Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Deep learning approaches for automated classification of neonatal lung ultrasound with assessment of human-to-AI interrater agreement

Noreen Fatima [a], Umair Khan [a], Xi Han [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, Trento, Italy
[b] Fondazione IRCCS San Gerardo Dei Tintori Monza, Italy

ARTICLE INFO

Keywords:
Lung ultrasound
Neonatal ultrasound
Deep learning
Classification
Interrater agreement

ABSTRACT

Neonatal respiratory disorders pose significant challenges in clinical settings, often requiring rapid and accurate diagnostic solutions for effective management. Lung ultrasound (LUS) has emerged as a promising tool to evaluate respiratory conditions in neonates. This evaluation is mainly based on the interpretation of visual patterns (horizontal artifacts, vertical artifacts, and consolidations). Automated interpretation of these patterns can assist clinicians in their evaluations. However, developing AI-based solutions for this purpose is challenging, primarily due to the lack of annotated data and inherent subjectivity in expert interpretations. This study aims to propose an automated solution for the reliable interpretation of patterns in LUS videos of newborns. We employed two distinct strategies. The first strategy is a frame-to-video-level approach that computes frame-level predictions from deep learning (DL) models trained from scratch (F2V-TS) along with fine-tuning pre-trained models (F2V-FT) followed by aggregation of those predictions for video-level evaluation. The second strategy is a direct video classification approach (DV) for evaluating LUS data. To evaluate our methods, we used LUS data from 34 neonatal patients comprising of 70 exams with annotations provided by three expert human operators (3HOs). Results show that within the frame-to-video-level approach, F2V-FT achieved the best performance with an accuracy of 77% showing moderate agreement with the 3HOs. while the direct video classification approach resulted in an accuracy of 72%, showing substantial agreement with the 3HOs, our proposed study lays down the foundation for reliable AI-based solutions for newborn LUS data evaluation.

1. Introduction

Preterm birth is defined as the delivery before 37 completed weeks of gestation. Worldwide, about 15 million babies are born preterm every year [1]. Preterm infants often face respiratory complications due to the immaturity of their respiratory system. At birth, the most frequent respiratory conditions in preterm infants are respiratory distress syndrome (RDS) [2] and Transient Tachypnea of the Newborn (TTN) [3]. Some patients gradually improve in the first weeks of life, while others require prolonged respiratory support and eventually develop chronic lung dysfunction [4]. Understanding and closely monitoring the respiratory condition of preterm infants over time is pivotal in tailoring the respiratory management strategy.

Lung ultrasound (LUS) is a bedside, radiation-free imaging technique that has become increasingly popular in neonatal intensive care units due to its safety, cost-effectiveness, and wide availability, and has been intensively studied in clinical research [5,6]. Specifically, LUS has been utilized to diagnose pneumothorax [7,8], support the differential diagnosis of neonatal respiratory distress [9], describe postnatal adaptation [10], and monitor the natural history of neonatal lung disease [11]. Cursomize RDS management including surfactant replacement in preterm infants [12–14], predict the need for non-invasive respiratory support [15] or mechanical ventilation [16], and early identify infants at risk of bronchopulmonary dysplasia [17].

Semi-quantitative LUS scores are based on the detection and analysis of visual patterns. Such patterns mainly appear as horizontal and vertical artifacts and small to large consolidation areas [18]. Artificial intelligence (AI) [19] has emerged as a promising tool to support healthcare workers by automating the interpretation of LUS patterns and identifying respiratory conditions in newborns [20]. While various AI methods have been developed for the adult patient population, [21–25] limited work has been done on the LUS data analysis for newborn

* Corresponding author.
E-mail address: libertario.demi@unitn.it (L. Demi).

https://doi.org/10.1016/j.compbiomed.2024.109315
Received 25 July 2024; Received in revised form 3 October 2024; Accepted 21 October 2024
Available online 5 November 2024

patients. In this regard, Bassiouny et al. [26], utilized the faster Region-Based Convolutional Neural Network (fRCNN) and RetinaNet models to detect 7 common neonates LUS features (A-lines, coalescent B-lines, separated B-lines, irregular pleural, thick pleural, normal pleural lines and irregular pleural with consolidation). Additionally, Aujla et al. [27] applied a feature extraction method that utilizes recurrence quantification analysis (RQA) on virtual scanlines extracted from LUS images. This method classifies the images into six common neonatal lung conditions (normal, pneumothorax, chronic lung disease, respiratory distress syndrome, tachypnea of the newborn, and consolidation). Building on this work with an expanded dataset, authors in study [28] isolated and extracted localized LUS line and texture patterns of the most common neonatal lung diseases using a two-dimensional (2D) Dual-Tree Complex Wavelet Transform (DTCWT). Gravina et al. [29] focused on classifying a subset of the most common neonatal lung pathologies, specifically distinguishing between RDS and TTN diagnoses using a pre-trained ImageNet model. Another study attempted to develop a fetal lung gestational age (GA) grading model using deep learning (DL) algorithms, where a convolutional neural network (CNN) was designed to identify different categories of fetal lung ultrasound images [30]. Furthermore, Jiao et al. [31] utilized machine learning algorithms such as Support Vector Machine (SVM) and AdaBoost algorithms to classify Neonatal Respiratory Morbidity (NRM) versus normal patients, involving a dataset of 210 fetal LUS images. Although these studies reported high sensitivity (0.82) and specificity (0.84) however, these studies focused on binary classification tasks, which do not fully address the complexity of neonatal lung diseases. Similarly, another study [30] analyzed Neonatal Respiratory Distress Syndrome (NRDS) in premature infants using a deep residual network (DRN), the study reported high accuracy in detecting specific pathological features, such as the disappearance of A-lines and the appearance of B-lines. Some of the studies have focused on fetal LUS image analysis, primarily using textural descriptors to assess respiratory status [32,33].

Many of these studies are constrained by small datasets [30,31] and have not yet fully leveraged advanced DL strategies such as transfer learning, fine-tuning, transformer-based architectures, attention mechanisms, or domain adaptation. Furthermore, these methods have been evaluated on data acquired without a standardized acquisition protocol or scoring system, which is attributed to a higher level of subjectivity. Addressing these shortcomings is crucial for developing more robust and generalizable models capable of accurately classifying a wider range of neonatal lung conditions. Due to the limited number of expert neonatologists in LUS, accurate manual evaluation of lung images is challenging, often leading to increased inter-observer variability. This issue is further compounded by the difficulty in obtaining high-quality annotations, which are crucial for training DL models. Moreover, variations in LUS interpretation, stemming from differences in clinical expertise and experience, can introduce subjectivity in the ground truth (GT) labels used in datasets [34]. Given these constraints, the relatively small dataset size, the subjective nature of interpretation, and the scarcity of precise annotations developing reliable DL models becomes challenging. These difficulties are particularly evident when assessing the models' ability to generalize effectively across diverse datasets.

Therefore, the motivation for this research is grounded in its potential to improve clinical decision-making and enhance patient outcomes, particularly in neonatal care. As we discussed before, manual interpretations of LUS often suffer from subjectivity and variability, leading to inconsistent assessments of neonatal respiratory conditions. By integrating AI tools, this study aims to reduce these inconsistencies, providing more standardized and reliable evaluations. This is especially important in resource-limited or high-demand neonatal intensive care units (NICUs), where time and expertise may be limited. The AI-driven analysis proposed in this study can rapidly and accurately identify key LUS patterns, enabling faster and more informed clinical decisions, which could streamline workflows, reduce diagnostic errors, and inter-observer variability. Moreover, the proposed AI solution could be adapted to other populations, age groups, or even different medical domains, broadening its overall applicability and impact in various healthcare settings.

To achieve the above-mentioned goals, in this study, we will analyze the use of different AI-based methods to assist clinicians in evaluating LUS patterns in newborns. Furthermore, we will also evaluate the interrater agreement between human operators (HOs) and the proposed AI solutions [35]. This offers a comprehensive assessment of the reliability and efficacy of both human and AI-driven interpretations in analyzing LUS patterns in neonates. The neonatal LUS data evaluation performed within the study is in line with new international guidelines and consensus on the use of lung ultrasound [36] and offers significant contributions to the field by:

- Developing AI-based solutions for the automated frame and video level scoring of LUS patterns in newborns.
- Evaluating the impact of the amount of training data on the performance of AI solutions.
- Assessing the interrater agreement between HOs and the proposed AI solutions.

The rest of the paper is organized as follows. The proposed techniques and strategies are presented in Section 2. Experimental settings and results are provided in Sections 3 and 4, respectively. Section 5 analyzes the interrater agreement between the HOs and the AI-based solutions. Finally, Section 6 comprehensively discusses the key findings and draws conclusions based on the results.

2. Methods

2.1. Neonatal LUS scoring

The LUS scores used in this study are an adaptation of the score definitions originally proposed in study [37] (see Fig. 1). For each patient, longitudinal scans were taken along the anterior (midclavicular) and lateral (midaxillary) lines of each hemothorax, defining six chest regions (upper anterior, lower anterior, and lateral for each hemothorax). Each video was scored as per [11], a minor modification of the score that Brat and colleagues originally created and validated for neonates [37]. The score 2 and 3 definitions for each scan are the same as [11]. This comprehensive LUS score represents characteristic signs associated with Transient Tachypnea of the Newborn (TTN) and Respiratory Distress Syndrome (RDS) which describes the most likely conditions of the included population [37].

Specifically, the LUS score was assigned as follows:

- Score 0: Presence of hyper-echoic horizontal artifacts, which appear due to the reverberation effect between the lung surface and the probe, indicating fully aerated state of the lung.
- Score 1: Presence of vertical artifacts, suggesting mild lung alterations as early signs of complications.
- Score 2: The presence of vertical artifacts extending through 100% of the pleural lines indicates more severe lung alterations.
- Score 3: Presence of consolidations with irregular pleural line, indicating severe lung pathology.

2.2. Data description

Data were collected in the Neonatal Intensive Care Unit of Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy, as part of a longitudinal study including preterm infants with a gestational age below 32 weeks and/or a birth weight <1500 g free from major congenital abnormalities. 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. Lung ultrasound was performed with the patient in the supine position using a single

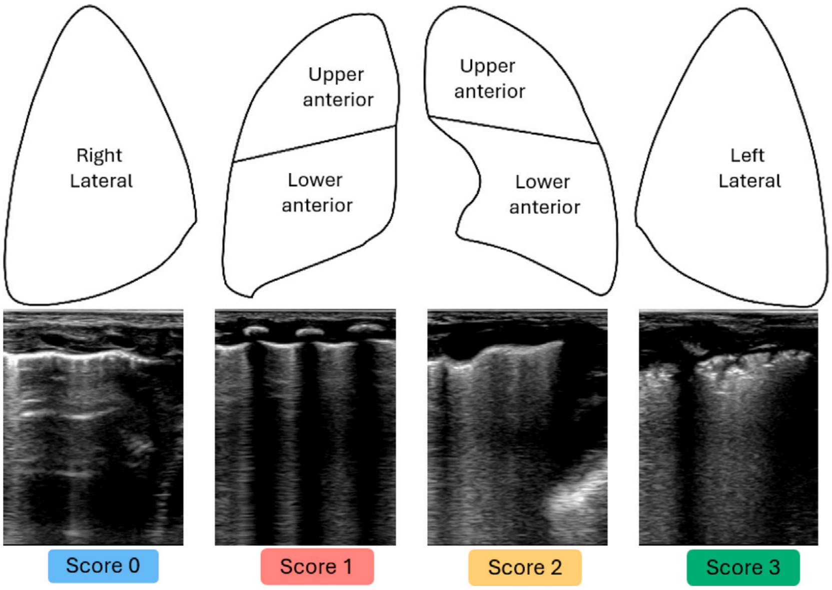


Fig. 1. The upper part of the figure illustrates the division of each lung into three areas. Each area has been assigned a score ranging from 0 to 3 based on the associated LUS patterns.

ultrasound machine (Philips Affiniti 70) and a high-resolution microlinear probe (7.0–15.0 MHz) named as hockey stick. Focal point is set at pleural line with an image depth of 3 cm and frame rate of 63 Hz. The dataset included 70 exams performed on 34 patients scored by three expert human operators (3HOs). The experience levels of the operators were as follows: HO1 with 4 years, HO2 with 7 years, and HO3 with 6 years. Table 1 summarizes the characteristics of study participants. At the time of the exam, infants had a median (Q1, Q3) gestational age was 28.43 (25.71, 30.43) weeks, birth weight was 1025 (720, 1485)g, postnatal age of 25 (9, 45) days, a postmenstrual age of 31.29 (28.29, 35.86) weeks, and a body weight of 1335 (910, 1950) g. Infants had a wide spectrum of prematurity-related lung disorders, namely acute lung disease (including both RDS and TTN), evolving and established BPD. In patients who received pulmonary surfactant, a lung ultrasound was performed afterwards. None of the infants had pulmonary hypoplasia, pneumonia, pneumothorax, or meconium aspiration at the time of assessment. In terms of respiratory support, the median (Q1, Q3) overall airway pressure was 7 (3, 11) cmH_2O, and the fraction of inspired oxygen was 25 (21, 35)%; infants required invasive respiratory support in 24(34%) occasions, nasal continuous positive airway pressure or non-invasive respiratory support on 17 (24%) occasions, high flow nasal cannula on 18 (26%) occasions and were on spontaneous breathing on the remaining 12 (17%) occasions.

Since each exam included 6 videos (one for each chest region), the dataset comprised a total of 420 videos and 78,439 frames. The videos varied in length with number of frames ranging from 607 to 188. The labeling was performed at the video-level for the entire dataset; a subset of 20 exams were additionally labeled at the frame-level.

The scoring distribution for 3HOs at the frame and video-level is shown in Fig. 2.

2.3. Data preparation

During the acquisition of neonates' LUS data, scanner information, and imaging parameters are also captured alongside the frames. This includes textual data regarding imaging settings, measurement lines, and focal point indicators within the field of view (FOV). To ensure accurate and unbiased analysis by an AI-based model, it is important to filter out this information that could lead to ambiguous interpretations. The presence of such information in LUS scans might cause the model to learn incorrect patterns, potentially resulting in wrong predictions [38]. To this extent, we applied pre-processing techniques to extract the FOV while maintaining the spatial resolution of each pixel (see Fig. 3). The removal of this redundant information from the LUS data makes it suitable for further analysis.

Table 1
Characteristics of clinical data utilized for this study.

Variable	Value
N	34
GA, weeks	28.43 (25.71, 30.43)
BW, g	1025 (720, 1485)
SGA, n (%)	5 (15%)
Male, n (%)	19 (56%)
Antenatal steroids, n (%)	29 (38%)
Cesarean section, n (%)	22 (65%)
Surfactant, n (%)	29 (85%)
Duration of O2 supplementation, days	28 (2, 60)
Duration of IMV, days	2 (0, 16)
Duration of respiratory support, days	50 (24, 96)
BPD, n (%)	13 (38%)

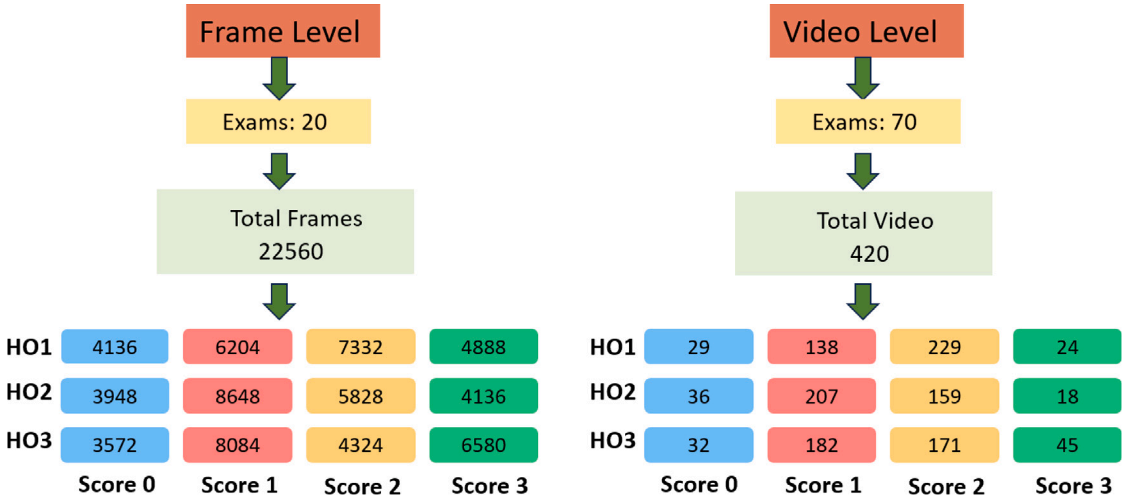
Note*: GA = gestational age; BW = birth weight, SGA = small for gestational age, IMV = invasive mechanical ventilation, BPD = bronchopulmonary dysplasia defined as the need for any respiratory support at 36 weeks postmenstrual age.

2.4. Lung ultrasound video classification

To evaluate LUS video data from newborns, we applied two video-level scoring strategies. These strategies are discussed in detail in subsequent sections.

2.4.1. Frame-to-video-level scoring

Frame-to-video-level (F2V) scoring is a multi-step approach that involves collecting frame-level predictions of a LUS video followed by an aggregation technique to compute the score at the video-level. Two methods are proposed in this regard. The first method (F2V-TS) involves training DL models from scratch for frame-level classification. To this extent, we employed models from traditional deep convolutional neural network (DCNN) [39], to transformer-based architectures such as Vision Transformer (ViT) [40]. In addition, we utilized ResNet-18



	Score 0	Score 1	Score 2	Score 3
HO1	4136	6204	7332	4888
HO2	3948	8648	5828	4136
HO3	3572	8084	4324	6580

	Score 0	Score 1	Score 2	Score 3
HO1	29	138	229	24
HO2	36	207	159	18
HO3	32	182	171	45

Fig. 2. Score-wise distribution of dataset label by 3HOs both at frame and video-level.

which consistently outperformed other models in previous state-of-the-art benchmark studies on adult LUS pattern analysis, demonstrating its success in this area [23,41].

To achieve a score at the video-level, we aggregated frame-level predictions using a threshold-based method [21]. The method involves assigning the worst score to a video if predicted for a given percentage of frames (threshold) within the video. The task hereby requires to find the optimal threshold $\left(TH_{opt}\right)$ that maximizes the overall agreement with the clinical evaluation, as ground truth, at video-level $\left(V_{agr}^{all}\left(TH_i\right)\right)$. The mathematical representation of its calculation is given as,

$$V_{agr}^{\text{all}}\left(TH_i\right) = \sum_{p=1}^{p=n}\sum_{a=1}^{a=m} \frac{100 \cdot V_{agr}\left(TH_i, p, a\right)}{N_{total}} \tag{1}$$

where N_{total} is the total number of videos, n represents the total number of exams, and m represents the total number of scanning areas within an exam.

The second method (F2V-FT) is a domain adaptation task that involves using a pre-trained model, originally trained on LUS data from adults (source domain) [23] and fine-tuning it for frame-level classification of neonatal LUS data (target domain).

To this extent, we used a ResNet-18 model initially trained on a large representative dataset comprising 58,924 LUS frames obtained from 35 adult COVID-19 patients. To adapt the pre-trained model for neonatal LUS analysis, we fine-tuned it on LUS data from newborns. Fine-tuning involves unfreezing all the layers of the DL model and back-propagating the loss throughout the model during training. This allows the model to adjust the learned features of adult data to accurately recognize and interpret the specific characteristics of LUS data from neonates. After obtaining the frame-level predictions, we employed the aggregation technique [21], as used for F2V-TS, to compute the score at the video-level.

2.4.2. Direct-video-level scoring

In this section, we introduce the second strategy based on direct video-level scoring of neonatal LUS patterns (DV). For this purpose, a CNN-LSTM and video vision transformer inspired architecture, transferred sequential lung ultrasound encoding based transformer (TranSLUCEnT) are utilized to classify the spatiotemporal features of neonatal LUS data into four scores, as shown in Fig. 3. The two methods involve a pre-trained CNN model used as a feature extractor, computing the spatial features x_i for each video frame i. Unlike using the models pre-trained on a large ImageNet dataset, we used the ResNet-18 model previously trained on LUS data from the adult patient population [23]. The extracted spatial features are then passed to the LSTM and the transformer encoder for predicting the score at the video level.

Long short-term memory (LSTM)

LSTMs are a special type of recurrent neural network (RNN), designed to learn information over long sequential data, in our case LUS videos. An LSTM unit, as shown in Fig. 4 proposed by Donahue et al. [42], is a memory cell c that encodes the information at the time step t of all the previous inputs till $t-1$. The cell functions by three types of gates, input gate (i), forget gate (f), and output gate (o). The input gate controls if the current input x_t is to be considered or not. The forget gate allows the LSTM to forget the previous memory cell c_{t-1}. Lastly, the output gate determines how much memory will be transferred to the hidden state (h_t).

Values for each gate at the time t are mathematically computed as follows,

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right), \tag{2}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right), \tag{3}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right), \tag{4}$$

$$g_t = \sigma\left(W_{xc}x_t + W_{hc}h_{t-1} + b_g\right). \tag{5}$$

The sum of dot products of previous cell state c_{t-1} with values of forget gate f_t and cell update g_t with input gate i_t results in the new cell state c_t. The output at the hidden state h_t is computed by the dot product of the output gate o_t and the cell state c_t. Mathematically given as

$$c_t = f_t \odot c_{t-1} + i_t \odot \phi g_t, \tag{6}$$

$$h_t = o_t \odot \phi\left(c_t\right), \tag{7}$$

where, σ represents the sigmoidal function, ϕ denotes the hyperbolic tangent and $\odot$ represents the product operation with the values of the gate and weight W_{ij}.

In our proposed CNN-LSTM approach, each input video is normalized to an equal length of 188 frames (maximum frames in a video). Shorter videos were appended with frames with all zeros towards the end. The pre-trained ResNet-18 extracts the distinct characteristics of the video by obtaining features of size 188 × 512. These features are then processed by the LSTM with 256 hidden units. Finally, the output

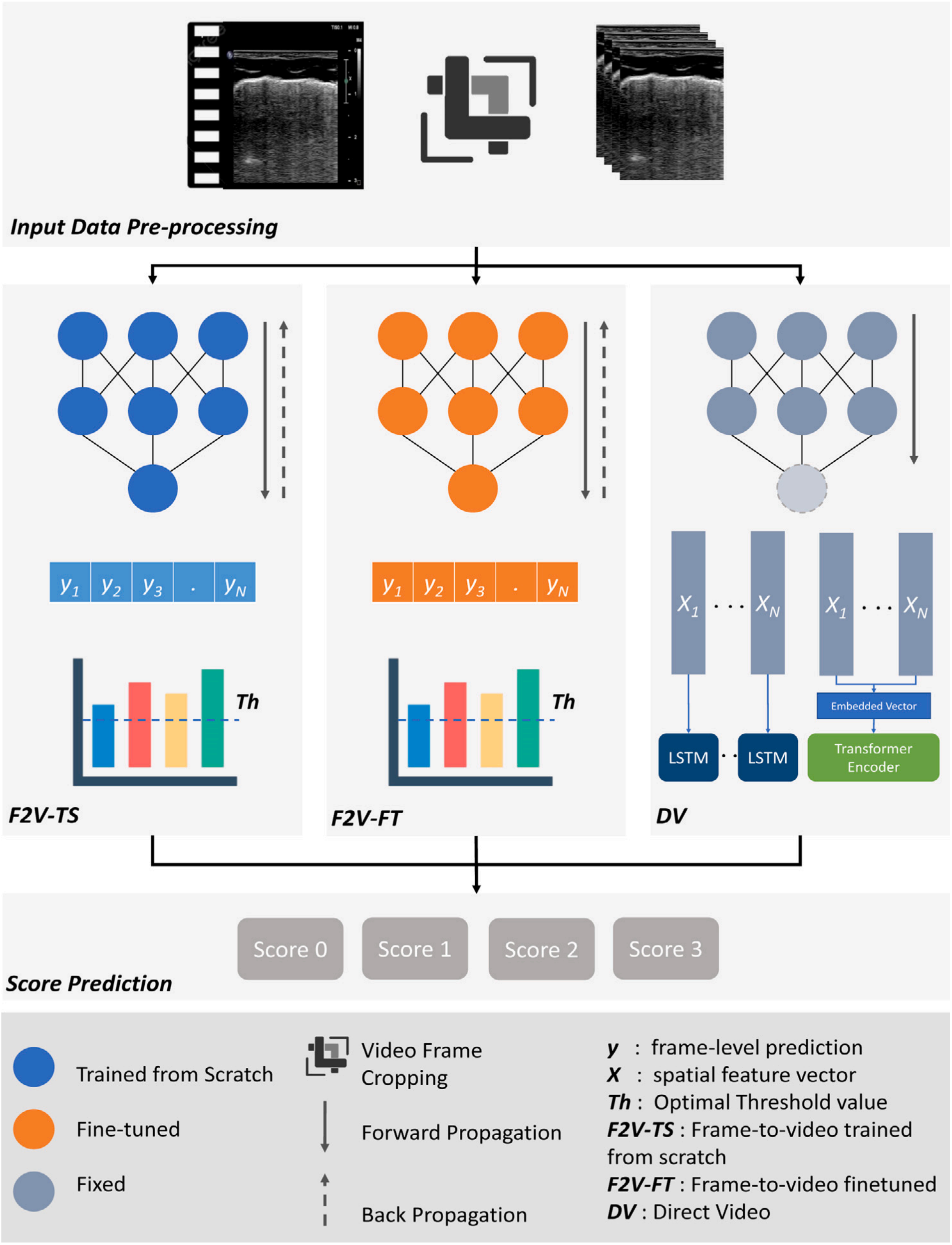


Fig. 3. The block diagram represents the framework for classifying Neonatal Lung Ultrasound (LUS) frame and video data. The initial block encompasses data preprocessing, primarily involving video cropping. The subsequent segment delineates three distinct methods. The first method involves training from scratch for frame-to-video-level classification (F2V-TS). The second method employs a pre-trained model originally trained on adults and fine-tuned for frame-to-video-level classification (F2V-FT). The final method entails direct video-level classification (DV).

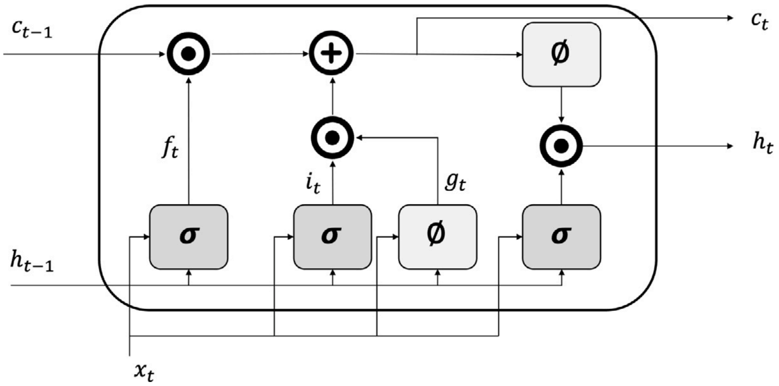


Fig. 4. Internal structure of an LSTM Cell with gates, inputs, and outputs.

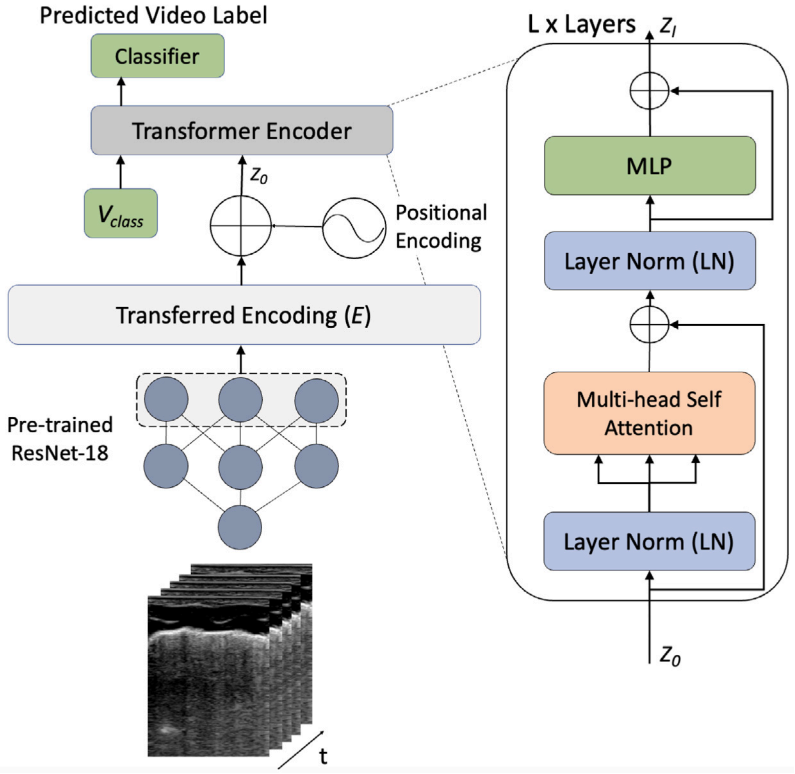


Fig. 5. Block-based representation of TranSLUCEnT.

stage consists of a dense layer of size 4 to classify the spatiotemporal information of a video in the four scores.

TranSLUCEnT

TranSLUCEnT is a video vision transformer-inspired architecture, consisting of an encoding layer that captures frame-level features using the pre-trained ResNet-18 model, followed by a transformer encoder and a classifier. The model leverages the knowledge from the pre-trained model to efficiently extract frame-level feature representations and learn the relationship between the sequence of these frames and the corresponding video label. For a video V consisting of T frames with dimensions $w \times h \times c$ (where h is the height, w is the width, and c is the number of channels), a sequence of frames $(x_1, x_2, \ldots, x_t)$ is extracted. The feature encodings E, extracted from the pre-trained ResNet-18, are combined with a learnable classification token v_{class} for the classification task. This encoded sequence of frames, along with the classification token, is represented as z_0 and passed through the transformer encoder.

As illustrated in Fig. 5, the transformer encoder consists of L identical layers, with each layer comprising two main components: a multi-head self-attention (MSA) block and a fully connected feed-forward (MLP) block. In the final layer of the encoder, TranSLUCEnT uses the first element of the sequence $z_L^{(0)}$ and passes it to the classifier for predicting the class label.

To address the issue of imbalanced distribution during training, CNN-LSTM and TranSLUCEnT model utilized class weighting technique. This method provides a straightforward yet effective way to manage class imbalance by assigning higher weights to samples from the minority class and lower weights to those from the majority class. As a result, the model focuses more on the minority class, improving its ability to make accurate predictions for that class.

3. Experimental setup

3.1. Data configurations

Based on the data annotations provided at the frame and video-levels (see Fig. 2), the entire data from 70 exams is utilized in three different configurations (see Fig. 6). For each configuration, the training, validation, and test splits are made at the exam-level. In the first configuration (config-1), we utilized the frame-level annotated data from 10 exams for training and validation purposes of the DL models in F2V-TS and F2V-FT. The remaining 60 exams, annotated at video-level, are utilized for testing purposes. In the second configuration (config-2), a similar approach is applied to the entire frame-level annotated data from 20 exams for training and validation purposes while the remaining data from 50 exams is utilized for testing purposes. In the third configuration (config-3) the entire data from 70 exams, annotated at video-level, is utilized for training and validation of the DV approach. The data distribution for each configuration is provided in the following sections.

3.1.1. Config-1

In Config-1, we utilized 8210 LUS frames from 10 exams, for training and validation. These frames were annotated at the frame level by 3HOs, with the same scores achieved 100% agreement in their labeling. The remaining 60 exams, consisting of 360 LUS videos, were used for testing and labeled by the 3HOs independently at the video level. The distribution of the dataset based on config-3 is mentioned in Fig. 6.

3.1.2. Config-2

In Config-2, we utilized 22,560 LUS frames from 20 exams, for training and validation. These frames were annotated at the frame level by 3HOs through majority voting (with at least two of the 3HOs assigned the same score). The remaining 50 exams, consisting of 300 LUS videos, were used for testing and labeled by the 3HOs independently at the video level. The distribution of dataset based on config-3 is mentioned in Fig. 6.

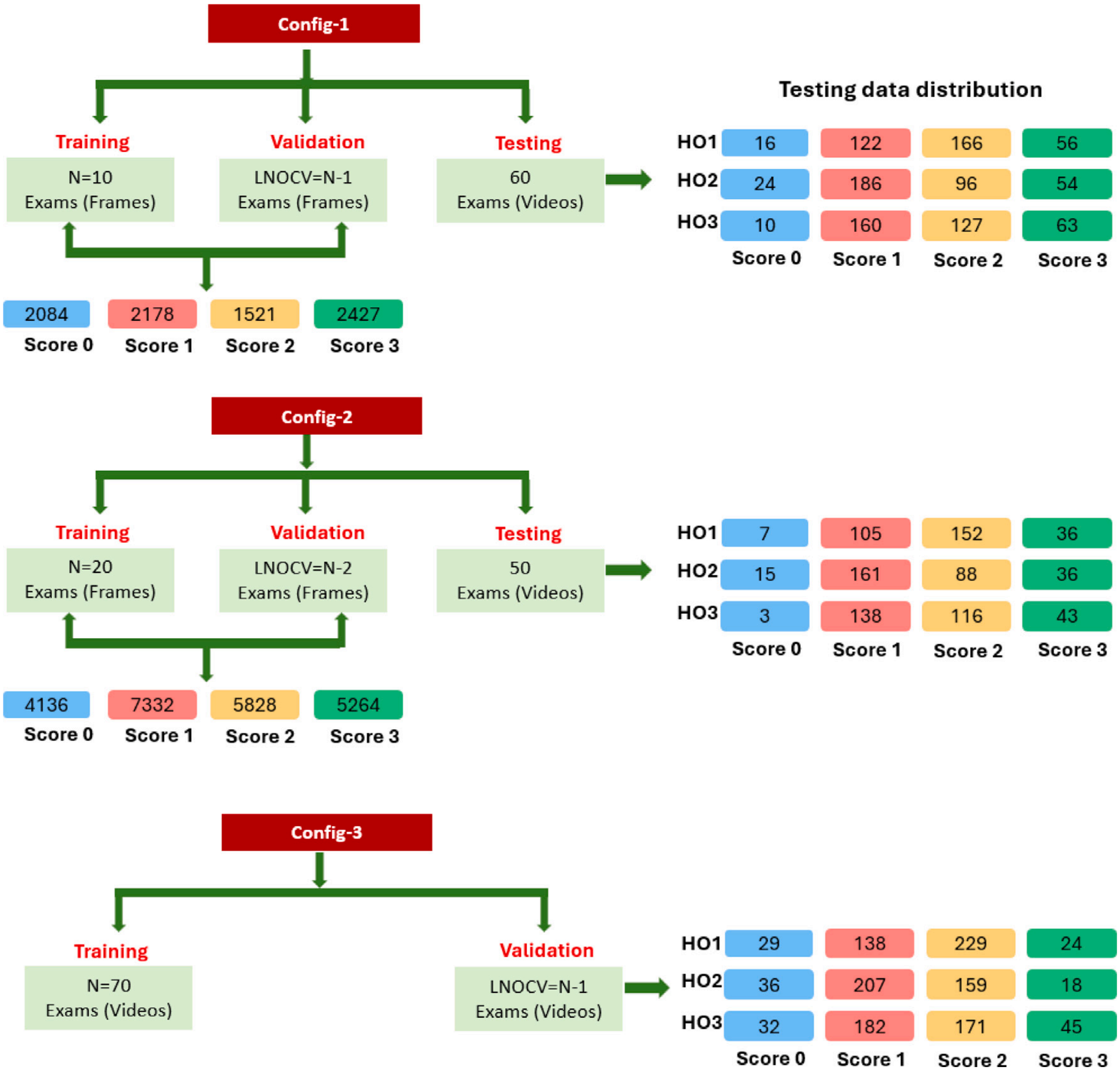


Fig. 6. Score-wise distribution of frame-level and video-level dataset configurations used for training, validation, and testing. F2V-TS and F2V-FT utilized config-1 and config-2 for frame-to-video-level classification, while DV utilized config-3 for direct video-level classification.

3.1.3. Config-3

In Config-3, we utilized the LUS video from 70 exams annotated at the video level. The score provided in the majority among the three HOs for each video was considered as the ground truth for training. This resulted in a dataset comprising a total of 417 LUS videos. The final model predictions were compared independently against the labels provided by the 3HOs at the video level. The distribution of dataset based on config-3 is mentioned in Fig. 6.

3.2. Training and validation strategy

Therefore, all the approaches, F2V-TS, F2V-FT, and DV, employed leave-N-out cross-validation (LNOCV). This technique involves training a model on $T_p - N$ exams and validating it on N exams, where T_p represents the total number of exams, and N varies within the range $N \in \{1, 2, 3, \dots, T_p - 1\}$.

For both the configurations, config-1, and config-2, F2V-TS and F2V-FT are trained and validated over ten folds. Particularly, for config-1, the data from 9 exams are used for training the model in each fold, with the remaining 1 exam designated for validation ($N = 1$). In the case of config-2, which comprises the data from 20 exams, 18 exams are used for training, and the remaining 2 are held out for validation in each fold ($N = 2$).

DL models, for the F2V-TS approach, are trained for 100 epochs with a batch size of 16, enabling iterative parameter updates and accuracy enhancement. Adam optimizer is used to update model parameters based on gradients calculated using backpropagation. Categorical cross-entropy is used as the loss function. To prevent overfitting, early stopping criteria are used in this regard. In the case of the F2V-FT method, the model has been trained in batches of 4 over 50 epochs, with stochastic gradient descent utilized as the optimizer. The learning rate 1e−4 is used. The details on the training settings and parameters can be found in the study [23].

DV strategy does not utilize frame-level labels for training and it is trained on the videos obtained from 70 exams (config-3) for video-level classification. In this regard, for each fold, 69 out of a total of 70 exams are allocated to the training set, and the remaining 1 exam is set aside for validation ($N = 1$). The CNN-LSTM and TranSLUCEnT architectures are trained using backpropagation with cross-entropy loss, calculated over a batch size of 16. The Adam optimizer, with a learning rate of 1e−4, is applied for 30 epochs.

The selection of LNOCV over Leave-One-Subject-Out Cross-Validation (LOSO-CV) is due to the unique nature of our dataset and the developmental variability between exams. Each exam was taken at a different time frame, capturing different stages of lung development in neonates. Specifically, the median interquartile range (Q1, Q3) for the

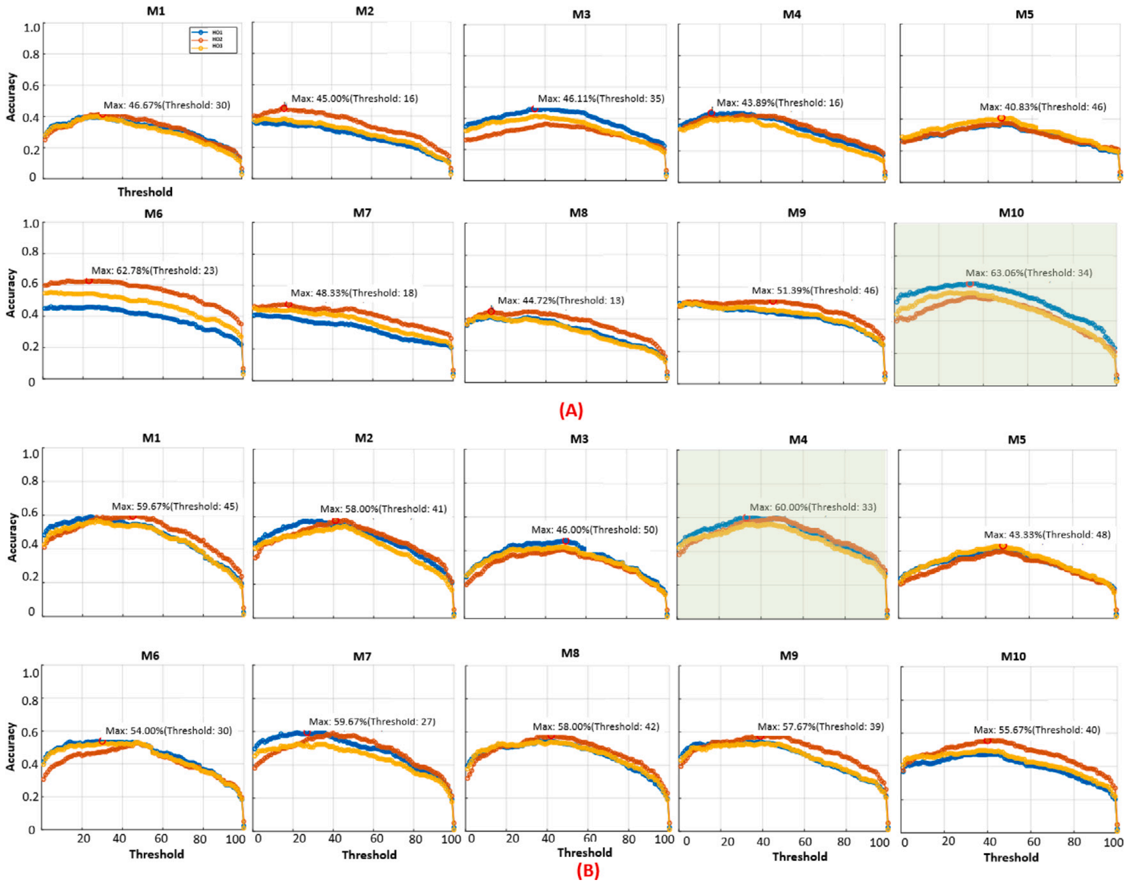


Fig. 7. Accuracy analysis of 10 plots (M1 to M10) ranging from 1% to 100% thresholds by utilizing the F2V-TS approach. The *x*-axis of each plot indicates the number of thresholds applied to predicted frames of each video, while the *y*-axis showcases the accuracy concerning three human operators (HOs): HO1(blue lines), HO2 (red lines), and HO3 (orange lines) representing the ground truth labels. (A), and (B) represent the results of F2V-TS approach utilizing the config-1 and config-2 respectively. The highlighted plots M10 (A) and M4 (B) represents the best performing model.

time gaps between measurements for the same subjects was 22 (9, 49) days, indicating significant temporal variation. For this reason, leaving out an entire subject (LOSO-CV) can result in the model training on data that may not capture the same variability seen within that subject. To this extent, this would limit the model's ability to generalize well when faced with new subjects at different developmental stages. In contrast, the LNOCV model is trained on a diverse range of developmental stages. This approach helps the model generalize better across different exams, as it captures the high variability in lung ultrasound patterns associated with different exam intervals.

3.3. Evaluation metrics and computational resources

In this section, we describe the evaluation metrics used to assess the performance of our proposed methods. These metrics are important for gauging the performance of frame-level classification methods in F2V-TS and F2V-FT and video-level classification of all three methods, F2V-FT, F2V-TS, and DV. The metrics used in our evaluation are accuracy, precision, recall, and F1-Score. The formulated representation of these metrics is given below.

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + TP + FN}, \tag{8}$$

$$\text{Precision} \;= \frac{TP}{TP + FP}, \tag{9}$$

$$\text{Recall} \;= \frac{TP}{TP + FN}, \tag{10}$$

$$\text{F1 Score} \;= 2 \times \frac{\text{Precision} \;\times\; \text{Recall}}{\text{Precision} \;+\; \text{Recall}}, \tag{11}$$

Table 2
Computational analysis of the models.

Method	Models	No. parameters	Training time (s)
F2V-TS	DCNN	1,778,708	config-1: 39.04 , config-2 : 16.29
F2V-TS	VIT	47,515,396	config-1: 3318.62, config-2 : 5098.51
F2V-TS	Resnet-18	11 689 512	config-1: 334 , config-2 :889
F2V-FT	Resnet-18	11 178 564	config-1: 775, config-2 :2236
DV	LSTM	789 508	config-3: 19
DV	TranSLUCEnT [43]	11 689 512	config-3:17

where, TP, TN, FP, and FN indicate true positive, true negative, false positive, and false negative, respectively.

All the experiments for the F2V approach were performed on NVIDIA GeForce RTX 3060, CUDA 10.1, a 16-core processor, and 44 GB of GPU memory (12 GB dedicated GPU memory and 32 GB shared GPU memory). Experiments for the DV approach were performed on NVIDIA A100 Tensor Core GPU, 80 GB. The computational details with respect to the model parameters and training time are given in Table 2.

4. Results

4.1. Lung ultrasound video scoring using F2V-TS

In the first step of F2V-TS approach, among the three employed models (DCNN, VIT, and ResNet-18), the best-performing model for frame-level classification is selected. To do so, each model is trained and validated using config-1. Based on the combined metric performance over the 10 folds, The DCNN model performed the best among

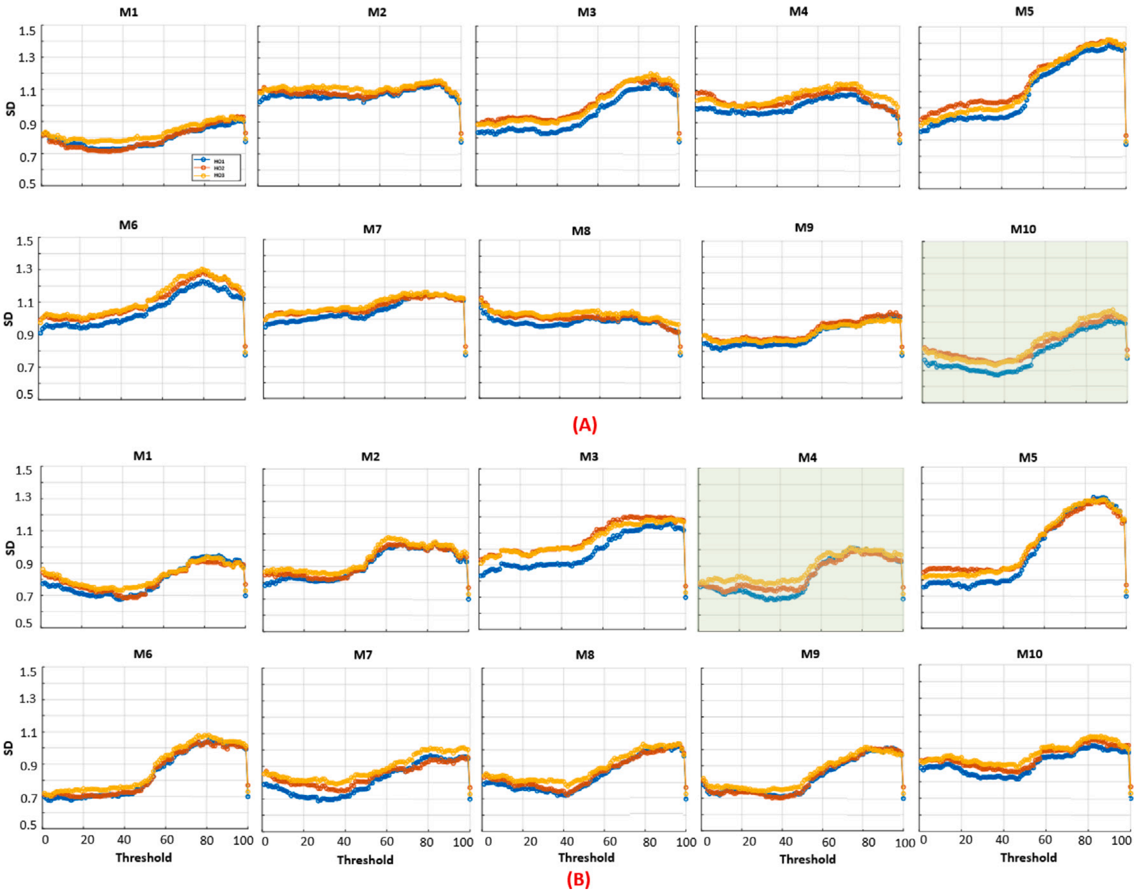


Fig. 8. Standard Deviation (SD) analysis of 10 plots (M1 to M10) on config-1 ranging from 1% to 100% thresholds by utilizing the F2V-TS approach. The *x*-axis of each plot indicates the number of thresholds applied to predicted frames of each video, while the *y*-axis showcases the accuracy concerning three human operators (HOs): HO1 (blue lines), HO2 (red lines), and HO3 (orange lines) are representing the ground truth labels. (A), and (B) represent the results of F2V-TS approach utilizing the config-1 and config-2 respectively. The highlighted plots M10 (A) and M4 (B) represents the best performing model.

the three models, achieving a mean validation accuracy of 54%, compared to VIT with 52% and ResNet-18 with 51%. In the second step, frame-level predictions for the test-set of config-1, from all the 10 DCNN models, (one model for each fold) are aggregated using the threshold-based method mentioned in Section 2.4.1. Video-level performance for each model (M1 to M10) at different levels of thresholds (Th) ranging from 1 to 100% is illustrated in Fig. 7(A). For each model, we saw a similar trend in accuracy across different Th levels. It is observed that, at lower Th levels, videos are likely to be scored accurately as score 3 compared to the other scores, since a lower percentage of frames is required to be predicted as the worst score by the DL model. In contrast, at higher Th levels, videos indicating a fully aerated lung show a higher percentage of frames predicted as score 0 by the DL model and are thus accurately predicted as score 0. As we increase the Th from 1% to 100%, the video-level scoring accuracy for scores 1 and 2 also improved, resulting in an optimal classification performance across all four scores at the given threshold. This observed trend across the Th levels is consistent with the GTs provided by all 3HOs. To determine the final AI solution using the F2V-TS approach, among the 10 models of DCNN, the corresponding model is selected, based on the highest agreement achieved across either of the 3HOs. In this regard, model M10 exhibited the highest accuracy of 63.06% at a Th value 34% w.r.t the GT labels given by HO1 (GT-HO1) (accuracy trend w.r.t Th levels shown in blue-lines of Fig. 7A). For the model M10, the highest accuracy achieved w.r.t GT labels given by HO2 (GT-HO2) and HO3 (GT-HO3) are reported as 54.72% at a Th of 33%, and 57.50% at a Th of 25%, respectively (accuracy trend w.r.t Th levels shown in red and orange lines of Fig. 7A). Conclusively, the final predicted scores (PredTS) from M10 at above-selected thresholds w.r.t the 3HOs (34%: GT-HO1, 33%: GT-HO2, and 25%: GT-HO3) are utilized to compute the video-level scores. The PredTS on the test-set of config-1 ($c1$) for GT-HO1, GT-HO2, and GT-HO3 are represented as PredTS$_{c1HO1}$, PredTS$_{c1HO2}$, and PredTS$_{c1HO3}$ respectively.

To analyze the impact of the amount of training data for frame-level classification models, Fig. 7(B) shows the performance of 10 DCNN models on the test-set of config-2. A similar trend is observed across all the 10 models as shown in Fig. 7(A). Among the 10 models, M4 provides the highest accuracy of 60.00% at Th of 33% w.r.t GT-HO1. Similarly, the highest accuracy w.r.t other GTs are reported as 59.33% (Th: 44%, GT-HO2) and 55.67% (Th: 41%, GT-HO3), respectively. The PredTS on testset of config-2 ($c2$) on the given threshold w.r.t GT-HO1, GT-HO2, and GT-HO3 are represented as PredTS$_{c2HO1}$, PredTS$_{c2HO2}$, and PredTS$_{c2HO3}$ respectively.

Standard deviation (SD) calculates the variation between the video-level scores computed across different Th values (1% to 100%) and GT labels assigned by the 3HOs, as provided in Fig. 8. The SD values w.r.t the increasing Th are analyzed for both configurations. In the case of config-1 (see Fig. 8A), the model M10 showed SD value of 0.68 (Th: 34%, GT-HO1), 0.75 (Th: 33%, GT-HO2) and 0.75 (Th: 25%, GT-HO3), whereas for config-2 (see Fig. 8B) M4 showed SD values of 0.69 (Th: 33%, GT-HO1), 0.75 (Th: 44%, GT-HO2) and 0.81 (Th: 41%, GT-HO3). Overall all the models trained on config-1, compared to config-2, demonstrated relatively lower and more consistent SD values w.r.t the GT from 3HOs. This behavior can be attributed to the fact that in config-1, the GT labels collected to train the DL model are based on the perfect (i.e. 100%) agreement among the 3HOs compared to the majority vote-based labels in config-2. Although the data increased in config-2, varying interpretations of LUS patterns among the 3HOs

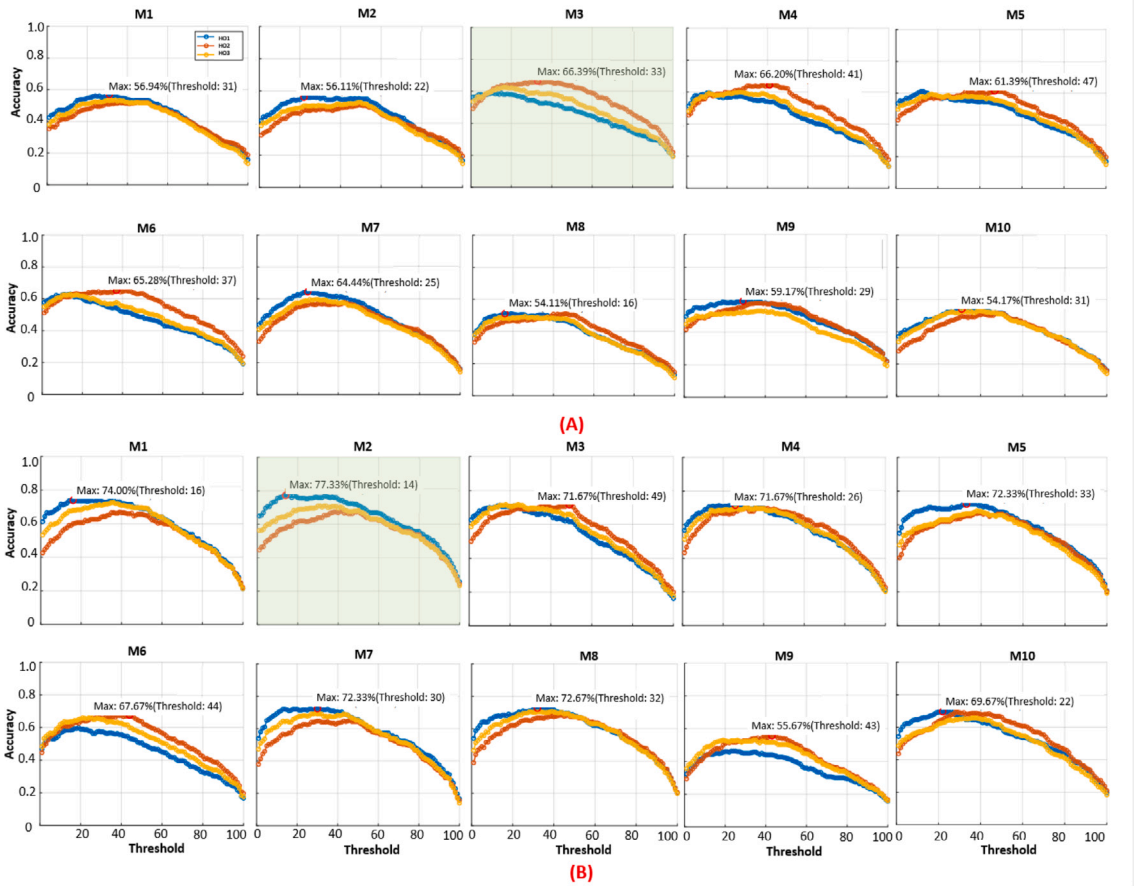


Fig. 9. Accuracy analysis of 10 plots (M1 to M10) ranging from 1% to 100% thresholds by utilizing the F2V-FT approach. The x-axis of each plot indicates the number of thresholds applied to predicted frames of each video, while the y-axis showcases the accuracy concerning three human operators (HOs): HO1(blue lines), HO2 (red lines), and HO3 (orange lines) representing the ground truth labels. (A), and (B) represent the results of F2V-FT approach utilizing the config-1 and config-2 respectively. The highlighted plots M3 (A) and M2 (B) represents the best performing model.

introduced an implicit subjectivity within the GT labels to train the DL models.

4.2. Lung ultrasound video scoring using F2V-FT

For F2V-FT approach, frame-level scores from all 10 models of ResNet-18, one model for each fold, using the config-1 are aggregated by the threshold-based method. Video-level performance for each model (M1 to M10) at different levels of Th (1 to 100%) are illustrated in Fig. 9(A). For each model, we observed a similar trend among 3HOs achieved w.r.t increasing Th levels. However, slight improvements in accuracy are observed in all 10 models. This is likely due to the models being pre-trained on a larger representative dataset from the adult patient population.

Among the 10 models of ResNet-18, M3 exhibited the highest accuracy of 66.39% (Th: 15%, GT-HO2) (accuracy trend w.r.t Th levels shown in red lines Fig. 9A). For the model M3, the highest accuracy achieved w.r.t GT-HO2 and GT-HO3 are reported as 65.83% at Th of 33%, and 61.94% at a Th of 16%, respectively. The final video-level predicted scores (PredFT) given by M3 at Th w.r.t 3HOs (15%: GT-HO1, 33%: GT-HO2, and 16%: GT-HO3) are utilized for video-level scores. The PredFT on $c1$ on the given threshold w.r.t GT-HO1, GT-HO2, and GT-HO3 are represented as PredFT$_{c1HO1}$, PredFT$_{c1HO2}$, and PredFT$_{c1HO3}$ respectively.

Similarly, for config-2 the performance of ten ResNet-18 models shown in Fig. 7(B). Among these models, M2 demonstrated the highest performance with an accuracy of 77.33% (Th: 14%, GT-HO1), see Table 3. Similarly, the highest accuracy achieved by M2 are reported as 68.00% (Th: 50%, GT-HO2), and 71.00% (Th: 30%, GT-HO3), respectively. It is noted that the scoring definitions of the LUS patterns in adults and neonates are not exactly aligned. This significant increase in the accuracies can be due to the increased training data available to adjust the model features from source to target domain more effectively. The PredFT on $c2$ based on the given threshold w.r.t GT-HO1, GT-HO2, and GT-HO3 are represented as PredFT$_{c2HO1}$, PredFT$_{c2HO2}$, and PredFT$_{c2HO3}$, respectively.

Looking at the SD across the different Th for config-1 (see Fig. 10A), the model M3 showed an SD value of 0.73 across the Th (Th: 15%, GT-HO1), while 0.77 across (Th: 33%, GT-HO2) and (Th: 16%, GT-HO3). Similarly for config-2, M2 showed relatively lower SD value of 0.64 (Th: 14%, GT-HO1), 0.57 (Th: 50%, GT-HO2), and 0.52 (Th: 30%, GT-HO3) (see Fig. 10B). We noted that unlike the overall trend of SD observed for the F2V-TS approach, F2V-FT models trained on config-2 demonstrated relatively lower and more consistent SD values w.r.t the GT from 3HOs. One possible explanation for this improvement can be attributed to the potential of the pre-trained model to mitigate the subjectivity of GT, in config-2, with an increase in the training data.

4.3. Lung ultrasound video scoring using DV

The *DV* approach involved training and validating both CNN-LSTM and TranSLUCEnT models using the config-3 setup. Among the two, the TranSLUCEnT model demonstrated superior performance in video-level classification compared to CNN-LSTM (see Table 3). The model

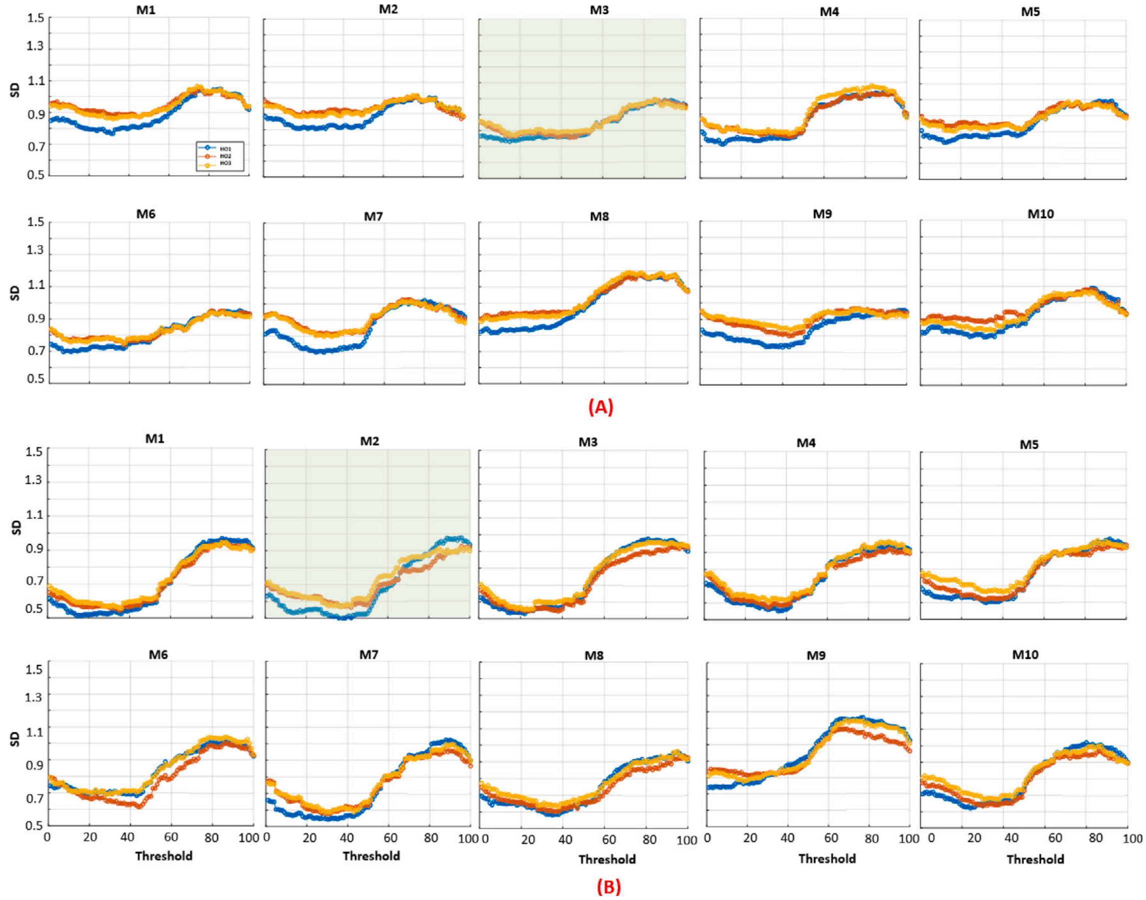


Fig. 10. Standard Deviation (SD) analysis of 10 plots (M1 to M10) on config-1 ranging from 1% to 100% thresholds by utilizing the F2V-FT approach. The *x*-axis of each plot indicates the number of thresholds applied to predicted frames of each video, while the *y*-axis showcases the accuracy concerning three human operators (HOs): HO1 (blue lines), HO2 (red lines), and HO3 (orange lines) are representing the ground truth labels. (A), and (B) represent the results of F2V-FT approach utilizing the config-1 and config-2 respectively. The highlighted plots M3 (A) and M2 (B) represents the best performing model.

predictions across the GT-HO1 achieved a mean accuracy of 71%. The model predictions (PredDV) across the GT-HO1 achieved a mean accuracy of 71%. However, across the GT-HO2 and GT-HO3, the model showed a slight improvement, reaching a mean accuracy of 72%. These results are based on the average metrics calculated over 69 folds. The final predicted scores from the TranSLUCEnT model in config-3 are denoted as PredDV$_{c3}$.

5. Inter-rater reliability analysis (IRR)

To determine inter-rater reliability (IRR), Fleiss's kappa (f) coefficient [44] is computed. The interpretation of f follows specific criteria with its value ranging from less than 0 to 1, indicating different levels of agreement. To this extent, an f value less than zero indicates poor agreement, whereas, a value between 0.00 and 0.20, indicates slight agreement. An f value between 0.21 and 0.40 indicates fair agreement whereas, a value between 0.41 and 0.60 represents moderate agreement. An f value between 0.61 and 0.80 shows substantial agreement whereas, a value above 0.81 indicates a perfect agreement. The kappa values are calculated between the 3HOs and each of the AI solutions PredTS, PredFT, and PredDV on different configurations, indicating the level of agreement between the 3HOs and AI at the video-level. The Fleiss' kappa analysis of these solutions with the HOs is elaborated in the subsequent sections.

5.1. HOs vs. AI (F2V-TS)

Fig. 11 shows the Fleiss' kappa analysis among the 3HOs and PredTS, obtained using config-1, as AI solutions. The *x*-axis indicates whether Fleiss' kappa is computed considering all the scores (overall shows mean kappa value) and only specific scores (from 0 to 3). Among the 3HOs (Fig. 11 3HOs), the mean kappa value is found as 0.61 (all the scores together). The score-wise kappa value shows a moderate agreement for score 1 and score 2, while substantial agreement is observed for score 0 and score 3. This suggests that for all the HOs, classifying LUS data into scores 0 and 3 is relatively simpler than scores 1 and 2. This is primarily due to the blurred boundaries between the neighboring scores resulting in a subjective nature of the analysis [35]. The Fleiss' kappa analysis in subsequent plots Fig. 11(A, B, C) illustrates the agreement between 3HOs and AI predictions. The overall mean kappa values are found as 0.48 for both 3HOs vs. PredTS$_{c1HO1}$, and 3HO vs. PredTS$_{c1HO2}$, and 0.47 for 3HOs vs. PredTS$_{c1HO3}$, indicates a moderate level of agreement across all cases. This suggests that the AI solutions generally aligns well with the 3HOs assessments. When analyzing specific scores, the agreement remains moderate for scores 0, 1, and 2 across all three cases, indicating that the AI solutions reliably replicates the variability observed among the HOs. Also, for score 3, which likely represents more severe or distinct lung conditions, the agreement is more substantial. This suggests that certain AI solutions are more adept at recognizing features associated with higher scores, such as lung consolidations. Overall, the findings suggest that introducing AI solutions as an operator alongside the 3HOs in kappa analysis reduces the overall variability (see Fig. 11A, B, C), which is notably higher among the human operators alone (see Fig. 11, 3HOs).

Fig. 12 shows the Fleiss' kappa analysis among the 3HOs and PredTS, obtained using config-2 as AI solutions. Among the 3HOs, the mean kappa value is found to be 0.58. The score-wise analysis, as

Table 3
Performance analysis of three different methods at video-level classification on different configurations. (Performance in Blue represents the best-performing strategy for video-level classification following the F2V-FT approach on config-2).

Methods	Models	Accuracy	Precision	Recall	F1 Score	Configurations	Ground Truth
		0.63	**0.63**	**0.63**	**0.63**	config-1	GT-HO1
F2V-TS	DCNN	0.54	0.60	0.54	0.55	config-1	GT-HO2
		0.57	0.59	0.57	0.57	config-1	GT-HO3
		0.60	**0.53**	**0.60**	**0.60**	config-2	GT-HO1
F2V-TS	DCNN	0.59	0.61	0.59	0.60	config-2	GT-HO2
		0.55	0.58	0.55	0.56	config-2	GT-HO3
		0.58	0.61	0.58	0.58	config-1	GT-HO1
F2V-FT	ResNet-18	**0.66**	**0.65**	**0.66**	**0.65**	config-1	GT-HO2
		0.62	0.62	0.62	0.62	config-1	GT-HO3
		0.77	**0.78**	**0.77**	**0.76**	config-2	GT-HO1
F2V-FT	ResNet-18	0.68	0.77	0.68	0.68	config-2	GT-HO2
		0.71	0.76	0.71	0.70	config-2	GT-HO3
		0.58	**0.60**	0.58	**0.57**	config-3	GT-HO1
DV	LSTM	0.53	**0.61**	0.53	0.55	config-3	GT-HO2
		0.55	0.59	0.55	0.55	config-3	GT-HO3
		0.71	0.72	0.71	0.71	config-3	GT-HO1
DV	TranSLUCEnT	**0.72**	**0.75**	**0.72**	**0.73**	config-3	GT-HO2
		0.72	**0.72**	0.72	0.72	config-3	GT-HO3

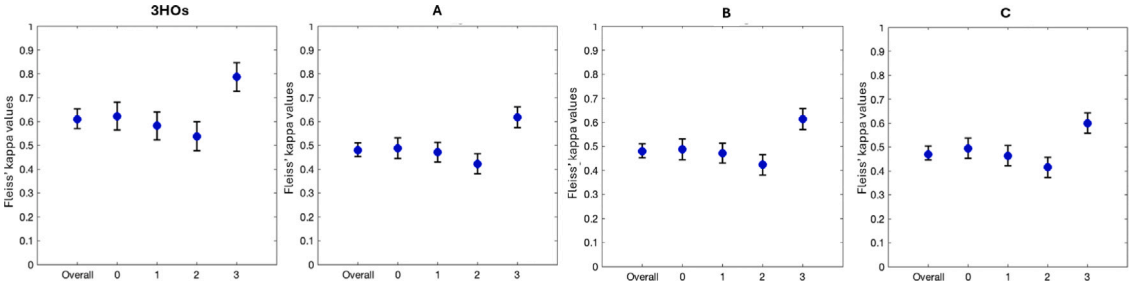


Fig. 11. (From Left to Right) illustrates Fleiss' kappa analysis between human operators (HOs) and AI solutions at video-level using config-1; graph (3HOs) illustrates the kappa values between 3HOs at video-level; graph (A) illustrates the kappa values between 3HOs vs. PredTS$_{c1HO1}$; graph (B) illustrate the kappa values between 3HOs vs. PredTS$_{c1HO2}$; graph (C) illustrate the kappa values between 3HOs vs. PredTS$_{c1HO3}$.

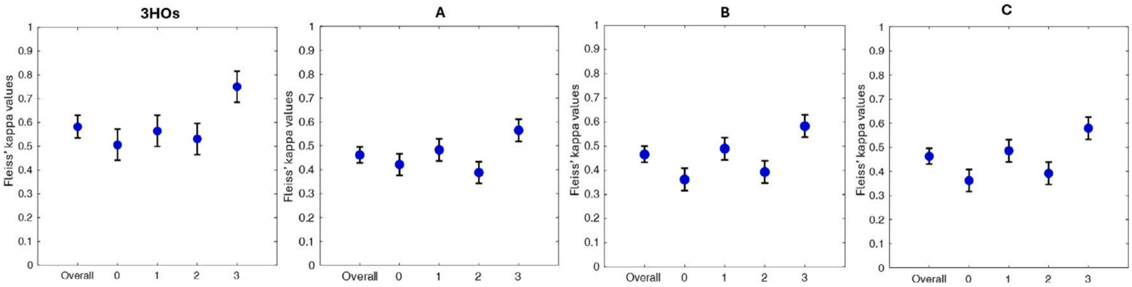


Fig. 12. (From Left to Right) illustrates Fleiss' kappa analysis between human operators (HOs) and AI solutions at video-level using config-2; graph (3HOs) illustrates the kappa values between 3HOs at video-level; graph (A) illustrates the kappa values between 3HOs vs. PredTS$_{c2HO1}$; graph (B) illustrate the kappa values between 3HOs vs. PredTS$_{c2HO2}$; graph (C) illustrate the kappa values between 3HOs vs. PredTS$_{c2HO3}$.

illustrated in Fig. 12 (3HOs), reveals moderate agreement among the 3HOs for scores 0, 1, and 2, with a substantial agreement for score 3. The lower kappa value for score 0 among the 3HOs can be attributed to the reduced number of videos classified as score 0 in the test set of config-2. Introducing AI solutions as an operator alongside the 3HOs results in a moderate agreement, with a mean kappa value of 0.46 across all three cases (see Fig. 12A, B, C). However, the score-wise agreement between 3HOs vs. PredTS$_{c2HO1}$ shows moderate agreement for scores 0, 1, and 3, but only fair agreement for score 2 (see Fig. 12A). On the other hand, 3HOs vs. PredTS$_{c2HO2}$ and 3HOs vs. PredTS$_{c2HO3}$, the results show fair agreement for scores 0 and 2 and moderate agreement for scores 1 and 3. The fair agreement for score 0 could be attributed to the limited number of score 0 instances in the test set, which may have affected the model's performance. Additionally, the model seems to struggle with distinguishing between the patterns associated with score 2 and score 3, leading to lower agreement for these scores. Overall, the trend in Fig. 12 (A, B, C) remains consistent with the 3HOs. This suggests that the high variability observed in the 3HOs is significantly reduced when AI is introduced as the operator.

5.2. HOs vs. AI (F2V-FT)

Fig. 13 shows the Fleiss' kappa analysis among the 3HOs and PredFT, obtained using config-1, as AI solutions. The agreement among the 3HOs is previously discussed in 5.1 (mean kappa value is 0.61) indicating the substantial agreement when it comes to kappa analysis among the 3HOs vs. PredFT$_{c1HO1}$ and 3HOs vs. PredFT$_{c1HO3}$ (see Fig. 13A, C), the mean kappa value is found as 0.51 for both cases.

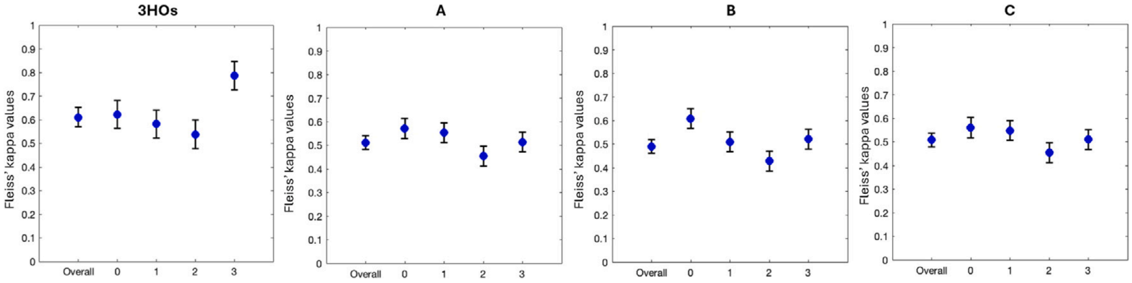


Fig. 13. (From Left to Right) illustrates Fleiss' kappa analysis between human operators (HOs) and AI solutions at video-level using config-1; graph (3HOs) illustrates the kappa values between 3HOs; graph (A) illustrates the kappa values between 3HOs vs. PredFT$_{c1HO1}$; graph (B) illustrate the kappa values between 3HOs vs. PredFT$_{c1HO2}$; graph (C) illustrate the kappa values between 3HOs vs. PredFT$_{c1HO3}$.

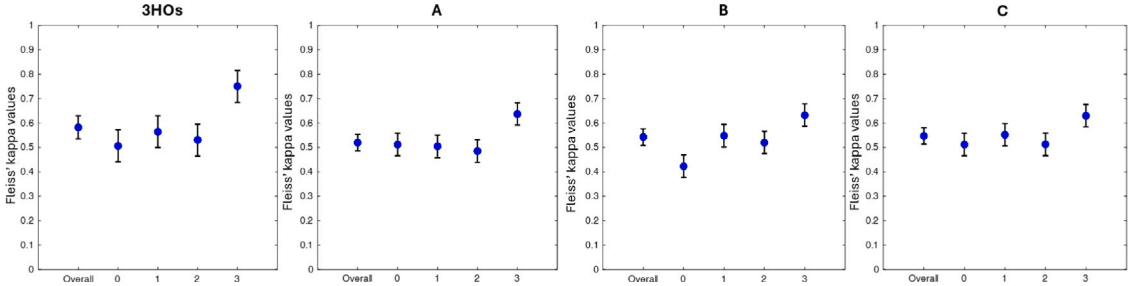


Fig. 14. (From Left to Right) illustrates Fleiss' kappa analysis between human operators (HOs) and AI solutions at video-level using config-2; graph (3HOs) illustrates the kappa values between 3HOs; graph (A) illustrates the kappa values between 3HOs vs. PredFT$_{c2HO1}$; graph (B) illustrate the kappa values between 3HOs vs. PredFT$_{c2HO2}$; graph (C) illustrate the kappa values between 3HOs vs. PredFT$_{c2HO3}$.

Similarly, 3HOs and PredFT$_{c1HO2}$ (see Fig. 13B) the mean kappa value is found as 0.49. These results show slightly lower kappa values compared to the ones among the 3HOs. Specifically, score-wise agreement (see Fig. 13 A, B, C) score 3 indicates moderate agreement among in all three cases. This is because the DL models are previously trained on adults LUS patterns slightly deviating from the ones of neonates [45]. When scoring LUS patterns in adults, small and large consolidations are labeled as score 2 and score 3 respectively. Whereas, in neonates, consolidations are only labeled as score 3. Consequently, this causes the potential misclassification of consolidation patterns into score 2 instead of score 3 by the AI solutions.

Fig. 14 represents the Fleiss' kappa analysis among the 3HOs and PredFT obtained using config-2 as AI solutions. The agreement among the 3HOs was previously discussed in Section 5.1 (mean kappa value is 0.58) indicating the moderate agreement. In the kappa analysis between the 3HOs and PredFT$_{c2HO1}$ (see Fig. 14A), the mean kappa value is 0.52. Similarly, for 3HOs and PredFT$_{c2HO2}$ (see Fig. 14B), the mean kappa value is 0.54. Among the 3HOs and PredFT$_{c2HO3}$ (see Fig. 14C), the mean kappa value is 0.51. All three cases demonstrate moderate agreement. The score-wise kappa analysis among the 3HOs and AI solutions (Fig. 14A, B, C) shows moderate agreement for score 0, 1, and 2, and substantial agreement for score 3. Unlike the trend observed for config-1 and config-2 w.r.t score 1, 2, and 3 are the same. However, since the model was trained on a smaller set of score 0, the kappa value has decreased in this case. The increased agreement indicates that the pre-trained model can learn the features from the neonatal LUS data more effectively due to the larger training data size. This potentially enabled the pre-trained model to align the feature representations of the LUS patterns from adults more accurately to neonates. Overall, the findings suggest that introducing AI as an operator alongside the 3HOs in kappa analysis reduces the overall variability (see Fig. 14A, B, C), which is notably higher among the 3HOs (see Fig. 14, 3HOs).

5.3. HOs vs. AI (DV)

Fig. 15 presents the Fleiss' kappa analysis 3HOs and PredDV, obtained using config-3 as the AI solution. Among the 3HOs (Fig. 15 3HOs), the analysis revealed a mean kappa value of 0.65. The score-wise analysis shows moderate agreement for scores 1 and 2, substantial agreement for score 0, and perfect agreement for score 3. The moderate agreement among the HOs when labeling scores 1 and 2 suggests that the similarity between these scores both exhibiting vertical artifacts, with score 1 showing mild alterations and score 2 showing severe ones made it difficult for the HOs to differentiate between them. In contrast, scores 0 and 3 had more distinct features, resulting in higher agreement among the HOs. These findings align with trends described in previous studies on adult LUS patterns [35].

The Fleiss' kappa analysis in Fig. 15 (3HOs vs. AI) further illustrates the agreement between the 3HOs and PredDV$_{c3}$, with a mean kappa value of 0.62. Score-wise, the AI solution exhibited moderate agreement for score 2 and substantial agreement for the other scores. The inclusion of an AI solution alongside the 3HOs shows that the AI performed similarly to the 3HOs and effectively distinguished the more distinct features associated with scores 0 and 3. Notably, the integration of AI improved the agreement for score 2, raising it from moderate (among 3HOs) to substantial (among 3HOs vs. AI). This improvement suggests that the AI learned the features obtained from the pre-trained model trained on a large population of adult LUS patterns, and is better at recognizing the subtle differences between scores 1 and 2. The AI model identified the distinctive features in score 2, where the HOs struggled due to the similarity of artifacts between score 1 and score 2. As a result, the variability in interpretation was reduced by introducing the AI as the operator with 3HOs, improving the overall consistency and reliability of the analysis. This demonstrates that the AI can complement HO interpretation by enhancing accuracy, especially in challenging cases.

Statistical analysis between the 3HOs and proposed AI solutions in terms of kappa value is given in Table 4.

6. Discussion and conclusion

In this study, we introduced two strategies for automated scoring of LUS patterns in newborns: frame-to-video-level (F2V-TS and F2V-FT)

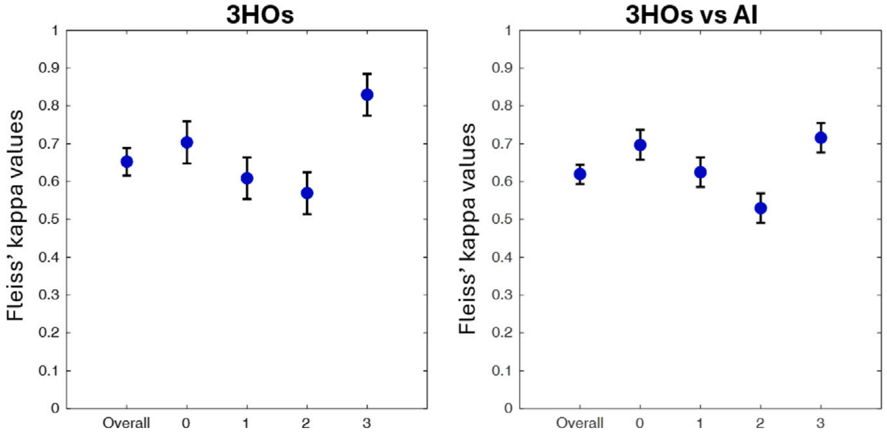


Fig. 15. (From Left to Right) illustrates Fleiss' kappa analysis between three human operators (3HOs) at video level and 3HOs vs. AI solution following the TranSLUCEnT model prediction.

Table 4
Statistical analysis using Fleiss's kappa, conducted between the three human operators and the proposed AI solutions (F2V-TS, F2V-FT, and DV) across different configurations.

Operators	Mean fleiss' kappa values	Score 0	Score 1	Score 2	Score 3	Configurations
3HOs	0.61	0.62	0.58	0.53	0.78	config-1
3HOs	0.58	0.50	0.56	0.53	0.75	config-2
3HOs	0.65	0.70	0.60	0.57	0.83	config-3
3HOs vs. PredTS$_{c1HO1}$	0.48	0.48	0.47	0.42	0.61	config-1
3HOs vs. PredTS$_{c1HO2}$	0.48	0.48	0.47	0.42	0.61	config-1
3HOs vs. PredTS$_{c1HO3}$	0.47	0.49	0.46	0.41	0.60	config-1
3HOs vs. PredTS$_{c2HO1}$	0.46	0.42	0.48	0.38	0.56	config-2
3HOs vs. PredTS$_{c2HO2}$	0.46	0.36	0.49	0.39	0.58	config-2
3HOs vs. PredTS$_{c2HO3}$	0.46	0.36	0.48	0.39	0.57	config-2
3HOs vs. PredFT$_{c1HO1}$	0.51	0.57	0.55	0.45	0.51	config-1
3HOs vs. PredFT$_{c1HO2}$	0.49	0.61	0.51	0.43	0.52	config-1
3HOs vs. PredFT$_{c1HO3}$	0.51	0.56	0.55	0.45	0.51	config-1
3HOs vs. PredFT$_{c2HO1}$	0.52	0.51	0.50	0.48	0.63	config-2
3HOs vs. PredFT$_{c2HO2}$	0.54	0.42	0.54	0.52	0.63	config-2
3HOs vs. PredFT$_{c2HO3}$	0.51	0.51	0.55	0.51	0.63	config-2
3HOs vs. PredDV$_{c3}$	0.62	0.69	0.62	0.53	0.71	config-3

and direct video-level (DV) approaches. Both strategies were trained on foundational deep learning models, such as DCNN, and widely used method pre-trained Resnet-18 which was outperforming in state-of-the-art studies [23,41], as well as on more advanced models, including VIT transformers, LSTM, and TranSLUCEnT with attention mechanisms. The selection of these models was aimed at establishing a benchmark for neonatal studies, ensuring a fair comparison with state-of-the-art research. Frame-to-video-level (F2V) scoring involves producing frame-level predictions of a LUS video followed by an aggregation technique to compute the score at the video level. Frame-level classification models of the two methods (F2V-TS and F2V-FT) were initially trained on a smaller dataset (config-1) labeled by 3HOs. The smaller dataset was chosen to determine if comparable results could be achieved without increasing the training size. Summarizing our findings, F2V-TS shows that DCNN performs the best (accuracy: 0.54) among the employed DL models (DCNN, ResNet-18, and VIT) for frame-level classification. The corresponding accuracy at the video-level was 0.63. This shows that classifying LUS frames using relatively less complex models like DCNN, can achieve good classification performance. In comparison, the mean Fleiss' kappa value among the 3HOs on testset of config-1 was found as 0.61 thus indicating substantial agreement, also F2V-TS as the AI solution with 3HOs showed a moderate agreement with the mean kappa value of 0.48. Although the kappa value between the 3HOs and the AI was lower, incorporating AI as an operator reduced the variability, which had been higher among the 3HOs alone. The second method F2V-FT was implemented using a pre-trained ResNet-18 model, initially trained on 58,924 frames of adults collected from multiple centers [23]. The pre-trained model (ResNet-18), used in the F2V-FT approach, outperformed DCNN in F2V-TS, achieving a relatively higher accuracy of 0.66 at the video-level. This slight improvement in accuracy (3%) can be due to the fact that, although the model in F2V-FT was previously trained on relatively larger LUS data from adults, the limited data in config-1 seems insufficient to adapt the differences in scoring definitions of LUS patterns between adults and neonates. F2V-FT as the AI solution with 3HOs showed moderate agreement with the mean kappa value of 0.51. A 3% increase in the kappa value, reflecting moderate agreement, was observed when transitioning from the F2V-TS to the F2V-FT approach. These results suggest that with the smaller training dataset available (config-1), no significant improvement in video-level scoring is observed while utilizing either of the two approaches (F2V-TS and F2V-FT).

To assess the impact of the amount of training data on overall model performances, both methods F2V-TS, and F2V-FT were evaluated on a relatively larger dataset (config-2) labeled by 3HOs. F2V-TS with the DCNN achieved the video-level performance (accuracy: 0.60), indicating a 3% decrease from config-1. In comparison mean kappa value among the 3HOs was 0.58, indicating a moderate agreement. The AI as the operator vs. 3HOs also showed a moderate agreement with the mean kappa value of 0.46, kappa value decreased. It is observed that in a scenario with a perfect agreement among the GT labels given by the HOs (config-1), F2V-TS performed well. Whereas, for a relatively larger training dataset (config-2), the results indicate that increasing the training data size may not necessarily enhance rather than negatively impact the model performance due to the subjectivity in the GT labels assigned by 3HOs with majority voting. In contrast, with the increasing training data (config-2), results show that F2V-FT is able to align the feature representation of LUS patterns from adults to neonates more effectively, resulting in the highest accuracy of 0.77 at the video-level. Furthermore, compared to F2V-TS, results show that the pre-trained model can potentially mitigate the variability in the GT provided by HOs. AI solution shows moderate agreement with 3HOs indicating the mean kappa value of 0.52. Thus, having relatively larger representative data available for training, F2V-FT is able to perform the best.

To bypass labeling at the frame-level and replicate the clinical practice of directly scoring at the video-level, the second strategy (DV) was utilized for direct video-level classification. The TranSLUCEnT classification model was used to classify the entire dataset (config-3), achieving an accuracy of 0.72. Compared to the Fleiss kappa value among the 3HOs was 0.65, indicating substantial agreement, the given AI solution with 3HOs showed a mean kappa value of 0.62, also indicating substantial agreement. These findings suggest that the TranSLUCEnT model demonstrated strong performance, while also reducing the variability in their interpretations, which was higher among the HOs. However, two key factors are also likely contributed to this improvement; the model was evaluated on a larger dataset, and the TranSLUCEnT model leverages domain knowledge from a pre-trained Resnet-18 model, originally trained on LUS data from COVID-19 adult patients [23]. The pre-trained Resnet-18 serves as a feature extractor, effectively transferring domain knowledge to provide a better representation of the LUS data from preterm neonates.

In conclusion, the choice of a suitable method from the above-mentioned strategies depends on factors such as dataset size, clinical preferences, and available resources. As clinical decision-making is generally based on the overall interpretation of the entire video rather than isolated frames. Therefore, Frame-to-video (F2V) scoring using the threshold-based aggregation technique is particularly useful in resource-constrained environments. In the scenario where we have a limited amount of frame-level labels for model training and aim to compute the video-level scores, the threshold-based aggregation technique is the best choice. As it is a clinically validated method utilized in state-of-the-art studies. For situations with limited frame-level labeling with no variability among HOs, training classification models from scratch (F2V-TS) appears to be a more suitable option. On the other hand, in cases where a relatively larger dataset is available and there is no absolute agreement among HOs, adapting deep learning models using the F2V-FT approach previously trained on a larger dataset proves a good choice. However, both F2V-TS and F2V-FT require a two-step process that depends on frame-level labels, which can be computationally intensive and prone to labeling errors. Alternatively, the DV strategy offers a more computationally efficient method by directly classifying videos, making it a better choice in some cases.

In contrast, the AI models developed in this study for prematurity-related lung disorders have the potential to be utilized in other domains beyond their original focus. These models have learned foundational features from neonatal LUS data, which are not limited to specific lung conditions. Because many lung disorders share common characteristics, the models can be adapted for use in diagnosing and managing other lung diseases. Additionally, these foundational features may be relevant across different age groups, such as older children or adults. To extend the use of these models to other domains whether it is other lung disorders or different age groups fine-tuning is essential. This involves retraining the model on new datasets that reflect the specific characteristics of the target population or condition. For example, if the model were to be applied to adult lung disorders, it would need to learn the differences in lung structure and disease patterns between neonates and adults. Fine-tuning allows the model to adjust its learned features to better match the new domain, making it more accurate and applicable to a broader range of clinical scenarios. By fine-tuning, the models become versatile tools that can perform well in various settings, making them useful not only for neonatal lung disorders but also for diagnosing and managing lung conditions in other populations.

7. Limitations and future work

A key limitation of this study is the small dataset size, especially in Config-1 and Config-2, where only a limited number of exams were labeled at the frame level for model training. Since no frame-level labels were available for testing, the remaining video-level exams were used as a substitute external dataset (termed as the test set) for evaluating the model's performance at the video level in both configurations. In contrast, Config-3 leverages all available exams. For this configuration, we employed LOOCV to train and validate the models. However, we lacked access to an external dataset for testing. As a result, the generalization performance of the models remains untested on unseen data. Addressing this limitation by acquiring a larger, more diverse dataset is a primary goal for future work.

Additionally, the training sets for Config-2 and Config-3 rely on majority voting, which introduces noise and variability, ultimately affecting the model's performance. Moreover, due to the limited availability of frame-level labeled data, we had to employ a larger test set for Config-1 and Config-2, which deviates from standard practice. This approach, while necessary, also restricts the generalizability of our findings. On the other hand, although the proposed AI solution reduces the inter-observer variability (IOV), we plan to hold calibration sessions to further improve inter-rater agreement among clinicians. In these sessions, all participating clinicians will review and discuss a subset of LUS images to align their interpretations before labeling the dataset. In addition, we aim to involve more expert neonatal clinicians with over 10 years of experience in LUS interpretation to enhance labeling consistency and reduce variability.

In future work, we plan to develop an interactive AI system based on federated learning, which will provide real-time feedback to raters. For instance, if a rater's label significantly deviates from the AI's prediction (based on prior training), the system could prompt the rater to reconsider or provide additional information, ultimately reducing IOV. We also plan to integrate gradient-weighted class activation map (Grad-CAM) and segmentation models for feature visualization, which could assist clinicians in making more informed decisions. Additionally, future work will explore the use of generative adversarial networks (GANs) to create synthetic images, addressing class imbalances and augmenting the dataset. These models will also facilitate automated labeling, improving consistency across a broader range of clinicians and reducing inter-observer variability.

CRediT authorship contribution statement

Noreen Fatima: Writing – review & editing, Writing – original draft, Supervision, Investigation, Formal analysis, Data curation. **Umair Khan:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Data curation. **Xi Han:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Data curation. **Emanuela Zannin:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation. **Camilla Rigotti:** Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation. **Federico Cattaneo:** Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation. **Giulia Dognini:** Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation. **Maria Luisa Ventura:** Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation. **Libertario Demi:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, 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

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.

References

[1] H. Blencowe, S. Cousens, M.Z. Oestergaard, D. Chou, A.-B. Moller, R. Narwal, A. Adler, C.V. Garcia, S. Rohde, L. Say, et al., National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications, Lancet 379 (9832) (2012) 2162–2172.

[2] R. Copetti, L. Cattarossi, The 'double lung point': an ultrasound sign diagnostic of transient tachypnea of the newborn, Neonatology 91 (3) (2007) 203–209.

[3] R. Copetti, L. Cattarossi, F. Macagno, M. Violino, R. Furlan, Lung ultrasound in respiratory distress syndrome: a useful tool for early diagnosis, Neonatology 94 (1) (2008) 52–59.

[4] E. Baraldi, M. Filippone, Chronic lung disease after premature birth, N. Engl. J. Med. 357 (19) (2007) 1946–1955.

[5] F. Raimondi, N. Yousef, F. Migliaro, L. Capasso, D. De Luca, Point-of-care lung ultrasound in neonatology: classification into descriptive and functional applications, Pediatr. Res. 90 (3) (2021) 524–531.

[6] Y. Singh, C. Tissot, M.V. Fraga, N. Yousef, R.G. Cortes, J. Lopez, J. Sanchez-de Toledo, J. Brierley, J.M. Colunga, D. Raffaj, et al., International evidence-based guidelines on Point of Care Ultrasound (POCUS) for critically ill neonates and children issued by the POCUS Working Group of the European Society of Paediatric and Neonatal Intensive Care (ESPNIC), Crit. Care 24 (2020) 1–16.

[7] L. Cattarossi, R. Copetti, G. Brusa, S. Pintaldi, Lung ultrasound diagnostic accuracy in neonatal pneumothorax, Can. Respir. J. 2016 (1) (2016) 6515069.

[8] F. Raimondi, J.R. Fanjul, S. Aversa, G. Chirico, N. Yousef, D. De Luca, I. Corsini, C. Dani, L. Grappone, L. Orfeo, et al., Lung ultrasound for diagnosing pneumothorax in the critically ill neonate, J. Pediatr. 175 (2016) 74–78.

[9] I. Corsini, N. Parri, E. Gozzini, C. Coviello, V. Leonardi, C. Poggi, M. Giacalone, T. Bianconi, L. Tofani, F. Raimondi, et al., Lung ultrasound for the differential diagnosis of respiratory distress in neonates, Neonatology 115 (1) (2019) 77–84.

[10] D.A. Blank, C.O.F. Kamlin, S.R. Rogerson, L.M. Fox, L. Lorenz, S.C. Kane, G.R. Polglase, S.B. Hooper, P.G. Davis, Lung ultrasound immediately after birth to describe normal neonatal transition: an observational study, Arch. Dis. Child.-Fetal Neonatal Ed. 103 (2) (2018) F157–F162.

[11] F. Raimondi, F. Migliaro, I. Corsini, F. Meneghin, P. Dolce, L. Pierri, A. Perri, S. Aversa, S. Nobile, S. Lama, et al., Lung ultrasound score progress in neonatal respiratory distress syndrome, Pediatrics 147 (4) (2021).

[12] L. Capasso, D. Pacella, F. Migliaro, S. Salomè, F. Grasso, I. Corsini, D. De Luca, P.G. Davis, F. Raimondi, Can lung ultrasound score accurately predict surfactant replacement? A systematic review and meta-analysis of diagnostic test studies, Pediatr. Pulmonol. 58 (5) (2023) 1427–1437.

[13] D. De Luca, L. Bonadies, A. Alonso-Ojembarrena, D. Martino, I. Gutierrez-Rosa, B. Loi, R. Dasani, L. Capasso, E. Baraldi, A. Davis, et al., Quantitative lung ultrasonography to guide surfactant therapy in neonates born late preterm and later, JAMA Netw. Open 7 (5) (2024) e2413446.

[14] D. De Luca, C. Autilio, L. Pezza, S. Shankar-Aguilera, D.G. Tingay, V.P. Carnielli, Personalized medicine for the management of RDS in preterm neonates, Neonatology 118 (2) (2021) 127–138.

[15] F. Raimondi, F. Migliaro, A. Sodano, A. Umbaldo, A. Romano, G. Vallone, L. Capasso, Can neonatal lung ultrasound monitor fluid clearance and predict the need of respiratory support? Crit. Care 16 (2012) 1–5.

[16] F. Raimondi, F. Migliaro, A. Sodano, T. Ferrara, S. Lama, G. Vallone, L. Capasso, Use of neonatal chest ultrasound to predict noninvasive ventilation failure, Pediatrics 134 (4) (2014) e1089–e1094.

[17] L. Pezza, A. Alonso-Ojembarrena, Y. Elsayed, N. Yousef, L. Vedovelli, F. Raimondi, D. De Luca, Meta-analysis of lung ultrasound scores for early prediction of bronchopulmonary dysplasia, Ann. Am. Thorac. Soc. 19 (4) (2022) 659–667.

[18] F. Mento, U. Khan, F. Faita, A. Smargiassi, R. Inchingolo, T. Perrone, L. Demi, State of the art in lung ultrasound, shifting from qualitative to quantitative analyses, Ultrasound Med. Biol. 48 (12) (2022) 2398–2416.

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

[20] R. Chioma, A. Sbordone, M.L. Patti, A. Perri, G. Vento, S. Nobile, Applications of artificial intelligence in neonatology, Appl. Sci. 13 (5) (2023) 3211.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] R. Bassiouny, A. Mohamed, K. Umapathy, N. Khan, An interpretable neonatal lung ultrasound feature extraction and lung sliding detection system using object detectors, IEEE J. Transl. Eng. Health Med. (2023).

[27] S. Aujla, A. Mohamed, N. Khan, K. Umapathy, Multi-level classification of lung pathologies in neonates using recurrence features, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2022, pp. 1531–1535.

[28] S. Aujla, A. Mohamed, R. Tan, K. Magtibay, R. Tan, L. Gao, N. Khan, K. Umapathy, Classification of lung pathologies in neonates using dual-tree complex wavelet transform, BioMed. Eng. OnLine 22 (1) (2023) 115.

[29] M. Gravina, D. Gragnaniello, L. Verdoliva, G. Poggi, I. Corsini, C. Dani, F. Meneghin, G. Lista, S. Aversa, F. Raimondi, et al., Deep learning in the ultrasound evaluation of neonatal respiratory status, in: 2020 25th International Conference on Pattern Recognition, ICPR, IEEE, 2021, pp. 10493–10499.

[30] Y. Wu, S. Zhao, X. Yang, C. Yang, Z. Shi, Q. Liu, Y. Wang, M. Qin, L. Zhang, [Retracted] ultrasound lung image under artificial intelligence algorithm in diagnosis of neonatal respiratory distress syndrome, Comput. Math. Methods Med. 2022 (1) (2022) 1817341.

[31] J. Jiao, Y. Du, X. Li, Y. Guo, Y. Ren, Y. Wang, Prenatal prediction of neonatal respiratory morbidity: a radiomics method based on imbalanced few-shot fetal lung ultrasound images, BMC Med. Imaging 22 (1) (2022) 2.

[32] A. Perez-Moreno, M. Dominguez, F. Migliorelli, E. Gratacos, M. Palacio, E. Bonet-Carne, Clinical feasibility of quantitative ultrasound texture analysis: a robustness study using fetal lung ultrasound images, J. Ultrasound Med. 38 (6) (2019) 1459–1476.

[33] Y. Du, Z. Fang, J. Jiao, G. Xi, C. Zhu, Y. Ren, Y. Guo, Y. Wang, Application of ultrasound-based radiomics technology in fetal-lung-texture analysis in pregnancies complicated by gestational diabetes and/or pre-eclampsia, Ultrasound Obstet. Gynecol. 57 (5) (2021) 804–812.

[34] C. Gomond-Le Goff, L. Vivalda, S. Foligno, B. Loi, N. Yousef, D. De Luca, Effect of different probes and expertise on the interpretation reliability of point-of-care lung ultrasound, Chest 157 (4) (2020) 924–931.

[35] N. Fatima, F. Mento, A. Zanforlin, A. Smargiassi, E. Torri, T. Perrone, L. Demi, Human-to-Ai interrater agreement for lung ultrasound scoring in COVID-19 Patients, J. Ultrasound Med. 42 (4) (2023) 843–851.

[36] 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.

[37] 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.

[38] N. Savage, Breaking into the black box of artificial intelligence, 2022.

[39] R. Chauhan, K.K. Ghanshala, R. Joshi, Convolutional neural network (CNN) for image detection and recognition, in: 2018 First International Conference on Secure Cyber Computing and Communication, ICSCCC, IEEE, 2018, pp. 278–282.

[40] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., A survey on vision transformer, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2022) 87–110.

[41] U. Khan, S. Afrakhteh, F. Mento, G. Mert, A. Smargiassi, R. Inchingolo, F. Tursi, V.N. Macioce, T. Perrone, G. Iacca, et al., Low-complexity lung ultrasound video scoring by means of intensity projection-based video compression, Comput. Biol. Med. 169 (2024) 107885.

[42] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[43] U. Khan, N. Fatima, X. Han, C. Rigotti, F. Cattaneo, G. Dognini, M.L. Ventura, E. Zannin, G. Iacca, L. Demi, TranSLUCEnt: Transferred sequential lung ultrasound characteristic encodings-based transformer for lung ultrasound pattern classification in premature neonates, in: 2024 Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS) (Accepted), IEEE, 2024.

[44] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (5) (1971) 378.

[45] 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.