Development of a Deep Learning Network to Classify Inferior Vena Cava Collapse to Predict Fluid Responsiveness

Michael Blaivas, MD, MBA, Laura Blaivas, BS, Gary Philips, MS, Roland Merchant, MD, ScD, Mitchell Levy, MD, Adeel Abbasi, MD, Carsten Eickhoff, PhD, Nathan Shapiro, MD, Keith Corl, MD, ScM

Received August 26, 2020, from the Department of Medicine, 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.); Michigan State University, East Lansing, Michigan, USA (L.B.); Center for Biostatistics, Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA (G.P.); Department of Emergency Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA (R.M.); Department of Medicine, Division of Pulmonary Critical Care and Sleep, Warren Alert Medical School of Brown University, Providence, Rhode Island, USA (M.L., A.A., K.C.); Brown Center for Biomedical Informatics, Brown University, Providence, Rhode Island, USA (C.E.); and Department of Emergency Medicine, Beth Israel Deaconess Hospital, Harvard Medical School, Boston, Massachusetts, USA (N.S.). Manuscript accepted for publication September 14, 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

AUROC, area under the receiver operating characteristic curve; CI, confidence interval; cIVC, collapsibility of the inferior vena cava; DL, deep learning; ICU, intensive care unit; IV, intravenous; IVC, inferior vena cava; LSTM, long short-term memory; POCUS, point-of-care ultrasound; US, ultrasound

doi:10.1002/jum.15527

Objectives—To create a deep learning algorithm capable of video classification, using a long short-term memory (LSTM) network, to analyze collapsibility of the inferior vena cava (IVC) to predict fluid responsiveness in critically ill patients.

Methods—We used a data set of IVC ultrasound (US) videos to train the LSTM network. The data set was created from IVC US videos of spontaneously breathing critically ill patients undergoing intravenous fluid resuscitation as part of 2 prior prospective studies. We randomly selected 90% of the IVC videos to train the LSTM network and 10% of the videos to test the LSTM network's ability to predict fluid responsiveness. Fluid responsiveness was defined as a greater than 10% increase in the cardiac index after a 500-mL fluid bolus, as measured by bioreactance.

Results—We analyzed 211 videos from 175 critically ill patients: 191 to train the LSTM network and 20 to test it. Using standard data augmentation techniques, we increased our sample size from 191 to 3820 videos. Of the 175 patients, 91 (52%) were fluid responders. The LSTM network was able to predict fluid responsiveness moderately well, with an area under the receiver operating characteristic curve of 0.70 (95% confidence interval [CI], 0.43–1.00), a positive likelihood ratio of infinity, and a negative likelihood ratio of 0.3 (95% CI, 0.12–0.77). In comparison, point-of-care US experts using video review offline and manual diameter measurement via software caliper tools achieved an area under the receiver operating characteristic curve of 0.94 (95% CI, 0.83–0.99).

Conclusions—We demonstrated that an LSTM network can be trained by using videos of IVC US to classify IVC collapse to predict fluid responsiveness. Our LSTM network performed moderately well given the small training cohort but worse than point-of-care US experts. Further training and testing of the LSTM network with a larger data sets is warranted.

Key Words—artificial intelligence; critical care; deep learning; emergency medicine; fluid responsiveness; inferior vena cava; long short-term memory; point-of-care ultrasound

For more than half a century, physicians caring for critically ill patients have sought to tailor intravenous (IV) fluid resuscitation to patients' physiologic states.[1] Routine use of invasive hemodynamic monitoring by pulmonary artery catheterization fell from favor in the early 2000s when high-

quality trial data demonstrated increased patient risk without improved outcomes.[2] Over the last 20 years, physicians have tested a variety of noninvasive technologies to guide IV fluid resuscitation, yet no single modality has attained widespread clinical acceptance.[3]

Although the risks of under-resuscitation have long been recognized, over the last decade, numerous studies have illustrated the harms of over-resuscitation.[4–6] After the decline of invasive hemodynamic monitoring, the evidence for using noninvasive technology to tailor IV fluid resuscitation to individual patient needs was largely limited to clinical anecdotal and observational data.[7] In 2019, Pontet et al[8] were the first to show in a randomized trial that ultrasound (US)-guided IV fluid resuscitation in the intensive care unit (ICU) reduced the patient fluid balance and decreased the time receiving mechanical ventilation. The subsequent Fluid Response Evaluation in Sepsis Hypotension and Shock trial demonstrated that a noninvasive fluid resuscitation protocol that used bioreactance to measure patient hemodynamics to direct IV fluid resuscitation reduced the amount of IV fluids given to patients with septic shock.[9] That trial observed decreased rates of acute kidney injury and respiratory failure associated with excess IV fluids.

It is in this backdrop that physicians have proposed the use of point-of-care ultrasound (POCUS) to guide IV fluid resuscitation.[10] Point-of-care US–guided IV fluid resuscitation offers the advantage of the near omnipresence of US machines in clinical environments and physician familiarity, which other noninvasive technologies such as bioreactance lack.[11] Although emergency and critical care medicine societies advise that POCUS proficiency requires sufficient residency or fellowship training that includes quality assessment teaching sessions and at least 150 US scans covering a range of organs, clinically, POCUS is being widely adopted to guide clinical decisions by a variety of clinicians with relatively limited US skills and training.[12,13]

Measurement of the collapsibility of the inferior vena cava (cIVC) predicts fluid responsiveness reasonably well when performed by expert sonologists in spontaneously breathing patients (area under the receiver operating characteristic curve [AUROC], 0.82).[14] However, work by our group showed that the test characteristics were not as favorable when POCUS examinations were performed by novice sonologists (AUROC, 0.69).[15] In addition, poor-to-moderate interrater agreement in cIVC measurement among nonexpert sonologists has been reported by other authors.[16–18] Training clinicians to a level at which they obtain the requisite skills necessary to accurately use POCUS-measured cIVC to guide IV fluid resuscitation might be the most substantial hurdle this approach faces before it can be widely adopted. Artificial intelligence using deep learning (DL) algorithms has been proposed as a solution to overcome skill gaps to make the widespread use of POCUS-directed IV fluid resuscitation more feasible, reliable, and useful.[19]

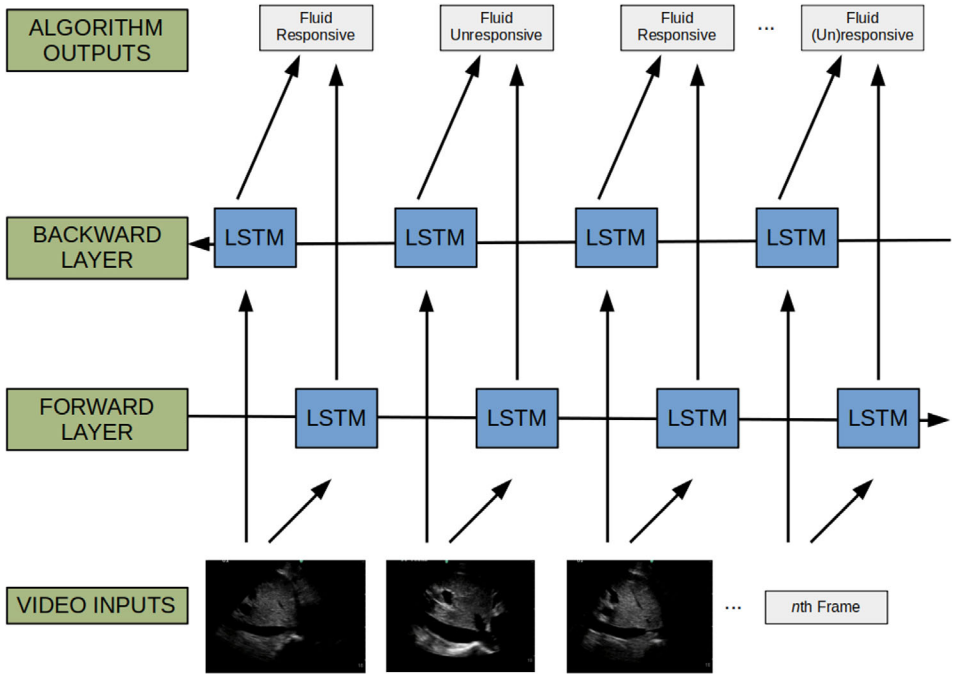
The primary aim of our study was to create a DL algorithm using videos of cIVC measurements to predict fluid responsiveness among spontaneously breathing critically ill patients. We then tested the performance of our DL algorithm in predicting fluid responsiveness compared to bioreactance-determined fluid responsiveness. In addition, we further compared the performance of the DL algorithm against POCUS experts who measured cIVC during offline video review and measurement to predict fluid responsiveness.

Materials and Methods

Study Design

We performed a secondary analysis of cIVC POCUS measurements recorded on video files from 2 previous prospective studies to create a DL algorithm.[15,20] We sought to determine whether the DL algorithm was able to predict fluid responsiveness via a VGG-16 bidirectional long short-term memory (LSTM) network. Details of the participant enrollment criteria have been previously described.[15,20] Briefly, patients were eligible for participation if they were spontaneously breathing, admitted to a medical ICU, and had signs of acute circulatory failure. A POCUS examination was performed before the IV fluid bolus. Fluid responsiveness was defined as a 10% increase in the cardiac index after a 500-mL bolus of normal saline detected by noninvasive cardiac output monitoring (Cheetah Medical, Tel Aviv, Israel). This approach is regarded as standard and safe when determining fluid

Figure 1. Long short-term memory network concept diagram (courtesy of Kira Blaivas).



responsiveness in critically ill patients.[14,15] Written informed consent for the original video data was obtained from each study participant or his or her surrogate. The Institutional Review Board waived repeated informed consent for the secondary analysis. All patient identifiers were removed from the study videos before analysis.

Study Data and Manipulation

A total of 211 videos of POCUS-measured cIVC from 175 critically ill patients were extracted from 210 Digital Imaging and Communications in Medicine files. Video data types were AVI and MP4. Based on the results of the bioreactance measured for each patient during the course of the study, participants were determined to be "fluid responders" or "nonresponders," and their POCUS videos were categorized accordingly. We randomly selected 191 videos for DL training (99 fluid responders and 92 nonresponders) and 20 videos for DL testing (9 fluid responders and 11 nonresponders). We used data augmentation, as described below, to amplify the available number of videos for the DL algorithm training. Data augmentation is a common practice in DL algorithm training that leads to improved training results.[21] In addition to each video being included in its original format, we used FFmpeg open-source software (https://ffmpeg.org/) to perform vertical and horizontal flips, mirroring, and 45° and 90° clockwise and anticlockwise rotations on each video. FFmpeg was used to complete a single rotation or flip manipulation on each of the training videos before initiating the next script to accomplish another manipulation on the original videos. After augmentation, a total of 3820 videos (1980 fluid responder and 1840 nonresponder videos) were available for DL training.

Algorithm Design

We used Python programming language version 3.72 (Python Software Foundation, Wilmington, DE) with the Anaconda (Austin, TX) package manager to manage packages and facilitate scripting and a publicly available Keras-based (a Python DL library or framework) VGG-16 bidirectional LSTM DL algorithm. The VGG-16 model is available from various public sources, including github.com (an online scripting repository). VGG-16 is an early convolutional neural network using 16 layers and has been shown to be

Table 1. Participant Characteristics Among the Training Set

Characteristic	Nonresponders by NICOM (n = 91)[a]	Fluid Responders by NICOM (n = 84)[a]	*P*[b]
Demographic and clinical characteristics			
Age, y	52 (35–71)	57.5 (44.0–69.5)	.38
Male	50 (55.0)	41 (48.8)	.45
BMI, kg/m^2	24.5 (21.4–29.0)	25.1 (22.4–29.7)	.54
APACHE II score	16 (12–22)	15 (11–24)	.89
Fluid and other resuscitation			
IV fluid before US, mL	4000 (2050–5000)	3550 (3000–4050)	.34
IV fluid administered during study, mL	500 (500–500)	500 (500–500)	.65
Duration of fluid bolus, min	8 (6–11)	7 (6–10)	.76
Required vasopressor	20 (22.0)	12 (14.3)	.24
Medical history			
Hypertension	49 (53.8)	44 (52.4)	.88
Diabetes mellitus	49 (53.8)	44 (52.4)	.88
Cardiomyopathy	27 (29.7)	18 (21.4)	.23
Alcohol abuse	11 (12.1)	18 (21.4)	.14
COPD	4 (4.4)	19 (22.6)	<.001
Pulmonary embolism	0 (0.0)	2 (2.4)	.23
Pulmonary hypertension	10 (11.0)	1 (1.2)	.01
Hospital discharge diagnosis			
Severe sepsis/septic shock	34 (37.4)	35 (41.7)	.64
DKA/HHS	36 (39.6)	26 (30.9)	.27
Gastrointestinal hemorrhage	11 (12.1)	7 (8.3)	.46
Alcohol withdrawal	1 (1.1)	2 (2.4)	.61
Respiratory failure secondary to pneumonia	3 (3.3)	4 (4.76)	.71
Outcomes			
ICU length of stay, d	2 (2–4)	2 (2–3)	.22
Hospital length of stay, d	5 (3–7)	5 (3–7)	.89
Alive at discharge	84 (92.3)	79 (94.1)	.77

Data are presented as median (interquartile range) and number (percent) where applicable. APACHE indicates Acute Physiology and Chronic Health Evaluation; BMI, body mass index; COPD, chronic obstructive pulmonary disease; DKA, diabetic ketoacidosis; HHS, hyperosmolar hyperglycemic syndrome; NICOM, noninvasive cardiac output monitoring.

[a]Note that some participants contributed multiple videos, so the total number of participants was not equal to the total number of videos for the training set.

[b]*P* values were based on the Wilcoxon rank sum test for continuous variables and the Fisher exact test for categorical variables.

superior for US DL applications.[22] By analyzing individual frames, LSTM can track temporal changes and relationships in real-time video to create an algorithm for US video analysis. Long short-term memory has additional layers in the network algorithm architecture, which track temporal changes in images (changes in the US image in a cine loop such as contraction of the heart). In contrast to standard LSTM unidirectional networks, the bidirectional aspect allows the flow of temporally related information in both forward and reverse directions through the algorithm. This feature makes the architecture more sensitive and specific at detecting changes in the image from frame to frame, thereby identifying changes in action. Additionally, it improves the network's understanding of the context of the motion (Figure 1). The VGG-16 bidirectional LSTM used weights trained on the UCF-101 action recognition data set (University of Central Florida, Orlando, FL). Weights are learnable parameters in neural networks responsible for their ability to interpret images.

We trained our bidirectional LSTM algorithm by manipulating optimizers, learning rates, and batch size (which are all adjustable settings that affect a convolutional neural network's training performance) during training for optimal training times and accuracies, while avoiding exploding gradients (dramatic changes in learning parameters of the convolutional

Table 2. Inferior Vena Cava Measurements and Noninvasive Cardiac Output Monitoring Results Before and After an IV Fluid Bolus

Parameter	Nonresponders (n = 91)	Fluid Responders (n = 84)	*P*[a]
Hemodynamic measurements			
Mean arterial pressure, mm Hg	60.9 (52.9–76.0)	63.5 (54.8–85.0)	.27
Heart rate, beats/min	109 (83–124)	112 (98.5–126.5)	.14
Baseline stroke volume, mL	76.8 (58.8–93.6)	66.3 (49.6–78.8)	.01
Post–IV fluid bolus stroke volume, mL	73.0 (53.0–88.5)	86.6 (64.8–96.1)	.01
Change in baseline stroke volume, mL	−1.4 (−6.4–3.5)	14.8 (9.2–21.7)	<.001
Baseline cardiac index, L/min/m^2	3.4 (2.9–4.0)	3.3 (2.6–3.7)	.10
Post–IV fluid bolus cardiac index, L/min/m^2	3.3 (2.6–3.7)	4.1 (3.3–4.5)	<.001
Change in cardiac index, L/min/m^2	−0.1 (−0.4–0.1)	0.7 (0.5–1.0)	<.001
IVC US measurement cIVC, %	13.3 (7.9–23.4)	37.2 (29.1–51.9)	<.001

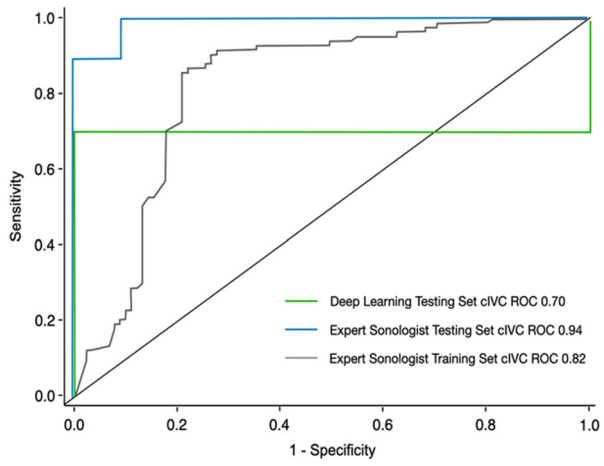
Data are presented as median (interquartile range)
[a]*P* values were based on the Wilcoxon rank sum test.

neural network during training), which can result in training failure. The number of epochs, defined as a single round of training through all of the data, was adjusted for optimal results while avoiding overfitting. The best performance was obtained with 60 epochs using a stochastic gradient descent optimizer with a learning rate of 0.001. A batch size of 100 videos ultimately produced the best algorithm training performance.

Algorithm Validation and Testing

The bidirectional LSTM algorithm was coded to automatically perform cross-validation with each epoch. Cross-validation accuracy, learning, and training losses were used to guide algorithm training adjustments. After results were optimized and no further adjustments improved performance, the algorithm was tested on the randomly selected 10% testing data set. We executed a testing script using our newly generated training weights on the 20 separate inferior vena cava (IVC) US examination videos through the VGG-16 bidirectional LSTM to predict fluid responsiveness. This step tested our algorithm's performance on videos that were not previously used to train our DL algorithm. We then compared these results to 2 POCUS experts who measured cIVC with calipers during review of the same videos. The expert physician sonologists were attending ICU physicians who had performed and reviewed more than 250 cIVC US measurements.

Figure 2. Area under the receiver operating characteristic curve for cIVC to detect fluid responsiveness for expert sonologists in the training set and expert sonologists and the DL algorithm in the testing set.



Statistical Analyses

Participant demographics, clinical characteristics, and clinical outcomes were summarized by descriptive statistics. Fisher exact and Wilcoxon rank sum tests were used to compare fluid responders versus nonresponders for dichotomous and continuous participant characteristics and study outcomes, respectively. Test performance characteristics, including the AUROC, sensitivity, specificity, negative and positive predictive values, and likelihood ratios, for distinguishing fluid responsiveness were calculated. The Wilson method was used to calculate 95% confidence intervals (CIs) for the performance characteristics. We used Stata 14.2 software (StataCorp, College Station, TX) to complete the above analyses.

Table 3. Test Performance Characteristics for Changes in cIVC to Determine Fluid Responsiveness

Method	cIVC, %	Fluid Responders	Nonresponders	Sensitivity, %	Specificity, %	PPV, %	NPV, %	+LR	−LR	Accuracy, %
DL	≥25	7	0	70 (34.8–93.3)	100 (69.2–100)	100 (69.2–100)	76.9 (49.7–91.8)	a	0.3 (0.12–0.77)	85 (62.1–96.8)
	<25	3	10							
Expert sonologist	≥25	8	0	88.9 (56.5–98.0)	100.0 (74.1–100.0)	100.0 (67.6–100.0)	91.7 (64.6–98.5)	a	0.11 (0.06–0.19)	95.0 (76.4–99.1)
	<25	1	11							

Values in parentheses are 95% CIs. LR indicates likelihood ratio; NPV, negative predictive value; and PPV, positive predictive value.
[a]Positive likelihood ratio could not be calculated because specificity was 100.0%.

Results

Study Participants

Table 1 provides participant demographics and clinical characteristics. Ninety-one (52%) participants were fluid responders. There were no differences between fluid responders and nonresponders at baseline, except fluid responders were more likely to have a history of chronic obstructive pulmonary disease and less likely to have a history of pulmonary hypertension compared to nonresponders. Table 2 displays the hemodynamic characteristics of fluid responders and nonresponders. Fluid responders had a median cIVC of 37.2% versus a cIVC of 13.3% for nonresponders.

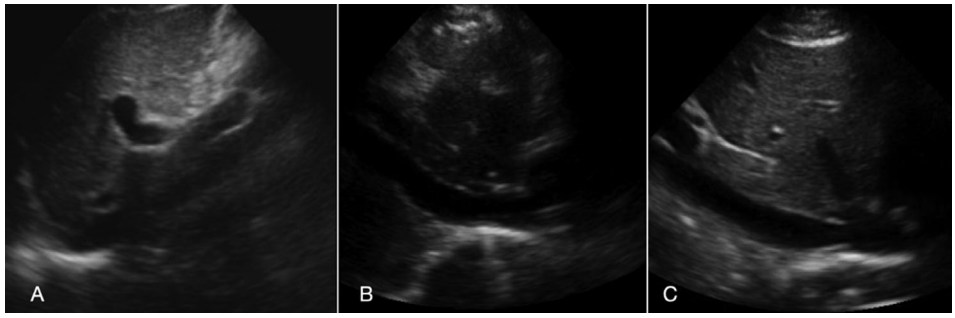
Outcomes

The best performance of the LSTM network was attained with training for 60 epochs (a range of 10–200 were tested during training). The network took 19 minutes 29 seconds to complete training and 10 seconds to review the test videos. Using the 20 test videos, the LSTM network was able to predict fluid responsiveness with an AUROC of 0.70 (95% CI, 0.43–1.00), correctly classifying 17 of 20 videos (Figure 2 and Table 3). All 3 of the misclassifications were in patients who were fluid responsive; the cIVC was measured to be greater than 25% (37%, 38%, and 66%, respectively) by the expert sonologists but was classified as nonresponsive by the LSTM network (Figure 3). Expert sonologists who reviewed the same POCUS videos were able to predict fluid responsiveness using a cIVC cutoff of 25%, with an AUROC of 0.94 (95% CI, 0.83–0.99; Table 3). Test characteristics for all participant videos used to train the DL algorithm as read by experts are displayed in Table 4. In the training sample, expert sonologists were able to predict fluid responsiveness, with an AUROC of 0.82 (95% CI, 0.76–0.89).

Discussion

In this pilot study, we built a DL algorithm using an LSTM network to classify cIVC from POCUS videos. Our study demonstrates that an LSTM network trained on single-view short sagittal proximal IVC videos can achieve moderately good initial test

Figure 3. Long short-term memory misclassifications: 3 participant videos of cIVC (**A–C**) that were fluid responsive by bioreactance but misclassified by the LSTM network.



characteristics (AUROC, 0.70) for predicting fluid responsiveness among spontaneously breathing critically ill patients.

Clinicians have widely adopted the use of cIVC measured by POCUS[23] to direct fluid resuscitation despite evidence that image acquisition and interpretation by novice sonologists often limit its accuracy.[24] Artificial intelligence has the potential to eliminate errors in both image acquisition and interpretation through automation, allowing novice sonologists to overcome their own skill limitations. To date, only a small number groups have studied automation or DL to facilitate POCUS cIVC measurement. None have compared it to an objective measure of fluid responsiveness. However, there is commercial interest in automating cIVC evaluations among US machine manufacturers.

In 2015, GE Healthcare (Waukesha, WI) released the Auto-IVC tool, which relies on edge technology to identify the IVC vessel walls and produce a real-time cIVC calculation. Although the Auto-IVC tool reportedly does not use DL, the automated diameter measurement eliminates the burden of manually calculating cIVC. Clinicians using this commercial feature must still obtain sufficient POCUS images of the IVC, so the major benefit of Auto-IVC is its potential to reduce measurement and interpretation errors made by novice sonologists. Internal company data are reported to show a good correlation with expert POCUS measurement of IVC diameters; however, publicly reported data supporting Auto-IVC's feasibility and accuracy has yet to be released. Similar work by another group has reported an initial proof-of-concept study on automation of cIVC measurement and interpretation.[25] That group's algorithm requires researchers to select points in the IVC vessel to allow for frame-by-frame processing that tracks IVC movement in the B-mode, making the technology user dependent and thereby limiting widespread use.

In an effort to automate IVC image acquisition, Chen et al[26] used a porcine model to build a DL algorithm capable of identifying and measuring the IVC diameter. The algorithm used 48 data sets of IVC images, including both static and dynamic US videos. It could successfully localize the IVC in 98% of the cases and produced most IVC diameter measurements that were within 15% of an expert sonologist's read. However, the stepwise process required considerable cine loop preprocessing, including use of various filters and image size adjustments. Color Doppler US had to be used to identify areas of interest based on the presence of blood flow. Although encouraging, the numerous steps in their approach suggest that automated IVC image acquisition is in the early stages of development, and it may be several years before any commercial prototypes are available.

In some of the most promising work to date, Belmont et al[27] used Kanade-Lucas-Tomasi feature tracking and pyramidal segmentation to build an artificial intelligence algorithm capable of measuring cIVC. Their complex algorithm analyzed 57 cine loops from 47 spontaneously breathing hemodialysis patients and demonstrated good agreement with manual measurements by POCUS experts (>95% of the artificial intelligence measurements differed by

Table 4. Test Performance Characteristics for Changes in cIVC to Determine Fluid Responsiveness Read by Expert Sonologists Among the Training Set

cIVC, %	Fluid Responders (n = 84)	Nonresponders (n = 91)	Sensitivity, %	Specificity, %	PPV, %	NPV, %	+LR	–LR	Accuracy, %
≥25	71	19	84.5 (75.3–90.7)	79.1 (69.7–86.2)	78.9 (69.4–86.0)	84.7 (75.6–90.8)	4.05 (3.88–4.23)	0.20 (0.19–0.21)	81.7 (75.3–86.7)
<25	13	72							

Values in parentheses are 95% CIs.

<10% from the expert reads). However, the algorithm's clinical application is limited because it requires a clear and consistent IVC vessel edge to facilitate tracking, which may be difficult to maintain in the clinical setting.

Deep learning is a branch of artificial intelligence that does not rely on prespecified parameters or rules to interpret data. Instead, it studies patterns within the data and draws unstructured associations.[28] We sought to use the ability of DL algorithms to find novel associations and to be able to make predictions where no actual anatomic localization, border identification, or diameter measurements are required for prediction of the outcome. Furthermore, rather than requiring operators to obtain individual IVC images or to break US cine loops into single images to analyze at a later date, we explored the potential of a real-time application, which could analyze a cine loop at the point of care. Future application of DL analysis might be capable of running in real time while a novice sonologist scans the patient's IVC. By embedding our DL algorithm into a US system or simply by having the application run on a real-time video feed from the system, a novice would only have to obtain a sagittal view of the proximal IVC. The DL algorithm would do the rest, providing a nearly instant prediction of fluid responsiveness. Introduction of such DL technology into the hands of novice sonologists (residents, attending physicians, nurses, advanced practice providers, and emergency medical technicians) could greatly increase health care workers' ability to individually tailor IV fluid resuscitation.

Our study had limitations. First our training sample size (191 initial videos and 3820 augmented) was relatively small to adequately train a DL algorithm, which may lead to an overestimation or underestimation of its performance. We that estimate more than 1000 videos before augmentation are needed for robust training based on the experience of authors working on similar networks but with access to massive US databases.[29] Second, our testing sample (20 patient videos) provided only an initial point estimate of the algorithm's ability to predict fluid responsiveness. Our DL algorithm test performance, with an AUROC of 0.70 (95% CI, 0.43–1.00), was similar to test characteristics we reported in a prior study for novice sonologists using cIVC to predict fluid responsiveness (AUROC,

0.69) but worse than expert sonologists (AUROC, 0.82).[15] The range of the DL CIs reflects the small testing sample. Given our current data, it remains to be seen whether further training will improve DL test characteristics to match or exceed the performance of expert sonologists or whether they will more closely resemble novice test characteristics. This may suggest that the LSTM network has difficulty interpreting images that are intermittently indistinct. Third, POCUS videos were obtained by several ICU fellows and an emergency medicine resident. Prospective testing of our algorithm is needed in a variety of clinical settings and by health care workers with a range of US expertise.

In conclusion, we have demonstrated that it is possible to construct an LSTM network capable of reading videos of respirophasic variation of the IVC in critically ill patients to predict fluid responsiveness. Our LSTM network performed moderately well with a small test sample but worse than expert sonologists. Further training and testing of the network are needed before clinical implementation.

References

1. Thakkar AB, Desai SP. Swan, Ganz, and their catheter: its evolution over the past half century. *Ann Intern Med* 2018; 169: 636–642.
2. Harvey S, Harrison DA, Singer M, et al. Assessment of the clinical effectiveness of pulmonary artery catheters in management of patients in intensive care (PAC-Man): a randomised controlled trial. *Lancet* 2005; 366:472–477.
3. Huygh J, Peeters Y, Bernards J, Malbrain MLNG. Hemodynamic monitoring in the critically ill: an overview of current cardiac output monitoring methods. *F1000Res* 2016; 5:F1000.
4. Boyd JH, Forbes J, Nakada TA, Walley KR, Russell JA. Fluid resuscitation in septic shock: a positive fluid balance and elevated central venous pressure are associated with increased mortality. *Crit Care Med* 2011; 39:259–265.
5. Marik PE, Linde-Zwirble WT, Bittner EA, Sahatjian J, Hansell D. Fluid administration in severe sepsis and septic shock, patterns and outcomes: an analysis of a large national database. *Intensive Care Med* 2017; 43:625–632.
6. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med* 2018; 24:1716–1720.
7. Latham HE, Bengtson CD, Satterwhite L, et al. Stroke volume guided resuscitation in severe sepsis and septic shock improves outcomes. *J Crit Care* 2017; 42:42–46.
8. Pontet J, Yic C, Díaz-Gómez JL, et al. Impact of an ultrasound-driven diagnostic protocol at early intensive-care stay: a randomized-controlled trial. *Ultrasound J* 2019; 11:24.
9. Douglas IS, Alapat PM, Corl KA, et al. Fluid Response Evaluation in Sepsis Hypotension and Shock: a randomized clinical trial. *Chest* 2020; S0012-3692(20):30768–6. https://doi.org/10.1016/j.chest.2020.04.025.
10. Evans D, Ferraioli G, Snellings J, Levitov A. Volume responsiveness in critically ill patients: use of sonography to guide management. *J Ultrasound Med* 2014; 33:3–7.
11. Kendall JL, Hoffenberg SR, Smith RS. History of emergency and critical care ultrasound: the evolution of a new imaging paradigm. *Crit Care Med* 2007; 35:S126–S130.
12. American College of Emergency Physicians. Ultrasound guidelines: emergency, point-of-care and clinical ultrasound guidelines in medicine. *Ann Emerg Med* 2017; 69:e27–e54.
13. Monti J. Revolution or evolution? A proposal for the integration of point-of-care ultrasound into physician assistant clinical practice. *J Physician Assist Educ* 2017; 28:27–32.
14. Orso D, Paoli I, Piani T, Cilenti FL, Cristiani L, Guglielmo N. Accuracy of ultrasonographic measurements of inferior vena cava to determine fluid responsiveness: a systematic review and meta-analysis. *J Intensive Care Med* 2020; 35:354–363.
15. Corl KA, Azab N, Nayeemuddin M, et al. Performance of a 25% inferior vena cava collapsibility in detecting fluid responsiveness when assessed by novice versus expert physician sonologists [published online ahead of print October 14, 2019]. *J Intensive Care Med*. https://doi.org/10.1177/0885066619881123.
16. Bussmann BM, Sharma S, McGregor D, Hulme W, Harris T. Observational study in healthy volunteers to define inter-observer reliability of ultrasound haemodynamic monitoring techniques performed by trainee doctors. *Eur J Emerg Med* 2019; 26:217–223.
17. Bowra J, Uwagboe V, Goudie A, Reid C, Gillett M. Interrater agreement between expert and novice in measuring inferior vena cava diameter and collapsibility index. *Emerg Med Australas* 2015; 27:295–299.
18. Akkaya A, Yesilaras M, Aksay E, Sever M, Atilla OD. The interrater reliability of ultrasound imaging of the inferior vena cava performed by emergency residents. *Am J Emerg Med* 2013; 31: 1509–1511.
19. 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; 38:1887–1897.

20. Corl KA, George NR, Romanoff J, et al. Inferior vena cava collapsibility detects fluid responsiveness among spontaneously breathing critically-ill patients. *J Crit Care* 2017; 41:130–137.
21. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. *ArXiv* 2017; 1712.04621.
22. 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: 1187–1194.
23. Sawe HR, Haeffele C, Mfinanga JA, Mwafongo VG, Reynolds TA. Predicting fluid responsiveness using bedside ultrasound measurements of the inferior vena cava and physician gestalt in the emergency department of an urban public hospital in sub-Saharan Africa. *PLoS One* 2016; 11:e0162772.
24. Lucas BP, D'Addio A, Clark J, et al. Reproducibility of point-of-care ultrasonography for central vein diameter measurement: separating image acquisition from interpretation. *J Clin Ultrasound* 2017; 45:488–496.
25. Mesin L, Pasquero P, Albani S, Porta M, Roatta S. Semi-automated tracking and continuous monitoring of inferior vena cava diameter in simulated and experimental ultrasound imaging. *Ultrasound Med Biol* 2015; 41:845–857.
26. Chen J, Li J, Ding X, Chang C, Wang X, Ta D. Automated identification and localization of the inferior vena cava using ultrasound: an animal study. *Ultrason Imaging* 2018; 40:232–244.
27. Belmont B, Kessler R, Theyyunni N, et al. Continuous inferior vena cava diameter tracking through an iterative Kanade-Lucas-Tomasi–based algorithm. *Ultrasound Med Biol* 2018; 44:2793–2801.
28. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2016.
29. Ghorbani A, Ouyang D, Abid A, et al. Deep learning interpretation of echocardiograms. *NPJ Digit Med* 2020; 3:10