Deep Learning in Medical Image Analysis

Heang-Ping Chan, Ravi K. Samala, Lubomir M. Hadjiiski, and Chuan Zhou

Abstract

Deep learning is the state-of-the-art machine learning approach. The success of deep learning in many pattern recognition applications has brought excitement and high expectations that deep learning, or artificial intelligence (AI), can bring revolutionary changes in health care. Early studies of deep learning applied to lesion detection or classification have reported superior performance compared to those by conventional techniques or even better than radiologists in some tasks. The potential of applying deep-learning-based medical image analysis to computer-aided diagnosis (CAD), thus providing decision support to clinicians and improving the accuracy and efficiency of various diagnostic and treatment processes, has spurred new research and development efforts in CAD. Despite the optimism in this new era of machine learning, the development and implementation of CAD or AI tools in clinical practice face many challenges. In this chapter, we will discuss some of these issues and efforts needed to develop robust deep-learning-based CAD tools and integrate these tools into the clinical workflow, thereby advancing towards the goal of providing reliable intelligent aids for patient care.

Keywords

Machine learning · Deep learning · Artificial intelligence · Computer-aided diagnosis · Medical imaging · Big data · Transfer learning · Validation · Quality assurance · Interpretable AI

Introduction

Medical imaging is an important diagnostic tool for various diseases. Roentgen discovered that X-rays could non-invasively look into the human body in 1895 and X-ray radiography became the first diagnostic imaging modality soon after. Since then many imaging modalities were invented, with computed tomography, ultrasound, magnetic resonance imaging, and positron emission tomography among the commonly used, and more and more complex imaging procedures have been developed. Image information plays a crucial role in decision making at many stages in the patient care process, including detection, characterization, staging, treatment response assessment, monitoring of disease recurrence, as well

H.-P. Chan (✉) · R. K. Samala · L. M. Hadjiiski
C. Zhou
Department of Radiology, University of Michigan, Ann Arbor, MI, USA
e-mail: chanhp@umich.edu

G. Lee, H. Fujita (eds.), *Deep Learning in Medical Image Analysis*, Advances in Experimental Medicine and Biology 1213,
https://doi.org/10.1007/978-3-030-33128-3_1

as guiding interventional procedures, surgeries, and radiation therapy. The number of images for a given patient case increases dramatically from a few two-dimensional (2D) images to hundreds with 3D imaging and thousands with 4D dynamic imaging. Application of multi-modality imaging further increases the amount of image data to be interpreted. The increasing workload makes it difficult for radiologists and physicians to maintain workflow efficiency while utilizing all the available imaging information to improve accuracy and patient care. With the advances in machine learning and computational techniques in recent years, developing effective and reliable computerized methods to assist radiologists and physicians in image analysis at various stages of disease diagnosis and management during the patient care process has been recognized as an important area of research in medical imaging.

The attempt of using computers to automatically analyze medical images emerged as early as the 1960s [1–4]. Several studies demonstrated the feasibility of applying computer to medical image analysis but the work did not attract much attention, probably because of the limited access to high quality digitized image data and computational resources. Doi et al. in the Kurt Rossmann Laboratory at the University of Chicago began systematic development of machine learning and image analysis techniques for medical images in the 1980s [5], with the goal to develop computer-aided diagnosis (CAD) as a second opinion to assist radiologists in image interpretation. Chan et al. developed a CAD system for detection of microcalcifications on mammograms [6] and conducted the first observer performance study [7] that demonstrated the effectiveness of CAD in improving breast radiologists' detection performance of microcalcifications. The first CAD commercial system was approved by the Food and Drug Administration (FDA) for use as a second opinion in screening mammography in 1998. CAD and computer-assisted image analysis have been a major area of research and development in medical imaging in the past few decades. CAD methods have been investigated for various applications including disease detection, characterization, staging, treatment response assessment, prognosis prediction, and risk assessment for various diseases and with various imaging modalities. The work in the CAD field has been steadily increasing as can be seen from the trend of publications in peer-reviewed journal articles found by literature search in the Web of Science (Fig. 1).

Although the research in CAD has been increasing, very few CAD systems are used routinely in the clinic. One of the major reasons may be that CAD tools developed with conventional machine learning methods may not have reached the high performance that can meet physicians' needs to improve both diagnostic accuracy and workflow efficiency. With the success of deep learning in many machine learning applications such as text and speech recognition, face recognition, autonomous vehicles, chess and Go game, in the past several years, there are high expectations that deep learning will bring breakthrough in CAD performance and widespread use of deep-learning-based CAD, or artificial intelligence (AI), to various tasks in the patient care process. The enthusiasm has spurred numerous studies and publications in CAD using deep learning. In this chapter, we will discuss some issues and challenges in the development of deep-learning-based CAD in medical imaging, as well as considerations needed for the future implementation of CAD in clinical use.

Deep Learning for Medical Image Analysis and CAD

CAD systems are developed with machine learning methods. Conventional machine learning approach to CAD in medical imaging used image analysis methods to recognize disease patterns and distinguish different classes of structures on images, e.g., normal or abnormal, malignant or benign. CAD developers design image processing and feature extraction techniques based on domain knowledge to represent the image characteristics that can distinguish the various states. The effectiveness of the feature descriptors often depends on the domain expertise of the CAD developers and the capability of the mathematical formulations or empirical image analysis techniques that are designed to translate the image

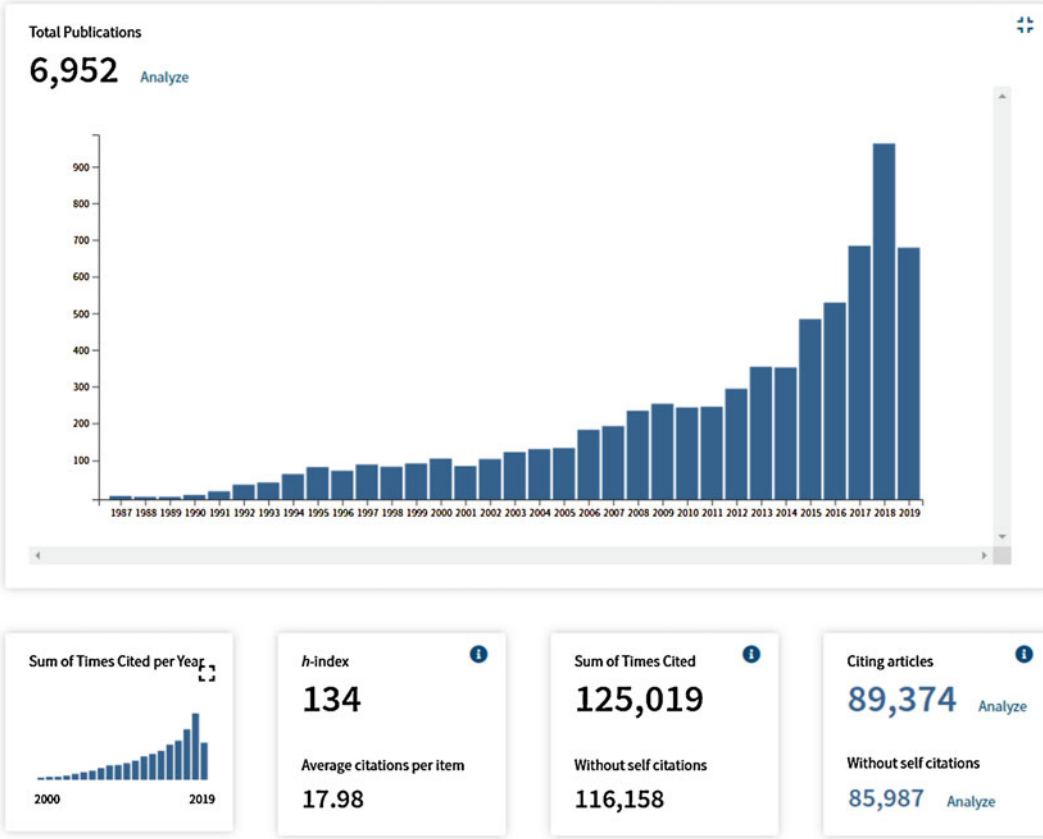

Fig. 1 Literature search for publications in peer-reviewed journals by Web of Science from 1900 to early July of 2019 using key words: ((imaging OR images) AND (medical OR diagnostic)) AND (machine learning OR deep learning OR neural network OR deep neural network OR convolutional neural network OR computer aid OR computer assist OR computer-aided diagnosis OR automated detection OR computerized detection OR computer-aided detection OR automated classification OR computerized classification OR decision support OR radiomic) NOT (pathology OR slide OR genomics OR molecule OR genetic OR cell OR protein OR review OR survey))

characteristics to numerical values. The extracted features are then used as input predictor variables to a classifier, and a predictive model is formed by adjusting the weights of the various features based on the statistical properties of a set of training samples to estimate the probability that an image belongs to one of the states. Conventional machine learning approach has limitations in that the human developer may not be able to translate the complex disease patterns into a finite number of feature descriptors even if they have seen a large number of cases from the patient population. The hand-engineered features may also have difficulty to be robust against the large variations of normal and abnormal patterns in the population. The performance of the developed CAD system is often limited in its discriminative power or generalizability, resulting in high false positive rate at high sensitivity or vice versa.

Deep learning has emerged as the state-of-the-art machine learning method in many applications. Deep learning is a type of representation learning method in which a complex multilayer neural network architecture learns representations of data automatically by transforming the input information into multiple levels of abstractions [8]. For pattern recognition tasks in images, deep convolutional neural networks (DCNN) are the most commonly used deep learning networks. With a sufficiently large training

set, DCNN can learn to automatically extract relevant features from the training samples for a given task by iteratively adjusting its weights with backpropagation. DCNN therefore discovers feature representations through training and does not require manually designed features as input. If properly trained with a large training set that are representative of the population of interest, the DCNN features are expected to be superior to hand-engineered features in that they have high selectivity and invariance [8]. Importantly, since the learning process is automated, deep learning can easily analyze thousands or millions of cases that even human experts may not be able to see and memorize in their lifetime. Deep learning can therefore be more robust to the wide range of variations in features between different classes to be differentiated as long as the training set is large and diverse enough for it to analyze.

CNN can trace its origin to the neocognitron proposed by Fukushima et al. in the early 1980s [9]. LeCun first trained a CNN by backpropagation to classify patterns of handwritten digits in 1990 [10]. CNN was used in many applications such as object detection, character, and face recognition in the early 1990s. Lo et al. first introduced CNN to the analysis of medical images in 1993 and trained a CNN for lung nodule detection in chest radiographs [11, 12]. Chan et al. applied CNN to microcalcification detection [13, 14] on mammograms in the same year and to mass detection in the following year [15–18]. Zhang et al. applied a similar shift-invariant neural network for the detection of clusters of microcalcifications in 1994 [19]. Although these early CNNs were not very deep, the pattern recognition capability of CNN in medical images was demonstrated.

Deep CNN was enabled by several important neural network training techniques developed over the years, including layer-wise unsupervised pre-training followed by supervised fine-tuning [20–22], use of rectified linear unit (ReLU) [23, 24] as activation function in place of sigmoid-type activation functions, pooling to improve feature invariance and reduce dimensionality [25], dropout to reduce overfitting [26], and batch normalization [27] that further reduces the risk of internal covariate shift, vanishing gradient and overfitting, as well as increases training convergence speed. These techniques allow neural networks with more and more layers and containing millions of weights to be trained. In 2012, Krizhevsky et al. [28] proposed a CNN with five convolutional layers and 3 fully connected layers (named "AlexNet") containing over 60 million weights and achieved breakthrough performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [29] that classified over 1000 classes of everyday objects on photographic images. AlexNet demonstrated the pattern recognition capability of the multiple layers of a deep structure. DCNNs with increasing depth were developed since AlexNet. He et al. [30] proposed residual learning and showed that a residual network (ResNet) with 110–152 layers could outperform several other DCNNs and won the ILSVRC in 2015. Sun et al. [31] showed that the learning capacity of a DCNN increased with depth but the capacity could be utilized only with sufficiently large training data.

The success of deep learning or AI in personal devices and social media, self-driving cars, chess, and Go game have raised unprecedented expectations of deep learning in medicine. Deep learning has been applied to many medical image analysis tasks for CAD [32–34]. The most common areas of CAD application using deep learning include classification of disease and normal patterns, classification of malignant and benign lesions, and prediction of high risk and low risk patterns of developing cancer in the future. Other applications included segmentation and classification of organs and tumors of different types, classification of changes in tumor size or texture for assessment of treatment response, or prediction of prognosis or recurrence. Because there are relatively large public data sets available for chest radiographs, thoracic CT, and mammograms, a large number of studies were conducted for lung diseases and breast cancer using the public data sets. Deep-learning-based image analysis has also been applied to fundus images or optical computed tomography for detection of eye diseases [35], or histopathological images for classification of cell types [36]. Most of the studies reported very promising results, further boosting

the hype of deep-learning-based CAD. This new generation of CAD is called AI although these CAD tools still behave like a very complex mathematical model that memorizes information in its millions of weights and far from being "intelligent."

Challenges in Deep-Learning-Based CAD

CAD or AI is expected to be useful decision support tools in medicine in the near future. Other than detection and characterization of abnormalities, applications such as pre-screening and triaging, cancer staging, treatment response assessment, recurrence monitoring, and prognosis or survival prediction are being explored. Although no CAD systems with new AI techniques have been subjected to large scale clinical trials to date, experiences from CAD use in screening mammography may provide some insights into what may be expected of CAD tools in the clinic [37].

The conventional machine-learning-based CAD for detection of breast cancer in screening mammography is the only CAD application in widespread clinical use to date. These systems have been shown to have sensitivity comparable to or higher than that of radiologists, especially for microcalcifications, but they also mark a few false positives per case on average [38]. Although the performances of CAD systems are moderate, they may detect lesions of different characteristics than those by radiologists. The complementary detections by the radiologist and CAD can improve the overall sensitivity when radiologist reads with CAD. Studies have shown that radiologists' accuracy was improved significantly when reading with CAD [5]. CAD systems were therefore approved by FDA for use as a second opinion but not as a primary reader or pre-screener. Early clinical trials [39, 40] to compare single reading with CAD to double reading showed promising results. In the CADET II study by Gilbert et al. [39], they conducted a prospective randomized clinical trial at three sites in the United Kingdom. A total of over 28,000 patients were included. The screening mammograms of each patient were independently read in two arms; one was single reading with CAD and the other was their standard practice of double reading. The experiences of the single readers in the CAD arm were matched to those of the first readers' in the double reading arm. Arbitration was used in cases of recall due to the second reader or CAD. They found that arbitration was performed in 1.3% of the cases in single reading with CAD. The average sensitivity in the two arms were comparable at 87.2% and 87.7%, respectively. The recall rates at two centers were comparable in the two arms, 3.7% versus 3.6% and 2.7% versus 2.7%, respectively, but one of the centers had a significantly higher recall rate for single reading with CAD, 5.2% versus 3.8%. The overall recall rate therefore increased in the single reading with CAD from 3.4% to 3.9%. Gromet et al. [40] performed a retrospective review of the sensitivity and recall rate by single reading with CAD after CAD implementation in comparison to those of double reading before CAD use as historical control for the same group of nine radiologists in a single mammography facility. The first reading in their double reading protocol was also analyzed and treated as single reading without CAD. The study cohort contained over 110,000 screening examinations in each group. Arbitration by a third subspecialty radiologist was a part of their standard double reading protocol. A second radiologist was consulted for 2.1% of the cases interpreted by single reading with CAD but the consult might or might not be related to CAD marks. They reported that the sensitivity of single reading with CAD was 90.4%, higher than the sensitivities of either single reading alone (81.4%) or double reading (88.0%). The recall rate was 10.6% for single reading with CAD, slightly higher than the recall rate of single reading alone (10.2%) but lower than that of double reading (11.9%). These relatively well-controlled studies showed that single reading with CAD is potentially an alternative to double reading, with a gain in sensitivity but at the expense of increased recalls, which can be reduced by arbitration similar to that in double reading.

Table 1 Odds ratios (95% confidence interval) of increase in cancer detection rate and increase in recall rate obtained by comparison of single reading with CAD and double reading to single reading alone by Taylor et al. [41]

	Odds ratio of increase in cancer detection rate	Odds ratio of increase in recall rate
Single reading with CAD		
Matched ($N = 5$)	1.09 (0.92, 1.29)	1.12 (1.08, 1.17)
Unmatched ($N = 5$)	1.02 (0.93, 1.12)	1.10 (1.08, 1.12)
Double reading		
Unilateral ($N = 6$)	1.13 (1.06, 1.19)	1.31 (1.29, 1.33)
Mixed ($N = 3$)	1.07 (0.99, 1.15)	1.21 (1.19, 1.24)
Arbitration ($N = 8$)	1.08 (1.02, 1.15)	0.94 (0.92, 0.96)

N: the number of studies included in each group
Matched studies: the assessment before and after using CAD was on the same mammograms
Unmatched studies: the performance of mammography facilities after the introduction of CAD was compared to that before CAD implementation as historical controls. Different mammograms were interpreted in the two conditions

Taylor et al. [41] conducted a meta-analysis of clinical studies comparing single reading with CAD or double reading to single reading alone. They compared the cancer detection rate per 1000 women screened (CDR) and the recall rate, and estimated the average odds ratios weighted by sample size over the studies in each group (Table 1). The results showed that double reading with arbitration improved the CDR without increasing the recall rates. Single reading with CAD for the matched studies increased the CDR but with a wide variation; however, without the benefit of arbitration, the recall rate increased significantly. The increase in recall rate for double reading without arbitration was more than twice of that for single reading with CAD.

Taylor et al. revealed that there are large variations in the impact of CAD on the cancer detection rate ranging from 0% to 19%, and the recall rate ranging from 0% to 37%. Other than the differences in the study designs and radiologists' experiences in the studies, the variations may also be attributed to the varied ways that radiologists used CAD in the clinic. Some users might have misunderstood the limitations and performance of the CAD systems. They might have over-relied on the CAD marks and thus did not maintain their vigilance in searching for lesions while increasing their recalls. Others might have used CAD as a pre-screener or first reader to increase workflow. Although there were no systematic studies of how CAD was used in the clinic, Fenton et al. [42] noted that "radiologists with variable experience and expertise may use CAD in a nonstandardized idiosyncratic fashion," and "Some community radiologists, for example, may decide not to recall women because of the absence of CAD marks on otherwise suspicious lesions." Lehman et al. [43] compared reading digital mammograms with and without CAD by 271 radiologists in 66 facilities of the Breast Cancer Surveillance Consortium (BCSC). They reported that the average sensitivity decreased by 2.3% and the recall rate increased by 4.5% with the use of CAD. The decrease in sensitivity was a clear indication that the radiologists did not use CAD as a second opinion, which require the users to maintain their vigilance in interpretation and thus their sensitivity, but over-relied on the CAD marks for recall decisions. The authors acknowledged that "Prior reports have confirmed that not all cancers are marked by CAD and that cancers are overlooked more often if CAD fails to mark a visible lesion" and that "CAD might improve mammography performance when appropriate training is provided on how to use it to enhance performance."

The study by Cole et al. [38] demonstrated another facet of using CAD. They conducted an observer study to compare single reading with and without CAD using two commercial CAD systems applied to 300 screening cases (150 cancers and 150 benign or normal) from the Digital Mammographic Imaging Screening Trial (DMIST). All participating

Table 2 Observer performance study by Cole et al. [38] comparing single reading with and without CAD using two commercial CAD systems and 300 screening mammography cases (150 cancer and 150 benign or normal) from DMIST

	CAD system A		CAD system B	
Standalone performance	75% sensitivity at 0.79 FPs/image		73% sensitivity at 0.77 FPs/image	
Radiologists	$N = 14$		$N = 15$	
	Without CAD	With CAD	Without CAD	With CAD
Average AUC	0.71	0.72	0.71	0.72
Average sensitivity	49%	51%	51%	53%
Average specificity	89%	87%	87%	86%

None of the changes were statistically significant. AUC = area under the receiver operating characteristic curve. N = number of radiologists in the study

readers were experienced breast radiologists and had been using CAD in their clinical practice. As summarized in Table 2, they found that the changes in the radiologists' sensitivity or specificity with CAD were only 1% to 2%. The standalone sensitivity of both CAD systems were 25% higher than the radiologists with or without CAD but had an average of more than 2 false positive marks per case. These results were very different from those observed in the early days of CAD development when radiologists were enthusiastic about CAD. They appeared to show that after radiologists used CAD in the clinic for a period of time, the many false positive CAD marks they have seen may have desensitized their attention and most of the marks were dismissed including true positives. In a screening setting, the time a radiologist has to spend to exclude over 2000 false positive marks in order to gain one or two cancers per 1000 examinations is considered not cost-effective by many radiologists. This study indicated that the specificity of a decision support tool has to be high to avoid inducing fatigue on clinicians' response to the computer's recommendations.

Although these clinical experiences of CAD were observed from screening mammography, they reveal the many challenges of implementing CAD or AI tools in the clinic and may provide some guidance on the development of the new generations of CAD for various applications in general. Accuracy and workflow efficiency are important considerations in clinical practice. User training is crucial to ensure their understanding of the limitations and capability of CAD and thus avoid improper use or disillusion. Clinicians' experiences and level of enthusiasm with CAD also strongly impact on whether they will accept a CAD tool and how they may respond to its recommendation. Performance standards and acceptance testing should be established to ensure the CAD tool can meet certain criteria before routine clinical use. Quality assurance is needed to monitor the consistency and accuracy of the CAD tool over time, as well as to prevent improper use that may impact patient safety. Fully automated medical decision systems are ideal, but experienced clinicians' supervision is vital as many clinical cases may not evolve following a statistical model and require human intelligence to determine the best course of action based on the individual patient's conditions and medical history. The AI community has recently scaled back the expectation and defines a less ambitious term as "narrow AI" or "Augmented Intelligence," recognizing the supporting role of machine learning algorithms. Regardless of CAD, narrow AI, or general AI, there are many challenges of developing machine-learning-based tools for medicine. Some of the challenges are discussed below.

Data Collection

The majority of the studies to date on the application of deep learning to medical imaging reported very promising results that often exceeded clinicians' performance, raising high expectations of AI tools. However, most of the studies used small training set and the trained models have not been subjected to rigorous validation with large real world test data. The generalizability of these deep

learning models to new patients or to different clinical settings is still unknown.

One of the basic requirements to develop a robust machine learning algorithm is a sufficiently large training sample set with verified reference truth that are representative of the characteristics of the population of interest. Training deep learning is even more demanding because of the extremely large number of weights in a DCNN structure. Even with effective regularization methods to reduce overfitting, how general the feature representations it has learned still depends on how much the training set covers. AlexNet has over 60 million weights and the "ImageNet" data set for training includes over 1.2 million images with annotations. Sun et al. [31] showed that the performance of a DCNN increased linearly with the orders of magnitude of the training data and the performance of a DCNN with large learning capacity continued to increase even when the training set increased to over 300 million images.

Collection of medical imaging data that are representative of the patient population and with reliable annotation or reference truth is costly. While it is relatively easy to collect a large number of normal cases for a screening modality, it is difficult to collect sufficient abnormal cases, especially that the different classes in the data set ideally should be balanced. For example, for disease such as breast cancer that is the most prevalent cancer in women, there are only several cases per thousand in the screening population. It is difficult to collect enough breast cancer mammograms or tomosynthesis that can cover the variabilities in image features due to factors such as patient age, breast density and size, habitus, race, ethnicity, imaging protocols, and processing methods. The collection of normal and abnormal cases with special imaging modalities such as MR or PET is even more challenging because a relatively small number of patients will have these examinations and the availability may depend on the protocols for different types of diseases in different health systems.

Studies have demonstrated the feasibility of collecting a large number of annotated cases by data mining and natural language processing of the electronic medical record (EMR) [44] and clinical annotations in picture archiving and communication system (PACS) [45]. The accuracy and usefulness of the labels or annotations obtained from these methods not only depend on the methods used but also how the information is generated and stored in the systems. It has been shown that automatically mined disease labels or annotations can include substantial noise in a data set, as in the large public set of chest radiographs [46]. In the Digital Mammography DREAM Challenge (2016–2017) that aimed at building a model to help reduce the recall rate for breast cancer screening [47], the participants were provided with over 640,000 training mammograms from over 86,000 women. The training set only included breast-level labeling without lesion annotation. The winning teams all used deep learning approach but the highest performance only reached an area under the receiver operating characteristic curve (AUC) of 0.8744, and a sensitivity of 80% at specificity of 80.8%. The false positive rate, and thus potentially the recall rate, was much higher than that of an experienced breast radiologist at comparable sensitivity even for the top deep learning model. This example illustrates that, although the total number of images appeared to be large, the lack of high quality labeling may reduce its effectiveness in deep learning training. In general, weakly supervised training, unsupervised training, or using training set with substantial labeling errors is not as effective as supervised training with well-curated cases for the same training sample size; a much larger sample size is required to achieve similar performance for a DCNN model as a well-curated training set.

Data mining of the unstructured text and non-standardized reporting in current EMR or PACS systems is challenging, especially for more complex CAD task such as treatment response monitoring, in which a case may include multiple stages of diagnosis and treatment involving multiple imaging examinations and clinical tests. To generate reference standards for CAD development, one needs to correlate the imaging and clinical test data with outcomes at the various stages. It is a difficult process even if performed

manually. Automation will be useful but it may require the development of an intelligent data mining tool. For patient cases that have been transferred between different hospitals, the incomplete prior or follow-up information may introduce errors into data curation. To facilitate collecting big data for development of AI towards precision medicine in the future, it will be prudent for the vendors and users to establish standardized reporting methods and structures among the various data archiving systems. In addition, establishing standardized protocols for secure electronic transmission of patient files among hospitals for referral patients will not only improve the health care of referral patients by transferring patient data accurately and efficiently, but also improve the accuracy of data mining for these cases. Ultimately, multi-institutional collaboration may be the best approach to building big database, which can cover the wide ranges of heterogeneous imaging protocols and equipment, clinical settings, and patient characteristics, to accelerate the development of robust deep learning models for each type of diseases that may be more readily applicable to various clinical environments.

Transfer Learning

Transfer learning is a common approach that deep learning developers use when the training set was small. In transfer learning, a DCNN that has been well trained with a large training set from a source domain is adapted to a new target task by fine-tuning the DCNN using a relatively small training set from the target domain. DCNN is considered a feature extractor that learns representation of the input data by extracting multiple levels of abstractions by its convolutional layers. Yosinski et al. [48] showed that the learned features in the shallow layers are more generic, whereas the learned features in the deeper layers become increasingly specific to the task that the DCNN is being trained for. Since the features are decomposed into numerous components in a DCNN, and most images are composed of some common basic elements, the knowledge learned by a trained DCNN in extracting features is shown to be transferrable to images from different domains. The transferability of features decreases as the differences between the source domain and the target domain increase. However, even for very different source and target tasks, transfer learning by initializing a DCNN with weights trained for another source task can outperform the same DCNN trained with randomly initialized weights for the target task.

For training deep learning models in medical imaging, the majority of studies used transfer learning due to the limited data available. To date, the largest annotated public data set available is the ImageNet data, which contained photographic images containing over 1000 classes of everyday life objects such as animals, vehicles, plants, ships, planes, etc. Most of the DCNN models in medical imaging were trained by transfer learning using models initialized with ImageNet-pretrained weights and fine-tuned by limited medical image data. Transfer learning was generally found to be useful in improving the training convergence and robustness of the DCNNs. In some cases, the pretrained DCNNs were used as feature extractor without fine-tuning; the deep features extracted from deploying the pretrained DCNN to the image data of the target domain were used as predictor variables to train an external classifier for the target task.

Although transfer learning can alleviate the problem of limited data to a certain degree, a large training set is still needed to achieve a high performance DCNN model for a given target task. Samala et al. [49] conducted a study to evaluate the effect of training set size on the performance of a transfer-trained DCNN for the target task of classifying malignant and benign breast masses in digital breast tomosynthesis (DBT). The ImageNet-pretrained AlexNet with 5 convolutional layers and 3 fully connected layers was appended with 2 additional fully connected layers (total of 5 fully connected layers) to reduce the classes from over 1000 to 2 (malignant and benign) and transfer-trained for the target task. Because the DBT data set was small and mammogram data were relatively abundant, the

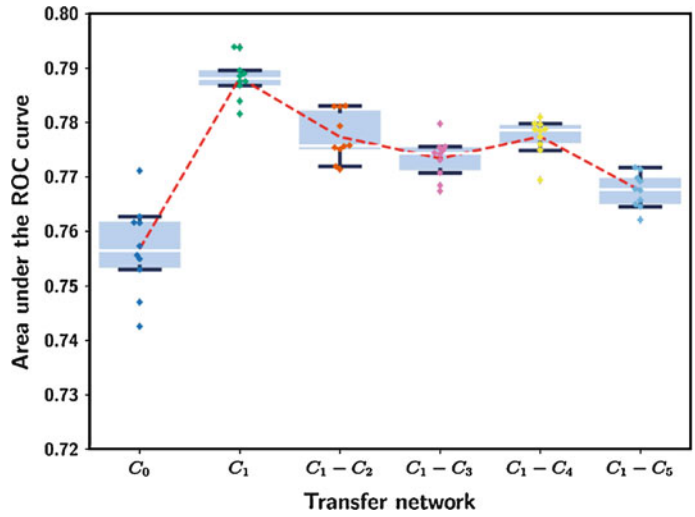

Fig. 2 The effect of different number of layers of the DCNN being frozen during transfer learning of ImageNet-pretrained AlexNet to classify malignant and benign masses on mammograms. The area under the receiver operating characteristic curve (AUC) for the test ROIs was plot as box-and-whisker plots of 10 repeated experiments under each condition. The training set and the test set consists of 12,360 and 7272 ROIs after augmentation, respectively. C_0 denotes no layer was frozen, i.e., the pretrained weights in all layers were allowed to be updated. C_1 denotes the first convolutional layer was frozen, C_1–C_i ($i = 2, 3, 4, 5$) denotes the C_1 to C_i convolutional layers were frozen during transfer training. The result shows that C_1-frozen training provided the best test AUC for this task (reprint with permission [49])

pretrained AlexNet was transfer-trained in the first stage for the classification of masses on mammograms, which brought the AlexNet from an unrelated classification task on non-medical ImageNet data to a task (mammography) much closer to the target task (DBT). A small DBT set was then used for a second-stage transfer training to the target task. Their mammography set contained 2242 unique views (craniocaudal or mediolateral oblique) with 2454 regions of interest (ROIs) containing breast masses. The DBT set contained 324 unique views with 1585 ROIs (5 slices or ROIs from each mass), which was partitioned by case into a training set of 1140 ROIs and an independent test set of 445 ROIs. Each ROI was flipped and rotated to obtain 8 augmented versions to reduce noise. To evaluate the training sample size effects on stage 1 and stage 2 fine-tuning, several transfer learning strategies were compared: (A) single-stage transfer learning with mammography data, in which the first convolutional layer (C_1) of AlexNet was frozen and all other layers were allowed to be fine-tuned, (B) two-stage transfer learning with mammography data in stage 1 and DBT training set in stage 2, in which C_1 was frozen in both stages, (C) two-stage transfer learning similar to (B) except that convolutional layers C_1 to F_4 were frozen in stage 2, and (D) single-stage transfer learning with DBT training set, in which C_1 was frozen, was also trained as a baseline for comparison. The results are summarized in Figs. 2, 3, 4, and 5.

Figure 2 shows the dependence of the test performance, in terms of AUC, of the transfer-trained AlexNet on the number of layers being frozen during transfer training for the classification of masses on mammograms. The AUC was the highest when only C_1 was frozen. However, if all layers were allowed to be re-trained (C_0), the transfer trained AlexNet did not perform well, probably because the mammography data was not large enough to fine-tune the large number of weights. Figure 3 shows the dependence of the test AUC on the sample size of the training mammography data. The test AUC was obtained by applying the AlexNet transfer-trained with mammography data directly to classify the masses on DBT with-

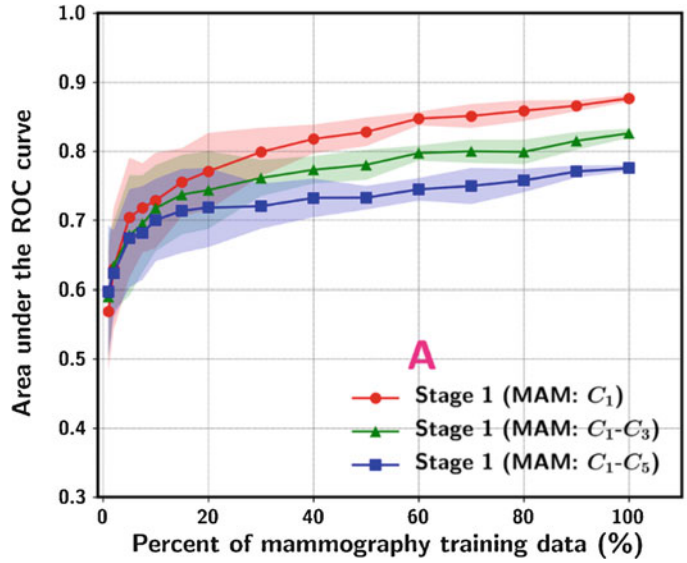

Fig. 3 Dependence of test AUC on mammography training sample size using strategy (A) transfer training. The varied training sample size was simulated by random drawing by case of a percentage (ranging from 1% to 100%) from the entire set of 19,632 mammography ROIs. The ROI-based AUC performance for classifying the 9120 DBT training ROIs (serve as a test set at this stage) for three transfer networks at Stage 1. The data point and the upper and lower range show the mean and standard deviation of the test AUC resulting from ten random samplings of the training set of a given size from the original set (reprint with permission [49])

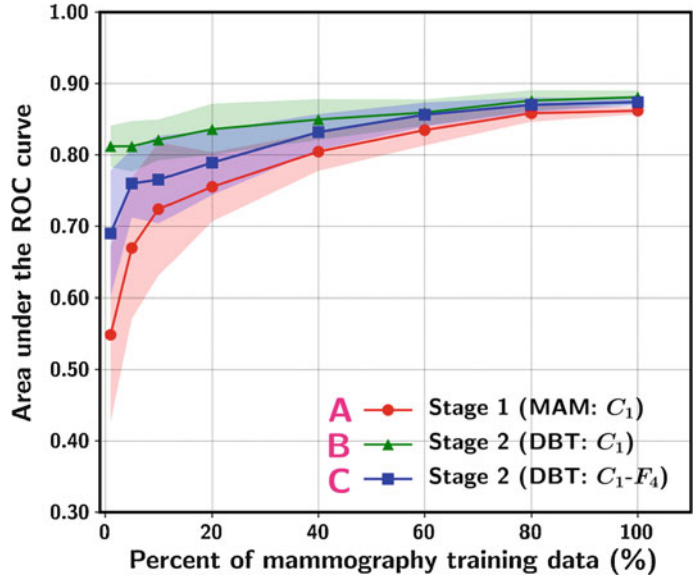

Fig. 4 ROI-based AUC on the DBT test set while varying the mammography sample size available for transfer training. The data point and the upper and lower range show the mean and standard deviation of the test AUC resulting from ten random samplings of the training set of a given size from the original set. "A. Stage 1 (MAM:C_1)" denotes single-stage training using mammography data and the C_1-layer frozen during transfer learning without stage 2. "B. Stage 2 (DBT:C_1)" denotes stage 2 C_1-frozen transfer learning at a fixed (100%) DBT training set size after Stage 1 transfer learning (curve A). "C. Stage 2 (DBT:C_1-F_4)" denotes Stage 2 C_1-to-F_4-frozen transfer learning at a fixed (100%) DBT training set size after stage 1 transfer learning (curve A) (reprint with permission [49])

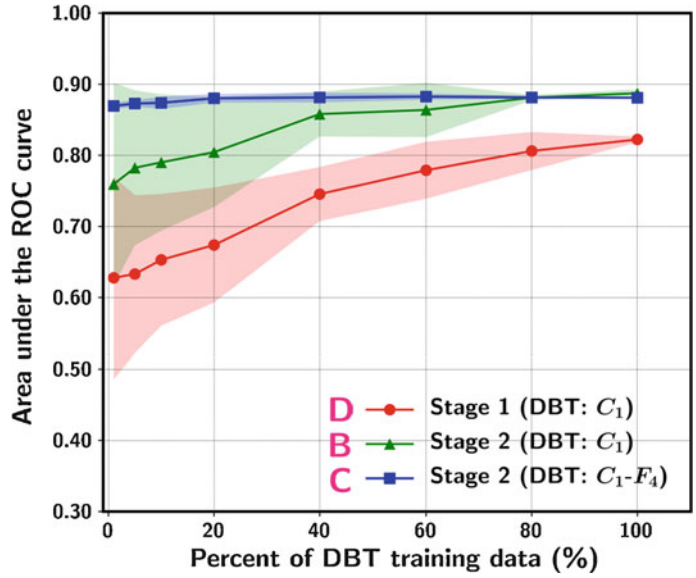

Fig. 5 ROI-based AUC on the DBT test set while varying the simulated DBT sample size available for transfer training. The data point and the upper and lower range show the mean and standard deviation of the test AUC resulting from ten random samplings of the training set of a given size from the original set. "D. Stage 1 (DBT:C_1)" denotes single-stage training using DBT data with the C_1-layer frozen during transfer learning without Stage 2. "B. Stage 2 (DBT:C_1)" denotes Stage 2 C_1-frozen transfer learning after Stage 1 transfer learning with a fixed (100%) mammography training set. "C. Stage 2 (DBT:C_1-F_4)" denotes Stage 2 C_1-to-F_4-frozen transfer learning after Stage 1 transfer learning with a fixed (100%) mammography training set (reprint with permission [49])

out the second-stage fine-tuning with DBT. The test AUC increased steadily as the training sample size increased. For a given training set size, the test AUC decreased as more and more layers were frozen, indicating that the learning capacity of the DCNN was restricted and insufficient knowledge was learned from the mammography data. Furthermore, the test AUC on the DBT set (Fig. 3) was higher than that on the mammography test set (Fig. 2) at the corresponding training sample size and frozen layers, indicating that mammography is an effective auxiliary domain for transfer training to DBT and that malignant and benign masses in DBT are easier to be distinguished by DCNN, similar to that by human vision.

Figures 4 and 5 compared the two-stage transfer learning to one-stage transfer learning on the classification of masses in the DBT test set. Several observations can be made. First, the test AUC increases with training sample size either in stage 1 or stage 2. Second, when the training set in the target domain is small, the additional stage of pre-training with data of auxiliary domain can improve the overall performance at all training sample sizes in the range studied (compare curves A and B in Fig. 4, and curves B and D in Fig. 5). Third, when too many layers are frozen during transfer learning, the performance of the DCNN after two-stage training may not reach the same level as that of the DCNN with less layers frozen using the same training sample sizes (compare curves B and C in Fig. 4), indicating that the DCNN cannot learn adequately from the training data if the learning capacity of the DCNN is overly restricted. Fourth, on the other hand, when the DCNN is well trained in the source domain and the training set in the target domain is very small, freezing most of the layers during transfer training may be beneficial by avoiding the loss of the pretrained knowledge without adequate learning from the target domain data (compare the small sample size region of curves B and C in Fig. 5). Although one can expect that the training sample size required for transfer training for a

given task will depend on many factors such as the complexity of the tasks and the DCNN structure, the differences in the characteristics between the source and the target domains, the relative training sample sizes between the tasks, the relative trends observed from this study will likely be applicable to many transfer learning applications, and multi-stage transfer training with data from similar domains should be helpful if the training data of the target domain is too scarce.

Data Augmentation

Data augmentation generates multiple slightly different versions of images from each image in the original training set. Data augmentation may use techniques such as flipping the image in various directions, translating the image within a range of distance, cropping the image in different ways, rotating the image within a range of angles, scaling the image over a range of factors, generating shape- and intensity-transformed images by linear or non-linear methods. Data augmentation can be implemented on-line or off-line and an augmentation operation in a specified range can be performed randomly or by fixed increments. For off-line augmentation, the augmented versions of the images are pre-generated and mixed with the original data into a larger training set, which is randomly grouped as mini-batches for the DCNN training. If the various techniques are applied in combinations, the apparent number of training images can increase easily to hundreds or thousands of times. For on-line augmentation, the various augmentation techniques are usually implemented as a part of the DCNN pipeline with user-selectable probability and range. The original training set is input in mini-batches but each image in a batch is randomly altered according to the pre-selected probability and range of the augmentation techniques. The number of times an image is augmented in a given training run will depend on the number of training epochs chosen and the pre-selected probabilities for the different augmentation techniques. The choice between off-line and on-line augmentation may depend on the tradeoffs between computational resources and storage space or memory; off-line augmentation is more practical if the available training set is small as it requires more space and memory for the augmented set, while on-line augmentation is preferred for large training sets if computational resource is plentiful. Data augmentation introduces variations or jittering to the original data, thereby reducing the risk of overfitting to a small training set and improving generalizability [28, 50, 51]. However, it is important to note that augmenting the training set to a certain size is not equivalent to having a set of independent training samples of comparable size. Since the features in the augmented versions of an image are highly correlated and the CNN learning is invariant to many of these small variations, the augmented images do not provide much new knowledge for the DCNN to learn in comparison to new independent images. In particular, if the original small training set does not include representative samples of certain characteristics, data augmentation will not generate samples of the missing types for the DCNN training; for example, if there are no spiculated nodules in the original data, augmentation cannot generate spiculated nodules for the DCNN to learn. Other more sophisticated data augmentation methods are also being considered, such as generative adversarial networks (GANs) that can generate images with mixed features learned from different images after training on the available sample images [52], digitally generate artificial lesions inserted into normal images [53, 54], or inserting real lesions to other locations of normal or abnormal images [55]. Investigations are needed to evaluate issues such as how effective the data augmentation methods are compared to one another for a given sample size, whether the features learned from the artificial lesions, especially texture features, may help or hinder DCNN learning of real features, whether real lesions inserted at other locations can contribute new features to learn, and whether the usefulness of the augmented data for deep learning depends on the target task.

Training, Validation, and Independent Testing

During training of a machine learning model including deep learning, a validation set is generally used for guiding the optimization of the parameters. The validation set is used to compare the performances of models with different parameter sets, or to monitor the changes in the performance or cost during iterative training of the model weights. The validation set may be split from the training set by cross validation or by hold-out. Regardless of the methods, the validation set is a part of the training process because it is repeatedly used to guide training, and the model structure and parameters are usually chosen to maximize the performance on the validation set. It is well known in machine learning that the training or validation performance is generally optimistically biased [56–61]. To estimate the true performance of the trained model in unknown cases, one has to use an independent test set that has not been seen by the model in the training process and is representative of the population to which the trained model will be applied. To date, most of the published studies only include cross validation results, and even in studies with a "hold-out" test set, the test set will be turned into a validation set if the same test set is used for evaluation many times during model development and eventually the best model is chosen based on the performance of the test set. The American Association of Physicists in Medicine (AAPM) CAD Subcommittee (renamed as Computer-Aided Image Analysis Subcommittee in 2018) has published an opinion paper to discuss the training and evaluation methodology for development of CAD systems [62]. The importance and the strategy of collecting a representative independent test set and the potential biases on the reported "test" result due to multiple repeated use of the same test set are discussed in more details. Deep-learning-based CAD or AI follows similar general principles as conventional machine learning methods, and the need for independent testing will be even more important due to the vast capacity of deep learning to extract and memorize information from the training set.

Acceptance Testing, Preclinical Testing, and User Training

If properly trained with a large data set, deep learning is expected to be more robust and more accurate than conventional machine learning approaches. However, studies showed that deep learning in medical imaging, or machine learning in general, can learn non-medical features that are not related to the medical conditions of the patient but other properties such as image acquisition protocols or equipment, image processing techniques, or even other markings and accessories related to the facilities or patient comorbidity that are recorded in the images [63]. As a result, a deep learning algorithm well trained and independently tested showing high accuracy using data collected from the same site(s) may not be generalizable to different clinical sites that may have different population or imaging characteristics. Even if a CAD or AI algorithm is approved by FDA for clinical use, a clinical site should conduct acceptance testing, similar to the installation of a new medical device or equipment, using a set of representative local data to verify that its performance for the local patient population can pass a certain standard or reference level before clinical implementation. In addition, after the AI tool is implemented in the clinical workflow, the users should allow for a test period in which they refrain from being influenced by the CAD output. The users should familiarize themselves with the output of the CAD tool and quantitatively, if possible, assess the performance of the CAD tool on a large number of consecutive clinical cases. The users should evaluate critically the strengths and weaknesses of the CAD tool based on follow-up review of the outcomes of the cases, so as to recognize the characteristics of cases that the CAD tool makes mistakes or the CAD tool makes correct recommendations whereas the clinician may have failed. The hands-on experience of the performance of the CAD tool will allow the users to learn how to reduce the risk of accepting erroneous recommendation while taking advantage of the recommendations for cases that the CAD tool is useful. The test period

will serve both as a real world evaluation of the CAD tool on the local population and user training. With better understanding of the AI's limitation and capability, the users may be able to establish proper expectation and confidence level on the CAD tool and thus reducing the risk of improper use or negative outcomes of using CAD.

Quality Assurance and Performance Monitoring

With the new generation of CAD, there are high expectations that they will be far more robust than the conventional CAD systems, especially that many of the studies reported performance higher than those of clinicians. Although the initial concerns of AI algorithms replacing radiologists have tamped down, the expectations of using AI to improve workflow efficiency or reduce workload are prevalent. A recent observer study [64] compared breast cancer detection in DBT by radiologist alone to radiologist using deep-learning-based CAD as a concurrent reader that marked suspected lesions and showed the confidence of malignancy on the DBT slices. A data set of 260 DBT cases including 65 cancer, 65 benign, and 130 normal cases were read by 24 radiologists. The experimental concurrent CAD had a case-based sensitivity of over 90% and a specificity of over 40%, which are higher than all of the CAD tools currently used in screening DM. They demonstrated that reading with CAD could provide all the benefits a radiologist would hope for: reducing the average reading time by more than 50% for a DBT case, increasing sensitivity and specificity, as well as reducing recall rate. In another study [65], researchers developed a DCNN to identify normal mammograms from screening cases. With 10-fold cross validation, they showed that the DCNN could identify 34% and 91% of the normal mammograms at a negative predictive value (NPV) of 0.99 for a cancer prevalence of 15% and 1%, respectively. The study showed the potential of using DCNN to improve radiologists' workflow efficiency by excluding the negative mammograms from reading.

For an AI model to be a useful routine clinical tool, it is crucial to validate that its performance in clinical settings can meet certain standards and is consistent over time, similar to other medical devices, especially for any AI model that is designed to operate as a decision maker, rather than as a decision support tool or a second opinion. The acceptance testing or preclinical testing described above can serve as the baseline performance on the local population. Since the performance of DCNN is affected by the properties of the input images, which may be determined by a number of factors such as the imaging techniques or equipment and the image processing or reconstruction software or parameters that may change intentionally or unintentionally due to many factors, periodic quality assurance (QA) procedures should be established to monitor the performance of the CAD tool as well as the performance of clinicians using CAD over time. The AAPM CAD Subcommittee has published an opinion paper on the quality assurance and user training on CAD devices in clinical use [66]. The discussions have not attracted much attention, probably because of the limited use of CAD in the clinic at that time. Currently FDA has no post-market monitoring and regulations on the consistency or accuracy of CAD software as second opinion in clinical use after it is approved and there is no control of off-label use. As CAD/AI tools are anticipated to have widespread use in health care in the future, either as second opinion or automated decision maker in some applications such as pre-screening or triaging, their impact on patient care or welfare can be much greater. It will be important for organizations such as the American College of Radiology (ACR), the Radiological Society of North America (RSNA), the European Society of Radiology, and AAPM to provide leadership to establish performance standards, QA and monitoring procedures, and compliance guidance, to ensure safety and effectiveness for implementation and operation of CAD tools in clinical practice.

Interpretability of CAD/AI Recommendations

The DCNN learns multiple levels of feature representations from the input data by using the deep architecture of convolutional layers. At present a DCNN model is mostly operated like a black-box as there is no easy way to explain how and what the DCNN has learned to perform a specific classification task. Researchers have developed methods to visualize the feature maps at each convolutional layer [67, 68] and to highlight the target objects recognized by the DCNN with a class activation map [69]. The feature maps illustrate the deep features [70] extracted by the DCNN and the class activation map may be correlated with the target location or the locations of the most important features for classification. These visualization tools are the first steps to explore the inner workings of deep learning but they are still far from being able to translate the deep learning output to interpretable clinical decisions, especially for tasks more complex than lesion detection. For CAD/AI to be more widely acceptable as a clinical decision support tool, it should be able to more intelligently present the recommendation to clinicians with reasons, correlating the findings with the medical conditions and data of the patient, and ideally, be able to present further explanations if the clinician has questions on the recommendation. Uncovering the relationship between the machine findings with medical conditions of the patient or even utilizing deep learning and big data analytics to discover new links between disease and clinical data or symptoms will be an important area of research to enable CAD to deliver interpretable diagnosis to clinicians and advance CAD towards true AI in medicine.

Summary

Deep learning is expected to revolutionize CAD and image analysis in medicine. Although machine learning has been applied to CAD and medical image analysis for over three decades, CAD has not been commonly used in the clinic due to the limited performance of conventional machine learning approaches. The recent success of deep learning technology spurs new efforts to develop CAD or AI tools for many applications in health care. Numerous studies have reported promising results. Amid the high expectations of the accuracy and efficiency that AI can bring to medicine, many challenges have yet to be overcome in order to integrate the new generation of CAD tools into clinical practice and to minimize the risk of unintended harm to patients. The discussion in this chapter is not limited to computer-aided lesion detection. Similar considerations are applicable to any CAD tools in general, such as those for disease characterization, staging, treatment planning, surgical guidance, treatment response assessment, recurrence monitoring, and prognosis or survival prediction. Big databases have to be collected to provide sufficient training and validation samples to develop robust deep learning models and independent testing with internal and external multi-institutional data to assess generalizability; performance standards, acceptance testing, and quality assurance procedures should be established for each type of applications to ensure the performance of a deep learning model can meet the requirements in the local clinical environment and remains consistent over time; adequate user training in local patient population is vital to allow users to understand the capability and limitations of the CAD tool, establish realistic expectations, and avoid improper use or disillusion; CAD recommendation has to be interpretable to allow clinicians to make informed decisions. More importantly, workflow efficiency and costs are major considerations in health care. A decision support tool will not be acceptable if it requires additional time and/or costs without significant clinical benefits. It is important for CAD researchers and developers to understand the preferred mode of assistance by clinicians for each type of clinical tasks, design effective CAD tools, and deliver interpretable outputs by taking into consideration the practical issues in clinical settings. If properly developed, validated, and implemented, it can be expected that the

efficient data analytics from CAD or AI tools can complement the human intelligence of clinicians to improve the accuracy and workflow and thus patient care.

Acknowledgment This work is supported by National Institutes of Health award number R01 CA214981.

Disclosures The authors have no conflicts to disclose.

References

1. Winsberg F, Elkin M, Macy J, Bordaz V, Weymouth W (1967) Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. Radiology 89:211–215
2. Kimme C, O'Laughlin BJ, Sklansky J (1977) Automatic detection of suspicious abnormalities in breast radiographs. Data structures, computer graphics and pattern recognition. Academic Press, New York
3. Spiesberger W (1979) Mammogram inspection by computer. IEEE Trans Biomed Eng 26:213–219
4. Semmlow JL, Shadagopappan A, Ackerman LV, Hand W, Alcorn FS (1980) A fully automated system for screening mammograms. Comput Biomed Res 13:350–362
5. Doi K (2015) Chapter 1. Historical overview. In: Li Q, Nishikawa RM (eds) Computer-aided detection and diagnosis in medical imaging. Taylor & Francis Group, LLC, CRC Press, Boca Raton, FL, pp 1–17
6. Chan H-P, Doi K, Galhotra S, Vyborny CJ, MacMahon H, Jokich PM (1987) Image feature analysis and computer-aided diagnosis in digital radiography. 1. Automated detection of microcalcifications in mammography. Med Phys 14:538–548
7. Chan H-P, Doi K, Vyborny CJ, Schmidt RA, Metz CE, Lam KL et al (1990) Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis. Investig Radiol 25:1102–1110
8. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444
9. Fukushima K, Miyake S (1982) Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recogn 15:455–469
10. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W et al (1990) Handwritten digit recognition with a back-propagation network. Proc Adv Neural Inf Process Syst:396–404
11. Lo SCB, Lin JS, Freedman MT, Mun SK (1993) Computer-assisted diagnosis of lung nodule detection using artificial convolution neural network. Proc SPIE 1898:859–869
12. Lo SCB, Chan H-P, Lin JS, Li H, Freedman M, Mun SK (1995) Artificial convolution neural network for medical image pattern recognition. Neural Netw 8:1201–1214
13. Chan H-P, Lo SCB, Helvie MA, Goodsitt MM, Cheng SNC, Adler DD (1993) Recognition of mammographic microcalcifications with artificial neural network. Radiology 189(P):318
14. Chan H-P, Lo SCB, Sahiner B, Lam KL, Helvie MA (1995) Computer-aided detection of mammographic microcalcifications: pattern recognition with an artificial neural network. Med Phys 22:1555–1567
15. Chan H-P, Sahiner B, Lo SCB, Helvie MA, Petrick N, Adler DD et al (1994) Computer-aided diagnosis in mammography: detection of masses by artificial neural network. Med Phys 21:875–876
16. Sahiner B, Chan H-P, Petrick N, Wei D, Helvie MA, Adler DD et al (1995) Image classification using artificial neural networks. Proc SPIE 2434: 838–845
17. Wei D, Sahiner B, Chan H-P, Petrick N (1995) Detection of masses on mammograms using a convolution neural network. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing. 95CH2431-5, pp 3483–3486
18. Sahiner B, Chan H-P, Petrick N, Wei D, Helvie MA, Adler DD et al (1996) Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging 15:598–610
19. Zhang W, Doi K, Giger ML, Wu Y, Nishikawa RM, Schmidt RA (1994) Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. Med Phys 21:517–524
20. Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
21. Bengio Y, Lamblin P, Popovici D, Larochelle H (2006) Greedy layer-wise training of deep networks. Proc Adv Neural Inf Process Syst 19:153–160
22. Erhan D, Bengio Y, Courville A, Manzagol P-A, Vincent P, Bengio S (2010) Why does unsupervised pre-training help deep learning? J Mach Learn Res 11:625–660
23. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning, pp 807–814
24. Glorot X, Bordes A, Yoshua B (2011) Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, pp 315–323

25. Ranzato MA, Huang FJ, Boureau Y-L, LeCun Y (eds) (2007) Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 17–22 June 2007. Minneapolis, MN, USA
26. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958
27. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning (ICML'15), vol 37, pp 448–456. arXiv:1502.03167
28. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst:1097–1105
29. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3): 211–252
30. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778. arXiv:1512.03385
31. Sun C, Shrivastava A, Singh S, Gupta A (2017) Revisiting unreasonable effectiveness of data in deep learning era. arXiv:1707.02968
32. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
33. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH et al (2019) Deep learning in medical imaging and radiation therapy. Med Phys 46(1):e1–e36
34. Mazurowski MA, Buda M, Saha A, Bashir MR (2019) Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. J Magn Reson Imaging 49(4):939–954
35. Fauw JD, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S et al (2018) Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 24(9):1342–1350
36. Janowczyk A, Madahushi A (2016) Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform 7:29
37. Chan H-P, Hadjiiski LM, Samala RK (2019) Computer-aided diagnosis in the era of deep learning. Med Phys (accepted)
38. Cole EB, Zhang Z, Marques HS, Hendrick RE, Yaffe MJ, Pisano ED (2014) Impact of computer-aided detection systems on radiologist accuracy with digital mammography. AJR Am J Roentgenol 203:909–916
39. Gilbert FJ, Astley SM, Gillan MGC, Agbaje OF, Wallis MG, James J et al (2008) Single reading with computer-aided detection for screening mammography. N Engl J Med 359(16):1675–1684
40. Gromet M (2008) Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. Am J Roentgenol 190(4):854–859. https://doi.org/10.2214/ajr.07.2812
41. Taylor P, Potts HWW (2008) Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer 44:798–807
42. Fenton JJ, Abraham L, Taplin SH, Geller BM, Carney PA, D'Orsi C et al (2011) Effectiveness of computer-aided detection in community mammography practice. J Natl Cancer Inst 103(15):1152–1161. https://doi.org/10.1093/jnci/djr206
43. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL (2015) Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 175(11):1828–1837
44. Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A et al (2018) Natural language-based machine learning models for the annotation of clinical radiology reports. Radiology 287(2):570–580. https://doi.org/10.1148/radiol.2018171093
45. Yan K, Wang X, Lu L, Summers RM (2018) DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J Med Imaging 5(3):036501
46. Oakden-Rayner L (2017) Exploring the ChestXray14 dataset: problems.https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/
47. The Digital mammograph DREAM Challenge (2017). https://www.synapse.org/#!Synapse:syn4224222/wiki/434547
48. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In: Proceedings of the Advances in neural information processing systems (NIPS'14), pp 3320–3328
49. Samala RK, Chan H-P, Hadjiiski LM, Helvie MA, Richter CD, Cha K (2019) Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. IEEE Trans Med Imaging 38(3): 686–696
50. Taylor L, Nitschke G (2017) Improving deep learning using generic data augmentation. arXiv:1708.06020
51. Wang J, Perez L (2017) The effectiveness of data augmentation in image classification using deep learning. arXiv:1712.04621
52. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozairy S et al (2014) Generative Adversarial Nets. arXiv:1406.2661v1
53. Badano A, Graff CG, Badal A et al (2018) Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial. JAMA Network Open 1(7):e185474. https://doi.org/10.1001/jamanetworkopen.2018.5474

54. Cha KH, Petrick N, Pezeshk A, Graff CG, Sharma D, Badal A et al (2019) Reducing overfitting of a deep learning breast mass detection algorithm in mammography using synthetic images. Proc SPIE 10950:1095004
55. Pezeshk A, Petrick N, Chen W, Sahiner B (2017) Seamless lesion insertion for data augmentation in CAD training. IEEE Trans Med Imaging 36(4):1005–1015
56. Efron B (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 78:316–331
57. Fukunaga K, Hayes RR (1989) Effects of sample size on classifier design. IEEE Trans Pattern Anal Mach Intell 11:873–885
58. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer-Verlag, New York
59. Bishop CM (1995) Neural networks for pattern recognition. Clarendon Press, Oxford
60. Chan H-P, Sahiner B, Wagner RF, Petrick N (1999) Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers. Med Phys 26:2654–2668
61. Sahiner B, Chan H-P, Petrick N, Wagner RF, Hadjiiski LM (2000) Feature selection and classifier performance in computer-aided diagnosis: the effect of finite sample size. Med Phys 27:1509–1522
62. Petrick N, Sahiner B, Armato SG, Bert A, Correale L, Delsanto S et al (2013) Evaluation of computer-aided detection and diagnosis systems. Med Phys 40(8):087001. https://doi.org/10.1118/1.4816310
63. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLOS Med 15(11):e1002683
64. Conant EF, Toledano AY, Periaswamy S, Fotin SV, Go J, Hoffmeister JW et al (2018) Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis screening. In: Radiological Society of North America Scientific Assembly and Annual Meeting. RC215-14
65. Kyono T, Gilbert FJ, van der Schaar M (2019) Improving workflow efficiency for mammography using machine learning. J Am Coll Radiol. https://doi.org/10.1016/j.jacr.2019.05.012
66. Huo ZM, Summers RM, Paquerault S, Lo J, Hoffmeister J, Armato SG et al (2013) Quality assurance and training procedures for computer-aided detection and diagnosis systems in clinical use. Med Phys 40(7):077001. https://doi.org/10.1118/1.4807642
67. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Lecture Notes in Computer Science Computer Vision – European Conference on Computer Vision (ECCV) 2014, vol 8689, pp 818–833
68. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. arXiv:1506.06579v1
69. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2921–2929
70. Samala RK, Chan H-P, Hadjiiski LM, Helvie MA, Cha KH, Richter CD (2017) Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. Phys Med Biol 62:8894–8908