Ultrasonics 140 (2024) 107251

Contents lists available at ScienceDirect

Ultrasonics

journal homepage: www.elsevier.com/locate/ultras

Deep learning for real-time multi-class segmentation of artefacts in lung ultrasound

Lewis Howell [a,b], Nicola Ingram [c], Roger Lapham [d], Adam Morrell [d], James R. McLaughlan [b,c,*]

[a] School of Computing, University of Leeds, Leeds, LS2 9JT, UK
[b] School of Electronic and Electrical Engineering, University of Leeds, Leeds, LS2 9JT, UK
[c] Leeds Institute of Medical Research, University of Leeds, St James' University Hospital, Leeds, LS9 7TF, UK
[d] Radiology Department, Leeds Teaching Hospital Trust, Leeds General Infirmary, Leeds, LS1 3EX, UK

ARTICLE INFO

Dataset link: https://doi.org/10.5518/1485, https://github.com/ljhowell/LUS-Segmentation-RT

Keywords:
Lung ultrasound
Deep learning
Artificial intelligence
Image segmentation
Point of care ultrasound
COVID-19

ABSTRACT

Lung ultrasound (LUS) has emerged as a safe and cost-effective modality for assessing lung health, particularly during the COVID-19 pandemic. However, interpreting LUS images remains challenging due to its reliance on artefacts, leading to operator variability and limiting its practical uptake. To address this, we propose a deep learning pipeline for multi-class segmentation of objects (ribs, pleural line) and artefacts (A-lines, B-lines, B-line confluence) in ultrasound images of a lung training phantom. Lightweight models achieved a mean Dice Similarity Coefficient (DSC) of 0.74, requiring fewer than 500 training images. Applying this method in real-time, at up to 33.4 frames per second in inference, allows enhanced visualisation of these features in LUS images. This could be useful in providing LUS training and helping to address the skill gap. Moreover, the segmentation masks obtained from this model enable the development of explainable measures of disease severity, which have the potential to assist in the triage and management of patients. We suggest one such semi-quantitative measure called the B-line Artefact Score, which is related to the percentage of an intercostal space occupied by B-lines and in turn may be associated with the severity of a number of lung conditions. Moreover, we show how transfer learning could be used to train models for small datasets of clinical LUS images, identifying pathologies such as simple pleural effusions and lung consolidation with DSC values of 0.48 and 0.32 respectively. Finally, we demonstrate how such DL models could be translated into clinical practice, implementing the phantom model alongside a portable point-of-care ultrasound system, facilitating bedside assessment and improving the accessibility of LUS.

1. Introduction

Artificial intelligence (AI) has significant potential in the healthcare industry, where data could be used to reduce costs, whilst improving patient care [1]. Early uses of AI in healthcare were aimed at supporting clinical decisions [2], before rapidly expanding into a wide range of applications, such as management and interpretation of patient data, predictive medicine, and allocation of health service provision [3]. Deep learning (DL), a subset of AI and machine learning (ML), uses hierarchical combinations of learnable feature extractors to automatically build high-level representations of data [4]. The most common DL techniques for vision problems are based on convolutional neural networks (CNNs), such as the U-Net architecture [5], which are often used for segmenting biomedical images. Recently, variants of the original U-Net, such as those which incorporate attention gates or vision transformers have been demonstrated to achieve positive results in semantic image segmentation tasks, where each pixel in an image is assigned a class [6–8]. Additionally, transfer learning may help DL models to achieve better performance when there are limited images in a dataset [9,10], leveraging the knowledge gained by a model trained on one dataset to inform the training of a model on a related dataset [11]. In the medical imaging domain, semantic segmentation networks have mainly been used with radiography datasets from magnetic resonance imaging (MRI) or X-ray Computed Tomography (CT) scanners, with significantly fewer from ultrasound [12]. The primary function of these networks is to efficiently segment features to aid in diagnosis, prognosis, or assist with image-guided therapies and/or interventions.

Lung ultrasound (LUS) is a technique that can be used to investigate the health of a patient's lungs in a range of clinical settings [13–15]. LUS is straightforward to perform but requires experienced ultrasound practitioners due to a steep learning curve for correct image interpretation [16–18]. This is because, unlike most soft tissue imaging with

[*] Corresponding author at: School of Electronic and Electrical Engineering, University of Leeds, Leeds, LS2 9JT, UK.
E-mail address: j.r.mclaughlan@leeds.ac.uk (J.R. McLaughlan).

https://doi.org/10.1016/j.ultras.2024.107251
Received 28 June 2023; Received in revised form 20 December 2023; Accepted 17 January 2024
Available online 29 January 2024

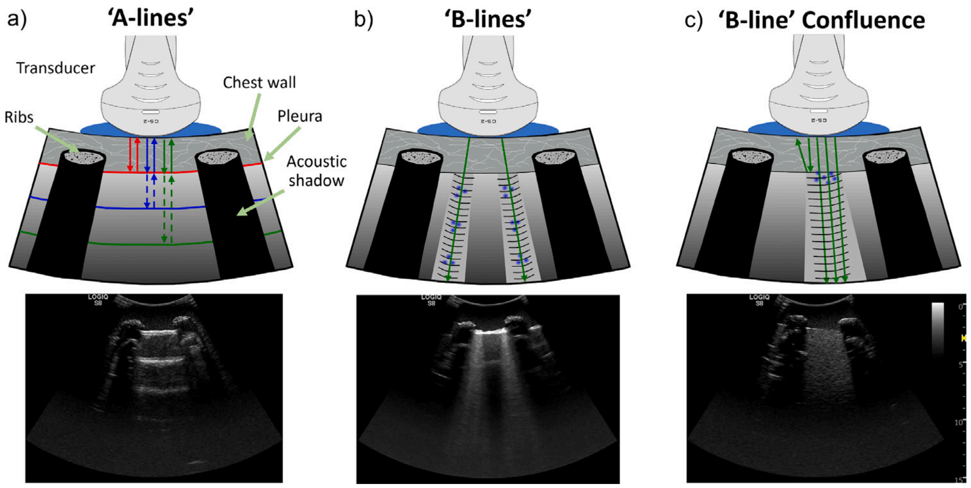

Fig. 1. A representation of lung ultrasound (LUS) artefacts and corresponding example B-mode images of these artefacts acquired using a lung phantom model. (a) Ultrasound waves are reflected at the pleura, creating horizontal reverberation artefacts (’A-lines‘) in a healthy lung. (b) As the amount of fluid increases, vertical artefacts (’B-lines‘) become more prevalent. (c) In more serious cases, B-lines can merge together and become confluent.

ultrasound, it relies on the observation and interpretation of ultrasound artefacts [19] (Fig. 1). A healthy lung would predominantly be air-filled and due to the high acoustic impedance mismatch between tissue and air, it results in near complete back reflection of the ultrasound waves. These reflected waves can reverberate between the ultrasound transducer and the pleural line, which produces B-mode images with characteristic hyperechoic horizontal lines at regular intervals. These horizontal artefacts appear parallel to the pleural line and are commonly referred to as ‘A-lines’ [20] (Fig. 1a). As the pathological state of the lung deteriorates, there is a decrease in the presence of air, replaced by fluid or other biological material. It is thought that as the air spaces in the lung are replaced with less attenuating medium the ultrasound field penetrates deeper into the lung, causing the greater appearance of vertical artefacts including ‘B-lines’ and ‘white lung’ artefacts [21]. B-lines are highly heterogeneous in their appearance, however, these vertical artefacts generally appear as discrete, vertical, hyperechoic, artefacts, which originate at the pleural line [20]. Recent work suggests that B-lines may result from the internal reflection of ultrasound by semi-aerated alveoli, acting as acoustic traps [22](Fig. 1b). In severe pathology cases, discrete B-lines can merge together and become confluent (‘confluent B-lines’), even covering the entire intercostal space as a ‘white lung’ artefact (Fig. 1c), which may be accompanied by a consolidation, where fluid build-up allows for direct imaging of the lung parenchyma [23]. Other indications include pleural effusions, where fluid built up in the pleural space can be directly imaged with ultrasound. Simple pleural effusions commonly present as a homogeneous anechoic region of fluid, while complex pleural effusions appear heterogeneous, representing a region of turbid fluid which may contain particles, debris, or clotted blood [24].

To date, a number of studies have investigated the application of DL to LUS, including image classification, artefact detection, and segmentation [10,25,26]. Among these, several have investigated the frame-wise classification of B-lines. In 2020, a study classified the presence, or absence of B-lines with 93% sensitivity and 96% specificity, additionally measuring B-line severity (multi-class image classification on a scale of 0–4) [27]. While this approach showed promise, it lacks transparency in terms of explaining the model’s prediction process, thereby limiting its trustworthiness for computer-assisted diagnosis [28]. Other studies have attempted to count and localise B-lines [29, 30]. In 2020, van Slaun and Demi used DL for automatic detection and localisation of B-lines images from phantoms and patients, with the potential for real-time implementation at an inference framerate of 276 frames per second (FPS) [30]. In this method, localisation was obtained using gradient-weighted Class Activation Maps (grad-CAM), which lack fine-grained detail compared to segmentation and have been shown to perform poorly in cases with multiple occurrences of the same class [31]. This may be important in the common case of multiple B-lines in a single image. Furthermore, since the appearance of B-lines is highly dependent on ultrasound imaging parameters [32, 33], and B-lines can merge together, B-line counting is unreliable in practice [34].

In 2018, Kulhare et al. trained a single-shot object detector for localisation of vertical artefacts (B-lines and confluent B-lines), horizontal artefacts (A-lines), pleural line, pleural effusion, and lung consolidations in B-mode images [35]. While this method achieved >85% sensitivity and specificity in all classes except B-lines, object detection provides only bounding boxes for features, limiting its usefulness for improving visualisation and feature quantification. Conversely, segmentation allows precise delineation of features and may be more useful in severity assessment. In 2020, Roy *et at.* segmented healthy lung features and markers of disease in LUS, using this to produce an early method of severity assessment [36]. Others have segmented specific artefacts and anatomy, such as Xue et al. who segmented the pleural line, A-line, B-line, and lung consolidation with a mean DSC of 0.44 [37] and Gare et al. who segmented A-line, B-line and pleural line using a pre-trained U-Net model to achieve a mean intersection over union of 0.63, excluding the background class [38].

To date, much of the literature in DL for LUS has focused on classification and secondary localisation of individual features, with few attempting the more challenging and information-rich task of multi-class anatomical and artefact semantic segmentation. Of the studies using segmentation, even fewer consider real-time performance, which is important since LUS is a highly dynamic and investigative imaging technique. Many of the existing DL approaches also rely heavily on large labelled datasets, which can be difficult to obtain in practice [25, 39]. In this work, we train DL models for the semantic segmentation of artefacts (A-lines, B-lines, confluent B-lines) and anatomy (ribs, pleural line) in phantom images, using fewer than 500 images in training. We optimise these models for segmentation performance and inference time, demonstrating the real-time performance of our models in live imaging with a Point-of-Care Ultrasound (PoCUS) system. To supplement this, we investigate the feasibility of transfer learning to train a clinical model with less than 60 images, improving model convergence. We discuss how such models enhance visualisation, enable automatic quantification of features, and may help address the skill gap in LUS interpretation by improving education and minimising operator variability in image interpretation [20,40–42].

2. Methods

2.1. Image datasets

A commercial lung ultrasound model training phantom (CAE Healthcare Inc., Blue Phantom COVID-19 Lung Ultrasound Simulator) was used as a controlled platform to investigate LUS segmentation. This anthropomorphic half-chest model was designed to provide a realistic platform to assess features from healthy to severely damaged lungs and has been suggested as a useful training tool for learning the pathological signs and progression of COVID-19 [43]. Anatomical landmarks of the phantom include the chest wall, ribs, lungs and pleural lining. Under ultrasound examination, it can demonstrate A-lines, B-lines, and B-line confluence in different regions as well as replicating lung sliding with an electric pump.

B-mode ultrasound videos of the lung phantom were acquired using clinical ultrasound systems (GE Healthcare, Logiq S8 & E10) using both a convex curved-array transducer (GE Healthcare, C1-6), and a linear transducer (GE Healthcare, 9L) with a focal depth of 2.5–3 cm (focussed on the pleural line), imaging depth of 10–15 cm, and a mechanical index (MI) in the range of 1.2–1.3. A total of 297 two-second video clips were recorded by a senior sonographer, performing sagittal scans through the anterior and lateral lungs, similar to a standard clinical examination [24]. To avoid systematic biases and maximise image diversity, approximately 560 images were randomly sampled from the videos for annotation.

The VGG Image Annotator (VIA) tool [44] was used to label all images, with polygon annotation of objects and artefacts in the image. To make these labels compatible with semantic segmentation, polygons were converted to segmentation masks using contour filling (OpenCV). For the phantom dataset images were annotated for 5 classes of object/artefact: Rib, Pleural line, A-line, B-line, and B-line confluence. Five individuals with varying levels of ultrasound experience (ranging from none to > 15 years of active research in the field) independently labelled 100 images each. Additionally, a senior sonographer labelled a further 64 images, with a peer review of the labels to ensure consistency between groups. This dataset is accessible online, https://doi.org/10.5518/1485.

Additionally, a retrospective clinical dataset of anonymised ultrasound images was requested from Leeds Teaching Hospitals Trust. These images were recorded between March 2020 and March 2021 in hospitals across the Leeds Teaching Hospital Trust and included patients diagnosed with COVID-19 pneumonia. A variety of point-of-care ultrasound systems and transducers were used to collect the images.

A total of 57 images were selected for annotation from 41 patients across 8 different hospitals. As the images were fully anonymised, it was impossible to split data at the patient level, so data was split randomly to minimise bias. Since these images were taken from patients with severe COVID-19 pneumonia, most contained examples of lung consolidation and/or simple pleural effusion, with no A-lines or individual B-lines present. Additionally, only two images contained clear examples of confluent B-lines which is too few for model training. Therefore, the classes identified for transfer learning were Ribs, Pleural line, Lung consolidation, Simple pleural effusion, and Complex pleural effusion. These images were labelled by two experienced sonographers and cross-verified for accuracy. This anonymised clinical dataset will be made available on reasonable request.

2.2. Models and training

2.2.1. Model architecture and augmentations

A lightweight version of the U-Net architecture [5] was developed for image segmentation to provide a balance between model speed and accuracy (https://github.com/ljhowell/LUS-Segmentation-RT). This was also compared with other variants of the U-Net including Residual U-Net, Attention U-Net, U-Net++, Inception U-Net, SE U-Net, and Dense U-Net [12]. These models all rely on an encoder–decoder structure for segmentation, using skip connections to improve information flow and better preserve both local detail and global context, improving prediction accuracy [5]. In the encoder, the input image is progressively downsampled, with intermediate convolutional blocks which capture information in the image as an increasing number of feature maps. The decoder then upsamples the feature maps back to the original image size over several steps, collating the information as it passes through more convolutional blocks before mapping it to a segmentation mask in the output layer, where each pixel in the mask is assigned a class.

The lightweight implementation of the U-Net model used in this study, keeps the core structure of the base U-Net, with two convolutional blocks in each of the four layers of the encoder and decoder, with 32, 64, 128, and 256 filters respectively. Alongside 3 × 3 convolutions, batch normalisation [45] and dropout layers (20%) [46] were added into the convolutional blocks to reduce overfitting and improve generalisability, especially when the dataset size is limited. Additionally, Leaky Rectified Linear Unit (Leaky ReLU) activation functions were employed to introduce non-linearity into the data transforms, and enhance training stability by addressing the issue of vanishing gradients [47]. Finally, bi-linear upsampling (un-pooling) was used instead of deconvolution in the decoder to help minimise computational complexity and reduce the number of learnable parameters.

To improve the model's ability to generalise to unseen data, an ultrasound-specific augmentation pipeline was used during the training phase (Fig. 2). This included geometric transformations (horizontal axis flip, random rotation) and ultrasound-specific augmentations (gain, time gain compensation, and depth). Geometric transformations increased data diversity, while ultrasound-specific image augmentations were designed to improve model robustness to changes in common imaging settings. To replicate the effect of changing gain, an augmentation was created to adjust the image brightness and contrast by a factor in the interval of +25 to −25 %. For the depth control, an augmentation was applied to randomly crop, centre, and pad images such that the scale and extent were modified similarly to depth adjustment. Finally, for time gain compensation (TGC) the brightness along the vertical axis was altered using shifted Gaussian functions at eight depths, the amplitude of which was sampled from a normal distribution.

Pre-processing was kept to a minimum to reduce the computation time during model inference. Images and masks were cropped to the approximate scan area (i.e. removing scan descriptors), resized to 256 × 256x1 and then normalised to the range 0–1.

2.2.2. Training and evaluation

Prior to training with the LUS phantom data, an independent 20% testing dataset was split from the training data to allow a fair assessment of model performance. This split was conducted at the lowest level (video level), reducing the likelihood of similar images appearing in both datasets, hence minimising the risk of information leakage. Additionally, a 20% validation set was used to monitor performance during model training and regulate the learning process. All reported scores are given as the average test set score ± the standard deviation for 5 splits.

A dynamic approach to training was used, utilising the Adam optimiser [48] with an initial learning rate of 1e-4, which decreased upon a plateau of the validation loss, with early stopping to prevent overfitting. A Combo loss function which combined the Dice loss and cross-entropy loss was used, weighted 2:1 in favour of the Dice term [49]. This helps to address the class imbalance (most pixels belong to the background class) whilst maintaining training stability. Augmentations were applied 'on the fly', meaning that each training image had random augmentations applied each epoch, maximising data diversity. All training was conducted on a single graphics processing unit (GPU) (NVIDIA RTX 3080 laptop), using CUDA 11.4, cuDNN 8.4.1

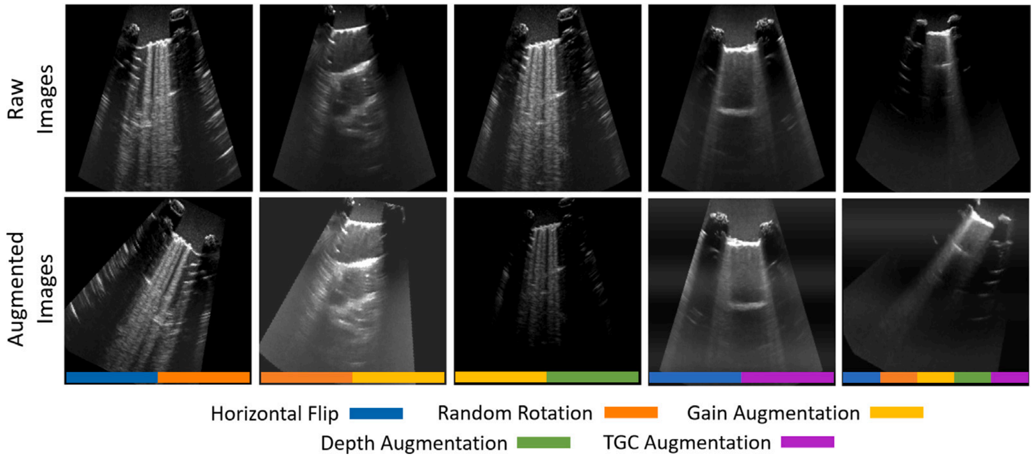

Fig. 2. Examples of lung ultrasound B-mode images from a model phantom (top row) and the same images with applied augmentations (bottom row) including horizontal flip, random rotation, gain, time gain compensation (TGC), and depth augmentations.

and models implemented in TensorFlow 2.9.1 for Python 3.9. Software implementation for network training and validation is made freely available online.

Models were evaluated using an independent test dataset not seen by the network during training. The per class and average Dice similarity coefficient (DSC) were used to quantify the similarity between the predicted masks (X) and manually labelled masks (Y). The DSC is given by the equation

$$\mathrm{DSC} = \frac{2\,|X \cap Y|}{|X| + |Y|} \quad (1)$$

where ∩ represents the intersection operator.

To demonstrate the real-time application of our model, the trained network was used to segment a live image feed from a cart-based point-of-care ultrasound system (GE Healthcare, Venue Fit R3). The LUS phantom was imaged using a convex curved-array transducer (GE Healthcare, C1-5-RS) and frames were streamed to the PC via an HDMI to USB 3.0 capture interface (Magewell, MAG-32060). These frames were pre-processed, segmented by the U-Net model, and the segmentation masks overlaid onto the B-mode image to be displayed on a screen using OpenCV for Python (Fig. 3).

2.2.3. Transfer learning

We explored the feasibility of transfer learning to train U-Nets for segmentation with clinical LUS. By choosing to refine the pre-trained LUS phantom model instead of training a new model from scratch, we leveraged the generic LUS features learned by the network to help the model learn to recognise new features in clinical LUS images. We trained this clinical model to segment five classes, including ribs and pleural line (common to both the phantom and clinical datasets), lung consolidation, simple pleural effusion, and complex pleural effusion (unique to the clinical dataset).

To this effect, we froze the encoder and bottleneck of the pre-trained phantom model and then fine-tuned the weights in the decoder and output layers. This approach ensured that the representations contained in the encoder and bottleneck were not destroyed during training, allowing the features learned to be used to make predictions on the new dataset. The training pipeline was similar to that of the phantom mode but with a 10% test dataset and no validation set due to the limited dataset size and a smaller learning rate of 5e-5 to mitigate overfitting risks.

3. Results

Training on a single GPU took approximately 12 min with 450 images. As expected, we observed a decrease in the training and validation losses as the model learns, converging on a point of stability with a negligible generalisation gap between training and validation loss (Fig. 4a). Model performance improved with an increase in the number of training images until around 300 images, after which we saw no significant improvement (Fig. 4b). Additionally, we found models trained with ultrasound-specific augmentations consistently outperformed those without, with a DSC increase of 0.04 ± 0.01. The accuracy and DSC for the model trained on the full dataset (450 images) can be seen in Table 1. A comparison with state-of-the-art U-Net variants was also conducted, but we found no benefit to accuracy (Fig. A.11) or speed (Fig. A.12) when using these more complex architectures with our dataset.

Table 1
Scores for segmentation of the thyroid phantom are given as the mean ± standard deviation across five training repeats. Timing performance metrics were calculated during real-time experiments and averaged across several hundred frames.

	Metric	Score
Dice similarity coefficient (DSC)	Background	0.98 ± 0.001
	Ribs	0.80 ± 0.01
	Pleural line	0.81 ± 0.01
	A-line	0.63 ± 0.03
	B-line	0.72 ± 0.01
	B-line confluence	0.73 ± 0.09
	Mean (ex. background)	0.74 ± 0.02
Accuracy (%)	Pixel-wise accuracy	95.7 ± 0.34
Timing performance (ms)	Inference	30.0 ± 2.77
	Pre-processing	3.36 ± 0.72
	Displaying	16.8 ± 0.39
Framerate (FPS)	Inference framerate	33.4 ± 2.86
	Total framerate	20.0 ± 1.22

Qualitative examples of U-Net segmentations are presented in Fig. 6a-l. The ribs and the pleural line were generally segmented accurately, even in cases where the pleural lining was thickened and irregular (a-l). B-lines were detected in the majority of instances (c-h, k); however, the extent to which they extended axially in the predictions was occasionally less than that of the manually-labelled images due to lower signals at greater depths (d, k). Most A-lines, including partially obscured ones, were detected by the model (a-f, h, l), although A-lines further from the pleural line had a lower likelihood of being segmented (d). B-line confluence was segmented accurately in most images (i, j), although, in some cases, confusion between individual B-lines and B-line confluence led to the misclassification of certain regions (l).

These observations were supported by an assessment of the pixel-wise agreement between the manually-labelled and model-predicted

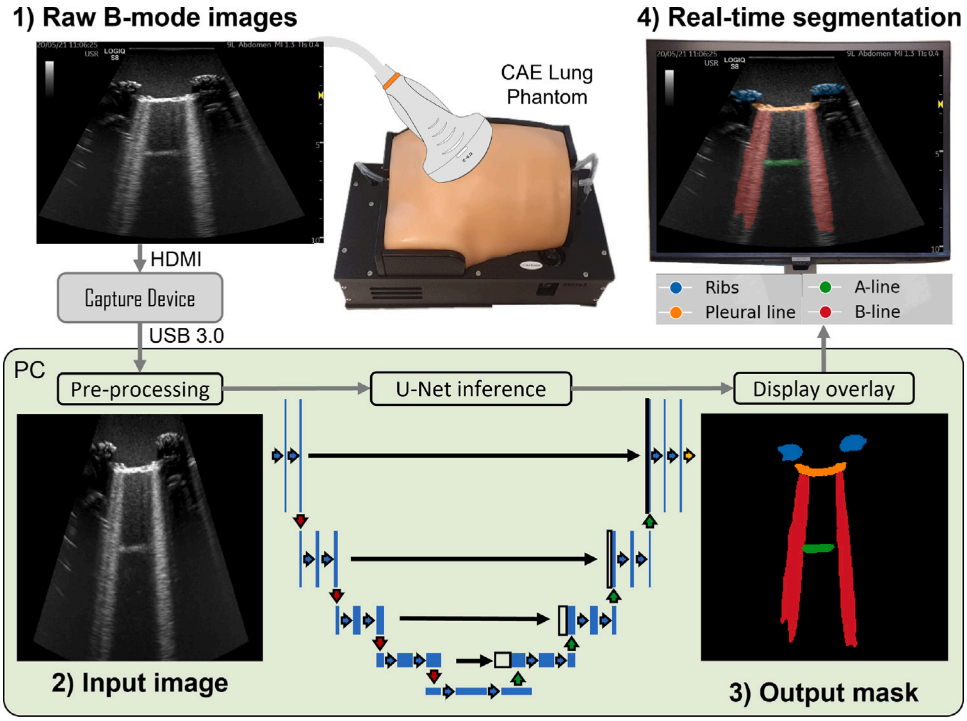

Fig. 3. Workflow diagram showing real-time lung ultrasound segmentation with U-Net.

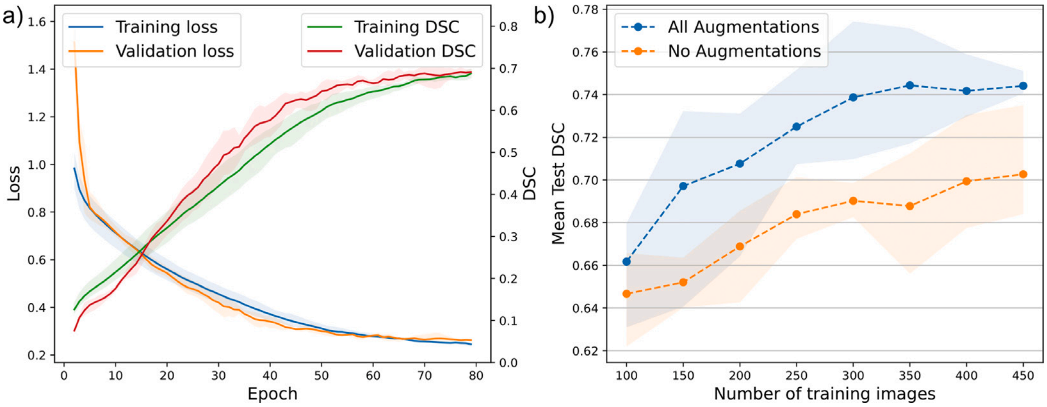

Fig. 4. (a) Lightweight U-Net Training loss and DSC during model training. (b) Mean DSC for models trained with 100 to 450 images, with and without image augmentations. Error regions show the best and worst scores over five training runs.

masks. The normalised confusion matrix for precision (Fig. 5) shows that of the manually-labelled images, 86.8% of pixels labelled as ribs, 85.4% of pleural line, and 72.2% of B-line confluence were predicted correctly by the model. However, only 57.7% of A-line and 57.9% of B-line pixels were correctly predicted. We can see that the majority of pixel-wise errors are false negatives, but some false positive predictions occurred where A-line pixels are predicted as B-line or B-line confluence (likely due to overlapping of A-lines and B-lines) and B-line confluence pixels are predicted as B-line.

Using the predicted segmentation masks, metrics related to the size and shape of objects in the image can also be considered. Here, we defined a semi-quantitative measure for the fraction of the intercostal window taken up by B-lines named the B-line artefact score (BLAS), showing how it can be automatically calculated during real-time imaging. To measure this, the intercostal region of interest (ROI) was determined as the area between the ribs and bounded by the bottom of the pleural line, extending to the maximum depth of the detected B-lines. The B-line fraction (percentage of pixels classified as B-line or B-line confluence) was then quantified in the ROI, with respect to depth (Fig. 7). From this, the percentage of the intercostal space occupied by vertical artefacts (B-lines or confluent B-lines) was calculated using Simpson's Rule on the B-line fraction to give a BLAS in the range 0–1. Empirically, BLAS values less than 0.5 generally suggest one or a few B-lines, values between 0.5 and 0.9 suggest multiple B-lines or some B-line confluence, and values greater than 0.9 indicate white lung artefact. Calculating the ROI dynamically for each image helped to ensure that the BLAS is independent of the imaging angle, depth and extent of the vertical artefacts. This is significant since B-lines cannot be defined by their spatial extent and may not extend to the bottom of the image [20]. Further improvements such as using a trapezoidal or sector-shaped ROI may be more appropriate for curved array transducers.

A Bland-Altman plot was used to assess the agreement between measurements of the BLAS by manual segmentation vs automatic segmentation (Fig. 8). In the phantom test dataset, the mean difference in BLAS was minimal (less than 0.01), indicating little or no systematic bias. The majority of points fall within the limits of agreement (±0.29), defined by the mean plus or minus 1.96 times the standard deviation of

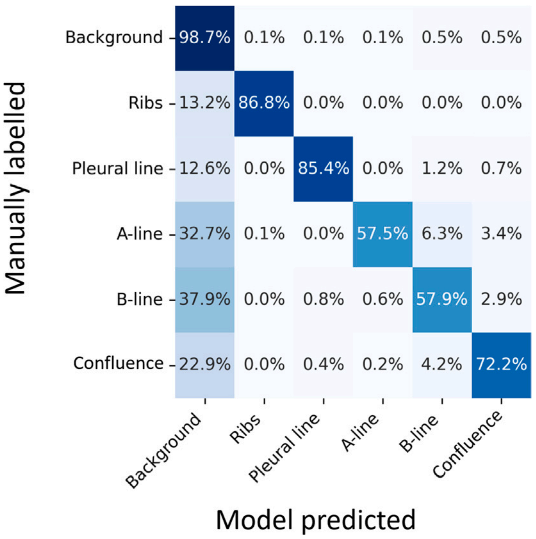

Fig. 5. Confusion matrix for semantic segmentation of LUS phantom images, normalised to give the pixel-wise precision.

the differences, suggesting agreement between the two methods, particularly for cases where the BLAS was greater than 0.9. Investigation of outliers revealed individual cases where a significant difference existed between the manually-labelled and model-predicted masks. Further research would be needed to evaluate the clinical significance of this approach.

To demonstrate live segmentation, the best model was implemented alongside the PoCUS system to image the LUS phantom and overlay masks in real time during scanning (Fig. 9). Segmentation showed recognition of most features, despite the domain shift to a different ultrasound scanner and transducer to that used in model training. Images from the PoCUS system showed greater speckle noise than those in the training dataset, yet the model could still accurately segment the ribs, pleural line and most A-lines, B-lines and B-line confluence. In some frames, the model struggled to segment the full extent of B-lines with depth, and some B-lines were misclassified.

In benchmark tests, the lightweight U-Net model achieved an inference framerate of 33.4 ± 2.86 FPS. However, in real-time experiments, the average framerate lowered to 20.0 ± 1.22 FPS. As a result, some frames were dropped, but the overall speed remained acceptable for real-time visualisation and the framerate variance was not noticeable to the eye. This discrepancy stems from the additional time required for pre-processing and image visualisation, where on average, pre-processing took 3.36 ms, model inference took 29.9 ms, and displaying the image took 16.8 ms. To enhance performance, hardware-accelerated rendering could be used to reduce the display time, or alternatively, further optimisation of the network architecture could be explored.

Finally, we investigated the feasibility of transfer learning to re-train our phantom model to segment clinical LUS images. This approach proved valuable as the clinical dataset had very limited images for training, but the domain was similar to that of the phantom model. The limited clinical LUS images were found to be too few to train a U-Net model from scratch, with poor convergence in initial experiments. Transfer learning improved convergence and allowed reasonable feature recognition of the pleural line, lung consolidations, and simple pleural effusions in some examples (Fig. 10a-d), but ribs and complex pleural effusions were generally poorly segmented (e-f). Predictions on the clinical dataset were generally of poorer quality than those with the phantom dataset, likely due to the limited size and larger variability in the observed anatomy, pathology and types of scanners in the clinical dataset. Variations in the quality of clinical images also meant that some were reported to be diagnostically ambiguous (f) by the two expert annotators, highlighting the difficulty of the task and introducing additional uncertainty to the segmentation labels. Furthermore, since we could not monitor training with a validation dataset there was a possibility of overfitting and more data would be needed to fully assess the effectiveness of transfer learning. The per class DSC for simple pleural effusion, lung consolidation, and pleural line are 0.48, 0.32, and 0.25 respectively, however, the model was unable to reliably segment the ribs (DSC = 0.01) or complex pleural effusion (DSC = 0.01). This is reflected in the mean DSC (across classes excluding the background) of 0.20 ± 0.08 as measured with the limited test dataset.

4. Discussion

ML and DL present an exciting opportunity to assist in the interpretation of LUS [10], and other pathologies imaged using ultrasound [50]. Automatic segmentation of anatomical features and artefacts in LUS B-mode images could aid interpretation, as well as provide new information that could be used to monitor disease severity. To be successful, these models must be both accurate and sufficiently fast to provide feedback for the operator in real-time during scanning. However, due to low signal-to-noise ratio, poor contrast, speckle noise, shadows, and signal dropouts, ultrasound segmentation is a challenging task [51]. Perhaps for this reason, much of the previous work into applying AI with LUS has focussed on image or video classification rather than segmentation [26]. Of these publications, fewer still focus on real-time performance or report the inference time of their models. Additionally, there is an over-reliance in methods which rely on datasets of thousands to hundreds of thousands of images [39]. This is particularly limiting in densely-labelled segmentation, where obtaining large, representative, and high-quality annotated datasets can be prohibitively expensive and time-consuming, so approaches which rely on fewer images are desirable [10,52].

In this study, multi-class segmentation of anatomical features (ribs and pleural line) and artefacts (A-line, B-line, and B-line confluence) in a LUS phantom was explored, considering both the speed and accuracy of the model. A lightweight U-Net model was trained for this task, which demonstrated a mean DSC >0.7 when trained with as few as 300 images. It was demonstrated that more complex U-Net variants did not yield any significant benefit on this dataset, which could be attributed to several factors including the limited number of training images and the uncertainty introduced by ambiguities in labelling.

Most significantly, the subjective nature of annotation is known to limit model performance in LUS [36,53]. In ML, data labels provided by human experts are typically regarded as the immutable ‘ground truth’ [28]. However, objects in ultrasound often cannot be precisely delineated due to speckle noise and limited spatial resolution. This is particularly pronounced for artefacts, which are dynamic, heterogeneous, and lack well-defined boundaries. Consequently, inter-annotator agreement in LUS segmentation may be low [36,54,55]. This has impacts on diagnoses, for example, since it is difficult to know if B-lines are separate or confluent, scores based on counting B-lines are ambiguous [34]. For these reasons, we refer to the labelled datasets as ‘manually-labelled’ rather than ground truth, acknowledging the uncertainty associated with the labels. Nevertheless, this inherent variability may increase the value of DL segmentation in practice, since clinically verified models trained on data from several expert annotators may be able to provide a second opinion without the confirmation bias of any one individual.

A commonly overlooked issue in the literature is the sensitivity of ML models to imaging parameters [20]. To help address this concern, we incorporated ultrasound-specific augmentations into model training, reducing the diversity requirements in the dataset. This helps to improve generalisability and reduce the likelihood of overfitting, which is particularly valuable when training with limited images. As a

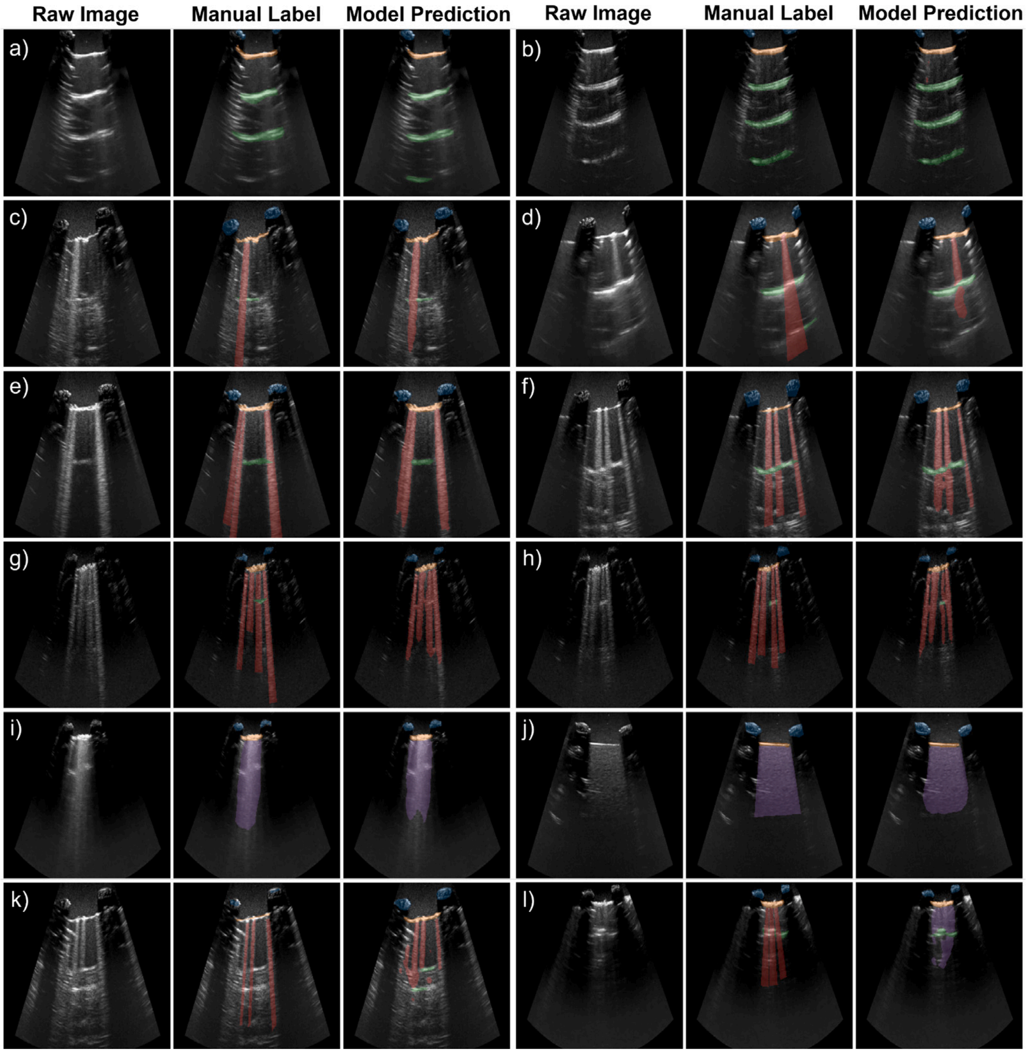

Fig. 6. Example LUS phantom images, overlaid with their manual label and corresponding model prediction. The colour overlay used is as follows, ribs: blue, pleural line: orange, A-lines: green, B-lines: red, and B-line confluence: purple.

result, we observed a mean DSC increase of 6.1% compared to training without augmentations.

Finally, many of the existing studies using AI to calculate prognostic scores from LUS lack explainability, relying on 'black box' methods which do not assist in the decision-making process or allow the clinician to independently review the basis for the recommendations [10]. This can form a major barrier to their adoption in practice, where trust can play a significant role in whether a model is useful [56]. Approaches such as grad-CAM [57] have been used previously in LUS to represent areas of the image that are important to an image's classification [36, 58], but do not fully explain model reasoning.

Although segmentation still relies on complex models, its results are inherently intelligible and verifiable when segmentation masks are overlaid with real-time images. Therefore, metrics based on segmentation masks are more explainable and interpretable than those based on image classifiers. Here we present one such metric, named the B-line Artefact Score (BLAS), which is a simple semi-quantitative measure of the percentage of each intercostal space occupied by B-lines. This provides an indication of the number and appearance of B-lines in an image, which is known to be associated with a number of respiratory diseases [59], including COVID-19 [21], where significant B-line confluence marks the presence of pneumonia deterioration. Compared to individual line counting, measuring the percentages of rib space covered by confluent B-lines has been shown to be more reliable [34]. Our score can be calculated automatically over one or more respiratory cycles, with the maximum score displayed immediately on-screen. This facilitates a rapid assessment of patients while eliminating the ambiguity in counting individual B-lines and selecting the frame to measure.

A similar method could potentially be used to assess pleural line irregularities, which are the second most frequent finding associated with COVID-19 diagnosis, or A-lines which are a marker of recovery from COVID-19 [26]. Once validated on clinical data, such methods could allow the automatic calculation of severity scores, such as those proposed for COVID-19 pneumonia [60], allowing improved triage and management of patients in clinical settings, or monitoring disease progression. In this case, it must also be acknowledged that since the appearance of artefacts varies with frequency, bandwidth, and beam angle [61], predictions or diagnoses based on these measurements must consider these as confounding factors [20].

To showcase the practical application of our models, their implementation alongside a point-of-care ultrasound system was demonstrated. This showed the real-time overlay of segmentation masks onto B-mode images at 20 FPS, which highlighted key features to the operator during scanning. A user-friendly graphical interface allowed clinicians to toggle the visibility of segmentation masks or adjust their transparency, providing an alternative representation which aids feature visualisation. Although in this example a separate PC was used for

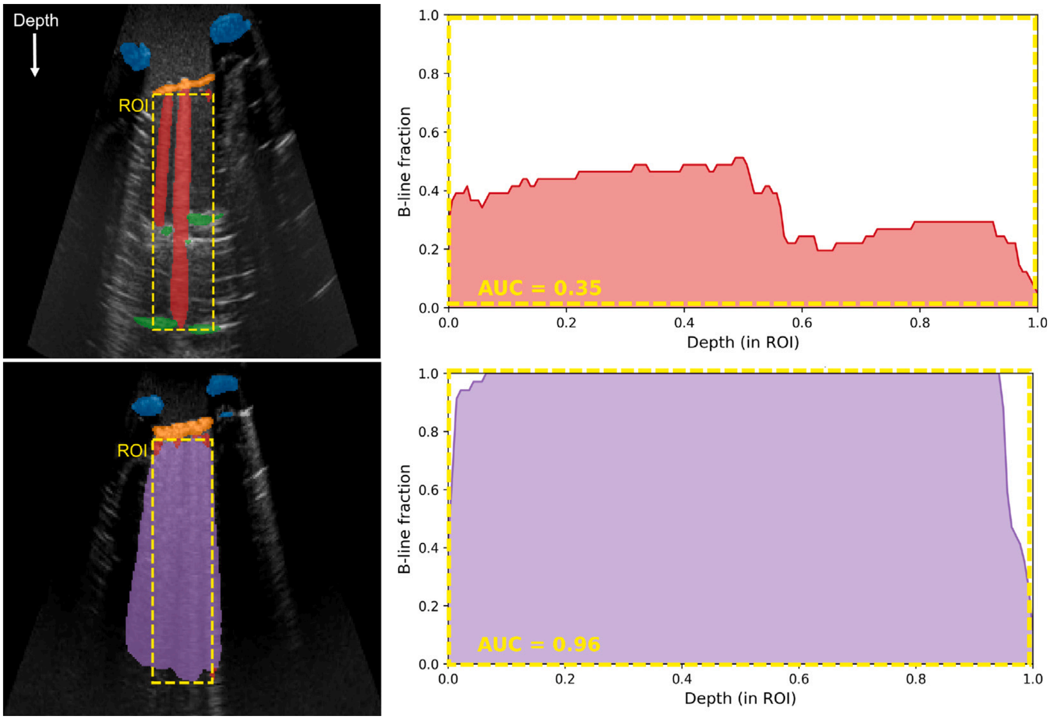

Fig. 7. Calculation of the B-line Artefact Score (BLAS). (Left) The region of interest (ROI) is determined by the boundaries of the pleural line and extends to the depth extent of the measured B-lines (red) or B-line confluence (purple). (Right) The BLAS is equal to the normalised area under the curve (AUC) for the B-line fraction with depth.

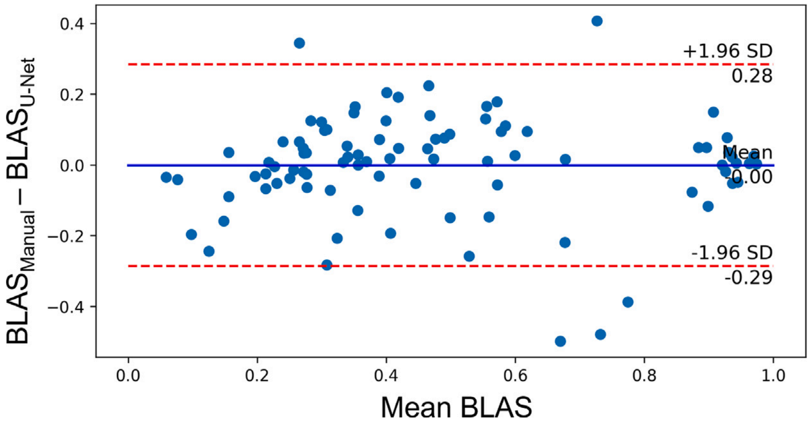

Fig. 8. Bland-Altman difference plot showing the difference between B-line Artefact Score (BLAS) calculated from manual vs. AI-generated labels of LUS images. The plot shows the score difference plotted against the mean score for the test dataset. The mean difference (solid blue line) and limits of agreement (red dashed lines), defined as the mean difference ±1.96 times the standard deviation (SD) of the differences are also displayed.

processing the imaging feed, future implementations may consider how DL methods could be integrated onboard or with embedded devices. For example, tools such as TensorFlow Lite that optimise models for mobile devices may allow lightweight models to be deployed with handheld ultrasound, greatly expanding their usefulness at the bedside. Such approaches have already gained commercial interest. For example, in April 2023 an AI-enabled tool for scoring B-lines received FDA 510(k) clearance for use with a handheld ultrasound device [62].

We also investigate the feasibility of transfer learning to train a model for clinical LUS images. By using transfer learning we leveraged the knowledge acquired from models trained on LUS phantom data to segment pathological features present in patients with severe COVID-19 pneumonia, such as lung consolidation and simple pleural effusions. Previous studies have used similar methods for segmentation of the ribs in LUS images, pre-training U-Net models on a large greyscale dataset of natural images (real-world objects) prior to transfer learning [63]. However, pre-training on ultrasound images may help models to better detect lung-specific patterns [10] and since phantom data is simple to collect and is free of data governance issues, this may present a viable alternative for model pre-training with transfer learning. While our model's DSC indicates it is unlikely to be suitable for implementation in clinical practice, this should be considered a baseline for comparison with future models. Here, training and evaluation of our models is limited by the small size and large heterogeneity of the available data and we are also unable to fully exclude residual information leakage in our model. Nevertheless, this method shows promise for future work in decision support and patient monitoring. With access to a larger and more diverse labelled dataset, future models will improve LUS interpretability in patients with varying disease severity.

Other limitations of this work include a lack of assessment for intra-annotator variability, which could not be conducted since there were no commonly labelled images available. Measuring this variability improves understanding of the data characterisation and subjectivity

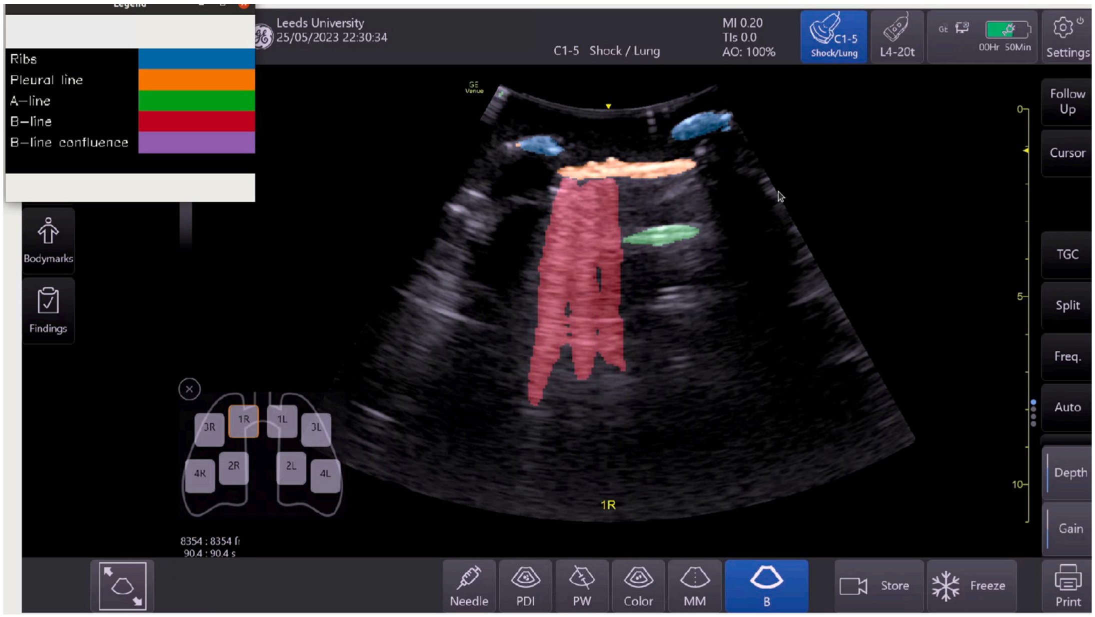

Fig. 9. Real-time segmentation of anatomy and artefacts in a lung ultrasound phantom. Segmentation masks are overlaid onto B-mode images from a point-of-care ultrasound system during scanning to assist with image interpretation. For a video recording please see the supplementary files of this article.

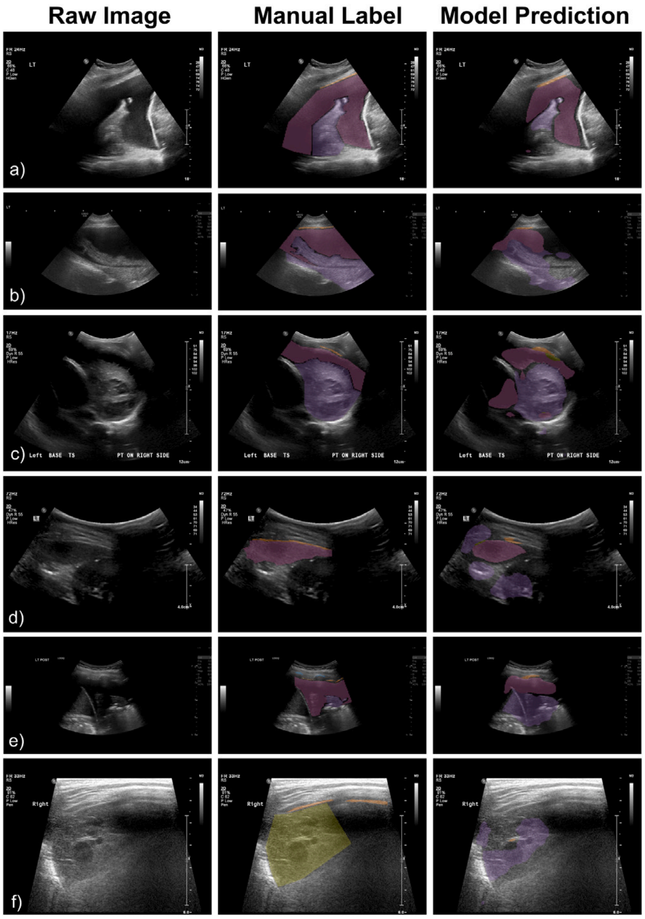

Fig. 10. Example clinical LUS images, overlaid with their manual labels and corresponding model prediction. The colour overlay used is as follows, ribs: blue, pleural line: orange, lung consolidation: purple, simple pleural effusion: pink, and complex pleural effusion: yellow.

from annotation [64], and should therefore be a priority in the curation of future datasets. Secondly, phantom segmentation uses images from a single subject with clearly defined features, so translating the methods employed here to clinical data, which typically exhibit a wider range of patients and pathologies, may pose additional challenges and necessitate a larger and more diverse dataset for optimal performance [28]. Transfer learning may assist in training such models, but does not guarantee better performance [9] and still requires good-quality labelled training data. This study found that transfer learning from a model pre-trained with phantom data provided a practical method to improve model convergence, but more clinical data was needed to assess its impact on segmentation performance. Finally, it would be useful to quantify the generalisability of our models to new ultrasound systems and transducers and to assess whether the inclusion of ultrasound-specific augmentations enhances robustness to changes in imaging settings.

Future works could investigate the application of DL for real-time multi-object LUS segmentation of clinical images, with the aim of deploying models with portable or handheld scanners for training and education, and longer-term in clinical decision support systems. This could have an impact on the diagnosis and management of numerous respiratory conditions [26], especially in low- and middle-income countries and rural areas, where cost-effective ultrasound systems are more widely available than CT [65], but the expertise needed to interpret them may be lacking [66]. More broadly, the methods for training and validation proposed here could be used to develop models for clinical examination of the thyroid, breast, heart or abdomen, among others. This may be especially useful in a variety of settings where real-time feedback is important, such as emergency assessment of trauma with ultrasound [67], ultrasound-guided biopsies [68], or for therapeutic systems such as ultrasound-guided focused ultrasound for tissue ablation [69].

5. Conclusion

This study demonstrated the application of deep learning to real-time, multi-class segmentation of objects and artefacts in lung ultrasound. Models trained with as few as 300 images were able to segment the ribs, pleural line, A-lines, B-lines, and B-line confluence in B-mode images of a COVID-19 lung tissue-mimicking phantom, with

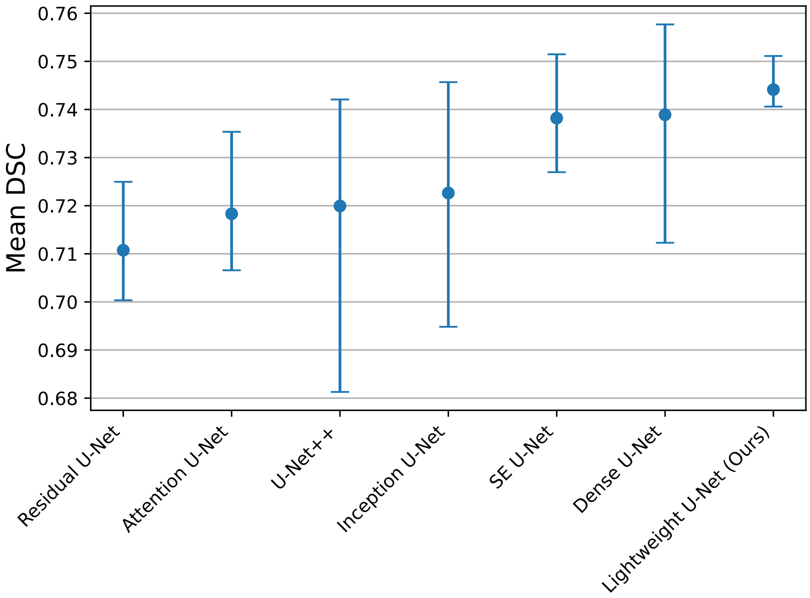

Fig. A.11. Comparison of the Dice similarity coefficient for various U-Net variants, trained on the phantom dataset (450 images). Error bars show the best and worst scores over five training repeats.

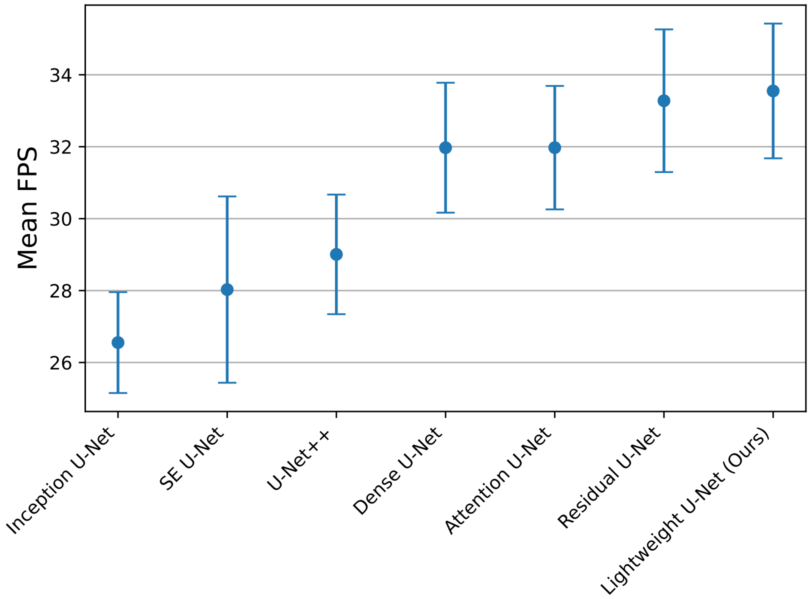

Fig. A.12. Comparison of the benchmark framerate for inference of various U-Net variants, trained on the phantom dataset (450 images). Error bars show the best and worst scores over five training repeats.

an average DSC of 0.74 across the five classes. When overlaid onto LUS B-mode images, segmentation masks offer an improved visual interpretation of the image. This could be incorporated into training systems for medical education of LUS in specialised programmes or as part of training new clinicians [20], addressing the existing skill gap in LUS. To maximise their clinical potential, these models must be deployable on PoCUS systems and suitable for real-time inference. We demonstrated the feasibility of this at 20 FPS, with the potential to achieve >30 FPS with hardware-accelerated rendering. Furthermore, segmentation masks provide an explainable method for scoring disease severity, which has the potential to assist in the triage and management of patients for a variety of respiratory conditions. On this theme, we propose a B-line artefact score (BLAS) which automatically measures the percentage of the intercostal space occupied by vertical artefacts (B-lines and confluent B-lines). Future work should consider the translation of these methods to clinical data, considering transfer learning as a viable method to build models which can assist in the interpretation of LUS and reduce inter-operator variability associated with this subjective imaging technique. For the LUS phantom dataset, a comparison with state-of-the-art was also conducted (Fig. A.11). We found no benefit in terms of DSC or inference time compared to our lightweight U-Net (Fig. A.12).

CRediT authorship contribution statement

Lewis Howell: Data curation, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Nicola Ingram:** Data curation, Project administration, Supervision, Writing – original draft. **Roger Lapham:** Data curation, Formal analysis, Writing – original draft. **Adam Morrell:** Data curation, Formal analysis, Writing – original draft. **James R. McLaughlan:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Phantom data is available through link in manuscript. Clinical data are available upon request. The data associated with this paper are openly available from the University of Leeds Data Repository. https://doi.org/10.5518/1485. Where the model can be accessed via, https://github.com/ljhowell/LUS-Segmentation-RT.

Acknowledgements

This work was supported by the following UK Research and Innovation (UKRI) grants, EP/S001069/1 and EP/V043714/1. The authors would also like to thank Stephen Wolstenhulme who aided with the acquisition and labelling of phantom data. The authors thank the patients for the use of their anonymised imaging data.

Appendix A. Comparison with the state-of-the-art

See Figs. A.11 and A.12.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ultras.2024.107251.

References

[1] K.-H. Yu, A.L. Beam, I.S. Kohane, Artificial intelligence in healthcare, Nat. Biomed. Eng. 2 (10) (2018) 719–731.

[2] R.J. Burton, M. Albur, M. Eberl, S.M. Cuff, Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections, BMC Med. Inf. Decis. Mak. 19 (1) (2019) 1–11.

[3] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, P. Biancone, The role of artificial intelligence in healthcare: a structured literature review, BMC Med. Inf. Decis. Mak. 21 (2021) 1–23.

[4] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, Deep learning for visual understanding: A review, Neurocomputing 187 (2016) 27–48.

[5] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III, Vol. 18, Springer, 2015, pp. 234–241.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[7] L. Karlinsky, T. Michaeli, K. Nishino, Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, vol. 13806, Springer Nature, 2023.

[8] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, UNETR: Transformers for 3D medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 574–584, URL https://openaccess.thecvf.com/content/WACV2022/html/Hatamizadeh_UNETR_Transformers_for_3D_Medical_Image_Segmentation_WACV_2022_paper.html.

[9] K. Sailunaz, T. Özyer, J. Rokne, R. Alhajj, A survey of machine learning-based methods for COVID-19 medical image analysis, Med. Biol. Eng. Comput. 61 (6) (2023) 1257–1297, http://dx.doi.org/10.1007/s11517-022-02758-y.

[10] T. Yang, O. Karakus, N. Anantrasirichai, A. Achim, Current advances in computational lung ultrasound imaging: A review, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 70 (1) (2023) 2–15, http://dx.doi.org/10.1109/TUFFC.2022.3221682, URL https://ieeexplore.ieee.org/document/9945990/.

[11] D. Karimi, S.K. Warfield, A. Gholipour, Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations, Artif. Intell. Med. 116 (2021) 102078, http://dx.doi.org/10.1016/j.artmed.2021.102078, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8164174/.

[12] N. Siddique, S. Paheding, C.P. Elkin, V. Devabhaktuni, U-net and its variants for medical image segmentation: A review of theory and applications, IEEE Access 9 (2021) 82031–82057.

[13] M.A. Pereda, M.A. Chavez, C.C. Hooper-Miele, R.H. Gilman, M.C. Steinhoff, L.E. Ellington, M. Gross, C. Price, J.M. Tielsch, W. Checkley, Lung ultrasound for the diagnosis of pneumonia in children: a meta-analysis, Pediatrics 135 (4) (2015) 714–722.

[14] L. Long, H.-T. Zhao, Z.-Y. Zhang, G.-Y. Wang, H.-L. Zhao, Lung ultrasound for the diagnosis of pneumonia in adults: a meta-analysis, Medicine 96 (3) (2017) e5713.

[15] M. Smith, S. Hayward, S. Innes, A. Miller, Point-of-care lung ultrasound in patients with COVID-19–a narrative review, Anaesthesia 75 (8) (2020) 1096–1104.

[16] D.A. Lichtenstein, BLUE-protocol and FALLS-protocol: two applications of lung ultrasound in the critically ill, Chest 147 (6) (2015) 1659–1670.

[17] S. Moore, E. Gardiner, Point of care and intensive care lung ultrasound: a reference guide for practitioners during COVID-19, Radiography 26 (4) (2020) e297–e302.

[18] S. Wolstenhulme, J.R. McLaughlan, Lung ultrasound education: simulation and hands-on, Br. J. Radiol. 93 (1119) (2021) 20200755.

[19] A. Miller, Practical approach to lung ultrasound, BJA Educ. 16 (2) (2016) 39–45.

[20] L. Demi, F. Wolfram, C. Klersy, A. De Silvestri, V.V. Ferretti, M. Muller, D. Miller, F. Feletti, M. Wełnicki, N. Buda, A. Skoczylas, A. Pomiecko, D. Damjanovic, R. Olszewski, A.W. Kirkpatrick, R. Breitkreutz, G. Mathis, G. Soldati, A. Smargiassi, R. Inchingolo, T. Perrone, New international guidelines and consensus on the use of lung ultrasound, J. Ultrasound Med. 42 (2) (2023) 309–344, http://dx.doi.org/10.1002/jum.16088, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jum.16088. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jum.16088.

[21] G. Soldati, A. Smargiassi, R. Inchingolo, D. Buonsenso, T. Perrone, D.F. Briganti, S. Perlini, E. Torri, A. Mariani, E.E. Mossolani, et al., Is there a role for lung ultrasound during the COVID-19 pandemic? J. Ultrasound Med. 39 (7) (2020) 1459.

[22] M. Demi, R. Prediletto, G. Soldati, L. Demi, Physical mechanisms providing clinical information from ultrasound lung images: Hypotheses and early confirmations, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 67 (3) (2020) 612–623, http://dx.doi.org/10.1109/TUFFC.2019.2949597, URL https://ieeexplore.ieee.org/document/8883203/.

[23] E.A. Fischer, T. Minami, I.W. Ma, K. Yasukawa, Lung ultrasound for pleural line abnormalities, confluent B-lines, and consolidation: Expert reproducibility and a method of standardization, J. Ultrasound Med. 41 (8) (2022) 2097–2107.

[24] T.J. Marini, D.J. Rubens, Y.T. Zhao, J. Weis, T.P. O'Connor, W.H. Novak, K.A. Kaproth-Joslin, Lung ultrasound: The essentials, Radiol.: Cardiothoracic Imag. 3 (2) (2021) e200564, http://dx.doi.org/10.1148/ryct.2021200564, URL https://pubs.rsna.org/doi/full/10.1148/ryct.2021200564. Publisher: Radiological Society of North America.

[25] L. Zhao, M.A. Lediju Bell, A review of deep learning applications in lung ultrasound imaging of COVID-19 patients, BME Front. 2022 (2022) http://dx.doi.org/10.34133/2022/9780173, URL https://spj.science.org/doi/10.34133/2022/9780173. Publisher: American Association for the Advancement of Science.

[26] J. Wang, X. Yang, B. Zhou, J.J. Sohn, J. Zhou, J.T. Jacob, K.A. Higgins, J.D. Bradley, T. Liu, Review of machine learning in lung ultrasound in COVID-19 pandemic, J. Imag. 8 (3) (2022) 65, http://dx.doi.org/10.3390/jimaging8030065, URL https://www.mdpi.com/2313-433X/8/3/65. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[27] C. Baloescu, G. Toporek, S. Kim, K. McNamara, R. Liu, M.M. Shaw, R.L. McNamara, B.I. Raju, C.L. Moore, Automated lung ultrasound B-line assessment using a deep learning algorithm, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 67 (11) (2020) 2312–2320, http://dx.doi.org/10.1109/TUFFC.2020.3002249, Conference Name: IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control.

[28] G. Choy, O. Khalilzadeh, M. Michalski, S. Do, A.E. Samir, O.S. Pianykh, J.R. Geis, P.V. Pandharipande, J.A. Brink, K.J. Dreyer, Current applications and future impact of machine learning in radiology, Radiology 288 (2) (2018) 318–328, http://dx.doi.org/10.1148/radiol.2018171820, URL https://pubs.rsna.org/doi/full/10.1148/radiol.2018171820. Publisher: Radiological Society of North America.

[29] X. Wang, J.S. Burzynski, J. Hamilton, P.S. Rao, W.F. Weitzel, J.L. Bull, Quantifying lung ultrasound comets with a convolutional neural network: Initial clinical results, Comput. Biol. Med. 107 (2019) 39–46, http://dx.doi.org/10.1016/j.compbiomed.2019.02.002, URL https://www.sciencedirect.com/science/article/pii/S0010482519300356.

[30] R.J.G. van Sloun, L. Demi, Localizing B-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results, IEEE J. Biomed. Health Inf. 24 (4) (2020) 957–964, http://dx.doi.org/10.1109/JBHI.2019.2936151.

[31] A. Chattopadhyay, A. Sarkar, P. Howlader, V. Balasubramanian, Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, 2017.

[32] F. Mento, L. Demi, Dependence of lung ultrasound vertical artifacts on frequency, bandwidth, focus and angle of incidence: An in vitro study, J. Acoust. Soc. Am. 150 (6) (2021) 4075, http://dx.doi.org/10.1121/10.0007482.

[33] F. Mento, L. Demi, On the influence of imaging parameters on lung ultrasound B-line artifacts, in vitro studya), J. Acoust. Soc. Am. 148 (2) (2020) 975–983, http://dx.doi.org/10.1121/10.0001797.

[34] K.L. Anderson, J.M. Fields, N.L. Panebianco, K.Y. Jenq, J. Marin, A.J. Dean, Inter-rater reliability of quantifying pleural B-lines using multiple counting methods, J. Ultrasound Med. 32 (1) (2013) 115–120, http://dx.doi.org/10.7863/jum.2013.32.1.115, URL https://onlinelibrary.wiley.com/doi/abs/10.7863/jum.2013.32.1.115. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.7863/jum.2013.32.1.115.

[35] S. Kulhare, X. Zheng, C. Mehanian, C. Gregory, M. Zhu, K. Gregory, H. Xie, J. McAndrew Jones, B. Wilson, Ultrasound-based detection of lung abnormalities using single shot detection convolutional neural networks, in: D. Stoyanov, Z. Taylor, S. Aylward, J.a.M.R. Tavares, Y. Xiao, A. Simpson, A. Martel, L. Maier-Hein, S. Li, H. Rivaz, I. Reinertsen, M. Chabanas, K. Farahani (Eds.), Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2018, pp. 65–73, http://dx.doi.org/10.1007/978-3-030-01045-4_8.

[36] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, I. Huijben, N. Chennakeshava, F. Mento, A. Sentelli, E. Peschiera, R. Trevisan, G. Maschietto, E. Torri, R. Inchingolo, A. Smargiassi, G. Soldati, P. Rota, A. Passerini, L. Demi, Deep learning for classification and localization of COVID-19 Markers in point-of-care lung ultrasound, IEEE Trans. Med. Imaging PP (2020) http://dx.doi.org/10.1109/TMI.2020.2994459.

[37] W. Xue, C. Cao, J. Liu, Y. Duan, H. Cao, J. Wang, X. Tao, Z. Chen, M. Wu, J. Zhang, H. Sun, Y. Jin, X. Yang, R. Huang, F. Xiang, Y. Song, M. You, W. Zhang, L. Jiang, Z. Zhang, S. Kong, Y. Tian, L. Zhang, D. Ni, M. Xie, Modality alignment contrastive learning for severity assessment of COVID-19 from lung ultrasound and clinical information, Med. Image Anal. 69 (2021) 101975, http://dx.doi.org/10.1016/j.media.2021.101975.

[38] G.R. Gare, A. Schoenling, V. Philip, H.V. Tran, B.P. deBoisblanc, R.L. Rodriguez, J.M. Galeotti, Dense pixel-labeling for reverse-transfer and diagnostic learning on lung ultrasound for COVID-19 and pneumonia detection, in: 2021 IEEE 18th International Symposium on Biomedical Imaging, ISBI, 2021, pp. 1406–1410, http://dx.doi.org/10.1109/ISBI48211.2021.9433826, URL http://arxiv.org/abs/2201.10166. arXiv:2201.10166 [cs, eess].

[39] F. Mento, U. Khan, F. Faita, A. Smargiassi, R. Inchingolo, T. Perrone, L. Demi, State of the art in lung ultrasound, shifting from qualitative to quantitative analyses, Ultrasound Med. Biol. 48 (12) (2022) 2398–2416, http://dx.doi.org/10.1016/j.ultrasmedbio.2022.07.007.

[40] C.-H. Tsai, J. van der Burgt, D. Vukovic, N. Kaur, L. Demi, D. Canty, A. Wang, A. Royse, C. Royse, K. Haji, et al., Automatic deep learning-based pleural effusion classification in lung ultrasound images for respiratory pathology diagnosis, Phys. Medica 83 (2021) 38–45.

[41] L. Zhao, T.C. Fong, M.A.L. Bell, COVID-19 feature detection with deep neural networks trained on simulated lung ultrasound B-mode images, in: 2022 IEEE International Ultrasonics Symposium, IUS, IEEE, 2022, pp. 1–3.

[42] U. Khan, S. Afrakhteh, F. Mento, N. Fatima, L. De Rosa, L.L. Custode, Z. Azam, E. Torri, G. Soldati, F. Tursi, et al., Benchmark methodological approach for the application of artificial intelligence to lung ultrasound data from COVID-19 patients: From frame to prognostic-level, Ultrasonics (2023) 106994.

[43] C. Milius, L. Jepson, D. Maclaskey, V. Pazdernik, T. Kondrashova, Development of hands-on skills in diagnostics of lung diseases using ultrasonography in undergraduate medical education, Missouri Med. 120 (2) (2023) 128–133, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10121122/.

[44] A. Dutta, A. Zisserman, The VIA Annotation software for images, audio and video, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, Nice France, 2019, pp. 2276–2279, http://dx.doi.org/10.1145/3343031.3350535, URL https://dl.acm.org/doi/10.1145/3343031.3350535.

[45] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating deep network training by reducing internal covariate shift, 2015, arXiv. URL http://arxiv.org/abs/1502.03167. arXiv:1502.03167 [cs].

[46] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, 2012, arXiv. URL http://arxiv.org/abs/1207.0580. arXiv:1207.0580 [cs].

[47] B. Xu, N. Wang, T. Chen, M. Li, Empirical evaluation of rectified activations in convolutional network, 2015, arXiv. URL http://arxiv.org/abs/1505.00853. arXiv:1505.00853 [cs, stat].

[48] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017, http://dx.doi.org/10.48550/arXiv.1412.6980, arXiv. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs].

[49] S.A. Taghanaki, Y. Zheng, S. Kevin Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, G. Hamarneh, Combo loss: Handling input and output imbalance in multi-organ segmentation, Comput. Med. Imaging Graph. 75 (2019) 24–33, http://dx.doi.org/10.1016/j.compmedimag.2019.04.005, URL https://www.sciencedirect.com/science/article/pii/S0895611118305688.

[50] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S.X. Li, D. Ni, T. Wang, Deep learning in medical ultrasound analysis: A review, Engineering 5 (2) (2019) 261–275, http://dx.doi.org/10.1016/j.eng.2018.11.020, URL https://www.sciencedirect.com/science/article/pii/S2095809918301887.

[51] J. Noble, D. Boukerroui, Ultrasound image segmentation: a survey, IEEE Trans. Med. Imaging 25 (8) (2006) 987–1010, http://dx.doi.org/10.1109/TMI.2006.877092, Conference Name: IEEE Transactions on Medical Imaging.

[52] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J.N. Chiang, Z. Wu, X. Ding, Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation, Med. Image Anal. 63 (2020) 101693, http://dx.doi.org/10.1016/j.media.2020.101693, URL https://www.sciencedirect.com/science/article/pii/S136184152030058X.

[53] H. Mason, L. Cristoni, A. Walden, R. Lazzari, T. Pulimood, L. Grandjean, C.A.G. Wheeler-Kingshott, Y. Hu, Z.M. Baum, Lung ultrasound segmentation and adaptation between COVID-19 and community-acquired pneumonia, 2021, http://dx.doi.org/10.48550/arXiv.2108.03138, arXiv. URL http://arxiv.org/abs/2108.03138. arXiv:2108.03138 [cs, eess].

[54] J.L. Herraiz, C. Freijo, J. Camacho, M. Muñoz, R. González, R. Alonso-Roca, J. Álvarez-Troncoso, L.M. Beltrán-Romero, M. Bernabeu-Wittel, R. Blancas, A. Calvo-Cebrián, R. Campo-Linares, J. Chehayeb-Morán, J. Chorda-Ribelles, S. García-Rubio, G. García-de Casasola, A. Gil-Rodrigo, C. Henríquez-Camacho, A. Hernandez-Píriz, C. Hernandez-Quiles, R. Llamas-Fuentes, D. Luordo, R. Marín-Baselga, M.C. Martínez-Díaz, M. Mateos-González, M. Mendez-Bailon, F. Miralles-Aguiar, R. Nogue, M. Nogué, B. Ortiz de Urbina-Antia, A.Á. Oviedo-García, J.M. Porcel, S. Rodriguez, D.A. Rodríguez-Serrano, T. Sainz, I.M. Sánchez-Barrancos, M. Torres-Arrese, J. Torres-Macho, A. Trueba Vicente, T. Villén-Villegas, J.J. Zafra-Sánchez, Y. Tung-Chen, Inter-rater variability in the evaluation of lung ultrasound in videos acquired from COVID-19 patients, Appl. Sci. 13 (3) (2023) 1321, http://dx.doi.org/10.3390/app13031321, URL https://www.mdpi.com/2076-3417/13/3/1321. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[55] J. Gullett, J.P. Donnelly, R. Sinert, B. Hosek, D. Fuller, H. Hill, I. Feldman, G. Galetto, M. Auster, B. Hoffmann, Interobserver agreement in the evaluation of B-lines using bedside ultrasound, J. Crit. Care 30 (6) (2015) 1395–1399, http://dx.doi.org/10.1016/j.jcrc.2015.08.021, URL https://www.sciencedirect.com/science/article/pii/S0883944115004566.

[56] T. Davenport, R. Kalakota, The potential for artificial intelligence in healthcare, Future Healthc. J. 6 (2) (2019) 94–98, http://dx.doi.org/10.7861/futurehosp.6-2-94, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/.

[57] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (2020) 336–359, http://dx.doi.org/10.1007/s11263-019-01228-7, URL http://arxiv.org/abs/1610.02391. arXiv:1610.02391 [cs].

[58] U. Khan, S. Afrakhteh, F. Mento, N. Fatima, L. De Rosa, L.L. Custode, Z. Azam, E. Torri, G. Soldati, F. Tursi, V.N. Macioce, A. Smargiassi, R. Inchingolo, T. Perrone, G. Iacca, L. Demi, Benchmark methodological approach for the application of artificial intelligence to lung ultrasound data from COVID-19 patients: From frame to prognostic-level, Ultrasonics 132 (2023) 106994, http://dx.doi.org/10.1016/j.ultras.2023.106994, URL https://www.sciencedirect.com/science/article/pii/S0041624X23000707.

[59] G. Soldati, M. Demi, A. Smargiassi, R. Inchingolo, L. Demi, The role of ultrasound lung artefacts in the diagnosis of respiratory diseases, Expert Rev. Respirat. Med. 13 (2019) http://dx.doi.org/10.1080/17476348.2019.1565997.

[60] T. Perrone, G. Soldati, L. Padovini, A. Fiengo, G. Lettieri, U. Sabatini, G. Gori, F. Lepore, M. Garolfi, I. Palumbo, R. Inchingolo, A. Smargiassi, L. Demi, E.E. Mossolani, F. Tursi, C. Klersy, A. Di Sabatino, A new lung ultrasound protocol able to predict worsening in patients affected by severe acute respiratory syndrome coronavirus 2 pneumonia, J. Ultrasound Med.: Official J. Am. Inst. Ultrasound Med. 40 (8) (2021) 1627–1635, http://dx.doi.org/10.1002/jum.15548.

[61] L. Demi, W. van Hoeve, R.J.G. van Sloun, G. Soldati, M. Demi, Determination of a potential quantitative measure of the state of the lung using lung ultrasound spectroscopy, Sci. Rep. 7 (1) (2017) 12746, http://dx.doi.org/10.1038/s41598-017-13078-9, URL https://www.nature.com/articles/s41598-017-13078-9. Number: 1 Publisher: Nature Publishing Group.

[62] US Food and Drug Administration, 510(k) premarket notification, 2023, URL https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K202406.

[63] D. Cheng, E.Y. Lam, Transfer learning U-Net deep learning for lung ultrasound segmentation, 2021, http://dx.doi.org/10.48550/arXiv.2110.02196, arXiv. URL http://arxiv.org/abs/2110.02196. arXiv:2110.02196 [cs, eess].

[64] F. Yang, G. Zamzmi, S. Angara, S. Rajaraman, A. Aquilina, Z. Xue, S. Jaeger, E. Papagiannakis, S.K. Antani, Assessing inter-annotator agreement for medical image segmentation, IEEE Access : Pract. Innov., Open Solut. 11 (2023) 21300–21312, http://dx.doi.org/10.1109/access.2023.3249759, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10062409/.

[65] K.A. Stewart, S.M. Navarro, S. Kambala, G. Tan, R. Poondla, S. Lederman, K. Barbour, C. Lavy, Trends in ultrasound use in low and middle income countries: A systematic review, Int. J. Matern. Child Health AIDS 9 (1) (2020) 103–120, http://dx.doi.org/10.21106/ijma.294, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7031872/.

[66] S. Shah, B.A. Bellows, A.A. Adedipe, J.E. Totten, B.H. Backlund, D. Sajed, Perceived barriers in the use of ultrasound in developing countries, Crit. Ultrasound J. 7 (1) (2015) 28, http://dx.doi.org/10.1186/s13089-015-0028-2.

[67] M.M. Leo, I.Y. Potter, M. Zahiri, A. Vaziri, C.F. Jung, J.A. Feldman, Using deep learning to detect the presence and location of hemoperitoneum on the focused assessment with sonography in trauma (FAST) examination in adults, J. Digit. Imag. (2023) http://dx.doi.org/10.1007/s10278-023-00845-6.

[68] A. Wijata, J. Andrzejewski, B. Pyciński, An automatic biopsy needle detection and segmentation on ultrasound images using a convolutional neural network, Ultrason. Imaging 43 (5) (2021) 262–272, http://dx.doi.org/10.1177/01617346211025267, Publisher: SAGE Publications Inc.

[69] E.S. Ebbini, G. Ter Haar, Ultrasound-guided therapeutic focused ultrasound: Current status and future directions, Int. J. Hyperth. 31 (2) (2015) 77–89, http://dx.doi.org/10.3109/02656736.2014.995238, Publisher: Taylor & Francis _eprint.