

Article

Automatic Lung Aeration Assessment for Lung Ultrasound Imaging in the Pediatric Intensive Care Unit

Sabien G. J. Heisterkamp ^{1,*},[†] , Tharanghi Logendran ^{1,†}, Ariane Willems ¹ and Can Ozan Tan ²

¹ Pediatric Intensive Care Unit, Leiden University Medical Center (LUMC), Postbus 9600, 2300 RC Leiden, The Netherlands

² Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, RAM Group, 7522 NH Enschede, The Netherlands

* Correspondence: s.g.j.heisterkamp@lumc.nl

[†] These authors contributed equally to this work.

Abstract

Imaging of the lungs is traditionally based on chest X-ray as a first-line imaging method for lung aeration assessment. However, radiation exposure limits its use for patients in the pediatric intensive care unit. Lung ultrasonography (LUS) is a suitable alternative, but its interpretation is highly observer-dependent and requires sufficient experience and skill. We sought to develop a model based on deep learning to assist the clinician in the interpretation of LUS observations. In this retrospective, single-center, proof-of-concept study, all patients, age 0–18 years old admitted at the PICU of the Leiden University Medical Center (LUMC) between January and May 2022 who underwent an LUS were included. LUS video frames were analyzed using a deep learning tool; a conditional generative adversarial network (cGAN) was developed to generate segmentation masks containing clinical features from individual LUS frames. A total of 31 patients, with a median age of 2.5 months (IQR 0–11 months), were analyzed. A total of 98 LUS assessments and 506 4-s videos were collected. The median LUS score was 12 (IQR 8–17). The two best-performing frame-based segmentation models achieved mean Dice similarity coefficients of 0.97 ± 0.03 and 0.96 ± 0.03 , with mean squared errors of 0.025 ± 0.025 and 0.030 ± 0.026 , respectively. These findings demonstrate that a pediatric-specific cGAN can segment key LUS features from individual frames. However, the results derive from a small, single-center cohort with a frame-level rather than patient-level data split, and no formal clinical validation; independent, prospectively collected validation cohorts are required before any clinical implementation.

Keywords: lung ultrasound; deep learning; pediatric; PICU



Academic Editors: Claus Jacob and Eleonore Fröhlich

Received: 30 March 2026

Revised: 28 May 2026

Accepted: 25 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Critically ill pediatric patients admitted to pediatric intensive care units (PICU) frequently require diagnosis and monitoring of acute lung and chest abnormalities such as pneumonia, bronchiolitis, and atelectasis. Chest X-ray (CXR) imaging has traditionally served as the first-line imaging modality for this purpose. However, the use of ionizing radiation in pediatric populations presents a significant clinical concern, as children demonstrate substantially higher radiosensitivity than adults due to their developing tissues, greater cellular proliferation, and longer remaining lifespan [1,2]. A single small radiation dose increases the lifetime cancer risk in infants by approximately an order of magnitude compared to adults, with estimated lifetime cancer mortality risks from a single CT in a

1-year-old being 0.18% for abdominal and 0.07% for head examinations [2,3], and cumulative radiation exposure from repeated chest radiographs in PICU settings further amplifying this risk.

Lung ultrasonography (LUS) represents a clinically attractive alternative to CXR, offering real-time, portable, bedside imaging without ionizing radiation. Implementation of LUS protocols in neonatal intensive care units (NICU) has demonstrated significant reductions in radiation exposure, with studies reporting decreased CXR examinations (median 0 vs. 2 CXRs per patient, $p < 0.001$) and lower median radiation doses per infant (30 μGy vs. 56 μGy , $p < 0.001$) following LUS protocol implementation [4]. LUS has consequently become a widespread diagnostic and monitoring tool across pediatric and neonatal populations. Beyond the absence of ionizing radiation, ultrasound offers additional advantages in critically ill children, including repeatability and the ability to assess dynamic respiratory mechanics without transporting unstable patients. Recent advances in ultrasound technology, such as high-frequency and portable devices and experimental wearable arrays, further underscore its potential for continuous or near-real-time bedside monitoring. Collectively, these developments highlight LUS as an attractive platform for integration with automated analysis tools, particularly in pediatric intensive care settings [5,6]. However, clinical adoption remains limited by two principal barriers. First, LUS interpretation exhibits high interobserver variability, even among experienced practitioners [7], limiting diagnostic consistency. Second, proficiency requires specialized training and substantial operator experience, creating a critical barrier to standardized implementation in routine clinical workflows [8].

Deep learning approaches, particularly convolutional neural networks (CNNs), have demonstrated promise for quantifying LUS image characteristics and reducing subjective interpretation variability in adult populations [8–10]. However, prior CNN models developed for LUS are fundamentally inappropriate for pediatric applications. Pediatric patients present unique anatomical and physiological characteristics—including relatively unossified thoracic structures, thinner subcutaneous tissues, and smaller thoracic dimensions—that substantially alter ultrasound image characteristics and acoustic propagation compared to adult thoraces. Consequently, models trained on adult LUS data do not generalize to pediatric populations. While a small number of recent studies have applied deep learning to pediatric LUS, these efforts remain limited in scope and clinical utility [11–13]. Critically, no prior work has developed an objective, automated, clinically validated approach to classify diverse lung pathologies in PICU patients using pediatric-specific models. This gap represents a significant unmet clinical need: a pediatric-specific, algorithm-based approach could simultaneously reduce radiation exposure, standardize interpretation across practitioners, and facilitate LUS integration into routine PICU workflows. We hypothesized that a pediatric-specific deep learning model could accurately segment clinically relevant LUS features from individual frames, thereby providing an objective and reproducible foundation for future automated classification of clinically relevant lung pathologies in critically ill pediatric patients.

2. Materials and Methods

2.1. Study Design and Data Collection

This proof-of-concept study evaluated deep learning models for automated pediatric lung ultrasound assessment in critically ill patients. Data were retrospectively collected from PICU patients (age 0–18 years) at Leiden University Medical Center between January and May 2022, yielding 506 four-second video clips. The study protocol (reference N22.019) was submitted to the Medical Ethics Committee Leiden Den Haag Delft (METC LDD). The committee concluded that the study does not fall within the scope of the Dutch Medical

Research Involving Human Subjects Act (WMO) and therefore did not require full ethical review. According to Dutch law (WGBO and GDPR) and the committee's guidance, the need to obtain specific informed consent from parents or legal guardians for the use of routinely collected clinical data was waived. This decision was based on the retrospective design, reliance on existing patient records, the impracticality of approaching all former patients, and the absence of more than minimal risk for the individuals involved. The research was carried out in line with the ethical standards set by the Declaration of Helsinki (1975, last amended in 2013).

All patients underwent systematic LUS examination as part of routine clinical care. Two independent trained observers acquired and compared videos anonymously. Systematic LUS examination was performed, in supine position, using two linear probes (3–20 MHz frequency range, Venue GO, GE Healthcare, Chicago, IL, USA) selected based on individual patient size. The thorax was divided into 12 standardized regions of investigation: anterior, lateral, and posterior areas of both the left and right chest walls (upper and lower divisions). Each region was scored by consensus of two or more observers on a 0–3 scale based on degree of pulmonary aeration loss, yielding a maximum total LUS score of 36. Videos showing pneumothorax were excluded, as underlying lung tissue cannot be assessed ultrasonographically. To ensure models could distinguish LUS from other ultrasound modalities, 20 arm ultrasound videos were collected and labeled as “indeterminable”.

2.2. LUS Feature Assessment and Scoring Criteria

Aeration status was classified using a standardized 4-point scale. Score 0 indicated normal aeration characterized by A-lines and ≤ 2 B-lines per view. Score 1 indicated moderate aeration loss with ≥ 3 well-separated B-lines or small focal consolidations. Score 2 indicated severe aeration loss with coalescent B-lines or consolidations occupying $< 50\%$ of visible parenchyma. Score 3 indicated complete consolidation with $\geq 50\%$ aerated parenchyma loss, with or without bronchograms (Figure 1).

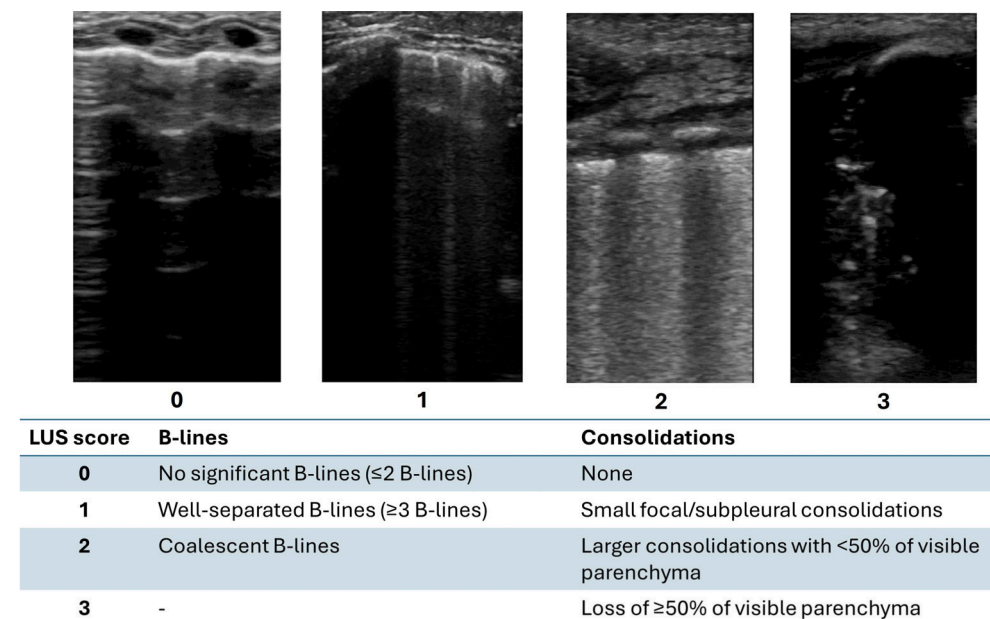


Figure 1. Lung aeration status classified by standardized 4-point scale.

2.3. Image Annotation and Data Preparation

From all collected videos, 800 frames were randomly selected for model training and evaluation. After scoring and consensus between the observers, each frame was manually

annotated for four clinical LUS features: A-lines, B-lines, consolidations, and atelectasis by a trained observer. Semantic segmentation was performed at the pixel level using Label Studio (Heartex Labs, San Francisco, CA, USA; <https://github.com/heartexlabs/label-studio>, accessed on 15 July 2022), generating separate binary masks for each clinical feature. All segmentation masks were created by one trained observer, reviewed by a second experienced observer, and reconciled by consensus. No formal interobserver variability analysis (e.g., κ or intraclass correlation coefficient) was performed.

For segmentation tasks, frames were resized to 256×512 pixels and pixel values normalized to $[-1, 1]$. No data augmentation was applied to the segmentation training set, reflecting a conservative design choice to preserve the exact spatial relationship between pleural structures and artefacts in pediatric LUS and to focus this proof-of-concept on baseline feasibility rather than optimization. The resulting dataset was randomly divided into training and test sets at the frame level, meaning that frames from the same video or patient could, in principle, be present in both sets (Table 1).

Table 1. Hyperparameter configurations for cGAN segmentation model.

cGAN (Segmentation)	
Dataset split	80% train/20% test
Training samples	640 frames
Testing samples	160 frames
Frames per video	Full 256×512
Input dimensions	256×512 pixels
Normalization range	$[-1, 1]$
Batch size	64
Total epochs	120
Optimizer	Adam
Learning rate (initial)	0.0002
Loss function	Binary cross-entropy (both G and D)
Data augmentation	None
Evaluation metrics	Dice Similarity Coefficient (DSC), Mean Squared Error (MSE)

2.4. Model Development and Architecture

Generative Adversarial Network for Semantic Segmentation

A conditional GAN (cGAN) was implemented to perform pixel-level semantic segmentation of the four clinical LUS features. The architecture consisted of a generator network that produced pixel-wise segmentation masks and a discriminator network that discriminated real annotations from generated masks.

The generator network received the input frame (256×512 pixels) and produced a multi-channel output, with one channel per clinical feature (A-lines, B-lines, consolidations, atelectasis). The architecture combined encoder–decoder components with connections between corresponding encoder and decoder levels to preserve spatial fine details.

The discriminator network classified whether patches or regions of the output masks were realistic (matching ground truth) or generated, enabling the model to distinguish high-frequency structural details. Both networks were trained in an adversarial manner, with the generator optimized to fool the discriminator while minimizing pixel-wise error against ground truth masks.

The frame-based segmentation model was trained to predict pixel-wise masks of clinically relevant lung ultrasound features from individual B-mode frames extracted from pediatric LUS videos. Video files were first decomposed into still frames, and corresponding expert-annotated segmentation masks were stored in a separate directory with strict filename ordering to preserve one-to-one correspondence between input frames and target masks. Using a dedicated preparation script, paired frame–mask images were resized to

a fixed resolution of 256×512 pixels, converted to three-channel arrays, and normalized from 0–255 to a range of -1 to 1 before being saved as compressed NumPy arrays for model training. The resulting dataset comprised source images (LUS frames) and target images (segmentation masks), which were subsequently partitioned into training and validation sets at the frame level for supervised learning.

For segmentation, we implemented a conditional generative adversarial network following the pix2pix image-to-image translation framework, using Keras with a TensorFlow backend. The discriminator received the concatenated LUS frame and corresponding mask as input and consisted of a PatchGAN architecture with successive convolutional blocks (C64–C128–C256–C512) using 4×4 kernels, stride 2, LeakyReLU activations, batch normalization, and a final 1-channel convolution with sigmoid activation to output a patch-level real/fake map; it was optimized with binary cross-entropy loss and the Adam optimizer (learning rate 0.0002, $\beta_1 = 0.5$). The generator was a U-Net-like encoder–decoder built from down-sampling encoder blocks (strided convolutions with LeakyReLU and batch normalization) and up-sampling decoder blocks (transposed convolutions with batch normalization, optional dropout, ReLU activation, and skip connections), mapping $256 \times 512 \times 3$ LUS inputs to $256 \times 512 \times 3$ segmentation outputs. During training, real frame–mask pairs and generated pairs were alternately used to update the discriminator, while the composite GAN (generator plus frozen discriminator) was trained to minimize a weighted sum of adversarial loss and pixel-wise reconstruction loss on the target masks; model performance over epochs was monitored using loss curves and periodic visual inspection of generated segmentation maps, and final models were evaluated by comparing generated masks to ground truth using mean squared error and Dice similarity coefficient, including analyses restricted to the central lung region via cropping.

To mitigate overfitting given the relatively small dataset, we employed multiple regularization strategies. Batch normalization layers were included throughout the generator and discriminator to stabilize training, and dropout was applied to selected decoder layers of the generator. The cGAN was trained using an 80/20 frame-level split, and model selection was based on performance on the held-out test set, assessed using the Dice similarity coefficient and mean squared error. Hyperparameters, including learning rate and network depth, were chosen based on prior pix2pix implementations and refined through empirical experimentation during preliminary training runs.

2.5. Computational Resources and Implementation

All models were implemented in Python (version 3.10.0; Python Software Foundation, Wilmington, DE, USA; <https://www.python.org>) using TensorFlow (version 2.8.0; Google LLC, Mountain View, CA, USA; <https://www.tensorflow.org>) and Keras (version 2.8.0; <https://keras.io>). Model training employed segmentation (cGAN): NVIDIA A40 (48 GB VRAM; NVIDIA Corp., Santa Clara, CA, USA). Complete source code, trained model weights, hyperparameter configurations, and training logs are publicly available at https://github.com/TharanghiLogendran/LUS_assessment (accessed on 24 June 2026), enabling full reproducibility of all reported results.

2.6. Model Evaluation and Performance Metrics

The segmentation model (cGAN) was evaluated on Dice Similarity Coefficient (DSC; measures spatial overlap between predicted and ground truth masks) and Mean Squared Error (MSE; pixel-wise error between predicted and ground truth masks). For each saved model, the DSC and MSE are reported as mean $\pm$ standard deviation across all images in the held out test set. The cGAN was trained exclusively on the training set (80% of the data), and model performance was evaluated on the held out 20% test set only.

2.7. Statistical Analysis

Statistical analyses were performed using SPSS® version 29 (IBM Corp., Armonk, NY, USA). Continuous data was tested for normal distribution with the Shapiro–Wilk test. The median and interquartile range (IQR) were reported for continuous variables. Categorical data were reported as numbers and percentages. Classification performance metrics were computed at the video level. Segmentation metrics were computed at the frame level by comparing the predicted and ground-truth masks over the entire image for each frame. Dice similarity coefficient and mean squared error were calculated as mean $\pm$ standard deviation across all frames in the held-out test set. No cross-validation was performed in order to preserve a single independent test set.

3. Results

3.1. Patients' Characteristics

For model development, images from 31 patients were used. Most were admitted for post-operative care after cardiac surgery. Median age of the patients at first LUS examination was 2.5 months (IQR 0–11 months). A total of 67 LUS examinations (69%) were performed while the patient was supported by invasive mechanical ventilation with a median LUS score of 12 (IQR 8–17). In 31% of patients, signs of a pulmonary infection were present at the time of LUS examination (Table 2). Across all patients, there were 98 LUS assessments, which resulted in 506 4-s videos that were used for model development.

Table 2. Patient characteristics.

Characteristics	
Total number of patients	n = 31
Age at LUS (months, IQR)	2.5 (0–11)
Gender (male, %)	22 (71%)
Diagnosis at admission (n, %)	
Respiratory insufficiency	6 (19%)
Circulatory insufficiency	5 (16%)
Post-operative after cardiac surgery	19 (61%)
Other	1 (3%)
Total LUS examinations	n = 98
Respiratory support at time of LUS (n, %)	
None	2 (2%)
Oxygen therapy	7 (7%)
HFNC	14 (14%)
CPAP	5 (5%)
IMV	67 (69%)
IMV—ECMO	3 (3%)
Pulmonary infection at time of LUS (yes, %)	30 (31%)
LUS—Lung Ultrasound Score (IQR)	12 (8–17)

LUS—lung ultrasound, IQR—interquartile range, HFNC—high flow nasal cannula, CPAP—continuous positive airway pressure, IMV—invasive mechanical ventilation, ECMO—extracorporeal membrane oxygenation.

3.2. Frame-Based Segmentation

The generator loss after each training batch is shown in Figure 2 and demonstrates convergence after 600 batches. Overall segmentation performance on the held-out test set was high. The two best-performing models achieved mean Dice similarity coefficients of 0.97 ± 0.03 and 0.96 ± 0.03 , with corresponding mean MSE values of 0.025 ± 0.025 and 0.030 ± 0.026 , respectively. For both models, generative loss at convergence was, respectively, 1.8 and 1.9. Figure 3 illustrates representative examples of LUS frames alongside

expert-annotated masks and model-generated masks, demonstrating that the cGAN captures the main clinical features with high spatial fidelity.

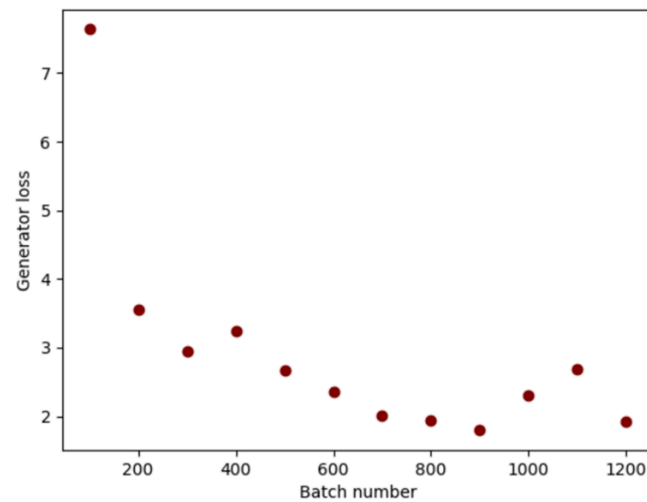


Figure 2. Generator loss of the cGAN after each training batch.

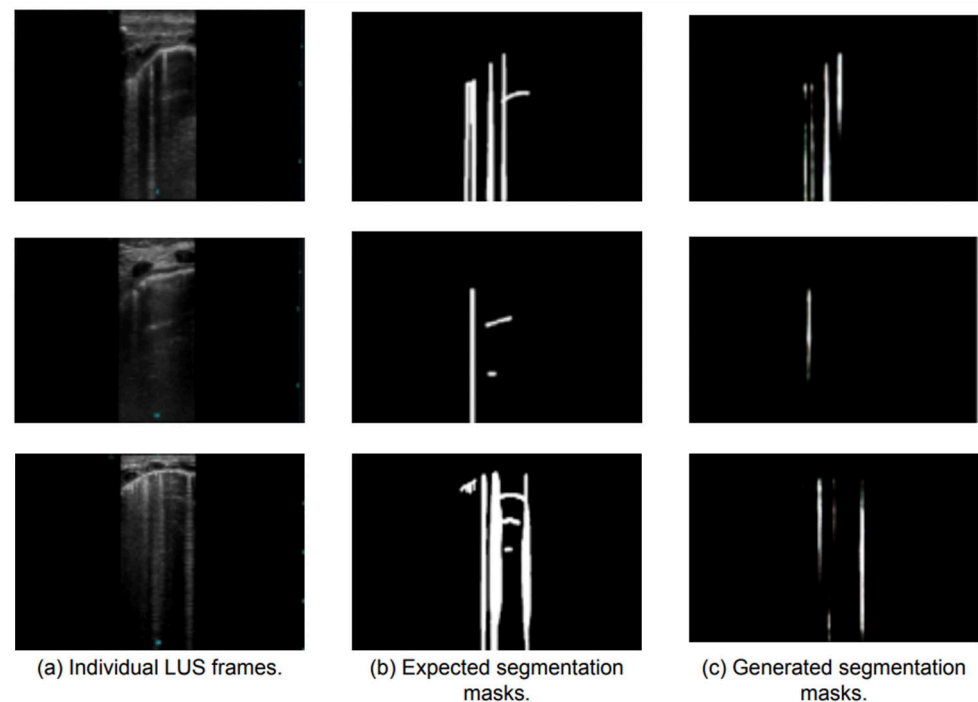


Figure 3. Examples of frame-based segmentation. (a) Original LUS frames. (b) Expert-annotated masks. (c) Model-generated masks. The first two rows illustrate a typical successful segmentation, whereas the third row shows a failure case in which the model underestimates the number of B-lines in a frame with motion artefacts and incomplete pleural visualization.

4. Discussion

In this study, we describe a generative adversarial network to identify A-lines, B-lines, consolidation, and atelectasis from individual pediatric LUS frames. The predictions made by the model were aimed to assist physicians in their interpretation of LUS images to assess the regional aeration of the lungs. A previous study by Nhat et al. describes this need and potential contribution of the use of artificial intelligence (AI)-assisted LUS in a resource-limited ICU setting [14].

The results of the frame-based segmentation were high: 0.97 ± 0.03 and 0.96 ± 0.03 Dice score and 0.025 ± 0.025 and 0.030 ± 0.026 mean squared error for the best per-

forming models. This high accuracy implies a high similarity between the desired vs generated segmentation masks. This is in accordance with adult data, as demonstrated by Cheng et al., with an average DSC between 78% and 86% using transfer learning methods on pre-trained networks, as well as earlier reported DSC's between 64–75% and 75–90% [15–17]. However, the higher DSC values observed in our study likely reflect methodological differences—namely, a smaller dataset and frame-level data splitting—rather than a genuine performance advantage over adult LUS models. Accordingly, the comparison to adult studies should be viewed as contextual rather than competitive, and our results interpreted as preliminary, upper-bound estimates of segmentation performance.

In addition, DSC and MSE primarily capture pixel-wise overlap and may overestimate clinical performance because isolated B-lines or small consolidations occupy only a small fraction of the image. Missing a few B-lines can leave the DSC largely unchanged while substantially altering the regional aeration score and downstream clinical interpretation. Consequently, high DSC values do not guarantee detection of all clinically relevant features. Visual review also revealed characteristic failure modes: performance deteriorated in the presence of poor image quality (motion artefacts, incomplete pleural visualization, suboptimal probe contact, rib shadowing) and postoperative anatomical alterations. The model therefore performs best under standardized acquisition conditions, and rigorous image-quality control is essential.

The anatomical and physiological characteristics of pediatric patients likely influenced model behavior and further underscore the need for pediatric-specific training data. The relatively unossified thorax and thin subcutaneous tissues in infants modify acoustic propagation and artefact patterns compared with adults, resulting in more prominent B-lines and distinct pleural signatures. Smaller thoracic dimensions and higher respiratory rates also produce more rapid motion and a larger proportion of lung parenchyma within a single view. Together, these factors help explain why models trained on adult data do not generalize well to children and support the development of dedicated pediatric LUS networks.

These results are promising for the future use of AI in pediatric LUS, as such tools may support more accurate and efficient interpretation in training settings and/or clinical settings, such as diagnosing pneumonia, bronchiolitis, and neonatal respiratory distress syndrome [18].

As mentioned above, LUS is a skill that should be trained, and therefore, it is only understandable that there is a certain learning curve in both collecting images and scoring them. In this proof-of-concept study, all LUS examinations were acquired and scored by trained observers, and disagreements in scoring or segmentation were resolved by consensus. However, we did not perform a formal inter- or intra-observer variability analysis using intraclass correlation coefficients or kappa statistics. Therefore, there is a possibility that there was an intra-observer variability when labeling frames. For example, Kumar et al. demonstrated an interobserver reliability of 79% when determining the total B-line count per scan and 56% for lateral consolidations and 86% for posterior consolidations [19]. The interobserver reliability was lowest for bilateral consolidations (28%). Another study demonstrated the intraclass correlation coefficient (ICC) for clinical experts and a neural network when counting B-lines from LUS images [20]. The agreement between their visual and automatic assessment was 79%, with an ICC among clinical experts between 62–99%. For the automatic system the ICC was 49–83%. Therefore, it is recommended to compare the identification of clinical features by the model with labeled masks of multiple clinicians—not only to measure the performance of the model, but also to depict where disagreements occur.

Compared to previous studies, we have found higher agreement between the expected (actual) and generated segmentation masks. Even though there is still the possibility that

we have missed clinically important features, such as B-lines, a single or a few isolated B-lines constitute only a small percentage of the overall LUS image. As a result, even if the model fails to localize single or multiple B-lines—which would possibly have clinical implications—DSC would still remain relatively high. A solution for this problem may be to not localize clinical features by segmentation, but by object detection with bounding boxes. In that case, the number of clinical features would be detected, rather than the number of pixels per frame. As the LUS score is dependent on B-line counts, feature extraction is a common method in studies on DL for LUS assessment. However, B-lines vary in thickness, and studies have shown that the underlying pathologies can be distinguished by B-line shapes. B-line appearances are heterogeneous based on the etiology: B-lines generated by a fibrotic or inflammatory lung have a different appearance from those generated by cardiogenic edema, and therefore, B-line detection alone is not sufficient, and the appearance should be analyzed [7]. Therefore, though the chosen method for identification of clinical LUS features is not ideal, it seems the most optimal and promising.

Deployment in a clinical setting is an important next step to investigate in the future. An achievable and safe start would be in a supervised training environment rather than immediate real-time bedside use. In such a setting, recorded LUS clips could be processed offline, with the GAN-generated segmentations shown alongside the original images to support education and standardization of scoring. Once performance and usability are established, integration into ultrasound devices or bedside workstations could enable near real-time overlays of detected features. This stepwise approach—from offline training use to real-time clinical decision support—may facilitate safe implementation while technical integration and regulatory approval are being pursued.

This study has several limitations. Firstly, it is a retrospective study which may have inherent biases. Secondly, the evaluation of LUS was performed on examinations from a relatively low number of unique patients in a single center. While our results are consistent with previously published findings, training and validating our proposed system on larger data sets would likely produce more accurate outcomes. We did not apply or systematically evaluate data augmentation, which likely limits robustness and should be addressed in future work. A further methodological limitation is the use of a frame-level rather than a patient-level data split. Frame-level splitting can yield overly optimistic estimates of generalization and is not the standard for clinical evaluation of medical imaging AI. Together with the fact that some artefacts (e.g., A-line-dominated regions or large non-lesion areas) occupy substantial portions of the image, this may inflate global DSC despite missed fine-grained features. We also report aggregate DSC across all annotated structures without a systematic feature-wise analysis (e.g., separate performance for A-lines, isolated B-lines, or small consolidations) or analysis in relation to class frequency. As a consequence, the reported Dice scores should be interpreted as upper-bound estimates of segmentation performance rather than definitive measures of clinical accuracy on unseen patients, and our ability to fully characterize clinical utility for sparse or subtle findings is limited. Future work will therefore employ strict patient-level splits, external validation cohorts, and detailed class-specific reporting to obtain more unbiased and clinically informative estimates of generalizability. Overall, the reported metrics should be interpreted as preliminary and likely optimistic, underscoring the need for larger, more diverse, prospectively collected cohorts to robustly assess performance across the full spectrum of PICU patients.

Thirdly, LUS images were obtained using two different ultrasound machines and probes due to differences in body size based on varying ages. However, this is routine in the current clinical setting, the preset settings were similar, and the inter-rater agreement for LUS is known to be high, even when using different probes [21]. Finally, we did not

correlate the automated assessments with traditional imaging modalities or final clinical diagnoses. Because this proof-of-concept study primarily focused on technical segmentation performance (DSC, MSE) in a small, heterogeneous PICU cohort, robust diagnostic validation was not feasible and should be addressed in future prospective studies.

5. Conclusions

In this proof-of-concept study, the cGAN-based model was able to segment key features from individual pediatric LUS frames with high global overlap metrics. The model can localize important features relevant for lung aeration assessment and may provide a technical foundation for future automated scoring and decision-support tools. However, given the small, single-center cohort, frame-level data split, and lack of clinical validation, these findings should be regarded as preliminary. Further advancements and validation in larger, prospectively collected cohorts will be required before clinical implementation can be considered and before the model can be used to facilitate the education of less experienced ultrasonographers.

Author Contributions: Conceptualization, S.G.J.H., T.L. and C.O.T.; methodology, T.L. and C.O.T.; software, T.L. and C.O.T.; validation, S.G.J.H., T.L. and A.W.; formal analysis, T.L. and S.G.J.H.; investigation, S.G.J.H., T.L. and A.W.; resources, S.G.J.H. and T.L.; data curation, S.G.J.H., T.L. and A.W.; writing—original draft preparation, S.G.J.H. and T.L.; writing—review and editing, S.G.J.H., T.L., A.W. and C.O.T.; visualization, T.L.; supervision, A.W. and C.O.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was approved by the Medical Research Ethics Committee Leiden Den Haag Delft (N22.019) on 6 May 2022 with waiver of informed consent.

Informed Consent Statement: Due to the retrospective character of the study, the study was exempted by the ethics committee.

Data Availability Statement: All models were implemented in Python version 3.10.0, and the code can be found at https://github.com/TharanghiLogendran/LUS_assessment (accessed on 24 June 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CXR	Chest X-ray
CNN	Convolutional Neural Network
DSC	Dice Similarity Coefficient
cGAN	Conditional Generative Adversarial Network
ICC	Intraclass Correlation Coefficient
IQR	Interquartile Range
LUMC	Leiden University Medical Center
LUS	Lung Ultrasound
NICU	Neonatal Intensive Care Unit
MSE	Mean Squared Error
PICU	Pediatric Intensive Care Unit

References

1. Brenner, D.; Elliston, C.; Hall, E.; Berdon, W. Estimated risks of radiation-induced fatal cancer from pediatric CT. *AJR Am. J. Roentgenol.* **2001**, *176*, 289–296. [[CrossRef](#)] [[PubMed](#)]

2. Mahendra, M.; Chu, P.; Amin, E.K.; Nawaytou, H.; Duncan, J.R.; Fineman, J.R.; Smith-Bindman, R. Associated radiation exposure from medical imaging and excess lifetime risk of developing cancer in pediatric patients with pulmonary hypertension. *Pulm. Circ.* **2023**, *13*, e12282. [[CrossRef](#)] [[PubMed](#)]
3. Pearce, M.S.; Salotti, J.A.; Little, M.P.; McHugh, K.; Lee, C.; Kim, K.P.; Howe, N.L.; Ronckers, C.M.; Rajaraman, P.; Sir Craft, A.W.; et al. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: A retrospective cohort study. *Lancet* **2012**, *380*, 499–505. [[CrossRef](#)] [[PubMed](#)]
4. Rodriguez-Fanjul, J.; Benet, N.; Rodrigo Gonzalo de Liria, C.; Porta, R.; Guinovart, G.; Bobillo-Perez, S. Lung ultrasound protocol decreases radiation in newborn population without side effects: A quality improvement project. *Med. Intensiv. (Engl. Ed.)* **2023**, *47*, 16–22. [[CrossRef](#)] [[PubMed](#)]
5. Kim, M.G.; Yoon, C.; Lim, H.G. Recent Advancements in High-Frequency Ultrasound Applications from Imaging to Microbeam Stimulation. *Sensors* **2024**, *24*, 6471. [[CrossRef](#)] [[PubMed](#)]
6. Weik, D.; Nauber, R.; Kaiser, E.; Kirsch, N.; Kunz, R.; Schierling, L.; Leitner, C.; Benini, L.; Liu, H.C.; Zhou, Q.; et al. Current Trends in Ultrasound Wearables: Spotlight on System Architecture. *IEEE Rev. Biomed. Eng.* **2026**, 1–21. [[CrossRef](#)] [[PubMed](#)]
7. Musolino, A.M.; Toma, P.; De Rose, C.; Pitaro, E.; Boccuzzi, E.; De Santis, R.; Morello, R.; Supino, M.C.; Villani, A.; Valentini, P.; et al. Ten Years of Pediatric Lung Ultrasound: A Narrative Review. *Front. Physiol.* **2021**, *12*, 721951. [[CrossRef](#)] [[PubMed](#)]
8. Hong, D.; Choi, H.; Hong, W.; Kim, Y.; Kim, T.J.; Choi, J.; Ko, S.B.; Park, C.M. Deep-learning model accurately classifies multi-label lung ultrasound findings, enhancing diagnostic accuracy and inter-reader agreement. *Sci. Rep.* **2024**, *14*, 22228. [[CrossRef](#)] [[PubMed](#)]
9. Arntfield, R.; VanBerlo, B.; Alaifan, T.; Phelps, N.; White, M.; Chaudhary, R.; Ho, J.; Wu, D. Development of a convolutional neural network to differentiate among the etiology of similar appearing pathological B lines on lung ultrasound: A deep learning study. *BMJ Open* **2021**, *11*, e045120. [[CrossRef](#)] [[PubMed](#)]
10. Wang, X.; Burzynski, J.S.; Hamilton, J.; Rao, P.S.; Weitzel, W.F.; Bull, J.L. Quantifying lung ultrasound comets with a convolutional neural network: Initial clinical results. *Comput. Biol. Med.* **2019**, *107*, 39–46. [[CrossRef](#)] [[PubMed](#)]
11. Correa, M.; Zimic, M.; Barrientos, F.; Barrientos, R.; Roman-Gonzalez, A.; Pajuelo, M.J.; Anticona, C.; Mayta, H.; Alva, A.; Solis-Vasquez, L.; et al. Automatic classification of pediatric pneumonia based on lung ultrasound pattern recognition. *PLoS ONE* **2018**, *13*, e0206410. [[CrossRef](#)] [[PubMed](#)]
12. Kessler, D.; Zhu, M.; Gregory, C.R.; Mehanian, C.; Avila, J.; Avitable, N.; Coneybeare, D.; Das, D.; Dessie, A.; Kennedy, T.M.; et al. Development and testing of a deep learning algorithm to detect lung consolidation among children with pneumonia using hand-held ultrasound. *PLoS ONE* **2024**, *19*, e0309109. [[CrossRef](#)] [[PubMed](#)]
13. Magrelli, S.; Valentini, P.; De Rose, C.; Morello, R.; Buonsenso, D. Classification of Lung Disease in Children by Using Lung Ultrasound Images and Deep Convolutional Neural Network. *Front. Physiol.* **2021**, *12*, 693448. [[CrossRef](#)] [[PubMed](#)]
14. Nhat, P.T.H.; Van Hao, N.; Tho, P.V.; Kerdegari, H.; Pisani, L.; Thu, L.N.M.; Phuong, L.T.; Duong, H.T.H.; Thuy, D.B.; McBride, A.; et al. Clinical benefit of AI-assisted lung ultrasound in a resource-limited intensive care unit. *Crit. Care* **2023**, *27*, 257. [[CrossRef](#)] [[PubMed](#)]
15. Dorothy Cheng, E.Y.L. Transfer Learning U-Net Deep Learning for Lung Ultrasound Segmentation. *arXiv* **2021**, arXiv:2110.02196. [[CrossRef](#)]
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 Trans. Med. Imaging* **2020**, *39*, 2676–2687. [[CrossRef](#)] [[PubMed](#)]
17. Frey, B.; Zhao, L.; Fong, T.C.; Lediju Bell, M. *Multi-Stage Investigation of Deep Neural Networks for COVID-19 B-Line Feature Detection in Simulated and In Vivo Ultrasound Images*; SPIE: Bellingham, WA, USA, 2022; Volume 12033. [[CrossRef](#)]
18. Chu, D.; Liteplo, A.; Duggan, N.; Hutchinson, A.B.; Shokoohi, H. Artificial Intelligence in Lung Ultrasound. *Curr. Pulmonol. Rep.* **2024**, *13*, 127–134. [[CrossRef](#)]
19. Kumar, A.; Weng, Y.; Graglia, S.; Chung, S.; Duanmu, Y.; Lalani, F.; Gandhi, K.; Lobo, V.; Jensen, T.; Nahn, J.; et al. Interobserver Agreement of Lung Ultrasound Findings of COVID-19. *J. Ultrasound Med.* **2021**, *40*, 2369–2376. [[CrossRef](#)] [[PubMed](#)]
20. Short, J.; Acebes, C.; Rodriguez-de-Lema, G.; La Paglia, G.M.C.; Pavon, M.; Sanchez-Pernaute, O.; Vazquez, J.C.; Romero-Bueno, F.; Garrido, J.; Naredo, E. Visual versus automatic ultrasound scoring of lung B-lines: Reliability and consistency between systems. *Med. Ultrason.* **2019**, *21*, 45–49. [[CrossRef](#)] [[PubMed](#)]
21. Biasucci, D.G.; Buonsenso, D.; Piano, A.; Bonadia, N.; Vargas, J.; Settanni, D.; Bocci, M.G.; Grieco, D.L.; Carnicelli, A.; Scoppettuolo, G.; et al. Lung ultrasound predicts non-invasive ventilation outcome in COVID-19 acute respiratory failure: A pilot study. *Minerva Anestesiol.* **2021**, *87*, 1006–1016. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.