IEEE Journal of Translational Engineering in Health and Medicine

Received 17 May 2023; revised 11 October 2023; accepted 20 October 2023.
Date of publication 25 October 2023; date of current version 24 November 2023.

*Digital Object Identifier 10.1109/JTEHM.2023.3327424*

# An Interpretable Neonatal Lung Ultrasound Feature Extraction and Lung Sliding Detection System Using Object Detectors

**RODINA BASSIOUNY[1], ADEL MOHAMED[2], KARTHI UMAPATHY[1], (Senior Member, IEEE), AND NAIMUL KHAN[1], (Senior Member, IEEE)**

[1]Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada
[2]Mount Sinai Hospital, University of Toronto, Toronto, ON M5S 1A1, Canada

CORRESPONDING AUTHOR: N. KHAN (n77khan@ryerson.ca)

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through an Alliance Grant.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Mount Sinai Hospital Research Ethics Board under Approval No. MSH REB 19-0065-E.

**ABSTRACT** The objective of this study was to develop an interpretable system that could detect specific lung features in neonates. A challenging aspect of this work was that normal lungs showed the same visual features (as that of Pneumothorax (PTX)). M-mode is typically necessary to differentiate between the two cases, but its generation in clinics is time-consuming and requires expertise for interpretation, which remains limited. Therefore, our system automates M-mode generation by extracting Regions of Interest (ROIs) without human in the loop. Object detection models such as faster Region Based Convolutional Neural Network (fRCNN) and RetinaNet models were employed to detect seven common Lung Ultrasound (LUS) features. fRCNN predictions were then stored and further used to generate M-modes. Beyond static feature extraction, we used a Hough transform based statistical method to detect ''lung sliding'' in these M-modes. Results showed that fRCNN achieved a greater mean Average Precision (mAP) of **86.57**% (Intersection-over-Union (IoU) = 0.2) than RetinaNet, which only displayed a mAP of 61.15%. The calculated accuracy for the generated RoIs was **97.59**% for Normal videos and **96.37**% for PTX videos. Using this system, we successfully classified **5** PTX and **6** Normal video cases with **100**% accuracy. Automating the process of detecting seven prominent LUS features addresses the time-consuming manual evaluation of Lung ultrasound in a fast paced environment. Clinical impact: Our research work provides a significant clinical impact as it provides a more accurate and efficient method for diagnosing lung diseases in neonates.

**INDEX TERMS** Lung ultrasound, object detection models, *faster* RCNN, RetinaNet, Hough transform, M-mode, automatic lung sliding detection.

## I. INTRODUCTION

Preterms are babies whose delivery is prior to 37 weeks' gestation [1]. According to WHO, approximately 2.4 million preterm fatalities were reported globally in 2019 [2]. Therefore, early identification of these lung disorders can save lives. Lung Ultrasound (LUS) has been lately used by healthcare providers instead of Chest X-Rays (CXRs) and Computed Tomography (CT)-chest in both neonatal and adult population in different clinical settings [3]. This is due to several significant advantages such as real-time imaging, non-ionizing radiation, lower equipment costs, portability, and bedside availability [4]. Manual evaluation of LUS scans necessitates sonography skills, trained operators, and a significant amount of time. Each LUS scan reveals certain patterns. Pediatricians compiled a list of seven LUS features that are associated with most common neonatal respiratory disorders. Due to the restricted number of expert neonatologists in lung ultrasound, accurate manual evaluation of lung images

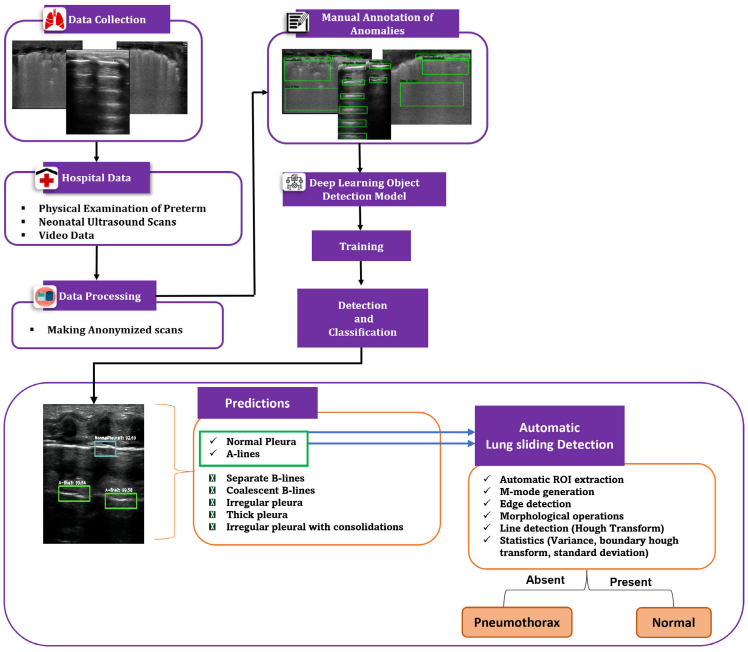


**FIGURE 1.** Pipeline of our proposed LUS feature detection system.

remains a challenge. Therefore, we propose an Artificial Intelligence approach to detect these LUS features with high accuracy.

The main contributions of this research work are listed below[1]:

- **Annotated LUS feature dataset:** Physicians indicated interest in a system capable of extracting lung features from LUS videos. Therefore, we annotated 17,491 LUS features under the supervision of a neonatology expert using a labeling tool called Dark Label [6]
- **Object Detection Models for LUS:** A single seven class fRCNN [7] was trained on the set of LUS features, and a RetinaNet [8] was implemented for comparison. They were evaluated based on their recognition speed and mAP.
- **Automatic Motion Mode (M-mode) Extraction:** M-Mode images are required to differentiate between normal lung and PTX. Expert neonatologists are required for M-mode interpretation. Allocation of cursor around the ROI is done manually by the physician which again requires expertise. Therefore, we proposed a system that can automatically generate these ROIs.
- **Automatic Lung Sliding Detection:** Further to the extracted M-Mode, we created an automated process to detect lung sliding, which is the differentiating feature between normal and PTX lungs. Lung sliding cannot be detected from images alone. We were successful in detecting this feature using Hough Transform and Morphological operations. Variance and the boundary Hough transform were further deduced to differentiate between PTX and Normal M-mode scans.

[1]A preliminary version of this work was published in [5].

The full workflow of our proposed LUS feature detection can be seen in Fig. 1.

## II. RELATED WORKS

Deep learning has shown popularity for assessing medical images and making fast decisions. It has replaced the need for hand-crafted features, due to its efficiency to generate both low-level and high-level features from data [9]. According to the most recent review of the main contributions in LUS image analysis [10], related works can be divided into two major classes: i) Model Based Techniques, ii) Deep Learning Models.

### A. MODEL BASED TECHNIQUES

Earlier, LUS scans were graded based on the severity of aeration loss. Authors in [11] assessed ExtraVascular Lung Water (EVLW) through quantitative LUS and CT intensities. LUS intensity was more accurate as it demonstrated the air to water ratio which is an excellent interpretation of the amount of echogenic interfaces. This is critical as this measurement is a powerful tool in early diagnosis of Pneumonia [12] which is a chronic pulmonary disorder. Brusasco et al. [13] developed a completely automated and quantitative scoring system to detect B-lines in Acute Respiratory Distress Syndrome (ARDS) patients.

### B. DEEP LEARNING NEURAL NETWORKS

Gravina et al. [14], in 2021, employed a deep learning approach in classifying ARDS, Transient Tachypnea (TTN) and healthy ultrasound scans. Five Neural Networks (NN) were deployed respectively. Awasthi et al. [15] proposed a mobile-friendly, yet efficient deep learning model for detecting COVID-19. Their constructed network "Mini-COVIDNet" surpassed other light weight neural network

IEEE Journal of Translational Engineering in Health and Medicine

models as well as a state-of-the-art heavy model. LUS scans were classified as COVID-19, Pneumonia, or a healthy condition. Apostolopoulos et al. [16], on the other hand, used a VGG-19 and a MobileNet-v2 to classify normal, COVID-19 Pneumonia, viral, and bacterial Pneumonia. They utilized transfer learning during their training process. Magrelli et al. [17] utilizes Gradient weighted class activation mapping (Grad-CAM) to add interpretability to Image classification models. VGG-19, Xception, Inception-v3, and Inception-ResNet-v2 were trained. This approach did not perform well on B-lines and could only detect 50% of A-lines. In our scenario, this approach was not sufficient to detect other LUS features. If we used those feature maps, as described in paper [17], the colours will overlap, hence, making the interpretation more difficult to read. Physicians require detections to be precise. As a result, we used object detection models to deliver precise detections of LUS features, allowing clinicians to spend less time making a diagnosis. Kulhare et al. [18] used Single-Shot Detectors (SSDs) for detecting LUS features in swine dataset. Another worth-mentioning work done by Xing et al. [19] 2022, they proposed a novel A-line detection method for LUS images for both linear and convex probes. An fRCNN model was implemented and enhanced with localization box selection strategy to accurately allocate pleural lines.

## III. LUS DATA MINING AND LABELING

### A. LUS IMAGING PROTOCOLS

All our LUS scans were performed using a portable Ultrasound machine **(Z.One PRO Ultrasound System, Mindray North America, CA, USA)**, a linear probe **(L20-5 MHz)** with a high-resolution to enhance image quality. Our spatial resolution was 20-30 millimetres. Depth was adjusted at 3 cm to ensure adequate penetration and to aid with image acquisition. Lungs were scanned in six zones, 3 zones on either side (Upper Anterior R1, L1, Lower Anterior R2, L2, and Lateral R3, L3 bilaterally). Neonates were scanned in supine position. Footprint for the linear probe was set to 2.5 cm to correlate with beam width. Probe was positioned vertical (perpendicular) on the chest wall with little to no angulation. All scans were saved as a video clip of approximately 6 seconds.

### B. LUS FEATURES

Physicians determined a list of seven key LUS features that are associated to respiratory diseases. These features are; normal pleural line, thick pleural, irregular plural, A-lines, separate B-lines, coalescent B-lines, and consolidations.

Normal pleural is one of the brightest (echogenic) feature in LUS scan. Regarding pleural line thickness, we are unable to provide a tight cut-off. But according to a previous study [20], a pleural line $>$ 2 mm is considered a thick pleural line. A-lines are repetitive reverberation artifacts that mimic the pleural line and are equally distant [21]. B-lines are discrete laser-like vertical hyperechoic reverberation artifacts that arise from the pleural line, extend to the bottom of the screen without fading, and move synchronously with lung sliding. Vertical hyperechoic reverberation are referred as Separate B-lines if they are $\geq$ 3 per frame. Otherwise, they are Coalescent B-lines. Consolidations are regarded as a critical reduction of aeration and appear as a hypo-echoic tissue. More details about the LUS features can be found in our preliminary work [5]. All LUS pathologies are displayed in Fig. 2.

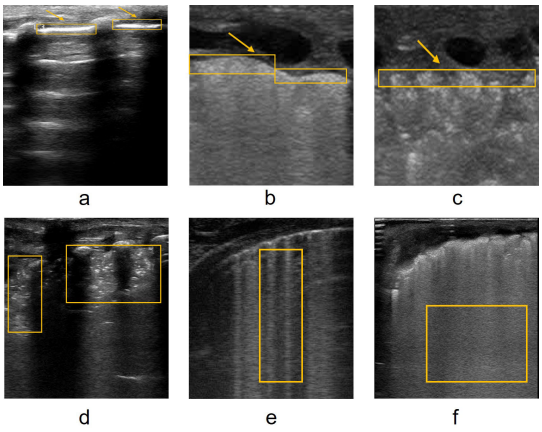


**FIGURE 2. LUS pathologies: (a) Normal pleural line (b) Thick pleural line (c) Irregular pleural line (d) Consolidations (e) Separate B-lines (f) Coalescent B-lines.**

Neonatologists outlined five common lung conditions. These include ARDS [22], TTN [23] of the newborn, Chronic Lung Disease (CLD) [24], PTX [25], Consolidation and normal lungs. These lung conditions are the most common respiratory causes for NICU admissions. They were selected manually from each frame and annotated under the supervision of a neonatologist (second author) who is an expert in lung ultrasound. Presence/absence of the seven lung features described before is associated to different lung conditions. Therefore, in this work we focus mainly on detecting the LUS features.

### C. DATA MINING

LUS scans were performed on a total of 45 neonates.[2] There were 6 Normal, 6 CLD, 7 Consolidation, 5 PTX, 7 ARDS, and 14 TTN neonatal case studies. Briefly, a longitudinal scan of the anterior and lateral chest walls was performed. A total of six zone areas of the chest were scanned. Ultrasonographic images were anonymized and transferred to a password protected external hard drive.

The exam was performed longitudinally using a linear probe, and at least one 6-second ultrasound video was captured from each zone at a frame rate of 13.7 fps. A total of 324 videos and 26,632 frames were collected from 45 patients. Each patient had a range of 1-10 videos. The full LUS dataset is shown in Table 1.

## IV. METHODOLOGY

We opted for detecting the LUS features from images rather than developing an image classification system. Only

[2]Study approved by Mount Sinai Hospital and Ryerson Research Ethics Board. Our REB study approval number is: MSH REB 19-0065-E.

**TABLE 1. Mount sinai lung ultrasound dataset.**

| Lung Condition | Cases/Patients | No. of videos | Annotated frames |
|---|---|---|---|
| *PTX* | 5 | 16 | 1279 |
| *Normal lungs* | 6 | 31 | 85 |
| *Consolidation* | 7 | 61 | 3096 |
| *CLD* | 6 | 31 | 959 |
| *ARDS* | 7 | 59 | 1690 |
| *TTN* | 14 | 126 | 4920 |

computer vision techniques such as object detection-based methods can meet this purpose. Object-based detection models are classified into two groups: One stage and two-stage detectors. We decided to choose one model from each group: RetinaNet [8] (a one-stage detector model) and the two-stage detector *faster* RCNN [7].

RetinaNet has a high recognition speed, and therefore, can work as a real-time model. It has demonstrated great detection performance over aerial/satellite images. This is due to its Feature Pyramid Network (FPN) which is designed to produce rich semantics at all levels. It is, however, not always the best when it comes to object localization. fRCNN, on the other hand, may be slower in recognition speed, but provides a higher localization performance, thus, better precision of its bounding boxes. It is due to the second stage unit that filters out bad proposals and fine-tunes positive proposals. Use of a pre-trained Image-Net model as a backbone network is fundamental to get rich semantic features. We opted for VGG-16 and ResNet-18 as backbone models for fRCNN and RetinaNet, respectively, as per the source papers [7], [8].

### A. LUNG SLIDING DETECTION USING AUTOMATIC CURSOR ALLOCATION ON EXTRACTED M-MODE IMAGES

*Why M-mode?* PTX and Normal lungs share the same visual LUS features such as A-lines. Detection of these features by object detection models is not sufficient to differentiate between these cases. According to physicians, the only distinguishing sign would be ''Lung sliding''. This is something not easily spotted by the physician on the LUS scan, especially if the baby is working hard to breath and the whole chest is moving. In order to accurately spot this feature, the scans had to be switched from 2-D mode to M-mode. M-mode shows the velocity of a certain organ along a specific line (motion cursor) in relation to the time axis.

PTX patients are typically classified as emergency situations that require prompt medical attention/intervention to save the patients' lives. There are not enough physicians in each NICU who know how to interpret these patients' LUS findings and distinguish between normal and PTX images confidently. Therefore, our aim was to help in making a quick differentiation between normal and PTX cases by first proposing a solution to replace the manual selection of the region of interest and to generate M-modes with excellent efficiency. Our second goal was to test the absence or presence of the lung sliding. It has been observed that PTX M-mode indicates the absence of lung sliding as a stratosphere pattern as opposed to Normal scans that show a seashore pattern. Therefore, we selected line detection techniques (Hough Transform) and other morphological operations to make a distinction between PTX and normal scans.

### B. AUTOMATIC EXTRACTION OF LUNG SLIDING FEATURE

For placing the line of motion and extracting a good ROI, the following requirements must be met: i) Line of motion should be set within the pleural line ii) Line should not include rib shadows iii) A slice of 15-25 pixel width to include enough information from the sub-pleural region. In order to exclude the rib shadow from the ROI, we made sure to place the cursor one-third way into the pleural line. Based on experiments made over the coral software used by Mount Sinai hospital, we picked an average ROI width of 20 pixels. It was enough to display a good enough M-mode image that can be further utilized to differentiate between PTX and normal case. To slice the ROIs automatically and without human in the loop we stored the fRCNN predictions imposed over the LUS video frames. We specifically used the detected normal pleurae as potential indicator to our ROI slice. Two cursors were defined for the ROI slice, $C_1$ and $C_2$. If the bounding box of the detected normal pleurae in a frame $f$, is $x_1$, $y_1$, $x_2$, $y_2$, then $C_1$ is set to one third way into the box and $C_2$ is set to 20 pixels away from $C_1$. This is seen in equations 1 and 2.

$$C_1 : x_1 + \frac{1}{3}(x_2 - x_1) \tag{1}$$

$$C_2 : C_1 + 20 \tag{2}$$

### C. GENERATION OF M-MODE IMAGES USING THE EXTRACTED ROIS

To generate M-mode images close to the reference M-mode images generated by the hospital, extracted ROIs had to be accurate. They were first evaluated before extracting the M-mode images. We were successful in achieving a high level of precision for the ROIs and hence, we concatenated them horizontally side-by-side to form a single M-mode image. This can be observed in Fig. 3.

We see that in a PTX clip with a run time of 6 seconds and a frame rate of 14.2 fps, 85 ROIs were combined to create one M-mode with a width of **w** (85*20 cursor width = 1700) and a height of **h**. Fig. 4 shows a generated PTX M-mode image as well as a normal M-mode image. Both constructed M-mode images look very close to the true reference M-mode image provided by the hospital.

### D. LINE DETECTION IN M-MODES USING HOUGH TRANSFORM

Through observation, PTX M-mode exhibits more lines because to its stratospheric pattern nature, whereas a Normal scan exhibits fewer lines due to its sea-shore pattern. Therefore, we considered detection of the horizontal lines and counting them to differentiate between Normal and PTX scans. For this purpose, Hough transform was utilized. Pre-processing of image included simple thresholding

IEEE Journal of Translational Engineering in Health and Medicine

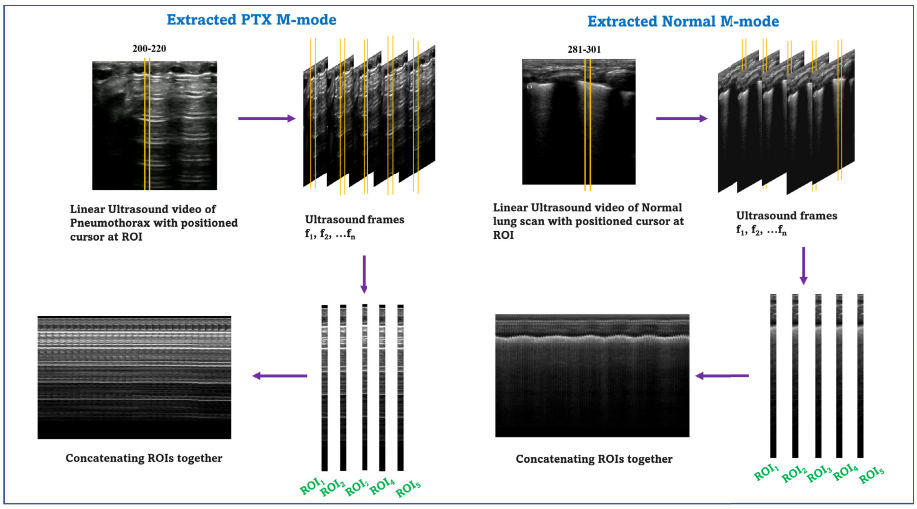


**FIGURE 3.** An example of extraction of M-mode from both PTX and Normal video. The figure shows how the slicing of ROIs in all frames is done and concatenating them to create an extracted M-mode image.

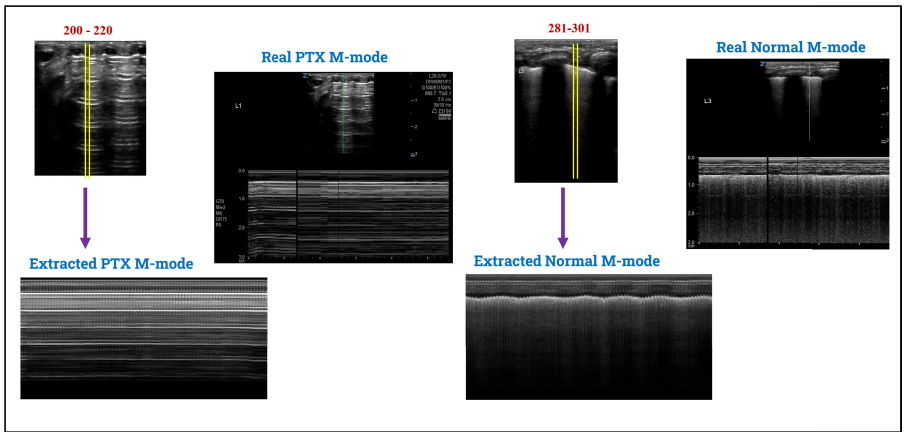


**FIGURE 4.** Extracted M-mode images from both PTX and normal LUS videos vs the acquired reference real M-mode image.

and morphological gradient which is the difference between Erosion and Dilation. This effect highly enriched the LUS features (A-lines and pleural line) which are vital in determining the lung sliding feature.

Next, the edge image is converted into an (r-$\theta$) plane, where r is the perpendicular distance in pixels from the line to the origin and $\theta$ is the angle between the line stretched from the origin and the line passing through the edge point. The complete edge image is represented by a sequence of sinusoidal curves, the largest of which has the same amplitude as the images diagonal length (-$d_{max}$ - $d_{max}$) and angles between ($-90°$, $90°$). Each line is expressed in terms of equations 3 and 4.

$$y = \left(\frac{-\cos\theta}{\sin\theta}\right)x + \left(\frac{r}{\sin\theta}\right) \tag{3}$$

$$r = x\cos\theta + y\sin\theta \tag{4}$$

All edge points that contribute to the formation of a line are called line candidates. They are stored in a matrix by "Hough Accumulator". Local maxima is deduced from the accumulator. It denotes the most prominent lines in the accumulator, which are the highest votes. Multiple thin horizontal lines can appear very close to each other due to the nature of ultrasound imaging, while they actually represent one single line morphologically. Therefore, a "fill gap" parameter value of 20 pixels was introduced to fill the gap between lines that appear close to each other, so that they are combined into a single line. Only lines bigger than 90% of the image's width were considered.

The full solution of lung sliding detection to distinguish PTX from Normal LUS scan is illustrated in Fig. 5.

Here, the trained fRCNN performs detections on a PTX study case. The values of the predicted pleural boundary box are saved in a csv file. If the box values were [222, 110, 314, 156], the computed cursor locations would be 253 and 273 according to equations 1 and 2. The accuracy of the automatically assigned line of motion vs. the manual ones is reported in the Results section. Only lines that were at least 7 pixels vertically apart were taken into consideration. Three probable line candidates were recognised and imposed across

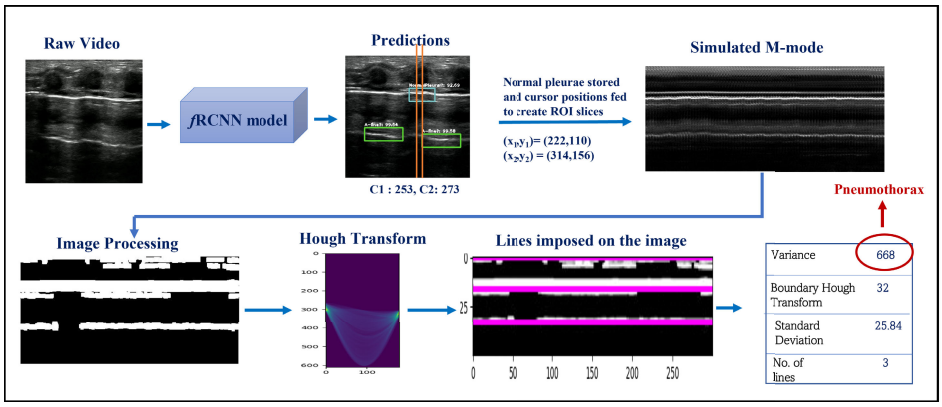


**FIGURE 5.** Classification of PTX using proposed automatic lung sliding feature detection and statistical measurements.

the processed image in the PTX case as shown in Fig. 5. These detected lines correlate to the subcutaneous tissue, pleural line, and one A-line. These lines effectively describe what can be seen in the input raw video. On the other hand, a normal case study exhibits two lines (one for the skin lesion and the other for the pleural line).

After experiment, it was impossible to find a right threshold for the number of lines. We noticed in most of the views the probe position was not guaranteed to be fixed in a specific chest area, and its movement altered the reading. Therefore, Line detection was not sufficient in classifying PTX from normal. As a result, we had to deduce other statistical information from the M-modes to provide accurate results.

### E. DIFFERENTIATING BETWEEN EXTRACTED PTX AND NORMAL M-MODE SCANS USING STATISTICS

Since line detection was not enough in differentiating PTX from Normal case, we decided to calculate the sample variance for processed M-mode images. This variance measures the variability of the pixel intensity values from the mean value. Almost all PTX processed videos displayed a relatively higher variance value than Normal scans. A total of 6 normal cases (31 videos) and 5 PTX (16 videos) cases were tested. 29/31 normal videos demonstrated a variance $\leq 500$, but the variance in the other two videos was fairly substantial as they did not show clear LUS features. However, the averaged variance for each Normal case stayed $\leq 500$. On the other hand, 14/16 PTX videos displayed a variance $\geq 500$. The averaged variance for each PTX case stayed $\geq 500$.

## V. RESULTS

### A. PARAMETERS FOR FRCNN

#### 1) ANCHOR RATIOS

The aspect ratio of the object classes in the training dataset can be preserved using anchors. Choosing these anchor ratios depend on the varied sizes of different object classes. Each LUS pattern (class) in our dataset displayed different shapes throughout the videos. Several anchor ratios were hence added as the sizes of the LUS features varied from one patient to another depending on the age. Not only this, but the angle with which the neonates chest is being scanned also affects the size. As stated in Table 2, a total of 28 anchor ratios were specified.

**TABLE 2.** LUS features with their respective anchor ratios.

| LUS features | Anchor Ratios |
|---|---|
| *A-lines* | [46,141], [33,125] , [44,134], [34,131] |
| *Coalescent B-lines* | [190,155],[235,144], [354,153] |
| *Separate B-lines* | [370,46], [320,51], [364,50], [372,48] |
| *Irregular pleural* | [109,54], [128,43], [85,41] |
| *Thick pleural* | [124,39], [118,37], [86,39] |
| *Normal pleural line* | [27,127], [22,108], [26,125], [24,113] |
| *Irregular pleural with consolidations* | [268,216], [316,144], [220,171], [208,205], [337,243], [337,243] [208,205] |

### B. OTHER PARAMETERS

32, 64, 128 and 256 were selected as anchor scales. A total of 112 anchors were introduced per spatial location. The model was trained to learn the offset difference between the predicted and ground truth bounding boxes. Smooth L1 was used for the regression loss. Early stopping was employed as a regulation technique for the neural network to avoid overfitting. As a result, the model was trained for 100 epochs with a step size of 100. The learning rate was set to $1 \times 10^{-5}$ and the stochastic gradient (SGD) momentum was set to 0.9. A 12GB NVIDIA Tesla K80 GPU was employed to ramp up the training of the object detection model. The models were implemented in the Keras framework, with a TensorFlow backend. To speed up the learning process, a pre-trained VGG-16 was used as the backbone of fRCNN. Non-maximum suppression was set to 0.15 to prevent overlapping between predicted boxes over one LUS feature. During testing, predicted bounding boxes with an overlapping bounding box threshold of 85% were accepted.

IEEE Journal of Translational Engineering in Health and Medicine

### C. EVALUATION: MAP

mAP is used to evaluate object detection models. Probability score is assigned to every predicted bounding box. Only boxes with a probability score and IoU greater than the set threshold are considered. IoU threshold values were set as [0.2, 0.3, and 0.4].

We noticed that the pleural line appeared as multiple pleurae separated by rib shadows. Whereas anterior views displayed a long continuous line. As an object detection model, it detects various parts of the continuous line correctly as a pleural line. But after non-maximum suppression overlapping boxes are discarded. The boxes remaining when matched with the ground truth box still provides us with a lower IoU (between 0.2 and 0.4) compared to IoU thresholds traditionally accepted by natural object detection models (typically 0.5 and above). According to physicians, the predictions are correct as long they capture the line itself. Since the objective is not to quantify the size of the line, rather than the presence/absence of it, IoU threshold was set lower (0.2) to accept these bounding boxes.

Considering two bounding boxes, C1 and C2, IoU will be calculated as in equation 5.

$$IoU = \frac{C1 \cap C2}{C1 + C2 - C1 \cap C2} \tag{5}$$

Recall and precision are calculated as in the following equation 6.

$$R = \frac{TP}{TP + FN}, P = \frac{TP}{TP + FP} \tag{6}$$

where FN, TP and FP are the false negatives, true positives, and false positives respectively. The model is trained to keep the offset difference between the predicted box and the ground truth box to a minimum. The average precision of the model is the area under the precision recall curve. Average precision is calculated per LUS frame, and is then averaged over all frames to get the mAP per LUS feature. Finally, the model's mAP was calculated by averaging the mAP of all LUS features.

### D. AUGMENTATION AND PRE-TRAINING

Deep learning models like fRCNN require a massive amount of labelled training data. Ultrasound images are limited, and manual annotation takes a long time. As a result, horizontal flipping and rotations of +5 degrees and +10 degrees were used to augment the data. Training the VGG-16 backbone of fRCNN from scratch produced worse results than using its pre-trained model already on natural images. We only report results from the pre-trained network for brevity.

### E. EFFECT OF HYPERPARAMETERS

To achieve the best performance, few parameters were tuned, including the region proposal network (rpn) overlapping threshold and the classifier maximum threshold. The rpn network proposes regions by scattering anchors of various ratios and scales across feature maps. As shown in Table 3, the overlapping threshold between the anchors and the ROI was set to 0.7, 0.6, and 0.5, and the trained models were named as $M_1$, $M_2$, and $M_3$. The models were evaluated by calculating mAP for various IoU values of 0.2, 0.3, and 0.4. It was discovered that the higher the threshold for the rpn network were kept, the more accurate the detections were. The model with the highest rpn overlapping threshold had the highest mAP of 86.6%, 84.3%, and 79.9% for IoUs of [0.2, 0.3, 0.4] respectively.

**TABLE 3. Tuning the rpn overlapping threshold for fRCNN.**

| Model | Training parameters | Anomalies | mAP IoU 0.4 | mAP IoU 0.3 | mAP IoU 0.2 |
|---|---|---|---|---|---|
| $M_1$ | rpn-overlap threshold = 0.7 classifier-max threshold = 0.5 | A-lines | 87.1 | 89.0 | 93.0 |
| | | Coalescent B-lines | 93.2 | 98.0 | 99.4 |
| | | Separate B-lines | 65.0 | 65.0 | 66.0 |
| | | Thick pleural | 68.0 | 68.0 | 68.5 |
| | | Irregular pleural with consolidations | 91.4 | 95.8 | 96.0 |
| | | Irregular pleural | 72.9 | 81.0 | 93.0 |
| | | Normal pleural | 81.6 | 87.0 | 90.1 |
| | | Total mAP | 79.9% | 83.4% | 86.6% |
| $M_2$ | rpn overlap threshold = 0.6 classifier-max threshold = 0.5 | A-lines | 80.1 | 88.0 | 89.7 |
| | | Coalescent B-lines | 90.2 | 98.0 | 98.0 |
| | | Separate B-lines | 63.7 | 65.3 | 65.4 |
| | | Thick pleural | 71.1 | 80.2 | 82.3 |
| | | Irregular pleural with consolidations | 85.6 | 92.7 | 92.1 |
| | | Irregular pleural | 61.4 | 70.2 | 73.3 |
| | | Normal pleural | 80.2 | 85.7 | 88.2 |
| | | Total mAP | 76.0% | 82.9% | 84.1% |
| $M_3$ | rpn overlap threshold = 0.5 classifier-max threshold = 0.5 | A-lines | 78.0 | 86.4 | 88.7 |
| | | Coalescent B-lines | 86.7 | 98.0 | 98.0 |
| | | Separate B-lines | 62.6 | 65.4 | 66.1 |
| | | Thick pleural | 81.8 | 87.2 | 87.6 |
| | | Irregular pleural with consolidations | 81.3 | 88.4 | 88.2 |
| | | Irregular pleural | 51.4 | 62.9 | 64.3 |
| | | Normal pleural | 74.6 | 80.3 | 83.8 |
| | | Total mAP | 73.8% | 81.2% | 82.4% |

According to Table 4, $M_3$ takes approximately 3 hours to train, which is roughly double the time the first model $M_1$ takes. This is due to the additional regions proposed with at least 50% overlap with the ground truth object. This takes longer during the detection phase, as observed by the slowest model $M_3$, which takes 16.11s vs. 8.75s for the fastest model $M_1$.

Another hyperparameter for fRCNN is the classifier max threshold, which is calculated after the rpn proposes the ROIs. To detect the presence of an object, proposed regions are passed through a classifier. As shown in Table 5, the classifier max threshold is set to 0.5 and 0.4, and the trained models were named $M_4$ and $M_5$, respectively.

The model with the higher classifier max threshold achieves 86.6%, 83.4%, and 79.9% for IoUs of 0.2, 0.3, and 0.4, respectively, whereas the second model achieves 73.3%, 79.5%, and 81.4% for the same IoUs.

**TABLE 4.** Effect of the tuning parameter rpn-overlap threshold on the training time and recognition speed.

| Tuned Models | Training time (hour) | Inference time per frame (s) |
|---|---|---|
| $M_1$ *Rpn-overlap-threshold = 0.7* | **1.72h** | **8.75s** |
| $M_2$ *Rpn-overlap-threshold = 0.6* | 2.75h | 14.89s |
| $M_3$ *Rpn-overlap-threshold = 0.5* | 2.98h | 16.11s |

**TABLE 5.** Tuning the classifier max threshold for fRCNN.

| Model | Training parameters | Anomalies | IoU 0.4 | mAP IoU 0.3 | IoU 0.2 |
|---|---|---|---|---|---|
| $M_4$ | classifier -max threshold =0.5 | A-lines | 87.1 | 89.0 | 93.0 |
| | | Coalescent B-lines | 93.2 | 98.0 | 99.4 |
| | | Separate B-lines | 65.0 | 65.0 | 66.0 |
| | | Thick pleural | 68.0 | 68.0 | 68.5 |
| | | Irregular pleural with consolidations | 91.4 | 95.8 | 96.0 |
| | | Irregular pleural | 72.9 | 81.0 | 93.0 |
| | | Normal pleural | 81.6 | 87.0 | 90.1 |
| | | Total mAP | 79.9% | 83.4% | 86.6% |
| $M_5$ | classifier -max threshold =0.4 | A-lines | 86.4 | 89.8 | 92.9 |
| | | Coalescent B-lines | 87.8 | 96.5 | 98.0 |
| | | Separate B-lines | 54.0 | 58.0 | 58.8 |
| | | Thick pleural | 64.2 | 66.9 | 68.2 |
| | | Irregular pleural with consolidations | 86.5 | 93.8 | 95.7 |
| | | Irregular pleural | 64.0 | 68.9 | 70.3 |
| | | Normal pleural | 70.6 | 83.2 | 85.6 |
| | | Total mAP | 73.3% | 79.5% | 81.4% |

### F. TRAINING AND TESTING PARAMETERS FOR RETINANET

The same anchor ratios used in Table 2 were used. The anchor scales were set to [0.25,2,5]. The number of potential anchors was set to 84. The model was trained for 100 epochs at $3.64 \times 10^{-4}$ learning rate. A bounding box threshold of 85%, a non-maximum suppression of 0.15, and a IoU of 0.2 were chosen. The models were trained on a 70%/30% split for train/test as shown in Table 6. The split was frame-based for each video from each patient. For example, if a patient has 6 videos and each video represents a certain LUS zone. Then each video zone was further divided into train and test frames. This was to ensure that the model learns the variation of each LUS feature across all views. Each patient had a range of 1-10 videos. The training and testing statistics of RetinaNet vs fRCNN are shown in Table 7.

We see that fRCNN outperforms RetinaNet, with an overall mAP of 86.57 % vs. RetinaNet's 61.15%. This mAP represents the average precision of all anomalies in the dataset. fRCNN is capable of capturing 5 out of 7 Lung pathologies with a mAP above 90%, while the other two pathologies have a mAP close to 65%. RetinaNet, on the other hand, was able to capture 4 out of 7 pathologies with a mAP greater than 60% and the remaining pathologies with a mAP less than 55%. Separate B-lines depicted low mAP for both models as they are rare relative to other pathologies. Thick pleural also showed low mAP for both models as its structure was visually similar to normal pleural. Despite RetinaNet's low efficiency, it is still closer to a real-time model, with an inference speed of only 2.53 s per frame. This model is roughly three times faster than the fRCNN model. It is due to the fRCNN model's complexity and the number of stages included in its architecture. RetinaNet is a one-stage detector as opposed to the two-stage fRCNN detector. That's why fRCNN takes twice as long to train as RetinaNet. fRCNN was selected as the efficient model for our LUS dataset due to its higher mAP despite its lower speed. Fig. 6 shows LUS frames from the six LUS study cases with their corresponding detected features.

**TABLE 6.** Train/Test frame split.

| Neonatal study cases | Training LUS frames | Testing LUS frames |
|---|---|---|
| *Normal* | 90 | 22 |
| *PTX* | 1086 | 271 |
| *TTN* | 2432 | 125 |
| *ARDS* | 910 | 228 |
| *CLD* | 768 | 192 |
| *Consolidations* | 2768 | 692 |

**TABLE 7.** Training and testing statistics of fRCNN and RetinaNet.

| Lung Pathology | RetinaNet mAP IoU=0.2 | fRCNN Model 1 mAP IoU=0.2 |
|---|---|---|
| *A-lines* | 45.8% | 93.0% |
| *Normal pleural* | 49.9% | 90.1% |
| *Irregular pleural* | 73.4% | 93.0% |
| *Coalescent B-lines* | 53.1% | 99.4% |
| *Separate B-lines* | 62.0% | 66.0% |
| *Thick pleural* | 61.0% | 68.5% |
| *Irregular pleural with consolidations* | 82.9% | 96% |
| *Total mAP* | 61.15% | **86.57**% |
| *Prediction Time Per frame* | 2.53s | 8.75s |
| *Time per epoch* | 35.8s | 61.92s |

Xing et al. [19] detected pleural line using a single class fRCNN reporting an accuracy of 95.41% vs a 90.10% by our seven class model as shown in Table 8. However, our model is a multiple class model vs their model which is only a single class focused on detecting only one LUS feature "Pleural line". Single-class models tend to have a narrower focus and can therefore make more precise predictions because they are only looking for one type of object. Our multi class model, however, is designed to detect more features. Our model proves to be more robust as it has been trained on a wider range of LUS features (six more prominent features). Interestingly, our model is within an acceptable margin from their work [19].

Moreover, while Xing et al. [19] utilized image processing techniques to identify A-lines, our approach relied solely on Faster RCNN and was fully automated. Their accuracy in detecting A-lines was 97.88%, whereas our model's accuracy

IEEE Journal of Translational Engineering in Health and Medicine

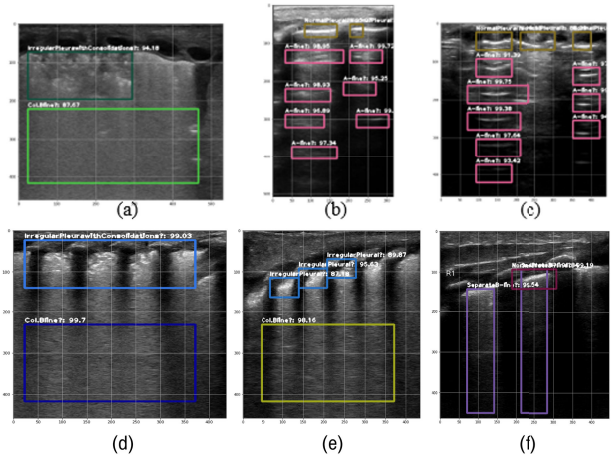


**FIGURE 6.** Screenshots from the Interpretable model: (a) CLD case with detected irregular pleural and consolidations as well as Coalescent B-lines. (b) Normal scan with normal pleural line and A-lines. (c) PTX scan with detected pleural line and A-lines (d) Consolidation scan with detected irregular pleural line with subpleural consolidations and Coalescent B-lines. (e) ARDS scan with its associated detected features: Coalescent B-lines and Irregular pleural lines. (f) TTN scan with detected normal pleural and separate B-lines.

**TABLE 8.** Comparison with the state-of-the-art for A-line detection.

| | Trained Deep Learning model | A-lines | Normal pleural |
|---|---|---|---|
| *Xing et al.* [19] | Single class fRCNN for Pleural line and Image Processing for A-line localization | 97.88% | 95.41% |
| *Ours* | Seven class fRCNN for seven LUS features | 93.00% | 90.10% |

**TABLE 9.** Performance results of Our model with other state of the art work on B-lines.

| | Trained Deep Learning model | All B-lines | Separate B-lines | Coalescent B-lines |
|---|---|---|---|---|
| *Sloun et al.* [26] | Single class 6-layer CNN | 89.20% | – | – |
| *Ours* | Seven class fRCNN | 82.70% | 66.00% | 99.40% |

was 93.00% as shown in Table 8. Yet, our model's results falls within relatively the same margin.

Sloun et al. [26] focused on the detection of B-lines using a single class CNN, scoring 89.20% vs 82.70% for our model as shown in Table 9. However, the paper localized B-lines using class activation maps (CAMs). These maps provide class scores and not precise localization as what our proposed model offers. Our model precisely gives information on the exact location of the object within an image through the use of bounding boxes, while CAMs only give a rough indication of where the object is located. Also, the paper has not focused on any characterization and differentiation of the detected B-lines. Our paper focused on detection of coalescent as well as Separate B-lines. We detect coalescent B-lines with an accuracy of 99.40%. Also, our model is a seven class feature model with a total mAP of 86.57% yet very close to the model proposed by [26].

**TABLE 10.** Automatic vs manual selection of line of motion.

| Normal | Accuracy of Automatic Cursor allocation | PTX | Accuracy of Automatic Cursor allocation |
|---|---|---|---|
| *Case 1* | 99.99% | *Case 1* | 99.99% |
| *Case 2* | 98.23% | *Case 2* | 92.45% |
| *Case 3* | 93.63% | *Case 3* | 98.06% |
| *Case 4* | 95.22% | *Case 4* | 92.95% |
| *Case 5* | 98.5% | *Case 5* | 98.42% |
| *Case 6* | 99.99% | – | – |
| *Averaged* | **97.59%** | | **96.37%** |

### G. EVALUATION OF AUTOMATIC VS MANUAL SELECTION OF LINE OF MOTION

In our work, we proposed an automatic cursor allocation to generate M-modes without involving a human in the loop setting. Cursors were manually positioned on 47 videos of both PTX and normal cases and compared to our extracted M-modes. The accuracy for PTX automatic cursor allocation was reported as 96.37%, and the accuracy for Normal was reported as 97.59% as shown in Table 10. The precision was calculated by dividing the manual cursor position by the predicted position. This high accuracy provides a reliable method for physicians to automatically generate M-modes for their neonatal study cases. We were able to diagnose 5 PTX and 6 Normal cases with 100% accuracy using these M-modes and the line detection technique.

### H. EVALUATION OF LINE DETECTION METHOD FOR DIFFERENTIATING PTX AND NORMAL M-MODE

Our proposed method for line detection was tested on LUS normal and PTX scans. Statistics such as the number of lines and the variance were calculated. One or two views in some cases did not meet the variance threshold. As a result, the variance for all views was averaged, and the final classification of the case was based on the average. Averaging variance was critical because, in the opinion of the physicians, it is not necessary for all 6-LUS-scan views to exhibit the typical features for the tested case. As a result, we see in PTX that 14 out of 16 videos have a variance greater than 500, whereas after averaging variance, we get a full 5/5 correct case evaluation. Similarly, in generated Normal processed M-modes, 29 out of 31 videos had variance lower than 500, resulting in a final 6/6 correct case evaluation.

Overall, the model has the potential to improve the efficiency and accuracy of Lung Ultrasound evaluation in healthcare settings.

## VI. CONCLUSION

Our research work focused on interpreting neonatal lung ultrasound scans. Rather than detecting the lung conditions

from direct image classification, physicians have recommended that an LUS feature extraction system will be more useful, since the ultimate diagnosis may depend on other factors than images alone. Therefore, we trained two object detection models and the excelling model was *faster* RCNN which achieved a mAP of **86.57**% using an IoU of 0.2.

We extracted M-mode images to enrich the details of the lung sliding movement shown in the LUS videos. We successfully extracted ROI's automatically with a high precision. The accuracy of our automatic line of motion was **97.59**% for **Normal** cases and **96.37**% for **PTX**, respectively. Image processing techniques such as edge detection and morphological procedures were applied over M-mode images. Hough transform approach was used to detect these lines in the M-mode images. Normal processed LUS M-modes showed a variance $\leq 500$, whereas PTX processed images showed higher values. We were able to effectively distinguish **6/6** Normal cases and **5/5** PTX scans.

### A. CLINICAL TRANSLATION PLANS

The highly applied nature of the project and the encouraging results sets the stage right for clinical translation. Currently, we are in the process of developing a software with the trained model to be installed on a computer on the premise of our clinical partner, the Mount Sinai Hospital in Toronto. First, we will conduct a usability testing of the software, where the usability will be evaluated using the System Usability Scale (SUS) [27]. Once the usability aspect is validated and up-to-standard, we will significantly expand the dataset through a retrospective study ($N > 500$) so that the obtained results are generalizable. Once the results have improved, we will conduct a validation study with held-out dataset. Additional funding sources will be secured to conduct the clinical translation.

## ACKNOWLEDGMENT

The authors thank Jenna Ibrahim from Mount Sinai Hospital for organizing the data collection process.

An earlier version of this paper was presented at the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) [DOI: 10.1109/EMBC46164.2021.9630169].

## REFERENCES

[1] J. E. Lawn et al., "Born too soon: Care for the preterm baby," *Reproductive Health*, vol. 10, no. S1, pp. 1–19, Nov. 2013.

[2] *Newborns: Reducing Mortality*, World Health Org., Geneva, Switzerland, 2019.

[3] A. Vezzani et al., "Diagnostic value of chest ultrasound after cardiac surgery: A comparison with chest X-ray and auscultation," *Crit. Care*, vol. 18, no. 1, pp. 1–182, 2014.

[4] H. M. El-Masry, M. A. Aladawy, T. M. Mansor, and H. A. A. E. Magd, "Comparative study between chest X-ray and lung ultrasound in neonatal respiratory distress," *Ann. Neonatol. J.*, vol. 3, no. 1, pp. 125–143, 2021.

[5] R. Bassiouny, A. Mohamed, K. Umapathy, and N. Khan, "An interpretable object detection-based model for the diagnosis of neonatal lung diseases using ultrasound images," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 3029–3034.

[6] (Sep. 4, 2021). *Darklabel: Video/Image Labeling and Annotation Tool Version 2.4*. [Online]. Available: https://darkpgmr.tistory.com/16

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[9] C. Wu, C. Luo, N. Xiong, W. Zhang, and T.-H. Kim, "A greedy deep learning method for medical disease analysis," *IEEE Access*, vol. 6, pp. 20021–20030, 2018.

[10] F. Mento et al., "State of the art in lung ultrasound, shifting from qualitative to quantitative analyses," *Ultrasound Med. Biol.*, vol. 48, no. 12, pp. 2398–2416, Dec. 2022.

[11] F. Corradi et al., "Assessment of extravascular lung water by quantitative ultrasound and CT in isolated bovine lung," *Respiratory Physiol. Neurobiol.*, vol. 187, no. 3, pp. 244–249, Jul. 2013.

[12] Y. Xia, Y. Ying, S. Wang, W. Li, and H. Shen, "Effectiveness of lung ultrasonography for diagnosis of pneumonia in adults: A systematic review and meta-analysis," *J. Thoracic Disease*, vol. 8, no. 10, pp. 2822–2831, Oct. 2016.

[13] C. Brusasco et al., "Quantitative lung ultrasonography: A putative new algorithm for automatic detection and quantification of B-lines," *Crit. Care*, vol. 23, no. 1, pp. 1–7, Dec. 2019.

[14] M. Gravina et al., "Deep learning in the ultrasound evaluation of neonatal respiratory status," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 10493–10499.

[15] N. Awasthi, A. Dayal, L. R. Cenkeramaddi, and P. K. Yalavarthy, "Mini-COVIDNet: Efficient lightweight deep neural network for ultrasound based point-of-care detection of COVID-19," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 68, no. 6, pp. 2023–2037, Jun. 2021.

[16] I. D. Apostolopoulos and T. A. Mpesiana, "COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, Jun. 2020.

[17] S. Magrelli, P. Valentini, C. De Rose, R. Morello, and D. Buonsenso, "Classification of lung disease in children by using lung ultrasound images and deep convolutional neural network," *Frontiers Physiol.*, vol. 12, Aug. 2021, Art. no. 693448.

[18] S. Kulhare et al., "Ultrasound-based detection of lung abnormalities using single shot detection convolutional neural networks," in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*. Granada, Spain: Springer, 2018, pp. 65–73.

[19] W. Xing et al., "Automatic detection of A-line in lung ultrasound images using deep learning and image processing," *Med. Phys.*, vol. 50, no. 1, pp. 330–343, Jan. 2023.

[20] H. Li et al., "High-resolution transthoracic ultrasonography for assessment of pleural lines in patients with dyspnea with CT comparison: An observational study," *J. Ultrasound Med.*, vol. 36, no. 4, pp. 707–716, Apr. 2017.

[21] M. Demi, G. Soldati, and L. Demi, "On the artefactual information of ultrasound lung images: A lines and B lines," in *Proc. Meetings Acoust.*, vol. 35. Melville, NY, USA: Acoustical Society of America, 2018, Art. no. 020003.

[22] R. Gregorio-Hernández, M. Arriaga-Redondo, A. Pérez-Pérez, C. Ramos-Navarro, and M. Sánchez-Luna, "Lung ultrasound in preterm infants with respiratory distress: Experience in a neonatal intensive care unit," *Eur. J. Pediatrics*, vol. 179, no. 1, pp. 81–89, Jan. 2020.

[23] Z. Alhassen, P. Vali, L. Guglani, S. Lakshminrusimha, and R. M. Ryan, "Recent advances in pathophysiology and management of transient tachypnea of newborn," *J. Perinatol.*, vol. 41, no. 1, pp. 6–16, Jan. 2021.

[24] N. Chowdhury, B. L. Giles, and S. D. Dell, "Full-term neonatal respiratory distress and chronic lung disease," *Pediatric Ann.*, vol. 48, no. 4, pp. e175–e181, Apr. 2019.

[25] U. Parekh, A. Maguire, J. Emery, and P. Martin, "Pneumothorax in neonates: Complication during endotracheal intubation, diagnosis, and management," *J. Anaesthesiol. Clin. Pharmacol.*, vol. 32, no. 3, p. 397, 2016.

[26] R. J. G. van Sloun and L. Demi, "Localizing B-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 4, pp. 957–964, Apr. 2020.

[27] J. Brooke, "SUS: A retrospective," *J. Usability Stud.*, vol. 8, no. 2, pp. 29–40, 2013.