighted sequences from a single institution were used. The 3D-DESS and T1rho sequences were used to test the algorithms separately. The Dice coefficients for predicting overall cartilage and meniscus in the DESS dataset were 0.87 and 0.83, respectively, while those for cartilage and meniscus in the T1rho-weighted dataset were 0.74 and 0.77, respectively.

Perkuhn et al. [108] evaluated a deep learning-based, automatic glioblastoma tumor segmentation algorithm and compared the results with ground truth, manual expert segmentation; Laukamp et al. [109] evaluated meningiomas in the same way. Both studies used the deep learning architecture of a pretrained DeepMedic network [110], which won the Ischemic Stroke Lesion Segmentation Challenge competition in 2015. DeepMedic was developed by applying a 3D network to a multi-scale CNN with CRFs and a pretrained model using 249 gliomas. Perkuhn et al. [108] and Laukamp et al. [109] prepared 62 glioblastomas and 56 meningiomas for testing, respectively. Each dataset was collected by different institutions. The former and latter algorithms achieved Dice coefficients of 0.86 and 0.81 for the whole tumor, respectively.

Montoya et al. [111] developed a deep learning angiography method using architecture based on ResNet to generate 3D cerebral angiograms from a single contrast-enhanced C-arm cone-beam CT image to reduce image artifacts and radiation dose. The developing and test dataset comprised 43 and eight examinations, respectively. The datasets were collected by a single institution. The Dice coefficient was 0.98 for vascular segmentation. No residual signal from osseous structures was observed for any of the 3D deep learning angiography testing cases, except for small regions in the optic capsule and nasal cavity, compared with 37% of the 3D rotational angiographies.

Tao et al. [112] applied U-net to develop a deep learning-based method for the fully automated quantification of left ventricular function from short-axis cine MR images and evaluate its performance in a multivendor and multicenter setting. The cine MR datasets were obtained from three major MR vendors in four medical centers. The developing and test datasets contained 400 and 196 examinations, respectively. The Dice coefficients were 0.88, 0.95, and 0.92

and 0.91, 0.96, and 0.94 in the apex, middle, and base of the endocardium and epicardium, respectively.

Image processing

Five papers on image-processing in radiology were found, including five on SR. Liu et al. [113] developed and evaluated the feasibility of deep learning approaches for MR imaging-based attenuation correction in brain positron emission tomography/MR imaging. SegNet was applied to convert MR to pseudo-CT images. The developing and test datasets comprised 30 and 10 examinations of a T1-weight image (T1WI) sequence. The datasets were collected by a single institution. The algorithm provided an accurate pseudo-CT scan with Dice coefficients of 0.97 for air, 0.94 for soft tissue, and 0.80 for bone.

Kim et al. [114] developed a deep learning algorithm based on DeepMedic that generates arterial spin labeling (ASL) perfusion images with higher accuracy and robustness using a smaller number of subtraction images. First, they adopted Hadamard-encoded pseudo-continuous ASL. The developing and test datasets comprised 114 and 26 examinations, respectively. Second, they adopted 3D pseudo-continuous ASL and cross-validation methods. The developing data comprised seven examinations; no test dataset was used. The datasets were collected by a single institution. The mean square errors were approximately 40% lower than those of the conventional averaging method for cross-validation with healthy subjects and patients and a separate test with patients who had experienced a stroke.

Ahn et al. [115] compared observer preferences for image quality and radiation doses between non-grid, grid-like, and grid images. They adopted SimGrid, which converts non-grid to grid-like images. SimGrid allows the distribution and degree of scatter radiation to be estimated using raw image data directly with a pretrained CNN; however, the development method remains unclear. The test dataset comprised 38 examinations. The dataset was collected in a single institution. Radiologists significantly preferred grid-like to non-grid images ($p < 0.001$).

Chen et al. [116] developed a deep learning reconstruction approach to improve the reconstruction speed and quality of highly undersampled variable-density single-shot fast spin-echo imaging using a variational network (VN) [117], which applied a variational method with deep learning. Next, they compared the ability of the VN with conventional parallel imaging and compressed sensing (PICS) reconstruction. The developing and test datasets comprised 130 and 27 examinations, respectively. The datasets were collected by a single institution. Compared with conventional PICS, the VN achieved improved the perceived signal-to-noise ratio ($p = 0.01$) and sharpness ($p = 0.001$), with no difference in image contrast ($p = 0.24$) or residual artifacts ($p = 0.07$). In terms of overall image quality, the VN performed significantly better than PICS ($p = 0.02$). The average reconstruction times (within six standard deviations) were 5.6 s per section for PICS and 0.19 s per section for VN.

Jiang et al. [118] tested whether a deep learning-based denoising CNN (DnCNN) method [119] could robustly denoise 3D MR images with Rician noise. The developing dataset comprised 20 examinations of a T1WI sequence from a single institution. To evaluate this method, three datasets from different institutions, including T1WI acquired at 1.5 and 3 T, as well as MR images simulated with a widely used MR simulator, were randomly selected and artificially added with different noise levels ranging from 1 to 15%. For comparison, four other denoising methods (Coupe-Block, WSM, ODCT3D, and PRI-NLM3D) were also tested using these datasets. DnCNN for a specific noise level showed the most robust denoising performance in all three datasets in terms of the highest perceived signal-to-noise ratio and global structure similarity index. The general noise-applicable model also performed better than the other four methods for the two datasets.

Natural language processing (NLP)

Two studies on NLP in radiology were found. Chen et al. [120] evaluated the performance of a CNN compared with a traditional NLP model in extracting pulmonary embolism (PE) findings from thoracic CT reports. The CNN model was developed to classify three categories: presence of PE (present or absent), chronicity of PE (acute or chronic), and location of PE (central or subsegmental). If a PE was present in the report, then it was considered a positive study for PE. They applied GloVe for word-embedding and then used a simple CNN with one layer of convolution [121]. The developing data of 3512 reports contained 1407 PE-positive reports, of which 1232 were acute and 250 were subsegmental. The testing data of 1000 reports contained 40 PE-positive reports, of which 34 were acute and 11 were subsegmental. The testing data of 859 reports contained 293 PE-positive reports, of which 261 were acute. The test dataset did not distinguish between central or subsegmental location. In the test dataset, the model correctly predicted all three classification dimensions with an accuracy of 99% and an F1 score of 0.94. In the test dataset, the model correctly predicted both classification dimensions with an accuracy of 92% and an F1 score of 0.89.

Zech et al. [122] compared different methods for generating features from radiology reports and developed a method to automatically identify the findings in these reports applying CBOW. The best-performing model had a held-out AUC of 0.97 for identifying the presence of any critical head CT findings and an average 0.96 AUC across all head CT findings.

Combination of deep learning techniques

Five studies combining deep learning techniques were found. Chang et al. [123] developed a deep learning algorithm to classify genetic mutations in gliomas. They first used a DeepMedic algorithm for the segmentation of gliomas. Second, they trained ResNet to predict the underlying molecular genetic mutation status in gliomas. The glioma dataset containing 259 examinations was applied to the segmentation. The processed dataset was then used for sequential development. The mean tumor size determined by the automated segmentation tasks was 105.6 cm^3. The classification had high accuracy: *IDH1* mutation status, 94%; 1p/19q codeletion, 92%; and *MGMT* promotor methylation status, 83%. However, the results were calculated using only a validation dataset; no independent test dataset was used.

Liu et al. [124] determined the feasibility of using a deep learning approach to detect cartilage lesions (including softening, fibrillation, fissuring, focal defects, diffuse thinning due to cartilage degeneration, and acute cartilage injury) within the knee joint in MR images. They used SegNet [35] for segmentation and VGGNet [18] with fine tuning by SegNet for classification. In the segmentation task, 175 examinations outlining cartilage by a radiologist were used to develop an algorithm applying a three-fold cross-validation method and no test dataset. The architecture provided good segmentation of the femur, tibia, femoral cartilage, and tibial cartilage. The mean Dice coefficients were 0.96 for the femur, 0.95 for the tibia, 0.81 for the femoral cartilage, and 0.82 for the tibial cartilage. In the classification task, based on the dataset used in the segmentation task, 16,075 small image patches were prepared for developing, including 1982 of cartilage, and 1320 small image patches were prepared for test, including 660 of cartilage. All images were collected by a single institution. The AUC of the cartilage lesion detection system was 0.91–0.92.

Choi et al. [125] developed and validated a deep learning system for staging liver fibrosis using CT images of the liver. They used U-net [107] for segmentation and the NAS model (GeNet [86] or NEMO [126]) for classification. In the segmentation task, 50 examinations outlining the liver by a radiologist were used to develop an algorithm applying a fivefold cross-validation method. The segmentation architecture provided a Dice coefficient of 0.92. In the classification task, 7461 examinations (3357, 113, 284, 460, and 3247 cases in categories F0, F1, F2, F3, and F4, respectively) and 891 examinations (118, 109, 161, 173, and 330 cases in categories F0, F1, F2, F3, and F4, respectively) outlining the liver by the developed algorithm were prepared for developing and tests. The developing dataset was collected from a single institution and the test dataset from five different institutions. The algorithm achieved staging ROCs of 0.96, 0.97, and 0.95 for diagnosing significant fibrosis (F2–F4), advanced fibrosis (F3–F4), and cirrhosis (F4), respectively.

Nam et al. [127] developed and validated a deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. They used a model based on ResNet for classification and segmentation. The algorithm was trained in a semi-supervised learning manner using all of the image-level labels, but only part of the pixel-level annotations in the training dataset: 37% (3213 of 8625) of nodule chest radiographs underwent pixel-level annotation. Next, 300 examinations outlining the nodule were used for the internal validation dataset. There were four external test datasets, including 181 examinations (119 nodules), 182 examinations (123 nodules), 181 examinations (111 nodules), and 149 examinations (89 nodules). The AUC-ROC and jackknife alternative free-response ROC of the algorithm ranged between 0.92–0.99 and 0.83–0.92, respectively.

Liang et al. [128] developed a fully automated deep learning-based method using CT images of the neck and investigated the model performance in automated detection and segmentation. Their model was based on Faster R-CNN. After using Faster R-CNN, an FCN was adopted for semantic segmentation. A total of 185 subjects from a single institution were included in the study and divided into developing and test datasets. The number of datasets is unclear. The model provides an accurate detection result with a sensitivity of 1.00 for most organs and a specificity of 0.98–1.00. Furthermore, segmentation results from the model correlated strongly with manual segmentation with a Dice coefficient of more than 0.85 in most organs.

Quality proof

Deep learning has been applied to clinical applications in radiology. The performance of these applications has achieved almost the same or even higher performance compared with radiologists. The use of clinical applications created by deep learning in medicine is an issue of great interest. Like every type of approved medical equipment, the applications developed by deep learning should be evaluated with prudent methods considering the peculiar characteristics of deep learning. The most important and interesting feature of deep learning is that it can extract features from developing data on its own. On one hand, it enables researchers to develop algorithms using only prepared data. On the other hand, it more or less fits to the characteristics in the developing data. More training data provide deep learning applications with more robust and higher performances, and vice versa. It is said that deep learning algorithms will generally achieve acceptable performance, with around 5000 labeled examples per category, and will match or exceed human performance when trained with a dataset

containing at least ten million labeled examples [3]. Most of the applications created by deep learning in radiology do not have plenty of data compared with general image recognition. However, unlike general images, it is not easy to obtain medical data from patients. Especially in radiology, data preparation means extracting hundreds of thousands of images, making a bounding box or segmentation one-by-one, and preparing the cases' demographic characteristics, including not only sex and age, but also the size or grade of diseases. This preparation requires substantial amounts of both time and effort. The lack of sufficient training data causes variations in accuracy between different cohort datasets. The most important process in the quality certification of applications based on deep learning is the prudent examination of test and training datasets.

Test dataset

As stated previously in the section of clinical overview, there are generally three datasets used for the development of deep learning algorithms. Training and validation datasets are used for developing the algorithms, and a test dataset is used to assessing their versatility. In principle, each dataset has no overlap. Algorithms developed by a lack of a test dataset have no proof of versatility. Park et al. [129] emphasized the use of a test dataset composed of newly recruited patients (referred to as a temporal test dataset), and a test dataset collected by independent investigators at a different site (referred to as a geographic dataset). Test datasets may also include a small fraction of data that were randomly split from the entire dataset and kept untouched for use as a test dataset while the main portion of the data was used for training (referred to as a split-sample test dataset). From the clinical standpoint, testing with the use of a temporal or geographic dataset is preferred to split-sample tests to confirm the generalizability of a diagnostic or predictive model for clinical practice. Split-sample validation may address the internal validity of a model, but does not accurately assess its generalizability. Geographic tests are also helpful for evaluating the generalizability of deep learning models

Fig. 6 Examples of overlooked aneurysms [103]. The images on the left are time-of-flight magnetic resonance angiography axial source images, and those on the right are maximum intensity projection images. **a** 15-mm (in long-axis diameter) aneurysm (arrowhead) at the bifurcation of the middle cerebral artery of a 67-year-old woman. **b** 10-mm (in long-axis diameter) aneurysm (arrowhead) in the vertebral artery of a 30-year-old woman

Table 1 The types of test datasets used for classification, object detection, and segmentation

Citation	Target		Modality	The deep learning model	The deep learning category	Development/evaluation	The type of test dataset
[12]	Abnormality		CR	GoogLeNet	Classification	Both	Split-sample test
[91]	Tb		CR	AlexNet, GoogLeNet	Classification	Both	Split-sample test[a]
[93]	Emergency	HMH	CT	GoogLeNet	Classification	Both	Split-sample test
		SAI	CT	GoogLeNet	Classification	Both	Split-sample test
[94]	Bone age		CR	Unknown	Classification	Evaluation	Geographic test
[95]	Liver mass		CT	Six convolutions	Classification	Both	Split-sample test
[96]	Bone age		CR	ResNet	Classification	Both	Geographic test
[97]	Liver fibrosis		CT	Four convolutions	Classification	Both	Split-sample test
[98]	Liver fibrosis		MR	Three convolutions	Classification	Both	Split-sample test
[99]	Image classification		MR	AlexNet, GoogLeNet	Classification	Both	Split-sample test
[100]	Elbow effusion		CR	DenseNet	Classification	Both	Split-sample test
[101]	Sinusitis		CR	ResNet	Classification	Both	Geographic test
[102]	Mammo density		CR	ResNet	Classification	Both	Temporal test
[103]	Aneurysms		MR	ResNet	Classification	Both	Geographic test
[104]	Hemorrhage		CT	mask R-CNN	Detection	Both	Temporal test
			CT	mask R-CNN	Segmentation	Both	Temporal test
[105]	Breast cancer	study1	MG	Unknown (ViDi Red)	Segmentation (not clear)	Both	Geographic test
		study2	MG	Unknown (ViDi Red)	Segmentation (not clear)	Both	Temporal test
[106]	Knee	DESS	MR	U-net	Segmentation	Both	Split-sample test
		T1rho	MR	U-net	Segmentation	Both	Split-sample test
[108]	Glioblastoma		MR	DeepMedic	Detection	Evaluation	Geographic test
			MR	DeepMedic	Segmentation	Evaluation	Geographic test
[109]	Meningioma		MR	DeepMedic	Detection	Evaluation	Geographic test
			MR	DeepMedic	Segmentation	Evaluation	Geographic test
[111]	Angiography		CT	ResNet	Segmentation	Both	Split-sample test
[112]	Left ventricular		MR	U-net	Segmentation	Both	Geographic test
[123]	Glioma		MR	DeepMedic	Segmentation	Evaluation	Geographic test
			MR	ResNet	Classification	Development	Cross-validation only
[124]	Cartilage		MR	SegNet	Segmentation	Both	Cross-validation only
			MR	VGG	Classification	Both	Split-sample test
[125]	Liver fibrosis		CT	U-net	Segmentation	Both	Cross-validation only
			CT	GeNet or NEMO	Classification	Both	Geographic test
[127]	Lung cancer		CR	ResNet	Segmentation	Both	Cross-validation only
			CR	ResNet	Classification	Both	Geographic test
[128]	Neck anatomy		CT	Faster R-CNN	Detection	Both	Split-sample test

[a]There is no independent test dataset but both developing and test datasets were collected from several institutions

that involve radiologic imaging across technical variations, because the same imaging examination may be performed with slightly different technical parameters at different sites. This procedure is crucial for avoiding an overestimation of the performance as a result of overfitting in a high-dimensional or overparameterized classification model and spectrum bias.

Training dataset

In almost all studies, the fewer the cases in the training dataset, the less the accuracy, sensitivity, or AUC. For example, examining correlations in the study by Chang et al. [104], the AUCs for EDH/SDH and SAH gradually decreased in accordance with the size reduction of the hematoma. Two hypotheses arise from this example. First, smaller lesions are

Table 2 Correlations between the number of developing data and sensitivities

Citation	Target	Category	No. of developing data	Sensitivity
[95]	Liver mass	A (classic HCCs)	240/1068	0.71
		B (malignant liver tumors other than classic and early HCCs)	121/1068	0.33
		C (indeterminate masses or mass-like lesions)	320/1068	0.94
		D (hemangiomas)	207/1068	0.90
		E (cysts)	180/1068	1.00
[97]	Liver fibrosis (CT) by Yasaka	Significant fibrosis (F2–F4)	247/396	0.76
		Advanced fibrosis (F3–F4)	191/396	0.75
		Chirrhosis (F4)	125/396	0.75
[98]	Liver fibrosis (MR)	Significant fibrosis (F2–F4)	427/534	0.84
		Advanced fibrosis (F3–F4)	346/534	0.78
		Chirrhosis (F4)	233/534	0.76
[103]	Aneurysm	Internal carotid artery area	444/748	0.94
		Middle cerebral artery area	90/748	0.86
		Anterior cerebral artery area	149/748	0.91
		Posterior artery area	26/748	1.00
		Basilar artery area	25/748	0.93
		Vertebral artery area	14/748	0.79
[104]	IPH	Large	192/358	1.00
		Medium	88/358	1.00
		Small	63/358	1.00
		Punctate	15/358	1.00
	EDH/SDH	Large	188/319	1.00
		Medium	79/319	0.93
		Small	49/319	0.75
		Punctate	3/319	NA
	SAH	Large	85/224	1.00
		Medium	53/224	1.00
		Small	52/224	0.83
		Punctate	34/224	0.67
[125]	Liver fibrosis (CT) by Choi	Significant fibrosis (F2–F4)	3991/7461	0.96
		Advanced fibrosis (F3–F4)	3707/7461	0.95
		Chirrhosis (F4)	3247/7461	0.85

difficult to diagnose for deep learning algorithms, similar to physicians. Second, the lack of the training data causes the decrease in the AUC. In the study by Ueda et al. [103] the algorithm overlooked larger aneurysms (Fig. 6), for which there were not enough training data; this supports the latter hypothesis. To estimate the weakness of the algorithm, it is, therefore, important to evaluate training datasets and reveal imbalances in the prevalence of data. Clinical applications developed by deep learning in radiology are summarized from these perspectives in Tables 1 and 2 and Fig. 7. Table 1 shows the types of test datasets in medical classification, detection, and segmentation applications. The most reliable test dataset is the geographic test dataset. In addition, the temporal test dataset is trustworthy. Table 2 summarizes the category classifications (not binary classifications), and Fig. 7 shows the correlation between the number of training data and sensitivities. A clear association can be seen between the number of training data and sensitivities. The higher the number of training data, the higher the sensitivity of the algorithms.

Ultimately, clinical verification of the diagnostic or predictive accuracy of every medical application (not only those developed by deep learning) requires a demonstration of an effect on patient outcomes [129].

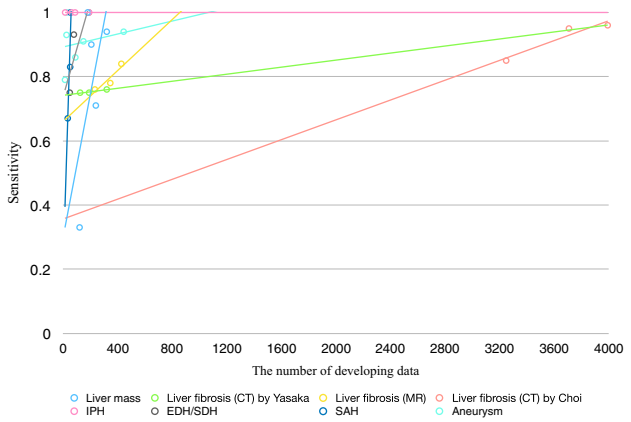

Fig. 7 Correlations between the number of developing data and sensitivities. Each line is a trendline corresponding to the dots that have the same color

Future of deep learning in medicine

Deep learning is a state-of-the-art technique currently being applied in the field of radiology. Deep learning emerged as the preferred machine learning approach in machine perception. As mentioned previously, deep learning techniques are divided into the following five categories in radiology: classification, object detection, segmentation, image processing, and NLP. In each area, deep learning has been performing increasingly better through integrating with existing methods and maximizing performance. The use of deep learning for clinical applications is an issue of great interest. Before using deep learning algorithms, it is important that they are assessed using various test datasets, and also that the training datasets are evaluated to reveal imbalances in the data distribution.

One recent study published in *Nature* [130] reported that deep learning applications can provide precise predictions about mortality rates, readmission rates, and prolonged hospital stays. Using the deep learning techniques, NLP and classification, to analyze all the patient data in electronic health records achieved high accuracy for tasks such as predicting in-hospital mortality (AUC-ROC of 0.93–0.94), 30-day unplanned readmission (AUC-ROC of 0.75–0.76), prolonged length of stay (AUC-ROC of 0.85–0.86), and a final discharge diagnoses (frequency-weighted AUC-ROC of 0.90). All patient data, including laboratory data, vital signs, and radiological reports, and even clinical handwriting, was processed in that study. Deep learning has been changing and improving not only radiology, but also all other areas of medicine. The introduction of deep learning applications into real clinical situations to detect and classify disease lesions in radiographs is expected in the near future. Moreover, deep learning applications are expected to be applied to the analysis of health records to predict patients' conditions.

As the use of artificial intelligence (AI)-assisted imaging diagnosis spreads in the future, what changes will it bring to the work of radiologists? Answering this question is not straightforward because it is a type of future prediction that will be influenced greatly by a number of complex factors. The answer will also vary from country to country because it will largely depend on medical and legal systems in each country. Therefore, here we present a scenario that is likely to occur in Japan. It has been reported by the news media that the Japanese government is planning to prepare inclusive rules for the use of AI in medical practice [131]. These regulations stipulate that physicians should bear the ultimate responsibility for AI-supported diagnosis and treatment planning [131]. Under such circumstances, it is estimated that radiologists will judge the suitability of AI for imaging diagnosis using modalities such as CT and MR imaging and will be responsible for the final decisions. Therefore, radiologists will be required to have the skills necessary to deal with AI, as well as adequate knowledge regarding the strengths and weaknesses of AI in imaging diagnosis. In addition, since ground truths are not shown in the judgment by AI using deep learning, it would be necessary for radiologists to be capable of showing a scientific basis in assessing the suitability of AI judgments. The workload of radiologists in Japan is an issue of concern [132–134]. The use of AI for imaging diagnosis in Japan could be expected to lead to a reduction in the workloads of radiologists in the future. Radiologists will need to have a solid understanding and sincere appreciation of deep learning in the AI era, and they will need to carefully consider the contributions to medical development that will be made possible by deep learning as an initial conductor.

Funding Another research about a deep learning for mammography received 10,000$ in 2017 from Wellness Open Living Labs. LLC, Osaka, Japan.

Compliance with ethical standards

Conflict of interest Daiju Ueda received a research grant from Wellness Open Living Labs, LLC.

Ethical considerations This article does not contain any research involving human participants or animals performed by any of the authors.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436.
2. Deng L, Yu D. Deep learning: methods and applications. Foundations and Trends®. Signal Processing. 2014;7:197–387.
3. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. Cambridge: MIT Press; 2016.

4. Hebb DO. The organization of behavior: a neurophysiological approach. New York: Wiley; 1949.
5. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5:115–33.
6. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65:386.
7. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323:533.
8. Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. In: Advances in neural information processing systems. 2007. p. 153–60.
9. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18:1527–54.
10. Poultney C, Chopra S, Cun YL. Efficient learning of sparse representations with an energy-based model. In: Advances in neural information processing systems. 2007. p. 1137–44.
11. Asada N, Doi K, MacMahon H, et al. Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases: pilot study. Radiology. 1990;177:857-60.
12. Cicero M, Bilbily A, Colak E, 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.
13. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 1–9.
14. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. Jpn J Radiol. 2018;36(4):257–72.
15. Fukushima K, Miyake S. Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recogn. 1982;15:455–69.
16. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol. 1962;160:106–54.
17. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012. p. 1097–105.
18. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556.
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–8.
20. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: CVPR. 2017. p. 3.
21. Zhao Z-Q, Zheng P, Xu S-t, Wu X. Object detection with deep learning: a review. 2018. arXiv:1807.05511.
22. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 580–7.
23. Girshick R. Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. 2015. p. 1440–8.
24. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. 2015. p. 91–9.
25. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: IEEE transactions on pattern analysis and machine intelligence. 2018.
26. Erhan D, Szegedy C, Toshev A, Anguelov D. Scalable object detection using deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 2147–54.
27. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 779–88.
28. Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector. In: European conference on computer vision. Springer; 2016. p. 21–37.
29. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: IEEE transactions on pattern analysis and machine intelligence. 2018.
30. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 3431–40.
31. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell. 2018;40:834–48.
32. Lin G, Milan A, Shen C, Reid ID. RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: Cvpr. 2017. p. 5.
33. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: IEEE conf on computer vision and pattern recognition (CVPR). 2017. p. 2881–90.
34. Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Computer vision and pattern recognition workshops (CVPRW), 2017 IEEE conference. IEEE; 2017. p. 1175–83.
35. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. 2015. arXiv:1511.00561.
36. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science; 1985.
37. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J. A review on deep learning techniques applied to semantic segmentation. 2017. arXiv:1704.06857.
38. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected crfs. 2014. arXiv:1412.7062.
39. Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. 2015. p. 2650–8.
40. Liu W, Rabinovich A, Berg AC. Parsenet: Looking wider to see better. 2015. arXiv:1506.04579.
41. Pinheiro PO, Lin T-Y, Collobert R, Dollár P. Learning to refine object segments. In: European conference on computer vision. Springer; 2016. p. 75–91.
42. Krähenbühl P, Koltun V. Parameter learning and convergent inference for dense random fields. In: International conference on machine learning. 2013. p. 513–21.
43. Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in neural information processing systems; 2011. p. 109–17.
44. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. 2015. arXiv:1511.07122.
45. Yang T, Wu Y, Zhao J, Guan L. Semantic segmentation via highly fused convolutional network with multiple soft cost functions. Cognit Syst Res. 2018. arXiv:1801.01317
46. Park SC, Park MK, Kang MG. Super-resolution image reconstruction: a technical overview. IEEE Signal Process Mag. 2003;20:21–36.
47. Dong C, Loy CC, He K, Tang X. Image super-resolution using deep convolutional networks. IEEE Trans Pattern Anal Mach Intell. 2016;38:295–307.
48. Dong C, Loy CC, He K, Tang X. Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. Springer; 2014. p. 184–99.

49. Dong C, Loy CC, Tang X. Accelerating the super-resolution convolutional neural network. In: European conference on computer vision. Springer; 2016. p. 391–407.
50. Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 1874–83.
51. Kim J, Kwon Lee J, Mu Lee K. Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 1646–54.
52. Kim J, Kwon Lee J, Mu Lee K. Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 1637–45.
53. Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR. 2017. p. 4.
54. Tai Y, Yang J, Liu X. Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 5.
55. Lim B, Son S, Kim H, Nah S, Lee KM. Enhanced deep residual networks for single image super-resolution. In: The IEEE conference on computer vision and pattern recognition (CVPR) workshops. 2017. p. 4.
56. Tong T, Li G, Liu X, Gao Q. Image super-resolution using dense skip connections. In: Computer vision (ICCV), 2017 IEEE international conference. IEEE; 2017. p. 4809–17.
57. Tai Y, Yang J, Liu X, Xu C. Memnet: A persistent memory network for image restoration. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 4539–47.
58. Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y. Residual dense network for image super-resolution. In: The IEEE conference on computer vision and pattern recognition (CVPR). 2018.
59. Haris M, Shakhnarovich G, Ukita N. Deep backprojection networks for super-resolution. In: Conference on computer vision and pattern recognition. 2018.
60. Shocher A, Cohen N, Irani M. Zero-Shot" super-resolution using deep internal learning. In: Conference on computer vision and pattern recognition (CVPR). 2018.
61. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. 2017. arXiv :1708.02709.
62. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013. p. 3111–9.
63. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. p. 1532–43.
64. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. 2016. arXiv:1607.04606.
65. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. 2016. arXiv:1607.01759.
66. Shannon CE. A mathematical theory of communication. In: ACM SIGMOBILE mobile computing and communications review, vol. 5. 2001. p. 3–55.
67. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Mach Learn Res. 2011;12:2493–537.
68. Elman JL. Finding structure in time. Cognit Sci. 1990;14:179–211.
69. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735–80.
70. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. 1999.
71. Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014. arXiv:1406.1078.
72. Goller C, Kuchler A. Learning task-dependent distributed representations by backpropagation through structure. Neural Netw. 1996;1:347–52.
73. Graves A, Wayne G, Danihelka I. Neural turing machines. 2014. arXiv:1410.5401.
74. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2014. arXiv:1409.0473.
75. Santoro A, Bartunov S, Botvinick M, Wierstra D, Lillicrap T. Meta-learning with memory-augmented neural networks. In: International conference on machine learning. 2016. p. 1842–50.
76. Hertel L, Barth E, Käster T, Martinetz T. Deep convolutional neural networks as generic feature extractors. In: Neural networks (IJCNN), 2015 international joint conference. IEEE; 2015. p. 1–4.
77. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Advances in neural information processing systems. 2014. p. 2672–80.
78. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. 2015. arXiv:1511.06434.
79. Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. 2017. arXiv:1611.07004
80. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017. arXiv:1703.10593
81. Zhang Y. XOGAN: one-to-many unsupervised image-to-image translation. 2018. arXiv:1805.07277.
82. Zhang Y, Gan Z, Fan K, et al. Adversarial feature matching for text generation. 2017. arXiv:1706.03850.
83. Yu L, Zhang W, Wang J, Yu Y. SeqGAN: sequence generative adversarial nets with policy gradient. In: AAAI. 2017. p. 2852–858.
84. Fedus W, Goodfellow I, Dai AM. Maskgan: better text generation via filling in the _. 2018. arXiv:180107736.
85. Mortazi A, Bagci U. Automatically designing CNN architectures for medical image segmentation. In: International workshop on machine learning in medical imaging. Springer; 2018. p. 98–106.
86. Xie L, Yuille AL. Genetic CNN. In: ICCV; 2017. p. 1388–97.
87. Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. 2017. p. 2. arXiv :1707.07012.
88. Pham H, Guan MY, Zoph B, Le QV, Dean J. Efficient neural architecture search via parameter sharing. 2018. arXiv :1802.03268.
89. Liu H, Simonyan K, Yang Y. DARTS: differentiable architecture search. 2018. arXiv:1806.09055.
90. Luo R, Tian F, Qin T, Liu T-Y. Neural architecture optimization. 2018. arXiv:1808.07233.
91. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284:574–82.
92. Dietterich TG. Ensemble methods in machine learning. International workshop on multiple classifier systems. Springer; 2000. p. 1–15.
93. Prevedello LM, Erdal BS, Ryu JL, et al. Automated critical test findings identification and online notification system using artificial intelligence in imaging. Radiology. 2017;285:923–31.
94. Kim JR, Shim WH, Yoon HM, et al. Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. Am J Roentgenol. 2017;209:1374–80.

95. 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. 2017;286:887–96.
96. 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;287:313–22.
97. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Deep learning for staging liver fibrosis on CT: a pilot study. Eur Radiol. 2018;28:440–51.
98. 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;287:146–55.
99. Noguchi T, Higa D, Asada T, et al. Artificial intelligence using neural network architecture for radiology (AINNAR): classification of MR imaging sequences. Jpn J Radiol. 2018;36(12):691–7.
100. England JR, Gross JS, White EA, Patel DB, England JT, Cheng PM. Detection of traumatic pediatric elbow joint effusion using a deep convolutional neural network. Am J Roentgenol. 2018;211(6):1361–8.
101. Kim Y, Lee KJ, Sunwoo L, et al. Deep learning in diagnosis of maxillary sinusitis using conventional radiography. Invest Radiol. 2018. https://doi.org/10.1097/RLI.0000000000000503
102. Lehman CD, Yala A, Schuster T, et al. Mammographic breast density assessment using deep learning: clinical implementation. Radiology. 2018:180694.
103. Ueda D, Yamamoto A, Nishimori M, et al. Deep learning for MR angiography: automated detection of cerebral aneurysms. Radiology. 2018:180901.
104. Chang P, Kuoy E, Grinband J, et al. Hybrid 3D/2D convolutional neural network for hemorrhage evaluation on head CT. Am J Neuroradiol. 2018;39(9):1609–16.
105. 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.
106. Norman B, Pedoia V, Majumdar S. Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry. Radiology. 2018;288(1):177–85.
107. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–41.
108. Perkuhn M, Stavrinou P, Thiele F, et al. Clinical evaluation of a multiparametric deep learning model for glioblastoma segmentation using heterogeneous magnetic resonance imaging data from clinical routine. Investig Radiol. 2018;53(11):647–54.
109. Laukamp KR, Thiele F, Shakirin G, et al. Fully automated detection and segmentation of meningiomas using deep learning on routine multiparametric MRI. Eur Radiol. 2018:1-9.
110. Kamnitsas K, Ledig C, Newcombe VF, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal. 2017;36:61–78.
111. Montoya J, Li Y, Strother C, Chen G-H. 3D deep learning angiography (3D-DLA) from C-arm conebeam CT. Am J Neuroradiol. 2018;39:916–22.
112. Tao Q, Yan W, Wang Y, et al. Deep learning–based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology. 2018:180513.
113. Liu F, Jang H, Kijowski R, Bradshaw T, McMillan AB. Deep learning MR imaging–based attenuation correction for PET/MR imaging. Radiology. 2017;286:676–84.
114. Kim KH, Choi SH, Park S-H. Improving arterial spin labeling by using deep learning. Radiology. 2017;287:658–66.
115. Ahn SY, Chae KJ, Goo JM. The potential role of grid-like software in bedside chest radiography in improving image quality and dose reduction: an observer preference study. Korean J Radiol. 2018;19:526–33.
116. Chen F, Taviani V, Malkiel I, et al. Variable-density single-shot fast spin-echo MRI with deep learning reconstruction by using variational networks. Radiology. 2018;289(2):180445.
117. Kobler E, Klatzer T, Hammernik K, Pock T. Variational networks: connecting variational methods and deep learning. In: German conference on pattern recognition. Springer; 2017. p. 281–93.
118. Jiang D, Dou W, Vosters L, Xu X, Sun Y, Tan T. Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network. Jpn J Radiol. 2018;36:566–74.
119. Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a gaussian denoiser: residual learning of deep cnn for image denoising. IEEE Trans Image Process. 2017;26:3142–55.
120. Chen MC, Ball RL, Yang L, et al. Deep learning to classify radiology free-text reports. Radiology. 2017;286:845–52.
121. Kim Y. Convolutional neural networks for sentence classification. 2014. arXiv:1408.5882.
122. Zech J, Pain M, Titano J, et al. Natural language–based machine learning models for the annotation of clinical radiology reports. Radiology. 2018;287:570–80.
123. Chang P, Grinband J, Weinberg B, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. Am J Neuroradiol. 2018;39(7):1201–7.
124. Liu F, Zhou Z, Samsonov A, et al. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. Radiology. 2018;289(1):160–9.
125. Choi KJ, Jang JK, Lee SS, et al. Development and validation of a deep learning system for staging liver fibrosis by using contrast agent–enhanced CT images in the liver. Radiology. 2018;289(3):688–97.
126. Kim Y-H, Reddy B, Yun S, Seo C. Nemo: Neuro-evolution with multiobjective optimization of deep neural network for speed and accuracy. In: ICML.
127. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology. 2018:180237.
128. Liang S, Tang F, Huang X, et al. Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. Eur Radiol. 2018. https://doi.org/10.1007/s00330-018-5748-9
129. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology. 2018;286:800–9.
130. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. npj Dig Med. 2018;1:18.
131. Japanese goverment to make inclusive rules for use of AI in medical practice. Nikkei. 2018.
132. Nakajima Y, Yamada K, Imamura K, Kobayashi K. Radiologist supply and workload: international comparison–Working Group of Japanese College of Radiology. Radiat Med. 2008;26:455–65.
133. Nishie A, Kakihara D, Nojo T, et al. Current radiologist workload and the shortages in Japan: how many full-time radiologists are required? Jpn J Radiol. 2015;33:266–72.
134. Kumamaru KK, Machitori A, Koba R, Ijichi S, Nakajima Y, Aoki S. Global and Japanese regional variations in radiologist potential workload for computed tomography and magnetic resonance imaging examinations. Jpn J Radiol. 2018;36:273–81.