Automated Lung Ultrasound B-Line Assessment Using a Deep Learning Algorithm

Cristiana Baloescu, Grzegorz Toporek, Seungsoo Kim, *Member, IEEE*, Katelyn McNamara, Rachel Liu, Melissa M. Shaw, Robert L. McNamara, Balasundar I. Raju, *Member, IEEE*, and Christopher L. Moore

***Abstract*— Shortness of breath is a major reason that patients present to the emergency department (ED) and point-of-care ultrasound (POCUS) has been shown to aid in diagnosis, particularly through evaluation for artifacts known as B-lines. B-line identification and quantification can be a challenging skill for novice ultrasound users, and experienced users could benefit from a more objective measure of quantification. We sought to develop and test a deep learning (DL) algorithm to quantify the assessment of B-lines in lung ultrasound. We utilized ultrasound clips ($n = 400$) from an existing database of ED patients to provide training and test sets to develop and test the DL algorithm based on deep convolutional neural networks. Interpretations of the images by algorithm were compared to expert human interpretations on binary and severity (a scale of 0–4) classifications. Our model yielded a sensitivity of 93% (95% confidence interval (CI) 81%–98%) and a specificity of 96% (95% CI 84%–99%) for the presence or absence of B-lines compared to expert read, with a kappa of 0.88 (95% CI 0.79–0.97). Model to expert agreement for severity classification yielded a weighted kappa of 0.65 (95% CI 0.56–074). Overall, the DL algorithm performed well and could be integrated into an ultrasound system in order to help diagnose and track B-line severity. The algorithm is better at distinguishing the presence from the absence of B-lines but can also be successfully used to distinguish between B-line severity. Such methods could decrease variability and provide a standardized method for improved diagnosis and outcome.**

***Index Terms*— Medical imaging, medical signal and image processing, medical ultrasonics, signal and image processing.**

I. Introduction

SHORTNESS of breath is among the top ten reasons patients visit the emergency department (ED), accounting for over 3.5 million ED visits in the U.S. annually [1]. There are diverse causes of dyspnea and point-of-care ultrasound (POCUS) has been shown to aid in establishing a diagnosis [2]–[4]. Alveolar interstitial syndrome (AIS) is a broad sonographic term indicating the presence of fluid in the alveolar and interstitial spaces of the lung parenchyma and is based on the presence of artifacts seen on the ultrasound image that extend from the pleural line to the bottom of the screen known as "B-lines" [5], [6]. Recent literature also describes B-lines in patients with respiratory symptoms of coronavirus disease 2019 (COVID-19), with progression of the B-line pattern as the disease advances.

While sometimes referred to as "comet tails," B-lines are technically a ring down artifact that form due to resonance of the air–fluid interface in the interstitial space [10]. B-lines appear as hyperechoic lines extending from the pleural surface to the bottom of the screen along the direction of the ultrasound beam. B-lines are dynamic and vary in location and quantity on the images obtained from frame to frame depending on the movement of tissue or ultrasound probe. Still images may miss their presence and are inadequate to judge the overall severity.

B-lines representing AIS may appear in several different pathologic conditions such as pulmonary edema in acute heart failure (HF) or volume overload, noncardiogenic pulmonary edema, pneumonia, pulmonary embolus, and acute respiratory distress syndrome (ARDS), including pneumonitis from COVID-19 [6], [11]. Establishing the presence or absence of B-lines aids in diagnosis, while a quantitative assessment of B-lines can help classify disease severity and prognosis [12]–[16]. For instance, a higher burden of B-lines on lung ultrasound at hospital discharge or in the ambulatory HF population identified patients at high risk for readmission or death [12], [14], [15], [17]. In patients hospitalized for dyspnea or chest pain, B-lines were found to be better predictors of all-cause mortality and complications such as myocardial

Manuscript received May 8, 2020; accepted June 8, 2020. Date of publication June 15, 2020; date of current version October 26, 2020. This work was supported in part by a grant from Philips Research North America. *(Corresponding author: Cristiana Baloescu.)*

Cristiana Baloescu, Rachel Liu, Melissa M. Shaw, and Christopher L. Moore are with the Department of Emergency Medicine, School of Medicine, Yale University, New Haven, CT 06511 USA (e-mail: cristiana.baloescu@yale.edu; rachel.liu@yale.edu; melissa.m.shaw@yale.edu; chris.moore@yale.edu).

Grzegorz Toporek and Balasundar I. Raju are with Philips Research North America, Cambridge, MA 02141 USA (e-mail: grzegorz.toporek@philips.com; balasundar.raju@philips.com).

Seungsoo Kim was with Philips Research North America, Cambridge, MA 02141 USA. He is now with Infraredx, Inc., Bedford, MA 01730 USA (e-mail: kim.seungsoo@gmail.com).

Katelyn McNamara was with the Department of Emergency Medicine, School of Medicine, Yale University, New Haven, CT 06511 USA. She is now with the Department of Internal Medicine, Yale University, New Haven, CT 06510 USA (e-mail: katiemac04@gmail.com).

Robert L. McNamara is with the Department of Cardiology, School of Medicine, Yale University, New Haven, CT 06511 USA.

Digital Object Identifier 10.1109/TUFFC.2020.3002249

infarction than recognized predictors such as left ventricular ejection fraction or end-stage renal disease [13], [16]. A pulmonary ultrasound scoring system based on B-line quantification in intensive care unit (ICU) patients was found to be predictive of mortality, length of stay, and time spent on the ventilator [18]. Severity rating for B-lines could also potentially be used to track the changes in B-line profile over time as a marker of disease severity and to evaluate the response to treatments such as intravenous fluids and medications.

Despite the potential for B-line identification to improve diagnosis and prognosis, challenges to ultrasound imaging include interoperator and intraoperator variability and image quality control. B-line identification and quantification can be a challenging skill for novice ultrasound users, while experienced users could benefit from a more objective measure of quantification.

Such challenges can be addressed by automated detection and quantification algorithms. Operator dependence in image acquisition and interpretation is one major reason for the need for automated detection. In addition to minimizing operator error, automated detection affords the possibility of rapid processing of a vast amount of data for research. A robust automated system might even be able to be used by patients themselves to self-report an objective level of alveolar congestion. Automated assessment can also be deployed in the absence of trained professionals, in scenarios where resources are scarce and trained personnel is unavailable. In particular, artificial intelligence methods such as machine learning could decrease variability and improve consistency with a potential for improved diagnosis and outcome [19]–[21]. Recently, the need for automated assessment of B-lines in the evaluation of COVID-19 patients has been raised [22].

We sought to develop and test a deep learning (DL) automated algorithm to assess the presence of B-lines on ultrasound clips. DL is a state-of-the-art method in medical image analysis [23]. It relies on automatic learning of complex patterns from the existing data and then reaching intelligent decisions based on learned behavior. It is recognized as an effective tool for medical applications, because it is suitable for the type of data encountered in medicine where the high dimensionality and variable environments pose problems for classical analytical solutions [24], [25]. In particular, clips containing B-lines can exhibit considerable heterogeneity across patients depending on the underlying pathology, image characteristics, and machine presets [26], [27]. Due to this heterogeneity, B-lines may be well assessed through DL approaches [24], [25].

II. Materials and Methods

Approval for the study was obtained from the Yale Human Research Protection Program. A waiver of Health Insurance Portability and Accountability Act (HIPAA) authorization was granted for the entire study. A full waiver of consent was also granted for the entire study.

We utilized ultrasound B-mode clips from an existing database of ED patients to provide training and test sets to develop and test a DL algorithm that would identify the presence and assess the severity of B-lines. The model for automated lung feature detection was developed using deep convolutional neural networks (CNNs). Interpretations of the clips by the algorithm were compared to expert sonographer interpretation.

A. Data Extraction

Ultrasound clips were extracted from an existing database that included all POCUSs performed in the Yale-New Haven Hospital Emergency Department system (QPath, Telexy Healthcare) from 2012 onward. Clips obtained with three different transducers (linear, curvilinear, and phased array) from patients presenting with dyspnea or chest pain where thoracic ultrasound views had been obtained as part of routine ED care were included. Yale New Haven Hospital Emergency Department uses Philips SPARQ (Philips Healthcare) ultrasound systems, and the probes used during the time frame the data were collected are linear L12-4 model, curvilinear C5-1 model, and phased-array S4-2 model.

Clips were extracted starting from January 17, 2017, working back through the Qpath database until a total of 400 consecutive thoracic ultrasound clips, each from a unique patient, were collected. The frame rate ranged from 20 to 48 frames/s, with an average duration of 2.6 s across the clips. Frame rate was not especially set for this project and was standard for the Philips SPARQ ultrasound system and its default factory settings.

All lung ultrasound clips were downloaded in both Digital Imaging and Communications in Medicine (DICOM) and MPEG-4 video file format (MP4). DICOM clips were deidentified using Dicom Cleaner (version 10.2, PixelMed Publishing), and MP4 clips were deidentified using Clip De-Identifier (Ben C. Smith, MD; https://www.ultrasoundoftheweek.com/clipdeidentifier/), both freeware software packages are available online. The DICOM data were used for algorithm development, while the MP4 data were used for viewing and annotation purposes.

B. Data Labeling and Organization

Each of the 400 clips was split into several subclips consisting of 12 consecutive frames each, yielding a total of 2415 subclips for analysis. Each subclip was around half a second duration. While not all the subclips from the same patient were truly independent, they often had different characteristics due to respiration-induced dynamic motion that changed the ultrasound imagery across the clip.

All the 2415 subclips were rated by two emergency physician POCUS experts with fellowship training for severity of B-lines, based on a predetermined ordinal scale from 0 (none) to 4 (severe). Examples of still images illustrating each severity level are shown in Fig. 1.

Subclips were rated as 0 if no B-lines were visible, 1 if there was an occasional B-line but the subclip was still thought to be consistent with a clinically normal result, 2 if the subclip was abnormal but contained relatively few B-lines, 3 if there was a large burden of B-lines, and 4 at the most severe end of the spectrum.

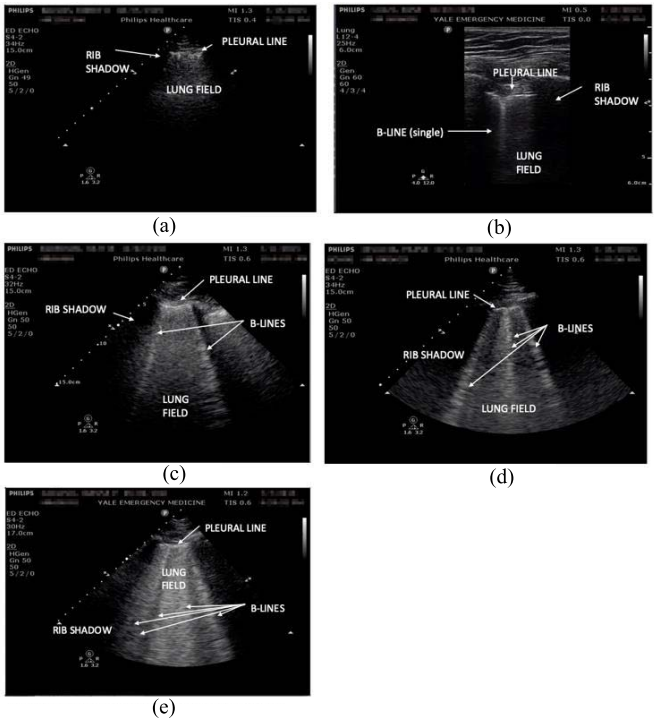


Fig. 1. Examples of clips illustrating each severity level. (a) Rating 0, no B-lines. (b) Rating 1, isolated B-lines. (c) Rating 2, a few B-lines. (d) Rating 3, many B-lines. (e) Rating 4, innumerable B-lines (only a few marked).

Custom software was developed in MATLAB (The MathWorks, Inc., Natick MA) that presented the subclips for rating in a randomized and blinded manner and recorded the responses of each reviewer. For cases where the two experts assigned different ratings to the subclips, the final rating was adjudicated by rereview and discussion between the two raters. In order to perform a binary classification, subclips with ratings of 0 and 1 were pooled into a normal category and the remaining data were pooled into an abnormal category.

For algorithm training and validation, data from 300 of the 400 unique patients yielding 1847 subclips were selected, and further separated in an approximately 85:15 ratio for training and validation. Several data augmentation steps were done that included left–right flip, time reversal of the frames, minor rotations, minor changes in aspect ratio, and changes in gain, resulting in a significant increase in data available for analysis for the training set. The remaining 100 unique patients were used to provide the test data sets where a random selection of a single subclip from each patient yielded 100 test data sets. The remaining subclips from these 100 patients were set aside and were not used for training or validation in order to maintain sample independence during the testing phase. There was no overlap patient wise among the training, validation, and test data sets. Approximately one month from the initial rating, experts relabeled the 100 subclips from the test data set blinded to their initial rating to provide intrarater reliability.

C. Data Preprocessing

The images from the three types of transducers (linear, curvilinear, and sector probes) yield distinct image formats that could potentially confound the algorithm when used as input. In order to standardize the data across these different formats, the images were processed to a consistent rectilinear format, where the B-lines would always present as vertical lines aligned with the ultrasound beam direction (Fig. 2).

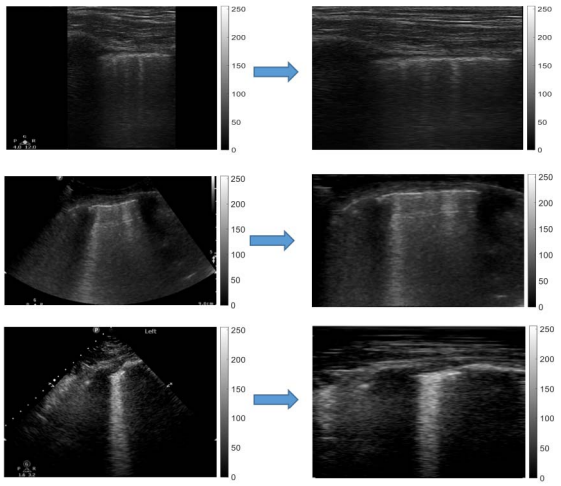


Fig. 2. Examples of clips illustrating preprocessing of data to rectilinear format.

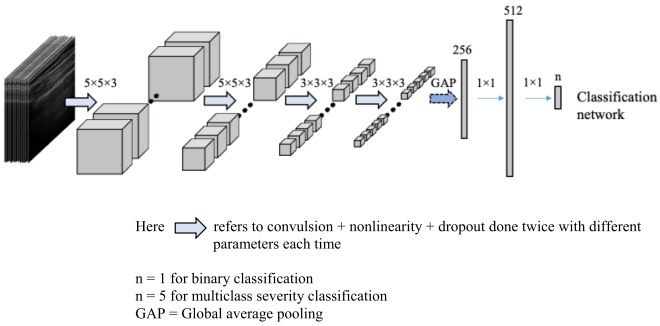


Fig. 3. Schematic of the DL model built for B-line rating.

For the curvilinear and sector probe data, the coordinates of the apex and image region end points were manually selected and fed into a custom-written program in MATLAB that performed the geometric transformation from polar to rectilinear format. This process also removed the textual information surrounding the image area. All frames were downsampled to a resolution of 75 × 75. The data sets were also normalized to have pixel intensities in the range of (0,1) by dividing the DICOM image data by 255.

D. Model(s) and Training

The DL model built for B-line rating was a supervised learning model and consisted of a CNN consisting of eight intermediate layers followed by two fully connected layers (Fig. 3).

Each intermediate layer consisted of a 3-D convolution operation, a rectified linear unit (ReLU) as a nonlinear activation function, and a dropout. Every second intermediate convolutional layer had a stride of (1,1,1) and a dropout of 0.1 and a stride of (2,2,1) and a dropout of 0.2 otherwise. The network used global average pooling after the output

of the last intermediate layer. Two fully connected layers follow the global average pooling. The first fully connected layer consisted of a 1 × 1 convolution followed by ReLU activation. The second layer, so-called task layer, consisted of a 1 × 1 convolution followed by sigmoid or softmax activation function for binary and multiclass classification tasks. Dimensions of the last task layer depended on the task: 1 for binary and 5 for multiclass classification problems. In total, the network had about 4M parameters. The above network parameter configuration (i.e., type and number of layers, type of activation functions, and other design decisions) was chosen experimentally after hyperparameter optimization and choosing the architecture that gave the best results on validation data. Separate models and training runs were performed for the binary and multiclass problems.

The network was optimized with a cross entropy loss. The model was trained in TensorFlow using RMSprop optimizer with a batch size of 32 and an initial learning rate of 0.0001 that decayed every 500 iterations with an exponential rate of 0.5. Early stopping criteria that stopped training when the model performance on validation data was not improving was used to prevent overfitting. The training time was approximately 57 min on a single GPU system (NVIDIA Titan Xp). The inferencing time for a single block of 12 frame data was about 5 ms on the same GPU and about 160 ms when only a single CPU was used (Intel Xeon CPU E5-1620 v4 at 3.50 GHz).

For the severity classification problem, the severity class that had the highest output value by the algorithm was selected as the predicted class. For the binary classification problem, a threshold for B-line detection was determined using the validation data that corresponded to equal sensitivity and specificity values. The test data were not accessed until the hyperparameters and threshold selection for the binary case were completed.

E. Comparison to State-of-the-Art DL Models

In this work, we chose a relatively shallow, custom-designed architecture rather than state-of-the-art networks such as AlexNet, ResNet, or DenseNet. This choice was driven by several requirements and the nature of the data. Typical state-of-the-art networks are pretrained on natural red green blue (RGB) color model images (e.g., ImageNet data set) and use a relatively large input image size. Ultrasound is a grayscale imaging modality and requires a degree of computational performance that is demanding for large input sizes, especially on mobile ultrasound devices. Pretraining often not only boosts the performance of deep networks but it also constrains input to a certain size (e.g., 224 × 224 × 3), whereas our data consisted of grayscale ultrasound images with an additional temporal information along the third dimension (12 temporal frames). To choose the most optimal network architecture, we performed additional experiments, in which we compared our custom-made network (3-D CsNet) with state-of-the-art DL models (3-D ResNet and 3-D DenseNet) and two variations of our network having either 2-D filters (2-D CsNet) or more convolutional layers (3-D CdNet). To enable initialization of weights with pretrained ImageNet parameters on all 12 temporal frames, we modified both the ResNet and DenseNet architectures by repeating the weights of 2-D filters 12 times along time dimension and rescaling them by dividing by 12 [28]. Comparison among the different architectures was only performed for the binary classification task. All architectures are summarized in Table I.

TABLE I
COMPARISON OF FIVE DIFFERENT MODELS THAT WERE EXPERIMENTALLY EVALUATED

MODEL	2D CSNET	3D CSNET	3D CDNET	3D RESNET	3D DENSENET
IMAGE SIZE	75 × 75 × 12	75 × 75 × 12	75 × 75 × 12	224 × 224 × 12	224 × 224 × 12
BATCH SIZE	32	32	32	16	8
NO. OF LAYERS	8	8	10	50	101
TOTAL PARAMS	1.3M	3.9M	14.6M	46.1M	11.1M
INITIALIZATION	R	R	R	P	P
SPEED ON GPU	4 ± 1 MS	5.1 ± 1 MS	11 ± 1 MS	18 ± 1 MS	52 ± 1 MS
AUC	0.95	0.97	0.93	0.92	0.91

R= RANDOM, P= PRE-TRAINED

F. Statistics

The algorithm rating of the 100 test subclips was compared to the gold standard represented by consensus expert rating using unweighted kappa for binary classification and weighted kappa for multiclass severity classification [29]. In addition, consensus expert ratings were evaluated for intrarater reliability using the initial ratings for the test 100 subclips.

III. RESULTS

Out of the 400 clips, 95 clips were obtained with the linear probe, 52 with the curvilinear probe and 253 with the sector probe. The spread of the subclips regarding B-line presence or absence and severity of B-lines according to consensus ratings is presented in Table II.

Fig. 4 summarizes the results comparing the DL models and expert review of the test clips. When compared to expert interpretation for the presence or absence of significant B-lines, the model for binary classification yielded a sensitivity of 93% (95% confidence interval (CI) 81%–98%) and a specificity of 96% (95% CI 84%–99%), with an area under the curve (AUC) of 0.97. Kappa for binary classification was 0.88 (95% CI 0.79–0.97).

For the multiclass classification of B-line severity, agreement between the (second) DL model and consensus expert ground truth was 93% when calculated within one score deviation of the training set. Agreement for severity classification yielded a linear weighted kappa of 0.65 (95% CI 0.56–0.74).

Intrarater agreement of consensus expert review of the same 100 test subclips measured by unweighted kappa for binarity classification was 0.89 (95% CI 0.81–0.99) and by weighted kappa for severity classification was 0.87 (95% CI 0.81–0.93).

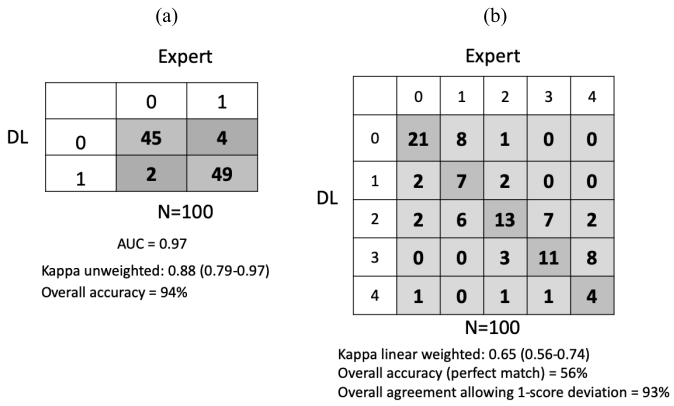


Fig. 4. Expert versus DL for (a) binary and (b) severity ratings.

TABLE II
SPREAD OF ALL SUBCLIPS REGARDING THE PRESENCE AND SEVERITY OF B-LINES ACCORDING TO RATING BY EXPERT CONSENSUS

BINARY		SEVERITY	
NEGATIVE FOR B-LINES	1095	CATEGORY 0	673
POSITIVE FOR B-LINES	1320	CATEGORY 1	422
		CATEGORY 2	669
		CATEGORY 3	485
		CATEGORY 4	166
TOTAL	2415	TOTAL	2415

As stated in Section II, we experimentally compared our model to other state-of-the-art methods. The analysis indicated that the custom network described in this study had smaller size, ran several times faster on GPU, and had a slightly higher accuracy than the other networks. The results of the comparison are summarized in Table I.

IV. DISCUSSION

A. Model Characteristics

We chose a relatively shallow custom-made model architecture with 3-D filters (3-D CsNet) for B-line assessment.

The 3-D CsNet architecture was shown to be approximately ten times and three times faster on GPU compared to 3-D DenseNet and 3-D ResNet architectures, respectively (Table I). Input size and number of convolutional operations could be a major factor for practical implementation especially on mobile point-of-care platforms. The 3-D CsNet architecture had 11 times and 2 times less trainable parameters than the 3-D ResNet and 3-D DenseNet architectures, respectively. The 3-D CsNet architecture outperformed all other models for a binary classification task with an AUC of 0.97 compared to 0.92 and 0.91 for the ResNet and DenseNet models, respectively. Finally, custom models are more flexible and easier to deploy because they lack the need for pretraining.

B. Comparison to Other B-Line Detection and Quantification Literature

To our knowledge, this is the largest study to date on automated assessment of B-lines. Our results show that the algorithm can distinguish between the presence and absence of B-lines with substantial agreement when compared to expert consensus. The algorithm can also distinguish B-line severity with moderate agreement compared to expert consensus.

While simple automated algorithms for processing a still image of a lung ultrasound and detecting or quantifying B-lines have been described, our approach uses DL which we believe is better suited at capturing variability in data across a spectrum of subjects. In clinical practice, it becomes apparent that clips obtained in real patients vary in terms of quality, B-line signal strength and appearance (broad versus narrow B-lines, B-lines associated with pneumonias, uneven pleural surface, etc.), and the amount of noise present in the clip. These variations make detection of B-lines using traditional methods difficult and limited in generalizability. Furthermore, clinicians interpret an ultrasound clip by evaluating a movie clip that is dynamic in nature and take into consideration elements such as clip quality, depth, width of pleural line visible, then appearance and apparent quantity of B-lines. In practice, clinicians do not simply count B-lines, as B-line count can vary between one frame and another depending on the movement of probe and respiratory movements, and the physician must make a judgment as to the overall B-line burden. Ultrasound trained experts often look at the dynamic clip in deciding how many B-lines exist as variation with respiration, shadowing from ribs brought in by thoracic movement can all influence frame B-line quantity. It can take significant training to obtain adequate expertise to appropriately and reliably detect and quantify B-line burden. A DL model has the advantage of learning and, thus, processing ultrasound clips to generate a solution capable of tackling the sophisticated patterns required for interpreting lung ultrasound clips from a quantification standpoint. Interpretation of clips with a traditional model would be difficult to standardize across varied clinical scenarios, ultrasound machines, and probe types.

While some prior studies have looked at automating B-line interpretation, they were limited by small sample size, narrow inclusion criteria, and lack of expert interpretation for agreement [30], [31]. Furthermore, most existing studies involve automated image processing algorithms as opposed to DL. For instance, one study used clips from 20 stable dialysis outpatients to develop an image processing algorithm for B-line detection and quantification in dynamic clips [31]. Quantification was associated with clinical parameters during dialysis such as blood pressure, age, and dialysis volume, but algorithm performance against an expert review was not assessed [31]. The methods by Weitzel *et al.* [31] processed distortion-corrected B-mode image loops by region of interest definition, spatial and temporal filtering, and comet energy evaluation in order to detect and count the B-lines. The number of B-lines (used for quantification in this study) could be misleading in that there are many clinical situations of fused B-lines that indicate a more severe level of pulmonary edema. Thus, in this study, we have used a severity metric based on the judgment of expert clinicians rather than the number of B-lines.

Brattain *et al.* [30] developed a feature detection algorithm for B-line quantification based on features from 50 ultrasound clips from patients enrolled prospectively as part of a separate trial of dyspnea. Agreement to emergency physician expert review was 0.9 when calculated within 1 score deviation on the training set. Validation on a separate 13-clip data set yielded

perfect agreement [30]. The main limitation of this study is the small sample size for training and testing sets, which is particularly problematic for feature detection. Appearance of ultrasound clips containing B-lines can be extremely varied, and it is possible that simple algorithms using feature detection would fail when tested on a larger sample.

van Sloun and Demi [32], [33] described a DL algorithm identifying the frames of an ultrasound video where B-lines are found, first in ultrasound phantoms, then *in vivo*. However, their research focuses on recognition, not quantification. Moshavegh *et al.* [34] described an automated image processing method relying on fit Gaussian models for detection and visualization of B-lines. This study is limited due to small patient numbers (four controls and four patients with clinical pulmonary edema) and absence of performance evaluation regarding the accuracy of identifying individual B-lines versus obtaining a score that was different between the two groups. Kulhare *et al.* [35] described single-shot detection CNNs for detection of several lung ultrasound features including B-lines. Their results also show promise in DL approaches for lung ultrasound feature assessment, but all data in their study were from animal models [35]. In contrast, this study used a large human subject database, used 3-D data sets as input to capture dynamic B-line behavior, and investigated prediction of multiple levels of severity.

A promising image processing method by Anantrasirichai *et al.* [36], using a simple local maxima technique in the Radon transform domain, associated with known clinical definitions of line artifacts. Importantly, the technique is evaluated using as ground truth lines identified by experts [36]. However, the authors do not test quantification and how it compares to clinical interpretation.

A computer-generated rating based on DL may ultimately yield more reliable and objective results providing information for the initial diagnosis and consistent analysis independent of user and to determine progression over time. For example, patients in a critical care environment may be serially monitored over time to determine the efficacy of treatment. Improved monitoring might prevent HF readmissions and decrease healthcare system cost.

Reliability of rating is essential for any diagnostic test. Anderson *et al.* [37] revealed substantial interrater reliability among trained emergency physicians for B-line quantification in a single intercostal space, but agreement between experts and novices may be considerably weaker and may also depend on the thoracic region examined. Gullett *et al.* [38] showed substantial agreement between experts and novices in the anterior superior lung zones but not in the lateral superior and particularly lateral inferior and posterior zones. The intrarater agreements of 0.89 and 0.87 in this study are very good and lend credence to the consensus expert rating used as a ground truth.

C. Limitations and Considerations for Software Development and Performance

A weaker performance of our methods on multiclass severity rating compared to binary rating could potentially be explained by the presence of many categories. To test whether the number of categories influences the kappa obtained by the algorithm, analysis was rerun with a severity scale of 0 (no significant B-lines, prior levels 0 and 1), 1 (some B-lines, former severity level 2), and severe (former severity levels 3 and 4). Categorizing the ratings in this way did improve DL method's classification performance (weighted Kappa 0.72, 95% CI 0.62–0.82) when comparing DL ratings to expert consensus. A higher number of categories may result in lower agreement when it comes to DL performance for the same amount of data. However, a large enough number of categories are actually necessary in order to use and track severity in a clinically relevant fashion and too few categories would hinder the usefulness of a severity scale. We believe that in the scale used in this study, a change of score of one would call for clinical attention and the change of more than one could call for some intervention (such as change in medications).

While the test data set contained comparable number of clips containing B-lines between the severity categories, the proportion of clips in each severity category was not equal. Categories 0, 1, 2, and 3 contained roughly a similar number of clips, but category 4 had a quarter of the number of clips of other categories. Disagreements in rating for category 4 clips could substantially affect overall agreement for the entire data set.

DL algorithms require substantial data for training, often thousands to tens of thousands of separate data points. To feasibly produce such a rich data set from single-clip clinical data would be challenging and time prohibitive as few hospital databases contain that quantity of point-of-care lung ultrasound clips as of this time. Increasing the number of separate clips used for training the algorithm beyond 400 was key for improving the performance of the DL software, which was addressed by splitting the data into subclips and the use of data augmentation methods. As noted in Section II, subclips from the same patient had different characteristics due to respiration-induced dynamic motion resulting in different appearance of the ultrasound image across the subclips. Thus, while not strictly independent clips, these subclips were different enough to be treated as independent data.

One limitation of the data set affecting generalizability of the prediction results is the fact that the learning data come from one brand and type of ultrasound machine. However, clips from three different probe types and different presets were used in an effort to display a range of clinical and physiological variations in order to improve the applicability to real-world data.

Care was taken to keep the 100 subclips used for testing separate from the data used to develop the algorithm. These subclips were truly independent, each originating from a separate patient. This avoided any potential overfitting problem due to similarity in consecutive subclips from the same patient.

In the future, further training the algorithm on whole clips rather than collection of frames would perhaps increase the ability of the software to more accurately detect severity levels for the whole clip. This whole clip evaluation would closely resemble how ultrasound lung clips are evaluated by experts for B-line severity.

Next steps to confirm algorithm validation and improve the algorithm would be testing of the model in real time, while obtaining lung ultrasound clips on patients.

V. Conclusion

In this work, we present a custom-designed DL network that operated on dynamic ultrasound data for automated assessment of sonographic lung B-lines. The network was developed using 2415 subclips of 12 frames each extracted from 400 patients using expert consensus ratings as ground truth. This DL algorithm showed promise for automated assessment for both the binary and severity ratings based on the test data of 100 subclips that were set aside and not used during the training. This method could be used to improve the reliability and objectivity of the presence and severity of B-lines for diagnosis and prognosis of patients with respiratory complaints. Clinical applications for this include HF, pneumonia, and ARDS. In addition, reliable identification of B-lines may aid in diagnosis and management of the ongoing COVID-19 pandemic.

References

[1] L. F. McCaig and C. W. Burt, "National hospital ambulatory medical care survey: 1999 emergency department summary," *Adv. Data*, vol. 320, pp. 1–34, Jun. 2001.

[2] M. Al Deeb, S. Barbic, R. Featherstone, J. Dankoff, and D. Barbic, "Point-of-care ultrasonography for the diagnosis of acute cardiogenic pulmonary edema in patients presenting with acute dyspnea: A systematic review and meta-analysis," *Academic Emergency Med.*, vol. 21, no. 8, pp. 843–852, Aug. 2014, doi: 10.1111/acem.12435.

[3] S. Laribi *et al.*, "Epidemiology of patients presenting with dyspnea to emergency departments in Europe and the Asia–Pacific region," *Eur. J. Emergency Med.*, vol. 26, no. 5, pp. 345–349, Oct. 2019, doi: 10.1097/MEJ.0000000000000571.

[4] C. L. Moore and J. A. Copel, "Point-of-care ultrasonography," *New England J. Med.*, vol. 364, no. 8, pp. 749–757, Feb. 2011, doi: 10.1056/NEJMra0909487.

[5] D. Lichtenstein and G. Meziere, "A lung ultrasound sign allowing bedside distinction between pulmonary edema and COPD: The comet-tail artifact," *Intensive Care Med.*, vol. 24, no. 12, pp. 1331–1334, Dec. 1998, doi: 10.1007/s001340050771.

[6] D. Lichtenstein, G. Méziere, P. Biderman, A. Gepner, and O. Barré, "The comet-tail artifact: An ultrasound sign of alveolar-interstitial syndrome," *Amer. J. Respiratory Crit. Care Med.*, vol. 156, no. 5, pp. 1640–1646, Nov. 1997, doi: 10.1164/ajrccm.156.5.96-07096.

[7] Y. Huang *et al.*, "A preliminary study on the ultrasonic manifestations of peripulmonary lesions of non-critical novel coronavirus pneumonia (COVID-19)," May 8, 2020, p. 14. [Online]. Available: https://ssrn.com/abstract=3544750, doi: 10.2139/ssrn.3544750.

[8] E. Poggiali *et al.*, "Can lung US help critical care clinicians in the early diagnosis of novel coronavirus (COVID-19) pneumonia?" *Radiology*, vol. 295, no. 3, p. E6, Jun. 2020.

[9] G. Soldati *et al.*, "Proposal for international standardization of the use of lung ultrasound for patients with COVID-19: A simple, quantitative, reproducible method," *J. Ultrasound Med.*, vol. 39, pp. 1413–1419, Mar. 2020, doi: 10.1002/jum.15285.

[10] F. C. Yue Lee, C. Jenssen, and C. F. Dietrich, "A common misunderstanding in lung ultrasound: The comet tail artefact," *Med. Ultrasonography*, vol. 20, no. 3, p. 379, Aug. 2018, doi: 10.11152/mu-1573.

[11] G. Volpicelli *et al.*, "Bedside lung ultrasound in the assessment of alveolar-interstitial syndrome," *Amer. J. Emergency Med.*, vol. 24, no. 6, pp. 689–696, Oct. 2006, doi: 10.1016/j.ajem.2006.02.013.

[12] S. Coiro *et al.*, "Prognostic value of pulmonary congestion assessed by lung ultrasound imaging during heart failure hospitalisation: A two-centre cohort study," *Sci. Rep.*, vol. 6, no. 1, p. 39426, Dec. 2016, doi: 10.1038/srep39426.

[13] F. Frassi, L. Gargani, P. Tesorio, M. Raciti, G. Mottola, and E. Picano, "Prognostic value of extravascular lung water assessed with ultrasound lung comets by chest sonography in patients with dyspnea and/or chest pain," *J. Cardiac Failure*, vol. 13, no. 10, pp. 830–835, Dec. 2007, doi: 10.1016/j.cardfail.2007.07.003.

[14] L. Gargani *et al.*, "Persistent pulmonary congestion before discharge predicts rehospitalization in heart failure: A lung ultrasound study," *Cardiovascular Ultrasound*, vol. 13, no. 1, Dec. 2015, doi: 10.1186/s12947-015-0033-4.

[15] E. Platz *et al.*, "Detection and prognostic value of pulmonary congestion by lung ultrasound in ambulatory heart failure patients," *Eur. Heart J.*, vol. 37, no. 15, pp. 1244–1251, Apr. 2016, doi: 10.1093/eurheartj/ehv745.

[16] C. Zoccali *et al.*, "Pulmonary congestion predicts cardiac events and mortality in ESRD," *J. Amer. Social Nephrol.*, vol. 24, no. 4, pp. 639–646, 2013, doi: 10.1681/ASN.2012100990.

[17] E. Platz, A. A. Merz, P. S. Jhund, A. Vazir, R. Campbell, and J. J. Mcmurray, "Dynamic changes and prognostic value of pulmonary congestion by lung ultrasound in acute and chronic heart failure: A systematic review," *Eur. J. Heart Failure*, vol. 19, no. 9, pp. 1154–1163, Sep. 2017, doi: 10.1002/ejhf.839.

[18] D. M. Tierney *et al.*, "Pulmonary ultrasound scoring system for intubated critically ill patients and its association with clinical metrics and mortality: A prospective cohort study," *J. Clin. Ultrasound*, vol. 46, no. 1, pp. 14–22, Jan. 2018, doi: 10.1002/jcu.22526.

[19] L. J. Brattain, B. A. Telfer, M. Dhyani, J. R. Grajo, and A. E. Samir, "Machine learning for medical ultrasound: Status, methods, and future opportunities," *Abdominal Radiol.*, vol. 43, no. 4, pp. 786–799, Apr. 2018, doi: 10.1007/s00261-018-1517-0.

[20] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *RadioGraphics*, vol. 37, no. 2, pp. 505–515, Mar. 2017, doi: 10.1148/rg.2017160130.

[21] H. Shokoohi, M. A. LeSaux, Y. H. Roohani, A. Liteplo, C. Huang, and M. Blaivas, "Enhanced point-of-care ultrasound applications by integrating automated feature-learning systems using deep learning," *J. Ultrasound Med.*, vol. 38, no. 7, pp. 1887–1897, Jul. 2019, doi: 10.1002/jum.14860.

[22] F. Corradi, G. Via, F. Forfori, C. Brusasco, and G. Tavazzi, "Lung ultrasound and B-lines quantification inaccuracy: B sure to have the right solution," *Intensive Care Med.*, vol. 46, no. 5, pp. 1081–1083, May 2020, doi: 10.1007/s00134-020-06005-6.

[23] S. Liu *et al.*, "Deep learning in medical ultrasound analysis: A review," *Engineering*, vol. 5, no. 2, pp. 261–275, Apr. 2019, doi: 10.1016/j.eng.2018.11.020.

[24] K. K. L. Wong, L. Wang, and D. Wang, "Recent developments in machine learning for medical imaging applications," *Computerized Med. Imag. Graph.*, vol. 57, pp. 1–3, Apr. 2017, doi: 10.1016/j.compmedimag.2017.04.001.

[25] G.-S. Fu, Y. Levin-Schwartz, Q.-H. Lin, and D. Zhang, "Machine learning for medical imaging," *J. Healthcare Eng.*, vol. 2019, pp. 1–2, Apr. 2019, doi: 10.1155/2019/9874591.

[26] C. F. Dietrich *et al.*, "Lung B-line artefacts and their use," *J. Thoracic Disease*, vol. 8, no. 6, pp. 1356–1365, Jun. 2016, doi: 10.21037/jtd.2016.04.55.

[27] D. A. Lichtenstein, "Current misconceptions in lung ultrasound," *Chest*, vol. 156, no. 1, pp. 21–25, Jul. 2019, doi: 10.1016/j.chest.2019.02.332.

[28] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Honolulu, HI, USA: IEEE, Jul. 2017, pp. 4724–4733.

[29] N. Gisev, J. S. Bell, and T. F. Chen, "Interrater agreement and interrater reliability: Key concepts, approaches, and applications," *Res. Social Administ. Pharmacy*, vol. 9, no. 3, pp. 330–338, May 2013, doi: 10.1016/j.sapharm.2012.04.004.

[30] L. J. Brattain, B. A. Telfer, A. S. Liteplo, and V. E. Noble, "Automated B-Line scoring on thoracic sonography," *J. Ultrasound Med.*, vol. 32, no. 12, pp. 2185–2190, Dec. 2013, doi: 10.7863/ultra.32.12.2185.

[31] W. F. Weitzel *et al.*, "Quantitative lung ultrasound comet measurement: Method and initial clinical results," *Blood Purification*, vol. 39, nos. 1–3, pp. 37–44, Jun. 2015, doi: 10.1159/000368973.

[32] R. J. G. van Sloun and L. Demi, "Deep learning for automated detection of B-lines in lung ultrasonography," *J. Acoust. Soc. Amer.*, vol. 144, no. 3, p. 1668, Sep. 2018.

[33] R. J. G. van Sloun and L. Demi, "Localizing B-lines in lung ultrasonography by weakly supervised deep learning, *in-vivo* results," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 4, pp. 957–964, Apr. 2020, doi: 10.1109/JBHI.2019.2936151.

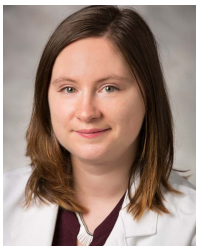
[34] R. Moshavegh, K. L. Hansen, H. Moller-Sorensen, M. B. Nielsen, and J. A. Jensen, "Automatic detection of B-lines in *in vivo* lung ultrasound," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 66, no. 2, pp. 309–317, Feb. 2019, doi: 10.1109/TUFFC.2018.2885955.

[35] S. Kulhare *et al.*, "Ultrasound-based detection of lung abnormalities using single shot detection convolutional neural networks," in *Proc. Int. Workshops Simulation, Image Process., Ultrasound Syst. Assist. Diagnosis Navigat. POCUS, BIVPCS, CuRIOUS, CPM, MICCAI*. Basel, Switzerland: Springer-Verlag, 2018, pp. 65–73. [Online]. Available: https://ohsu.pure.elsevier.com/en/publications/ultrasound-based-detection-of-lung-abnormalities-using-single-sho

[36] N. Anantrasirichai, W. Hayes, M. Allinovi, D. Bull, and A. Achim, "Line detection as an inverse problem: Application to lung ultrasound imaging," *IEEE Trans. Med. Imag.*, vol. 36, no. 10, pp. 2045–2056, Oct. 2017, doi: 10.1109/TMI.2017.2715880.

[37] K. L. Anderson, J. M. Fields, N. L. Panebianco, K. Y. Jenq, J. Marin, and A. J. Dean, "Inter-rater reliability of quantifying pleural B-lines using multiple counting methods," *J. Ultrasound Med.*, vol. 32, no. 1, pp. 115–120, Jan. 2013, doi: 10.7863/jum.2013.32.1.115.

[38] J. Gullett *et al.*, "Interobserver agreement in the evaluation of B-lines using bedside ultrasound," *J. Crit. Care*, vol. 30, no. 6, pp. 1395–1399, Dec. 2015, doi: 10.1016/j.jcrc.2015.08.021.



Cristiana Baloescu was born in Bucharest, Romania, in 1986. She received the M.D. degree from the Geisel School of Medicine, Dartmouth, Hanover, NH, USA, in 2013 and the Advanced Professional MPH degree from the Yale School of Public Health, New Haven, CT, USA, in 2019.

In 2017, she completed the Emergency Medicine Residency Program at Yale New-Haven Hospital, New Haven. She completed an Emergency Ultrasound and Research Fellowship at the Department of Emergency Medicine, Yale School of Medicine, New Haven, where she is currently a Clinical Instructor, and practicing clinically with the Emergency Department, Yale New Haven Hospital, New Haven. Her research interests include applications of machine learning to point-of-care ultrasound, ultrasound implementation in trauma care in low-resource settings, and ultrasound-guided regional anesthesia.

Dr. Baloescu serves leadership positions at the American Institute of Ultrasound in Medicine, where she currently chairs the Community of Ultrasound in Global Health. She is also a member of the Academy of Women in Emergency Medicine and the Society for Academic Emergency Medicine.

Grzegorz Toporek was born in Kraków, Poland, in 1987. He received the B.Sc. and M.Sc. degrees in biomedical engineering and computer science from the AGH University of Science and Technology, Kraków, in 2010 and 2011, respectively, and the Ph.D. degree from the University of Bern, Bern, Switzerland, in 2015, by developing and clinically validating an image-guided therapy system.

He is currently a Senior Scientist with Philips Research North America, Cambridge, MA, USA, leading projects devising machine learning methods for ultrasound imaging. He has authored ten publications and holds over 35 filed patents. The goal of his research is to expand the usage of ultrasound in point-of-care and emergency settings by providing intuitive solutions to assist both acquisition and interpretation of ultrasound images. He also contributed to the development of an image-guided robot for spine surgery for hybrid operating room equipped with Augmented Reality Surgical Navigation, Philips.

Seungsoo Kim (Member, IEEE) was born in Seoul, South Korea, in 1979. He received the B.S. and M.S. degrees in electronic engineering from Sogang University, Seoul, in 2005 and 2007, respectively, and the Ph.D. degree in biomedical engineering from the University of Texas at Austin, Austin, TX, USA, in 2011. His Ph.D. dissertation entitled Ultrasound and Photoacoustic Imaging for Cancer Detection and Therapy Guidance.

From 2011 to 2016, he was an Ultrasound Systems Engineer with Siemens Medical Solutions USA, Inc., Malvern, PA, USA, developing ultrasound elastography. From 2016 to 2018, he was a Research Scientist with Philips Research North America, Cambridge, MA, USA, developing mobile ultrasound applications. He is currently an Ultrasound Imaging Engineer with Infraredx, Inc., Bedford, MA, USA, developing a near-infrared spectroscopy intravascular ultrasound imaging system.

Katelyn McNamara received the B.S. degree in behavioral neuroscience with a minor in political science from Northeastern University, Boston, MA, USA, in 2019.

She worked as a Research Intern, with Dr. Christopher Moore and Dr. Cristiana Baloescu, with the Section of Ultrasound, Department of Emergency Medicine, Yale University, New Haven, CT, USA, studying the use of ultrasound technology in emergency medicine. She is currently a Postgraduate Research Associate with the Section of Infectious Disease, Department of Internal Medicine, Yale University. She works in clinical research involving treatments for opioid use disorder, specifically among persons with human immunodeficiency virus (HIV) and other infections and/or justice involvement. She will soon be attending the Stritch School of Medicine, Loyola University, Maywood, IL, USA, as a member of the class of 2024.

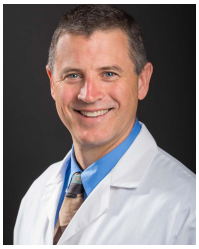
Rachel Liu received the MB BcH BAO degree in medicine from Trinity College Dublin, Dublin, Ireland, in 2007.

She completed the Emergency Medicine Residency from New York Medical College, Manhattan, NY, USA, in 2011 and the Emergency Ultrasound Fellowship from the Yale School of Medicine, New Haven, CT, USA, in 2012. Since 2012, she has been a Clinical and Education Faculty Member with the Department of Emergency Medicine, Yale School of Medicine, where she is currently an Associate Professor of emergency medicine and also the Fellowship Director of emergency ultrasound and the Director of point-of-care ultrasound education for students and faculty.

Dr. Liu has held leadership roles in the major emergency medicine organizations nationally. She is currently the President of the Society of Clinical Ultrasound Fellowships and has been a Past Chair of the American College of Emergency Physicians Ultrasound Section and a Past President of the Academy of Emergency Ultrasound with the Society for Academic Emergency Medicine. She is also a Board Member of the American Institute of Ultrasound in Medicine. She has received a number of teaching awards from the Yale School of Medicine and distinguished service awards from each of these national organizations.

Melissa M. Shaw received the B.S. degree from Southern Connecticut State University, New Haven, CT, USA, in 2002.

She has been employed with the Yale School of Medicine, New Haven, within various departments involved as a Research Coordinator. While with Yale Pediatric Endocrinology, New Haven, her research has been focused in identifying effective treatments for pediatric obesity. Since moving to Yale Emergency Medicine, New Haven, area of focus involved the promotion of reduced-radiation dose current transformers (CTs) for suspected kidney stone patients.



Robert L. McNamara was born in Atlanta, GA, USA, in 1964. He received the Bachelor of Science degree in chemical engineering from the University of Notre Dame, Notre Dame, IN, USA, in 1986, the M.D. degree from Washington University, St. Louis, MO, USA, in 1991, and the master's degree in clinical epidemiology from the Johns Hopkins School of Hygiene and Public Health, Baltimore, MD, USA, in 1997.

He trained in internal medicine and clinical cardiology at the Hospital of the University of Pennsylvania, Philadelphia, PA, USA, in 1995. He completed his fellowship in echocardiography at The Johns Hopkins Hospital, Baltimore. He continued as an Instructor of epidemiology with the Johns Hopkins School of Public Health, Baltimore, and an Instructor of medicine in cardiology with the Johns Hopkins School of Medicine, Baltimore, from 1997 to 1999. He practiced as a Staff Physician with the Chinle Comprehensive Health Care Facility, Navajo Nation, Chinle, AZ, USA, from 1999 to 2002. He then became a Faculty Member of cardiovascular medicine with Yale University, New Haven, CT, USA, where he was an Assistant Professor of medicine from 2002 to 2006 and has been an Associate Professor of medicine since 2006. He was the Director of echocardiography from 2003 to 2012 and the Program Director of the Cardiovascular Fellowship from 2007 to 2010. He has been serving as the Medical Director of the Yale New Haven Hospital Ultrasound School, New Haven, since 2003. Nationally, he has worked with the development of guidelines, data standards, and performance measures for the management of patients with atrial fibrillation and acute coronary syndromes. He works with Yale-New Haven Health System Center for Outcomes Research and Evaluation, American Heart Association, and the American College of Cardiology to assess and improve the outcomes for patients with acute coronary syndromes and other cardiovascular conditions through registry research, risk modeling, performance measures, accreditation, and clinical trials. Internationally, he works with Yale Global Health Institute and collaborates with colleagues in U.K., Sweden, China, Russia, Rwanda, Uganda, Colombia, Argentina, Dominican Republic, and Honduras on medical education and clinical research projects. Finally, he is the President of a nonprofit group dedicated to medical education across the world, the International Team of Educators Advancing Cardiovascular Health (ITEACH).

Mr. McNamara is a fellow of the American Heart Association, the American College of Cardiology, and the American Society of Echocardiography.

Balasundar I. Raju (Member, IEEE) received the B.Tech. degree in mechanical engineering from IIT Madras, Chennai, India, in 1994 and the M.S. degrees in mechanical engineering and in electrical engineering and computer science and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1998 and 2002, respectively.

In 2003, he joined Philips Research, Briarcliff Manor, NY, USA, as a Senior Scientist, where he has lead research efforts in the area of Doppler ultrasound, high-intensity-focused ultrasound, and ultrasound-mediated drug and gene delivery. He is currently a Principal Scientist with Philips Research North America, Cambridge, MA, USA, responsible for directing activities in point-of-care ultrasound including technology development, collaborating with academic and clinical institutions, developing intellectual property, and overall project management involving multisite research teams, where he is also a Group Leader of the Ultrasound Applications Group, focused on expanding the use of medical ultrasound to new users and uses with a focus on cardiac, lung, vascular, and transcranial ultrasound applications. His research interests include signal and image processing and machine learning applied to healthcare and medical ultrasound.

Christopher L. Moore was born in Washington, DC, USA, in 1969. He received the bachelor's degree from Amherst College, Amherst, MA, USA, in 1992 and the M.D. degree from the University of Virginia, Charlottesville, VA, USA, in 1998.

He completed the Emergency Medicine Residency from Carolinas Medical Center, Charlotte, NC, USA, in 2001 and the Emergency Ultrasound Fellowship from Resurrection Medical Center, Chicago, IL, USA, in 2002. He joined the Yale School of Medicine, New Haven, CT, USA, and is currently the Chief of the Section of Emergency Ultrasound, Department of Emergency Medicine. His research interests include medical imaging and point-of-care testing.