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

Baloescu Cristiana[1] MD MPH, Toporek Grzegorz[2], PHD, Kim Seungsoo[2] PhD, McNamara Katelyn[1], Liu Rachel[1] BAO MBBCh, Shaw Melissa M.[1], McNamara Robert L[1] MD MHS, Raju Balasundar[2] PhD, and Moore Christopher L.[1] MD

***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 (scale of 0 to 4) classifications. Our model yielded sensitivity of 93% (95% CI 81%-98%) and specificity 96% (95% CI 84%-99%) for presence or absence of B-lines compared to expert read, with 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 presence from 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, signal and image processing, medical signal and image processing, medical ultrasonics**

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 COVID-19 infection with progression of the B-line pattern as the disease advances.[7-9]

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 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, non-cardiogenic pulmonary edema, pneumonia, pulmonary embolus, and acute respiratory distress syndrome (ARDS), including pneumonitis from COVID-19.[6 11] Establishing 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 heart failure 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 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 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 changes in B-line profile over time as a marker of disease severity, and to evaluate 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 inter-operator and intra-operator variability and image quality control. B-line identification and quantification can be a challenging skill for novice ultrasound

Manuscript submitted for review on January 17th 2020. This work was supported in part by a grant from Philips Research North America.

The authors Cristiana Baloescu, Melissa Shaw, Rachel Liu, Chris Moore are with Yale University School of Medicine, Department of Emergency Medicine, New Haven, CT 06511 USA (e-mail: cristiana.baloescu@yale.edu, chris.moore@yale.edu, melissa.m.shaw@yale.edu, rachel.liu@yale.edu). The author Katelyn McNamara was with Yale University School of Medicine, Department of Emergency Medicine when the study was conducted (email: katiemac04@gmail.com).

The authors Toporek Grzegorz, Balasundar Raju are with Philips Research North America, Cambridge, MA 02141 USA (email: balasundar.raju@philips.com, grzegorz.toporek@philips.com).

The author Seungsoo Kim was with Philips Research North America when the research was conducted (email: kim.seungsoo@gmail.com).

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

[1] Yale University School of Medicine
[2] Philips Research North America

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 automated algorithm to assess the presence of B-lines on ultrasound clips. Deep learning 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 deep learning approaches.[24 25]

II. Materials and methods

Approval for the study was obtained from the Yale Human Research Protection Program. A waiver of 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 deep learning algorithm that would identify the presence and assess severity of B-lines. The model for automated lung feature detection was developed using deep convolutional neural networks. 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 point-of-care ultrasounds 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 was collected are: linear L12-4 model, curvilinear C5-1 model, and phased-array S4-2 model, respectively.

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/second, with an average duration of 2.6 seconds 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.

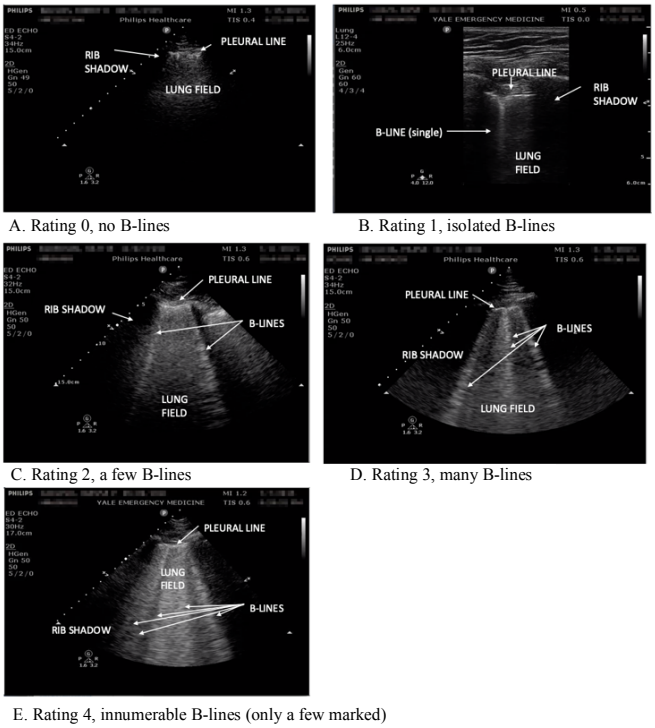
All lung ultrasound clips were downloaded in both DICOM and MP4 format. DICOM clips were de-identified using Dicom Cleaner™ (version 10.2, PixelMed Publishing™), and MP4 clips were de-identified using Clip De-Identifier (Ben C. Smith, MD; https://www.ultrasoundoftheweek.com/clipdeidentifier/), both freeware software packages 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 sub-clips consisting of 12 consecutive frames each, yielding a total of 2415 sub-clips for analysis. Each sub-clip was around half a second duration. While not all the sub-clips 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 of the 2415 sub-clips were rated by two emergency physician point-of-care ultrasound experts with fellowship training for severity of B-lines, based on a pre-determined ordinal scale from 0 (none) to 4 (severe). Examples of still images illustrating each severity level can be seen in Fig. 1.

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)

Sub-clips were rated as 0 if no B-lines were visible, 1 if there was an occasional B-line but the sub-clip was still thought to be consistent with a clinically normal result, 2 if the sub-clip 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.

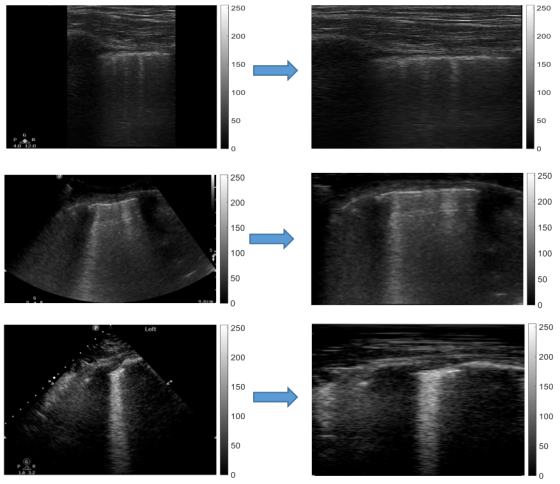
Custom software was developed in Matlab (The MathWorks, Inc., Natick MA) that presented the sub-clips for rating in a randomized and blinded manner and recorded the responses of each reviewer. For cases where the two experts assigned different ratings to the sub-clips, a final rating was adjudicated by re-review and discussion between the two raters. In order to perform a binary classification, sub-clips 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 sub-clips were selected, and further separated in approximately an 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 sub-clip from each patient yielded 100 test data sets. The remaining sub-clips 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 train, validation, and test data sets. Approximately one month from initial rating, experts re-labeled the 100 sub-clips from the test dataset blinded to their initial rating, to provide intra-rater 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 pre-processing of data to rectilinear format.

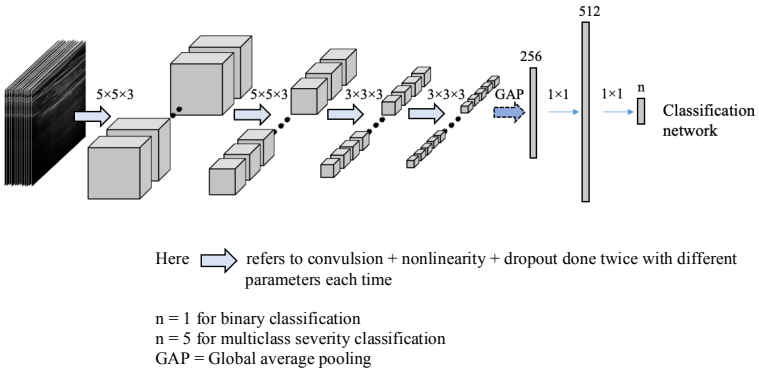


For the curvilinear and sector probe data, the coordinates of the apex and image region end points were manually selected and fed into a custom-written program in Matlab that performed the geometric transformation from polar to rectilinear format. This process also removed the textual information surrounding the image area. All frames were downsampled to a resolution of 75x75. 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 deep learning model built for B-line rating was a supervised learning model and consisted of a convolutional neural network (CNN) consisting of 8 intermediate layers followed by two fully-connected layers (Fig. 3).

Fig. 3. Schematic of deep learning model built for B-line rating



Each intermediate layer consisted of a 3D 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 1x1 convolution followed by ReLU activation. The second layer, so-called task layer, consisted of a 1x1 convolution followed by sigmoid or softmax activation function for binary and multiclass classification tasks respectively. Dimensions of the last task layer depended on the task: 1 for binary and 5 for multiclass classification problems respectively. In total, the network had about 4M parameters. The above network parameters configuration (i.e. type and number of layers, type of activation functions and other design decisions) were 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 over-fitting. The training time was approximately 57 minutes on a single GPU system (NVIDIA

Titan Xp). The inferencing time for a single block of 12 frame data was about 5 milliseconds on the same GPU, and about 160 milliseconds when only a single CPU was used (Intel Xeon CPU E5-1620 v4 @ 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 deep learning 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 as well as the nature of the data. Typical state-of-the-art networks are pre-trained on natural RGB images (e.g., ImageNet data set) and use a relatively large input image size. Ultrasound is a grayscale imaging modality and computational performance requirements, especially on mobile ultrasound devices are demanding for large input sizes. Pre-training often boosts the performance of deep networks but it also constrains input to a certain size (e.g. 224x224x3), whereas our data consisted of grayscale ultrasound images with an additional temporal information along a third dimension (12 temporal frames). To choose the most optimal network architecture we performed additional experiments, in which we compared our custom-made network (3D CsNet) with state-of-the-art deep learning models (3D ResNet and 3D DenseNet) as well as two variations of our network having either 2D filters (2D CsNet) or more convolutional layers (3D CdNet). To enable initialization of weights with pre-trained ImageNet parameters on all 12 temporal frames we modified both the ResNet and DenseNet architectures by repeating the weights of 2D 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 II.

F. Statistics

The algorithm rating of the 100 test sub-clips 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 intra-rater reliability using the initial ratings for the test 100 sub-clips.

III. RESULTS

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

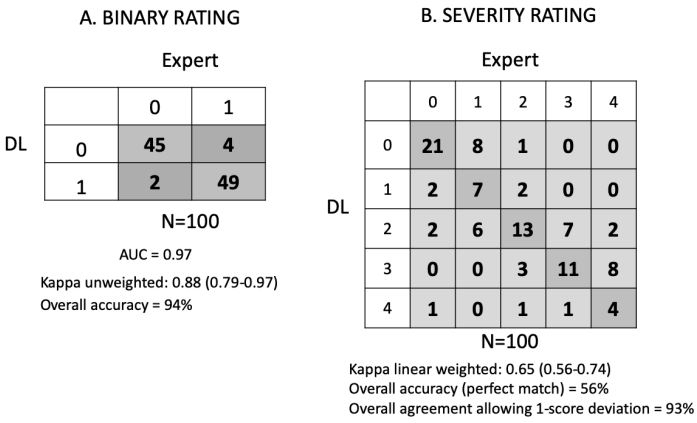
TABLE I
SPREAD OF ALL SUBCLIPS ACCORDING TO PRESENCE AND SEVERITY OF B-LINES, ACCORDING TO RATING BY EXPERT CONSENSUS.

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

Fig. 4 summarizes the results comparing the DL models and expert review of the test clips. When compared to expert interpretation for presence or absence of significant B-lines the model for binary classification yielded a sensitivity of 93% (95% 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) deep learning model and consensus expert ground truth was 93% when calculated within one score deviation of the training set. Agreement for severity classification yielded linear weighted kappa of 0.65 (95% CI 0.56-0.74).

Fig. 4. Expert versus DL for binary and severity ratings



Intra-rater agreement of consensus expert review of the same 100 test sub-clips, 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).

As stated in the Methods section, 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. Results of the comparison are summarized in Table II.

TABLE II
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

IV. DISCUSSION

A. Model characteristics

We chose a relatively shallow custom-made model architecture with 3D filters (3D CsNet) for B-line assessment. The 3D CsNet architecture was shown to be approximately 10 times and 3 times faster on GPU compared to 3D DenseNet and 3D ResNet architectures respectively (Table II). Input size as well as number of convolutional operations could be a major factor for practical implementation especially on mobile point-of-care platforms. The 3D CsNet architecture had 11 times and 2 times less trainable parameters than the 3D ResNet and 3D DenseNet architectures respectively. The 3D 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 ResNet and DenseNet models respectively. Finally, custom models are more flexible and easier to deploy because they lack the need for pre-training.

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 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 deep learning 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 judgement 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 deep learning 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 stand-point. 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 deep learning. 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. processed distortion corrected B-mode image loops by region of interest definition, spatial and temporal filtering as well as comet energy evaluation in order to detect and count the B-lines.[31] 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 our study we have used a severity metric based on the judgement of expert clinicians rather than the number of B-lines.

Brattain et al. 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.[30] 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 dataset 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.

Demi and van Sloun describe a deep learning algorithm identifying the frames of an ultrasound video where B-lines are found, first in ultrasound phantoms[32], then in vivo.[33] However, their research focuses on recognition, not quantification. Moshavegh et al. describe an automated image processing method relying on fitted Gaussian models for detection and

visualization of B-lines.[34] This study is limited due to small patient numbers (4 controls and 4 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. describe a single-shot detection CNNs for detection of several lung ultrasound features including B-lines.[35] Their results also show promise in deep learning approaches for lung ultrasound feature assessment, but all data in their study were from animal models.[35] In contrast, our study used a large human subjects database, used 3D 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., using a simple local maxima technique in the Radon transform domain, associated with known clinical definitions of line artifacts.[36] 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 deep learning may ultimately yield more reliable and objective results providing information for 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 efficacy of treatment. Improved monitoring might prevent heart failure re-admissions, and decrease healthcare system cost.

Reliability of rating is essential for any diagnostic test. Anderson et al. revealed substantial inter-rater 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.[37] Gullett et al. 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.[38] The intra-rater agreements of 0.89 and 0.87 in our 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 re-run 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 deep learning 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 is 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 dataset 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 dataset.

Deep learning algorithms require substantial data for training, often thousands to tens of thousands of separate data points. To feasibly produce such a rich dataset 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 the 400 was key for improving the performance of the deep learning software, which was addressed by splitting the data into sub-clips, as well as the use of data augmentation methods. As noted in the methods section, sub-clips from the same patient had different characteristics due to respiration induced dynamic motion resulting in different appearance of the ultrasound image across the sub-clips. Thus, while not strictly independent clips, these sub-clips were different enough to be treated as independent data.

One limitation of the dataset affecting generalizability of prediction results is the fact that the learning data comes 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 applicability to real-world data.

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

In the future, further training the algorithm on whole clips rather than collections 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 deep learning network that operated on dynamic ultrasound data for automated assessment of sonographic lung B-lines. The network was developed using 2415 sub-clips of 12 frames each extracted from 400 patients using expert consensus ratings as ground truth. This deep learning algorithm showed promise for automated assessment for both the binary and severity ratings

based on a test data of 100 sub-clips that were set aside and not used during the training. This method could be used to improve reliability and objectivity of presence and severity of B-lines for diagnosis and prognosis of patients with respiratory complaints. Clinical applications for this include heart failure, pneumonia, and ARDS. In addition, reliable identification of B-lines may aid in diagnosis and management of the ongoing COVID-19 pandemic.

REFERENCES AND FOOTNOTES

1. McCaig LF, Burt CW. National Hospital Ambulatory Medical Care Survey: 1999 emergency department summary. *Adv Data* 2001(320):1-34. [published Online First: 2003/04/02]
2. Al Deeb M, Barbic S, Featherstone R, et al. Point-of-care ultrasonography for the diagnosis of acute cardiogenic pulmonary edema in patients presenting with acute dyspnea: a systematic review and meta-analysis. *Acad Emerg Med* 2014;21(8):843-52. doi: 10.1111/acem.12435 [published Online First: 2014/09/02]
3. Laribi S, Keijzers G, van Meer O, et al. Epidemiology of patients presenting with dyspnea to emergency departments in Europe and the Asia-Pacific region. *Eur J Emerg Med* 2019;26(5):345-49. doi: 10.1097/MEJ.0000000000000571 [published Online First: 2018/09/01]
4. Moore CL, Copel JA. Point-of-care ultrasonography. *N Engl J Med* 2011;364(8):749-57. doi: 10.1056/NEJMra0909487 [published Online First: 2011/02/25]
5. Lichtenstein D, Meziere G. A lung ultrasound sign allowing bedside distinction between pulmonary edema and COPD: the comet-tail artifact. *Intensive Care Med* 1998;24(12):1331-4. doi: 10.1007/s001340050771 [published Online First: 1999/01/14]
6. Lichtenstein D, Meziere G, Biderman P, et al. The comet-tail artifact. An ultrasound sign of alveolar-interstitial syndrome. *Am J Respir Crit Care Med* 1997;156(5):1640-6. doi: 10.1164/ajrccm.156.5.96-07096 [published Online First: 1997/12/31]
7. Wang S, Liu Y, Zhang Y, et al. A preliminary study on the ultrasonic manifestations of peripulmonary lesions of noncritical novel coronavirus pneumonia (COVID-19), 2020.
8. Poggiali E, Dacrema A, Bastoni D, et al. Can Lung US Help Critical Care Clinicians in the Early Diagnosis of Novel
Coronavirus (COVID-19) Pneumonia? *Radiology* 2020 [published Online First: 03/13/2020]
9. Soldati G, Smargiassi A, Inchingolo R, 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* 2020 doi: 10.1002/jum.15285 [published Online First: 2020/04/01]
10. Yue Lee FC, Jenssen C, Dietrich CF. A common misunderstanding in lung ultrasound: the comet tail artefact. *Med Ultrason* 2018;20(3):379-84. doi: 10.11152/mu-1573 [published Online First: 2018/09/01]
11. Volpicelli G, Mussa A, Garofalo G, et al. Bedside lung ultrasound in the assessment of alveolar-interstitial syndrome. *Am J Emerg Med* 2006;24(6):689-96. doi: 10.1016/j.ajem.2006.02.013 [published Online First: 2006/09/21]
12. Coiro S, Porot G, Rossignol P, et al. Prognostic value of pulmonary congestion assessed by lung ultrasound imaging during heart failure hospitalisation: A two-centre cohort study. *Sci Rep* 2016;6:39426. doi: 10.1038/srep39426 [published Online First: 2016/12/21]
13. Frassi F, Gargani L, Tesorio P, et al. Prognostic value of extravascular lung water assessed with ultrasound lung comets by chest sonography in patients with dyspnea and/or chest pain. *J Card Fail* 2007;13(10):830-5. doi: 10.1016/j.cardfail.2007.07.003 [published Online First: 2007/12/11]
14. Gargani L, Pang PS, Frassi F, et al. Persistent pulmonary congestion before discharge predicts rehospitalization in heart failure: a lung ultrasound study. *Cardiovasc Ultrasound* 2015;13:40. doi: 10.1186/s12947-015-0033-4 [published Online First: 2015/09/05]
15. Platz E, Lewis EF, Uno H, et al. Detection and prognostic value of pulmonary congestion by lung ultrasound in ambulatory heart failure patients. *Eur Heart J* 2016;37(15):1244-51. doi: 10.1093/eurheartj/ehv745 [published Online First: 2016/01/29]
16. Zoccali C, Torino C, Tripepi R, et al. Pulmonary congestion predicts cardiac events and mortality in ESRD. *J Am Soc Nephrol* 2013;24(4):639-46. doi: 10.1681/ASN.2012100990 [published Online First: 2013/03/02]
17. Platz E, Merz AA, Jhund PS, et al. Dynamic changes and prognostic value of pulmonary congestion by lung ultrasound in acute and chronic heart failure: a systematic review. *Eur J Heart Fail* 2017;19(9):1154-63. doi: 10.1002/ejhf.839 [published Online First: 2017/05/31]
18. Tierney DM, Boland LL, Overgaard JD, 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* 2018;46(1):14-22. doi: 10.1002/jcu.22526 [published Online First: 2017/10/07]
19. Brattain LJ, Telfer BA, Dhyani M, et al. Machine learning for medical ultrasound: status, methods, and future opportunities. *Abdom Radiol (NY)* 2018;43(4):786-99. doi: 10.1007/s00261-018-1517-0 [published Online First: 2018/03/02]
20. Erickson BJ, Korfiatis P, Akkus Z, et al. Machine Learning for Medical Imaging. *Radiographics* 2017;37(2):505-15. doi: 10.1148/rg.2017160130 [published Online First: 2017/02/18]
21. Shokoohi H, LeSaux MA, Roohani YH, et al. Enhanced Point-of-Care Ultrasound Applications by Integrating Automated Feature-Learning Systems Using Deep Learning. *J Ultrasound Med* 2019;38(7):1887-97. doi: 10.1002/jum.14860 [published Online First: 2018/11/15]
22. Corradi F, Via G, Forfori F, et al. Lung ultrasound and B-lines quantification inaccuracy: B sure to have the right solution. *Intensive Care Med* 2020 doi: 10.1007/s00134-020-06005-6 [published Online First: 2020/03/20]
23. Liu S, Yang X, Lei B, et al. Deep Learning in Medical Ultrasound Analysis: A Review. *Engineering* 2019;5(2):11. doi: https://doi.org/10.1016/j.eng.2018.11.020
24. Wong KKL, Wang L, Wang D. Recent developments in machine learning for medical imaging applications. *Comput Med Imaging Graph* 2017;57:1-3. doi: 10.1016/j.compmedimag.2017.04.001 [published Online First: 2017/05/01]
25. Fu GS, Levin-Schwartz Y, Lin QH, et al. Machine Learning for Medical Imaging. *J Healthc Eng* 2019;2019:9874591. doi: 10.1155/2019/9874591 [published Online First: 2019/06/12]
26. Dietrich CF, Mathis G, Blaivas M, et al. Lung B-line artefacts and their use. *J Thorac Dis* 2016;8(6):1356-65. doi: 10.21037/jtd.2016.04.55 [published Online First: 2016/06/14]
27. Lichtenstein DA. Current Misconceptions in Lung Ultrasound: A Short Guide for Experts. *Chest* 2019;156(1):21-25. doi: 10.1016/j.chest.2019.02.332 [published Online First: 2019/03/16]
28. Carreira J, Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017:4724-33.
29. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Res Social Adm Pharm* 2013;9(3):330-8. doi: 10.1016/j.sapharm.2012.04.004 [published Online First: 2012/06/15]
30. Brattain LJ, Telfer BA, Liteplo AS, et al. Automated B-line scoring on thoracic sonography. *J Ultrasound Med* 2013;32(12):2185-90. doi: 10.7863/ultra.32.12.2185 [published Online First: 2013/11/28]
31. Weitzel WF, Hamilton J, Wang X, et al. Quantitative lung ultrasound comet measurement: method and initial clinical results. *Blood Purif* 2015;39(1-3):37-44. doi: 10.1159/000368973 [published Online First: 2015/02/11]
32. van Sloun RJ, Demi L. Deep learning for automated detection of B-lines in lung ultrasonography. *The Journal of the Acoustical Society of America* 2018;144:1668.
33. van Sloun RJ, Demi L. Localizing B-lines in Lung Ultrasonography by Weakly-Supervised Deep Learning, in-vivo results. *IEEE J Biomed Health Inform* 2019 doi: 10.1109/JBHI.2019.2936151 [published Online First: 2019/08/20]
34. Moshavegh R, Hansen KL, Moller-Sorensen H, et al. Automatic Detection of B-Lines in In Vivo Lung Ultrasound. *IEEE Trans Ultrason Ferroelectr Freq Control* 2019;66(2):309-17. doi: 10.1109/TUFFC.2018.2885955 [published Online First: 2018/12/12]
35. Kulhare S, Zheng X, Mehanian C, et al. Ultrasound-Based Detection of Lung Abnormalities Using Single Shot Detection Convolutional

Neural Networks: International Workshops, POCUS 2018, BIVPCS 2018, CuRIOUS 2018, and CPM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16–20, 2018, Proceedings. Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation2018:65-73.

36. Anantrasirichai N, Hayes W, Allinovi M, et al. Line Detection as an Inverse Problem: Application to Lung Ultrasound Imaging. *IEEE Trans Med Imaging* 2017;36(10):2045-56. doi: 10.1109/TMI.2017.2715880 [published Online First: 2017/07/07]
37. Anderson KL, Fields JM, Panebianco NL, et al. Inter-rater reliability of quantifying pleural B-lines using multiple counting methods. *J Ultrasound Med* 2013;32(1):115-20. doi: 10.7863/jum.2013.32.1.115 [published Online First: 2012/12/28]
38. Gullett J, Donnelly JP, Sinert R, et al. Interobserver agreement in the evaluation of B-lines using bedside ultrasound. *J Crit Care* 2015;30(6):1395-9. doi: 10.1016/j.jcrc.2015.08.021 [published Online First: 2015/09/26]