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

Michael Blaivas, MD, MBA, Robert Arntfield, MD, Matthew White, MD

Received January 6, 2020, from the University of South Carolina School of Medicine, Columbia, South Carolina, USA (M.B.); Department of Emergency Medicine, St Francis Hospital, Columbus, Georgia USA (M.B.); and Department of Critical Care Medicine, Western University, London, Ontario, Canada (R.A., M.W.). Manuscript accepted for publication February 27, 2020.

All of the authors of this article have reported no disclosures.

Address correspondence to Michael Blaivas, MD, MBA, PO Box 769209, Roswell, GA 30076 USA.

E-mail: mike@blaivas.org

Abbreviations

AUC, area under the curve; DL, deep learning; POCUS, point-of-care ultrasound; UE, upper extremity; US, ultrasound

doi:10.1002/jum.15270

Objectives—We sought to create a deep learning (DL) algorithm to identify vessels, bones, nerves, and tendons on transverse upper extremity (UE) ultrasound (US) images to enable providers new to US-guided peripheral vascular access to identify anatomy.

Methods—We used publicly available DL architecture (YOLOv3) and deidentified transverse US videos of the UE for algorithm development. Vessels, bones, tendons, and nerves were labeled with bounding boxes. A total of 203,966 images were generated from videos, with corresponding label box coordinates in a YOLOv3 format. Training accuracy, losses, and learning curves were tracked. As a final real-world test, 50 randomly selected images from unrelated UE US videos were used to test the DL algorithm. Four different versions of the YOLOv3 algorithm were tested with varied amounts of training and sensitivity settings. The same 50 images were labeled by 2 blinded point-of-care ultrasound (POCUS) experts. The area under the curve (AUC) was calculated for the DL algorithm and POCUS expert performance.

Results—The algorithm outperformed POCUS experts in detection of all structures in the UE, with an AUC of 0.78 versus 0.69 and 0.71, respectively. When considering vessels, only one of the POCUS experts attained an AUC of 0.85, just ahead of the DL algorithm, with an AUC of 0.83.

Conclusions—Our DL algorithm proved accurate at identifying 4 common structures on cross-sectional US imaging of the UE, which would allow novice POCUS providers to more confidently and accurately target vessels for cannulation, avoiding other structures. Overall, the algorithm outperformed 2 blinded POCUS experts.

Key Words—artificial intelligence; emergency medicine; deep learning; peripheral venous access; point-of-care ultrasound; vascular access

Deep learning (DL), a branch of artificial intelligence, is beginning to come to commercial reality in medical imaging. Multiple companies have sought and obtained

Food and Drug Administration clearance for various DL imaging applications, including computed tomography, chest radiography, and magnetic resonance imaging.[1–4] Deep learning applications can also be found in ultrasound (US) but have been mostly limited to large-scale, costly imaging platforms.[5,6] Recently, we have seen the introduction of DL applications in point-of-care ultrasound (POCUS) devices, including automated left ventricular ejection fraction assessment.[6] Although quite early in deployment, results for POCUS artificial intelligence functions have been mixed, with one automatic software application undergoing a class 2 Food and Drug Administration recall in early 2019.[7]

Although DL applications in traditional imaging promise improved work flow, inter-rater reliability, and faster turnaround times, the rapidly expanding and heterogeneous domain of POCUS may benefit most from DL innovations. As POCUS is increasingly spreading to new medical specialties beyond initial adopters such as emergency medicine and critical care providers, a familiar barrier to broader use persists. This barrier, now magnified by larger numbers of potential users, relates to the scarcity of training available for new POCUS providers.[8,9]

Deep learning holds the promise of POCUS automation to bridge the training gap with automated interpretation of images. Aside from functional determinations (eg, left ventricular function), a pressing area of broad uptake is US-assisted peripheral venous access. This application has merit across providers from all backgrounds: physicians, nurses, emergency medical technicians, and physician assistants.[10] We therefore sought to develop a DL application to assist in the image interpretation required for US-guided peripheral vascular access through the identification of 4 key anatomic landmarks on transverse US examinations of the upper extremity (UE).

Materials and Methods

Study Design

This was a study of DL algorithm development to automatically label blood vessels, nerves, and tendons on transverse UE US images. The study was Institutional Review Board exempted, as no patients or any patient data were used in the creation or testing of the DL algorithm.

Data

Ultrasound image data were obtained from US videos of cross-sectional scans of human UEs. Images were obtained from public domain open-access sources with no accompanying patient information, including anonymized image bank repositories, Internet posted videos and images, stock images and videos, and US vendor images and videos covering musculoskeletal, soft tissue, vascular access, and regional anesthesia categories. No patient identifiers were present on any of the image sources. Video data types included mov, avi, mp4, wmv, mp1, and mp2. Extracted single frames were all JPEG. A total of 183,522 images were imported into the training data set, out of the 203,966 total.

Data Manipulation and Labeling

All videos were kept in their original size and aspect ratio and imported into open-source video-labeling software (CVAT). CVAT was downloaded from the GitHub.com website (http://github.com/opencv/cvat). CVAT is a free online interactive video and image annotation tool for computer vision. CVAT stands for "Powerful and Efficient Computer Vision Annotation Tool." It was originally developed for labeling various objects on video for autonomous car operation. It allows users to draw bounding boxes around objects of interest. CVAT was used to label blood vessels, nerves, bones, and tendons approximately every 10 frames. The CVAT software allowed interpellation of bounding boxes in between the labeled frames. A researcher with extensive US experience was tasked with labeling and would review the entire video and adjust bounding boxes when interpellation failed to properly propagate their location between key frames. This researcher was not used as an evaluator of images later in the study. The CVAT software produced files with individual image frames comprising the videos with corresponding bounding box coordinates and labels.

Algorithm Design

We used Python programming language version 3.72 (Python Software Foundation, Wilmington, DE) with Anaconda (Anaconda, Inc, Austin, TX) to manage packages and help in scripting and use of the YOLOv3 DL

algorithm. Code for YOLOv3 is available from various public sources, including GitHub.com. YOLOv3 is a computationally efficient DL architecture designed to analyze real-time video of street scenes and similar settings. It allows object detection and, once trained to recognized them, will place bounding boxes around objects of interest such as cars, people, trees, and other objects. YOLOv3 is fed frames from video with bounding boxes around key objects and associated labels. The YOLOv3 architecture was adapted to US video through code manipulation and trial and error. Trial and error was used in finding the most effective settings for the YOLOv3 architecture for increased accuracy in training the model. For example, variables such as batch size were increased to speed training; different optimizers were tested to see which gave the best learning results (eg, momentum, adam, sgd, and rmprop); and the nms_topk was decreased from 150 to 7 to limit high numbers of meaningless boxed validation predictions that only served to confuse results.

We trained our YOLOv3 algorithm on a personal computer with an 11-GB GeForce RTX 2080 Ti graphics-processing unit (Nvidia Corporation, Santa Clara, CA), and 64 GB of RAM. Researchers adjusted the batch size and learning rates during training for optimal training times while avoiding overfitting and exploding gradients, which result in training failure. Batch sizes of 75 were ultimately the most effective as determined by testing a range of batch sizes from 10 to 100 by increments of 5 and observing training times and training validation accuracies. Two schools of thought exist regarding the optimal epochs used in YOLOv3 training. An epoch simply refers to a single cycle of training a convolutional neural network through the full data set; multiple epochs are typically required to train a convolutional neural network. One school advocates 273 epochs for optimal learning, whereas another suggests that at 68 epochs, most training reaches an inflection point beyond which there is little improvement in algorithm accuracy when tested.[11] Researchers manipulated the number of epochs and produced 4 variants: 25 epochs, 25 epochs with a low threshold for object detection, 54 epochs, and 60 epochs. Our images contained far fewer items that needed to be identified compared to

Figure 1. Area under the curve results are presented for POCUS experts and DL algorithms. DL 1, YOLOv3, 25 epochs; DL 2, YOLOv3, 25 epochs, lower threshold; DL 3, YOLOv3, 54 epochs; DL 4, YOLOv3, 60 epochs.

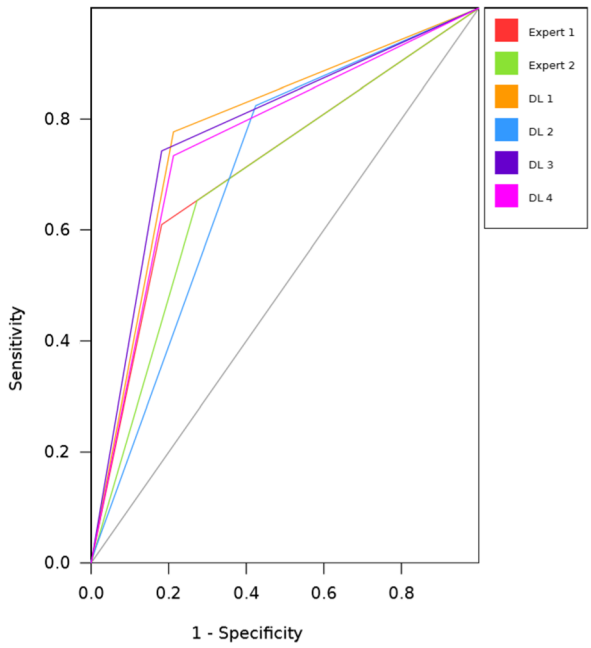


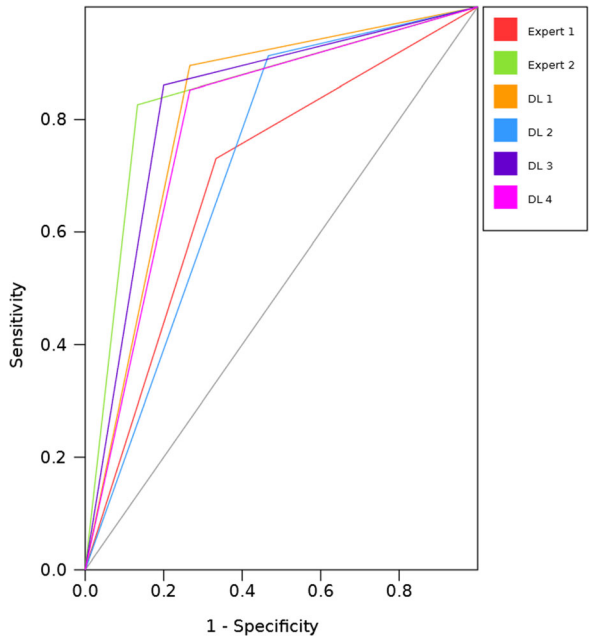
Table 1. Area Under the Curve Results for POCUS Experts and DL Algorithms for Identification of Blood Vessels, Nerves, Tendons, and Bone

				Asymptotic 95% CI	
Variable Under Test	**AUC**	**SE**	**Asymptotic *P***	**Lower Bound**	**Upper Bound**
Expert 1	0.71	0.04	<.001	0.64	0.79
Expert 2	0.69	0.05	<.001	0.61	0.77
DL 1	0.78	0.04	<.001	0.71	0.85
DL 2	0.70	0.05	<.001	0.61	0.79
DL 3	0.78	0.04	<.001	0.71	0.85
DL 4	0.76	0.04	<.001	0.69	0.83

DL 1, YOLOv3, 25 epochs; DL 2, YOLOv3, 25 epochs, lower threshold; DL 3, YOLOv3, 54 epochs; DL 4, YOLOv3, 60 epochs. CI indicates confidence interval.

original convolutional neural network scripting, so we reduced this from 400 to 4. Additionally, US images appeared to present considerable challenges to the prediction capabilities of the model. When the threshold was higher, the model appeared less likely to make a prediction. Reducing the threshold allowed the model to make predictions even when it was "less certain" or had a lower probability, yet this was required in the context of our application. We found these reduced threshold predictions to be worth including, as they were frequently correct. As an example, we reduced the score threshold from 0.3 to 0.2 and the nms threshold from 0.45 to 0.35.

Figure 2. Area under the curve results for vessel identification are presented for POCUS experts and DL algorithms. DL 1, YOLOv3, 25 epochs; DL 2, YOLOv3, 25 epochs, lower threshold; DL 3, YOLOv3, 54 epochs; DL 4, YOLOv3, 60 epochs.



Algorithm Validation and Testing

The YOLOv3 algorithm performs cross-validation automatically every 2 epochs. However, studies suggest that the final cross-validation accuracy at the end of training may not reflect algorithm performance on new data as would be encountered in real-world applications, in which patient US examinations not previously encountered by the algorithm during training may contain differently appearing anatomy and present a challenge. Additionally, it has been established that simply testing an algorithm on images generated on different US equipment can lead to poorer-than-predicted performance.[12,13] Therefore, a robust algorithm may need to be trained on a variety of US images from different subjects as well as different US equipment.

To test our algorithm's performance in a real-world–like scenario, we obtained additional US video of UE cross-sectional views that were not previously used for algorithm training. These additional videos were also sourced from the Internet. The videos were broken into 6822 single frames, and 50 frames were randomly selected for algorithm testing. Additionally, we compared the algorithm's performance to that of 2 fellowship-trained POCUS experts with 7 and 15 years of experience, respectively. The ground truth was assigned by a third fellowship-trained POCUS expert with 25 years of experience. The third POCUS expert reviewed the source US video, corresponding to each randomly selected US test image, to trace and label vessels, nerves, tendons, and bones when present. These labels were used for comparison to POCUS expert and YOLOv3 performance on the 50 test images. The third POCUS expert did not have

Table 2. Area Under the Curve Results for POCUS Experts and DL Algorithms for Identification of Blood Vessels

				Asymptotic 95% CI	
Variable Under Test	**AUC**	**SE**	**Asymptotic *P***	**Lower Bound**	**Upper Bound**
Expert 1	0.70	0.07	.013	0.58	.82
Expert 2	0.85	0.05	<.001	0.76	.94
DL 1	0.81	0.07	<.001	0.70	.93
DL 2	0.72	0.08	.005	0.59	.86
DL 3	0.83	0.06	<.001	0.73	.93
DL 4	0.79	0.07	<.001	0.68	.91

DL 1, YOLOv3, 25 epochs; DL 2, YOLOv3, 25 epochs, lower threshold; DL 3, YOLOv3, 54 epochs; DL 4, YOLOv3, 60 epochs. CI indicates confidence interval.

the results of the other 2 POCUS experts or the algorithm results before labeling the source images used as the reference standard.

Statistical Analysis

The 4 different training regimens for YOLOv3 and their resultant algorithms and the 2 blinded POCUS reviewers were compared to ground truth labels created by the unblinded POCUS reviewer. Video review was necessary to confirm structure identity when not obvious from still images. We calculated the area under the curve (AUC) using PSPP 1.0.1 statistical software (GNU Project, Boston, MA) to evaluate the performance of all 4 algorithms and 2 POCUS experts for identification of all 4 structures as well as for identification of just blood vessels.

Results

Table 1 shows the AUC results for POCUS experts and YOLOv3 algorithm variants. Figure 1 shows the actual AUC curves. Two variants of the YOLOv3 algorithm performed best, with an AUC of 0.78 for all 4 anatomic structures. The POCUS experts scored AUCs of 0.71 and 0.69, respectively. When broken down by performance in identifying blood vessels, one of the POCUS experts performed the best, with an AUC of 0.85. YOLOv3 with 54 epochs of training was slightly worse, with an AUC of 0.83 (Figure 2 and Table 2).

Discussion

Our study demonstrates that a neural network trained on US images of the UE can readily distinguish between various anatomic structures, including vessels, nerves, bones, and tendons. Peripheral vascular access is becoming increasingly challenging as patients live longer (exhausting conventional sites for intravenous access) and as the rates of obesity and chronic illness have increased steadily in recent decades.[14] In certain complex patient groups, up to 59% are classified as having difficult peripheral vascular access.[15] Difficulty in obtaining vascular access typically stems from lack of visible or palpable veins available for cannulation. When vascular access cannot be readily obtained, delays in medication administration, laboratory testing, and fluid delivery can result. Additionally, multiple attempts at peripheral intravenous access result in patient discomfort and poor patient satisfaction.

When no peripheral vascular access is available, some providers turn to central venous access.[16] Central venous catheter placement is an invasive procedure with numerous potential serious complications.[17] Lack of vascular access is the reason behind as many as 40% of central lines placed in emergency departments.[16] However, one alternative is placement of peripheral venous catheters under US guidance. Initially described almost 20 years ago, it was rapidly introduced to emergency nurses as the vascular access first responders and proved to be highly effective.[16] However, adequate training with regular quality assurance is required, and this is even more challenging in the modern era of nursing shortages and high nursing turnover, resulting in a large pool of nurses not trained in US use.

Identification of key anatomic landmarks is a major challenge faced by novice POCUS users when attempting to use US guidance for peripheral vascular access. Nurses and other novices find it difficult to identify potential blood vessels as well as other structures that can sometimes be confused for vessels but should be avoided, such as tendons, nerves, and sometimes even bones. Deep learning in medical imaging has proved well suited for identification of anatomic structures and even pathologic findings on chest radiography, head and body computed tomography, and magnetic resonance imaging.[18,19] Companies have been slower to deploy US DL applications, and initial products were found only on high-end consultative US equipment, limiting access for POCUS users who are likely to benefit most from DL automation.

Our DL algorithm performed better than 2 fellowship-trained POCUS experts in identifying key anatomic structures on transverse US images of the UE. When only blood vessels were considered, one of the POCUS experts slightly outperformed all 4 YOLOv3 variants, but one algorithm was just behind the expert. This performance by the DL algorithm was somewhat surprising, given that researchers used publicly available US video sources rather than a large database, as can be obtained for DL studies of chest radiography, computed tomography, and magnetic resonance imaging.[20,21] In fact, to our knowledge, no such databases are available, which makes POCUS DL algorithm development more challenging for researchers who do not have access to large US

databases. Our results suggest that even sourcing videos from highly varied public domain open-access sources can deliver good results. Additionally, unlike data from just a single medical center, which is likely to have a limited variety of US machines, we were forced to use data from a broad range of US equipment, which resulted in a more robust and widely applicable algorithm.

Surprisingly, longer training did not necessarily translate into better performance for the DL algorithms. The best overall performance was attained by a YOLOv3 algorithm with 25 epochs of training and one with 54 epochs of training. The most extensively trained algorithm, with 60 epochs of training, came in second behind these two. In the case of blood vessel identification, the 54-epoch–trained algorithm performed the best but was second to one of the POCUS experts. Overall, the DL performance suggests that it can be used in real clinical settings for vessel identification to aid novice providers in finding cannulation targets. This will need to be explored further in prospective randomized studies. However, the YOLOv3 algorithm we developed was able to label structures in real-time video, suggesting that clinical use is not a far stretch.

This study had a number of limitations, including having to source data from a variety of public domain open-access sources with highly varied equipment, image quality, and US machine types. However, this is also one of the potential strengths of our algorithm and increases its robustness for real-world use. We did not test our DL algorithm performance in actual patients or in the hands of real novice providers. However, this work lays the foundation for actual clinical testing and application.

In conclusion, a DL algorithm in this study outperformed 2 POCUS experts in identifying 4 key anatomic structures in transverse US images of the UE. One POCUS expert narrowly beat the algorithm in blood vessel identification. The AUCs for the DL algorithms suggested good performance compared to ground truth anatomic labeling. Future studies should evaluate real-time algorithm performance in actual clinical settings.

References

1. Singh R, Kalra MK, Nitiwarangkul C, et al. Deep learning in chest radiography: detection of findings and presence of change. *PLoS One* 2018; 13:e0204155.
2. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med* 2018; 15:e1002699.
3. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* 2018; 392:2388–2396.
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; 25:44–56.
5. Zhang J, Gajjala S, Agrawal P, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation* 2018; 138:1623–1635.
6. Asch FM, Poilvert N, Abraham T, et al. Automated echocardiographic quantification of left ventricular ejection fraction without volume measurements using a machine learning algorithm mimicking a human expert. *Circ Cardiovasc Imaging* 2019; 12:e009303.
7. US Food and Drug Administration. Class 2 device recall Vscan extend. US Food and Drug Administration website. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRes/res.cfm?ID=173162. Published June 18, 2019.
8. Olgers TJ, Ter Maaten JC. Point-of-care ultrasound curriculum for internal medicine residents: what do you desire? A national survey. *BMC Med Educ* 2020; 20:30.
9. Leschyna M, Hatam E, Britton S, et al. Current state of point-of-care ultrasound usage in Canadian emergency departments. *Cureus* 2019; 11:e4246.
10. Gottlieb M, Sundaram T, Holladay D, Nakitende D. Ultrasound-guided peripheral intravenous line placement: A narrative review of evidence-based best practices. *West J Emerg Med* 2017; 18: 1047–1054.
11. GitHub. Train YOLOv3-SPP from scratch to 61.5 mAP@0.5 at 608. GitHub website. https://github.com/ultralytics/yolov3/issues/310. Published May 31, 2019.
12. Perone CS, Ballester P, Barros RC, Cohen-Adad J. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *Neuroimage* 2019; 194:1–11.
13. Liu S, Wang Y, Yang X, et al. Deep learning in medical ultrasound analysis: a review. *Engineering* 2019; 5:261–275.
14. Rodriguez-Calero MA, Fernandez-Fernandez I, Molero-Ballester LJ, et al. Risk factors for difficult peripheral venous cannulation in hospitalised patients: protocol for a multicentre case-control study in 48 units of eight public hospitals in Spain. *BMJ Open* 2018; 8:e020420.
15. Armenteros-Yeguas V, Gárate-Echenique L, Tomás-López MA, et al. Prevalence of difficult venous access and associated risk factors in highly complex hospitalised patients. *J Clin Nurs* 2017; 26: 4267–4275.
16. Shokoohi H, Boniface K, McCarthy M, et al. Ultrasound-guided peripheral intravenous access program is associated with a marked reduction in central venous catheter use in noncritically ill emergency department patients. *Ann Emerg Med* 2013; 61:198–203.

17. Akmal A, Hasan M, Mariam A. The incidence of complications of central venous catheters at an intensive care unit. *Ann Thorac Med* 2007; 2:61–63.
18. Cha MJ, Chung MJ, Lee JH, Lee KS. Performance of deep learning model in detecting operable lung cancer with chest radiographs. *J Thorac Imaging* 2019; 34:86–91.
19. Abiyev RH, Ma'aitah MKS. Deep convolutional neural networks for chest diseases detection. *J Healthc Eng* 2018; 2018:4168538.
20. Kim TK, Yi PH, Wei J, et al. Deep learning method for automated classification of anteroposterior and posteroanterior chest radiographs. *J Digit Imaging* 2019; 32: 925–930.
21. Yuh EL, Gean AD, Manley GT, Callen AL, Wintermark M. Computer-aided assessment of head computed tomography (CT) studies in patients with suspected traumatic brain injury. *J Neurotrauma* 2008; 25:1163–1172.