DOI 10.5162/Ultrasonic2025/P1.20

A Deep Learning Segmentation Approach for Lung Ultrasound Scoring Classification

M. Muñoz[1,2], X. Han[3], L. Demi[3], T. Perrone[4], A. Smargiassi[5], R. Inchingolo[5], Y. Tung Chen[6], A. Trueba Vicente[7], and J. Camacho[1]

[1]Institute for Physical and Information Technologies, Spanish National Research Council, Madrid, Spain

[2]Electronic Department,Universidad de Alcalá, Madrid,Spain

[3]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

[4]Department of Internal Medicine, IRCCS San Matteo, Pavia, Italy

[5]UOC Pneumologia, Dipartimento Neuroscienze, Organi di Senso e Torace, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy

[6]Department of Internal Medicine, Hospital Universitario La Paz, Madrid, Spain

[7]Department of Internal Medicine, Hospital de Emergencias Enfermera Isabel Zendal, Madrid, Spain

mario.munoz.prieto@csic.es

Abstract: This work presents a Deep Learning method to automate the interpretation of lung ultrasound (LUS), aiming to reduce diagnostic subjectivity. A segmentation model was developed to first identify key artifacts, such as vertical artifacts or consolidations, and then calculate a corresponding severity score. Its performance was benchmarked against a classification model across two video datasets. The segmentation model achieved comparable accuracy to the traditional classification method. Furthermore, the approach proved robust to variations in the ultrasound probe's orientation.

Keywords: Lung Ultrasound (LUS), Deep Learning, Image Segmentation, Computer-aided diagnosis, Automated Scoring.

Introduction

Lung ultrasound (LUS) has rapidly evolved into an essential, non-invasive imaging tool for assessing a variety of pulmonary conditions. Its interpretation, however, is not based on a direct anatomical view but on sonographic artifacts that arise from the interaction between ultrasound waves and the lung parenchyma. These artifacts include horizontal A-lines, which are reverberations of the pleura indicative of a normal lung, and vertical B-lines, which appear when alveolar air is displaced by fluid, suggesting pathological conditions like pneumonia. As a disease progresses, tissue can solidify, appearing as hypoechoic consolidations. The correct identification of these patterns requires significant expertise and is susceptible to inter-observer variability, which limits the broader clinical application of the technique.

To address these challenges, Artificial Intelligence (AI) has emerged as a powerful tool to aid diagnosis by helping less experieced clinicians and reducing subjectivity. In clinical practice, physicians quantify lung involvement using a scoring system based on the presence and extent of key sonographic artifacts. This study adopts the 4-level scoring criteria proposed by Soldati et al. [1]:

- **Score 0 (Normal):** Characterized by a continuous, regular pleural line with the presence of horizontal A-line artifacts.
- **Score 1 (Mild):** Defined by the presence of vertical artifacts (B-lines) while the pleural line remains intact and unbroken.
- **Score 2 (Moderate):** Corresponds to a broken or irregular pleural line with confluent vertical artifacts affecting less than 50% of its length. Small consolidations may also be present.
- **Score 3 (Severe):** Indicates severe abnormalities, characterized by widespread, confluent vertical artifacts affecting more than 50% of the visible pleura, which may be accompanied by extensive consolidations.

Automating this process with AI has primarily followed two real-time paradigms:

classification models (CM), which are trained to directly predict a severity score from an image, and segmentation models (SM), which first delineate the

DOI 10.5162/Ultrasonic2025/P1.20

artifacts themselves and then, calculates the score based on these findings. In this work, we propose and validate a novel workflow that translates the rich output of an artifact segmentation model into a clinical severity score. The performance of this segmentation-to-score approach is evaluated against expert clinician annotations and benchmarked against a classification model trained specifically for the scoring task on a multi-center, multi-scanner dataset.

Datasets and Acquisition Protocols

This international, multi-center study was a collaborative effort analyzing a general dataset of 2219 LUS videos from COVID-19 patients in Italy and Spain. All data was acquired in accordance with the Declaration of Helsinki and approved by the respective institutional ethical committees. The analysis was performed on two distinct datasets:

- **Dataset-1:** This dataset comprises 1530 videos from 83 patients, following a 14-region acquisition protocol [2]. The data was acquired using three different ultrasound scanners (Esaote Mylab50, Philips IU22, CerberoATL) with varying imaging configurations, including frequencies from 2.5 to 10 MHz and both convex and linear probes depending on the patient as ins explained in [3].
- **Dataset-2:** This dataset consists of 689 videos from 30 patients, following a 12-region acquisition protocol [4]. All acquisitions were performed with a single scanner (UltraCOV) using a 3.5 MHz convex probe and a standardized scanning criteria to minimize variability [5]. For each patient, examinations included both longitudinal and transversal probe orientations (337 and 352 videos, respectively).

AI Models and Scoring Methodology

Two deep learning models were evaluated.

Classification Model (CM). The classification model (CM) utilizes a ResNet18 architecture [6], a convolutional neural network known for its effectiveness in image classification. It was pre-trained on a large dataset of 58,924 LUS images acquired from the same scanner models present in Dataset-1 [7]. The model is designed to classify LUS images directly into the 4-level severity score (0-3)

Segmentation Model (SM). The segmentation model (SM) approach is a complete workflow that translates segmented artifacts into a clinical score (Fig. 1). It consists of the following steps:

- *Input Data & Model Architecture:* The model uses an Attention U-Net architecture [8] which was trained on 9,159 LUS images contained in Dataset-2. A key methodological choice was to use B-scan images (raw data of 128×256 pixels) as input instead of conventional sector images (Pre-processing block in Fig. 1), making the model more robust to scanner-specific geometries.
- *From Segmentation to Score:* The workflow translates the SM output into a score by mimicking clinical reasoning. First, the SM delineates the pleural line. Then, a post-processing algorithm calculates the percentage of the pleura affected by B-lines, by dividing the number of scan lines where B-lines are detected by the total number of scan lines where the pleura is visible [9]. Finally, this metric is mapped to the 4-level clinical severity score based on established guidelines avobe mentioned.

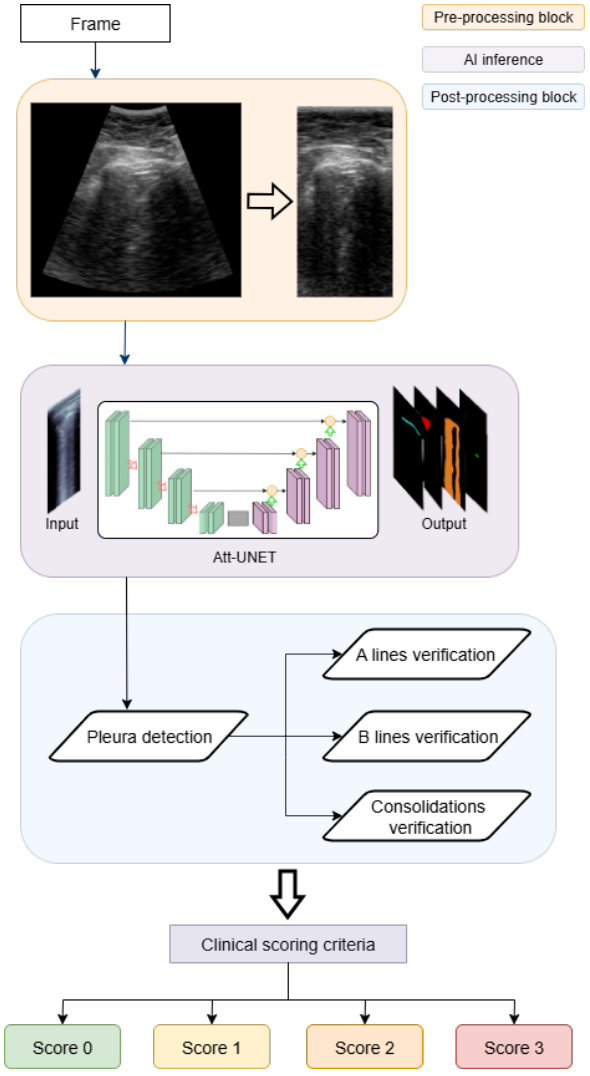

Fig. 1: From segmentation to score workflow.

DOI 10.5162/Ultrasonic2025/P1.20

Tab. 1: Video-level performance comparison of CM and SM methods. For Dataset-2, overall results are presented alongside a breakdown by probe orientation (longitudinal and transversal).

	Dataset-1		Dataset-2					
			CM			SM		
Metric	**CM**	**SM**	**General**	**Long**	**Trans**	**General**	**Long**	**Trans**
Accuracy	0.53	0.46	0.55	0.56	0.54	0.71	0.72	0.70
Accuracy ($\pm$1 tol.)	0.86	0.87	0.88	0.89	0.88	0.92	0.91	0.93
Cohen's Kappa (K_{qwc})	0.63	0.58	0.66	0.70	0.63	0.79	0.78	0.81

Experimental Setup and Evaluation Metrics

To assign a single score to each video, a 1% thresholding technique was employed [10]. This method identifies the highest severity score present in at least 1% of the video frames and assigns it to the entire video. The analysis was conducted at video level using as primary metric for assessing agreement with expert clinicians the Quadratic Weighted Cohen's Kappa (K_{qwc}), which measures inter-rater agreement while accounting for chance [11]. Also 1 error tolerance accuracy is performed to account for potential inter-observer variability in the annotations.

Results

The video-level performance of the Classification Model (CM) and Segmentation Model (SM) method are summarized in Table 1. Both method demonstrated good performance on the standardized Dataset-2, achieving substantial agreement than the CM, particularly in transversal acquisitions. A key finding is the robustness of both models to probe orientation, with performance on longitudinal and transversal views being highly comparable, confirming the method's flexibility in a clinical setting.

Discussion

The primary finding of this study is that the segmentation-based workflow (SM) demonstrates a prognostic capability comparable to a dedicated classification model (CM) in both datasets. The significance of this result lies in the inherent interpretability of the SM approach. Different from a "black box" classification model, the SM workflow provides a severity score based on quantifiable metrics, such as the percentage of the pleura affected by B-lines, which directly mimics the diagnostic reasoning a clinician would apply. Furthermore, the equivalent performance of both models on longitudinal and transversal acquisitions indicates a successful generalization of the problem, confirming that the AI is robust to this key acquisition variable. This would have positive implications for clinical practice, offering greater flexibility during the examination.

A key finfing from our analysis is the fundamental impact of data acquisition standardization on AI performance. A clear contrast in results was observed between the heterogeneous, multi-scanner Dataset-1 and the standardized, single-scanner Dataset-2. Although the CM was trained on a dataset with same scanners to those in Dataset-1, its significant performance improvement on Dataset-2 strongly suggests that variations in image quality, likely caused by different hardware and post-processing filters, are significant barriers to generalizability. This leads to the conclusion that standardizing image acquisition protocols may be as crucial as the AI architecture itself for achieving reliable and clinically translatable results.

Despite these promising findings, several limitations of this study must be acknowledged. Its retrospective nature means we could not control for potential selection bias. Another significant methodological limitation is the potential for data leakage in the SM results on Dataset-2, as the segmentation model was trained on images from 27 of the 30 patients that comprise this dataset, however, these results are presented to ensure a comprehensive study that includes all possible model and dataset combinations despite this known constraint.

Conclusions

This study demonstrates that a segmentation-based workflow can be effectively repurposed for severity scoring in lung ultrasound, achieving a prognostic accuracy comparable to that of a model trained specifically for classification. Furthermore, the results indicate that while these AI algorithms are robust to variations in probe orientation, their reliable performance is fundamentally dependent on high-quality, standardized image acquisition. This underscores that the successful clinical translation of AI-assisted diagnosis in LUS depends not only on algorithmic innovation, but also on the promotion of consistent clinical protocols.

References

[1] G. Soldati, A. Smargiassi, R. Inchingolo, et al. "Proposal for International Standardization of the Use of Lung Ultrasound for Patients With COVID-19: A Simple, Quantitative, Reproducible Method". In: *Journal of Ultrasound in Medicine* 39.7 (2020), pp. 1413–1419. DOI: 10.1002/jum.15285.

[2] L. Demi, F. Mento, A. Di Sabatino, et al. "Lung Ultrasound in COVID-19 and Post-COVID-19 Patients, an Evidence-Based Approach". In: *Journal of Ultrasound in Medicine* 41.9 (2022), pp. 2203–2215. DOI: 10.1002/jum.15902.

[3] L. Demi et al. "Lung Ultrasound in COVID-19 and Post-COVID-19 Patients, an Evidence-Based Approach". In: *Journal of Ultrasound in Medicine* 41.9 (Sept. 2022), pp. 2203–2215. DOI: 10.1002/jum.15902.

[4] Y. Tung-Chen, S. Ossaba-Vélez, K. S. Acosta Velásquez, et al. "The Impact of Different Lung Ultrasound Protocols in the Assessment of Lung Lesions in COVID-19 Patients: Is There an Ideal Lung Ultrasound Protocol?" In: *Journal of Ultrasound* 25.4 (2022), pp. 483–491. DOI: 10.1007/s40477-021-00610-x.

[5] J. Camacho, M. Muñoz, V. Genovés, et al. "Artificial Intelligence and Democratization of the Use of Lung Ultrasound in COVID-19: On the Feasibility of Automatic Calculation of Lung Ultrasound Score". In: *International Journal of Translational Medicine* 2.1 (2022), pp. 17–25. DOI: 10.3390/ijtm2010002.

[6] K. He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[7] U. Khan, S. Afrakhteh, F. Mento, et al. "Benchmark methodological approach for the application of artificial intelligence to lung ultrasound data from COVID-19 patients: From frame to prognostic-level". In: *Ultrasonics* 132 (2023), p. 106994. DOI: 10.1016/j.ultras.2023.106994.

[8] O. Oktay et al. "Attention U-Net: Learning Where to Look for the Pancreas". In: *arXiv preprint arXiv:1804.03999* (2018).

[9] M. Muñoz et al. "Deep Learning-Based Algorithms for Real-Time Lung Ultrasound Assisted Diagnosis". In: *Applied Sciences* 14.24 (2024), p. 11930. DOI: 10.3390/app142411930.

[10] F. Mento et al. "Deep learning applied to lung ultrasound videos for scoring COVID-19 patients: A multicenter study". In: *The Journal of the Acoustical Society of America* 149.5 (2021), p. 3626. DOI: 10.1121/10.0004855.

[11] J. Cohen. "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit". In: *Psychological Bulletin* 70.4 (1968), pp. 213–220. DOI: 10.1037/h0026256.