Computers in Biology and Medicine 198 (2025) 111244

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Lung ultrasound video scoring using a novel motion-aware segmentation technique: Toward automated neonatal LUS scoring

Hamed Jalilian [a,1], Sajjad Afrakhteh [a,1], Federico Mento [a], Emanuela Zannin [b], Camilla Rigotti [b], Federico Cattaneo [b], Giulia Dognini [b], Maria Luisa Ventura [b], Libertario Demi [a,*]

[a] Department of Information Engineering and Computer Science, University of Trento, Italy
[b] Fondazione IRCCS San Gerardo Dei Tintori Monza, Italy

ARTICLE INFO

Keywords:
K-means clustering
Lung ultrasound
Motion estimation
Scoring
Segmentation
Ultrasound

ABSTRACT

Lung ultrasound (LUS) is an essential tool for diagnosing lung diseases. However, its effectiveness is often limited by its reproducibility, making interpretation challenging for clinicians. LUS diagnosis typically relies on subjective assessments of pleural line and vertical artifacts. To address this limitation, we introduce a novel quantitative approach aimed at reducing the need to rely on human operators (HOs) for LUS data assessment (i.e., improving the reproducibility). In the first phase of our study, we propose a hybrid method that integrates motion estimation and K-means clustering for automated segmentation of LUS images. The technique utilizes K-means clustering to identify pleural line based on intensity variations, while motion estimation detects vertical artifacts by analyzing motion vectors between consecutive frames. Rather than employing a conventional learning-based classification model, we develop an interpretable scoring framework that assigns scores to individual video frames according to standard scoring criteria. A threshold-based approach is then applied to aggregate frame-level scores, determining the final score for each video. We evaluated our method on a clinical dataset comprising 420 neonatal LUS videos from 70 patients, with annotations provided by three HOs. When using the majority vote among HOs as the reference standard, our method achieved a video-level accuracy of 0.72. For cases with full agreement among HOs, accuracy improved to 0.77. These results demonstrate that our approach offers comparable or superior performance to state-of-the-art deep learning (DL)-based methods in terms of scoring consistency, while reducing the need for a huge training dataset.

1. Introduction

Ultrasound imaging is gaining increased attention due to its affordability, safety, real-time capabilities, and widespread availability. Among its applications, lung ultrasound (LUS) is becoming a key tool in point-of-care settings for detecting and managing acute respiratory conditions [1,2]. In certain cases, LUS has shown greater sensitivity than chest X-rays [3]. Recently, its use in emergency rooms for the prognosis of COVID-19 patients has been explored, with studies identifying LUS characteristics and imaging biomarkers associated with COVID-19 patients [4,5]. These biomarkers can aid both in early detection and in monitoring the respiratory effectiveness of mechanical ventilation [6]. The versatility and cost-effectiveness of ultrasound imaging make it particularly valuable in situations where patient demand exceeds hospital imaging capacity. Additionally, its affordability enhances accessibility in low- and middle-income countries. However, interpreting ultrasound images remains challenging due to the steep learning curve, which increases the risk of diagnostic errors [7].

In standard ultrasound imaging, a quasi-homogeneous speed of sound is typically assumed. However, in the lungs, this assumption does not hold due to the presence of air, which significantly disrupts the propagation of ultrasound waves. This interaction generates distinctive artifacts, including horizontal (A-lines) and vertical (B-lines) patterns, which are characteristic of LUS imaging and provide valuable diagnostic insights [8,9]. Horizontal artifacts appear as parallel lines to the pleura that repeat throughout the image due to ultrasound wave reflections from the pleural surface and their repeated return to the probe. In contrast, vertical artifacts manifest as bright lines extending from the pleura [2]. These phenomena arise when ultrasound waves interact with fluid in the interstitial space or thickened tissue due to swelling or inflammation. Additionally, some LUS images reveal

* Corresponding author.
E-mail address: libertario.demi@unitn.it (L. Demi).
[1] Equal contribution.

https://doi.org/10.1016/j.compbiomed.2025.111244
Received 25 April 2025; Received in revised form 30 September 2025; Accepted 18 October 2025
Available online 23 October 2025

a consolidation zone, which occurs when air in the small airways is replaced by fluid. The presence of consolidation suggests abnormal lung tissue or fluid accumulation, which is associated with conditions like pneumonia, pulmonary hemorrhage, or certain lung tumors. In such cases, ultrasound waves generate specific artifacts that aid in identifying affected areas [8]. However, the variability in experience among clinicians analyzing ultrasound images makes reproducibility challenging. Therefore, an automated system that assists clinicians by leveraging the statistical properties of these artifacts could be useful.

In the last decade, several valuable research works have been conducted for the semi-quantitative analysis of lung-related pathologies based on LUS imaging. Their techniques rely on visual analysis of LUS patterns, where a score is assigned based on the detected patterns, reflecting the condition of the lung. The adoption of these approaches has rapidly increased, particularly following the COVID-19 pandemic, which led to a significant portion of the literature on semi-quantitative LUS focusing on COVID-19 applications [5,10–14]. However, these techniques often exhibit variability and are affected by confounding factors such as imaging frequency, focal depth, and the type of probes used. These limitations were addressed by implementing a standardized imaging protocol and developing a 4-level scoring system (Score 0 to 3) that incorporates technical factors such as imaging parameters [5].

Furthermore, artificial intelligence (AI) has been utilized to automatically classify LUS data based on scores, leading to a more reliable and reproducible semi-quantitative method [7,15–17]. In particular, [15] introduced a support vector machine (SVM)-based method for frame-level classification of pleural line features to detect lung alterations associated with COVID-19. Their approach incorporated pleural line detection and feature extraction using a Hidden Markov Model and the Viterbi algorithm. However, the non-linear SVM model may be susceptible to overfitting. Earlier, Brattain et al. [18] developed one of the first automated vertical artifacts scoring systems in thoracic sonography. [7] suggested a regularized spatial transformer network to classify LUS frames into four scores [5]. Furthermore, a uninorm-based aggregation technique was employed to compute the video-level score by aggregating the predicted frame-level scores. Using the frame-level scores provided by the network [7], a benchmark study was performed to evaluate various aggregation techniques, such as a threshold approach [16], grammatically evolved decision trees [19], and cross-correlation [20], for determining video and exam-level scores with prognostic value [21]. Although the cross-correlation technique produced the best results, it sacrificed the simplicity and interpretability of the features used. [22] proposed deep learning-based methods for the automated classification of neonatal LUS image frames to support the diagnosis of respiratory conditions in newborns. They assessed human-to-AI interrater agreement by evaluating their proposed approach on LUS data from 34 neonatal patients. Their results demonstrated about 72% agreement with expert evaluations and highlighted the potential of AI for improving neonatal respiratory assessment.

Differently, [23] introduced a deep learning-based approach for automatically detecting and localizing vertical artifacts in ultrasound scans. They trained a fully connected convolutional neural network (CNN) on B-mode images, using both in vitro ultrasound phantoms and in vivo patient data. Their results confirm the capability of their proposed model to identify and localize the presence of vertical artifacts in clinical LUS images. [24] proposed another method to automatically detect and localize vertical artifacts in LUS videos using deep learning networks trained with weak labels. Their approach combined a convolutional neural CNN with a long short-term memory (LSTM) network and a temporal attention mechanism, which was evaluated on LUS scans. Their proposed technique led to a localization accuracy of 67.1%. [25] employed a CNN to quantify vertical artifacts in a LUS dataset comprising 4,864 clinician-labeled images. They also investigated the correlation between the automated counts and clinical parameters. Their intra-class correlation (ICC) analysis indicated a strong agreement between the human count and the neural network's output, with an ICC value of 0.791. Baloescu et al. proposed both deep learning and classical machine learning frameworks for the automated detection of vertical artifacts [26,27]. Moreover, the authors in [28] proposed a deep learning pipeline for multi-class segmentation of lung objects (ribs, pleural line) and artifacts (horizontal and vertical artifacts) in LUS images of a lung training phantom. Although they claim promising segmentation results, their lung data type is restricted to the phantom type, and the small number of image frames in the training reduces the generalizability of their approach in clinical settings. Chen et al. [29] proposed a contrastive self-supervised learning framework to capture spatio-temporal patterns in LUS videos, reinforcing the importance of dynamic information in video-level analysis.

Indeed, the application of the aforementioned AI-based techniques requires a large training dataset to improve adaptability and generalizability. The high computational complexity of deep learning methods, along with their lack of explainability, limits their use in real-time and clinical applications. To address these challenges, this study proposes a direct video-level scoring approach based on frame-level scores derived from segmentation. The proposed method is applied to neonatal LUS data. The process begins with segmenting the pleural line and vertical artifacts. For segmentation, we introduce a novel segmentation approach utilizing motion estimation for detecting the vertical artifacts. The key hypothesis behind segmenting vertical artifacts through motion estimation is that these artifacts exhibit greater motion magnitudes than other lung structures. Additionally, pleural line extraction is performed using K-means clustering on intensity values, leveraging the prior knowledge that the pleural line typically has the highest intensity. After completing the segmentation, we introduce a novel rule-based and explainable model, utilizing the key definitions in the scoring of the data introduced in the state of the art (i.e., in [22]), for the frame level scoring. Finally, we employ a well-known, simple threshold-based technique [16] to aggregate frame-level scores and compute the final video-level score. The proposed method provides a low-complexity and straightforward approach, eliminating the need for a huge training dataset and parameter settings, as required in machine learning and deep learning-based strategies. To the best of our knowledge, this research is the first study to leverage dynamic information for the segmentation of LUS image components and subsequently use the segmentation results for scoring.

The remainder of the paper is structured as follows: Section 2 presents the proposed methodology, detailing the datasets used and the hybrid approach for segmentation and the scoring protocol. Section 3 provides results obtained from the LUS dataset, specifically focusing on segmentation and scoring. A discussion of these findings is included in Section 4. Finally, Section 5 concludes with a summary of key remarks.

2. Methodology

In the following, first, we introduce the main elements of the proposed segmentation technique, including the motion estimation for detecting the vertical artifacts and K-means clustering for pleural line detection. Before estimating motion, the input video was preprocessed using a 2D median filter with a window size of [50, 2], applied over all frames. This step reduces frame to frame noise and stabilizes motion patterns while preserving vertically structured features relevant to artifact detection. Next, we provide the proposed technique for the frame-level scoring based on the segmentation results from the previous stage. Finally, we utilize the threshold-based technique [16] to aggregate the frame level scores to obtain the video level score. Fig. 1 provides the block diagram of the whole framework.

2.1. Motion estimation

Motion estimation is a key technique in computer vision used to track movement in image sequences [30]. One of the earliest approaches for motion estimation is the Lucas–Kanade optical flow algorithm [31]. Ideally, for motion estimation, the image pixels in the

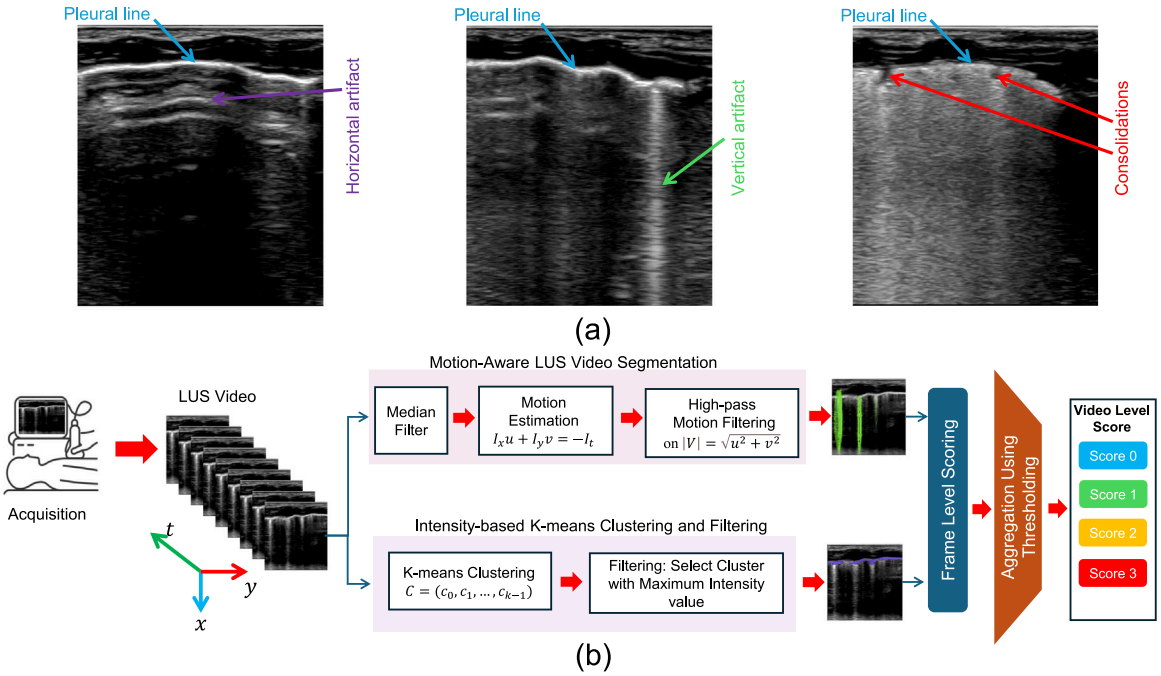

Fig. 1. (a) Examples of different components in LUS images, including pleural line, horizontal artifacts, vertical artifacts, and consolidations (b) Block diagram of the proposed hybrid pipeline for LUS video segmentation and scoring.

sequence adhere to two main assumptions: intensity constancy and small motion. We utilize these two main assumptions to formulate our proposed motion estimation technique. To this end, consider an imaging pixel $p(x, y)$ with an intensity value $I(x, y, t)$, where x and y indicate the lateral and axial positions of the image pixel, and t representing the slow time (i.e., temporal frame index). As indicated in Fig. 2, the first assumption implies that the image pixel's intensity value in frame t, $I(x, y, t)$, and frame $t+\delta t$, $I(x+u\delta t, y+v\delta t, t+\delta t)$, should be equal. Thus, the statement can be formulated as follows:

$$I(x, y, t) = I(x + u\delta t, y + v\delta t, t + \delta t), \tag{1}$$

where δt represents the time difference between consecutive temporal frames in motion estimation. Additionally, u and v denote the lateral and axial velocity components, respectively, while their displacements are given by $\delta x = u\delta t$ and $\delta y = v\delta t$. Under the second assumption, i.e., the assumption of small motion (δx, δy, and $\delta t \to 0$), we can expand the right-hand side of Eq. (1) using a two-dimensional Taylor series. Neglecting higher-order terms (> 1), we obtain:

$$\begin{aligned} & I(x, y, t) = I(x, y, t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t + \cdots \\ & \Rightarrow \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t = 0 \\ & \Rightarrow \frac{\partial I}{\partial x}u\delta t + \frac{\partial I}{\partial y}v\delta t + \frac{\partial I}{\partial t}\delta t = 0. \end{aligned} \tag{2}$$

By dividing both sides of Eq. (2) by δt and defining $I_x = \frac{\partial I}{\partial x}$, $I_y = \frac{\partial I}{\partial y}$, and $I_t = \frac{\partial I}{\partial t}$, we have the following fundamental relationship, referred to as the constraint equation:

$$I_x u + I_y v + I_t = 0. \tag{3}$$

It is important to note that solving Eq. (3) requires at least two neighboring image pixels around the $p(x, y)$ to estimate the possible displacement in the next frame. However, for a more reliable estimation of u and v, we consider a window of size $n \times n$ centered at the target pixel $p(x, y)$. This approach allows us to incorporate additional pixels within the window to form a more complete and stable system for motion estimation. By analyzing the image pixels within the selected window

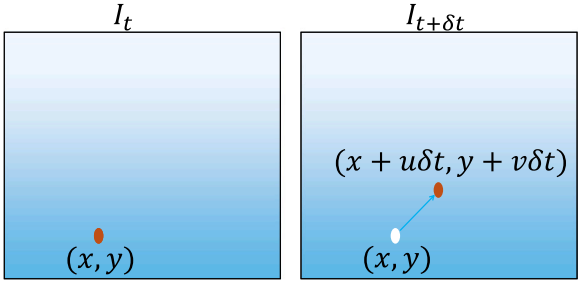

Fig. 2. A visualization of pixels displacement in the proposed motion estimation technique. Two consecutive frames at time window [t, $t+\delta t$] is used in the proposed model. u and v indicate the velocity components along lateral and axial directions in LUS image, respectively.

and applying Eq. (3) to each pixel, we obtain the following system of equations:

$$\begin{cases} I_x(p_1)u + I_y(p_1)v = -I_t(p_1) \\ I_x(p_2)u + I_y(p_2)v = -I_t(p_2) \\ \vdots \\ I_x(p_{n^2})u + I_y(p_{n^2})v = -I_t(p_{n^2}), \end{cases} \tag{4}$$

where $p_1, p_2, \ldots, p_{n^2}$ are the pixel intensities inside the window. The above system can be written in the matrix form $\mathbf{A}_{n^2\times 2} \times V_{2\times 1} = \mathbf{I}_{n^2\times 1}$, where

$$\mathbf{A} = \begin{bmatrix} I_x(p_1) & I_y(p_1) \\ I_x(p_2) & I_y(p_2) \\ \vdots & \vdots \\ I_x(p_{n^2}) & I_y(p_{n^2}) \end{bmatrix}, \mathbf{V} = \begin{bmatrix} u \\ v \end{bmatrix}, \mathbf{I} = \begin{bmatrix} -I_t(p_1) \\ -I_t(p_2) \\ \vdots \\ -I_t(p_{n^2}) \end{bmatrix}. \tag{5}$$

Solving the abovementioned equation, we may reach the infinite number of solutions for $\mathbf{V}$. However, to calculate the best possible solution, we propose using the least square minimization problem as follows:

$$\mathbf{V}^* = arg\min_{\mathbf{V}} \|\mathbf{AV} - \mathbf{I}\|^2. \tag{6}$$

To solve the minimization problem, we multiply $\mathbf{AV} = \mathbf{I}$ by $\mathbf{A}^T$ as Eq. (7).

$$\mathbf{A}^T\mathbf{AV} = \mathbf{A}^T\mathbf{I} \Rightarrow \mathbf{V}^* = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{I}. \tag{7}$$

Therefore, we can compute the vector $\mathbf{V} = [u, v]^T$ as the expanded version as follows:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n^2} I_x^2(p_i) & \sum_{i=1}^{n^2} I_x(p_i)I_y(p_i) \\ \sum_{i=1}^{n^2} I_y(p_i)I_x(p_i) & \sum_{i=1}^{n^2} I_y^2(p_i) \end{bmatrix}^{-1} \times \begin{bmatrix} -\sum_{i=1}^{n^2} I_x(p_i)I_t(p_i) \\ -\sum_{i=1}^{n^2} I_y(p_i)I_t(p_i) \end{bmatrix}. \tag{8}$$

2.1.1. Error analysis

The pair of equations that should be solved to obtain $V = [u, v]^T$ for the point $p(x_i, y_i)$ at temporal frame t_i can be expressed as Eq. (9).

$$\begin{cases} I_x^{(i)}u + I_y^{(i)}v = -I_t^{(i)} \\ I_x^{(j)}(u + \Delta u) + I_y^{(j)}(v + \Delta v) = -I_t^{(j)} \end{cases}. \tag{9}$$

Here, j refers to the neighbors of image pixel $p(x, y)$. However, the method assumes that the velocity is constant within the window, so the system of Eq. (9) can be changed to the following system:

$$\begin{cases} I_x^{(i)}u + I_y^{(i)}v = -I_t^{(i)} \\ I_x^{(j)}u + I_y^{(j)}v = -I_t^{(j)} \end{cases}. \tag{10}$$

However considering the uncertainty in the approximation of intensity gradients I_x, I_y, and I_t, we are essentially dealing with the following system:

$$\begin{cases} \hat{I}_x^{(i)}u + \hat{I}_y^{(i)}v = -\hat{I}_t^{(i)} \\ \hat{I}_x^{(j)}u + \hat{I}_y^{(j)}v = -\hat{I}_t^{(j)}. \end{cases} \tag{11}$$

To calculate the error related to the approximation $\hat{I}_x$, we expand the $I(x + \delta x, y, t)$ with respect to x and producing:

$$I(x + \delta x, y, t) = I(x, y, t) + I_x\delta x + I_{xx}\delta^2 x + \cdots, \tag{12}$$

where I_x and I_{xx} are the first and second partial derivatives of intensity in the x direction. By rearranging the above relation, we obtain the following relationship:

$$\hat{I}_x = \frac{I(x + \delta x, y, t) - I(x, y, t)}{\delta x} = I_x + I_{xx}\delta x + \cdots. \tag{13}$$

Then, Eq. (13) defines a unique approximation, and the resulting error $e_{I_x} = \hat{I}_x - I_x$ satisfies:

$$\|e_{I_x}\| = \|\hat{I}_x - I_x\| \le \|I_{xx}\delta x\| = \|I_{xx}\|\delta x. \tag{14}$$

Where $\|.\|$ denotes the norm. In a similar way, the error obtained from approximation I_y and I_t is as follows:

$$\|e_{I_y}\| \le \|I_{yy}\|\delta y, \qquad \|e_{I_t}\| \le \|I_{tt}\|\delta t. \tag{15}$$

Taking partial derivatives of the constraint equation, Eq. (3), with respect to x, y and t, we obtain the following equations:

$$I_{xx}u + I_x\frac{\partial u}{\partial x} + I_{yx}v + I_y\frac{\partial v}{\partial x} = -I_{tx} \tag{16}$$

$$I_{xy}u + I_x\frac{\partial u}{\partial y} + I_{yy}v + I_y\frac{\partial v}{\partial y} = -I_{ty} \tag{17}$$

$$I_{xt}u + I_x\frac{\partial u}{\partial t} + I_{yt}v + I_y\frac{\partial v}{\partial t} = -I_{tt}. \tag{18}$$

Assume that the second-order partial derivatives are continuous and motion is constant in the neighborhoods of each pixel, by substituting the left side of Eqs. (16) and (17) in Eq. (18), we have:

$$\begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \approx I_{tt}. \tag{19}$$

Without loss of generality, we can adjust our coordinate system by evaluating Eq. (19) along the x-axis ($v = 0$). Thus, this system can be reduced to the following equation:

$$u^2 I_{xx} \approx I_{tt}. \tag{20}$$

These analysis help us to calculate the upper bound of the residual error in estimation of Eq. (3). More specifically, considering relations in Eqs. (14), (15), and (20), we obtain the inequality as follows:

$$\|E\| = \|\hat{I}_x - I_x\||u| + \|\hat{I}_y - I_y\||v| + \|\hat{I}_t - I_t\| \le \|I_{xx}\||u|\delta x + \|I_{yy}\||v|\delta y + \|I_{xx}\|u^2\delta t. \tag{21}$$

By taking $\delta = \max\{\delta x, \delta y, \delta t\}$, and $M = \|I_{xx}\||u| + \|I_{yy}\||v| + \|I_{xx}\|u^2$ we conclude that

$$\|E\| \le M\delta. \tag{22}$$

Eq. (22) states that the estimation error's upper bound for the proposed motion estimation method is a factor of δ. In other words, the accuracy of approximated motion vector depends on how small are δx, δy and δt.

2.1.2. Vertical artifact segmentation using motion estimation

After performing motion estimation on the LUS video, we use the estimated u and v components to compute the velocity magnitude as $|V| = \sqrt{u^2 + v^2}$. To enhance the detection of vertical artifacts, which typically correspond to higher velocity magnitudes, we apply a velocity-based high pass filtering to the velocity magnitude values. This filtering process helps eliminate regions with lower velocity magnitudes, such as the pleural line and intercostal tissue. The cutoff velocity magnitude is determined based on the global average of the velocity magnitude at a specific temporal frame.

2.2. K-means clustering

K-means clustering is a technique for partitioning data into k clusters by minimizing intra-cluster variance [32]. The algorithm begins by selecting k initial centroids, $c_0, c_1, \ldots, c_{k-1}$, assigns each data point to its nearest centroid, and iteratively refines the centroids until convergence. This approach is widely applied in pattern recognition, noise reduction, and image segmentation tasks [33,34]. In this study, we explore its application for segmenting the pleural line, which exhibits high-intensity values in lung ultrasound (LUS) images. To illustrate, consider an image I where the pixel intensities are given by $\{I_{p_1}, I_{p_2}, \ldots, I_{p_n}\}$. The goal is to partition I into k clusters such that minimize the summation of inter cluster distances from each sample point I_{p_i} to the centroid c_l as follows:

$$D_{tot} = \sum_{l=0}^{k-1} \sum_{I_{p_i} \in c_l} d(I_{p_i}, c_l), \tag{23}$$

where d is distance between I_{p_i} and centroid c_l.

After clustering, we can detect different areas of the image based on the intensity characteristics of each part by selecting the corresponding cluster.

2.3. Frame level scoring

The LUS scores utilized in this study are an adapted interpretation of the original scoring system established by [35] and later applied in [22]. This detailed LUS scoring framework effectively captures essential characteristics linked to Transient Tachypnea of the Newborn (TTN) and Respiratory Distress Syndrome (RDS) [35]. This LUS scoring was determined according to the following criteria:

Score 0: Presence of horizontal artifacts without vertical artifacts, indicating a fully aerated lung state.

Score 1: Presence of multiple vertical artifacts along with an extended and continuous pleural line, suggesting mild lung alterations.

Score 2: Vertical artifacts spanning the entire pleural line, indicative of significant interstitial syndrome.
Score 3: Presence of lung consolidations with a disrupted pleural line, signifying severe lung pathology.

According to the aforementioned definitions, we have proposed a conditional algorithm for frame-level scoring. In this technique, frames are assigned a score by analyzing the segmented horizontal artifacts, pleural line, and vertical artifacts. The scoring system follows a hierarchical structure, beginning with cases where no vertical artifacts are present. Specifically, score 0 is assigned if at least one horizontal line is detected and no vertical artifacts are present. Score 1 is assigned if at least two vertical artifacts are detected and the pleural line extends across more than 80% of the total frame width. Score 2 is assigned when the relative width difference between the pleural and vertical artifacts' extensions is less than a small error threshold ϵ (In our evaluation, we set ϵ to 0.1, corresponding to a 10% difference relative to the larger width). Finally, Score 3 is given when the pleural line is fragmented into multiple sections, with at least four broken pleural lines detected, subject to the following condition: In cases where 4 or 5 fragmentations of the pleural line are observed, the interruptions may be due to rib shadows or pathology; to distinguish between these, we perform an additional check for the presence or absence of A-lines and vertical artifacts (absence indicating strong pathology), although some misclassification may still occur in this range. In such cases, if both A-lines and vertical artifacts are absent, we assign a score of 3. Additionally, if more than five pleural line fragmentations are found, we also assign a score of 3.

If none of the primary conditions are met, a nested conditional structure refines the assignment. Specifically, if no vertical artifacts are detected but at least three pleural line segments are present, the score is assigned as 3; otherwise, it is set to 0. If vertical artifacts are detected, an additional check compares the width of the pleural line and vertical artifacts. If their absolute difference is less than a small error tolerance 2ϵ, the score is set to 2; otherwise, it is set to 1.

This structured approach ensures a comprehensive evaluation of segmented frames based on pleural and vertical artifact characteristics. The complete scoring logic, including the primary and nested conditional structure, is presented in Algorithm 1. To validate the proposed scoring algorithm, we performed a frame-level evaluation using a subset of 12,000 manually annotated frames by expert clinicians from 20 patients.

Algorithm 1 Frame_Score

Input:
$W \leftarrow$ Image width
$HA \leftarrow$ Number of horizontal artifacts detected
$VA \leftarrow$ Number of vertical artifacts detected
$PL \leftarrow$ Number of pleural line detected
$PL\_ext \leftarrow$ Width of pleural line extension
$VA\_ext \leftarrow$ Width of vertical artifacts extension
$\epsilon \leftarrow 0.1$
Primary scoring algorithm:
if $VA = 0$ **and** $HA \geq 1$ **then**
    return 0
else if $VA \geq 2$ **and** $PL\_ext > 0.8 \times W$ **then**
    return 1
else if $\frac{|PL\_ext - VA\_ext|}{\max(PL\_ext, VA\_ext)} < \epsilon$ **then**
    return 2
else if $PL \geq 4$ **then**
    if $PL \leq 5$ **then**
        if $HA = 0$ **and** $VA = 0$ **then**
            return 3
        end if
    else
        return 3
    end if
end if
Nested conditional refinement:
if none of the above conditions met **then**
    if $VA = 0$ **then**
        if $PL \geq 3$ **then**
            return 3
        else
            return 0
        end if
    else
        if $\frac{|PL\_ext - VA\_ext|}{\max(PL\_ext, VA\_ext)} < 2 \times \epsilon$ **then**
            return 2
        else
            return 1
        end if
    end if
end if

2.4. Frame to video level scoring

Finally, after extraction of the frame level scores, we should aggregate them to reach a final video level score, which is of diagnostic value in the clinical LUS imaging. To this end, we utilize the simple and pre-existing approach in the state of the art, referred to the threshold (TH)- based technique [16]. The threshold-based technique works by assigning a video with the highest score found at a given percentage of frames (threshold) composing the video. Finally, the video level score is compared to the clinical evaluation to obtain the video level agreement.

In the end, by analyzing the video level agreement at various TH levels (in range 1% to 100%), the optimal TH can be identified to see which fraction of the video frames are most informative and useful for frame level score aggregation.

2.5. Data description

The data used to evaluate the method presented in this manuscript consist of clinical and specialized care data from the Fondazione IRCCS San Gerardo dei Tintori in Monza, Italy. The data involved preterm infants with a gestational age of less than 32 weeks and/or a birth weight under 1500 grams, excluding those with significant congenital abnormalities. The infants being infected a range of pulmonary disorders associated with prematurity, including acute respiratory conditions (such as RDS and TTN), evolving, and established broncho pulmonary dysplasia (BPD). LUS imaging was performed on patients who had received pulmonary surfactant therapy. None of the infants had pulmonary hypoplasia, pneumonia, pneumothorax, or meconium aspiration at the time of evaluation. Ultrasound imaging was conducted with the patient positioned supine, using a Philips Affiniti 70 ultrasound system equipped with a high-frequency micro linear probe (7.0–15.0 MHz), commonly referred to as a hockey stick transducer. The focal zone was aligned with the pleural line, with imaging parameters set to a depth of 3 cm and a frame rate of 63 Hz to ensure optimal resolution. The dataset comprises 70 examinations evaluated by human operators (HOs) with expert level of proficiency. Three HOs, each with 4 to 7 years of experience, performed video-level scoring to establish the ground truth (GT). Each of these 70 exams includes six videos (a total of 420 videos and 78,439 frames), each video corresponding to a specific area of the chest. The number of frames in each video varies from 188 to 607. Video-level labeling was performed for the entire dataset based on the same 4-level scoring scheme described in Section 2.3. In addition to video level labeling, a subset of 12,000 frames extracted from 20 patients was manually annotated at the frame level by clinical experts. This subset was used for quantitative evaluation of

the frame level scoring performance. The study was approved by the local ethical committee (protocol nr. 3804/21), and written informed consent was obtained from all parents prior to enrollment.

2.6. Evaluation strategy

We evaluate the proposed technique on the abovementioned datasets consisting of 70 exams, with each exam consisting of six videos. Our analysis covers three distinct cases: Case-1 and Case-2 include the same 50 and 60 exams used in [22], respectively; Case-3 considers the entire dataset of 70 exams. It should be noted that since three HOs labeled the videos, we extracted the results using different ground truth labeling strategies. The first strategy involves comparing the results with each HO's label individually. The second strategy uses majority voting among the three HOs' assessments. The third strategy considers only the videos where all three HOs provided the same score, ensuring complete agreement. This third approach resulted in filtering out some videos, leading to a final selection of 270 videos extracted from all 70 exams. This comprehensive evaluation ensures a robust assessment of the technique across different scenarios.

2.7. Evaluation metrics

To evaluate the performance of the proposed technique, we used key evaluation metrics including, *accuracy*, *recall*, *precision*, and F_1 score which are crucial to assessing model classification performance. These metrics are based on the following definitions: true positives (TP) refer to the number of videos correctly predicted as a specific class; false positives (FP) are the number of videos incorrectly predicted as that class; false negatives (FN) are the number of videos that belong to that class but were predicted as another; and true negatives (TN) refer to the number of videos correctly predicted as not belonging to the specific class. These metrics are defined as follows:

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \tag{24}$$

$$Recall = \frac{TP}{TP + FN}, \tag{25}$$

$$Precision = \frac{TP}{TP + FP}, \tag{26}$$

$$F_1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{27}$$

2.8. Cross-validation

As the proposed technique involves three main parameter settings (k,WS,TH), we perform a cross validation through the whole data to extract the optimal parameters. To this end, we divided 70 exams to 7 different parts, each of them includes 10 exams. Following that, we use 7 fold cross validation by training the model on 60 exams in each fold and testing on the rest of 10 exams. For each fold, we extract the accuracy and finally extract the optimal parameters based on the maximum average accuracy on the test set. The parameter space is considered as $k \in \{2,3,4,5,6\}$, $WS \in \{3,5,7\}$, and $TH \in \{1\%, 2\%, \ldots, 100\%\}$. It should be noted that this process is repeated for different ground truth strategies, i.e., HO1, HO2, HO3, and majority.

3. Results

We applied the proposed segmentation technique with different k values for pleural line detection and motion estimation with varying window sizes for vertical artifact segmentation. It should be noted that a window size of WS means we used a window of size $WS \times WS$. In our implementation, we compute motion vectors between frame pairs that are temporally separated by a gap of three frames (e.g., frame 1 with frame 4). This temporal spacing allows for the detection of meaningful motion patterns, particularly those associated with vertical artifacts. Qualitative examples of the segmentation of pleural line and vertical artifacts are presented in Fig. 3 for different k and WS parameter values. Then, we use the segmentations results for our frame level scoring. As shown in the figure, the pleural line and vertical artifacts segmentations revealed the best results with $k = 5$ and $WS = 3$. Even in cases where there are discontinuities in the pleural line and several vertical artifacts, the proposed method achieves promising segmentation accuracy.

In addition to the qualitative evaluation, we conducted a cross-validation study based on video-level agreement provided in Section 2.8. The results of the cross-validation process are presented in Fig. 4. The figure presents the best accuracies and parameters achieved in each fold for different ground truth cases. For each ground truth, the first column shows the best accuracy obtained in each fold that corresponds to the optimal pairs (k, WS). To simplify the visualization, the optimal threshold obtained in each pair of (k, WS) is shown in the right part of the figure. In addition, Fig. 5 summarizes the highest test accuracies obtained across the seven folds for each ground truth. Table 1 shows the optimal values of k, WS, and TH obtained in each when evaluated against the majority of the ground truth.

To compare the proposed method against the state-of-the-art study [22], we adopted the exact same train-test split used in their evaluation. For Case-1 and Case-2, the training sets consisted of 20 and 10 exams, respectively, from which we determined the optimal parameters k, WS, and TH. Figs. 6 and 7 present the highest accuracy values for each (k, WS) pair in the first column, along with the corresponding optimal threshold in the second column, for Case-1 and Case-2, respectively. The optimal parameters for both cases are summarized in Table 2.

In the next step, following the segmentation of the LUS images, we utilize our rule-based frame-level scoring technique designed based on the standard definitions in the standardized protocol of neonatal LUS scoring [22,35], for the frame-level scoring. To assess the reliability of this scoring approach, we conducted a frame-level evaluation using 12,000 manually annotated frames from 20 patients. The predicted frame-level scores were compared against expert annotations, and the results are presented in a confusion matrix (Fig. 8), showing a classification accuracy of 51.19%. This evaluation confirms that the proposed rule-based algorithm captures meaningful diagnostic features at the frame level. Next, the threshold-based aggregation technique with TH ranges from 1% to 100% was used to calculate the video-level score from the vector of frame level scores.

For a quantitative assessment, we evaluated the performance of our proposed technique at the video-level scoring by comparing it with the labels provided by HOs. Specifically, we extracted the evaluation metrics discussed in Section 2.7, including accuracy, precision, recall, and F1 Score. These metrics were computed for different cases, with results presented in Tables 3 and 4 for Case-1 and Case-2, and Case-3, respectively, along with a comparison against state-of-the-art deep learning techniques from [22]. It should be noted that, for the proposed technique, the results are reported for the best-performing threshold TH based on the observation (i.e., 37%). More specifically, the results of Case-1 (the same 50 exams used in [22]) show that while ResNet-18 achieved the highest accuracy of 0.77, our proposed technique demonstrated a promising accuracy of 0.66, outperforming DCNN. This performance difference in Case-1 could be due to the presence of more challenging cases, where the ResNet-18 deep learning-based technique demonstrates better capacity. Differently, in Case-2 (the same 60 exams used in [22]), our proposed method outperformed both ResNet-18 and DCNN, achieving a maximum accuracy of 0.68. In Case-3, reported in Table 4 which corresponds to using all the 70 exams (not presented in the state of the art), the proposed method continues to show strong performance, with accuracy values ranging from 0.67 to 0.72.

Finally, when using the full agreement strategy for determining the ground truth, the proposed method achieves its highest overall performance, with an accuracy of 0.77, precision of 0.77, recall of 0.74, and an F1 score of 0.75.

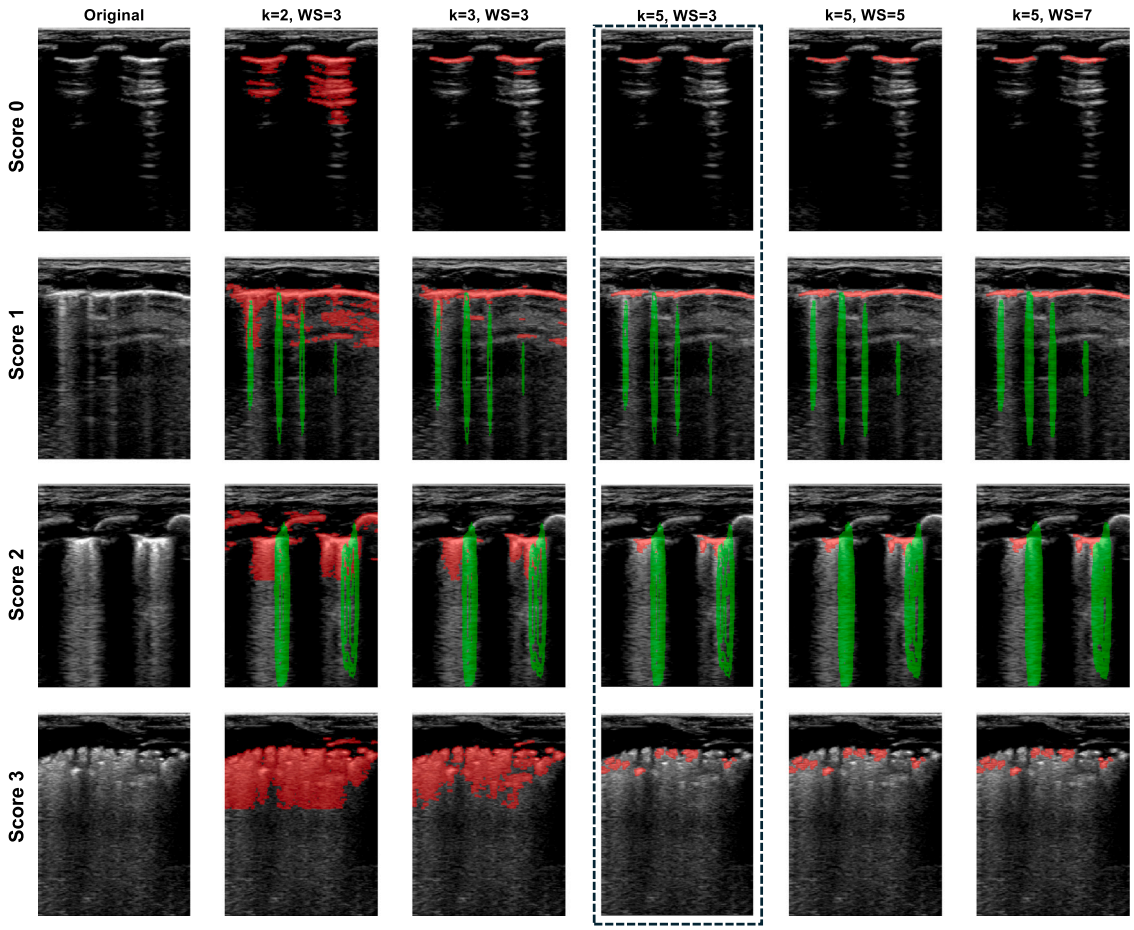

Fig. 3. Examples of segmented image frames of LUS videos from different scores using the proposed method. pleural line and vertical artifacts segmentations are shown in red and green color, respectively (see the link to the segmented videos). The dashed box indicates the best segmentation results.

Table 1
Optimal parameter values for each fold in cross validation against the Majority as ground truth.

Case	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7
Best k value	5	5	5	5	5	5	5
Best WS value	3	3	3	3	3	3	3
Best TH value	24	31	31	31	31	31	36

Table 2
Optimal parameter values for Case-1 and Case-2 against the Majority as ground truth.

Parameter	Best k value	Best WS value	Best TH value
Case-1	5	3	37
Case-2	5	3	37

Table 3
Performance analysis of the proposed method for video-level classification in case-1 (50-exam configuration in [22]) and case-2 (60-exam configuration in [22]) considering different ground truth (GT) labels. The values in bold represent the best results.

Methods	Case-1				Case-2				GT
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	
DCNN [22]	0.60	0.53	0.60	0.60	0.63	0.63	0.63	0.63	HO1
	0.59	0.61	0.59	0.60	0.54	0.60	0.54	0.55	HO2
	0.55	0.58	0.55	0.56	0.57	0.59	0.57	0.57	HO3
ResNet-18 [22]	**0.77**	**0.78**	**0.77**	**0.76**	0.58	0.61	0.58	0.58	HO1
	0.68	0.77	0.68	0.68	0.66	0.65	**0.66**	**0.65**	HO2
	0.71	0.76	0.71	0.70	0.62	0.62	0.62	0.62	HO3
Proposed Method	0.64	0.60	0.54	0.55	0.67	**0.69**	0.63	0.64	HO1
	0.65	0.58	0.55	0.56	0.67	0.65	0.61	0.62	HO2
	0.66	0.63	0.51	0.52	**0.68**	**0.69**	0.59	0.62	HO3

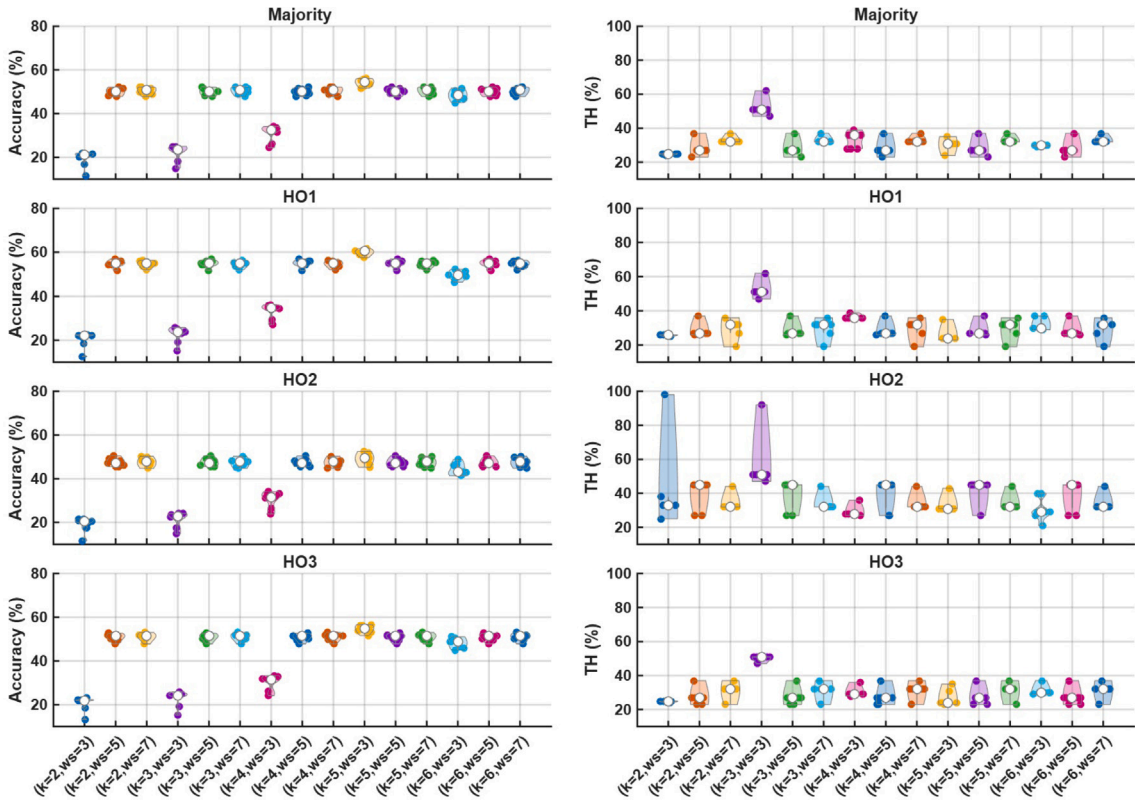

Fig. 4. Parameter optimization results for the four ground truth strategies (rows: Majority, HO1, HO2, HO3). First column: accuracy for different (k, WS) pairs. Second column: the best video-level threshold (TH) corresponding to the highest accuracy for each (k, WS) pairs.

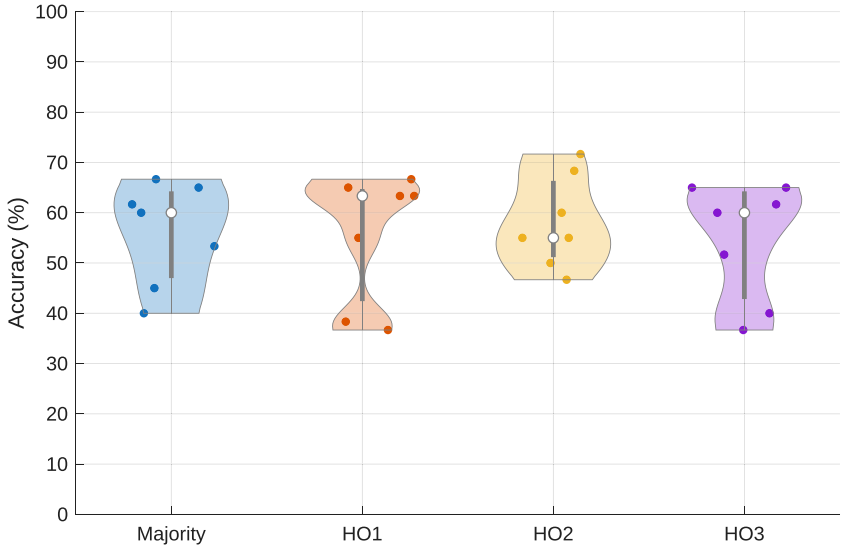

Fig. 5. Distribution of best test accuracies for each ground truth strategy (Majority, HO1, HO2, HO3) obtained across the seven fold cross validation. For each ground truth, seven accuracy values are shown, each corresponding to one fold.

Table 4
Performance analysis of the proposed method for video-level classification in case-3 (including all 70 exams) considering different ground truth (GT) labels.

Accuracy	Precision	Recall	F1 Score	GT
0.67	0.68	0.65	0.65	HO1
0.68	0.67	0.63	0.64	HO2
0.69	0.70	0.63	0.65	HO3
0.72	0.71	0.68	0.69	Majority
0.77	0.77	0.74	0.75	Full agreement

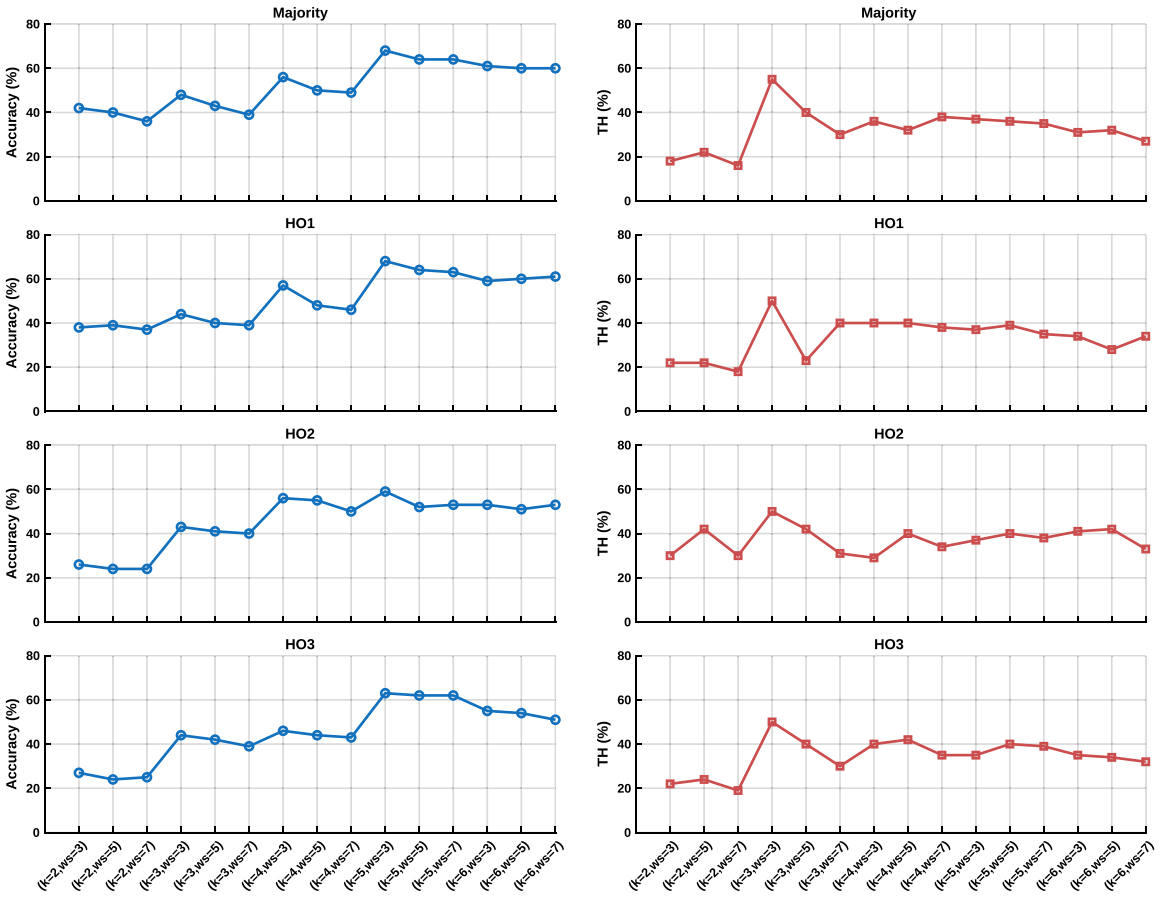

Fig. 6. Training results for each ground truth (Majority, HO1, HO2, and HO3) obtained from the parameter optimization set in Case-1. Each row corresponds to each ground truth. The first column shows video-level accuracies for all (k, WS) parameter combinations while the second column corresponds optimal threshold values for each (k, WS) pair.

Fig. 9 shows the confusion matrices obtained using different strategies to define the ground truth labels, including majority voting (first column) and the individual human operators HO1 (second column), HO2 (third column), and HO3 (fourth column). The results are provided for the best-performing threshold TH (i.e., 37%), which demonstrates strong agreement with majority voting, achieving higher accuracy. Moreover, among human operators, the method exhibits the closest alignment with HO3, showing fewer classification errors compared to HO1 and HO2. Overall, despite some variations, the method remains consistent across different ground truths.

As stated, the results in Tables 3 and 4 were reported based on the optimal TH value in the aggregation technique. However, we also analyzed accuracy by varying the TH from 1% to 100%, as illustrated in Fig. 10. The figure demonstrates similar trends across different cases and aligns with the observations reported in [22]. As TH increases from 1% to 100%, accuracy initially improves before gradually declining. Notably, the optimal TH for all cases is 37%. Furthermore, the highest accuracy is observed in the full consensus case (Case-3), highlighting the importance of robust ground truth for reliable evaluations.

Finally, it is worth to assess the proposed technique by allowing a one-error tolerance [16,21]. Specifically, this evaluation considers whether the absolute difference between the predicted score and the ground truth is within 1. Fig. 11 presents the one-error tolerance accuracy for video-level scoring using the proposed method. Under this criterion, the proposed technique achieves an accuracy of 0.97, demonstrating strong performance within a ±1 margin of error.

All implementations and experiments were conducted using MATLAB software (MathWorks, Natick, MA, USA) on a system equipped with an Intel(R) Core(TM) i5-10500 processor (3.10 GHz) and 32 GB of RAM.

4. Discussion

The proposed segmentation and scoring framework provides a robust and interpretable alternative to deep learning-based approaches for LUS analysis. The segmentation results (Fig. 3) confirm that K-means clustering successfully identifies pleural line, while motion estimation reliably detects vertical artifacts. The effectiveness of k = 5 in pleural line segmentation highlights the ability of clustering-based methods to enhance boundary delineation, even in cases where discontinuities exist. The motion estimation method performed optimally with a window size of 3, demonstrating its capacity to differentiate motion-rich vertical artifacts from background structures, which is crucial for reliable artifact segmentation. The use of optical flow for vertical artifact detection was also motivated by its pixel level resolution and computational efficiency, which are particularly valuable for real-time neonatal LUS analysis compared to existing methods such as speckle tracking [30].‘

Figs. 4 and 5, together with Table 1 show the results of the 7-fold cross validation through the whole 70 exams. Across the combinations of evaluated parameters, $(k, WS) = (5, 3)$ consistently yielded the highest precisions for the four ground-truth strategies in the 7-fold cross-validation, as further confirmed by the majority ground-truth results in Table 1, where all folds selected $k = 5$ and $WS = 3$. For this optimal configuration, the best video-level threshold (TH) for majority ranged between 24% and 36% across folds, while inspection of the other strategies in Fig. 4 shows corresponding TH values of approximately 37% for HO1 and HO3, and 20% for HO2. These findings indicate that while $(k, WS) = (5, 3)$ provides a robust

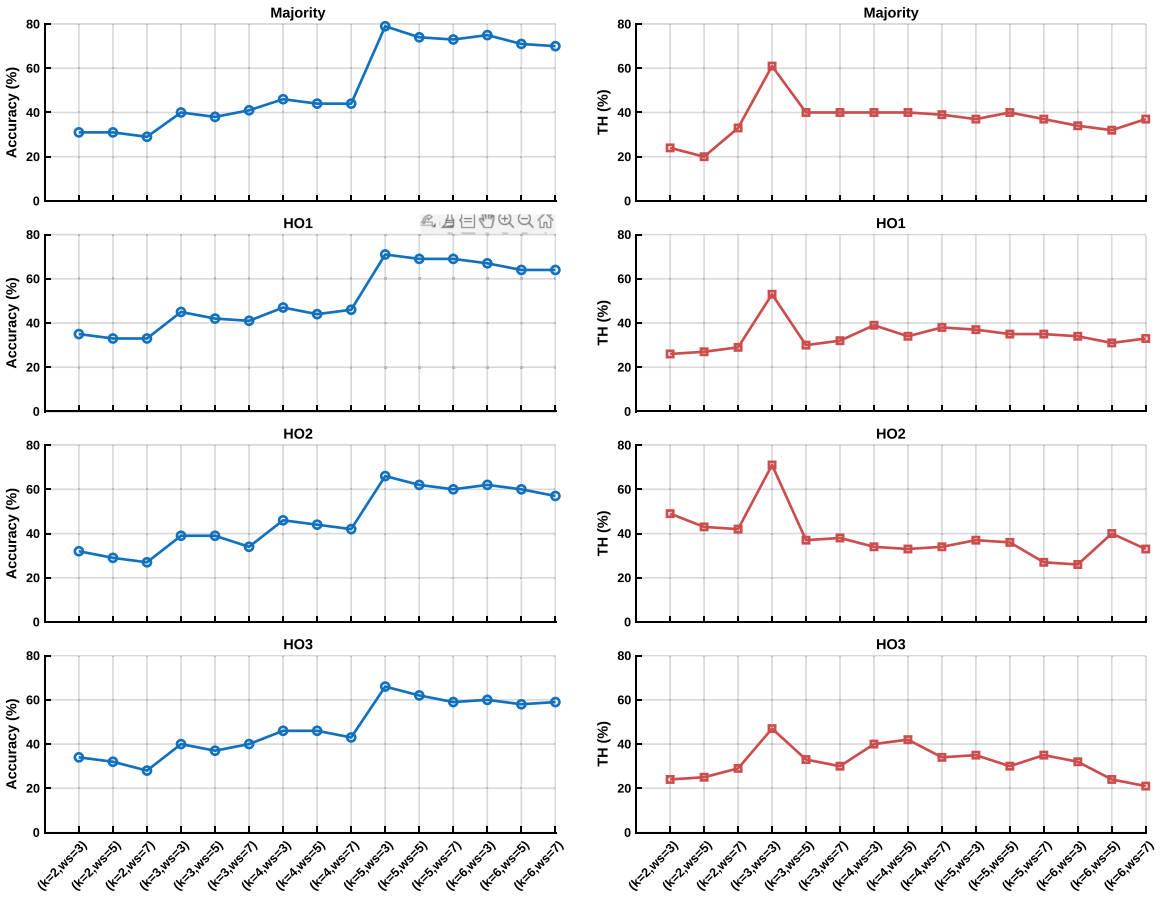

Fig. 7. Training results for each ground truth (Majority, HO1, HO2, and HO3) obtained from the parameter optimization set in Case-2. Each row corresponds to each ground truth. The first column shows video-level accuracies for all (k, WS) parameter combinations while the second column corresponds optimal threshold values for each (k, WS) pair.

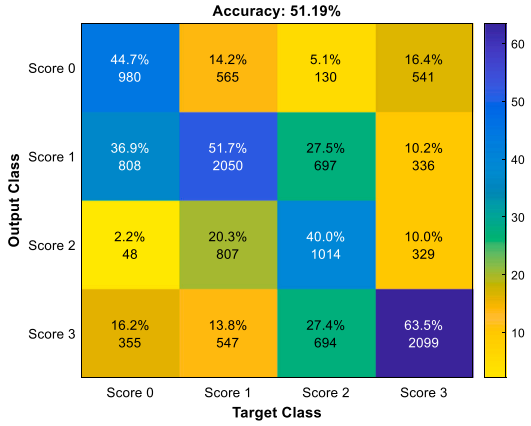

Fig. 8. Confusion matrix of the frame-level scoring algorithm when applied to a subset of 12,000 frames from 20 patients.

and generalizable setting across strategies, TH requires fine-tuning per strategy to maximize performance.

Furthermore, we applied the proposed method to the same training sets used in the state-of-the-art study—namely, Case-1 and Case-2 to determine the optimal parameters. As shown in Figs. 6 and 7, and Table 2, the best performance across all ground truth strategies (Majority, HO1, HO2, and HO3) was consistently achieved with $K = 5$, $WS = 3$, and $TH = 37\%$. This configuration delivered the highest video-level accuracies for both Case-1 and Case-2. The stability of this parameter set across different annotation schemes indicates strong generalization ability, making it an appropriate choice for fair comparison with the state-of-the-art. Since both cross-validation and the training-set evaluation for Case-1 and Case-2 yielded the same $(K, WS) = (5, 3)$ for the proposed segmentation method, we adopted this parameter set for all subsequent analyses. However, we continued to assess performance with varying TH values, as it changes during cross-validation.

The scoring accuracy results (Fig. 9) demonstrate that the proposed method aligns well with human expert assessments. For example in Case-2 (majority voting), the method achieved its highest accuracy in Score 1 (78.9%) and Score 3 (71.4%), suggesting that the segmentation process effectively distinguishes moderate and severe lung conditions. Slightly lower accuracies for Score 0 (68.8%) and Score 2 (63.0%) indicate that differentiating between normal lung patterns and mild interstitial syndromes remains challenging. These results suggest that while the method successfully captures pleural and vertical artifact characteristics, further refinements in feature selection may improve performance in borderline cases.

A comparative analysis with deep learning-based approaches in Table 3 highlights the advantages of the proposed method. In Case-2, the proposed approach outperformed DCNN and ResNet-18 by 9%, demonstrating its superior generalizability. The accuracy range of 67% to 72% in Case-3 further validates the robustness of the method across diverse clinical conditions. The highest accuracy of 77.4% in Case-3, where complete agreement among HOs was present, confirms that the proposed method performs optimally when the reference standard is highly reliable.

The results in Fig. 10 show the impact of selecting a threshold in the aggregation technique. It reveals a clear trend across all cases:

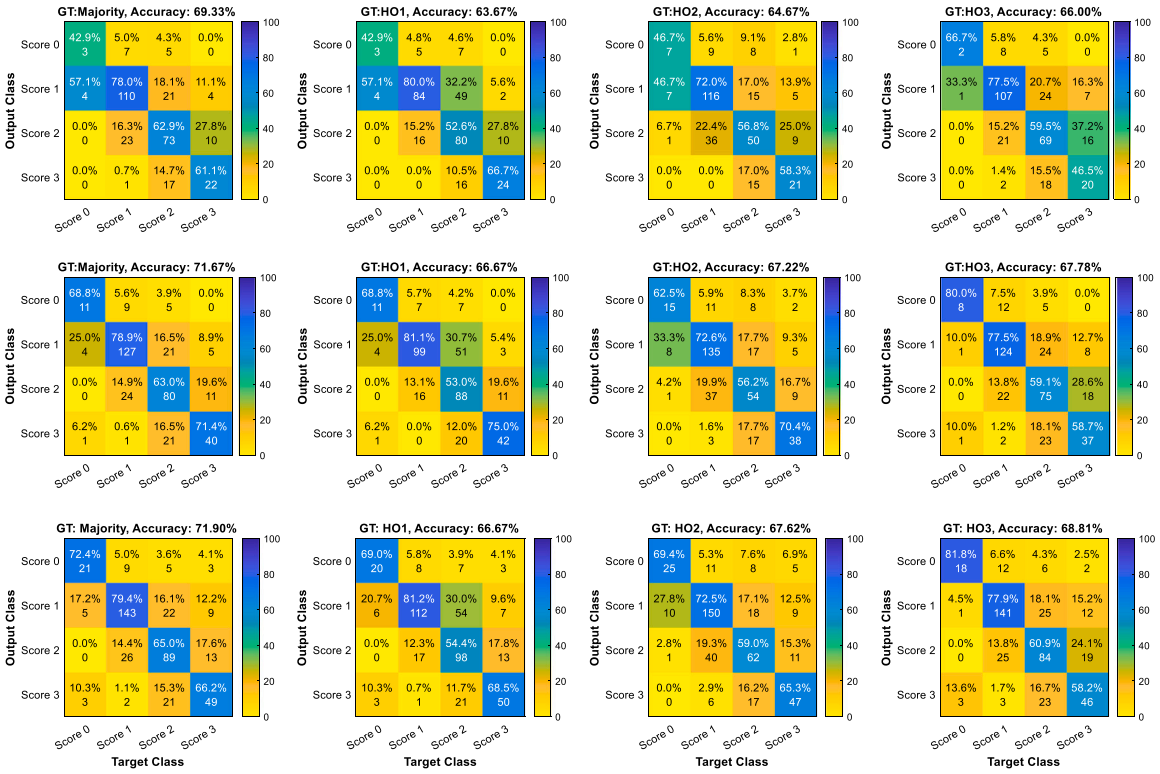

Fig. 9. Confusion matrices illustrating the model's performance using the proposed technique with different ground truth references: majority voting (column 1), HO1 (column 2), HO2 (column 3), and HO3 (column 4). The first, second, and third rows correspond to Case-1 (50-exam configuration from [22]), Case-2 (60-exam configuration from [22]), and Case-3 (all 70-exam configuration), respectively.

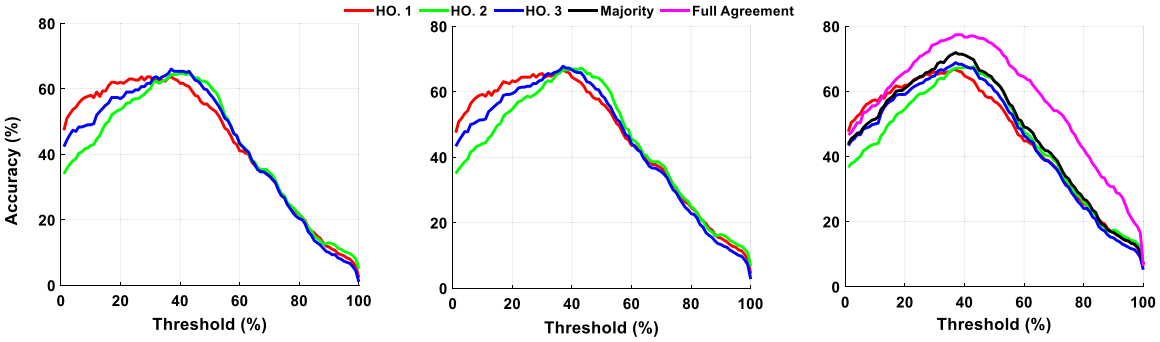

Fig. 10. Video level accuracy vs threshold (TH) in the threshold-based aggregation technique. The evaluation was performed with THs ranging from 1% to 100%. The horizontal axis represents the threshold value while the vertical axis represents the video level accuracy. The first, second, and third columns respectively represent Case-1, Case-2, and Case-3.

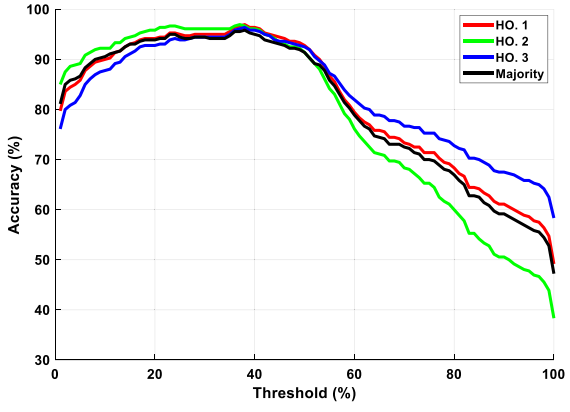

Fig. 11. One error tolerance accuracy evaluation across various THs ranging from 1% to 100%.

accuracy initially increases with the threshold, reaches a peak, and then declines as TH continues to rise. This behavior suggests that lower thresholds may allow noisy frame-level scores to influence the final video score, while excessively high thresholds may discard too much relevant information. The evaluation of video-level scoring revealed that an optimal threshold of 37% maximized accuracy, achieving 72% agreement with majority-voted ground truth and increasing to 77.4% when all three HOs agreed on the diagnosis. The improvement in accuracy for cases with full consensus suggests that inter-observer variability significantly impacts LUS interpretation. Additionally, the study inherently includes variability caused by probe movement, and respiratory motion, as the dataset was acquired in real clinical conditions. Such motion artifacts could potentially affect the detection of vertical artifacts. This finding underscores the necessity of automated scoring techniques to standardize assessments and reduce observer-dependent inconsistencies.

Fig. 11 illustrates the accuracy evaluation of the aggregation threshold-based technique under a one-error tolerance setting, where predictions are considered correct if their absolute difference from the ground truth score is within one. The overall trend shows that accuracy is high for low TH values, peaks around 20%–40% with accuracy $\geq$ 95%, and then declines as TH increases toward 100%. This behavior indicates that moderate threshold values strike a balance between flexibility in score aggregation and maintaining consistency with the ground truth. More specifically, at moderate TH values, the method includes a broad range of frames in determining the final video score, which enhances robustness under the one-error tolerance metric, leading to high accuracy. However, as TH increases, the requirement for a higher percentage of frames to hold a specific score makes the method more rigid, which reduces accuracy when slight disagreements exist. Among the different approaches, HO3 (blue) maintains the highest accuracy across most of the TH range, particularly between 40%–80%, indicating the observer's reliability in scoring. HO1 (red) follows a similar pattern, whereas HO2 (green) experiences the most pronounced decline at higher TH values, suggesting higher variability in this operator's scoring patterns. The sharp drop in accuracy beyond 80% for all methods indicates that enforcing extremely high thresholds results in overly rigid aggregation, which leads to reduced alignment with the ground truth.

This finding suggests that the method consistently assigns scores within a reasonable diagnostic range, supporting its potential application in real-world clinical workflows where minor disagreements between experts are common.

Computational analysis demonstrates the method's feasibility for clinical use. More specifically, the segmentation process, particularly motion estimation, accounted for the majority of processing time, averaging 49.41 $\pm$ 1.21 s per video, while K-means clustering required 10.11 $\pm$ 0.56 s. The scoring step was highly efficient, taking only 0.33 $\pm$ 0.01 s, showing the low computational complexity of the approach. In the proposed method, the total processing time per frame, including both segmentation and scoring, is 126 ms. In comparison, the state-of-the-art deep learning approach in [22] achieves a scoring time of 17 ms per frame. While this indicates that the deep learning method is faster for the scoring step, it should be noted that it performs only scoring and does not provide a frame-by-frame segmentation of the ultrasound video. On the other hand, our method delivers both segmentation and scoring within a single pipeline, offering greater interpretability and transparency compared to the black-box nature of deep learning scoring models. Moreover, the proposed approach uses only three parameters, whereas state-of-the-art deep learning such as ResNet-18 and ViT require extensive training times and involve millions of parameters (up to ~47 million) [22]. This makes deep learning models challenging to train on standard CPU-based systems.

It should be noted that the proposed method delivers interpretable LUS scoring without requiring a large annotated training set, which is very challenging in deep learning based techniques. It successfully segments pleural and vertical artifacts, achieves competitive scoring accuracy w.r.t. state of the art techniques, and operates efficiently in quasi-real-time settings. Future research should focus on enhancing segmentation precision for borderline cases and validating the approach on larger, multi-center datasets to further establish its clinical reliability.

5. Conclusion

This study introduced a novel hybrid segmentation and scoring framework for LUS analysis that effectively bridges the gap between interpretability and automation. By integrating motion estimation for vertical artifact detection and K-means clustering for pleural line segmentation, the proposed method reducing the need for large annotated datasets and huge parameter settings while achieving accuracy comparable to deep learning-based approaches. The scoring framework demonstrated strong agreement with expert evaluations, especially when full consensus among human operators was achieved. However, the computational efficiency analysis confirmed that the method operates within feasible time constraints for real-time application in neonatal care settings. Given its transparency using visual segmentation, efficiency, and reliability, the proposed framework provides a viable solution for automated LUS scoring that can aid clinicians in decision-making.

Although this study focuses on neonatal LUS data, which is more challenging, the proposed method is general in design and can be applied to other pathologies. In future work, we have plan to study the possibility of generalizing the presented method to adult LUS data.

CRediT authorship contribution statement

Hamed Jalilian: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Sajjad Afrakhteh:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Federico Mento:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Emanuela Zannin:** Data curation. **Camilla Rigotti:** Data curation. **Federico Cattaneo:** Data curation. **Giulia Dognini:** Data curation. **Maria Luisa Ventura:** Data curation. **Libertario Demi:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

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.

Acknowledgments

Part of this work was supported by a research grant provided in compliance with Mission 6/component 2/Investment: 2.1 "Strengthening and enhancing biomedical research in the NHS", financed by the European Union-NextGenerationEU, CUP: E63C24000660007. This study was partially funded by the European Union-NextGenerationEU, in the framework of the iNEST - Interconnected Nord-Est Innovation Ecosystem (iNEST ECS00000043–CUP E63C22001030007). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.111244.

Data availability

The data of this study can be available based on reasonable request from the corresponding author.

References

[1] R. Raheja, M. Brahmavar, D. Joshi, D. Raman, Application of lung ultrasound in critical care setting: a review, Cureus 11 (7) (2019).

[2] L. Demi, F. Wolfram, C. Klersy, A. De Silvestri, V.V. Ferretti, M. Muller, D. Miller, F. Feletti, M. Wełnicki, N. Buda, et al., New international guidelines and consensus on the use of lung ultrasound, J. Ultrasound Med. 42 (2) (2023) 309–344.

[3] Y. Amatya, J. Rupp, F.M. Russell, J. Saunders, B. Bales, D.R. House, Diagnostic use of lung ultrasound compared to chest radiograph for suspected pneumonia in a resource-limited setting, Int. J. Emerg. Med. 11 (2018) 1–5.

[4] E. Poggiali, A. Dacrema, D. Bastoni, V. Tinelli, E. Demichele, P. Mateo Ramos, T. Marcianò, M. Silva, A. Vercelli, A. Magnacavallo, Can lung US help critical care clinicians in the early diagnosis of novel coronavirus (COVID-19) pneumonia? Radiology 295 (3) (2020) E6–E6.
[5] G. Soldati, A. Smargiassi, R. Inchingolo, D. Buonsenso, T. Perrone, D.F. Briganti, S. Perlini, E. Torri, A. Mariani, E.E. Mossolani, et al., Proposal for international standardization of the use of lung ultrasound for patients with COVID-19: a simple, quantitative, reproducible method, J. Ultrasound Med. 39 (7) (2020) 1413–1419.
[6] K. Stefanidis, S. Dimopoulos, E.-S. Tripodaki, K. Vitzilaios, P. Politis, P. Piperopoulos, S. Nanas, Lung sonography and recruitment in patients with early acute respiratory distress syndrome: a pilot study, Crit. Care 15 (2011) 1–8.
[7] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, I. Huijben, N. Chennakeshava, F. Mento, A. Sentelli, et al., Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound, IEEE Trans. Med. Imaging 39 (8) (2020) 2676–2687.
[8] 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) (2019) 612–623.
[9] G. Soldati, M. Demi, A. Smargiassi, R. Inchingolo, L. Demi, The role of ultrasound lung artifacts in the diagnosis of respiratory diseases, Expert. Rev. Respir. Med. 13 (2) (2019) 163–172.
[10] A. Smargiassi, G. Soldati, E. Torri, F. Mento, D. Milardi, P. Del Giacomo, G. De Matteis, M.L. Burzo, A.R. Larici, M. Pompili, et al., Lung ultrasound for COVID-19 patchy pneumonia: extended or limited evaluations? J. Ultrasound Med. 40 (3) (2021) 521–528.
[11] G. Soldati, G. Giannasi, A. Smargiassi, R. Inchingolo, L. Demi, Contrast-enhanced ultrasound in patients with COVID-19: pneumonia, acute respiratory distress syndrome, or something else? J. Ultrasound Med. 39 (12) (2020) 2483–2489.
[12] F. Mento, T. Perrone, V.N. Macioce, F. Tursi, D. Buonsenso, E. Torri, A. Smargiassi, R. Inchingolo, G. Soldati, L. Demi, On the impact of different lung ultrasound imaging protocols in the evaluation of patients affected by coronavirus disease 2019: how many acquisitions are needed? J. Ultrasound Med. 40 (10) (2021) 2235–2238.
[13] T. Perrone, G. Soldati, L. Padovini, A. Fiengo, G. Lettieri, U. Sabatini, G. Gori, F. Lepore, M. Garolfi, I. Palumbo, et al., A new lung ultrasound protocol able to predict worsening in patients affected by severe acute respiratory syndrome coronavirus 2 pneumonia, J. Ultrasound Med. 40 (8) (2021) 1627–1635.
[14] L. Demi, F. Mento, T. Perrone, A. Fiengo, A. Smargiassi, R. Inchingolo, G. Soldati, Agreement Between Expert Sonographers and Artificial Intelligence in the Evaluation of Lung Ultrasound Data Acquired from COVID-19 Patients, European Respiratory Society, 2021.
[15] L. Carrer, E. Donini, D. Marinelli, M. Zanetti, F. Mento, E. Torri, A. Smargiassi, R. Inchingolo, G. Soldati, L. Demi, et al., Automatic pleural line extraction and COVID-19 scoring from lung ultrasound data, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 67 (11) (2020) 2207–2217.
[16] F. Mento, T. Perrone, A. Fiengo, A. Smargiassi, R. Inchingolo, G. Soldati, L. Demi, Deep learning applied to lung ultrasound videos for scoring COVID-19 patients: A multicenter study, J. Acoust. Soc. Am. 149 (5) (2021) 3626–3634.
[17] O. Frank, N. Schipper, M. Vaturi, G. Soldati, A. Smargiassi, R. Inchingolo, E. Torri, T. Perrone, F. Mento, L. Demi, et al., Integrating domain knowledge into deep networks for lung ultrasound with applications to COVID-19, IEEE Trans. Med. Imaging 41 (3) (2021) 571–581.
[18] L.J. Brattain, B.A. Telfer, A.S. Liteplo, V.E. Noble, Automated b-line scoring on thoracic sonography, J. Ultrasound Med. 32 (12) (2013) 2185–2190.
[19] L.L. Custode, F. Mento, F. Tursi, A. Smargiassi, R. Inchingolo, T. Perrone, L. Demi, G. Iacca, Multi-objective automatic analysis of lung ultrasound data from covid-19 patients by means of deep learning and decision trees, Appl. Soft Comput. 133 (2023) 109926.
[20] S. Afrakhteh, F. Mento, U. Khan, L. De Rosa, N. Fatima, Z. Azam, F. Tursi, A. Smargiassi, R. Inchingolo, T. Perrone, et al., Automatic scoring of covid-19 lus videos using cross-correlation-based features aggregated from frame-level confidence levels obtained by a pre-trained deep neural network, in: 2022 IEEE International Ultrasonics Symposium, IUS, IEEE, 2022, pp. 1–3.
[21] 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 132 (2023) 106994.
[22] N. Fatima, U. Khan, X. Han, E. Zannin, C. Rigotti, F. Cattaneo, G. Dognini, M.L. Ventura, L. Demi, Deep learning approaches for automated classification of neonatal lung ultrasound with assessment of human-to-AI interrater agreement, Comput. Biol. Med. 183 (2024) 109315.
[23] R.J. Van Sloun, L. Demi, Localizing B-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results, IEEE J. Biomed. Heal. Informatics 24 (4) (2019) 957–964.
[24] H. Kerdegari, N.T.H. Phung, A. McBride, L. Pisani, H.V. Nguyen, T.B. Duong, R. Razavi, L. Thwaites, S. Yacoub, A. Gomez, et al., B-line detection and localization in lung ultrasound videos using spatiotemporal attention, Appl. Sci. 11 (24) (2021) 11697.
[25] 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.
[26] C. Baloescu, A.A. Rucki, A. Chen, M. Zahiri, G. Ghoshal, J. Wang, R. Chew, D. Kessler, D.K. Chan, B. Hicks, et al., Machine learning algorithm detection of confluent B-lines, Ultrasound Med. Biol. 49 (9) (2023) 2095–2102.
[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.
[28] L. Howell, N. Ingram, R. Lapham, A. Morrell, J.R. McLaughlan, Deep learning for real-time multi-class segmentation of artefacts in lung ultrasound, Ultrasonics 140 (2024) 107251.
[29] L. Chen, J. Rubin, J. Ouyang, N. Balaraju, S. Patil, C. Mehanian, S. Kulhare, R. Millin, K.W. Gregory, C.R. Gregory, et al., Contrastive self-supervised learning for spatio-temporal analysis of lung ultrasound videos, in: 2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI, IEEE, 2023, pp. 1–5.
[30] A. Alfarano, L. Maiano, L. Papa, I. Amerini, Estimating optical flow: A comprehensive review of the state of the art, Comput. Vis. Image Underst. (2024) 104160.
[31] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: IJCAI'81: 7th International Joint Conference on Artificial Intelligence, vol. 2, 1981, pp. 674–679.
[32] R.C. de Amorim, V. Makarenkov, On k-means iterations and Gaussian clusters, Neurocomputing 553 (2023) 126547.
[33] P. Sarker, M.M.H. Shuvo, Z. Hossain, S. Hasan, Segmentation and classification of lung tumor from 3D CT image using K-means clustering algorithm, in: 2017 4th International Conference on Advances in Electrical Engineering, ICAEE, IEEE, 2017, pp. 731–736.
[34] D. Zhang, Y. Yan, Y. Huang, B. Liu, Q. Zheng, J. Zhang, N. Xia, Unsupervised cryo-EM images denoising and clustering based on deep convolutional autoencoder and K-means++, IEEE Trans. Med. Imaging 42 (5) (2022) 1509–1521.
[35] R. Brat, N. Yousef, R. Klifa, S. Reynaud, S.S. Aguilera, D. De Luca, Lung ultrasonography score to evaluate oxygenation and surfactant need in neonates treated with continuous positive airway pressure, JAMA Pediatr. 169 (8) (2015) e151797–e151797.