ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/335578068

A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence–Powered Ultrasound for Improving Clinical Workflow

Article *in* Journal of the American College of Radiology · September 2019

DOI: 10.1016/j.jacr.2019.06.004

CITATIONS
255

READS
7,641

7 authors, including:

Zeynettin Akkus
Mayo Clinic
58 PUBLICATIONS **4,396** CITATIONS
SEE PROFILE

Arunnit Boonrod
Khon Kaen University
28 PUBLICATIONS **609** CITATIONS
SEE PROFILE

Atefeh Zeinoddini
The University of Texas Medical Branch at Galveston
51 PUBLICATIONS **1,616** CITATIONS
SEE PROFILE

Alexander Weston
Mayo Clinic - Rochester
25 PUBLICATIONS **1,250** CITATIONS
SEE PROFILE

All content following this page was uploaded by Zeynettin Akkus on 07 September 2019.

The user has requested enhancement of the downloaded file.

Check for updates

A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence–Powered Ultrasound for Improving Clinical Workflow

Zeynettin Akkus, PhD[a], Jason Cai, MD[a], Arunnit Boonrod, MD[a,b], Atefeh Zeinoddini, MD[a], Alexander D. Weston, BSc[a], Kenneth A. Philbrick, PhD[a], Bradley J. Erickson, MD, PhD[a]

Abstract

Ultrasound is the most commonly used imaging modality in clinical practice because it is a nonionizing, low-cost, and portable point-of-care imaging tool that provides real-time images. Artificial intelligence (AI)–powered ultrasound is becoming more mature and getting closer to routine clinical applications in recent times because of an increased need for efficient and objective acquisition and evaluation of ultrasound images. Because ultrasound images involve operator-, patient-, and scanner-dependent variations, the adaptation of classical machine learning methods to clinical applications becomes challenging. With their self-learning ability, deep-learning (DL) methods are able to harness exponentially growing graphics processing unit computing power to identify abstract and complex imaging features. This has given rise to tremendous opportunities such as providing robust and generalizable AI models for improving image acquisition, real-time assessment of image quality, objective diagnosis and detection of diseases, and optimizing ultrasound clinical workflow. In this report, the authors review current DL approaches and research directions in rapidly advancing ultrasound technology and present their outlook on future directions and trends for DL techniques to further improve diagnosis, reduce health care cost, and optimize ultrasound clinical workflow.

Key Words: Artificial intelligence in ultrasound, deep learning in ultrasound, thyroid nodule, breast lesion, liver lesion

J Am Coll Radiol 2019;16:1318-1328. *Copyright © 2019 American College of Radiology*

INTRODUCTION

Artificial intelligence (AI)–powered ultrasound is becoming more mature and coming closer to routine clinical applications in recent years because of an increased need for efficient and objective acquisition and evaluation of ultrasound images. Because ultrasound is an operator-dependent imaging modality, it is important to develop deep-learning (DL) models that assess image quality and provide feedback to sonographers; providing guidance during data acquisition and measurement makes ultrasound more intelligent and less operator dependent.

In this review, we intend to give an overview of the advances in AI-powered ultrasound that create opportunities to objectively evaluate ultrasound data, improve clinical workflow, and reduce health care costs. We also provide a brief background on DL and ultrasound imaging to give a comprehensive insight into the field. Afterward, we present currently available DL applications in ultrasound, discuss challenges, and present our outlook on future directions in AI-powered ultrasound.

DL

DL is a subset of machine learning (ML) and AI that extracts a complex hierarchy of features from images by its self-learning ability, as opposed to the handcrafted feature extraction in classical ML algorithms [1]. DL involves neural networks with many layers that extract a hierarchy of features from raw input images. The rapid increase in the processing power of graphics processing units has enabled the development of state-of-the-art DL algorithms that can be trained with

[a]Radiology Informatics Lab, Department of Radiology, Mayo Clinic, Rochester, Minnesota.

[b]Radiology Department, Khon Kaen University, Khon Kaen, Thailand.

Corresponding author and reprints: Zeynettin Akkus, PhD, Mayo Clinic, Department of Radiology, Radiology Informatics Lab, 200 First Street SW, Rochester, MN 55904; e-mail: akkus.zeynettin@mayo.edu.

This work was supported by Mayo Clinic Ultrasound Center Pilot Grant and Radiology Informatics Lab. The authors state that they have no conflict of interest related to the material discussed in this article.

© 2019 American College of Radiology

1546-1440/19/$36.00 ■ https://doi.org/10.1016/j.jacr.2019.06.004

millions of images and are robust to variations in images. DL has become popular because of recent successes especially in image segmentation and classification applications. Compared with DL, classical ML approaches that are hand designed in decomposable pipelines are more interpretable because each component has an explanation, but they are usually not very accurate or robust. By using DL models we sacrifice interpretability for robust and complex imaging features with greater generalization ability.

Several types of DL approaches have been developed for different purposes, such as object detection and segmentation in images, speech recognition, genotype and phenotype detection, and classification of diseases. Some of the popular DL algorithms are stacked autoencoders [2], deep Boltzmann machines, deep-belief neural networks [3], and convolutional neural networks (CNNs) [4]. CNNs are the algorithms most commonly applied to images. Since their first introduction in 1989 [5], CNNs have been widely applied to classification and segmentation of photographic images with great success [4,6,7].

DL techniques achieve impressive results and robustness by training on large amounts of data. They are also gaining popularity in many areas of medical image analysis [8], such as tissue and lesion segmentation [1,9-13], lesion diagnosis [1,14-18], and histopathologic analysis [19,20]. CNN architectures are increasingly complex, with some systems having more than 100 layers, which means millions of weights and billions of connections among neurons. A typical CNN architecture contains multiple convolution, max-pooling, and activation layers. Convolutional layers produce feature maps by convolving a convolutional kernel across the input image. Max-pooling is used to downsample the output of convolutional layers by passing the maximum value of a defined neighborhood to the next layer. Rectified linear unit is one of the most commonly used activation functions. It nonlinearly transforms data by clipping any negative input values to zero, while positive input values are passed as output [21]. To perform a prediction from input data, the output scores of final CNN layer are connected to a softmax nonlinearity function that normalizes scores into multinomial distribution over labels. Also, an optimizer that minimizes the error between prediction and ground-truth labels through a loss function and a gradient backpropagation method that updates weights at each iteration are used to train CNN architectures until they converge to steady state (see Fig. 1).

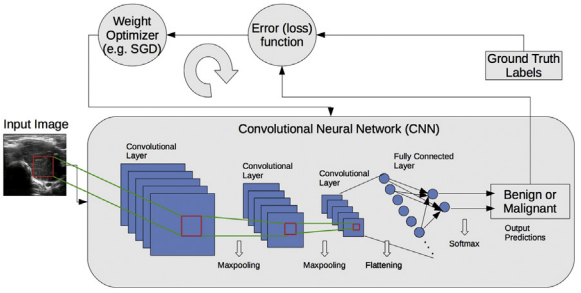

Fig 1. A schematic illustration of a deep-learning training process for a classification task. White arrows show operations between layers. CNN = convolutional neural network; SGD = stochastic gradient descent.

Ultrasound Imaging

Diagnostic ultrasonography is an ultrasound-based imaging technique used for visualizing and diagnosing pathological changes of internal organs such as liver, heart, and vessels and superficial structures such as thyroid, breast, and muscles. Ultrasound has several advantages compared with other medical imaging techniques. It is safe because it does not use harmful ionizing radiation, like radiography and CT, it is considerably lower in cost, it is portable for point-of-care applications, and it provides real-time imaging. Because it is portable, it can be transported to a patient's bedside and is useful for patient screening and follow-up. The disadvantages of ultrasound include its strong operator dependence and inability to examine areas of the body containing gas and bones. The most common types of ultrasound images are shown in Figure 2.

METHODS

We performed a thorough analysis of the literature using the Google Scholar and PubMed search engines. We included 31 peer-reviewed journal publications and conference proceedings in this field (*Medical Image Analysis*, *IEEE Transactions on Medical Imaging*, *IEEE Journal of Biomedical and Health Informatics*, *Medical Physics*, *Ultrasonics*, and conference proceedings from SPIE, the Medical Image Computing and Computer Assisted Intervention Society, the Institute of Electrical and Electronics Engineers, and others) that describe the application of DL to ultrasound before January 15, 2019 (see Fig. 3 for the identification and selection procedure). We divided reports into four groups on the basis of the frequency of studies in the literature, namely, studies on thyroid, breast, liver, and other areas.

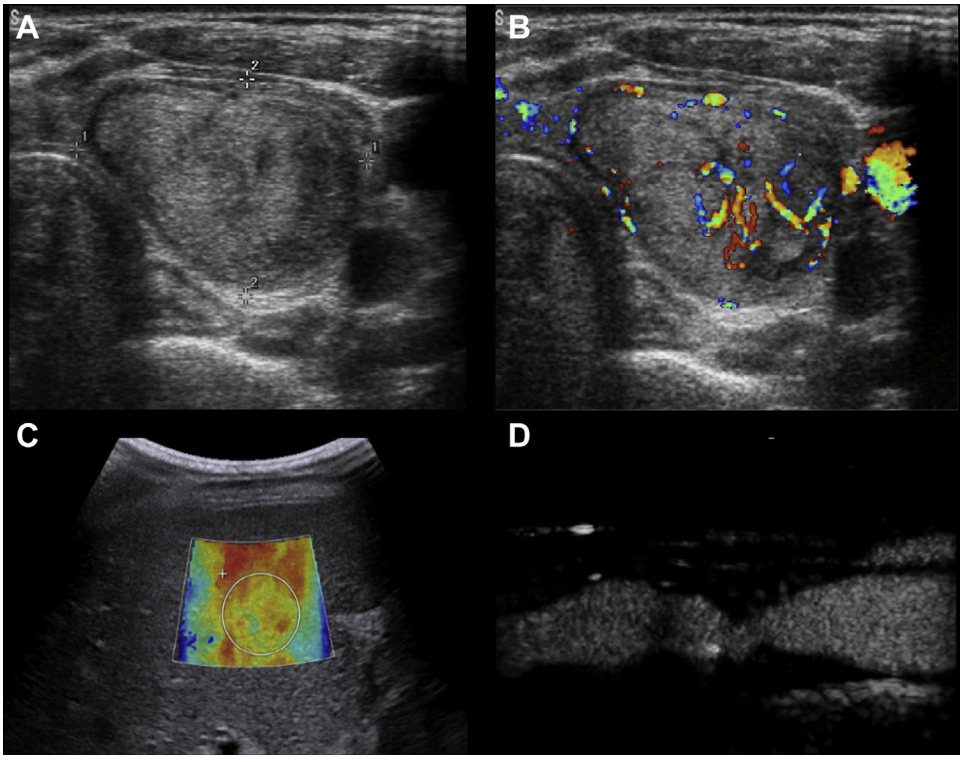

Fig 2. Examples of ultrasound images. (A) B-mode image of a thyroid nodule. (B) Color flow Doppler image of a thyroid nodule overlaid on B-mode image. (C) Shear-wave elastographic map overlaid on B-mode image of a liver with stage F3 fibrosis. (D) Contrast-enhanced ultrasound image of a carotid artery with stenosis.

Training, Validation, and Evaluation

In DL, data are divided into training, validation, and test sets to learn from examples, establish the soundness of learning results, and evaluate the generalizability of a developed algorithm on unseen data, respectively. When there are limited data, cross-validation methods (eg, k-fold) are preferred. Training is typically done with a supervised approach, which requires ground truth for the task. Most DL applications involve supervised learning, in which a DL model is trained on a data set of images with provided ground-truth labels or segmentation maps. Ground truth is usually obtained with manual delineations of lesions or structures by experts for segmentation tasks.

The optimal way to compare the performance of DL models presented in each application is to evaluate them on data sets that are publicly available with well-accepted ground truth. However, this is a big challenge in this field. We have presented the performance of each DL model on the reported validation and test data sets for each application in the "Results" section.

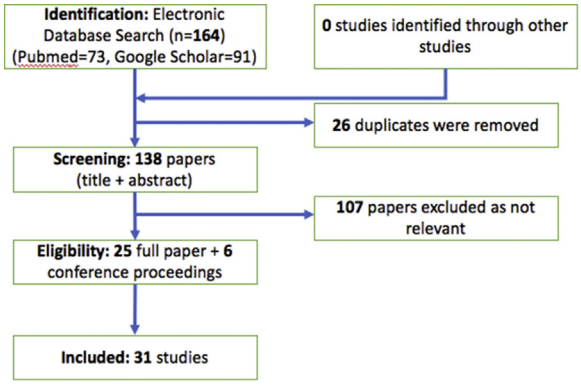

Fig 3. Flowchart for identification, screening, eligibility, and inclusion of studies.

AI-Powered Detection, Segmentation, and Diagnosis in Ultrasound

Computer-aided diagnosis (CAD) has become one of the major research subjects in diagnostic radiology. CAD provides a second opinion to assist radiologists in image interpretation by improving the accuracy and consistency of radiologic diagnosis and also by reducing image reading time. Many ultrasound CAD systems that are mainly for breast, thyroid, and liver disease detection and

classification have been proposed in the past 2 decades. The majority of these CAD systems are based on classical ML using texture features that include several processing steps: preprocessing, feature extraction and selection, and classification. Even though these studies show the promise of ultrasound CAD, they include several limitations: subjective selection of a region of interest that includes pathology, using limited training and test data sets, and having data that are collected only from a single ultrasound device at a single medical center. These limitations make it difficult to integrate ML algorithms on the basis of textures into clinical workflow to assist radiologists. In recent years, researchers have proposed DL-based ultrasound CAD systems following the success of DL in image classification and segmentation tasks [15,22]. Knowledge gained from one data set could be easily transferred to a new data set obtained from another center with another ultrasound device by fine-tuning the DL model on a new data set, which is called transfer learning.

Thyroid Nodule Detection and Classification

Thyroid nodules are extremely common lesions and are found in 50% of the adult population on the basis of autopsy studies [23-26]. The only nonsurgical test that is proven to differentiate a benign from a malignant nodule is fine-needle aspiration biopsy. Because the majority of thyroid nodules selected for fine-needle aspiration (~90%) are benign [27,28], a noninvasive and reliable method is necessary to identify nodules that do not require biopsy or surgery. This will significantly reduce health care costs and patient anxiety. To be acceptable, sensitivity at or near 100% is needed to make sure cancers are not missed with a reasonable specificity so that a good fraction of benign lesions is not biopsied.

Ma et al [29] presented a method based on the fusion of two customized CNN models that includes 7 × 7, 5 × 5, and 3 × 3 convolutional kernels to classify benign and malignant nodules. It was validated on 15,000 images collected from two local hospitals. In another study, Ma et al [30] presented a cascaded CNN model that includes two consecutive CNN models with 15 and 4 convolutional layers. It automatically detects nodules from ultrasound B-mode images in two steps. They evaluated their model with 10-fold cross-validation. Li et al [31] used Faster R-CNN [32] (which achieved best performance on the The PASCAL Visual Object Classes data set) to automatically detect papillary carcinoma and benign regions of nodules from ultrasound B-mode images. Li et al [33] presented a retrospective and multicohort study with a large population from three hospitals (332,180 images from 45,644 patients in total). They used Resnet50 [34] and Darknet [35] CNN models pretrained on the ImageNet dataset. Akkus et al [16] presented a CNN model [36] using an attention map to predict nodules from ultrasound B-mode images. Choi et al [22] used a proprietary CAD system (S-Detect for Thyroid; Samsung Medison, Seoul, South Korea) to classify 102 nodules from 89 patients as benign or malignant. Pereira et al [37] compared DL approaches for thyroid nodule characterization from shear wave elastographic (SWE) images. Chi et al [17] fine-tuned a GoogleNet [36] model to classify thyroid nodules and tested it on a data set of 61 cases.

Breast Lesion Detection and Classification

Breast cancer is the most frequent cancer among women and also the leading cause of cancer-related deaths among women [38]. The BI-RADS [39] score is used to standardize reporting and reduce confusion in breast imaging interpretations. However, significant intra- and interobserver variability on the basis of BI-RADS scoring has been reported in several studies [40,41]. CAD systems with high sensitivity and negative predictive value can be used to provide radiologists a second opinion in a cost-effective way and can help reduce unnecessary false-positive biopsies. To date, several DL approaches have been explored for objective and reproducible classification of breast lesions from ultrasound images. Byra et al [42] presented a CNN model based on transfer learning to classify breast lesions as benign or malignant. They used the VGG19 [43] CNN model pretrained on the ImageNet data set and fine-tuned it on 882 ultrasound images of breast masses. Han et al [44] trained a GoogleNet [36] model on 7,408 breast ultrasound images and tested on 829 images. This model is a component algorithm of S-Detect technology, which is implemented in RS80A (Samsung Medison). Cheng et al [15] used a stacked denoising autoencoder model [2] to classify breast lesions. Zhang et al [45] used a two-layer DL model that includes a fully connected neural network as the first layer to extract features and a restricted Boltzmann machine as the second layer to provide better feature representation. A support vector machine (SVM) classifier was connected to the restricted Boltzmann machine to predict breast lesions. Several other studies also used DL to detect breast lesions from ultrasound images [46,47].

Liver Lesion Classification

Ultrasound is the preferred imaging modality to evaluate liver diseases because it provides information about the appearance of the liver and portal venous blood flow. Although liver biopsy is sensitive in assessing cirrhosis, it is invasive and limited by sampling errors, interobserver variability, and various potential complications such as damage to the lung and gallbladder, bleeding, and infection. Wang et al [49] presented a multicenter study to assess liver fibrosis stage with DL from ultrasound SWE images. They concluded that DL-based elastography is more accurate than 2-D SWE imaging in assessing cirrhosis and advanced fibrosis and more accurate than biomarkers in assessing all three liver fibrosis stages in patients with chronic hepatitis B. Meng et al [50] used a fine-tuned VGGNet [43] and fully connected network based on transfer learning to predict the stage of liver fibrosis. Similarly, Liu et al [51] presented a pretrained CNN model that extracts features of liver capsule from ultrasound images and uses an SVM to classify a liver as normal or abnormal from extracted CNN features. Wu et al [52] trained a deep-belief network model [3] on time-intensity curves extracted from contrast-enhanced ultrasound for the classification of focal liver lesions. They showed that their method outperforms classical ML methods. Biswas et al [53] assessed fatty liver disease from ultrasound images using DL and achieved a performance superior to that of ML approaches.

Other Applications

In addition to three main applications of DL in ultrasound, several other AI applications in ultrasound have been explored by researchers in the field. Yu et al [55] used a customized CNN that contains 16 convolutional layers with 3×3 kernels and three fully connected layers to classify the fetal ultrasound plane. Wu et al [56] assessed the quality of fetal ultrasound images using two cascaded CNN models for accurate measurements. The first CNN (pretrained AlexNet) finds the region of interest of the abdominal region from ultrasound images. The second CNN that receives the region of interest from the first CNN as input evaluates the image quality by assessing the key structures of stomach bubble and umbilical vein. Chen et al [57] also used a composite CNN framework that includes a CNN model for in-plane feature extraction and long short-term memory model to classify fetal standard planes. Menchón-Lara et al [58] used an autoencoder [2] to segment intima-media thickness in a user-independent and reproducible manner. Lekadir et al [59] used a CNN that includes four convolutional and three fully connected layers to classify atherosclerotic plaque components including lipid core, fibrous tissue, and calcified tissue. Hetherington et al [60] presented a spine-level vertebra identification system using pretrained CNN models. Cheng and Malhi [61] used the VGG model [43] to classify the anatomic location and plane of abdominal ultrasound images. DL has been also used for ultrasound beamforming [62,63], image recovery [64], image reconstruction from subsampled RF data [65], and elastographic image reconstruction [66].

RESULTS

Thyroid Nodule Detection and Classification

Ma et al [29] validated their CNN model on 15,000 images collected from two local hospitals and achieved accuracy of $83.02 \pm 0.72\%$ for classifying benign and malignant nodules. However, their receiver operating characteristic (ROC) curve indicated that their model always missed some malignant cases and never reached 100% sensitivity. Ma et al [30] evaluated their nodule detection model with 10-fold cross-validation and achieved an average area under the ROC curve (AUC) of 98.51%. The CNN model of Li et al [31] for detecting papillary carcinoma and benign regions of nodules from ultrasound B-mode images obtained 93.5% sensitivity and 81.5% specificity. Li et al [33] concluded on the basis of a retrospective and multicohort study with a large population from three hospitals (a total of 332,180 images from 45,644 patients) that the CNN model showed similar sensitivity and improved specificity in identifying patients with thyroid cancer compared with a group of skilled radiologists. The performance of the CNN model by Akkus et al [16] on a test set of 100 transverse and longitudinal images of 50 nodules was 86% (sensitivity) and 90% (specificity). When the threshold was set for maximum sensitivity (zero missed cancers), their ROC curve suggests that the number of biopsies may be reduced by 52% without missing patients with malignant thyroid nodules. Choi et al [22] concluded on the basis of 102 nodules from 89 patients that the sensitivity of the CAD system using AI for malignant thyroid nodules was as good as that of the experienced radiologist. Pereira et al [37] obtained the highest accuracy of 83% on 20% of a data set of 964 images from 165 patients. Chi et al [17] achieved 98.29% accuracy on a test data set of 61 cases for classification of thyroid nodules with

a fine-tuned GoogleNet model [36]. Characteristics of each model are summarized in Table 1.

Breast Lesion Detection and Classification

Byra et al [42] achieved an AUC of 93.6% over a test data set of 150 cases for classifications of breast lesions into benign or malignant lesions and concluded that the model has potential to assist radiologists with breast mass classification in ultrasound. Han et al [44] achieved accuracy of 90%, sensitivity of 86%, and specificity of 96% on the test data set (n = 829 images). Cheng et al [15] achieved an AUC of 89.6% for classification of breast lesions, which outperforms conventional ML-based methods. Zhang et al [45] achieved better performance with their two-layer DL model when classifying breast lesions from SWE images (accuracy 93.4% versus 84%-87%). Characteristics of each model are summarized in Table 2.

Liver Lesion Classification

Several studies used DL to diagnose liver diseases from ultrasound images (see Table 3 for a summary). Wang et al [49] concluded that DL-based elastography is more accurate than 2-D SWE imaging in assessing cirrhosis and advanced fibrosis and more accurate than biomarkers in assessing all three liver fibrosis stages in patients with chronic hepatitis B. The AUCs of the proposed method were 97% for F4 (cirrhosis), 98% for F3 or higher (advanced fibrosis), and 85% for F2 or higher (significant fibrosis). Meng et al [50] achieved accuracy of 93.90% on 30% of their data set (n = 279 cases: 79 normal, 89 early-stage fibrosis, 111 late-stage fibrosis) for predicting the stage of liver fibrosis. The AUC of the method proposed by Liu et al [51] for classifying a liver as normal or abnormal reached 96.8%, which is superior to the accuracy of low-level features. Wu et al [52] showed that their method outperforms

Table 1. Model characteristics for thyroid nodule detection and classification

Study	Total Number of Images (Number of Patients)	Test Set Number of Images (Number of Patients)	Task	Method	Performance Metrics
Ma et al (2017) [29]	15,000 (4,782)	10-fold cross-validation	Classification	Two fused CNN models	Accuracy: 83.02% TPR: 82.41% TNR: 84.96% AUC: 89.30%
Ma et al (2017) [30]	21,523 (5,842)	10-fold cross-validation	Detection	Two cascaded CNN models	AUC: 98.51%
Li et al (2018) [31]	4,670 (300)	1,027 (100)	Detection	Faster R-CNN [32]	TPR: 93.5% TNR: 81.5% AUC: 93.8%
Li et al (2018) [33]	332,180 (45,644)	19,781 (2,692)	Classification	Resnet50 [34] and Darknet-19 [35]	Accuracy: 88.9% TPR: 92.2% TNR: 87.1% AUC: 94.7%
Akkus et al (2019) [16]	300 (300)	100 (50)	Classification	Inception [36]	TPR: 86% TNR: 90%
Choi et al (2017) [22]	102 (89)	102 (89)	Classification	Proprietary	Accuracy: 81.4% TPR: 90.7% TNR: 74.6% AUC: 83%
Pereira et al (2018) [37]	964 (165)	193 (33)	Classification	CNN based on AlexNet [4]	Accuracy: 83% TPR: 95% TNR: 40% AUC: 80%
Chi et al (2017) [17]	428 + 208 (NA)	549 (61)	Classification	GoogleNet [36]	Accuracy: 98% TPR: 99% TNR: 94%

Note: AUC = area under receiver operating characteristic curve; CNN = convolutional neural network; NA = not available; TNR = true negative rate (specificity); TPR = true positive rate (sensitivity).

Table 2. Model characteristics for breast lesion detection and classification

Study	Total Number of Images (Number of Patients)	Test Set Number of Images (Number of Patients)	Task	Method	Performance Metrics
Byra et al (2018) [42]	882 (NA)	150 (150)	Classification	CNN based on VGG19 [43]	Accuracy: 88.7% TPR: 84.8% TNR: 89.7% AUC: 93.6%
Han et al (2017) [44]	7,408 (5,151)	829 (NA)	Classification	CNN based on GoogLeNet [36]	Accuracy: 91.23% TPR: 84.29% TNR: 96.07% AUC: 96.01%
Cheng et al (2016) [15]	520 (520)	10-fold cross-validation	Classification	Stacked denoising autoencoder [2]	Accuracy: 82.4% TPR: 78.7% TNR: 85.7% AUC: 89.6%
Zhang et al (2016) [45]	227 (121)	5-fold cross-validation	Classification	Proprietary	Accuracy: 93.4% TPR: 88.6% TNR: 97.1%
Yap et al (2018) [46]	306 + 163 (NA)	10-fold cross-validation	Detection	FCN-AlexNet [4]	TPF: 98%+92%
Kumar et al (2018) [47]	433 (258)	61 (NA)	Segmentation	U-Net [48]	Dice: 82% TPF: 84%

Note: AUC = area under receiver operating characteristic curve; CNN = convolutional neural network; NA = not available; TNR = true negative rate (specificity); TPF = true positive fraction; TPR: true positive rate (sensitivity).

classical ML methods (accuracy 86.36% versus 66.67%-81.86%). Biswas et al [53] achieved performance superior to that of ML approaches (accuracy 82% [SVM] versus 92% [extreme learning machine (ELM)] versus 100% [DL]) in the assessment of fatty liver disease from ultrasound images using DL. According to the studies presented here, DL can significantly improve the diagnosis of liver diseases and could potentially reduce unnecessary liver biopsies and related health care costs.

Table 3. Model characteristics for liver lesion classification

Study	Total Number of Images (Number of Patients)	Test Set Number of Images (Number of Patients)	Task	Method	Performance Metrics
Wang et al (2018) [49]	1,990 (398)	One-third of total	Classification	CNN on radiomic features from SWE	TPR: 69.1%-96.9% TNR: 88.0%-98.3% AUC: 85%-98%
Meng et al (2017) [50]	1,674 (279)	30% of total	Classification	CNN based on VGGNet [43]	Accuracy: 93.90%
Liu et al (2017) [51]	91 (91)	3-fold cross-validation	Classification	CNN + SVM classification	Accuracy: 89.2% AUC: 96.8%
Wu et al [52]	22 (26)	10-fold cross-validation	Classification	Restricted Boltzmann machine	Accuracy: 86.36% TPR: 83.33% TNR: 87.50%
Byra et al [54]	55 (55)	Leave-one-out cross-validation	Classification	ResNet [34]	AUC: 97.7%
Biswas et al 2018 [53]	63 (63)	10-fold cross-validation	Classification	Inception [36]	Accuracy: 99%

Note: AUC = area under receiver operating characteristic curve; CNN = convolutional neural network; SVM =support vector machine; TNR = true negative rate (specificity); TPR = true positive rate (sensitivity).

Other Applications

The method proposed by Yu et al [55] for classification of fetal ultrasound plane obtained accuracy of 93.03%, which is superior to that of traditional ML methods. Wu et al [56] compared their CNN model with the subjective image quality assessment of three physicians and concluded that the performance of their model is comparable with the physicians' ratings. The T-RNN model of Chen et al [57] achieved an AUC of 0.95 for detecting fetal standard plane from ultrasound videos. Menchón-Lara et al [58] obtained intima-media thickness measurements using a radial basis function network with acceptable errors compared with manual measurements. Lekadir et al [59] presented Pearson correlation coefficients of 0.92, 0.87, and 0.93 for lipid core, fibrous tissue, and calcified tissue, respectively, between areas calculated by an expert clinician and the proposed CNN model. The experiment indicates that the classification accuracy of CNNs is much better than that of SVMs. Hetherington et al [60] concluded that their model could accurately detect the vertebral level so that the anesthesiologist can find the right site to inject anesthetic. The trained VGGNet [43] of Cheng and Malhi [61] classified 77.9% of abdominal ultrasound images correctly on the test data set (1,109 of 1,423 images), which is comparable with a radiologist's performance (71.7%).

Discussion and Outlook

Although DL methods for ultrasound provide promising results, AI-powered ultrasound is still far behind the progress in AI-powered CT and MRI because of high intra- and inter-reader variability in ultrasound image acquisition and interpretation.

Most of the DL applications in ultrasound were trained and evaluated on limited, single data sets obtained from a single medical center and a single ultrasound device. To overcome this limitation, transfer learning and fine-tuning a DL model previously trained on optical or natural-world images as previously proposed [42,50,61] can be applied. However, it would be more appropriate to apply transfer learning on a DL model trained natively on ultrasound images and fine-tune the model on a new data set obtained from a different medical center and/or a different ultrasound device. Another approach to overcome the limitation of having a limited data set is to use data augmentation (eg, tissue deformation, translations, horizontal flipping, adding noise, and enhancing images) to improve the generalization ability of DL models. However, data augmentation parameters should be chosen cautiously to realistically mimic variations in ultrasound images. For example, vertically flipping ultrasound images will not be a realistic transformation, because shadowing never appears in the opposite direction of the ultrasound beam.

To build trust in an AI system designed for disease prediction from medical images, we must build transparent models that explain how and what they predict. Understanding the inner workings of a CNN requires interpreting the feature activity in each layer [67-69]. However, these activities become complex and abstract in deep CNN layers, and therefore it is more difficult to interpret them but usually results in more robust and generalizable features. By using DL models, we sacrifice interpretability for robustness and complex imaging features with greater generalization ability. The convolutional properties of CNN layers have been projected back to the input pixel space, showing what input pattern originally resulted in an activation in the feature maps. This explains which regions of the image play an important role in maximizing the classification accuracy. Several techniques have been used to investigate what DL sees and make CNN understandable, including deconvolutional networks [70], gradient back-propagation [71], class activation maps [72], gradient-weighted class activation maps [73], and saliency maps [74,75] for multiple CNN architectures [43,48,76-78]. This is an active area of research in the DL community.

Current DL models for ultrasound diagnosis use only 2-D cross-sectional images for making predictions. However, the information in 2-D cross-sections is limited and does not represent lesions completely. DL models trained on 3-D ultrasound data, ultrasound cine clips with multiple views of lesions, or spatiotemporal data could potentially improve the diagnostic accuracy of models and consider complete lesions. Furthermore, developing DL models that are trained on multimodal (B-mode, Doppler, contrast-enhanced ultrasound, and SWE) images, which provide complementary information to one another, could also improve the diagnostic accuracy of DL models.

In summary, AI-powered ultrasound systems that evaluate multimodal data, guide sonographers, and provide objective qualifications (eg, standard view of an organ and acceptable image quality), measurements, and diagnosis will not only assist with decision making but

also improve ultrasound clinical workflow and reduce health care costs.

TAKE-HOME POINTS

- DL models that are trained on multimodal (B-mode, Doppler, contrast-enhanced ultrasound, and SWE) images, which provide complementary information to one another, could also improve the diagnostic accuracy of DL models.
- To build trust in an AI system designed for disease prediction from medical images, we must build transparent models that explain how and what they predict.
- The generalization ability of DL-based diagnosis approaches have been proved to be superior than traditional ML approaches.
- DL models trained on 3-D ultrasound data, ultrasound cine clips with multiple views of lesions, or spatiotemporal data could potentially improve the diagnostic accuracy of models and consider complete lesions.
- Providing guidance to operators with AI during data acquisition and measurement would make ultrasound more intelligent and less operator dependent.

REFERENCES

1. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging 2017;30:449-59.
2. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 2010;11:3371-408.
3. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput 2006;18:1527-54.
4. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. Advances in Neural Information Processing Systems 25. Red Hook, New York: Curran Associates; 2012:1097-105.
5. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten ZIP code recognition. Neural Comput 1989;1:541-51.
6. Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. Available at: https://ieeexplore.ieee.org/document/5206848. Accessed June 18, 2019.
7. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211-52.
8. Lin D, Vasilakos AV, Tang Y, Yao Y. Neural networks for computer-aided diagnosis in medicine: a review. Neurocomputing 2016;216:700-8.
9. Akkus Z, Kostandy P, Philbrick AK, Erickson BJ. Robust brain extraction tool for CT head images. Neurocomputing. In press.
10. Akkus Z, Kostandy P, Phillbrick K, Erickson BJ. Extraction of brain tissue from CT head images using fully convolutional neural networks. Proc SPIE 10574, Medical Imaging 2018: Image Processing, 1057420. Available at: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10574/1057420/Extraction-of-brain-tissue-from-CT-head-images-using-fully/10.1117/12.2293423.short?SSO=1. Accessed June 18, 2019.
11. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60-88.
12. Milletari F, Navab N, Ahmadi S. V-Net: fully convolutional neural networks for volumetric medical image segmentation. Available at: https://arxiv.org/abs/1606.04797. Accessed June 18, 2019.
13. Weston AD, Korfiatis P, Kline TL, et al. Automated abdominal segmentation of CT scans for body composition analysis using deep learning. Radiology 2018;290:669-79.
14. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303-12.
15. Cheng J-Z, Ni D, Chou Y-H, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep 2016;6:24454.
16. Akkus Z, Boonrod A, Stan M, Castro R, Erickson BJ. Reduction of thyroid nodule biopsies using deep learning. Proc SPIE 10949, Medical Imaging 2019: Image Processing, 109490W. Available at: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10949/109490W/Reduction-of-unnecessary-thyroid-biopsies-using-deep-learning/10.1117/12.2512574.short. Accessed June 18, 2019.
17. Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. J Digit Imaging 2017;30:477-86.
18. Brattain LJ, Telfer BA, Dhyani M, Grajo JR, Samir AE. Machine learning for medical ultrasound: status, methods, and future opportunities. Abdom Radiol (NY) 2018;43:786-99.
19. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.
20. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29.
21. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: IEEE International Conference on Computer Vision (ICCV). Available at: https://ieeexplore.ieee.org/document/7410480. Accessed June 18, 2019.
22. Choi YJ, Baek JH, Park HS, et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment. Thyroid 2017;27:546-52.
23. Burman KD, Wartofsky L. Thyroid nodules. N Engl J Med 2015;373:2347-56.
24. Jemal A, Murray T, Ward E, et al. Cancer statistics 2005. CA Cancer J Clin 2005;55:10-30.
25. Hegedüs L. The thyroid nodule. N Engl J Med 2004;351:1764-71.
26. American Thyroid Association Guidelines Taskforce on Thyroid Nodules and Differentiated Thyroid CancerCooper DS, Doherty GM, et al. Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid 2009;19:1167-214.
27. Frates MC, Benson CB, Charboneau JW, et al. Management of thyroid nodules detected at US: Society of Radiologists in Ultrasound consensus conference statement. Ultrasound Q 2006;22:231-8.
28. Guille JT, Opoku-Boateng A, Thibeault SL, Chen H. Evaluation and management of the pediatric thyroid nodule. Oncologist 2015;20:19-27.
29. Ma J, Wu F, Zhu J, Xu D, Kong D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics 2017;73:221-30.

30. Ma J, Wu F, Jiang T, Zhu J, Kong D. Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images. Med Phys 2017;44:1678-91.
31. Li H, Weng J, Shi Y, Gu W, Mao Y, Wang Y, et al. An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images. Sci Rep 2018;8:6600.
32. Girshick R. Fast r-cnn. arXiv. Available at: https://arxiv.org/abs/1504.08083?context=cs.CV. Accessed June 18, 2019.
33. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol 2019;20:193-201.
34. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Available at: https://ieeexplore.ieee.org/document/7780459. Accessed June 18, 2019.
35. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. ProcarXiv. Available at: https://arxiv.org/abs/1612.08242. Accessed June 18, 2019.
36. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Available at: https://ieeexplore.ieee.org/document/7298594. Accessed June 18, 2019.
37. Pereira C, Dighe M, Alessio AM. Comparison of machine learned approaches for thyroid nodule characterization from shear wave elastography images. Medical Imaging 2018: Computer-Aided Diagnosis, vol 10575, International Society for Optics and Photonics; 2018, p. 105751X. Available at: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10575/105751X/Comparison-of-machine-learned-approaches-for-thyroid-nodule-characterization-from/10.1117/12.2294572.short. Accessed June 18, 2019.
38. World Health Organization. Breast cancer. Available at: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/. Accessed June 18, 2019.
39. Shikhman R, Keppke AL. Breast, Imaging, Reporting and Data System (BI RADS). Treasure Island, Florida: StatPearls; 2018.
40. Schwab F, Redling K, Siebert M, Schötzau A, Schoenenberger C-A, Zanetti-Dällenbach R. Inter- and Intra-observer agreement in ultrasound BI-RADS classification and real-time elastography Tsukuba score assessment of breast lesions. Ultrasound Med Biol 2016;42:2622-9.
41. Grimm LJ, Anderson AL, Baker JA, et al. Interobserver variability between breast imagers using the fifth edition of the BI-RADS MRI lexicon. AJR Am J Roentgenol 2015;204:1120-4.
42. Byra M, Galperin M, Ojeda-Fournier H, 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.
43. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv. Available at: https://arxiv.org/abs/1409.1556. Accessed June 18, 2019.
44. Han S, Kang H-K, Jeong J-Y, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys Med Biol 2017;62:7714-28.
45. Zhang Q, Xiao Y, Dai W, et al. Deep learning based classification of breast tumors with shear-wave elastography. Ultrasonics 2016;72:150-7.
46. Yap MH, Pons G, Marti J, et al. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform 2018;22:1218-26.
47. Kumar V, Webb JM, Gregory A, et al. Automated and real-time segmentation of suspicious breast masses using convolutional neural network. PLoS ONE 2018;13:e0195816.
48. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, vol 9351. Cham, Switzerland: Springer International; 2015:234-41.
49. Wang K, Lu X, Zhou H, et al. Deep learning radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study. Gut 2019;68:729-41.
50. Meng D, Zhang L, Cao G, Cao W, Zhang G, Hu B. Liver fibrosis classification based on transfer learning and FCNet for ultrasound images. IEEE Access 2017;5:5804-10.
51. Liu X, Song JL, Wang SH, Zhao JW, Chen YQ. Learning to diagnose cirrhosis with liver capsule guided ultrasound image classification. Sensors 2017;17:149.
52. Wu K, Chen X, Ding M. Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound. Optik 2014;125:4057-63.
53. Biswas M, Kuppili V, Edla DR, et al. Symtosis: a liver ultrasound tissue characterization and risk stratification in optimized deep learning paradigm. Comput Methods Programs Biomed 2018;155:165-77.
54. Byra M, Styczynski G, Szmigielski C, et al. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. Int J Comput Assist Radiol Surg 2018;13:1895-903.
55. Yu Z, Tan E-L, Ni D, et al. A deep convolutional neural network-based framework for automatic fetal facial standard plane recognition. IEEE J Biomed Health Inform 2018;22:874-85.
56. Wu L, Cheng J-Z, Li S, Lei B, Wang T, Ni D. FUIQA: fetal ultrasound image quality assessment with deep convolutional networks. IEEE Trans Cybern 2017;47:1336-49.
57. Chen H, Wu L, Dou Q, et al. Ultrasound standard plane detection using a composite neural network framework. IEEE Trans Cybern 2017;47:1576-86.
58. Menchón-Lara R, Sancho-Gómez J. Ultrasound image processing based on machine learning for the fully automatic evaluation of the carotid intima-media thickness. In: 12th International Workshop on Content-Based Multimedia Indexing (CBMI). Available at: https://ieeexplore.ieee.org/document/6849839. Accessed June 18, 2019.
59. Lekadir K, Galimzianova A, Betriu A, et al. A convolutional neural network for automatic characterization of plaque composition in carotid ultrasound. IEEE J Biomed Health Inform 2017;21:48-55.
60. Hetherington J, Lessoway V, Gunka V, Abolmaesumi P, Rohling R. SLIDE: automatic spine level identification system using a deep convolutional neural network. Int J Comput Assist Radiol Surg 2017;12:1189-98.
61. Cheng PM, Malhi HS. Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. J Digit Imaging 2017;30:234-43.
62. Nair AA, Tran TD, Reiter A, Bell MAL. A deep learning based alternative to beamforming ultrasound images. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Available at: https://pulselab.jhu.edu/wp-content/uploads/2018/04/Nair_ICASSP_2018.pdf. Accessed June 18, 2019.
63. Luchies AC, Byram BC. Deep neural networks for ultrasound beamforming. IEEE Trans Med Imaging 2018;37:2010-21.
64. Perdios D, Besson A, Arditi M, Thiran J. A deep learning approach to ultrasound image recovery. In: IEEE International Ultrasonics Symposium. Available at: https://ieeexplore.ieee.org/document/8092746. Accessed June 18, 2019.
65. Yoon YH, Yoon, Khan S, Huh J, Ye JC. Efficient B-mode ultrasound image reconstruction from sub-sampled RF data using deep learning. IEEE Trans Med Imaging. Available at: https://ieeexplore.ieee.org/document/8432500. Accessed June 18, 2019.
66. Wu S, Gao Z, Liu Z, Luo J, Zhang H, Li S. Direct reconstruction of ultrasound elastography using an end-to-end deep neural network. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, eds. Medical Image Computing and Computer Assisted Intervention—MICCAI 2018, vol 11070. Cham, Switzerland: Springer International; 2018:374-82.
67. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. arXiv. Available at: https://arxiv.org/abs/1311.2901. Accessed June 18, 2019.
68. Zeiler MD, Taylor GW, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning. In: International Conference

on Computer Vision. Available at: https://ieeexplore.ieee.org/document/6126474. Accessed June 18, 2019.
69. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Available at: https://ieeexplore.ieee.org/document/7780688.
70. Zeiler MD, Krishnan D, Taylor GW, Fergus R. Deconvolutional networks. Piscataway, New Jersey: IEEE Computer Society; 2010.
71. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. arXiv. Available at: https://arxiv.org/abs/1412.6806. Accessed June 18, 2019.
72. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. arXiv. Available at: https://arxiv.org/abs/1512.04150. Accessed June 18, 2019.
73. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV). Available at: https://ieeexplore.ieee.org/document/8354201. Accessed June 18, 2019.
74. Li G, Yu Y. Visual saliency detection based on multiscale deep CNN features. IEEE Trans Image Process 2016;25:5012-24.
75. Philbrick KA, Yoshida K, Inoue D, et al. What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. AJR Am J Roentgenol 2018;211:1184-93.
76. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv. Available at: https://arxiv.org/abs/1512.03385. Accessed June 18, 2019.
77. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. arXiv. Available at: https://arxiv.org/abs/1512.00567. Accessed June 18, 2019.
78. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:2481-95.

View publication stats