PROCEEDINGS OF SPIE

Multi-stage investigation of deep neural networks for COVID-19 B-line feature detection in simulated and in vivo ultrasound images

Benjamin Frey, Lingyi Zhao, Tiffany Clair Fong, Muyinatu Lediju Bell

SPIE.

Multi-stage investigation of deep neural networks for COVID-19 B-line feature detection in simulated and *in vivo* ultrasound images

Benjamin Frey[a], Lingyi Zhao[b], Tiffany Clair Fong[c], and Muyinatu A. Lediju Bell[b,d,e]

[a]Department of Physics, University of St. Thomas, Saint Paul, Minnesota 55105-1094, USA
[b]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA
[c]Department of Emergency Medicine, Johns Hopkins Medicine, Baltimore, Maryland 21287, USA
[d]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA
[e]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218, USA

ABSTRACT

COVID-19 is a highly infectious disease with high morbidity and mortality, requiring tools to support rapid triage and risk stratification. In response, deep learning has demonstrated great potential to quicklyand autonomously detect COVID-19 features in lung ultrasound B-mode images. However, no previous work considers the application of these deep learning models to signal processing stages that occur prior to traditional ultrasound B-mode image formation. Considering the multiple signal processing stages required to achieve ultrasound B-mode images, our research objective is to investigate the most appropriate stage for our deep learning approach to COVID-19 B-line feature detection, starting with raw channel data received by an ultrasound transducer. Results demonstrate that for our given training and testing configuration, the maximum Dice similarity coefficient (DSC) was produced by B-mode images (DSC = 0.996) when compared with three alternative image formation stages that can serve as network inputs: (1) raw in-phase and quadrature (IQ) data before beamforming, (2) beamformed IQ data, (3) envelope detected IQ data. The best-performing simulation-trained network was tested on *in vivo* B-mode images of COVID-19 patients, ultimately achieving 76% accuracy to detect the same (82% of cases) or more (18% of cases) B-line features when compared to B-line feature detection by human observers interpreting B-mode images. Results are promising to proceed with future COVID-19 B-line feature detection using ultrasound B-mode images as the input to deep learning models.

Keywords: COVID-19, deep learning, lung ultrasound imaging, segmentation, B-line

1. INTRODUCTION

COVID-19 is an infectious disease caused by the SARS-CoV-2 virus. The disease was first identified in December 2019 and was declared by the World Health Organization as a pandemic in March 2020.[1] After the initial identification of COVID-19, multiple testing techniques have been employed to diagnose the disease including the reverse transcriptase quantitative polymerase chain reaction (RT-qPCR),[2] chest x-ray,[3] chest computed tomography (CT),[4] and lung ultrasound.[5,6] While RT-qPCR tests are considered accurate,[2] they require of biomolecular processing facilities, associated sample processing logistics, and an average 24-hour delay between test administration and result delivery.[7] Chest x-rays and chest CT had been used during times of surge and/or resource limited settings when access to RT-qPCR testing was limited. Both chest x-rays and chest CT have utility to identify and quantify pulmonary involvement, which may predict a poorer prognosis in patients with COVID infection. Although chest x-rays are a low-cost imaging option deemed as a first-line diagnosis tool

---

*Address all correspondence to Lingyi Zhao, lzhao45@jhu.edu; Muyinatu A. Lediju Bell, mledijubell@jhu.edu

in urgent care centers, the procedure has been reported to have poor sensitivity.[8] While chest CT is a higher-sensitivity imaging option that is recommended for hospitalized, symptomatic COVID-19 patients,[9] associated equipment is traditionally immobile, requiring patients to be relocated and the equipment to be disinfected.[10] In addition, chest x-rays and chest CTs require ionizing radiation.

Ultrasound is a mobile imaging option that does not require ionizing radiation, allowing operators to safely perform tests at the location of the patient, thus decreasing patient-to-patient exposure. A limitation of lung ultrasound is its operator dependence and the potential for misdiagnosis especially among inexperienced users. When imaging lungs, possible features that can be detected include A-lines[11] (i.e., horizontal reverberation artifacts of pleura) and B-lines[12] (i.e., well-defined vertical lines that arise from the pleural line and extend indefinitely through the lung tissue). When using lung ultrasound images to determine the presence of COVID-19, the most frequent abnormality is interstitial involvement depicted as three or more B-lines.[13] Multiple groups have demonstrated that deep neural networks (DNNs) can be employed for automated and rapid real-time detection of B-line features,[14–17] as summarized in a recent review article.[18] These demonstrations promise to speed up patient throughput in high-volume clinics and to assist less experienced users with feature detection.

There are multiple possible stages of ultrasound image formation inputs for feature detection with a DNN. While previous work uses B-mode images, our group has demonstrated that raw channel data contains richer information content for learning features of interest from ultrasound images (e.g., anechoic or hypoechoic cysts).[19–21] The objective of this work is to determine which stage of image formation is most appropriate for B-line feature detection from lung ultrasound images when utilizing a modified version of our previously reported deep learning architecture.[19]

In this paper, we investigate four possible processing stages in ultrasound imaging as the input to our U-Net DNN: (1) raw in-phase and quadrature (IQ) data before beamforming, (2) beamformed IQ data, (3) envelope detected IQ data, and (4) B-mode images. We describe our simulation training and testing parameters for each image formation stage together with our method for quantifying DNN performance. In addition, we demonstrate the performance of our simulation-trained DNNs on real *in vivo* B-mode images from COVID-19 patients.

2. METHODS

2.1 Training and testing with simulated data

After simulating raw channel data containing B-line features with the MATLAB Ultrasound Toolbox,[22] the resulting data were processed to yield the following four stages: (1) raw IQ data (2) beamformed IQ data prior to envelope detection, (3) envelope detected IQ data, and (4) B-mode images. These data were utilized to test
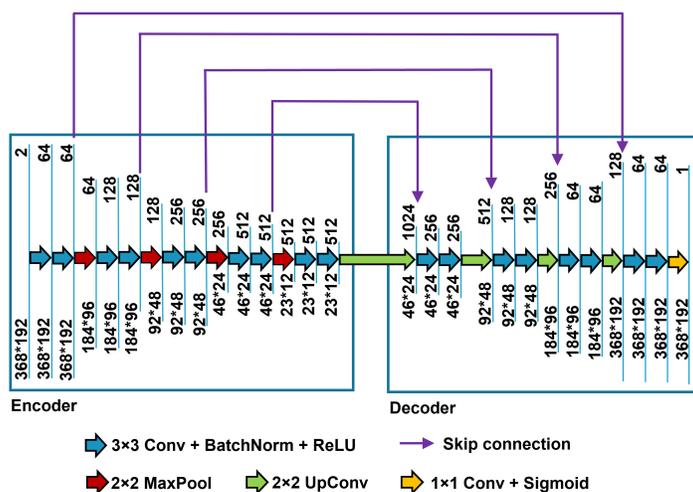
Figure 1. Deep learning architecture utilized with raw or beamformed IQ data as the input and with 10 cm imaging depth.

Table 1. Training and testing parameters for DNN models

Training Parameters	Testing Parameter	Value	Variation in Result
DNNs 1-20: - 10 cm imaging depth - B-line amplitudes: 0.2, 0.4, 0.6, 0.8, 1.0 (5 amplitudes total) - 5 DNNs per stage	Stage	1, 2, 3, 4	Yes (Stage 4 is generally best)
	B-line amplitude	0.2, 0.4, 0.6, 0.8, 1.0	No
	Imaging depth	8 cm, 10 cm, 12 cm, 14 cm, 16 cm	Yes
DNNs 21-24: - 16 cm imaging depth - B-line amplitude of 1 - One DNN per stage	Stage	1, 2, 3, 4	Yes (Stage 4 is generally best)
	Imaging depth	8 cm, 10 cm, 12 cm, 14 cm, 16 cm	Yes

and train DNNs that were derived from a modified version of the deep learning architecture previously reported by our group,[19] which is based on U-Net architecture.[23] Instead of using two decoders (one for segmentation, one for reconstruction) as in the previously reported deep learning architecture,[19] we only used one decoder for segmentation. We also changed the input and output dimensions to match the imaging depths (10 cm or 16 cm) and probe parameters in our simulation. Fig. 1 shows a deep learning architecture utilized with raw IQ data or beamformed IQ data as the input when the imaging depth was 10 cm. For other processed stages and imaging depths, the input and output dimensions were modified accordingly, and the remainder of the DNN architecture was unchanged.

A total of 24 DNNs were trained, as detailed in Table 1, with each reported training parameter being fixed at one value per network. For each DNN, the training set included 8,000 samples and the testing set included 2,000 samples, with half of the testing set generated by flipping the simulated images along the center lateral image dimension. Each DNN was trained for 80 epochs using the Adam optimizer[24] with a learning rate of 1e-5. The training loss was the Dice similarity coefficient (DSC) loss as described in previous work.[19] The DSC score for each test set was calculated to measure the performance of the DNN.

To investigate the effect of B-line amplitude, DNNs 1-20 in Table 1 were each tested using the testing set containing the same B-line amplitude as the corresponding training set and with the image depth fixed at 10 cm. To investigate the effect of imaging depth, DNNs 21-24 were each tested as the image depth was independently varied from 8 cm to 16 cm, as reported in Table 1. In addition, four of the networks in DNNs 1-20 (i.e., trained with 10 cm image depth and B-line amplitude of 1 for each image formation stage) were tested as the image depth was independently varied from 8 cm to 16 cm.

2.2 Testing with *in vivo* data

To test our simulation-trained DNNs on *in vivo* images of COVID-19 patients, B-mode lung ultrasound videos were first sourced from an online lung ultrasound repository used in previous work by Born *et al.*[14] Of the 44 videos indicated in this repository as belonging to confirmed COVID-19 patients, 51 still image frames were selected for testing. These image selections represent a variety of test-case scenarios, ranging from no B-lines present in the B-mode image to multiple B-lines present. These B-line features were manually labeled to generate the ground truth for each sample in our 51-image *in vivo* dataset.

This *in vivo* dataset was tested with the DNN trained on B-mode images containing a B-line amplitude of 1 with 10 cm image depth. After obtaining segmentation maps from the DNN, a custom computer vision algorithm was applied to the predicted segmentation maps to remove apparent false positives, such as segmentation above the pleural line, segmentations along the B-mode image border beyond the pleural line, and segmentation noise

Table 2. Binary Classification definitions

Binary classification	Definition
TP	At or above the DSC threshold, or when DSC equals 0, the B-line number counted with reference to the DNN segmentation was at least the B-line number counted without any reference
FP	At or above the DSC threshold, or when DSC equals 0, the B-line number counted with reference to the DNN segmentation was less than the B-line number counted without any reference
TN	Below the DSC threshold, the B-line number counted with reference to the DNN segmentation was less than the B-line number counted without any reference
FN	Below the DSC threshold, the B-line number counted with reference to the DNN segmentation was at least the B-line number counted without any reference

beyond the pleural line in dark regions of the B-mode image. The number of B-lines was counted with and without referring to these post-processed segmentation maps. The accuracy of the DNN prediction was defined as the percentage of cases where the number of B-lines present counted with reference to DNN segmentation maps was equivalent to or greater than the number of B-lines counted without this reference. While the human observer annotations for DNN segmentation were provided by the first author of this manuscript (B.F.), the DNN results chosen for display in this manuscript were confirmed to be consistent with clinical practice after professional consultation with co-author T.F.

To further quantify DNN performance on the *in vivo* dataset, we developed a binary classification system using DSC score as a discrimination threshold. The definition of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) in this binary classification system is outlined in Table 2. This binary classification system was used to produce the receiver operating characteristics (ROC) curves by first counting the number of B-lines in each *in vivo* image without referring to the DNN segmentation maps. The DNN trained for either 1 epoch or 80 epochs was then utilized to generate segmentation maps for the 51 *in vivo* images described above. The post-processed segmentation maps were then overlaid on the original image as a reference and the number of B-lines present were recounted. Each sample was then classified as TP, FP, TN, or FN using the definitions outlined in Table 2. The true positive rate (TPR) and false positive rate (FPR) for each sample was calculated using the following equations:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2}$$

Each result from Eqs. (1) and (2) was plotted with coordinates (FPR, TPR) to develop two ROC curves (i.e., one for the DNN trained for 1 epoch, another for the DNN trained for 80 epochs ROC curve). The area under the curve (AUC) for each ROC curve was then calculated using a left-endpoint Riemann sum.

3. RESULTS

3.1 DSC scores when testing with simulated data

Fig. 2 shows plots of test DSC scores for the simulation results. Fig. 2(a) plots DSC as a function of epoch number. Fig. 2(b) plots DSC as a function of B-line amplitude. DSC as a function of the imaging depth of test sets for DNNs trained with 10 cm-depth images and 16 cm-depth images are plotted in Figs. 2(c) and 2(d), respectively. These results are reported for each of the four image formation stages, and they demonstrate that
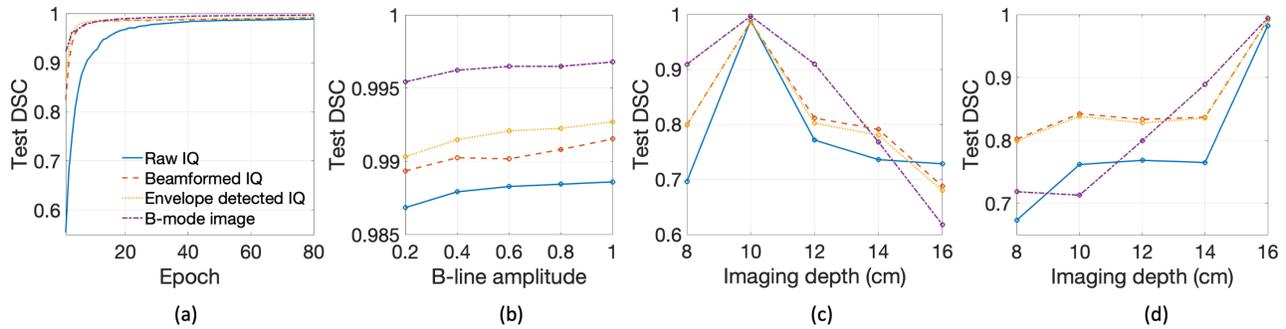
Figure 2. Test DSC scores obtained for each image formation stage (a) as a function of the number of training epochs investigated with a fixed imaging depth of 10 cm and a B-line amplitude of 1, (b) as a function of B-line amplitude with a fixed imaging depth of 10 cm and a fixed training epoch of 80, and as functions of the test image depth with a fixed B-line amplitude of 1, a fixed training epoch of 80, and fixed training image depths of (c) 10 cm and (d) 16 cm.
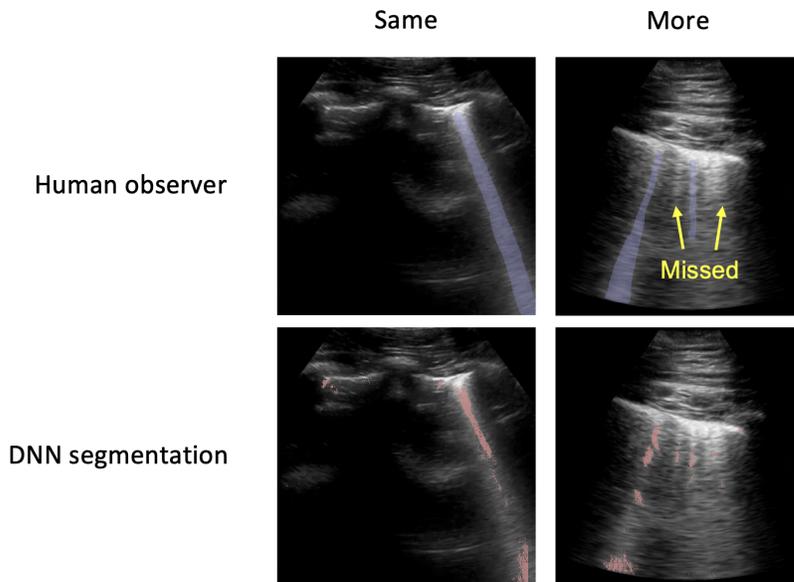
Figure 3. Example cases where the DNN trained for 80 epochs detected the same or more B-line features than the human observer. The ground truth segmentations (blue) and post-processed DNN segmentations (red) were overlaid on input B-mode images, with example results demonstrating (left) the DNN segmentation detecting the same B-line count of 1 as the human observer, and (right) the DNN segmentation detecting more B-lines than the human observer.

the greatest test DSC scores were generally produced by B-mode images and when the training and testing image depths were the same. DSC scores (obtained at epoch 80) were 0.1-0.25 lower when imaging depths were $\pm$ 2 cm from the imaging depth used during training.

3.2 Translation to B-mode images of COVID-19 patients

Fig. 3 shows two example cases where the DNN trained for 80 epochs on simulated B-mode images containing a B-line amplitude of 1 with 10 cm image depth detected the same or more B-line features than a human observer, with segmentations from the human observer indicated in blue and DNN segmentations indicated in red.

Out of 51 cases, the DNN trained for 80 epochs detected the same or more B-line features than the human observer in 39 cases and detected no B-line features in 6 cases for which the human observer detected at least one B-line feature. Overall, the DNN trained on simulated B-mode images (with B-line amplitude of 1 and image
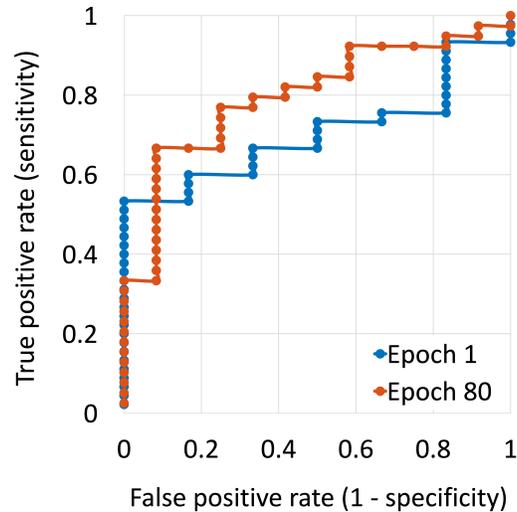
Figure 4. The ROC curves using DSC as a discrimination threshold.

depth of 10 cm) for 80 epochs achieved 76% accuracy to detect the same (82% of cases) or more (18% of cases) B-line features when compared to a human observer.

Fig. 4 shows the ROC curves for the binary classification system outlined in Table 2, reported for the DNNs trained with 1 and 80 epochs. The AUCs determined from these curves are 0.70 when testing with the DNN trained for 1 epoch and 0.80 when testing with a DNN trained for 80 epochs. These results indicate that the DNN trained with 80 epochs is more efficient when detecting a number of B-line features that is equal to or greater than that detected by human observers (when using DSC score as a discrimination threshold). The ideal operating point is a DSC threshold of 0.36 and 0.12 for the DNNs trained with 1 and 80 epochs, respectively.

4. DISCUSSION

This work is the first to investigate the performance of multiple image formation stages when detecting B-line features in the lung ultrasound images of COVID-19 patients. Although the presented U-Net implementation was previously more successful at identifying cyst-like features in raw IQ ultrasound channel data than in ultrasound B-mode images,[19] this same finding was not achieved when detecting B-line features. Instead, when identifying B-line features with our U-Net implementation, the B mode stage was determined to be the most efficient processing stage among the four image formation stages investigated (i.e., raw IQ data, beamformed IQ data, envelope detected IQ data, and B-mode images). This finding was obtained with a systematic simulation study that incorporated multiple imaging parameters (e.g., feature amplitude, imaging depth). One possible reason for the deviation of this result from previous results obtaining good cyst segmentation performance on raw data[19] is that cysts are defined structures, whereas B-lines represent an image artifact from tissue, rather than a feature in tissue. Therefore, the appearance of this feature in raw channel data may be less well-defined. Future work will investigate suitable network inputs for other lung ultrasound features of the COVID-19 disease.

Overall, we achieved an important insight that allows us to proceed with testing our DNNs on B-mode images from COVID-19 patients. Considering that the DNNs detected the same or more number of B-lines than human observers in 88% of the examples, this result highlights a benefit of our DNN approach to assist less experienced physicians with identifying B-line features for COVID-19 detection and diagnosis. In the future, we will use a similar approach to proceed with identification of other findings of COVID-19, such as pleural line abnormalities and sub-pleural consolidations. Finally, our novel binary classification system provides a new DNN benchmark and evaluation method for future investigations.

5. CONCLUSION

When identifying B-line features with our U-Net implementation, we demonstrated that the greatest DSC values were generally produced with B-mode images. The trained DNNs also performed best when tested on imaging depths that were used during the training process, and DSC values were 0.1-0.25 lower when imaging depths were within $\pm$ 2 cm from the training image depth. These DNNs, which were trained on simulated B-mode images, successfully identified B-line features when tested on *in vivo* images of COVID-19 patients. Results are sufficiently insightful to proceed with future COVID-19 B-line feature detection using B-mode images, with promising implications for assisting less experienced physicians with identifying B-line features for COVID-19 detection and diagnosis.

ACKNOWLEDGMENTS

REFERENCES

[1] Ghebreyesus, T., "WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020," (2020). <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020> (accessed 5 December 2021).

[2] Wang, W., Xu, Y., Gao, R., Lu, R., Han, K., Wu, G., and Tan, W., "Detection of SARS-CoV-2 in Different Types of Clinical Specimens," *JAMA* **323**(18), 1843–1844 (2020).

[3] Ng, M.-Y., Lee, E. Y., Yang, J., Yang, F., Li, X., Wang, H., Lui, M. M.-s., Lo, C. S.-Y., Leung, B., Khong, P.-L., et al., "Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review," *Radiology: Cardiothoracic Imaging* **2**(1), e200034 (2020).

[4] Bernheim, A., Mei, X., Huang, M., Yang, Y., Fayad, Z. A., Zhang, N., Diao, K., Lin, B., Zhu, X., Li, K., et al., "Chest CT Findings in Coronavirus Disease 2019 (COVID-19): Relationship to Duration of Infection," *Radiology* , 200463 (2020).

[5] Buda, N., Segura-Grau, E., Cylwik, J., and Wełnicki, M., "Lung ultrasound in the diagnosis of COVID-19 infection-A case series and review of the literature," *Advances in Medical Sciences* **65**(2), 378–385 (2020).

[6] Soldati, G., Smargiassi, A., Inchingolo, R., Buonsenso, D., Perrone, T., Briganti, D. F., Perlini, S., Torri, E., Mariani, A., Mossolani, E. E., et al., "Is There a Role for Lung Ultrasound During the COVID-19 Pandemic?," *Journal of Ultrasound in Medicine* (2020).

[7] "Testing Information and News for Individuals, Healthcare Providers, Payers, and Organizations," (2020). <https://www.labcorp.com/coronavirus-disease-COVID-19> (accessed 5 December 2021).

[8] Weinstock, M. B., Echenique, A., Russell, J., Leib, A., Miller, J., Cohen, D., Waite, S., Frye, A., and Illuzzi, F., "Chest X-Ray Findings in 636 Ambulatory Patients with COVID-19 Presenting to an Urgent Care Center: A Normal Chest X-Ray Is no Guarantee," *Journal of Urgent Care Medicine* **14**(7), 13–18 (2020).

[9] "ACR Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection," (2020). <https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection> (accessed 5 December 2021).

[10] Qu, J., Yang, W., Yang, Y., Qin, L., and Yan, F., "Infection Control for CT Equipment and Radiographers' Personal Protection During the Coronavirus Disease (COVID-19) Outbreak in China," *American Journal of Roentgenology* **215**(4), 940–944 (2020).

[11] Demi, M., Prediletto, R., Soldati, G., and Demi, L., "Physical Mechanisms Providing Clinical Information From Ultrasound Lung Images: Hypotheses and Early Confirmations," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **67**(3), 612–623 (2019).

[12] Soldati, G., Demi, M., Smargiassi, A., Inchingolo, R., and Demi, L., "The role of ultrasound lung artifacts in the diagnosis of respiratory diseases," *Expert Review of Respiratory Medicine* **13**(2), 163–172 (2019).

[13] Mohamed, M. F., Al-Shokri, S., Yousaf, Z., Danjuma, M., Parambil, J., Mohamed, S., Mubasher, M., Dauleh, M. M., Hasanain, B., AlKahlout, M. A., et al., "Frequency of abnormalities detected by point-of-care lung ultrasound in symptomatic COVID-19 patients: systematic review and meta-analysis," *The American Journal of Tropical Medicine and Hygiene* **103**(2), 815 (2020).

[14] Born, J., Wiedemann, N., Cossio, M., Buhre, C., Brändle, G., Leidermann, K., Aujayeb, A., Moor, M., Rieck, B., and Borgwardt, K., "Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis," *Applied Sciences* **11**, 672 (Jan 2021).

[15] Baum, Z. M., Bonmati, E., Cristoni, L., Walden, A., Prados, F., Kanber, B., Barratt, D. C., Hawkes, D. J., Parker, G. J., Wheeler-Kingshott, C. A. G., et al., "Image quality assessment for closed-loop computer-assisted lung ultrasound," in [*Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*], **11598**, 115980R, International Society for Optics and Photonics (2021).

[16] Roy, S., Menapace, W., Oei, S., Luijten, B., Fini, E., Saltori, C., Huijben, I., Chennakeshava, N., Mento, F., Sentelli, A., et al., "Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound," *IEEE Transactions on Medical Imaging* **39**(8), 2676–2687 (2020).

[17] Hu, Z., Liu, Z., Dong, Y., Liu, J., Huang, B., Liu, A., Huang, J., Pu, X., Shi, X., Yu, J., et al., "Evaluation of lung involvement in COVID-19 pneumonia based on ultrasound images," *BioMedical Engineering OnLine* **20**(1), 1–15 (2021).

[18] Zhao, L. and Bell, M. A. L., "A review of deep learning applications in lung ultrasound imaging of COVID-19 patients," *BME Frontiers* (2022).

[19] Nair, A. A., Washington, K. N., Tran, T. D., Reiter, A., and Bell, M. A. L., "Deep Learning to Obtain Simultaneous Image and Segmentation Outputs From a Single Input of Raw Ultrasound Channel Data," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **67**(12), 2493–2509 (2020).

[20] Bhatt, M., Nair, A. A., Kempski, K. M., and Bell, M. A. L., "Multi-task learning for ultrasound image formation and segmentation directly from raw in vivo data," in [*2020 IEEE International Ultrasonics Symposium (IUS)*], 1–4, IEEE (2020).

[21] Wiacek, A., González, E., and Bell, M. A. L., "Coherenet: A Deep Learning Architecture for Ultrasound Spatial Correlation Estimation and Coherence-Based Beamforming," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **67**(12), 2574–2583 (2020).

[22] Shahriari, S. and Garcia, D., "Meshfree simulations of ultrasound vector flow imaging using smoothed particle hydrodynamics," *Physics in Medicine and Biology* **63**(20), 205011 (2018).

[23] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional Networks for Biomedical Image Segmentation," in [*International Conference on Medical image computing and computer-assisted intervention*], 234–241, Springer (2015).

[24] Kingma, D. P. and Ba, J., "Adam: A Method for Stochastic Optimization," in [*3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*], Bengio, Y. and LeCun, Y., eds. (2015).