BJR

https://doi.org/10.1259/bjr.20190580

Received: 02 July 2019 | Revised: 13 November 2019 | Accepted: 17 November 2019

Cite this article as:
Chan H-P, Samala RK, Hadjiiski LM. CAD and AI for breast cancer—recent development and challenges. *Br J Radiol* 2019; **92**: 20190580.

REVIEW ARTICLE

CAD and AI for breast cancer—recent development and challenges

HEANG-PING CHAN, PhD, RAVI K. SAMALA, PhD and LUBOMIR M. HADJIISKI, PhD

Department of Radiology, University of Michigan, Ann Arbor, MI, United States

Address correspondence to: Dr Heang-Ping Chan
E-mail: *chanhp@umich.edu*

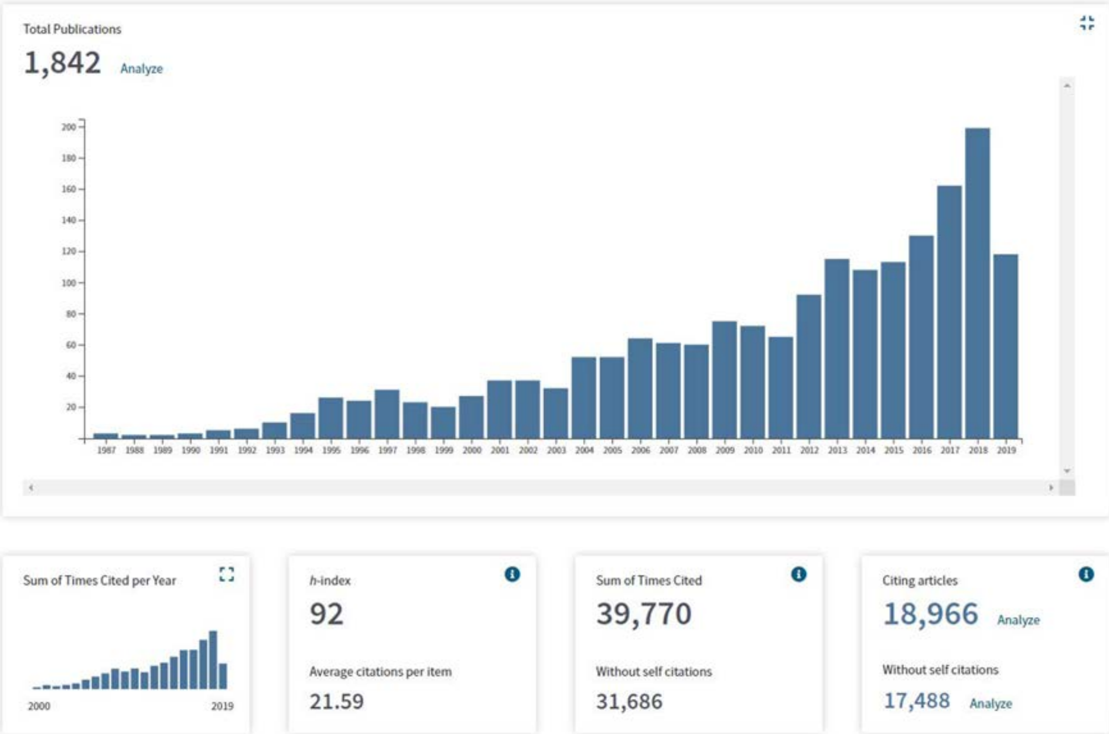
ABSTRACT

Computer-aided diagnosis (CAD) has been a popular area of research and development in the past few decades. In CAD, machine learning methods and multidisciplinary knowledge and techniques are used to analyze the patient information and the results can be used to assist clinicians in their decision making process. CAD may analyze imaging information alone or in combination with other clinical data. It may provide the analyzed information directly to the clinician or correlate the analyzed results with the likelihood of certain diseases based on statistical modeling of the past cases in the population. CAD systems can be developed to provide decision support for many applications in the patient care processes, such as lesion detection, characterization, cancer staging, treatment planning and response assessment, recurrence and prognosis prediction. The new state-of-the-art machine learning technique, known as deep learning (DL), has revolutionized speech and text recognition as well as computer vision. The potential of major breakthrough by DL in medical image analysis and other CAD applications for patient care has brought about unprecedented excitement of applying CAD, or artificial intelligence (AI), to medicine in general and to radiology in particular. In this paper, we will provide an overview of the recent developments of CAD using DL in breast imaging and discuss some challenges and practical issues that may impact the advancement of artificial intelligence and its integration into clinical workflow.

INTRODUCTION

The success of deep learning (DL) in many applications such as speech and text recognition, natural language processing, chess and Go game, object detection and classification in recent years opens a new era of machine learning and computer vision. The DL approach has since raised unprecedented enthusiasm in various fields of pattern recognition and artificial intelligence (AI) including computer-aided diagnosis (CAD) in medicine. CAD makes use of machine learning methods and multidisciplinary knowledge and techniques to analyze medical imaging data and/or non-imaging data and provides the analyzed results to clinicians as second opinion or decision support in the various stages of the patient care process such as lesion detection, characterization, disease risk prediction, cancer staging, treatment planning and response assessment, recurrence and prognosis prediction. CAD has been a major field of research and development in medical imaging. CAD tools developed with conventional machine learning methods mainly use hand-engineered features based on the domain knowledge and expertise of human developers, who translate the perceived image characteristics to descriptors that can be implemented with mathematical functions or conventional image processing techniques. The manually designed descriptors may not be able to capture the intricate differences between the normal and abnormal clinical conditions, and therefore may not generalize well to the wide range of variations in the patient population. The performance of CAD tools often can reach high sensitivity but at the cost of a relatively high false-positive rate. There are high expectations that the recent advances in machine learning techniques will overcome some of these challenges and bring significant improvement in the performance of CAD in medical imaging. There are also expectations that DL-based CAD or AI may advance to a level that it may automate some processes such as triaging cases for clinical care or identify negative cases in screening to help improve the efficiency and workflow. A previous article has reviewed the early CAD systems for breast cancer using DL, explained their superiorities relative to previously established systems, defined the methodologies including algorithmic developments, described remaining challenges in breast cancer screening and diagnosis, and discussed possible future directions for new CAD models.[1] In this paper, we will review the advances in CAD from past experiences to the promises brought about by DL, discuss the challenges in CAD development and clinical

Figure 1. The number of publications per year obtained from searching the Web of Science, "Science Citation Index Expanded," "Book Citation Index-Science," and "Emerging Sources Citation Index," with keywords: (breast imaging) AND (machine learning OR deep learning OR convolutional neural network OR deep neural network OR computer aid OR computer assist OR computer-aided diagnosis OR automated detection OR computerized detection OR automated classification OR computerized classification OR decision support OR radiomic), search period 1900 to 6/2019.



implementation, and consider some practical issues to assure the generalizability and reliability of CAD as decision support tools for clinicians in breast imaging applications.

CAD in breast imaging—past experience and future goals

Studies of automated analysis of radiographic images with computers emerged in the 1960's. Several investigators have attempted to automatically detect breast abnormalities.[2–5] These early attempts demonstrated the feasibility but did not attract much attention, probably because the accuracy was limited by computational resources and access to high quality digitized image data. Systematic development of machine learning techniques for medical imaging began in the 1980's,[6] with a more realistic goal to develop CAD systems as a second opinion to assist radiologists in image interpretation rather than automation. The first observer performance study conducted by Chan et al[7] using a CAD system developed by the same investigators[8] showed that breast radiologists' detection performance of microcalcifications was significantly improved when reading with CAD. The study demonstrated the potential of CAD in improving the detection of early stage breast cancer. The Food and Drug Administration (FDA) approved the first commercial CAD system as a second opinion for screening mammography in 1998. The research and development of CAD methods for various diseases and imaging modalities have been steadily growing over the years. Many retrospective observer studies demonstrated that radiologists' performance improved with CAD.[6] Figure 1 shows the number of peer-reviewed journal publications related to CAD and machine learning for all breast imaging modalities obtained by searching the Web of Science up to mid-2019, including the work for various CAD applications such as detection, characterization, risk prediction and radiomics. The growing trend in computer-aided image analysis related to breast imaging is evident and the growth speeds up in the last few years, probably spurred by DL.

CAD was introduced into screening mammography two decades ago. A number of prospective studies have been conducted to compare radiologists reading with and without CAD, or compare single radiologist reading with CAD to double reading in screening mammography. The reported effects of CAD in screening mammography varied. Taylor et al[9] conducted a meta-analysis of studies comparing single reading with CAD or double reading to single reading (Table 1). They concluded that double reading with arbitration increased cancer detection rate per 1000 females screened (CDR) and CAD did not significantly increase the CDR. Double reading with arbitration reduced recall rate but double reading with unilateral or a mixed strategy had much higher recall rates than single reading with CAD. These results indicate that double reading, regardless of with another radiologist or with a computer aid, will increase FP recalls unless the additional detections are properly scrutinized to dismiss potential lesions of low suspicion.

Table 1. Meta-analysis of pooled odds ratios of increase in cancer detection rate per 1000 females screened and the increase in recall rate obtained by comparing single reading with CAD or double reading to single reading

	Single reading with CAD		Double reading		
	Matched $N = 5$	Unmatched $N = 5$	Unilateral $N = 6$	Mixed $N = 3$	Arbitration $N = 8$
Odds ratio of increase in cancer detection rate	1.09	1.02	1.13	1.07	1.08
Odds ratio of increase in recall rate	1.12	1.10	1.31	1.21	0.94

CAD, computer-aided diagnosis.
Matched studies: the assessment before and after using CAD was on the same mammogram.
Unmatched studies: the performance of mammography facilities was compared before and after the introduction of CAD. Different mammograms are interpreted in the two conditions.
N is the number of studies included in each group.[9]

Although the pooled results by Taylor et al[9] did not show significant improvement in CDR for single reading using CAD, the study revealed that the performance of radiologists using CAD varied over a wide range. The change in CDR ranged from 0 to 19% and the increase in recall rate varied from 0 to 37%. These variations may be attributed to factors such as differences in study design (Table 1), user training, the experience and confidence of the radiologists in differentiating true and false CAD marks, and whether the radiologists used CAD properly as second reader as it was designed and approved to be. In two prospective clinical trials[10,11] that had better controls for comparing single reading with CAD to double reading (Table 2), Gilbert et al found that the sensitivities of the two approaches were comparable but the recall rate of the former was higher (3.9% *vs* 3.4%), while Gromet found that single reading with CAD was superior with higher sensitivity and lower recall rate. Both studies concluded that single reading with CAD may be an alternative to double reading.

Although CAD was approved by FDA as a second opinion, there is no monitoring of how radiologists use CAD in the clinic. Fenton et al analyzed the data from 43 mammography sites in three states before and after CAD implementation in 2007[12] and a follow-up study in 2011.[13] They found that the increase in recall rate decreased from 30 to 6%, while the increase in CDR decreased from 4.5 to 1.8% between the two studies. They observed that "radiologists with variable experience and expertise may use CAD in a nonstandardized idiosyncratic fashion," and "Some community radiologists, *e.g.* may decide not to recall females because of the absence of CAD marks on otherwise suspicious lesions." Lehman et al[14] compared 271 radiologists in 66 facilities before and after implementation of CAD. They found that the average sensitivity decreased by 2.3% and the recall rate increased by 4.5% with the use of CAD. They noted 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." These comments indicated that some radiologists may have used CAD prematurely as a concurrent reader to speed up reading while CAD was approved only as a second opinion. On the other hand, some studies showed that radiologists may overlook true positive CAD marks amid the large number of false positives they have to dismiss per 1000 cases to detect an additional cancer as the breast cancer prevalence is generally less than 1%.[15–17] These clinical experiences of CAD reveal that, useful CAD tools in the clinic should be either those significantly increasing workflow efficiency without reducing sensitivity or specificity, or those significantly improving clinical efficacy without impeding workflow, although ideally delivering both. A mismatch of the performance levels of CAD with the expectations and the need of the clinicians will increase the risk of misuse and negative outcomes. The recent success of DL over conventional machine learning approaches in many AI applications may offer new opportunities to improve the performance of CAD tools and meet the high expectations of achieving these goals.

Deep learning driven CAD development in breast imaging

DL is a type of representation learning method that can discover representations of data automatically by transforming the input

Table 2. Two prospective clinical trials that compared double reading to single reading with CAD

	Single reading (first read of double reading)				Double reading				Single reading with CAD			
	Sens	Recall rate	PPV3	CDR	Sens	Recall rate	PPV3	CDR	Sens	Recall rate	PPV3	CDR
Gromet[11] *2008*	81.4%	10.2%	30.6%	4.12	88.0%	11.9%	29.8 %	4.46	90.4%	10.6%	27.8%	4.20
Gilbert et. al[10] *2008*					87.7%	3.4%	21.1%	7.06	87.2%	3.9%	18.0%	7.02

CAD, computer-aided diagnosis; CDR = cancer detection rate per 1000 females screened.
Gromet et al.[11]: single center, nine radiologists, $N_{\text{single, double}}$ = 112,413, $N_{\text{single+CAD}}$ =118,808. CAD system alone: 81.7% sensitivity at 2.8 FPs/case
Gilbert et al.[10]: *CADET II* study, three centers, two arms reading matched cases N_{double} = $N_{\text{single+CAD}}$ = 28,204. CAD system alone: sensitivity (mass) 88% at 1.5 FPs/case, sensitivity (calcification) = 95% at 1 FPs/case

information into multiple layers of abstractions in a deep neural network architecture.[18] By training with a large data set and an appropriate cost function, the multiple layers of weights in the deep neural network are iteratively updated, resulting in a complex mathematical model that can extract relevant features from the input data with high selectivity and invariance. DL has led to significant advancements in many automated or computer-assisted tasks such as target detection and characterization, speech and text recognition, face recognition, autonomous vehicles, smart devices, and robotics.

Deep learning convolutional neural networks (DCNN) are the most popular method for pattern recognition and computer vision applications in image analysis at present. Convolutional neural networks (CNN) originated from neocognitron proposed in the early 1980's.[19] CNN was introduced into medical image analysis in 1993[20,21] and applied to microcalcification detection on mammograms in the same year,[22,23] and subsequently to mass detection.[24–27] A similar shift-invariant neural network was applied to the detection of clustered microcalcifications in 1994.[28] These early CNNs were relatively shallow but they demonstrated the feasibility of using CNN in medical images. In 2012, Krizhevsky et al[29] designed a DCNN with five convolutional layers (called "AlexNet"). Using the "ImageNet" data set containing over 1.2 million photographic images for training, the AlexNet achieved top performance and outperformed all previous methods at the ImageNet Large Scale Visual Recognition Challenge for classification of over 1000 classes of everyday objects (cars, animals, planes, etc). The performance of DCNN was shown to increase with depth for some tasks[30] and deeper and deeper DCNNs have been proposed since then.

DCNN has been applied to CAD for breast imaging in recent years; the main areas to date include detection and classification of microcalcifications or masses, characterization of cancer subtypes, breast density estimation and classification. The majority of the studies were conducted with mammographic images, a substantial number of studies used ultrasound images, but only a few studies used magnetic resonance (MR) images, likely because of the differences in the availability of imaging data. We summarize the studies reported in peer-reviewed journals for the three modalities in (Tables 3–5) except for some papers that appeared too preliminary with very few training samples. In the tables, we include the number of training and validation samples, and whether there was independent test set for performance evaluation. The training sample size is an important factor that impacts the robustness of the trained model, while testing with a true independent set is an important step to evaluate the generalizability of the trained model to unseen cases. Many of the studies have multiple comparisons with traditional methods or different DCNN approaches. To keep this paper concise, we tabulated the main proposed method and key results in the tables and do not discuss the approach of individual papers. Interested readers may refer to the original paper for the detailed description of each study. We will briefly summarize some observations for DL studies in each modality in the following.

Deep learning in mammography

There have been a number of studies applying DCNN to mammography for detection[31–33] or classification[34–36] of microcalcifications (Table 3(A)), and detection[37–39] and classification[40–63] of masses (Table 3(B)). Another common application of DCNN is the segmentation of breast density and classification of the breast in terms of BI-RADS density categories or dense-*vs*-non-dense[64–72] (Table 3(C)). Although most of the DCNNs used in these studies adapted the structural framework from the AlexNet,[29] the VGG nets by the Visual Geometry Group,[93] different versions of Inception by Google,[94,95] and different versions of ResNet by Microsoft,[30] there were variations in how the hyperparameters or the kernels and layers in the original structure were modified, especially the number of fully connected layers near the output for a specific classification task. Some studies proposed more complex structures by adding parallel channels or branches of networks to perform auxiliary functions. Many of the modifications were designed based on the image characteristics of the specific task of interest ("target task"). In some studies, a DCNN pre-trained in other image domain, with or without being fine-tuned with the target domain images, was used as feature extractor and the extracted deep features were trained with an external classifier such as support vector machine (SVM) or random forest for the target task. The studies show that different DCNN approaches can be trained to accomplish the same task, and generally obtain good performance for the specific data sets used.

Digital breast tomosynthesis (DBT) is increasingly being used for breast cancer screening, either standalone or in combination with two-dimensional mammography. A few studies have been conducted with DBT to detect microcalcifications or masses, and classification of masses as malignant or benign using DCNN. Because of the similarity between DBT and mammography and that mammographic images are more readily available, Samala et al[37,56] showed that an intermediate stage of fine-tuning with mammographic images was useful for transfer learning in DBT tasks. Contrast-enhanced spectral mammography or dual-energy contrast-enhanced digital mammography is a relatively new modality for diagnostic work-up, especially for dense breasts, but it has not been commonly implemented in the clinics so that data are scarce. Only two studies have been reported, both had a data set of only about 50 cases,[53,55] to demonstrate the feasibility of using contrast-enhanced spectral mammography or contrast-enhanced digital mammography in DCNN training for mass classification.

Deep learning in breast ultrasound

Ultrasound is an important breast imaging modality for diagnostic work-up to distinguish solid masses from cysts, and for screening in dense breasts. Machine learning methods have been applied to breast ultrasound in various applications.[96–100] An increasing number of DL applications in breast ultrasound has been reported in the past 2 years. We summarize these studies in Table 4. The majority of the studies were related to breast mass characterization,[44,76–82] followed by mass segmentation,[83–85] and detection.[86,87] The most commonly used DL models for ultrasound were again AlexNet, VGG-19, ResNet, GoogLeNet, and

Table 3(A). Studies using deep learning approach for microcalcification detection and classification in mammography and digital breast tomosynthesis

Journal article	Year	Training set	Validation set	Independenttest set	Convolutional neural network (CNN) structure	Performance* (validation or independent test)
Microcalcification detection						
Samala et al.[31]	2014	78 DBT vols with MC clus (DBT:21PVs, 60° scan)	49 DBT vols with MC clus	104 DBT vols with MC clus, 76 no MC	CNN with two convolution layers	FROC: 85% sens. at 0.71 FP/vol. (view-based), at 0.54 FP/vol (case-based)
Samala et al.[32]	2015	78 DBT vols with MC clus (DBT:11PVs, 30° scan)	49 DBT vols with MC clus	104 DBT vols with MC clus, 76 no MC	CNN with two convolution layers	FROC 85% sens. at 1.72 FP/vol. (view-based), at 0.49 FP/vol (case-based)
Wang et al.[33]	2018	167 cases (300 images)	67 cases (117 images)	158 cases (292 images)	Context-sensitive DNN: 7-conv-layer global CNN and 3-conv-layer local CNN (indiv MC 9 × 9, clus 95 × 95 ROIs) compared to clus-based CNN	FROC cluster-based 85% sens: DCNN with 10 conv layers 0.40 FPI; cluster-based CNN 0.44 FPI; SVM 0.52 FPI
Microcalcification classification						
Wang et al.[34]	2016	1000 images (677B, 323M); 10-fold CV		204 images (97B, 107M): 110 MC, 35 mass, 59 both	Stacked autoencoder (SAE) as feature extractor. SVM feature classifier	AUC(MC)=0.87, AUC(mass)=0.61, AUC(MC&mass)=0.90
Cai et al.[35]	2019	891 images (486M, 405B); 10-fold CV		99 images (54M, 45B)	Fine-tuning of ImageNet-pretrained AlexNet as deep feature extractor. SVM classification of deep features with and without handcrafted features	AUC(M *vs* B)=0.93–0.94
Shi et al[36]	2018	99 mag DMs DCIS (25 upstaged to invasive): 80% training, 20% validation			ImageNet-pretrained VGG16 as feature extractor, logistic regression classifier with feature selection	AUC (DCIS vs-upstaged)=0.70

Table 3(B). Studies using deep learning approach for mass detection and classification in mammography and other breast X-ray modalities.

Journal article	Year	Training set	Validation set	Independenttest set	CNN structure	Performance (validation or independent test)
Mass detection						
Samala et al.[37]	2016	2282 SFM & DM (2461 masses, 3173 FPs), 230 DBT vols (228 masses, 28330 FPs), 4-fold CV		94 DBT vols (89 masses)	Cuda-convNet: Stage 1 training with mammograms, Stage 2 fine-tuning with DBT	AUC(stage1 mam)=0.81; AUC(stage2 DBT)=0.90; FROC: Breast-based 91% sens at 1FP/vol
Kim et al.[38]	2017	154 cases (616 DBT vol, 185M, 431N); 5-fold CV			ImageNet pre-pretrained VGG16 and LSTM depth directional long-term recurrent learning	AUC(DCNN)=0.871, AUC(DCNN + LSTM)=0.919
Jung et al.[39]	2018	Private set: 350 pts (222 DMs) for second pretraining. INbreast: 115 pts (410 DMs), 5-fold CV			ImageNet-pretrained ResNet50 with a feature pyramid network (class subnet, box subnet)	FROC:Sens 0.94 at 1.3 FPI, Sens 0.97 at 3 FPI
Mass Classification						

(Continued)

Table 3 (Continued)

Table 3(B). Studies using deep learning approach for mass detection and classification in mammography and other breast X-ray modalities.

Journal article	Year	Training set	Validation set	Independenttest set	CNN structure	Performance (validation or independent test)
Arevalo et al.[40]	2016	344 cases (736 images, 426B, 310M): 50% training, 10% validation		40%	CNN with one or two conv layers. Also ImageNet-pretrained DeCAF	AUC(CNN)=0.822; AUC(combined with hand-crafted features)=0.826; AUC(DeCAF)=0.79
Jiao et al.[41]	2016	300 images (150B, 150M)	300 images (150B, 150M)		Fine-tuning of ImageNet-pretrained AlexNet as feature extractor. Two SVM classifiers for mid-level level and hi-level features.	Accuracy = 96.7%
Dhungel et al.[42]	2017	INBreast 115 cases, Detection: 410 images, Segmentation & classification: 40 cases (41B, 75M masses); 60% training, 20% validation		20%	Detection: multiscale deep belief network, a cascade of R-CNNs and random forest classifiers	FROC: 90% at 1FPI; AUC(DCNN features)=0.76; AUC(Manually marked mass)=0.91.
Sun et al.[43]	2017	2400 ROIs (100 labeled, 2300 unlabeled)		758 ROIs	DCNN with three convolution layers	AUC = 0.8818, Accuracy = 0.8234
Antropova et al.[44]	2017	DM: 245 masses (113B, 132M); 5-fold CV			ImageNet-pretrained VGG19 as feature extractor, SVM classifier	AUC(maxpool features)=0.81AUC(Fused with radiomic features)=0.86
Samala et al.[45]	2017	SFM & DM 1335 views (ROI: 604M, 941B); 4-fold CV		SFM 907 views (ROI:453M, 456B)	ImageNet-pretrained AlexNet	AUC = 0.82
Kooi et al.[46]	2017	Set 1: (1487M, 73102N); Set 2: (1108M, 696 cysts)	Set 1: (342M, 21913N), Set 2: nested CV		VGG-like DCNN pretrained with Set 1, used as feature extractor on Set 2. Gradient boosting trees classifier.	Malignant-vs-cysts classif. (CC + MLO): AUC(DCNN features)=0.78, AUC(with contrast features)=0.80
Jiao et al.[47]	2018	DDSM 300 images	DDSM 150 images	DDSM 150 images; MIAS set	Joint model of ImageNet-pretrained AlexNet and fine-tuned as feature extractor and parasitic metric learning net.	Accuracy(DDSM)=97.4%; Accuracy(MIAS)=96.7%
Samala et al.[48]	2018	SFM & DM 2242 views (ROI: 1057M, 1397B), DBT 230 vols (ROI: 590M, 550B); 4-fold CV		DBT 94 vols (ROI: 150M, 295B)	ImageNet-pretrained AlexNet, 2-stage transfer learning, pruning	AUC(with pruning)=0.90; AUC(without pruning)=0.88
Chougrad et al.[49]	2018	1529 cases (6116 images) from DDSM, INbreast, BCDR; 5-fold CV		MIAS (113 images)	Compare ImageNet-pretrained VGG16, ResNet, InceptionV3	InceptionV3: AUC = 0.99, Accuracy = 98.23%
Al-masni et. al.[50]	2018	DDSM 600 images (300M, 300B); 5-fold CV			ImageNet-pretrained DCNN with 24 convolutional layers (You-Only-Look-Once detection & classification)	AUC = 0.9645; Accuracy = 97%
Wang et al.[51]	2018	BCDR 736 images; 50% training, 10% validation		40%	Multiview-DCNN: ImageNet-pretrained InceptionV3 as feature extractor with attention map, Recurrent NN for classification	MV-DNN: AUC = 0.882, Accuracy = 0.828; MV-DNN + Attention map: AUC = 0.886, Accuracy = 0.846.
Al-antari et al.[52]	2018	INbreast: 115 cases (410 DMs, 112 masses); 4-fold CV: 75% training, 6.25% validation		18.75%	Detection DCNN (Al-masni et al); segmentation by second DCNN, Classification by simplified AlexNet.	Detection accuracy = 98.96%,AUC(M-vs-B classification)=0.9478

(Continued)

Table 3 (Continued)

Table 3(B). Studies using deep learning approach for mass detection and classification in mammography and other breast X-ray modalities.

Journal article	Year	Training set	Validation set	Independenttest set	CNN structure	Performance (validation or independent test)
Gao et al.[53]	2018	SCNN: 49 CEDM cases; DCNN ResNet50: INbreast 89 cases; 10-fold CV			Shallow-deep CNN (SD-CNN): SCNN generated virtual CEDM of mass. Pretrained ResNet50 as feature extractors for 2-view virtual CEDM and DM, Gradient boosting trees classifier	AUC(DM)=0.87; AUC(DM + virtual CEDM)=0.92
Kim et al.[54]	2018	DDSM (178M. 306B)	DDSM (44M, 77B)	DDSM (170M, 170B)	BI-RADS guided diagnosis network: ImageNet-pretrained VGG16, plus BI-RADS critic network and relevance score	AUC(with B-RADS critic network)=0.841; AUC(without BI-RADS critic network)=0.814
Perek et al.[55]	2019	54 CESM cases with 129 lesions (56M, 73B); 5-fold CV			Fine-tuning (FT) ImageNet-pretrained AlexNet, RawNet without pretraining	Using deep features and BI-RADS features: AUC(FT-AlexNet)=0.907; AUC(RawNet)=0.901
Samala et al.[56]	2019	SFM & DM 2242 views (ROI: 1057M, 1397B), DBT 230 vols (ROI: 590M, 550B); 4-fold CV		DBT 94 vols (ROI: 150M, 295B)	ImageNet-pretrained AlexNet, 2-stage transfer learning	AUC(one-stage fine-tuning with DBT)=0.85; AUC (two-stage fine-tuning with mammo then DBT)=0.91
Mendel et al.[57]	2019	76 cases (2-view DM, DBT, synthetic SM) with 78 lesions (30M, 48B) including 34 masses, 15 ADs, 30 MC clusters; Leave-one-out CV			ImageNet-pretrained VGG19 as feature extractor, SVM classifier	Two-view AUC: all lesions DBT = 0.89, SM = 0.86, DM = 0.81; mass&AD DBT = 0.98; MC DBT = 0.97
Cancer detection (any lesion types)						
Becker et al.[58]	2017	Study 1: (95M, 95N); Study 2: (83M, 513N)	Study 1: (48M, 48N); Study 2: (42M, 257N)	Study 1: BCDR (35M, 35N); Study 2: (18M, 233N)	dANN from commercial “ViDi” image analysis software	AUC(Study 1)=0.79; AUC(Study 2)=0.82;
Carneiro et al.[59]	2017	(1) classification: DDSM 86 cases; (2) detection & classif: INbreast 115 cases	(1) DDSM 86 cases; (2) INbreast 5-fold CV		ImageNet-pretrained ConvNet	Two-view AUC: (1) M-vs-B>0.9 or M-vs-(B + N)>0.9. (2) M-vs-B 0.78; M-vs-(B + N) 0.86
Kim et al.[60]	2018	3101M, 23,530 normal cases (four views/case)	1238 cases (619M)	1238 cases (619M)	DIB-MG: (ResNet with 19 convolutional layers + 2-stage global-average-pooling layer)	AUC(M-vs-(B + N))=0.906
Ribli et al.[61]	2018	DDSM 2620 cases and private DM set 174 cases		INbreast 115 cases	Faster R-CNN: ImageNet-pretrained VGG16 with region proposal network for localizing target	Detection FROC: 90% sensitivity at 0.3 FPI; Classification AUC = 0.95
Aboutalib et al.[62]	2018	DDSM 3294 images, private DM set 1734 images; 6-fold CV		private DM 100 images	ImageNet-pretrained AlexNet, pretrained with DDSM then fine-tuned with DM (best among other variaitons)	AUC(M-vs-recalled B)=0.80; AUC(M-vs-negative&recalled B)=0.74.
Akselrod-Ballin et al.[63]	2019	9611 cases (1049M, 1903 biopsy negative, 247 BI-RADS3, 6412 normals)	1055 cases + 31 FNs	2548 cases + 71 FNs	InceptionResnetV2 without pretraining	AUC(predict M per breast with clinical data)=0.91; AUC(identify normal case per breast with clinical data)=0.85; Identify M in 48% of FNs of radiologists

(Continued)

Table 3 (Continued)

Table 3(C). Studies using deep learning approach for breast density segmentation and classification in mammography.

Journal article	Year	Training set	Validation set	Independenttest set	CNN structure	Performance (validation or independent test)
Breast density segmentation						
Kallenberg et al.[64]	2016	Set1: 493N views; Set2: (226 cases, 442 controls); Set3: (394 cases, 1182 controls); 5-fold CV			Convolutional sparse autoencoder (CSAE). (1) density segmentation (MD); (2) case-vs-control classification (MT)	Correlation coeff. (MD)=0.85; Set3: AUC(MT-CSAE)=0.57; AUC(MT-density)=0.59
Li et al.[65]	2018	478 DMs; 10-fold CV		183 DMs	DCNN with three convolutional layers	DSC = 0.76; Correl coeff. = 0.94
Mohamed et al.[66]	2018	BI-RADS density B and C: 7000 DMs each; 6-fold CV		BI-RADS density B and C: 925 images each	Modified AlexNet: ImageNet-pretrained *vs* training from scratch.	AUC(scratch)=0.94, AUC(pretrained)=0.92
Mohamed et al.[67]	2018	963N cases with 15,415 DMs. BI-RADS density from clinical reports; 70% training, 15% validation		15%	Modified AlexNets for two tasks: (1) BI-RADS B-vs-C, (2) Dense (A&B)-vs-nondense (C&D)	(1) AUC (CC&MLO)=0.92; (2) AUC(CC&MLO)=0.95
Lee et al.[68]	2018	455 DM cases	58 DM cases	91 DM cases	ImageNet-pretrained VGG16	Correl coeff. % density-vs-BI-RADS (radiologist): CC=0.81, MLO=0.79, average=0.85
Wanders et al.[69]	2018	394 cancers, 1182 controls (DMs)		51,400 (301 cancer, 51,099 controls; DMs)	DCNN by Kallenberg et al.[64]	C-index: Texture + vol density = 0.62, vol density = 0.56
Gastounioti et al.[70]	2018	200 pts (1:3 case:control; DMs)	100 pts (1:3 case:control; DMs)	124 pts (1:3 case:control; DMs)	LeNet-like CNN with 29 input channels with texture images, two convolutional layers. DCNN with DM input, five convolutional layers	Case-vs-control: AUC(DCNN-multichannel texture) = 0.90, AUC(DCNN-DM) = 0.63
Ciritsis et al.[71]	2018	70% of 12,392 views (6470 RMLO, 6462 RCC)	30% of 12,392	Set 1: (850 MLO, 882 CC); Set 2: (100 MLO, 100 CC, 2 radiologists' consensus)	DCNN with 13 convolutional layers, four dense layers, output 4 BI-RADS density (A, B, C, D)	Accuracy: Set 1: BI-RADS: 71.0%-71.7%; dense-vs-nondense: 88.6%-89.9%; Set 2: BI-RADS 87.4%-92.2%; dense-vs-nondense 96%-99%
Lehman et al.[72]	2019	27684 cases (41,479 DMs)	8738 DMs	5741 cases (8677 DMs); Clinic test: 10,763 cases	ImageNet-pretrained ResNet18	Test set BI-RADS: Accuracy=77%, kappa=0.67; Dense-vs-nondense: 87%. Clinic test: BI-RADS: Accuracy = 90%, kappa=0.85

AD, architectural distortion; AUC(condition), area under the receiver operating characteristic (ROC) curve for the condition in parenthesis; B, benign; CC, craniocaudal view; CEDM, contrast-enhanced digital mammogram; CESM, contrast-enhanced spectral mammogram; CV, cross validation; DBT, digital breast tomosynthesis; DCIS, ductal carcinoma in situ; DM, digital mammography; DSC, Dice similarity coefficient; FPI, false positives/image; FROC, free-response ROC curve; LSTM, long short-term memory; M, malignant; MC, microcalcification; MLO, mediolateral oblique view; N, normal; SFM, screen-film mammogram; SVM, support vector machine; clus, cluster; indiv, individual; pts, patients; vol, volume.

ImageNet: training data set containing over 1.2 million photographic images from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) for classification of over 1000 classes of everyday objects (cars, animals, planes, etc)

DDSM= Digital Database for Screening Mammography,[73] BCDR= Breast Cancer Digital Repository,[74] INbreast – data set of digital mammograms[75]

Table 4. Studies using deep learning approach for mass segmentation, detection and classification on breast ultrasound (US) images

Journal article	Year	Training set	Validation set	Independent test set	Convolutional neural network (CNN) structure	Performance (validation or independent test)
Mass classification						
Antropova et al.[44]	2017	1125 cases (158M, 967B), 2393 ROIs (415M, 1098B cystic, 880B solid), 5-fold CV			ImageNet-pretrained VGG19 as feature extractor, SVM classifier	AUC(maxpool features)=0.872; AUC(fused with radiomics features)=0.902
Han et al.[76]	2017	6579 masses: (2814M, 3765B)	6579 masses: 10-fold CV	829 masses: (340M, 489B)	ImageNet-pretrained GoogLeNet	AUC = 0.958
Xiao et al.[77]	2018	2058 images (688M, 1370B): 80% training	10 %	10%	ImageNet-pretrained ResNet50, Xception, and InceptionV3	AUC(ResNet50)=0.91, AUC(InceptionV3)=0.91, AUC(combined)=0.93
Zhou et al.[78]	2018	Shear-wave elastography 400 images	45 images	95 images	16-layer DCNN	Accuracy: 95.8%
Lee et al.[79]	2018	Images: Study 1: 143 Study 2: 210	Images: Study 1: 27 Study 2: 40		Stacked Denoising Autoencoder (SDAE) network	Accuracy: Study 1: 82% Study 2: 83%
Huang et al.[80]	2019	Images of BI-RADS categories: (3) 531, (4A) 443, (4B) 376, (4C) 565, (5) 323			ImageNet-pretrained modified VGG-16	Accuracy: 0.734 to 0.998 for the five classes
Byra et al.[81]	2019	582 masses (23% M)	150 masses (23% M)	150 massdetection & classif: INbreastes (23% M)	ImageNet-pretrained VGG19 with fine-tuning (FT) and matching layer (ML) at input	AUC(VGG19 +FT + ML)=0.936, AUC(four radiologists)=0.806 to 0.882
Fujioka et al.[82]	2019	240 masses (144M, 96B) 947 images (467M, 480B)	120 masses (72M, 48B), 120 images (72M, 48B)		ImageNet-pretrained Inception v2	AUC(DCNN)=0.913, AUC(three radiologists)=0.728 to 0.845
Mass segmentation						
Lei et al.[83]	2018	Automated whole breast US, 16 cases: 3134 images; Leave-one-case-out CV			ConvEDNet with deep boundary supervision	Jaccard index: 72.2 to 86.8%
Hu et al.[84]	2019	570 images (400 training, 170 validation) from 89 patients.			ImageNet pre-trained VGG16, U-Net, DFCN, DFCN +active contour model	DSC(DFCN + active contour)=88.97 %
Yap et al.[85]	2019	Total: 469 masses (113 M, 356 B); 70% training, 10% validation; 5-fold CV		20%	ImageNet-pretrained FCN-AlexNet, FCN-32, FCN-16, and FCN-8	B mass: DSC(FCN-16)=0.7626; M mass: DSC(FCN-8)=0.5484
Mass detection						
Yap et al.[86]	2018	Study 1: 306 images (60M, 246B), 10-fold CV Study 2: 163 images (53M, 110B), 10-fold CV			GoogLeNet, U-Net, ImageNet-pretrained FCN-AlexNet,	FROC: Sens 77 to 98% at 0.28 to 0.10 FPI, FCN-AlexNet: best performance
Shin et al.[87]	2019	800 strongly & 4224 weakly annotated images	600 strongly annotated images		ImageNet-pretrained VGG16, ResNet	FROC: 84.5% at 1 FPI

FCN, fully convolutional network.

U-Net.[101] Due to the relatively small available breast ultrasound image sets, transfer learning was used to train the DCNNs and the DCNNs were most commonly pre-trained with the ImageNet data. The DCNN models were often used directly as classifiers but were also used as feature extractors, where the extracted deep features were merged by machine learning classifiers such as SVM, logistic regression or linear discriminant classifiers. Most of the studies used only training and validation sets without an independent test set. The reported performances were therefore preliminary and further development is needed.

Table 5. Studies using deep learning approach for breast MRI

Journal article	Year	Training set	Validation set	Independent test set	Convolutional neural network (CNN) structure	Performance (validation or independent test)
Classification						
Rasti et al.[88]	2017	112 pts (53M, 59B); 5-fold CV			CNN (three convolutional layers); mixture ensemble of CNNs (ME-CNN) with three CNNs and a convolutional gating network	AUC(CNN)=0.95 AUC(ME-CNN)=0.99
Antropova et al.[44]	2017	690 pts (478M, 212B); 5-fold CV			ImageNet-pretrained VGG19 as feature extractor, SVM classifier	AUC(Maxpool features)=0.87; AUC(fused with radiomic features)=0.89
Antropova et al.[89]	2018	690 pts (478M, 212B); maximum intensity projection (MIP), 5-fold CV			ImageNet-pretrained VGG19 as feature extractor, SVM classifier	AUC(MIP features)=0.88
Segmentation/Classification						
Zhang et al.[90]	2019	224 pts (combination of CV and using all data for training different U-Nets)		48 pts	Multiple U-Nets for breast segmentation, landmark detection and mass segmentation	DSC(mass segment.)=71.8 AUC(Luminal A *vs* others)=0.69 for variance of time-to-peak kinetic feature
Truhn et al.[91]	2019	447 pts (787M from 341 pts, 507B from 237 pts); 10-fold CV in outer loop and 5-fold CV in innerloop			CNN: ImageNet-pretrained ResNet18 Radiomics: PCA and L1 regularization 562 radiomics	AUC(CNN)=0.88 AUC(radiomics-PCA)=0.78 AUC(radiomics-L1 regularization)=0.81 AUC(radiologist)=0.98
Segmentation of fibroglandular tissue/breast density assessment						
Dalmış et al.[92]	2017	39 pts	5 pts	22 pts	3-class U-Net (non-breast, fatty tissue, fibroglandular tissue)	DSC(breast)=0.93; DSC(FGT)=0.85; Correlation(manual *vs* U-Net segmented FGT)=0.97

PCA, principal component analysis.

Deep learning in breast MRI

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) measures the properties of tissue microvasculature by imaging the small excess in the Boltzmann distribution of the spins within the magnetic field.[102,103] DCE-MRI provides functional and structural characteristics of the disease[104] and is routinely used to assess the tumor extent and detect multifocal and multicentric breast cancer. The apparent diffusion coefficient from diffusion-weighted MRI (DW-MRI) can be correlated with the macromolecular and microstructural changes at the cellular level, providing a useful biomarker during cancer treatment.[105] Breast MRI is used for screening females at high risk of breast cancer, treatment response monitoring of neoadjuvant chemotherapy, detection of residual disease and as supplement to other imaging modalities.[106–108] Machine learning methods have been applied to breast MRI for segmentation of fibroglandular tissue for breast density assessment, detection and diagnosis of breast cancer or cancer subtypes, identifying radiomics features as biomarkers and identifying the mapping between radiomics and genomics using radiogenomics analysis.[109–111]

DCNNs have the potential to replace or improve over the conventional machine learning methods in analysis of MRI. Unlike mammography, only a few studies have been conducted to develop DL methods for breast MRI to date, and transfer learning is generally used in these studies. The limited breast MR data available is the major factor limiting its development. The few studies that applied DCNNs to breast MRI are shown in Table 5, which include the use of U-Net for breast tumor segmentation,[90] VGG for feature extraction,[44,89] classification of malignant and benign breast lesions,[88,91] and U-Nets for segmentation of the breast and the fibroglandular tissue for breast density assessment.[92]

MRI has been shown to have a wide range of clinical applications as mentioned above. Some of these tasks involve multimodality, multiparametric imaging and diagnostic tests, where data fusion and quantitative biomarkers may provide important information to support precision medicine. This information is currently underutilized because manual processing is too complex or too time-consuming and thus difficult to conduct large clinical validation studies. Computer-assisted image analysis with machine learning techniques will be most helpful for these tasks. However, the development of DL tools in breast MRI is hindered not only by the difficulty to collect big data for training, but also by the large variations in image characteristics due to differences in acquisition protocols and scanner types among clinical sites.[112] Collecting big data from multi-institutional studies for quantitative DCE-MRI analysis or DL training requires standardized calibration of the scanners and/or robust image homogenization methods. The Quantitative Imaging Biomarkers Alliance has proposed performance standards and tools for MRI.[113] Before widespread implementation of the standardization for MRI in clinical practice, current DL application of MRI data will rely on post-processing techniques to reduce the variations. Development of AI in breast MRI is at an early stage and much more collaborative effort should be devoted to compile big data so that investigators can explore the potentials of the DL approach and the fusion of multidomain deep features with radiomics features and/or other patient data for the various stages of the diagnosis and management of breast cancer.

Promises of deep learning in medical imaging applications

As the development of DL and AI methods for various CAD applications is still ongoing, no large-scale clinical studies have been conducted to evaluate the impact of the new generation of AI-based CAD on clinicians. One application of strong interest in breast imaging is to use AI to reduce radiologists' workload in screening mammography, which is the highest volume exam in breast imaging but with a low cancer prevalence of less than 1%. A few retrospective studies have investigated the feasibility of using AI-based CAD to triage screening mammograms as having low risk or high risk of breast cancer so that radiologists can prioritize their reading and improve workflow.

Rodriguez-Ruiz et al[114] evaluated the standalone performance of an AI-based CAD system for breast cancer detection in 9 data sets used in observer studies from 7 countries which totaled to be 2652 mammography examinations with 653 cancers. Their system achieved an area under the receiver operating characteristic curve (AUC) of 0.840 which was statistically non-inferior to the average AUC of 0.814 by 101 radiologists from the observer studies, and was higher than 61.4% of the radiologists. In another study by the same group using the same data set,[115] the AI-based CAD system was used to assess the risk of malignancy of an exam by a score (1–10 scale). By selecting a risk score >2 and >5 as decision threshold for high risk cases, they could exclude 17 and 47% of the cases from radiologists' reading but missed 1 and 7% of the cancers, respectively.

Kyono et al[116] developed a machine learning method to identify normal cases in screening mammography. A DCNN in conjunction with multitask learning was trained to extract imaging features to predict diagnosis, another deep network was trained to merge the multiview predictions with the patient's non-imaging data into an assessment of whether the case is normal. With 2000 cases and 10-fold cross-validation, their DCNN model achieved a negative predictive value of 0.99 to identify 34 and 91% of the normal mammograms for test sets with a cancer prevalence of 15 and 1%, respectively. They concluded that machine learning could be used for patient triage to reduce the normal mammograms the radiologists need to read without degrading diagnostic accuracy. These results were superior to those reported by Rodriguez-Ruiz et al but the generalizability has yet to be validated with independent testing.

Conant et al[117] conducted a retrospective, fully crossed, multireader, multicase observer performance study on using an AI-based CAD system as a concurrent reader on radiologists' accuracy and reading time for cancer detection in DBT. 24 radiologists including 13 breast subspecialists and 11 general radiologists participated to read 260 DBT cases (65 cancer, 65 benign, 130 normal) with and without AI-CAD in different reading sessions. They found that the mean AUC, the sensitivity, and the specificity increased, while the reading time per case and the recall rate decreased. All improvements by concurrent use of AI-CAD were statistically significant ($p < 0.01$). They also showed that the improvements persisted in the analysis of the subgroups of breast and general radiologists. In another study, Benedikt et al[118] found that using concurrent CAD which showed the AI-detected lesion blended onto the synthetic mammograms of DBT could reduce radiologists' reading time significantly ($p < 0.01$) without significantly affecting the other performance measures.

Yala et al[119] trained a DL model with mammograms of over 56,831 females to triage screening mammograms to predict whether or not that breast would develop breast cancer within 1 year, and selected a threshold to triage mammograms as cancer-free and not needing radiologists' reading. On an independent test set of 7176 females, they showed that although the DL model obtained an AUC of only 0.82, it could triage 19% of the cases as cancer-free with only one false negative. The radiologists had a specificity of 93.6% and a sensitivity of 90.6% in the original test set, and would have obtained an improved specificity of 94.3% and a non-inferior sensitivity of 90.1% in a retrospective simulation of reading the remaining mammograms.

These studies show that AI-based CAD has the potential to reach sufficiently high sensitivity and specificity such that it may be used as a concurrent reader to reduce reading time in DBT or as a pre-screener to exclude some low risk mammograms from radiologists' reading in screening mammography. In general, for AI tools to play these roles beyond providing second opinion or decision support in patient care, they should be subjected to rigorous validation in clinical environment and demonstrate robustness before integration into the routine workflow. It is also important to ensure the stability of their performance over time. Although clinicians and developers are enthusiastic about the potential benefits amid the hype of AI, there are many challenges to achieve these goals, as discussed next.

CHALLENGES FROM THE LABORATORY TO THE CLINIC

Big data for CAD development

The major challenge of developing a robust DCNN for a specific task is to collect a large well-curated data set for training and validation of the model. In addition, a representative independent test set sequestered from the training process should be used to evaluate the generalizability of the trained model in unseen cases.[120] Each class in the data sets has to be representative of the population to which the DCNN is intended to be applied. In particular, the abnormal class in the training set has to be sufficiently large and cover the range of subtleties for the target lesions or diseases that may be encountered in clinical practice to enable adequate learning of the variations in lesion characteristics and thus ensure robustness during real-world deployment, which make data collection even more challenging for tasks such as screening in which the abnormal class is only a small fraction of the population. Collecting data in medical imaging with clinicians' annotation or biopsy truth is costly and such resources may not be available to DCNN developers. Data mining and natural language processing of the electronic health record may be useful for extracting clinical data and diagnosis from the physicians' and pathology reports[121] to correlate with images collected from the picture archiving and communication system. However, the accuracy of the retrieved labels depends on the methods used,[122] and the automatically mined disease labels can contain substantial noise[123] and most do not contain image-level or lesion-level annotations.[124] In the Digital Mammography DREAM Challenge (2016–2017),[124] the participants could access a training set of over 640,000 mammograms from 86,000 females but the cases only had breast-level labeling without lesion annotation. The winning teams all used DL approach but the highest performance only reached an AUC of 0.8744 and a sensitivity of 80% at a specificity of 80.8%.

It may be noted that many of the studies to date as cited in Tables 1–3 had very small training set. For mammography, the publicly accessible Digital Database for Screening Mammography[73] that contains only digitized screen-film mammograms, was used as the only or the main data set in many studies. The other two accessible digital mammography sets, Breast Cancer Digital Repository[74] and INbreast data set,[75] are relatively small. Most of the studies only included training and validation sets or by cross-validation without an independent test set. The reported results are likely optimistically biased because the validation set is usually used to guide the selection of hyperparameters during DCNN training. Without the evaluation using a large, representative independent test set, the generalizability of the reported trained DCNN models is uncertain. Furthermore, it has been shown that DCNN training can be biased to the specific characteristics of the training images acquired with certain imaging protocols or vendors' machines and thus independent testing with external data is necessary in addition to that with internal data to identify these potential biases.[123,125]

To alleviate the problems of limited data available for DCNN training, the commonly used approach at present is to use transfer learning with fine-tuning and data augmentation. Although these techniques can greatly improve DCNN training, they cannot compensate for the lack of adequate representations of disease patterns from the patient population in a sparse training set. Transfer learning takes advantage of the property of DCNN that learns from the input images multiple levels of feature representations from generic to specific and embedded the information in its layers of convolutional kernels and weights. Since many image features are composed of common basic elements, a DCNN initialized with weights well-trained in a different source domain can outperform a DCNN trained from randomly initialized weights,[126] especially when the data set from the target domain is small. Samala et al[56] showed that the performance of the pre-trained DCNN increased with fine-tuning in the target domain and it steadily increased with increasing training sample size. Transfer learning can therefore complement but not replace the need for a large data set to achieve high performance in the target task. Data augmentation generates a number of slightly different versions of a given training image using techniques such as scaling, flipping, rotation, translation, cropping, intensity or shape transformation and combinations of these techniques. Data augmentation can easily increase the apparent number of training samples by thousands of times. However, the augmented images are highly correlated with the original image so that they carry little new information or features for the DCNN to learn. Data augmentation can reduce the risk of overfitting to the training data[29,127,128] by introducing some variations to the images but cannot fill in the missing information if the original small training set does not contain samples that cover the wide range of disease characteristics in the real-world population. Other methods are also being considered for data augmentation, such as generative adversarial networks that can create new images from the learned features after training with an available set of images[129] and digitally inserting artificial lesions into normal images.[130] Whether these methods can mimic the pathological characteristics of real lesions other than structural similarity, especially the texture features inside and surrounding the lesion, and produce useful samples for training DL models to classify real patient cases remain to be studied.

Clinical implementation—acceptance testing and quality assurance

If properly trained with a large data set, AI-based CAD is expected to be more robust and more accurate than conventional CAD tools. However, studies showed that the large learning capacity of DL allows it to even learn non-medical features such as imaging protocols or the presence of accessories related to a patient's comorbidity to estimate the risk of certain disease.[125] As a result, an AI-based CAD tool well trained and independently tested using data collected from some clinical sites may not translate to other sites. Similar to installation of new clinical equipment, acceptance testing should be performed to verify that its performance can pass a certain reference level using a data set representative of the local patient population. In addition, given the current high expectation that DL technologies are "intelligent," it will be even more important for clinicians to understand the capabilities and limitations of a CAD tool and what it is designed for before clinical use. After the installation, the clinic and the users should allow for a test period in which the users refrain from being influenced by the CAD output. Rather, the users should review the correct and incorrect recommendations by CAD and assess its performance on a large number of consecutive clinical cases. By learning the characteristics of the cases and understanding the strengths and weaknesses of the CAD tool, the users may be able to establish proper expectation and confidence level and reduce the risk of improper use and adverse outcomes. The test period therefore serves both as a real-world evaluation of the CAD tool on the local population and user training.

The performance of a CAD tool may be affected by the properties of the input image, which depend on many factors such as the imaging protocol or equipment and the image processing or reconstruction software that may change or upgrade from time to time. As AI-based CAD tools are expected 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, CAD tool can directly impact clinical decision and thus patient management. It is important to establish quality assurance (QA) program and appropriate metrics to monitor the standalone CAD performance as well as the effectiveness and efficiency of CAD use in the clinic over time. The need for QA and user training on CAD devices has been discussed in an opinion paper by the American Association of Physicists in Medicine CAD Subcommittee.[131] Professional organizations should take the lead to establish performance standards, QA and monitoring procedures, and compliance guidance, to ensure the safety and effectiveness for implementation and operation of CAD/AI tools in clinical practice.

INTERPRETABILITY

A DCNN extracts layers of feature representations from the input data, merges them with a highly complex model and predicts the probability that the input belongs to a certain class. It is difficult to decipher the process of how the DCNN derives its prediction. Researchers have developed visualization tools to display the deep feature layers in the DCNN[132,133] and to visualize the relative importance of regions on the input image that contribute to the DL output by a heat map, such as the class activation map.[134] These visualization tools are the first steps to gain some understanding of the deep features in relation to the input data but still far from explaining why and how specific features are connected and weighted to make a clinical decision. For clinicians to be convinced of the recommendation by the AI model, especially for clinical tasks more complicated than lesion detection, the DL model has to provide reasonable interpretations of how its extracted features and output are correlated with the patient's medical conditions or other clinical data. Ideally, an AI tool should be able to convey the interpretation to clinicians in direct medical languages and can even provide deeper level of explanation if the recommendation is questioned. The level of interpretation and the method of presenting the interpretation will depend on the specific purpose of each type of AI tools. Much more research and development efforts are needed to determine clinicians' preferences on each type of applications and to advance the DL models to be truly intelligent decision support tools.

SUMMARY

DL technology has the potential of bringing the performance of CAD tools to a level far beyond those developed with conventional machine learning methods. However, the development of DL-based CAD tools including those for breast imaging are still at an early stage due to the lack of large data sets for training the DCNNs to date. Collaborative efforts from multiple institutions to compile big patient data for various diseases is the most urgent step to allow effective utilization of DL technology for the development of practical CAD or AI tools. With sufficiently large well-curated data for a given task, DL technology can build a robust predictive model based on the cumulative experiences from a large number of previous cases collected from the patient population, much greater than those human clinicians can ever learn from or memorize. It is likely that AI tools, if properly developed and integrated into the clinical workflow, can deliver performance comparable to or even exceeding clinicians' in some routine tasks. However, medical decision making is a highly complex process, which often cannot rely solely on statistical prediction but may vary based on individual patient's conditions and medical history, as well as some unpredictable physiological processes or reactions of the human body. A well-developed

CAD or AI tool can merge patient data from multiple resources efficiently and provide a reliable and hopefully interpretable assessment to clinicians, who should then play the key role as the final decision maker on the best course of management for a specific patient based on the CAD information, together with his/her experience and judgment. 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 in the clinic and thus patient care in this new era of machine learning.

ACKNOWLEDGMENT

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

REFERENCES

1. Burt JR, Torosdagli N, Khosravan N, RaviPrakash H, Mortazi A, Tissavirasingham F, et al. Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. *Br J Radiol* 2018; **2**: 20170545. doi: https://doi.org/10.1259/bjr.20170545
2. Winsberg F, Elkin M, Macy J, Bordaz V, Weymouth W. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology* 1967; **89**: 211–5. doi: https://doi.org/10.1148/89.2.211
3. Kimme C, O'Laughlin BJ, Sklansky J. *Automatic detection of suspicious abnormalities in breast radiographs*. eds., New York: Academic Press; 1977.
4. Spiesberger W. Mammogram inspection by computer. *IEEE Trans Biomed Eng* 1979; **BME-26**: 213–9. doi: https://doi.org/10.1109/TBME.1979.326560
5. Semmlow JL, Shadagopappan A, Ackerman LV, Hand W, Alcorn FS. A fully automated system for screening xeromammograms. *Computers and Biomedical Research* 1980; **13**: 350–62. doi: https://doi.org/10.1016/0010-4809(80)90027-0
6. Doi K. Chapter 1. Historical overview. In: Li Q, Nishikawa R. M, Raton B, eds. *Computer-Aided Detection and Diagnosis in Medical Imaging*: FL: Taylor & Francis Group, LLC, CRC Press; 2015. pp. 1–17.
7. Chan HP, Doi K, Vyborny CJ, Schmidt RA, Metz CE, Lam KL, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis. *Invest Radiol* 1990; **25**: 1102–10. doi: https://doi.org/10.1097/00004424-199010000-00006
8. Chan HP, Doi K, Galhotra S, Vyborny CJ, MacMahon H, Jokich PM. Image feature analysis and computer-aided diagnosis in digital radiography. I. automated detection of microcalcifications in mammography. *Med Phys* 1987; **14**: 538–48. doi: https://doi.org/10.1118/1.596065
9. Taylor P, Potts HWW. 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* 2008; **44**: 798–807. doi: https://doi.org/10.1016/j.ejca.2008.02.016
10. Gilbert FJ, Astley SM, Gillan MGC, Agbaje OF, Wallis MG, James J, et al. Single reading with computer-aided detection for screening mammography. *New England Journal of Medicine* 2008; **359**: 1675–84. doi: https://doi.org/10.1056/NEJMoa0803545
11. Gromet M. Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. *AJR Am J Roentgenol* 2008; **190**: 854–9. doi: https://doi.org/10.2214/AJR.07.2812
12. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med* 2007; **356**: 1399–409. doi: https://doi.org/10.1056/NEJMoa066099
13. Fenton JJ, Abraham L, Taplin SH, Geller BM, Carney PA, D'Orsi C, et al. Effectiveness of computer-aided detection in community mammography practice. *J Natl Cancer Inst* 2011; **103**: 1152–61. doi: https://doi.org/10.1093/jnci/djr206
14. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015; **175**: 1828–37. doi: https://doi.org/10.1001/jamainternmed.2015.5231
15. Boyer B, Balleyguier C, Granat O, Pharaboz C. Cad in questions/answers review of the literature. *Eur J Radiol* 2009; **69**: 24–33. doi: https://doi.org/10.1016/j.ejrad.2008.07.042
16. Cole EB, Zhang Z, Marques HS, Edward Hendrick R, Yaffe MJ, Pisano ED. Impact of computer-aided detection systems on radiologist accuracy with digital mammography. *AJR Am J Roentgenol* 2014; **203**: 909–16. doi: https://doi.org/10.2214/AJR.12.10187
17. Katzen J, Dodelzon K. A review of computer aided detection in mammography. *Clin Imaging* 2018; **52**: 305–9. doi: https://doi.org/10.1016/j.clinimag.2018.08.014
18. LeCun Y, Bengio Y, Hinton G, learning D. Deep learning. *Nature* 2015; **521**: 436–44. doi: https://doi.org/10.1038/nature14539
19. Fukushima K, Miyake S. Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognit* 1982; **15**: 455–69. doi: https://doi.org/10.1016/0031-3203(82)90024-3
20. Lo CSB, Lin JS, Freedman MT, Mun SK. Computer-Assisted diagnosis of lung nodule detection using artificial convolution neural network. *Proc SPIE* 1898; **1993**: 859–69.
21. Lo SCB, Chan HP, Lin JS, Li H, Freedman M, Mun SK. Artificial convolution neural network for medical image pattern recognition. *Neural Networks* 1995; **8**: 1201–14.
22. Chan H-P, Lo SCB, Helvie MA, Goodsitt MM, Cheng SNC, Adler DD. Recognition of mammographic microcalcifications with artificial neural network. *Radiology* 1993; **189**(P): 318.
23. Chan HP, Lo SCB, Sahiner B, Lam KL, Helvie MA. Computer-Aided detection of mammographic microcalcifications: pattern recognition with an artificial neural network. *Med Phys* 1995; **22**: 1555–67. doi: https://doi.org/10.1118/1.597428
24. Chan H-P, Sahiner B, Lo SCB, et al. Computer-Aided diagnosis in mammography: detection of masses by artificial neural network. *Medical Physics* 1994; **21**: 875–6.
25. Sahiner B, Chan H-P, Petrick N, et al. Image classification using artificial neural networks. *Proc SPIE* 1995; **2434**: 838–45.
26. Wei D, Sahiner B, Chan H-P, Petrick N. Detection of masses on mammograms using a convolution neural network. *Proceedings of International Conference on Acoustics,*

Speech and Signal Processing 1995; **5**: 3483–6.
27. Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA, Adler DD, et al. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE Trans Med Imaging* 1996; **15**: 598–610. doi: https://doi.org/10.1109/42.538937
28. Zhang W, Doi K, Giger ML, Wu Y, Nishikawa RM, Schmidt RA. Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. *Med Phys* 1994; **21**: 517–24. doi: https://doi.org/10.1118/1.597177
29. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 2012;: 1097–105.
30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Conference on Computer Vision and Pattern Recognition* 2015: 770–8. arXiv:1512.03385.
31. Samala RK, Chan H-P, Lu Y, Hadjiiski LM, Wei J, Helvie MA. Digital breast tomosynthesis: computer-aided detection of clustered microcalcifications on planar projection images. *Phys Med Biol* 2014; **59**: 7457–77. doi: https://doi.org/10.1088/0031-9155/59/23/7457
32. Samala RK, Chan H-P, Lu Y, Hadjiiski LM, Wei J, Helvie MA. Computer-Aided detection system for clustered microcalcifications in digital breast tomosynthesis using joint information from volumetric and planar projection images. *Phys Med Biol* 2015; **60**: 8457–79. doi: https://doi.org/10.1088/0031-9155/60/21/8457
33. Wang J, Yang Y. A context-sensitive deep learning approach for microcalcification detection in mammograms. *Pattern Recognit* 2018; **78**: 12–22. doi: https://doi.org/10.1016/j.patcog.2018.01.009
34. Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Sci Rep* 2016; **6**: 27327. doi: https://doi.org/10.1038/srep27327
35. Cai H, Huang Q, Rong W, Song Y, Li J, Wang J, et al. Breast Microcalcification diagnosis using deep Convolutional neural network from digital mammograms. *Comput Math Methods Med* 2019; **2019**: 2717454: 10: doi: https://doi.org/10.1155/2019/2717454
36. Shi B, Grimm LJ, Mazurowski MA, Baker JA, Marks JR, King LM, et al. Prediction of occult invasive disease in ductal carcinoma in situ using deep learning features. *J Am Coll Radiol* 2018; **15**(3 Pt B): 527–34. doi: https://doi.org/10.1016/j.jacr.2017.11.036
37. Samala RK, Chan H-P, Hadjiiski L, Helvie MA, Wei J, Cha K. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. *Med Phys* 2016; **43**: 6654–66. doi: https://doi.org/10.1118/1.4967345
38. Kim DH, Kim ST, Chang JM, Ro YM. Latent feature representation with depth directional long-term recurrent learning for breast masses in digital breast tomosynthesis. *Phys Med Biol* 2017; **62**: 1009–31. doi: https://doi.org/10.1088/1361-6560/aa504e
39. Jung H, Kim B, Lee I, Yoo M, Lee J, Ham S, et al. Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. *PLoS One* 2018; **13**: e0203355. doi: https://doi.org/10.1371/journal.pone.0203355
40. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA, Lopez MAG. Representation learning for mammography mass lesion classification with convolutional neural networks. *Comput Methods Programs Biomed* 2016; **127**: 248–57. doi: https://doi.org/10.1016/j.cmpb.2015.12.014
41. Jiao Z, Gao X, Wang Y, Li J. A deep feature based framework for breast masses classification. *Neurocomputing* 2016; **197**: 221–31. doi: https://doi.org/10.1016/j.neucom.2016.02.060
42. Dhungel N, Carneiro G, Bradley AP. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Med Image Anal* 2017; **37**: 114–28. doi: https://doi.org/10.1016/j.media.2017.01.009
43. Sun W, Tseng T-LB, Zhang J, Qian W. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Comput Med Imaging Graph* 2017; **57**: 4–9. doi: https://doi.org/10.1016/j.compmedimag.2016.07.004
44. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med Phys* 2017; **44**: 5162–71. doi: https://doi.org/10.1002/mp.12453
45. Samala RK, Chan H-P, Hadjiiski LM, Helvie MA, Cha KH, Richter CD. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. *Phys Med Biol* 2017; **62**: 8894–908. doi: https://doi.org/10.1088/1361-6560/aa93d4
46. Kooi T, van Ginneken B, Karssemeijer N, den Heeten A. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. *Med Phys* 2017; **44**: 1017–27. doi: https://doi.org/10.1002/mp.12110
47. Jiao Z, Gao X, Wang Y, Li J. A parasitic metric learning net for breast mass classification based on mammography. *Pattern Recognit* 2018; **75**: 292–301. doi: https://doi.org/10.1016/j.patcog.2017.07.008
48. Samala RK, Chan H-P, Hadjiiski LM, Helvie MA, Richter C, Cha K. Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Phys Med Biol* 2018; **63**: 095005. doi: https://doi.org/10.1088/1361-6560/aabb5b
49. Chougrad H, Zouaki H, Alheyane O. Deep Convolutional neural networks for breast cancer screening. *Comput Methods Programs Biomed* 2018; **157**: 19–30. doi: https://doi.org/10.1016/j.cmpb.2018.01.011
50. Al-Masni MA, Al-Antari MA, Park J-M, Gi G, Kim T-Y, Rivera P, et al. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Comput Methods Programs Biomed* 2018; **157**: 85–94. doi: https://doi.org/10.1016/j.cmpb.2018.01.017
51. Wang H , Feng J , Zhang Z , Su H , Cui L , He H , et al . Breast mass classification via deeply integrating the contextual information from multi-view data . *Pattern Recognit*2018 ; **80**: 42 – 52 . doi: https://doi.org/10.1016/j.patcog.2018.02.026
52. Al-Antari MA, Al-Masni MA, Choi M-T, Han S-M, Kim T-S. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. *Int J Med Inform* 2018; **117**: 44–54. doi: https://doi.org/10.1016/j.ijmedinf.2018.06.003
53. Gao F, Wu T, Li J, Zheng B, Ruan L, Shang D, et al. SD-CNN: a shallow-deep CNN for improved breast cancer diagnosis. *Comput Med Imaging Graph* 2018; **70**: 53–62. doi: https://doi.org/10.1016/j.compmedimag.2018.09.004
54. Kim ST, Lee J-H, Lee H, Ro YM. Visually interpretable deep network for diagnosis of breast masses on mammograms. *Phys. Med. Biol.* 2018; **63**: 235025. doi: https://doi.org/10.1088/1361-6560/aaef0a
55. Perek S, Kiryati N, Zimmerman-Moreno G, Sklair-Levy M, Konen E, Mayer A. Classification of contrast-enhanced spectral mammography (CESM) images. *Int J Comput Assist Radiol Surg* 2019; **14**: 249–57.

doi: https://doi.org/10.1007/s11548-018-1876-6
56. Samala RK, Chan HP, Hadjiiski L, Helvie MA, Richter CD, Cha KH. 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* 2019; **38**: 686–96. doi: https://doi.org/10.1109/TMI.2018.2870343
57. Mendel K, Li H, Sheth D, Giger M. Transfer learning from Convolutional neural networks for computer-aided diagnosis: a comparison of digital breast Tomosynthesis and Full-Field digital mammography. *Acad Radiol* 2019; **26**: 735–43. doi: https://doi.org/10.1016/j.acra.2018.06.019
58. 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. doi: https://doi.org/10.1097/RLI.0000000000000358
59. Carneiro G, Nascimento J, Bradley AP. Automated analysis of Unregistered Multi-View mammograms with deep learning. *IEEE Trans Med Imaging* 2017; **36**: 2355–65. doi: https://doi.org/10.1109/TMI.2017.2751523
60. Kim E-K, Kim H-E, Han K, Kang BJ, Sohn Y-M, Woo OH, et al. Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. *Sci Rep* 2018; **8**: 2762. doi: https://doi.org/10.1038/s41598-018-21215-1
61. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. *Sci Rep* 2018; **8**: 4165. doi: https://doi.org/10.1038/s41598-018-22437-z
62. Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep learning to distinguish recalled but benign mammography images in breast cancer screening. *Clin Cancer Res* 2018; **24**: 5902–9. doi: https://doi.org/10.1158/1078-0432.CCR-18-1115
63. Akselrod-Ballin A, Chorev M, Shoshan Y, Spiro A, Hazan A, Melamed R, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* 2019; **292**: 331–42. doi: https://doi.org/10.1148/radiol.2019182622
64. Kallenberg M, Petersen K, Nielsen M, Ng AY, PengfeiDiao, Igel C, et al. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Trans Med Imaging* 2016; **35**: 1322–31. doi: https://doi.org/10.1109/TMI.2016.2532122
65. Li S, Wei J, Chan H-P, Helvie MA, Roubidoux MA, Lu Y, et al. Computer-Aided assessment of breast density: comparison of supervised deep learning and feature-based statistical learning. *Phys Med Biol* 2018; **63**: 025005. doi: https://doi.org/10.1088/1361-6560/aa9f87
66. Mohamed AA, Berg WA, Peng H, Luo Y, Jankowitz RC, Wu S. A deep learning method for classifying mammographic breast density categories. *Med Phys* 2018; **45**: 314–21. doi: https://doi.org/10.1002/mp.12683
67. Mohamed AA, Luo Y, Peng H, Jankowitz RC, Wu S. Understanding clinical mammographic breast density assessment: a deep learning perspective. *J Digit Imaging* 2018; **31**: 387–92. doi: https://doi.org/10.1007/s10278-017-0022-2
68. Lee J , Nishikawa RM . Automated mammographic breast density estimation using a fully convolutional network . *Med Phys*2018 ; **45**: 1178 – 90 . doi: https://doi.org/10.1002/mp.12763
69. Wanders JOP, van Gils CH, Karssemeijer N, Holland K, Kallenberg M, Peeters PHM, et al. The combined effect of mammographic texture and density on breast cancer risk: a cohort study. *Breast Cancer Res* 2018; **20**: 36. doi: https://doi.org/10.1186/s13058-018-0961-7
70. Gastounioti A, Oustimov A, Hsieh M-K, Pantalone L, Conant EF, Kontos D. Using Convolutional neural networks for enhanced capture of breast parenchymal complexity patterns associated with breast cancer risk. *Acad Radiol* 2018; **25**: 977–84. doi: https://doi.org/10.1016/j.acra.2017.12.025
71. Ciritsis A, Rossi C, Vittoria De Martini I, Eberhard M, Marcon M, Becker AS, et al. Determination of mammographic breast density using a deep convolutional neural network. *Br J Radiol* 2018; **6**: 20180691. doi: https://doi.org/10.1259/bjr.20180691
72. Lehman CD , Yala A , Schuster T , Dontchos B , Bahl M , Swanson K , et al . Mammographic breast density assessment using deep learning: clinical implementation . *Radiology*2019 ; **290**: 52 – 8 . doi: https://doi.org/10.1148/radiol.2018180694
73. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer WP. The digital database for screening mammography. *Paper presented at: Proc. 5th international workshop on digital mammography* 2000.
74. Breast cancer digital Repository (BCDR). https://bcdr.eu/information/about.
75. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. *Acad Radiol* 2012; **19**: 236–48. doi: https://doi.org/10.1016/j.acra.2011.09.014
76. Han S, Kang H-K, Jeong J-Y, Park M-H, Kim W, Bang W-C, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Phys Med Biol* 2017; **62**: 7714–28. doi: https://doi.org/10.1088/1361-6560/aa82ec
77. Xiao T, Liu L, Li K, Qin W, Yu S, Li Z. Comparison of transferred deep neural networks in ultrasonic breast masses discrimination. *Biomed Res Int* 2018; **2018**: 1–9. doi: https://doi.org/10.1155/2018/4605191
78. Zhou Y, Xu J, Liu Q, Li C, Liu Z, Wang M, et al. A Radiomics approach with CNN for shear-wave elastography breast tumor classification. *IEEE Trans Biomed Eng* 2018; **65**: 1935–42. doi: https://doi.org/10.1109/TBME.2018.2844188
79. Lee C-Y, Chen G-L, Zhang Z-X, Chou Y-H, Hsu C-C. Is intensity inhomogeneity correction useful for classification of breast cancer in Sonograms using deep neural network? *J Healthc Eng* 2018; **2018**: 1–10. doi: https://doi.org/10.1155/2018/8413403
80. Huang Y, Han L, Dou H, Luo H, Yuan Z, Liu Q, et al. Two-Stage CNNs for computerized BI-RADS categorization in breast ultrasound images. *Biomed Eng Online* 2019; **18**. doi: https://doi.org/10.1186/s12938-019-0626-5
81. Byra M, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, et al. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. *Med Phys* 2019; **46**: 746–55. doi: https://doi.org/10.1002/mp.13361
82. Fujioka T, Kubota K, Mori M, Kikuchi Y, Katsuta L, Kasahara M, et al. Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network. *Jpn J Radiol* 2019; **37**: 466–72. doi: https://doi.org/10.1007/s11604-019-00831-5
83. Lei B, Huang S, Li R, Bian C, Li H, Chou Y-H, et al. Segmentation of breast anatomy for automated whole breast ultrasound images with boundary regularized convolutional encoder–decoder network. *Neurocomputing* 2018; **321**: 178–86. doi: https://doi.org/10.1016/j.neucom.2018.09.043
84. Hu Y, Guo Y, Wang Y, Yu J, Li J, Zhou S, et al. Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model. *Med Phys* 2019;

46: 215–28. doi: https://doi.org/10.1002/mp.13268
85. Yap MH, Goyal M, Osman FM, Martí R, Denton E, Juette A, et al. Breast ultrasound lesions recognition: end-to-end deep learning approaches. *J Med Imaging* 2019; **6**: 011007. doi: https://doi.org/10.1117/1.JMI.6.1.011007
86. Yap MH, Pons G, Marti J, Ganau S, Sentis M, Zwiggelaar R, et al. Automated breast ultrasound lesions detection using Convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics* 2018; **22**: 1218–26. doi: https://doi.org/10.1109/JBHI.2017.2731873
87. Shin SY, Lee S, Yun ID, Kim SM, Lee KM, Weakly J. Joint weakly and Semi-Supervised deep learning for localization and classification of masses in breast ultrasound images. *IEEE Trans Med Imaging* 2019; **38**: 762–74. doi: https://doi.org/10.1109/TMI.2018.2872031
88. Rasti R, Teshnehlab M, Phung SL. Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks. *Pattern Recognit* 2017; **72**: 381–90. doi: https://doi.org/10.1016/j.patcog.2017.08.004
89. Antropova N, Abe H, Giger ML. Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks. *Journal of Medical Imaging* 2018; **5**: 1. doi: https://doi.org/10.1117/1.JMI.5.1.014503
90. Zhang J, Saha A, Zhu Z, Mazurowski MA. Hierarchical convolutional neural networks for segmentation of breast tumors in MRI with application to radiogenomics. *IEEE Trans Med Imaging* 2019; **38**: 435–47. doi: https://doi.org/10.1109/TMI.2018.2865671
91. Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus Convolutional neural networks analysis for classification of Contrast-enhancing lesions at multiparametric breast MRI. *Radiology* 2019; **290**: 290–7. doi: https://doi.org/10.1148/radiol.2018181352
92. Dalmış MU, Litjens G, Holland K, Setio A, Mann R, Karssemeijer N, et al. Using deep learning to segment breast and fibroglandular tissue in MRI volumes. *Med Phys* 2017; **44**: 533–46. doi: https://doi.org/10.1002/mp.12079
93. Simonyan K, Zisserman A. Very deep convolutional networks for large-scal image recognition. 2014; arXiv:1409.1556.
94. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *Proc IEEE Conf on Computer Vision and Pattern Recognition* 2015;: 1–9.
95. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* 2017; arXiv:1602.07261.
96. Huang Q, Zhang F, Li X. Machine learning in ultrasound computer-aided diagnostic systems: a survey. *Biomed Res Int* 2018; **2018**: 10. doi: https://doi.org/10.1155/2018/5137904
97. Xian M, Zhang Y, Cheng HD, Xu F, Zhang B, Ding J. Automatic breast ultrasound image segmentation: a survey. *Pattern Recognit* 2018; **79**: 340–55. doi: https://doi.org/10.1016/j.patcog.2018.02.012
98. Wu G-G, Zhou L-Q, Xu J-W, Wang J-Y, Wei Q, Deng Y-B, et al. Artificial intelligence in breast ultrasound. *World J Radiol* 2019; **11**: 19–25. doi: https://doi.org/10.4329/wjr.v11.i2.19
99. Le EPV, Wang Y, Huang Y, Hickman S, Gilbert FJ. Artificial intelligence in breast imaging. *Clin Radiol* 2019; **74**: 357–66. doi: https://doi.org/10.1016/j.crad.2019.02.006
100. Mendelson EB. Artificial intelligence in breast imaging: potentials and limitations. *AJR Am J Roentgenol* 2019; **212**: 293–9. doi: https://doi.org/10.2214/AJR.18.20532
101. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation.. 2015; arXiv:1505.04597v1.
102. Frangioni JV. New technologies for human cancer imaging. *J Clin Oncol* 2008; **26**: 4012–21. doi: https://doi.org/10.1200/JCO.2007.14.3065
103. Hylton N. Dynamic contrast-enhanced magnetic resonance imaging as an imaging biomarker. *J Clin Oncol* 2006; **24**: 3293–8. doi: https://doi.org/10.1200/JCO.2006.06.8080
104. Kuhl CK, Schild HH. Dynamic image interpretation of MRI of the breast. *J Magn Reson Imaging* 2000; **12**: 965–74. doi: https://doi.org/10.1002/1522-2586(200012)12:6<965::AID-JMRI23>3.0.CO;2-1
105. Thoeny HC, Ross BD. Predicting and monitoring cancer treatment response with diffusion-weighted MRI. *J Magn Reson Imaging* 2010; **32**: 2–16. doi: https://doi.org/10.1002/jmri.22167
106. Saslow D, Boetes C, Burke W, Harms S, Leach MO, Lehman CD, et al. American cancer Society guidelines for breast screening with MRI as an adjunct to mammography. *CA Cancer J Clin* 2007; **57**: 75–89. doi: https://doi.org/10.3322/canjclin.57.2.75
107. Turnbull LW. Dynamic contrast-enhanced MRI in the diagnosis and management of breast cancer.. In: *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance*. **22**; 2009. pp. 28–39.
108. Lee CH, Dershaw DD, Kopans D, Evans P, Monsees B, Monticciolo D, et al. Breast cancer screening with imaging: recommendations from the Society of breast imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. *J Am Coll Radiol* 2010; **7**: 18–27. doi: https://doi.org/10.1016/j.jacr.2009.09.022
109. Dorrius MD, Jansen-van der Weide MC, van Ooijen PMA, Pijnappel RM, Oudkerk M. Computer-Aided detection in breast MRI: a systematic review and meta-analysis. *Eur Radiol* 2011; **21**: 1600–8. doi: https://doi.org/10.1007/s00330-011-2091-9
110. Fusco R, Sansone M, Filice S, Carone G, Amato DM, Sansone C, et al. Pattern recognition approaches for breast cancer DCE-MRI classification: a systematic review. *J Med Biol Eng* 2016; **36**: 449–59. doi: https://doi.org/10.1007/s40846-016-0163-7
111. Pinker K, Shitano F, Sala E, Do RK, Young RJ, Wibmer AG, et al. Background, current role, and potential applications of radiogenomics. *J Magn Reson Imaging* 2018; **47**: 604–20. doi: https://doi.org/10.1002/jmri.25870
112. Jackson A, O'Connor JPB, Parker GJM, Jayson GC. Imaging tumor vascular heterogeneity and angiogenesis using dynamic contrast-enhanced magnetic resonance imaging. *Clin Cancer Res* 2007; **13**: 3449–59. doi: https://doi.org/10.1158/1078-0432.CCR-07-0238
113. Shukla-Dave A, Obuchowski NA, Chenevert TL, Jambawalikar S, Schwartz LH, Malyarenko D, et al. Quantitative imaging biomarkers alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. *J Magn Reson Imaging* 2019; **49**: e101–21. doi: https://doi.org/10.1002/jmri.26518
114. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019; **111**: 916–22. doi: https://doi.org/10.1093/jnci/djy222
115. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol*

2019; **29**: 4825–32. doi: https://doi.org/10.1007/s00330-019-06186-9

116. Kyono T, Gilbert FJ, van der Schaar M. Improving workflow efficiency for mammography using machine learning. *J Am Coll Radiol* 2019;30 May 2019. doi: https://doi.org/10.1016/j.jacr.2019.05.012
117. Conant EF, Toledano AY, Periaswamy S, et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast Tomosynthesis screening. *Radiological Society of North America Scientific Assembly and Annual Meeting* 2018: RC215–214.
118. Benedikt RA, Boatsman JE, Swann CA, Kirkpatrick AD, Toledano AY. Concurrent computer-aided detection improves reading time of digital breast Tomosynthesis and maintains interpretation performance in a Multireader Multicase study. *AJR Am J Roentgenol* 2018; **210**: 685–94. doi: https://doi.org/10.2214/AJR.17.18185
119. Yala A, Schuster T, Miles R, Barzilay R. Lehman C. a deep learning model to triage screening mammograms: a simulation study. *Radiology*: 182908.
120. Petrick N, Sahiner B, Armato SG, Bert A, Correale L, Delsanto S, et al. Evaluation of computer-aided detection and diagnosis systems. *Med Phys* 2013; **40**: 087001. doi: https://doi.org/10.1118/1.4816310
121. Shin H, Le L, Kim L, Seff A, Yao J, Summers RM. Interleaved text/image deep mining on a large-scale radiology database. *Proc IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2015: 1090–9.
122. Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A, et al. Natural language-based machine learning models for the annotation of clinical radiology reports. *Radiology* 2018; **287**: 570–80. doi: https://doi.org/10.1148/radiol.2018171093
123. Oakden-Rayner L. Exploring the ChestXray14 dataset: problems. 2017. Available from: https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/.
124. The digital mammograph DREAM challenge. https://www.synapse.org/#!Synapse:syn4224222/wiki/434547.
125. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018; **15**: e1002683. doi: https://doi.org/10.1371/journal.pmed.1002683
126. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? *Advances in neural information processing systems. Proc Advances in neural information processing systems* 2014: 3320–8.
127. Taylor L, Nitschke G. Improving deep learning using generic data augmentation. 2017; arXiv:1708.06020.
128. Wang J, Perez L. The effectiveness of data augmentation in image classification using deep learning. 2017; arXiv:1712.04621.
129. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative Adversarial nets. 2014; arXiv:1406.2661v1.
130. Badano A, Graff CG, Badal A, Sharma D, Zeng R, Samuelson FW, et al. Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial. *JAMA Netw Open* 2018; **1**: e185474. doi: https://doi.org/10.1001/jamanetworkopen.2018.5474
131. Huo Z, Summers RM, Paquerault S, Lo J, Hoffmeister J, Armato SG, et al. Quality assurance and training procedures for computer-aided detection and diagnosis systems in clinical use. *Med Phys* 2013; **40**: 077001. doi: https://doi.org/10.1118/1.4807642
132. Zeiler MD, Fergus R. Visualizing and understanding Convolutional networks. *Lecture Notes in Computer Science Computer Vision – European Conference on Computer Vision (ECCV) 2014* 2014; **8689**: 818–33.
133. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization. 2015: arXiv:1506.06579v1.
134. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. *Proc IEEE Conference on Computer Vision and Pattern Recognition* 2016: 2921–9.