Deep Learning in Kidney Ultrasound: Overview, Frontiers, and Challenges

Hector J. De Jesus-Rodriguez, Matthew A. Morgan, and Hersh Sagreiya

Ultrasonography is a practical imaging technique used in numerous health care settings. It is relatively inexpensive, portable, and safe, and it has dynamic capabilities that make it an invaluable tool for a wide variety of diagnostic and interventional studies. Recently, there has been a revolution in medical imaging using artificial intelligence (AI). A particularly potent form of AI is deep learning, in which the computer learns to recognize pixel or written data on its own without the selection of predetermined features, usually through a specific neural network architecture. Neural networks vary in architecture depending on their task, and key design considerations include the number of layers and complexity, data available, technical requirements, and domain knowledge. Deep learning models offer the potential for promising innovations to workflow, image quality, and vision tasks in sonography. However, there are key limitations and challenges in creating reliable and safe AI models for patients and clinicians.

Key Words: Ultrasound, Machine learning, Deep learning, Artificial intelligence, Kidney

Ultrasonography (US) relies on transducers that transmit and receive ultra-high-frequency sound wave pulses via a piezoelectric element. Its relative low cost, portability, lack of ionizing radiation, relative safety, and real-time/dynamic examination make ultrasound a practical modality for a wide variety of diagnostic and interventional procedures.[1] Recent advances in point-of-care ultrasound (POCUS), which make use of ultrasound-on-a-chip technology, have further decreased the cost and increased the portability of ultrasound, allowing image display on mobile devices using pocket probes.[2] Ultrasound can also serve as an alternative to computed tomography (CT) and magnetic resonance imaging (MRI) when there are contraindications to these studies.

Despite its advantages, however, there are a few drawbacks: The potential for bioeffects from mechanical and thermal mechanisms may limit scan time; the quality and accuracy of the examination depends on an operator's level of training; and technical factors such as large body habitus, limited patient mobility, imaging artifacts, and acoustic impedance can degrade image quality, sometimes severely.[3] In this review, we look further into how deep learning (DL) tools can be a game changer for ultrasound, as well as the current challenges to the adoption of DL methods.

From the Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, PA.

Financial Disclosure: The authors declare that they have no relevant financial interests.

Support: Hersh Sagreiya was supported by an RSNA Scholar Grant from the Radiological Society of North America.

Address correspondence to Hersh Sagreiya, MD, Hospital of the University of Pennsylvania, Department of Radiology, Division of Abdominal Imaging, 3400 Spruce Street, Philadelphia, PA 19104-4283. E-mail: hersh.sagreiya@pennmedicine.upenn.edu

1548-5595/$36.00

https://doi.org/10.1053/j.ackd.2021.07.004

AI, MACHINE LEARNING, AND THE BASICS OF TRAINING ALGORITHMS

Artificial intelligence involves the development of computer algorithms capable of performing complex tasks that have traditionally been accomplished by humans; in medicine, these include image recognition, clinical decision-making, disease diagnosis, and treatment effectiveness monitoring, among many other tasks.[4] With modern advances in computer hardware and an unprecedented (and ever increasing) volume of data available, machine learning (ML) algorithms can extract relevant information from input data and train a computer to predict a known or unknown output.[5] ML seeks to optimize the performance of a particular task; in doing so, it employs the use of training, validation, and testing data sets.[6] The algorithms are trained using a specified training data set, and as the computer is being trained, its performance improvement is tracked using a separate validation data set. For final determination of model performance, many ML strategies employ the use of a separate test set, new data that the algorithm has never seen before. If the algorithm fits the training data extremely well but does not generalize to other data, this is known as overfitting, a major problem in ML.

Regardless of the complexity of the algorithm used, both the accuracy and legitimacy of AI models directly correlate with the quality of the data sets that are used. Biases may be unintentionally introduced through a low-quality data set, either because it lacks adequate representation of the patient population or because it was inappropriately curated. This is a practical problem—deficiencies introduced into an AI model by bias may cause harm.

DL AND CONVOLUTIONAL NEURAL NETWORKS

Modern graphics processing units allow deep learning algorithms to process a large amount of data efficiently because of their ability to perform many computations in parallel. Consequently, challenging applications that used to require costly mathematical computations for real-time results are now possible even on relatively inexpensive systems. Neural networks consist of mathematical functions (represented as artificial neurons) arranged in layers, taking inspiration from the biological neural system (Fig 1).[7] The layers are arranged in a hierarchical order. The most superficial layers have the lowest functionality, reading the input data. The output layer is the result of the series of computations performed by the neural network. In between the input and output layers are the hidden layers, which is where most of the processing occurs and where memory of certain data representations is stored. Each subsequent layer extracts particular information from its input data, and each neuron is connected to the previous layer and to subsequent layers by a series of weights, resulting in multiple distinctive features.

DL is a class of ML that consists of a neural network with many layers. Convolutional neural networks (CNNs) are a particular type of deep learning that is specialized for analyzing image data (Fig 2). Deep learning takes in raw data input, integrates it through several intermediate layers that offer increasingly complex abstractions of that initial data, and then integrates it into a meaningful output layer.[8] The most common AI applications related to radiologic imaging include segmentation, detection, and classification.

The order and the methods by which each layer communicates with the others in the network is termed the architecture. There are several types of architectures that are ideal for various kinds of applications. Segmentation and classification functions for unanimated images often use a 2D or U-Net architecture, whereas videos or cine clips may make use of a 3D V-Net or more complex architectures.[9] Increasing layers and complexity within a CNN can make it more powerful and more accurate, if appropriate measures have been put in place to prevent overfitting; however, more layers can lead to increased processing time and computational cost. Recent studies, however, have suggested that modified shallow neural networks can perform similar or even better than deeper networks.[10] A typical workflow for a deep learning study is shown in Figure 3.

CLINICAL SUMMARY

- Ultrasound is a relatively inexpensive and widely available imaging modality.
- The recent revolution in deep learning promises opportunities to improve clinical workflow and increase imaging quality in ultrasound.
- Key challenges associated with deep learning in ultrasound include reliability, generalizability, and bias.
- Deep learning has the potential to dramatically change ultrasound as an imaging modality.

IMAGE QUALITY

Image quality is one potential application for deep learning. It is no surprise that image quality increases the accuracy of a radiologic study. Deep learning tools are being developed to improve data-processing, decreasing artifact and power requirements. For example, Luijten and associates discuss the application of deep learning for beamforming in ultrasound images.[11] Ultrasound machines need to reconstruct images using the raw data acquired from the sensor/transducer. Many commercial probes use a low-complexity beamformer for this task. The authors showed that they could perform high-quality ultrasound beamforming using deep learning, which could maintain a high image quality, even while measuring at low data rates, which poses less of a computational burden and could be a boon for inexpensive ultrasound systems.

Another potential application for deep learning consists of the elimination of imaging artifacts. Color flow imaging is a critical tool in kidney ultrasound for analyzing blood flow to the kidneys, and it is particularly valuable for transplant evaluation. However, Doppler ultrasound can lead to a phenomenon known as "aliasing" if the vessel flow velocity exceeds a limit that occurs when the ultrasound transducer's pulse repetition frequency is insufficient for that velocity of flow. Aliasing can make it difficult to properly interpret the pattern of vascular flow. Nahas and colleagues used a set of 2 CNNs to resolve flow artifacts, which identified and segmented pixels with aliasing, and it corrected aliasing using a phase-unwrapping algorithm.[12] Consequently, the algorithm was able to render the triphasic dynamics of the femoral artery bifurcation, an area susceptible to aliasing because of its geometry; for dealiased pixels, the root-mean-squared difference was 2.51% or less than that of manually performed dealiasing, a time-consuming process.

It is also possible to use deep learning to exploit certain intrinsic characteristics of raw ultrasound data, making it more informative. For instance, one group was motivated by research indicating that longitudinal speed of sound carries information of similar diagnostic utility to shear wave imaging, which requires specialized ultrasound hardware capable of performing elastography.[13] The authors were able to recover sound speed maps from the raw ultrasound channel data. They validated their technique on simulated data, with promising results on a small set of real-world data.

FUTURE APPLICATIONS IN NEPHROLOGY

POCUS is ultrasound performed at the bedside and interpreted in real time by the treating provider rather than ultrasound performed by a technologist and interpreted remotely by a radiologist. In general, the diagnostic efficacy of US imaging depends heavily on the operator holding the probe, and this is even more of a problem with POCUS. Intricacies of ultrasound physics and anatomical variability can make ultrasound challenging to a non–imaging-trained or inexperienced professional. Deep learning may help with this limitation and guide less-experienced ultrasound imagers. Automated POCUS could assist medical learners by removing distracting background noise, helping them to discern the most relevant pathology; moreover, deep learning algorithms could label images for learners, while providing real-time feedback.[14] This could be useful for both diagnostic ultrasound

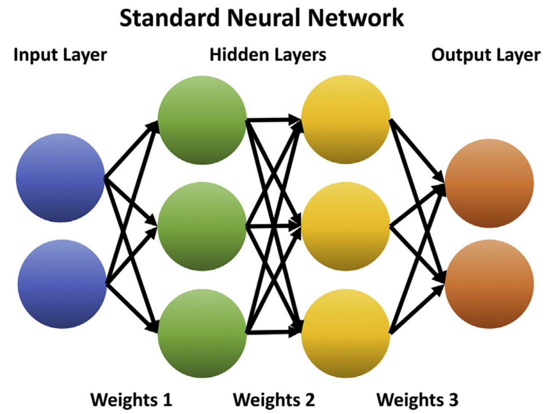

Figure 1. Standard neural network architecture. The data are input into the *input layer*. It is then transformed as it goes through a number of *hidden layers* in the network via mathematical operations that are affected by the current set of *weights* in the network. The data finally reaches the *output layer*, which is the current set of predictions for the model. As the model is run in successive steps, it uses a process known as gradient descent to adjust the weights to minimize the difference between the network's predictions and the actual ground truth. When the model is finally trained, the final set of weights are optimized for future predictions.

and procedures, effectively decreasing their learning curve.

Several segmentation, classification, and localization AI models have been developed to make the modality safer and more operator friendly, as well as decrease procedure time. A summary of the studies mentioned in this section is included in Table 1. Given how recent most published studies are, the literature on this topic will likely grow in the coming years.

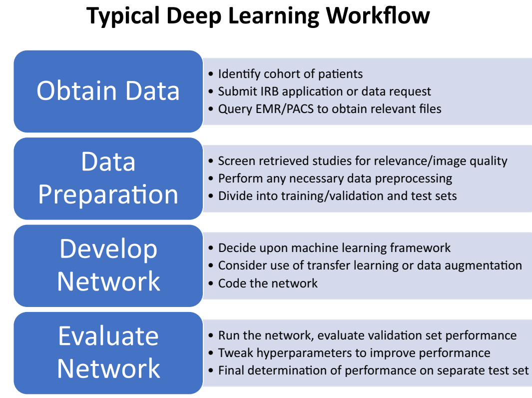

Figure 3. Typical workflow for a deep learning study. Abbreviations: EMR, electronic medical records; IRB, institutional review board; PACS, picture archiving and communication system.

Applications Directly Related to the Kidneys

Early work on deep learning in ultrasound has shown some intriguing results predicting kidney function (in terms of estimated glomerular filtration rate) based on

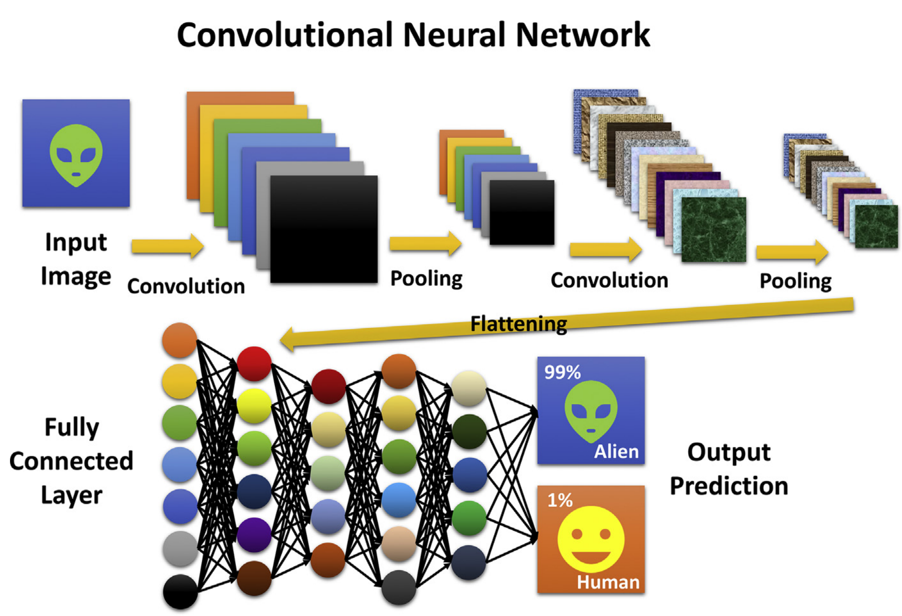

Figure 2. Convolutional neural network architecture. In a standard convolutional neural network, the input consists of an image, which is a collection of pixels. During the process of *convolution*, multiple filters are mathematically applied to the image, and these filters identify certain characteristics of the image, such as edges or spatial frequencies. During *pooling*, the spatial size of the images is reduced. The network will often involve multiple rounds of convolution and pooling, and each round will successively capture more abstract aspects of the image. Finally, the process of *flattening* takes imaging data and places it into a column or vector of data that is run through a neural network in a process similar to that outlined in Figure 1. The final output of the network is a prediction.

sonographic imaging.[15] The authors used transfer learning in conjunction with the ResNet model on kidney ultrasound images to predict CKD as determined by estimated glomerular filtration rate. The model achieved a CKD classification accuracy of 85.6%, higher than that of experienced nephrologists (60.3% to 80.1%). Deep learning is also being tested with CT and MRI for better volume measurement of the kidneys in autosomal dominant polycystic kidney disease without the need of complicated or laborious manual segmentations, and this could potentially be brought to ultrasound as well.[16] With algorithm-based segmentations, a machine can automatically calculate organ volume.

In fact, noting that segmenting the kidneys on ultrasound can be particularly challenging compared with CT or MRI, investigators used images from the Children's Hospital of Philadelphia to segment images of pediatric kidneys on ultrasound.[17] Their technique used transfer learning, as well as an image segmentation network architecture derived from DeepLab, and their technique resulted in a Dice coefficient of 0.94 ± 0.03, indicating a high level of overlap between manual and computer-derived segmentations. A similar technique could be extended for kidney volume estimation. Along a similar vein, another group used CNNs to segment fetal kidneys.[18] In order to improve segmentation accuracy, they fused both 3D B-mode images and power Doppler volumes, and they were able to achieve a Dice similarity coefficient of 0.81. Finally, Ravishankar and associates developed a fully CNN that incorporated prior shape information to better segment kidneys, and their data set included both children and adults.[19] This technique improved upon the traditional U-Net architecture by approximately 5%, with an average Dice coefficient of 83% to 84%.

A study using images from the Children's Hospital of Philadelphia looked at the use of multi-instance deep learning to distinguish between children with mild unilateral hydronephrosis and posterior urethral valves.[20] A classifier built using images in both the transverse and sagittal views obtained an area under the curve (AUC) of 0.96 ± 0.03. Another study sought to grade the degree of hydronephrosis in children using the 5-point Society for Fetal Radiology classification system.[21] Their CNN classified 94% of patients within one grade of the clinically provided label, which was reasonable performance given the slight subjectivity between adjacent grades. Many of the principles used in developing these networks in children could be extended to kidney applications for adults.

A different study sought to distinguish between 4 common categories of kidney ultrasound images: normal, cyst, kidney stone, and tumor.[22] Their technique used an ensemble of different networks (ResNET-101, ShuffleNet, and MobileNet-v2), which is a technique that combines the results of multiple models. They were able to achieve a classification accuracy of 96.5% when analyzing high-quality images and 93.7% to 95.6% when analyzing images with varying added levels of noise. While some of these are rather straightforward entities for radiologists to recognize, one could imagine a future system that analyzes categories that may present a greater diagnostic dilemma, such as renal cell carcinoma vs angiomyolipoma. Moreover, for all these applications, one could develop a system that flags certain abnormalities, either prioritizing them for radiology reads on a worklist or alerting clinicians regarding potentially time-sensitive findings.

Applications Related to Volume Status

Deep learning has also been brought to bear with lung ultrasound and evaluation of volume status. Some early work shows that DL improves classification of B-lines on POCUS lung ultrasound, and this would potentially differentiate the severity of pulmonary congestion better than a qualitative assessment.[23] Their technique resulted in a sensitivity of 93% and specificity of 96% for the presence of B-lines compared with expert readers, with kappa of 0.88. For the rating of severity on a scale of 0 to 4, kappa for agreement between the model and experts was 0.65. Similarly, DL has been applied to grade volume status based on changes in the inferior vena cava over the cardiac cycle.[24] The authors trained a deep network known as a Long Short Term Memory network, with 220 videos of the inferior vena cava serving as inputs to the model. The Fleiss Kappa measure of inter-reader agreement was 0.65 between 3 experts in POCUS, while it was moderate between experts and the model at 0.45.

Applications Related to Procedures

DL has also made inroads into guiding procedures. As nephrologists frequently perform kidney biopsies, DL techniques that can assist with ultrasound-guided procedures are of utility. Groves and colleagues collected in vitro raw US data from a carefully calibrated US probe, with a magnetic tracking system that was fixed and submerged in water.[25] A needle was imaged in various orientations at a depth of 6 cm. These images were used as ground truth to train a CNN, whose architecture was motivated by a facial KeyPoint detection algorithm. The trained model was able to perform visually validated US probe calibrations at imaging depths between 4 and 8 cm, with an average root-mean-squared error of 0.62 mm (axial) and 0.72 mm (sagittal). This could potentially one day automate needle calibration and localization. Blaivas and associates experimented with a publicly available DL algorithm (YOLOv3) and trained it on a data set of 203,966 video images to automatically label blood vessels, nerves, and tendons in transverse upper extremity ultrasound.[26] The performance of their testing data set was also compared with that of 2 fellowship-trained POCUS experts. The trained algorithm outperformed the POCUS experts in detecting all structures in the upper extremity, with an AUC of 0.78 vs 0.69 and 0.71 for the experts; with vessels, one POCUS expert achieved an AUC of 0.85, slightly ahead of the DL algorithm at 0.83. One could imagine that an algorithm developed with a similar technique that labeled key structures during kidney biopsies such as vessels and bowel could be of use during the procedure.

Other studies also analyzed deep learning for image guidance. Gillies and colleagues trained a CNN based on

Table 1. Performance of Different Deep Learning (DL) Algorithms

Author	Year	DL Model	Size*	Endpoint	Performance
Kuo CC[15]	2019	ResNet	1299	CKD estimation	Accuracy: 85.6%
Sharma K[16]	2017	fCNN	165 training, 79 testing	CT kidney segmentation	DSC: 0.86 ± 0.07
Yin S[17]	2019	VGG-16, DeepLab	105 training, 80 testing	Pediatric kidney segmentation	DSC: 0.94 ± 0.03
Weerasinghe N[18]	2020	U-Net, U-Net++	60 training, 20 validation, 20 testing	Fetal kidney segmentation	DSC: 0.81 ± 0.11
Ravishankar H[19]	2017	SR-UNet	100 training, 131 testing	Pediatric/adult kidney segmentation	Average DSC: 83.48%
Yin S[20]	2020	Multi-instance DL	86 cases, 71 controls	Posterior urethral valves vs mild hydronephrosis	AUC: 0.96 ± 0.03
Smail LC[21]	2020	5-layer CNN	2420	Grading hydronephrosis	Accuracy: 51% same grade, 94% within one grade
Sudharson S[22]	2020	Ensemble of models	4940 after data augmentation	Kidney ultrasound classification	Accuracy: 96.54%, 93.65 - 95.58% with noisy images
Baloescu C[23]	2020	10-layer CNN	300 training, 100 testing movie clips	Automated B-Line Assessment	Presence of B-lines: Kappa: 0.88, Sensitivity: 93%, Specificity: 96%, Severity: Kappa: 0.65
Blavias M, Adhikari S[24]	2020	LSTM	220 videos	Assessment of inferior vena cava collapsibility	Fleiss Kappa: 0.45
Groves LA[25]	2019	CNN	3825 training, 1519 validation, 614 testing	Out-of-plane US needle localization	Root-mean-squared error: 0.62 mm axial, 0.74 mm lateral
Blaivas M, Arntfield R[26]	2020	YOLOv3	183,555 training, 50 testing (images extracted from video)	Detection of US structures: vessels, bones, tendons, nerves	AUC: 0.78 (versus 0.69 and 0.71 for POCUS experts)
Gillies DJ[27]	2020	U-Net	917 training, 315 testing	Segmenting interventional tools	Tip error: 3.5 mm Angular error: 0.8° DSC: 73.3%

Other than Sharma (2017), all studies were on ultrasound.
Abbreviations: AUC, area under the curve; DSC, Dice similarity coefficient; fCNN, fully convolutional neural network; LSTM, Long Short Term Memory; POCUS, point-of-care ultrasound.
*The way in which articles explained data set size varied; some studies used cross-validation, whereas others explicitly listed the sizes of training, validation, or testing sets.

a U-Net architecture to segment interventional tools nearly in real time.[27] This was performed for a variety of use-cases including prostate, gynecologic, liver, and kidney applications, as well as in phantom experiments spanning a total of 917 training images with associated manual segmentations. When the authors compared the predictions to manual segmentations on a 315-image testing set, the overall median (with first and third quartiles) tip error, angular error, and Dice similarity coefficient were 3.5 (1.3, 13.5) mm, 0.8 (0.3, 1.7) degrees, and 73.3% (56.2%, 82.3%), respectively. While the authors were able to successfully show the ability of their tool to segment interventional tools in a variety of settings, further research is necessary to determine whether such a tool could safely be used in real-world settings.

CHALLENGES

Ethical Challenges and Bias

Most of the AI tools that have been developed need further validation for clinical approval and implementation. Limited data sets, latent bias, and racial bias are among the most recognized challenges that raise concerns for how these algorithms will execute in large and diverse patient populations without causing undue harm. Although there has recently been greater recognition of bias within models, it is possible that an apparently fair model could have latent biases, or biases waiting to happen. DeCamp and Lindvall discuss 3 challenges with regard to AI models: (1) a hypothetically fair algorithm, as an adaptive model, could become increasingly biased over time; (2) the interaction of AI and clinical environments, which have their own set of implicit and explicit biases, can lead to challenges; (3) the choice of the goal that the model is intended to promote could be problematic if it does not reflect the interests of the community.[28] It is important that the models built are both safe and fair for patients before they are implemented in clinical practice.

A recent article highlighted a particularly egregious example of this kind of bias. The authors studied a commercial prediction algorithm whose intent was to identify patients with complicated health care needs.[29] However, because the algorithm predicted health care costs rather than illness, unequal access to care meant that less money was spent caring for Black patients than for White patients. As a result, at any given risk score, Black patients were considerably sicker than White patients. The authors concluded that remedying this disparity would increase the percent of Black patients receiving additional help from 17.7% to 46.5%.

Regarding the topic of ethics as applied to AI in radiology, multiple prominent radiologic societies in the United States and Europe recently issued a joint statement that AI should promote well-being, minimize harms, and ensure that both benefits and harms are distributed among key stakeholders justly.[30] They further state that AI should be transparent, dependable, and unbiased, ensuring that accountability remains with its human designers and operators and promulgating the need for a new code of ethics for AI. Overall, the future use of AI algorithms presents ethical challenges, and it is crucial that these algorithms are implemented with care to prevent undue harm to vulnerable populations.

Practical Concerns Related to Training and Implementing Algorithms

There are other practical challenges for AI algorithms. First, DL typically requires a large amount of data to be successful. There are a few ways to at least partially ameliorate this issue. Data augmentation is a technique that manipulates real data (by performing various transformations upon it such as scaling, shearing, and rotating) to create synthetic data to be fed into the model.[31] Transfer learning takes in model weights from a pretrained network into a new network. Without measures such as these, DL algorithms that are presented with little data should be viewed with skepticism. This issue is related to the problem of overfitting. An algorithm that has been trained with too little data may fit the training data very well, but that does not ensure that it will generalize well to separate data in a validation set or a testing set. Potential strategies for addressing overfitting include adding more data, reducing the complexity of the model, and adding terms that prevent the model weights from getting too out of hand, known as regularization.

A second challenge relates to the time and costs of assembling the large data sets necessary for deep learning. ImageNet, one of the most famous data sets in all of ML and the catalyst for the development of several seminal ML algorithms, contains millions of labeled 2D images, and its labeling process was crowdsourced.[32] Manually labeling radiology images by clinical experts, on the other hand, would be substantially more expensive, and, unlike images of cats or bicycles, sharing medical images presents significant privacy concerns. Although radiology images are typically associated with reports, from a data perspective, the report contents are often free text and not structured in a way that is easily machine-computable. One potential solution for this challenge has been the idea of weak supervision, which can take advantage of noisy labels to train algorithms. For instance, one study used weak supervision combined with labeling functions on multiple large clinical data sets that consisted of images and their accompanying reports, effectively decreasing the requisite labeling effort from person-years to person-days.[33]

Another important practical problem for AI is the concern about generalizability.[34] Different institutions will have different machines with varying scan parameters, as well as different patient populations. This problem is exacerbated for ultrasound, which is fundamentally an operator-dependent modality that does not have the same level of consistency for image acquisition procedure as CT or MRI.[35] As a result of these factors, an algorithm trained at one institution may not work well if deployed elsewhere. Potential solutions for this problem include increasing the diversity of the training set, incorporating data from different machines and institutions, fine-tuning a model before introducing it to new clinical scenarios, and continuously monitoring an algorithm even after it has been deployed. Overall, there are significant human, ethical, and practical concerns with AI.

CONCLUSION

Although relatively early in its development, DL has begun to be tested on ultrasound applications, including those relevant to nephrologists, and it holds significant practical promise. Potential applications include disease diagnosis, kidney volume determination, image quality improvement, and evaluation of the patient's volume status, among others. In addition, AI and DL can also serve a role in POCUS education and training, increasing the confidence of learners and decreasing the learning curve. Nevertheless, the effective clinical adoption of DL algorithms presents significant challenges, both practical and ethical, and further validation is necessary before widespread clinical implementation in kidney imaging. If developed and implemented with caution and care, AI tools have innovative features that have the potential to transform ultrasound imaging and improve clinical workflow, with the eventual goal of making sonography even more reliable and accessible for kidney imaging than it is today.

REFERENCES

1. Liu S, Wang Y, Yang X, et al. Deep learning in medical ultrasound analysis: a review. *Engineering.* 2019;5(2):261-275.
2. Baribeau Y, Sharkey A, Chaudhary O, et al. Handheld point-of-care ultrasound probes: the new generation of POCUS. *J Cardiothorac Vasc Anesth.* 2020;34(11):3139-3145.
3. Hangiandreou NJ. AAPM/RSNA physics tutorial for residents. Topics in US: B-mode US: basic concepts and new technology. *Radiographics.* 2003;23(4):1019-1033.
4. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol.* 2019;19(1):64.
5. Deo RC. Machine learning in medicine. *Circulation.* 2015;132(20):1920-1930.
6. Hammad I, El-Sankary K. Practical considerations for accuracy evaluation in sensor-based machine learning and deep learning. *Sensors (Basel).* 2019;19(16):3491.
7. Zhou H, Jin Y, Dai L, et al. Differential diagnosis of benign and malignant thyroid nodules using deep learning radiomics of thyroid ultrasound images. *Eur J Radiol.* 2020;127:108992.
8. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. *Nat Genet.* 2019;51(1):12-18.
9. Hesamian MH, Jia W, He X, Kennedy P. Deep learning techniques for medical image segmentation: achievements and challenges. *J Digit Imaging.* 2019;32(4):582-596.
10. Blaivas M, Blaivas L. Are all deep learning architectures alike for point-of-care ultrasound? Evidence from a cardiac image classification model suggests otherwise. *J Ultrasound Med.* 2020;39(6):1187-1194.
11. Luijten B, Cohen R, de Bruijn FJ, et al. Adaptive ultrasound beamforming using deep learning. *IEEE Trans Med Imaging.* 2020;39(12):3967-3978.
12. Nahas H, Au JS, Ishii T, Yiu BYS, Chee AJY, Yu ACH. A deep learning approach to resolve aliasing artifacts in ultrasound color flow imaging. *IEEE Trans Ultrason Ferroelectr Freq Control.* 2020;67(12):2615-2628.
13. Feigin M, Freedman D, Anthony BW. A deep learning framework for single-sided sound speed inversion in medical ultrasound. *IEEE Trans Biomed Eng.* 2020;67(4):1142-1151.
14. Shokoohi H, LeSaux MA, Roohani YH, Liteplo A, Huang C, Blaivas M. Enhanced point-of-care ultrasound applications by integrating automated feature-learning systems using deep learning. *J Ultrasound Med.* 2019;38(7):1887-1897.
15. Kuo CC, Chang CM, Liu KT, et al. Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. *NPJ Digit Med.* 2019;2:29.
16. Sharma K, Rupprecht C, Caroli A, et al. Automatic segmentation of kidneys using deep learning for total kidney volume quantification in autosomal dominant polycystic kidney disease. *Sci Rep.* 2017;7(1):2049.
17. Yin S, Zhang Z, Li H, et al. *Fully-automatic segmentation of kidneys in clinical ultrasound images using a boundary distance regression network.* Paper presented at: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). Venice, Italy: IEEE; 2019:8-11.
18. Weerasinghe N, Lovell NH, Welsh AW, Stevenson G. Multi-parametric fusion of 3D power Doppler ultrasound for fetal kidney segmentation using fully convolutional neural networks. *IEEE J Biomed Health Inform.* 2020;25(6):2050-2057.
19. Ravishankar H, Venkataramani R, Thiruvenkadam S, Sudhakar P, Vaidya V. Learning and incorporating shape models for semantic segmentation, *Medical Image Computing and Computer Assisted Intervention − MICCAI 2017. MICCAI 2017. Lecture Notes in Computer Science.* Quebec City, QC, Canada: Springer, Cham. 2017;10433:203-211.
20. Yin S, Peng Q, Li H, et al. Multi-instance deep learning of ultrasound imaging data for pattern classification of congenital abnormalities of the kidney and urinary tract in children. *Urology.* 2020;142:183-189.
21. Smail LC, Dhindsa K, Braga LH, Becker S, Sonnadara RR. Using deep learning algorithms to grade hydronephrosis severity: toward a clinical adjunct. *Front Pediatr.* 2020;8:1.
22. Sudharson S, Kokil P. An ensemble of deep neural networks for kidney ultrasound image classification. *Comput Methods Programs Biomed.* 2020;197:105709.
23. Baloescu C, Toporek G, Kim S, et al. Automated lung ultrasound B-line assessment using a deep learning algorithm. *IEEE Trans Ultrason Ferroelectr Freq Control.* 2020;67(11):2312-2320.
24. Blaivas M, Adhikari S, Savitsky EA, Blaivas LN, Liu YT. Artificial intelligence versus expert: a comparison of rapid visual inferior vena cava collapsibility assessment between POCUS experts and a deep learning algorithm. *J Am Coll Emerg Physicians Open.* 2020;1(5):857-864.
25. Groves LA, VanBerlo B, Peters TM, Chen ECS. Deep learning approach for automatic out-of-plane needle localisation for semi-automatic ultrasound probe calibration. *Healthc Technol Lett.* 2019;6(6):204-209.
26. Blaivas M, Arntfield R, White M. Creation and testing of a deep learning algorithm to automatically identify and label vessels, nerves, tendons, and bones on cross-sectional point-of-care ultrasound scans for peripheral intravenous catheter placement by novices. *J Ultrasound Med.* 2020;39(9):1721-1727.
27. Gillies DJ, Rodgers JR, Gyacskov I, et al. Deep learning segmentation of general interventional tools in two-dimensional ultrasound images. *Med Phys.* 2020;47(10):4956-4970.
28. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. *J Am Med Inform Assoc.* 2020;27(12):2020-2023.
29. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453.
30. Geis JR, Brady AP, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *Can Assoc Radiol J.* 2019;70(4):329-334.
31. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a radiologist's guide. *Radiology.* 2019;290(3):590-606.

32. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. *ImageNet: a large-scale hierarchical image database*. Paper presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE; 2009:20-25.
33. Dunnmon JA, Ratner AJ, Saab K, et al. Cross-modal data programming enables rapid medical machine learning. *Patterns (N Y)*. 2020;1(2):100019.
34. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. *Yearb Med Inform*. 2019;28(1):16-26.
35. Akkus Z, Cai J, Boonrod A, et al. A survey of deep-learning applications in ultrasound: artificial intelligence-powered ultrasound for improving clinical workflow. *J Am Coll Radiol*. 2019;16(9 Pt B):1318-1328.