Deep Learning Video Classification of Lung Ultrasound Features Associated with Pneumonia

Daniel E Shea[1*], Sourabh Kulhare[1*], Rachel Millin[2], Zohreh Laverriere[1], Courosh Mehanian[1,3,4], Charles B Delahunt[1], Dipayan Banik[1], Xinliang Zheng[2], Meihua Zhu[4], Ye Ji[1], Travis Ostbye[2], Martha-Marie S Mehanian[2], Atinuke Uwajeh[5], Adeseye M Akinsete[6], Fen Wang[4] and Matthew P Horning[1]

[1]Global Health Labs, Inc, Bellevue, WA. [2]formerly at Global Health Labs, Inc, Bellevue, WA. [3]University of Oregon, Knight Campus, Eugene, OR. [4]Oregon Health and Science University, Portland, OR. [5]iMedReads, Lagos, Nigeria. [6]College of Medicine, University of Lagos, Lagos, Nigeria.
daniel.shea@ghlabs.org, sourabh.kulhare@ghlabs.org

Abstract

Ultrasound (US) imaging holds promise as a low-cost versatile, non-invasive point-of-care diagnostic modality in low- and middle-income countries (LMICs). Still, lung US can be challenging to interpret because air bronchograms are anechoic and the US images mostly contain artifacts rather than lung anatomy. To help overcome these barriers, advances in computer vision and machine learning (ML) provide tools to automatically recognize abnormal US lung features, offering valuable information to healthcare workers for point-of-care diagnosis. This paper describes deep learning algorithms that target three key US features associated with lung pathology: pleural effusion, lung consolidation, and B-lines. The algorithms were developed and validated using a large and varied dataset of 22,400 US lung scans (videos) from 762 patients of all ages (newborn to adult) in Nigeria and China. The architectures include effective methods for leveraging frame-level and video-level annotations, are light enough to deploy on mobile or embedded devices and have high accuracy (e.g., AUCs ≈0.9). Coupled with portable US devices, we demonstrate that they can provide expert-level clinical assistance for diagnosis of pneumonia, which is the leading cause of both childhood mortality and adult hospitalization in LMICs. We also discuss some of the challenges associated with determining ground truth for pneumonia, which impact the question of how to leverage ML models for lung US to support clinical diagnosis of pneumonia.

1. Introduction

Ultrasound (US) imaging offers several key advantages over X-ray, computed tomography (CT), and magnetic resonance imaging (MRI). These advantages include real-time imaging, non-ionizing radiation, low cost, ease of sterilization, and portability, making US imaging well suited to point-of-care applications [1]. However, US suffers from issues such as noise, limited field of view, artifacts, and skeletal obstruction of organs. Lung US presents a fundamental challenge because the lung is generally filled with air, which does not propagate US waves, precluding direct visualization of lung tissue. Lung US images mostly consist of artifacts generated by wave interactions at the interface between the pleural cavity and the lung, which require expertise to interpret and which lead to inter-reader variability [2]. Scarcity of interpretive expertise limits the usability of US for respiratory disease monitoring, especially in low-resource settings such as rural areas in LMICs, where pneumonia is the primary reason for child mortality and the most common cause of adult hospitalization [3].

Advances in artificial intelligence can potentially help fill this expertise gap. Recent research efforts have applied deep neural networks (DNNs) to medical US image analysis tasks including classification [4], segmentation [5], detection [6], registration [7], biometric measurement [8], and quality assessment [9], as well as image-guided interventions and therapy, on body parts such as breast, prostate, liver, heart, brain, fetus, and kidney. For a review, see [10].

In this paper we seek to lower the barriers to using lung US for assisting respiratory disease diagnosis by developing DNN models that provide accurate and objective evaluation of lung US scans. We describe DNNs that identify three

[*] These authors contributed equally.

important clinical features in lung US videos - consolidation, pleural effusion, and B-lines - associated with abnormal lung conditions including pneumonia and COVID-19.

Contributions:

(1) Three video-level deep learning algorithms detect, with high accuracy (AUCs ≈0.9), key lung US features associated with pneumonia. These algorithms can be deployed to mobile devices in low resource settings.

(2) The algorithms were developed and validated on a highly diverse and expertly annotated data set of 22,404 US scans from 732 human subjects, representing various ages, genders, demographic backgrounds, and lung pathologies, with strong ground truth as to pneumonia. This is the biggest known lung US dataset targeting pneumonia.

(3) We detail specific challenges in pneumonia diagnosis that are central to both defining ground truth to train AI models and incorporating such models into clinical workflows.

2. Background

Lung US is becoming increasingly common in clinical practice for multiple conditions including pneumonia, pulmonary embolism, asthma, pulmonary edema, and pneumothorax [11]. Recent research [12, 13] shows that US can detect pneumonia at higher sensitivity than chest X-ray [14]. In addition, lung US has played a key role in clinical management of patients with COVID-19 associated lung abnormalities. A review [15] of 66 articles, with a total patient population of 4687, found that the most consistent findings between COVID-19 patients were multiple B-lines, sub-pleural consolidation, and pleural effusion. Our work is centered around these same lung US features—pleural effusion, consolidation, and B-lines (merged and single). These features are described in appendix 1 of the supplementary technical report, and examples are shown in Figure 1.

3. Related Work

Over the last several years, DNNs have been increasingly applied to the automatic interpretation of US imagery [1, 10, 16, 17]. Computer-aided analysis of endobronchial US has been used to identify benign and malignant lesions in patients with lung cancer [18]. DNN-based algorithms have detected pleural effusion and consolidation in lung US by training feature-specific models on a swine lung dataset [6], and have automatically detected and localized B-lines [19]. These studies focused on detecting features within individual US frames while ignoring temporal patterns that can be learned from a video loop.

COVID-19 involves similar lung pathologies to pneumonia, and multiple studies [20–25] demonstrated the effectiveness of DNNs on lung US for COVID-19, such as quantitatively analyzing the severity of COVID-19 pneumonia by characterizing patterns related to pleural lines and B-lines [26]. Releasing lung US datasets collected from several Italian hospitals during the early stage of the COVID-19 pandemic, including frame-level, video-level, and pixel-level annotations indicating the severity of disease, [27] proposed an end-to-end deep learning framework to predict disease severity score at the frame-level and aggregated frame-level scores to generate a video score. A Spatial Transformer Network [28] was used to highlight the spatial pattern of pathology. A multi-modal approach to assess COVID-19 infection severity combined US data with clinical information [29], recognizing that lung US is not a stand-alone diagnostic. In [30], 202 US videos were released and a Convolutional Neural Network (CNN) method to differentiate COVID-19, bacterial pneumonia, non-COVID-19 viral pneumonia, and healthy lung was described. We omit the customary comparison of our approach with other methods, because the following complexities limit our ability to make a meaningful comparison:

(i) Our focus is on video-level detection since it is more relevant to clinical needs, whereas existing methods typically concentrate on frame-level detection [6, 26, 27].

(ii) Since other methods use different criteria to determine the clinical significance of small lung feature instances [31–33], defining positive labels for them can vary between studies. The status of these small features strongly impacts accuracy as shown in Tables 3 and 4.

(iii) Challenges in reproducing and comparing methods arises from the absence of open datasets and universally accepted benchmark measures.

4. Method

4.1. Data collection and annotation

Large datasets are critically important for the development of deep learning algorithms. They provide the necessary training data to represent the wide variety of presentations that pathology gives rise to, as well as data to comprehensively evaluate the accuracy, robustness, and effectiveness of such methods. The present article benefits from an extensive library of 22,404 lung US videos representing 762 human subjects varying in age, gender, demographic background, and pathology.

Two data collection studies, with institutional review board approval and informed consent, provided data for this work. One study enrolled pediatric patients (age 0-18) in Nigeria, and the other enrolled adult patients (age > 18) in China. A *Mindray DP-10* system equipped with a convex array transducer (part 35C50EB) was used to capture the US scans. Data consisted of US videos and associated demographic and clinical data (all de-identified).

Each pediatric patient also had a diagnostic quality chest X-ray and each adult had a CT scan. Diagnosis was based

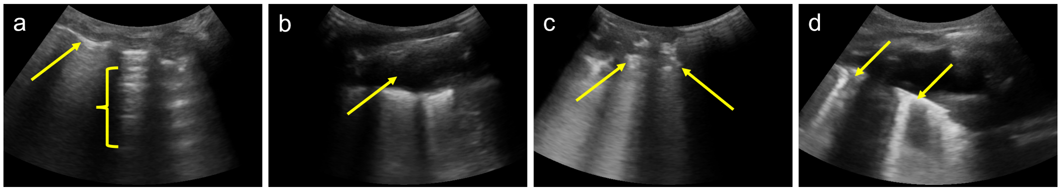

Figure 1. Example lung ultrasound frames and features: (a) Normal Lung, with arrow pointing to pleural line and bracket showing A-lines (b) Pleural Effusion (c) Lung Consolidation (d) Left arrow pointing to B-line, right arrow pointing to Merged B-lines

on all available clinical information, including radiography, enabling reliable pneumonia ground truth labels. (For challenges of definitive pneumonia diagnosis see section 6.) All but 20 patients had a diagnosis of pneumonia.

Ten lung zones per patient were imaged, with multiple ~ 3 second loops collected in each zone to capture variations in position, transducer/operator movement, and respiratory cycle. The average number of frames was 151 for pediatric videos and 80 for adult videos. In all, there were 378 adult and 384 pediatric patients and 22,404 usable videos.

Lung radiologists rated the overall quality of the video loop. Videos with inadequate quality (image or acquisition) were excluded. Every video was annotated by expert radiologists to note the presence of effusion, consolidation, and B-lines (merged and single). Videos annotated as containing B-lines and consolidation were further annotated on the frame-level, since these features are typically only visible in a subset of frames. Lung radiologists also labeled tiny effusions and sub-pleural changes (SPC, a very small consolidation or abnormal pleural line). These features have indeterminate clinical meaning; they are generally considered sub-clinical (not directly indicative of pneumonia) but still notable. Some research [32, 34] suggests that tiny effusions do not indicate health concerns warranting clinical action. Due to their equivocal clinical importance and their indefinite label boundaries, tiny effusions and SPC require special consideration for inclusion in training sets and their assessment in test sets. Examples are noted later.

4.1.1 Data cohorts for each model

Separate adult and pediatric models were trained for each feature. Videos were distributed among the training, validation, and testing sets by feature as shown in Table 1. All data splits occurred at the *patient* level; data from any one patient was placed in exactly one of the splits (train/validation/test). As seen in Table 1, not all of the $\sim$ 22,000 videos were used for training every model. These reduced video counts were due to various screening criteria and constraints which varied between the different models:

(i) For the consolidation model, two specific requirements were imposed on the data. First, videos where SPC was present were withheld from the training set, because doing so reduced specificity. Second, a subset of videos that were recorded with a non-standard device gain setting were withheld from the training set.
(ii) For B-line models, videos that were recorded with a non-standard gain setting were withheld from the training set.
(iii) For the adult pleural effusion model, pediatric pleural effusion videos were added to the training set to add to the limited number of adult pleural effusion videos available. However, the validation and holdout sets for the adult model included videos only from the adult population.

4.2. Deep learning architectures

We used two high-level architectures for video classification, driven by the type of annotations available. For pleural effusion, which only had video-level annotations, an LSTM-CNN classified videos as pleural effusion positive or negative (section 4.2.1). For consolidation and B-lines (merged and single), where frame-level annotations were available, a two-step classifier was used to leverage the more granular annotations. The first step computed classification confidence scores for each input frame independently. The second step took the vector of frame confidence scores as input to determine a video-level binary classification.

4.2.1 Pleural effusion one-step classifier

Pleural effusion is fluid in the pleural cavity which appears as a dark, anechoic or hypoechoic area. An ultrasound video clip captures the lateral motion between lung and the chest cavity during respiration. The pleural effusion models were trained specifically to identify pleural effusion of at least 0.5 cm depth. 'Tiny' effusions (< 0.5 cm) may not be clinically relevant if not associated with other abnormal features.

Because the pleural effusion annotations were video-level only, the model has a one-step, CNN + LSTM architecture which takes a video as input and outputs a video classification. A sequence of 60 regularly spaced frames are stacked to represent a video clip. Other frame counts were explored as detailed in appendix 3 of supplementary report, but 60 frames represented a good balance between computational load and performance; validation accuracy improved with increasing frame count but leveled off at 60 frames. Frames are centrally cropped with fixed cropping points to include

Adult	Training			Validation			Testing		
	neg	pos	exam	neg	pos	exam	neg	pos	exam
Pleural Effusion	4487	178	209	969	25	45	1170	34	45
Consolidation	3094	441	151	986	97	37	420	261	25
Merged B-line and B-line	2854	1797	153	631	581	39	288	546	22
Pediatric	Training			Validation			Testing		
	neg	pos	exam	neg	pos	exam	neg	pos	exam
Pleural Effusion	3683	461	272	952	105	72	208	58	18
Consolidation	1387	629	106	1944	644	119	2022	685	125
Merged B-line and B-line	1833	1189	122	973	680	239	811	406	238

Table 1. Counts of negative videos, positive videos, and patient exams in training, validation, and testing sets for each model.

only the fan-shaped segment of the image. Pixel values are normalized to [0, 1].

The model architecture is illustrated in Figure 2. Every frame is processed through a series of modules comprising a 2D convolutional filter, a ReLU activation function, a batch-normalization layer, followed by a 2D max-pooling operation. A flattening layer then a dense layer follow the last convolutional module. The weights of the convolutional modules and the dense layer were optimized in a time-distributed manner, by applying the same set of parameters to every temporally sliced video frame. The dense layer output is sequentially processed through two LSTM layers, of 128 units each. The first returns a sequence, which is processed by the second into a spatiotemporal feature vector that serves as input to a stack of two dense layers with softmax activation function for the final classification output. There are 831k parameters in the architectures for both pediatric and adult algorithms.

The LSTM [35] is a type of recurrent neural network (RNN) that models temporal patterns in sequential data. LSTMs accept past variable states as feedback–allowing information to flow in time–thus making them capable of learning long-term dependencies [36]. We found that a stack of two LSTM layers performed better than a single LSTM layer. Stacked LSTMs allows hidden states at each level to operate on different timescales [37] and adds an extra level of temporal abstraction. We used the Adam optimizer [38], learning rate of 0.0001, and categorical cross entropy loss with a mini batch size of 3 videos. Dropout [39] of 0.5 was used with LSTM and dense layers to forestall overfitting. Standard data augmentations (blurring, random pixel intensity adjustment, zero padding, frame averaging, and contrast adjustment) were used to combat overfitting and enhance generalization. The pleural effusion algorithms return a probability score (0-1) for a given video loop. A minimum score threshold of 0.525 was selected based on tuning the performance on the validation set.

4.2.2 Consolidation and B-line two-step classifiers

Models for consolidation, B-lines, and merged B-lines were two-step cascades, consisting of a frame classifier followed by a video classifier. The CNN frame classifier is trained to distinguish between feature-positive frames and feature-negative frames, treating each frame independently. The video classifier takes confidence score outputs from the frame classifier and outputs a video-level prediction.

The consolidation frame classifier returns a score indicating the probability of consolidation, and the video classifier returns a binary positive/negative prediction. B-lines require a more nuanced approach, both because of their visual similarity (B-lines and merged B-lines are on a continuum) and because of their uncertain clinical import: Merged B-lines are generally considered abnormal; B-lines also occur in healthy patients, but higher numbers can indicate illness. The B-line models seek to capture this complexity by using two binary frame classifiers and a decision matrix (based on their outputs) for a ternary video-level classification. At the frame level, one model distinguishes presence or absence of merged B-lines (with no regard for single B-lines), while the other detects the presence of any B-line (merged or single).

The next two sections describe the frame-level and the video-level stages.

4.2.3 Frame classifiers for two-step models

The three frame classifiers have the same CNN architecture, similar to the well-known VGG-16 but with fewer learned weights due to more aggressive pooling and one fewer convolutional layer. The architecture is presented in Figure 3. The input to the frame classifier is a batch of grayscale (single-channel) images with a batch tensor dimensionality of $(n, 256, 256, 1)$. The frame classifier outputs a vector of confidence scores $\boldsymbol{\eta} = [\eta_1, ..., \eta_i, ..., \eta_n]; \eta_i \in [0, 1]$. The confidence score is related to the probability that the frame is "positive". The models were trained with the RMSProp

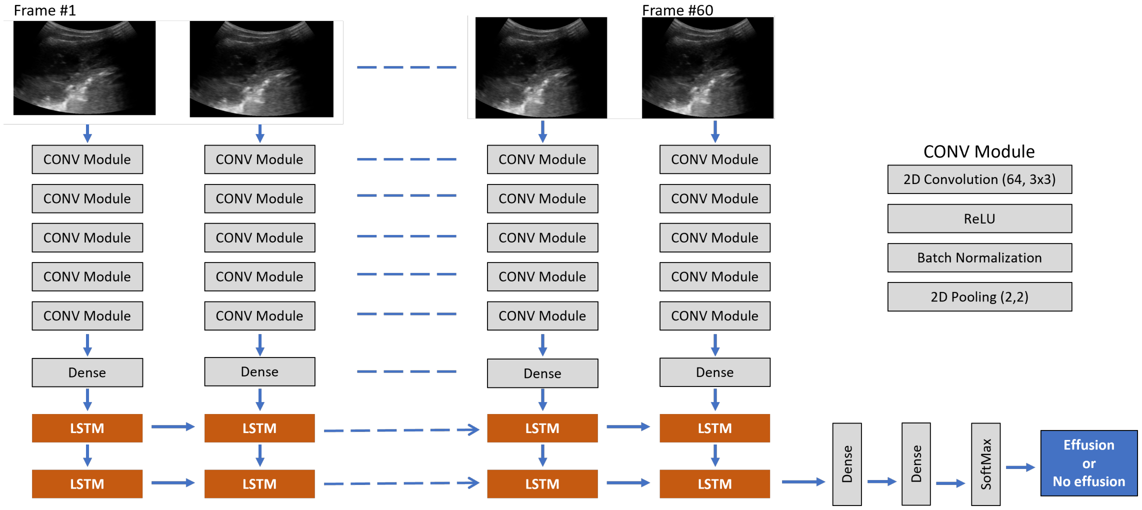

Figure 2. End-to-end architectural diagram for the pleural effusion models.

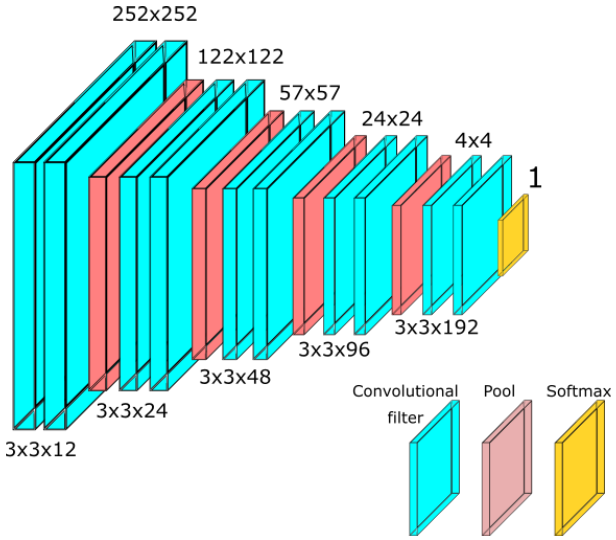

Figure 3. Convolutional network architecture used in consolidation, B-line, and merged B-line frame classifiers. The resolution is indicated along the top of the feature maps, while the convolutional filter size and number of channels are indicated below. The same filter size and channel count are used for both of the filters in the 2-filter pair. The top number indicates the size of the first feature map generated by the convolutional filter in the 2-filter pair.

optimizer, constant learning rates of $9e-4$ for consolidation models and $5e-4$ for merged B-line and B-line models, batch size of 16 images, and dropout of 0.56 for consolidation and 0.35 for the B-line and merged B-line models. Both models had fixed maximum number of training epochs set to 25 epochs, and both models enabled early stopping of training if cross-validation performance did not increase significantly over 5 epochs.

4.2.4 Video classifiers for two-step models

Consolidation model The consolidation video classifier takes the confidence scores from the frame classifier as input: $\boldsymbol{\eta} = [\eta_1, ..., \eta_i, ..., \eta_n]$, where η_i is the confidence score for the i-th frame and n frames are considered. The video classifier applies a threshold that is determined based on distributions of confidence scores in positive and negative training samples. A video is considered positive if the median confidence score is above this threshold and negative otherwise. Other techniques for classifying videos, such as hidden Markov models were explored but did not produce an improvement in video classification accuracy.

To minimize inference time, only a fraction of frames are processed. For the pediatric model, every 10th frame is classified; for adult, every 5th frame. This results in a minimal decrease in accuracy.

B-lines and merged B-line models Video classification for B-lines and merged B-lines is performed using hidden Markov models (HMMs), which are trained in a supervised manner. Let the set of classes be denoted by C. For each class $c \in C$, a generative HMM model, $\mathbf{H}_c(\cdot)$, is trained with m_c samples $\bar{\boldsymbol{\eta}}_{\boldsymbol{c}} = \{\boldsymbol{\eta}_{cj}\}$, $1 \leq j \leq m_c$, where $\boldsymbol{\eta}_{cj}$ is a sequence of confidence scores for the frames of the j^{th} video from class c (output by the CNN classifier described above). At inference time, a video with frame confidence sequence $\boldsymbol{\eta}$ is assigned to the class whose HMM has the maximum posterior probability of having emitted the sequence:

$$c^* = \underset{c\,\in\,C}{\arg\max}\ p(\boldsymbol{\eta}|\mathbf{H}_c). \tag{1}$$

In practice, the length of the sequence $\boldsymbol{\eta}$ (*i.e.*, number of confidence scores, equal to number of frames processed) does not matter because each HMM evaluates the same sequence. To avoid underflow, most HMM libraries compute the log-likelihood rather than the probability directly.

The B-line models that classify videos are binary, and the models use the same category class definitions as the frame classifiers: For one HMM classifier "positive" indicates the presence of either merged B-lines or B-lines and "negative" indicates the absence of both B-lines and merged B-lines; for the other HMM classifier "positive" indicates the presence of merged B-lines while "negative" indicates their absence.

The assumptions for an HMM are: (i) the system is modeled as a Markov process, which is a sequence of possible events, (ii) the system emits an observable parameter, and (iii) underlying, unobserved *hidden* states control the emission of the observable parameter. In this setting, the observable parameter is the sequence of frame confidence scores output by the frame classifier. The hidden states that emit these confidence scores could indicate the presence (or absence) of features in the US images, but the situation could be more complex than that.

HMMs are primarily characterized by the number of hidden states and the distribution of the observed variable(s) emitted from the hidden states. The hidden states in our HMMs are assumed to emit univariate confidence scores with a distribution described by a Gaussian mixture model. The HMMs $\mathbf{H}_c$ have the same number of Gaussian components and hidden states for all classes $c \in C$. The number of Gaussian components and hidden states for each model were chosen empirically, by selecting the best-performing models through cross-validation. Models with up to 4 hidden states and up to 4 Gaussian components in the mixture model were considered, and the best-performing model was selected based on the weighted F1 score of the model on the validation set. The optimal values are shown in Table 2.

Three of the models performed best with only 1 hidden state, which represents a simplification of a typical multiple hidden-state HMM wherein the underlying system has minimal or low-frequency temporal dynamics. The sufficiency of an HMM with minimal temporal dynamics mirrors the sufficiency, in the consolidation video classifier, of a simple threshold on the median frame score. This suggests that temporal dynamics were not very important in this problem.

Decision matrix for multi-class B-line classification As noted earlier, B-lines (merged and single) present a complex situation both clinically and visually. Here we describe how the two B-line model outputs are combined:

The outcome from the HMM classification models described above are two classifications indicating (i) whether a video is believed to contain merged B-lines **or** B-lines, and (ii) whether a video contains merged B-lines. The outputs from these two models can be further refined into a multi-class classification problem where a video could have B-lines, merged B-lines, both, or neither.

In this work, the final multi-class determination is made using rule-based logical operations rather than additional models. This simplifies the tuning of the algorithm to particular clinical needs and requirements. From a clinical perspective, merged B-lines are known to be correlated with adverse pulmonary conditions and thus are most concerning. From a modeling perspective, however, it can be challenging to differentiate between B-lines and merged B-lines. Our proposed approach for multi-class video classification aims to strike a balance between the complexity of frame and video classifiers and the simplicity of tuning the final video model. The performance of the algorithms can be tuned to minimize false positives, minimize false negatives, maximize true positives, or maximize true negatives by applying biases to the video-level classification log-likelihood.

4.2.5 Image preprocessing

All images were cast to gray-scale and resized to a specific input size prior to use by the deep learning networks. Images were resized to 384×384 for the pleural effusion models. Images were resized to 256×256 for the consolidation, merged B-line, and B-line models.

5. Results and Discussion

Tuning model performance for all these features is complicated by a lack of well-defined criteria based on clinical needs, and further by a lack of clear clinical import of SPC and tiny effusions. As a default, our tuning aimed for at least 85% sensitivity and specificity on the features with known clinical importance.

5.1. Pleural Effusion

Table 3 presents the pleural effusion results evaluated on a holdout set. Detailed composition of the holdout set can be found in Table 1. Sensitivity is reported at the video level for videos (i) with any type of effusion (including tiny); and (ii) with larger effusions only (excluding tiny) . The model effectively identifies large (clinically significant) effusions but often fails to detect tiny (clinically uncertain/insignificant) effusions. The majority of false positive videos contained dark, anechoic consolidation (example in Figure 4) without effusion in any rib space. False negatives were largely videos where effusion was transient due to transducer movement or patient movement/respiration.

The input image size of 384×384, 60 frames, and the chosen CNN + LSTM architecture yields an algorithm file size of 6.5 MB. The model processes one US video clip in 0.67 second on an Ubuntu system with 128 GB RAM and AMD Ryzen Threadripper 3960X processor.

Model	Gaussian Components	Hidden States
Adult (Merged B-Line or B-Line vs Negative)	4	4
Adult (Merged B-Line vs Non-Merged B-Line)	3	1
Pediatric (Merged B-Line or B-Line vs Negative)	1	1
Pediatric (Merged B-Line vs Non-Merged B-Line)	1	1

Table 2. Number of hidden states in each HMM model

	Sensitivity	Sensitivity	Specificity	AUC
	(w/ tiny,%)	(w/o tiny, %)	(%)	(w/o tiny)
Adult	40.1	94.0	82.3	.93
Pediatric	61.7	88.6	90.0	.94

Table 3. Video-level effusion results

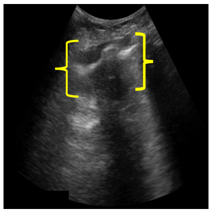

Figure 4. Anechoic consolidation

5.2. Consolidation

Video-level results of the consolidation algorithm on hold-out test sets are shown in Table 4. Specificity in reported for consolidation-negative videos (i) including those with SPC; and (ii) excluding those with SPC. Similarly to the tiny effusion case, the clinically uncertain/insignificant SPCs produce most false positives, especially in the pediatric population.

With an input image size of 256 × 256, the chosen CNN + median threshold classifier model architecture yields an algorithm file size of 0.5 MB. The consolidation CNN classifier contains 82k parameters. The model processes one US video clip in 55 milliseconds on an Ubuntu system with 128 GB RAM and AMD Ryzen Threadripper 3960X processor.

	Sensitivity	Specificity	Specificity	AUC
	(%)	(w/ SPC,%)	(w/o SPC,%)	(w/ SPC)
Adult	89.8	83.1	89.0	0.87
Ped.	89.2	84.3	98.0	0.95

Table 4. Video-level consolidation results

5.3. B-lines and merged B-lines

The video-level performance of the B-line and merged B-line models are presented in Table 5. The models are generally more specific than sensitive; this could be tuned by either adjusting the frame classifier or adding biases to the video classifier, as described in the Methods section 4.2.4.

	Sensitivity (%)	Specificity (%)	AUC
Model 1: MBL vs non-MBL			
Adult	84.0	86.6	0.92
Pediatric	77.0	91.6	0.91
Model 2: MBL or BL vs neither			
Adult	80.7	91.3	0.93
Pediatric	79.3	89.7	0.92

Table 5. Video-level merged B-line and B-line results

The video-level outputs from the two B-line models are combined to yield the ternary video-level classification described in Table 6, assigning the final label from the following classes: (i) negative, (ii) B-line positive, and (iii) merged B-line positive. As shown, the scenario with contradictory outputs from the two models (positive for merged B-lines, but negative for B-lines and merged B-lines) results in an overall negative label using these rules.

The rules-based classification method was optimized to minimize the rate of false positives. The three-class confusion matrix for the ternary classification scheme is shown in Figure 7 for both the adult and pediatric models. The confusion matrices demonstrate the preference for false negatives over false positives; less than 5.5% of negative videos are falsely labeled as containing merged B-lines for both adult and pediatric populations in the test data sets.

With an input image size of 256 × 256, the chosen CNN + HMM classifier architecture yields an algorithm file size of 0.6 MB. The merged B-line and B-line CNN classifiers each contain 82k parameters. The models typically process one US video clip in 60-80 milliseconds on an Ubuntu system with 128 GB RAM and AMD Ryzen Threadripper 3960X processor.

5.4. Age-based performance analysis

The pediatric patients in this work span the age range from <1 month old to 18 years old. We evaluated how the pediatric models performed for different age groups, with the results summarized and discussed in appendix 2 of supplementary technical report. Briefly, the models perform relatively consistently across age groups, though limited

(MBL) vs (no MBL)	(MBL or BL) vs (neither)	Final Classification
Negative	Negative	Negative (neither feature)
Negative	Positive	B-line positive
Positive	Negative	Negative (neither feature)
Positive	Positive	Merged B-line positive

Table 6. Merged B-Line (MBL) and B-Line (BL) Ternary Classification

		Predicted Label					
		Pediatric			Adult		
		Negative	B-Line	Merged B-Line	Negative	B-Line	Merged B-Line
True Label	Negative	0.91	0.033	0.053	0.9	0.062	0.04
	B-Line	0.36	0.42	0.22	0.39	0.43	0.18
	Merged	0.15	0.06	0.79	0.12	0.14	0.74

Table 7. Ternary B-line and merged B-line classification confusion matrix for pediatric (left) and adult (right) test populations.

patient numbers within each group make it difficult to draw firm conclusions.

6. Conclusions and Future Work

This article presents an initial exploration of automated video-level analysis of US for lung applications, and provides a baseline for future work. Although the presented work is evaluated for consolidation, pleural effusion, and B-lines, the frameworks defined here can be re-tuned to identify other pathologies. The architectures are efficient, and capable of processing videos in real-time on mobile devices such as smartphones or low-cost portable US systems. The algorithms allow a user with limited training in lung US interpretation to identify the abnormal lung conditions outlined here. This work is an important step towards developing an US device with on-board AI assistance for healthcare workers in LMICs, rural areas, and military settings.

US data collection for this project was performed by experienced medical professionals, and the algorithms in their current state “expect” good quality US video. An integrated image quality algorithm component could help to ensure high quality video input to these algorithms, irrespective of operator training. Since the video length was fixed at 3 seconds, and the respiratory cycle can be somewhat longer, lung features may be transient and appear in only a few frames. Alternatively, longer and variable length sequences could be included to increase the relevant temporal information.

In its current form, our framework provides only video-level output, and the models have been trained with ground truth based on video- and frame-level interpretations of expert radiologists. An “end-to-end" model that computes a patient-level diagnosis is challenging for multiple reasons:
(i) In a clinical setting, patient diagnosis is commonly based on a combination of numerous modalities, including clinical exam, history, laboratory tests, and radiography (X-ray, CT, and/or lung US). So an end-to-end model would need to be multi-modal, with model outputs for videos from multiple lung zones serving as one subset of inputs.
(ii) The diagnostic importance of the various features, especially tiny effusions, SPC, and single B-lines, are currently not well-defined. This complicates the use and interpretation of model results.
(iii) Defining ground truth is complicated because only moderate correlation exists between clinical conclusions drawn from X-ray, CT, and lung US imagery. CT is generally considered the gold standard for evaluation of lung abnormalities, but has the drawbacks of excessive cost and exposure to radiation, the latter precluding its use for pediatric patients. Thus, although CT provides the most reliable diagnostic modality, it is rarely used. X-ray is the usual standard of care for adults despite having relatively poor sensitivity for pneumonia. Lung US has been shown to have higher sensitivity to pneumonia than X-ray. This, along with its other positive characteristics, makes lung US an attractive alternative to X-ray, in both the adult and pediatric setting. An issue to contend with, however, is the lack of correlation between X-ray and lung US; patients with negative X-ray exams often show lung US findings. In the absence of CT imaging, it is likely that the entire clinical and laboratory records will be required to establish ground truth for these patients, which is a necessary step towards the development of an end-to-end patient-level assessment model.

7. Acknowledgements

Funding provided by Global Health Labs, Inc., www.ghlabs.org. We thank Dr. Peter Small and Dr. Amy Ginsburg for important discussions. We also thank Dr. Zihong Guo, Sheana Creighton, Matt O’Laughlin, Renee Rivers, and Stephen Burkot for coordinating data collection efforts.

References

[1] L. J. Brattain, B. A. Telfer, M. Dhyani, J. R. Grajo, and A. E. Samir, "Machine learning for medical ultrasound: status, methods, and future opportunities," *Abdominal Radiology*, 2018. 1, 2

[2] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, 2022. 1

[3] H. Zar, S. Madhi, S. Aston, and S. Gordon, "Pneumonia in low and middle income countries: progress and challenges," *Thorax*, 2013. 1

[4] Q. Guan and Y. Huang, "Multi-label chest x-ray image classification via category-wise residual attention learning," *Pattern Recognition Letters*, 2020. 1

[5] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of Digital Imaging*, 2019. 1

[6] S. Kulhare, X. Zheng, C. Mehanian, C. Gregory, M. Zhu, K. Gregory, H. Xie, J. M. Jones, and B. Wilson, "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*, 2018. 1, 2

[7] B. D. D. Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Iš̆gum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical Image Analysis*, 2019. 1

[8] Z. Sobhaninia, S. Rafiei, A. Emami, N. Karimi, K. Najarian, S. Samavi, and S. R. Soroushmehr, "Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019. 1

[9] C. T. Jensen, X. Liu, E. P. Tamm, A. G. Chandler, J. Sun, A. C. Morani, S. Javadi, and N. A. Wagner-Bartak, "Image quality assessment of abdominal ct by use of new deep learning image reconstruction: initial experience," *American Journal of Roentgenology*, 2020. 1

[10] S. Liu, Y. Wang, X. Yang, S. Li, T. Wang, B. Lei, D. Ni, and L. Liu, "Deep learning in medical ultrasound analysis: A review," *Engineering*, 2019. 1, 2

[11] T. Marini, D. Rubens, Y. Zhao, J. Weis, T. O'Connor, W. Novak, and K. Kaproth-Joslin, "Lung ultrasound: The essentials," *Radiology: Cardiothoracic Imaging*, 2021. 2

[12] A. Reissig and C. Kroegel, "Sonographic diagnosis and follow-up of pneumonia: a prospective study," *Respiration*, 2007. 2

[13] D. Lichtenstein, I. Goldstein, E. Mourgeon, P. Cluzel, P. Grenier, and J.-J. Rouby, "Comparative diagnostic performances of auscultation, chest radiography, and lung ultrasonography in acute respiratory distress syndrome," *The Journal of the American Society of Anesthesiologists*, 2004. 2

[14] S. Parlamento, R. Copetti, and S. D. Bartolomeo, "Evaluation of lung ultrasound for the diagnosis of pneumonia in the ed," *The American Journal of Emergency Medicine*, 2009. 2

[15] J. Gil-Rodríguez, J. P. de Rojas, P. Aranda-Laserna, A. Benavente-Fernández, M. Martos-Ruiz, J.-A. Peregrina-Rivas, and E. Guirao-Arrabal, "Ultrasound findings of lung ultrasonography in covid-19: A systematic review," *European Journal of Radiology*, 2022. 2

[16] N. Anantrasirichai, M. Allinovi, W. Hayes, and A. Achim, "Automatic b-line detection in paediatric lung ultrasound," in *2016 IEEE International Ultrasonics Symposium (IUS)*, IEEE, 2016. 2

[17] C. F. Dietrich, G. Mathis, M. Blaivas, G. Volpicelli, A. Seibel, D. Wastl, N. S. Atkinson, X.-W. Cui, M. Fan, and D. Yi, "Lung b-line artefacts and their use," *Journal of Thoracic Disease*, 2016. 2

[18] C.-H. Chen, Y.-W. Lee, Y.-S. Huang, W.-R. Lan, R.-F. Chang, C.-Y. Tu, C.-Y. Chen, and W.-C. Liao, "Computer-aided diagnosis of endobronchial ultrasound images using convolutional neural network," *Computer Methods and Programs in Biomedicine*, 2019. 2

[19] R. J. Van Sloun and L. Demi, "Localizing b-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results," *IEEE Journal of Biomedical and Health Informatics*, 2019. 2

[20] W. Xing, C. He, J. Li, W. Qin, M. Yang, G. Li, Q. Li, D. Ta, G. Wei, W. Li, *et al.*, "Automated lung ultrasound scoring for evaluation of coronavirus disease 2019 pneumonia using two-stage cascaded deep learning model," *Biomedical Signal Processing and Control*, 2022. 2

[21] Q. Huang, Y. Lei, W. Xing, C. He, G. Wei, Z. Miao, Y. Hao, G. Li, Y. Wang, Q. Li, *et al.*, "Evaluation of pulmonary edema using ultrasound imaging in patients with covid-19 pneumonia based on a non-local channel attention resnet," *Ultrasound in Medicine & Biology*, 2022. 2

[22] M. La Salvia, G. Secco, E. Torti, G. Florimbi, L. Guido, P. Lago, F. Salinaro, S. Perlini, and F. Leporati, "Deep learning and lung ultrasound for covid-19 pneumonia detection and severity classification," *Computers in Biology and Medicine*, 2021. 2

[23] J. Diaz-Escobar, N. E. Ordóñez-Guillén, S. Villarreal-Reyes, A. Galaviz-Mosqueda, V. Kober, R. Rivera-Rodriguez, and J. E. L. Rizk, "Deep-learning based detection of covid-19 using lung ultrasound imagery," *PLOS ONE*, 2021. 2

[24] 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 Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2021. 2

[25] G. Muhammad and M. S. Hossain, "Covid-19 and non-covid-19 classification using multi-layers fusion from lung ultrasound images," *Information Fusion*, 2021. 2

[26] Y. Wang, Y. Zhang, Q. He, H. Liao, and J. Luo, "Quantitative analysis of pleural line and b-lines in lung ultrasound images for severity assessment of covid-19 pneumonia," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2021. 2

[27] 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 Transactions on Medical Imaging*, 2020. 2

[28] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, 2015. 2

[29] W. Xue, C. Cao, J. Liu, Y. Duan, H. Cao, J. Wang, X. Tao, Z. Chen, M. Wu, J. Zhang, *et al.*, "Modality alignment contrastive learning for severity assessment of covid-19 from lung ultrasound and clinical information," *Medical Image Analysis*, 2021. 2

[30] J. Born, N. Wiedemann, M. Cossio, C. Buhre, G. Brändle, K. Leidermann, J. Goulet, A. Aujayeb, M. Moor, B. Rieck, *et al.*, "Accelerating detection of lung pathologies with explainable ultrasound image analysis," *Applied Sciences*, 2021. 2

[31] D. Lichtenstein, G. Meziere, P. Biderman, A. Gepner, and O. Barre, "The comet-tail artifact: an ultrasound sign of alveolar-interstitial syndrome," *American Journal of Respiratory and Critical Care Medicine*, 1997. 2

[32] V. SKarkhanis and J. M. Joshi, "Pleural effusion: diagnosis, treatment, and management," *Open Access Emergency Medicine: OAEM*, 2012. 2, 3

[33] A. Smargiassi, R. Inchingolo, G. Soldati, R. Copetti, G. Marchetti, A. Zanforlin, R. Giannuzzi, A. Testa, S. Nardini, and S. Valente, "The role of chest ultrasonography in the management of respiratory diseases: document ii," *Multidisciplinary Respiratory Medicine*, 2013. 2

[34] S. Afsharpaiman, M. Izadi, R. Ajudani, and M. H. Khosravi, "Pleural effusion in children: A review article and literature review," *International Journal of Medical Reviews*, 2022. 3

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997. 4

[36] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, 1994. 4

[37] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," *arXiv:1312.6026*, 2013. 4

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014. 4

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, 2014. 4