Automatic Deep Learning-Based Consolidation/Collapse Classification in Lung Ultrasound Images for COVID-19 Induced Pneumonia

Nabeel Durrani*,[2], Damjan Vukovic*,[1,3], Jeroen van der Burgt[1], Maria Antico[2,3], Ruud JG van Sloun[9], Libertario Demi[5], David Canty[6,10], Andrew Wang[6], Alistair Royse[6], Colin Royse[6,7], Kavi Haji[6], Jason Dowling[8], Girija Chetty[4], Davide Fontanarosa[1,3]

***Abstract*— Our automated deep learning-based approach identifies consolidation/collapse in LUS images to aid in the diagnosis of late stages of COVID-19 induced pneumonia, where consolidation/collapse is one of the possible associated pathologies. A common challenge in training such models is that annotating each frame of an ultrasound video requires high labelling effort. This effort in practice becomes prohibitive for large ultrasound datasets. To understand the impact of various degrees of labelling precision, we compare labelling strategies to train fully supervised models (frame-based method, higher labelling effort) and inaccurately supervised models (video-based methods, lower labelling effort), both of which yield binary predictions for LUS videos on a frame-by-frame level. We moreover introduce a novel sampled quaternary method which randomly samples only 10% of the LUS video frames and subsequently assigns (ordinal) categorical labels to all frames in the video based on the fraction of positively annotated samples. This method outperformed the inaccurately supervised video-based method of our previous work on pleural effusions. More surprisingly, this method outperformed the supervised frame-based approach with respect to metrics such as precision-recall area under curve (PR-AUC) and F1 score that are suitable for the class imbalance scenario of our dataset despite being a form of inaccurate learning. This may be due to the combination of a significantly smaller data set size compared to our previous work and the higher complexity of consolidation/collapse compared to pleural effusion, two factors which contribute to label noise and overfitting; specifically, we argue that our video-based method is more robust with respect to label noise and mitigates overfitting in a manner similar to label smoothing. Using clinical expert feedback, separate criteria were developed to exclude data from the training and test sets respectively for our ten-fold cross validation results, which resulted in a PR-AUC score of 73% and an accuracy of 89%. While the efficacy of our classifier using the sampled quaternary method must be verified on a larger consolidation/collapse dataset, when considering the complexity of the pathology, our proposed classifier using the sampled quaternary video-based method is clinically comparable with trained experts and improves over the video-based method of our previous work on pleural effusions.**

Submission date for review and to journal 31/12/2021.

*** Nabeel Durrani and Damjan Vukovic are shared first authors**
1 School of Clinical Sciences, Queensland University of Technology, Gardens Point Campus, 2 George St, Brisbane, QLD 4000, Australia.
2 Faculty of Engineering, Queensland University of Technology, Gardens Point Campus, 2 George St, Brisbane, QLD 4000, Australia.
3 Centre for Biomedical Technologies (CBT), Queensland University of Technology, Brisbane, QLD 4000, Australia.
4 School of IT & Systems, Faculty of Science and Technology, University of Canberra, 11 Kirinari Street, Bruce, ACT 2617, Australia.
5 Ultrasound Laboratory Trento, Department of Information Engineering and Computer Science, University of Trento, Italy.
6 Department of Surgery (Royal Melbourne Hospital), University of Melbourne, Royal Parade, Parkville, VIC 3050, Australia.
7 Outcomes Research Consortium, Cleveland Clinic, Cleveland Ohio, USA.
8 CSIRO Health and Biosecurity, The Australian eHealth Research Centre, Australia.
9 Dept. of Electrical Engineering, Eindhoven University of Technology, The Netherlands.
10 Department of Medicine and Nursing, Monash University, Wellington Road, Clayton, Victoria 3800, Australia.

***Index Terms*— Lung ultrasound; inaccurate supervision; label smoothing; weakly supervised deep learning; pulmonary consolidation/collapse diagnosis; machine learning, COVID-19 induced pneumonia.**

I. INTRODUCTION

Lung ultrasound (LUS) imaging has been used for the detection of highly contagious respiratory infections resulting from COVID-19 [1-6], as it has proven to outperform X-ray imaging and to be on par with computed tomography (CT) [7]. The portability of ultrasound imaging allows for diagnosing patients with contagious illnesses or restricted mobility, which makes it particularly useful for bed-side examinations or even point-of-care testing in particular in the late stages of COVID-19 induced respiratory disease. At the same time, the non-ionizing nature of ultrasound imaging allows for monitoring the progression of a given disease over time without exposing the patient to harmful radiation.

COVID-19 can result in multiple pathologies and imaging patterns being present as the disease progresses from early to late stages. These pathologies/imaging patterns vary from slight irregular and/or thickened pleural lining combined with the interstitial syndrome (i.e. lung scarring) in early stages, to consolidated lung regions with more pronounced pleural irregularities/thickening, interstitial syndrome, and pleural effusion (i.e. lung filled with fluid instead of air) [8].

The workflow for COVID-19 diagnosis using ultrasound typically consists of two steps: (1) a highly trained sonographer acquires the LUS images using a well-defined protocol (for example the one described in [9]), (2) followed by pathology interpretation. Both steps require considerable time and resources due to the extensive image acquisition training required [10] and the complexity of LUS images interpretation. These challenging aspects of the labelling and interpretation effort are time and resource expensive and can be alleviated by using machine learning-based automatic approaches to facilitate the training of a novice sonographer (or non-medical user) in acquiring LUS images using a standardised, reproducible method by guiding the user, for example, via displayed visual clues; and aiding the LUS images interpretation for pathology diagnosis.

Ultrasound findings in COVID-19 respiratory infection are initially pleural irregularities, followed by B-lines, pleural thickening and then development of small sub-pleural consolidation, which are typical of an interstitial pneumonitis from other contagions. If progression occurs an acute respiratory distress syndrome pattern results and other respiratory complications may occur such as pleural effusion, lobar pneumonia and pneumothorax, all of which can be identified readily with ultrasound [5].

Lung ultrasound may also be used to predict clinical response to intensive therapies such as prone ventilation or high positive end-expiratory pressure [11]. One of the principal hindrances to increased use of ultrasound in assessment of patients with acute respiratory illness is the reliance on the considerable training and experience required to perform and interpret ultrasound. This has been compounded during the COVID-19 pandemic as learning to perform ultrasound requires direct supervision and frequent contact with patients. Operator training and experience may be alleviated by using machine learning-based automatic approaches to facilitate the training of a novice sonographer (or non-medical user) in acquiring LUS images using a standardised, reproducible method by guiding the user, for example, via displayed visual clues; and aiding the LUS images interpretation for pathology diagnosis.

The related literature on automatic diagnosis of COVID-19 using LUS images revolves around using Deep Learning (DL) algorithms trained on COVID-19 images [12-15] and on imaging patterns such as B-lines and pleural thickening that are associated with COVID-19 [16-20]. These DL algorithms consist of convolutional neural networks (CNNs), which are presently considered the state-of-the-art for automated image analysis given their capability to extract low and high-level image features automatically. La Salvia M et al. [16] implemented a 4- (0-3) and a novel 7- class approach containing additional classes variations (namely 0, 0*, 1, 1*, 2, 2*, 3) to train a residual CNN where the classes vary from containing only A-line and B-lines (score 0) to artefacts resulting from the pleura and consolidated or tissue-like patterns (score 3). Alternatively, Arntfield et al. [17] used B-lines from patients diagnosed either healthy, hydrostatic pulmonary oedema, or COVID to train a residual CNN similar to [16] to automatically detect COVID-19. In contrast, Baloescu C et al. [18] created a custom supervised CNN to automatically assess and diagnose B-lines in non-COVID patients and compared their algorithm results to well-known algorithms such as ResNet and DenseNet.

Sadik F et al. [12], instead, proposed an approach where spectral mask enhancement (SpecMEn) and contrast-limited-adaptive histogram equalisation (CLAHE) pre-processing steps were used to enhance LUS images by reducing the noise present, before being implemented into a DL CNN for COVID-19 classification. Alternatively, Muhammad G et al. [13] made adjustments to a modified ResF CNN by fusing or combining the multiple layers of the CNN and directing them into their own classifier, which was then trained on the publicly available POCUS dataset, consisting of healthy, COVID, and pneumonia LUS images. The POCUS dataset was initially used in the approach proposed by Born J. et al. [21] where their custom 3 class DL CNN (POCOVID-Net) consisting of a modified VGG-16 NN was used to detect COVID-19 in LUS images.

Other works include Roy S et al. [22] who developed a COVID-19 severity scoring algorithm trained on a LUS dataset of patients. This dataset consisted of patients with mild (label = 1) to severe (label = 3) COVID-19 pathology which was validated in a frame and video method by trained sonographers. These images and their associated labels were used to train a DL algorithm in a weakly supervised way by providing segmented and image-based annotated ground truth labels to a Spatial Transform Network (STN) to automatically localise and classify severity of COVID-19 on a frame-by-frame basis. A class was assigned to each severity by providing a segmented ground truth label and was fed into a STN that determined the spatial relationship between the pathology and its location in each video and associated frame. Finally, our group in a previous work [23] focused on automatic identification and classification of pleural effusion using a modified DL COVID severity algorithm implemented initially by [22].

The approach proposed here further develops the pathology classification algorithm of [23] by identifying consolidation/collapse in patients that only exhibit or contain this pathology and are representative of COVID-19 respiratory issues in late stages. The novelty of this work includes the application to a unique consolidation/collapse dataset and the development of the video-based method of our previous work [23].

This development of the video-based method is based on the sampling of frames from LUS videos and bears similarity to label smoothing [24], a method that often improves the performance of classifiers trained on noisy labels (i.e. labels that may be incorrect) [25]. The opposing points of view for label smoothing are that (1) uniform noise is being injected to the labels hence accentuating the problem of noisy labels and that (2) the aforementioned smearing of label noise may reduce overconfidence in any one training example [25]. In practice, however, label smoothing has been demonstrated to be effective at improving classifier performance [25].

More concretely, the sampled video-based method is similar to label smoothing in the sense that it “smears” the noise such that the noise distribution is more uniform. One way this is done is through sampling frames from a video then assigning a score from 0-3 (i.e. quaternary labelling) depending on the number of frames containing the pathology (i.e. 0 label: no frames with pathology, 1-3 labels: increasing percentage of frames with pathology). The label selected is then assigned to each frame within the video, thus significantly reducing labelling time. Since a single label is being assigned to all frames of the video, the labelling noise becomes more uniform.

II. MATERIALS AND METHODS

A. Dataset

This study was approved by The Melbourne Health Human Research Ethics Committee (HREC/66935/MH-2020). Lung ultrasound images used in this study were previously acquired from a previous study [26] (Melbourne Health Human Research Ethics Committee approval HREC/18/MH/269, trial registration: http://www.ANZCTR.org.au/ACTRN12618001442291.aspx of patients admitted to hospital under an internal medicine unit with an initial diagnosis of cardiorespiratory disease. Lung ultrasound was performed using a Sonosite X-Porte portable ultrasound imaging system (Fujifilm, Bothell, WA, USA) with a 1-5 MHz phased array transducer. Lung ultrasound was standardized and followed the iLungScan protocol as established by The University of Melbourne, Ultrasound Education Group [27] and was performed by a physician trained and experienced in point of care lung ultrasound (XC) [26] and reviewed for diagnostic accuracy by an expert in lung ultrasound (DC, AR or CR). Patients were in a supine position for the examination, which was performed on all 3 anatomical zones of both lungs (Figure 1).

This dataset was collected following an in-house clinical protocol [28]. The protocol involved acquiring LUS imaging sequences of patients from 6 distinguished scanning regions as shown in Figure 1. The patients consisted of 10 unhealthy and 18 healthy patients that were admitted to the internal medicine department at the Royal Melbourne hospital where a cardiac, lung, femoral and vein ultrasound sequence were taken to determine a cardiopulmonary diagnosis.

All images (125 patients) were stored in DICOM format and interpretation by the physician recorded on a standardized form. The available dataset of images (125 patients) were reviewed for selection of images for inclusion in this study by experts in lung ultrasound (AR, CR and DC) that contained normal lungs, collapse and consolidation. A normal lung pattern was identified by the presence of normal lung sliding or lung pulse, reverberation artifacts from the pleura, and absence of atelectasis (collapse) or consolidation (Figure 2). Discrimination between lung atelectasis (collapse) and consolidation using lung ultrasound is not always possible and when present they are usually both present. Hence, for this preliminary study they were

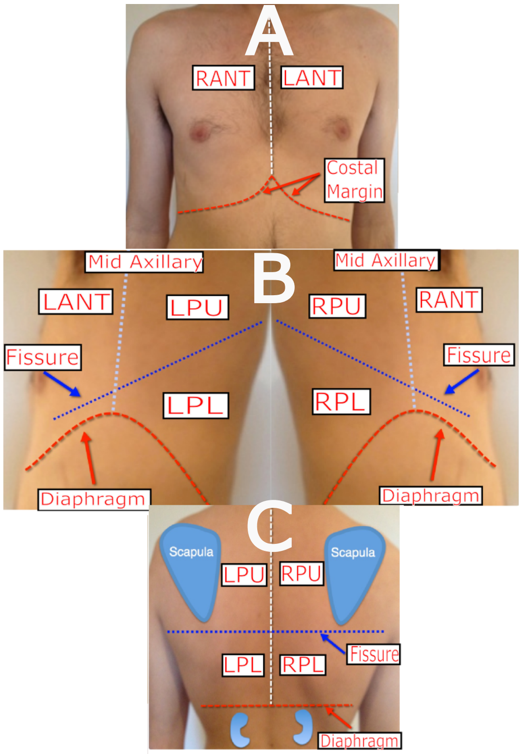


Fig. 1. Describes the scanning locations: (A) Right Anterior (RANT) and Left Anterior (LANT); (B) Lateral Posterior Upper (LPU), Lateral Posterior Lower (LPL), Right Posterior Upper (RPU), and Right Posterior Lower (RPL); (C) Lateral view for LANT, LPU, and LPL.
Source: Adapted from [5].

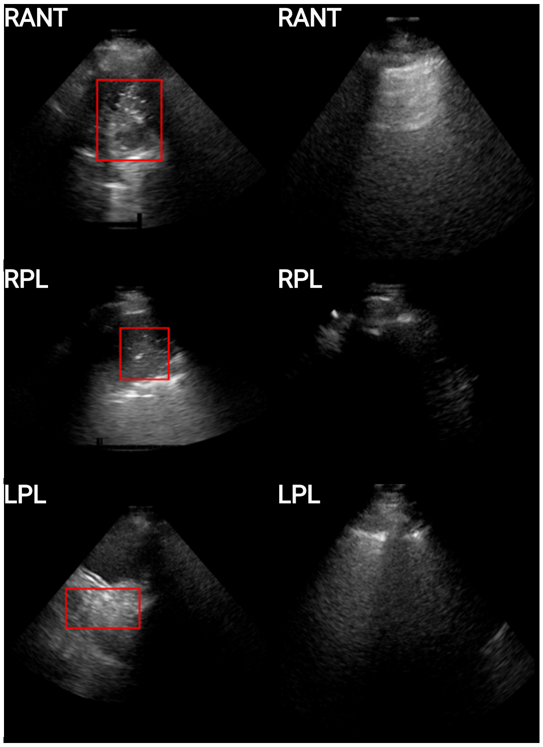


Fig. 2. Examples of unhealthy (left column) and healthy (right column) patients for 3 available scanning regions (viz. RANT, RPL, LPL). The unhealthy patients are those for which consolidation is present, and are depicted here with a red bounding box encompassing the pathology, while the healthy patients are those for which no pathology is present.

both considered as one pathology: collapse/consolidation. Discrimination between collapse and consolidation may be a future project. Collapse/consolidation was defined as an area of increased tissue density (tissue pattern) in the lung space that has the appearance of a solid organ, such as the liver ('hepatization'). Other features used to assist in diagnosis of collapse/consolidation include air bronchograms (hyperechoic dots) and loss of lung volume, however these were not required for diagnosis.

Representative cases of LUS frames are given by Figure 2. Only three of the six regions are depicted since the dataset only has consolidation present in those regions.

A total of 28 patient were scanned resulting in 51 videos and 34318 frames and the distribution of unhealthy and healthy patients in the dataset and the pathologies present are shown in Table II. An unhealthy patient is defined when a video or frame that contains signatures of consolidation and possibly of other lung pathologies is present in the corresponding dataset; while a healthy patient is defined as having neither consolidation or any other known respiratory issue present imaging patterns (anatomical or artefactual) in the LUS images and videos provided by Royal Melbourne Hospital [28].

In Table I, with respect to the first column (*number of patients*), the columns under the heading *pathology diagnosis* show the pathologies present, while the last 2 columns show the number of videos and frames associated with the patients of each row. The last row of the table shows the number of patients that do not have any of the presented pathologies in the *pathology diagnosis* columns nor any other pathology. These patients are considered the control group (i.e. the healthy patients).

The class imbalance of the dataset is reflected in the fact that 89% of the test set ground truths across folds are negatives, with true negatives comprising the disproportionate majority of negatives for each labelling method (Table II).

1) Dataset exclusion criteria: The image quality for unhealthy and healthy LUS videos/frames contributed to what patients and their associated videos and frames would be included for both the training of the algorithm and the ground truth labelling done by the trained sonographers. Clinical experts, namely, an experienced sonographer and a certified medical doctor, labelled the dataset described in Section A to provide the DL model with ground truth labels for training purposes. Each LUS frame was assigned a binary label indicating if it contained clinical signs of consolidation/collapse (Score 1) or not (Score 0).

An initial analysis of the dataset contained inconsistencies between the diagnosis presented in the medical report upon validation from an expert and provided a categorical division of a pathology given per patient to a pathology given per patient per video. These images along with images acquired from improper US probe placement or containing heavy imaging artefacts were excluded from the consolidation/collapse dataset until they could be further reviewed and validated. When the medical reports provided could be validated and accounted for, these images were reintroduced into the consolidation/collapse overall dataset or if they did not meet the criteria mentioned above, they were excluded from use for training the algorithm.

A further exclusion criterion we refer to as the *clinical criteria* presented in Section A2 takes the consolidation/collapse dataset after the initial exclusion of images and provides a clinical labelling that is further divided into 3 categories (Y/Y*/N). The original consolidation/collapse dataset consisted of 11 patients or 41 videos (5450 frames), from there, 10 patients or 36 videos (4910 frames) remained from the initial exclusion criteria and finally after the

TABLE I
DESCRIBES THE PATHOLOGY DISTRIBUTION AMONG PATIENTS.

	Pathology Diagnosis						
Number of patients	**Consolidation**	**Collapse**	**PE**	**APO**	**Interstitial Syndrome**	**Number of videos**	**Number of frames**
2	✓					7	1107
3	✓	✓	✓			10	1233
3	✓			✓	✓	10	1486
1	✓	✓				2	291
1	✓		✓			4	449
18	Healthy patients (no pathology present)					18	29167

Abbreviations: Acute Pulmonary Oedema (APO), Pleural Effusion (PE)

clinical criteria this led to a final consolidation/collapse training dataset of 9 patients or 27 videos (3827 frames) and is shown in Section C.

2) Criteria for clinical significance of algorithm performance: Once the performance of the algorithm was calculated these same videos were once again given to a trained sonographer with the intent of providing a video-based evaluation on a per patient per scanning region following a set of criteria. These criteria are based on the sonographer's confidence level in determining if a scanning region contains consolidation/collapse or not based on the video and frames image quality that can be affected by many contributing factors.

This criterion along with the pathology identification protocol [28] provided by the Royal Melbourne hospital has given a 3-scoring system for identifying consolidation/collapse and anatomical landmarks to determine LUS scanning position (ex. RANT, LANT, RPL, LPL, etc) in already consolidation/collapse identified videos and is shown in Table IV. These labels include Y, Y* and N representing the level of confidence in identifying frames associated with each scanning region on a per patient per video level. A label of N represents an inconclusive decision in determining both the scanning region and consolidation/collapse identification due to varying factors. Y* represents frames in the associated video having both a high confidence of frames containing consolidation/collapse with the inclusion of frames that are inconclusive in identifying anatomical landmarks. Finally, Y represents frames with a clear determination of consolidation/collapse being present and its associated anatomical markers.

B. Pre-processing

An open-source DICOM processing package (DICOM package used in python called Pydicom) was used to extract the original pixel data from the compressed DICOM format. Next, the various overlays inside and outside the ultrasound sector, including text, watermarks, and trademarks from the ultrasound imaging system, were replaced with black background pixels. The final step included cropping the images from 960 × 720 pixels to a size of 806 × 550 pixels which contained the ultrasound sector to reduce the dataset size before being input into the DL model and only include the relevant information contained in the image.

C. Frame-based labelling strategy

In the case of unhealthy patients, if the presence of consolidation/collapse could not be confidently identified in each frame using the provided protocol [28], these frames were labelled indecisive or inconclusive and were further examined by additional multiple trained experts to determine whether these frames are representative of the pathology. If an indecisive or inconclusive label was given and caused by poor image quality arising from artefacts overlapping/overshadowing the pathology due to inadequate ultrasound probe placement, these images were not included in the training of the algorithm. Table III represents the number of frames and their associated scanning region locations of the consolidation/collapse dataset. The training dataset for the unhealthy or consolidation/collapse consists of the frames and scanning positions after the process of the exclusion and clinical criteria has been applied. Where the healthy patients consist of equivalent number of frames and respective scanning regions.

In Table III, the number of frames considered, and their respective anatomical scanning regions are shown before and after any criteria were applied. These frames are used to train the algorithm by providing the ideal representation of the imaging patterns presented in LUS of patients/frames diagnosed with consolidation/collapse.

D. Video-based labelling strategies

Besides the standard frame-based labelling approach described in Section C, video-based labelling strategies were also explored, as described Section D, to reduce labelling time.

While the training of the frame-based model is considered supervised learning, since frame-level predictions are evaluated using frame-level labels during training, the training of the video-based models is considered inaccurately supervised learning, a form of weakly supervised learning. This is because it is possible for the frame-level labels used for training to have errors [29]. Such errors result because each video-based labelling strategy involves using a subset of frames from each video to determine the single label that is given to all frames of the video (Figure 3).

The all-or-nothing binary method takes a label/score of the video given by a de-anonymised medical report and is used for all the associated frames in that video. Whereas the sampled binary (0-1) and quaternary (0-3) video-based methods take a random 10% sample of all the labelled frames done by trained sonographers and give that same label to the overall video that these frames were taken from.

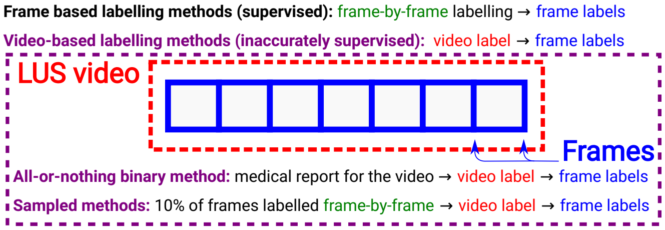

Frame based labelling methods (supervised): frame-by-frame labelling → frame labels
Video-based labelling methods (inaccurately supervised): video label → frame labels
LUS video
Frames
All-or-nothing binary method: medical report for the video → video label → frame labels
Sampled methods: 10% of frames labelled frame-by-frame → video label → frame labels

Fig. 3. Video-based labelling strategies: the all-or-nothing binary video-based method (Section II.D1) assigns the video label to all frames of the video, and the sampled binary and quaternary video-based methods (Section II.D2) take the label of a random 10% sample of the frames associated with the video and assign that label to each frame of the respective video.

TABLE II

THE CLINICAL CRITERIA USED TO SCORE LUS VIDEOS THAT HAVE BEEN EVALUATED BY THE ALGORITHM BEFOREHAND. THE LABELLING SYSTEM DETERMINES A CERTAIN CONFIDENCE OF AN EXPERIENCED SONOGRAPHER TO IDENTIFY THE OVERALL RATING OF A VIDEO BASED ON THE CORRESPONDING CONSOLIDATION/COLLAPSE IMAGING PATTERN PER FRAME USING ANATOMICAL MARKERS, IMAGE QUALITY, AND OTHER POSSIBLE OBSTRUCTIONS FOR APPROPRIATE IDENTIFICATION.

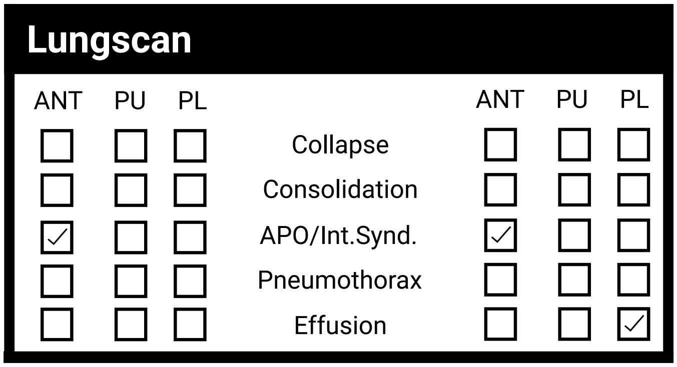
	Y	**Y***	**N**
Obstructions to identification	Little to no ambiguity or obstructions to view	Possible causes: - Artefacts overlapping - Poor ultrasound probe placement	
Consolidation/collapse identified	Conclusive frames	Conclusive frames Inconclusive frames	Inconclusive frames
Anatomical landmarks/scanning regions identified	Conclusive frames	Conclusive frames Inconclusive frames	Inconclusive frames
Image quality	Good to high	Not optimal to normal	Not optimal / poor

TABLE III

DESCRIBES THE CONSOLIDATION/COLLAPSE DATASET DISTRIBUTION BEFORE AND AFTER APPLICATION OF THE DATA EXCLUSION AND CLINICAL CRITERIA.

	Number of frames for training			
	RANT	**RPL**	**LPL**	**Total**
Original data	807	3065	1578	5450
After exclusion criteria	507	2825	1578	4910
After clinical criteria	507	2071	1249	3827

1) All-or-nothing video-based labelling: For the all-or-nothing video-based approach, each patient was provided with iLungScan™ (Heartweb Pty Ltd, ITeachU Ltd, ACN 146184812) reports from LUS experts, developed by the Ultrasound Education Group at the University of Melbourne and validated by other experts from the University of Melbourne and QUT. These reports state the severity of consolidation/collapse present (as well as other pathologies) in the six scanning positions and are marked with a checkmark as shown in Figure 4.



Lungscan

ANT	PU	PL		ANT	PU	PL
☐	☐	☐	Collapse	☐	☐	☐
☐	☐	☐	Consolidation	☐	☐	☐
☑	☐	☐	APO/Int.Synd.	☑	☐	☐
☐	☐	☐	Pneumothorax	☐	☐	☐
☐	☐	☐	Effusion	☐	☐	☑

Fig. 4. Example of the information provided by the medical report where each LUS scanning region consists of a video that has been checked marked if it contains a pathology.

This video-based labelling method was used in a previous work of our group on pleural effusion [23] and has been implemented here with a binary label based on the initial medical reports. Specifically, a single binary label indicating whether consolidation/collapse was present in an LUS video was obtained from the medical report, and this label was used to label each of the frames within the video (Figure 3).

Effectively, if a video frame contained a single video framed exhibited signatures of consolidation/collapse, all the remaining frames of the video would be incorrectly labelled as having consolidation/collapse. The all-or-nothing method was named after this limitation, whereby videos containing few frames with consolidation/collapse present produce more frames that are incorrectly labelled than are correctly labelled. The shortcoming is addressed by the sampled video-based labelling methods described in the following section.

2) Sampled binary and sampled quaternary video-based labelling: For the sampled binary and sampled quaternary labelling methods (Figure 5), a label was assigned to a video and extended to its constituent frames based on the number of frames exhibiting signatures of consolidation/collapses as a percentage of the total number of frames (Figure 3). For these sampled labelling methods, given an LUS video, 10% of its frames were randomly sampled. In our case, labels were readily available for these sampled frames from the frame-based labelling method (Section C). However, in practice, a clinical expert would review the LUS video before stepping through the sampled frames, assigning each frame a binary label of 0 (healthy) or 1 (unhealthy).

After the 10% sampling of frames from an LUS video, all frames of the video were then assigned the same label depending on the proportion of unhealthy sampled frames, i.e. frames containing signatures of pulmonary consolidation/collapse (Figure 5). For the binary method, a label of 0 was given to all video frames if less than half the sampled frames were unhealthy. Otherwise, a label of 1 was given. For the quaternary method, on the other hand, if the proportion of unhealthy sampled frames was lower than 25%, a label of 0 was given. Otherwise, a label of 1 was given for a proportion less than 50%, and otherwise a label of 2 was given for a proportion less than 75%. A label of 3 was given for the remaining interval of 75%-100% inclusive.

E. Cross-validation

As per [22] and [23], the train-test splits were performed at the patient level during cross-validation, i.e. all the images of a given patient were either included in the training or in the test set. The patient split was performed by assigning a binary label of healthy

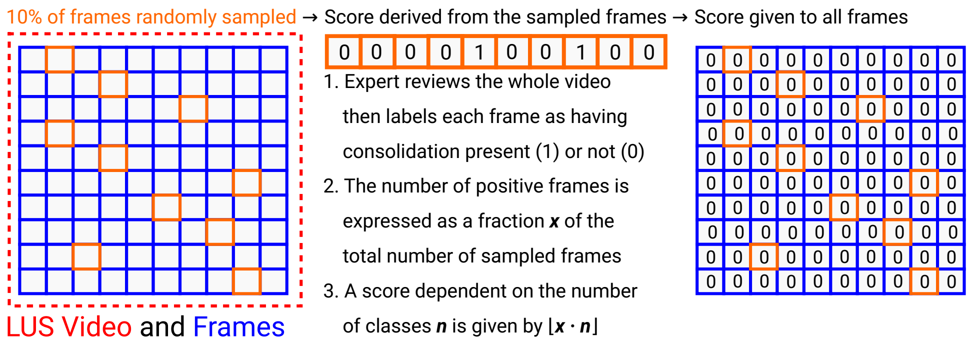

10% of frames randomly sampled → Score derived from the sampled frames → Score given to all frames
0 0 0 0 1 0 0 1 0 0
1. Expert reviews the whole video then labels each frame as having consolidation present (1) or not (0)
2. The number of positive frames is expressed as a fraction x of the total number of sampled frames
3. A score dependent on the number of classes n is given by ⌊x · n⌋
LUS Video and Frames

Fig. 5. Illustrates the flowchart for the sampled binary and quaternary video-based labelling methods.

or unhealthy (i.e. 0 or 1, respectively) to each of the 28 patients using the medical reports described in Section A. This could be done unambiguously because there were no patients for whom a mixture of healthy and unhealthy LUS videos was collected, as determined by the medical reports.

Stratified 10-fold cross-validation was used. Each test set contained video frames from exactly one patient for which consolidation/collapse was present, with the remaining video frames belonging to healthy patients for which no pathology was present. Since there was a total of 9 patients for which consolidation/collapse was present, one of the 10 folds had a test set containing only healthy patient video frames and was hence excluded. Note that for the sampled labelling methods of Section D2, the 10% sampling used to generate labels was performed once initially, rather than once per fold.

The trade-off between the quantity and quality of training examples is managed through the use of the N/Y/Y* video categories from Section A2. Unlike the training examples from the videos labelled N, the Y* videos were deemed of high enough quality to be used for training in addition to the Y videos. However, since they did not possess the near-perfect inter-observer agreement of the Y videos, they were not considered of sufficient quality to be included in the test set to evaluate the algorithm performance. Hence ultimately the test set consisted solely of healthy videos and Y videos.

During the cross-validation process for training the algorithm, there was an equivalent number of frames between healthy and unhealthy frames that consisted of the same scanning region. Therefore, per training fold the scanning regions used for healthy and unhealthy frames was kept relatively the same to represent a training done on a balanced dataset.

F. Deep learning model

A DL architecture consisting of a CNN and Spatial Transformer Network (STN) [30] was employed. Specifically, it used a Regularised Spatial Transformer Network (Reg-STN) [22] to localise signatures of pulmonary consolidation/collapse. The Reg-STN uses ordinal labels (i.e. binary or quaternary labels in our case) as opposed to the explicit consolidation/collapse locations per frame [22]. It creates an image crop used by the CNN for feature extraction to ultimately produce a prediction [22].

The algorithm was optimised using the same loss function as Roy S et al. [22], who employed it for COVID-19 severity score estimation, an ordinal regression [31] problem. This overall loss function, taking the form of a sum of terms, incorporated as one of its terms a soft ordinal regression (SORD) [32] cross-entropy loss function to allow long-distance errors to be penalised harsher than low-distance errors [22].

G. Training approach

Since the training and test sets were formed using stratified cross-validation, their class distributions reflected that of the whole dataset, with far healthier LUS frames than unhealthy frames. Hence, following [23], a batch-level class balancing was implemented using the weighted random sampler from PyTorch [33, 34]. As in [22], the DL model was trained using an Adam optimiser with an learning rate decay of 1×10^{-4}, early stopping on the training loss, and online data augmentation [22]. This training was run up to a maximum of 80 epochs, using a batch size of 32 and an initial learning rate of 1×10^{-5}. For the frame-based labelling, all-or-nothing video-based labelling, and binary video-based labelling the number of classes of the SORD loss function described in Section F was set to $n = 2$, while for the quaternary video-based labelling method $n = 4$ was used.

The network was trained on a single Nvidia Titan RTX GPU with 24 GB of memory installed on a workstation running Linux with 128GB of memory. The GPU workstation used an Intel i9-9820X CPU with 20 cores running at 3.30 GHz (Lambda Labs, San Francisco, CA, USA).

H. Evaluations

1) Evaluation metrics: The models trained using both the frame-based labelling approach and the video-based labelling approaches produced frame-level predictions, which were evaluated against the frame-based binary ground truths as in [23].

To evaluate the quaternary method in a manner comparable to the all-or-nothing and sampled binary methods, quaternary labels 0 and 1 were considered negatives (i.e. healthy) with the rest being considered positives (i.e. unhealthy). That is, letting $\mathbf{a}$ denote the test set label and $\mathbf{p}$ denote the prediction, a linear projection from the 4×4 quaternary confusion matrix $[Q_{\mathbf{ap}}]$ to the binary confusion matrix was defined by the equations

$$\begin{aligned}
\mathrm{TP} &= Q_{22} + Q_{23} + Q_{32} + Q_{33} \\
\mathrm{FP} &= Q_{02} + Q_{03} + Q_{12} + Q_{13} \\
\mathrm{FN} &= Q_{20} + Q_{21} + Q_{30} + Q_{31} \\
\mathrm{TN} &= Q_{00} + Q_{01} + Q_{10} + Q_{11}
\end{aligned}$$

which respectively define the true negatives TN, false positives FP, false negatives FN, and true positives TP for the quaternary method. Note that $Q_{\mathbf{ap}} = 0$ when $a = 1$ or $a = 2$ since the frame-based ground truths are binary. This method allowed metrics such as accuracy, precision, recall, and F-score to be defined using binary formulae for the quaternary method and thereby compared to the same set of metrics applied to the frame-based method and the all-or-nothing and sampled binary video-based methods. A classification threshold of 0.5 was used to separate the positive and negative classes to evaluate these metrics and was not calibrated because the calibration would require us to reduce the size of our already small training and test set sizes to afford a validation set for threshold tuning.

Given this imbalance, the appropriate classification threshold-independent measure of skill is PR-AUC score, which does not account for true negatives and thereby exaggerate classifier performance, unlike ROC-AUC score [35]. Additionally, precision-recall curves [35] were also used for evaluation. Each point on a precision-recall curve corresponds to a possible value for the classification threshold that separates negative and positive classes. This threshold is applied to the predicted score for the positive class, or the sum of the predicted scores for the positive classes (viz. 2 and 3) in the case of the quaternary method. Precision-recall curves are summarised by the Precision-Recall curve Area Under Curve score (PR-AUC), which corresponds to the average precision across the precision-recall curve.

2) Evaluation of statistical significance: To evaluate the significance of PR-AUC scores, the statistical procedure suggested by [36], stratified bootstrapping, was used. Stratified bootstrapping involves drawing n positive samples and m negative samples from the dataset with replacement for each of $\mathcal{I}$ iterations to produce a distribution of $\mathcal{I}$ bootstrapped PR-AUC scores. Stratification is necessary because PR-AUC scores are sensitive to class imbalance [36], with the $n : m$ ratio giving the vertical centre point for a horizontal line, which is the precision-recall curve for a classifier with no skill.

Specifically, for each fold, given the test set predictions for a pair of labelling methods to be compared, a pair of PR-AUC scores was obtained, whose difference we refer to as the observed difference. To test if the two scores for each fold were significantly different from each other, and hence test the null hypothesis that the performances of the pair of labelling methods were insignificantly different, stratified bootstrapping with 10,000 iterations was employed. For each fold, two sets of bootstrapped PR-AUC scores, corresponding to a pair of labelling methods, were obtained and used to form the distribution of 10,000 PR-AUC score differences. Under the null hypothesis, the two sets of PR-AUC scores would have been sampled from the same distribution, so that the distribution of differences would have mean zero; the alternate hypothesis, on the other hand, is that the respective means of the two sets of bootstrapped scores are different. Hence, to form a distribution able to test the null hypothesis, the observed difference was subtracted from each of the 10,000 PR-AUC score differences to form a mean-shifted distribution of differences, from which a p-value was finally obtained by performing a t-test. Note that an identical p-value, for a fold and pair of labelling methods, could have been obtained by performing a paired-samples t-test on the two mean-shifted distributions of PR-AUC scores, as opposed to differences, corresponding to each labelling method.

Ultimately, recalling from Section E that one of the 10 folds was excluded, for each pair of labelling methods, 9 p-values corresponding to 9 folds were obtained. Each p-value indicated whether the pair of PR-AUC scores corresponding to the pair of labelling methods for a given fold were significantly different. A Bonferroni correction [37, 38] was used to correct for the multiple comparisons problem, whereby the chance of falsely rejecting the null hypothesis by chance alone increases with the number of repetitions of a family of hypothesis tests testing the same hypothesis. Therefore, a 5% chosen significance level was divided by 9 folds to yield a 0.56% Bonferroni-corrected significance level. Using this Bonferroni-corrected significance, if the null hypothesis were true, 5% of the tests performed are expected to have their null hypothesis rejected by chance alone. Hence, out of the 9 p-values obtained to compare labelling methods, it is sufficient that one of them (i.e. 11% of the p-values) is below the Bonferroni-corrected significance level of 0.56% to conclude that the labelling methods are significantly different.

3) Inter/Intra-observer Tests: To perform the inter/intra-observer test metrics, two independent (1 MD from Royal Melbourne, 1 clinically trained LUS sonographer) experts were tasked with performing clinical labelling of the original consolidation/collapse (as described in Table II of the original patient dataset) and the healthy patient data was taken as is based on the provided medical reports. The labelling done by each expert comprised of a per frame binary scoring system where a score of 0 (no consolidation/collapse present) or a score of 1 (consolidation/collapse present) was assigned to all 12 patients from Table II. The scope of a given score of 0 includes frames that are conclusive for no pathology present and inconclusive or indeterminate for pathology not being present and is further described (on a per video basis) in Section A.

The inter/intra-observer agreement was calculated using a percent agreement given by the Cohen kappa score [39]. This metric is calculated by dividing the number of agreements between the observers with the total number of the scores (simpler percent agreement) but it takes into account with the consideration of taking into account the probability of chance agreement between observers.

III. RESULTS

For the specific classification threshold of 0.5 that was used, with respect to accuracy, the frame-based method performed best (accuracy: 90.1%), with the video-based methods performing from best (accuracy: 88.7%) to worst (accuracy: 87.2%) in the following order: sampled quaternary, sampled binary, all-or-nothing binary (Table V). However, given the class imbalance of the dataset (Table IV), accuracy reflects the ability of each method to produce true negatives. Indeed, the order of accuracies from the highest performing frame-based method to the lowest performing all-or-nothing binary method is identical to the order from highest to lowest of the percentage of true negatives produced (Table IV). Hence F1 score is a more suitable metric for evaluation. For the specific classification used, with respect to F1 score, the video-based methods (inaccurately supervised learning methods) outperformed the frame-based method (supervised learning method) with the sampled quaternary video-based method (F1 score: 67%) performing best and the frame-based method (F1 score: 55%) performing worst (Table V).

When considering the class imbalance of the dataset, it is particularly important to calibrate the classification threshold [40]. Hence, other Table V metrics are less informative than the PR-AUC score, a classification threshold-independent metric. With respect to PR-AUC score, the Table V methods performed from best to worst in the same order as they performed for F1 score with the sampled quaternary video-based method performing best (PR-AUC score: 73%) and the frame-based method performing worst (PR-AUC score: 60%). Additionally, with respect to PR-AUC score, the sampled quaternary method outperformed the sampled binary method by 11% and the all-or-nothing binary method by 9% (Table V).

The precision-recall curves of the best and worst folds are judged using PR-AUC score and given by Figure 6. Recall that PR-AUC score is sensitive to the ratio between consolidation containing ground truth frames to the total number of the frames, which defines the precision-recall curve for a classifier with no skill. This ratio across folds (mean/std (%): 11.6 ± 8.1) has maximum 29.1% and minimum 1.4%.

As described in Section H2, a Bonferroni-corrected significance level of 0.56% was used to assess for each pair of labelling methods,

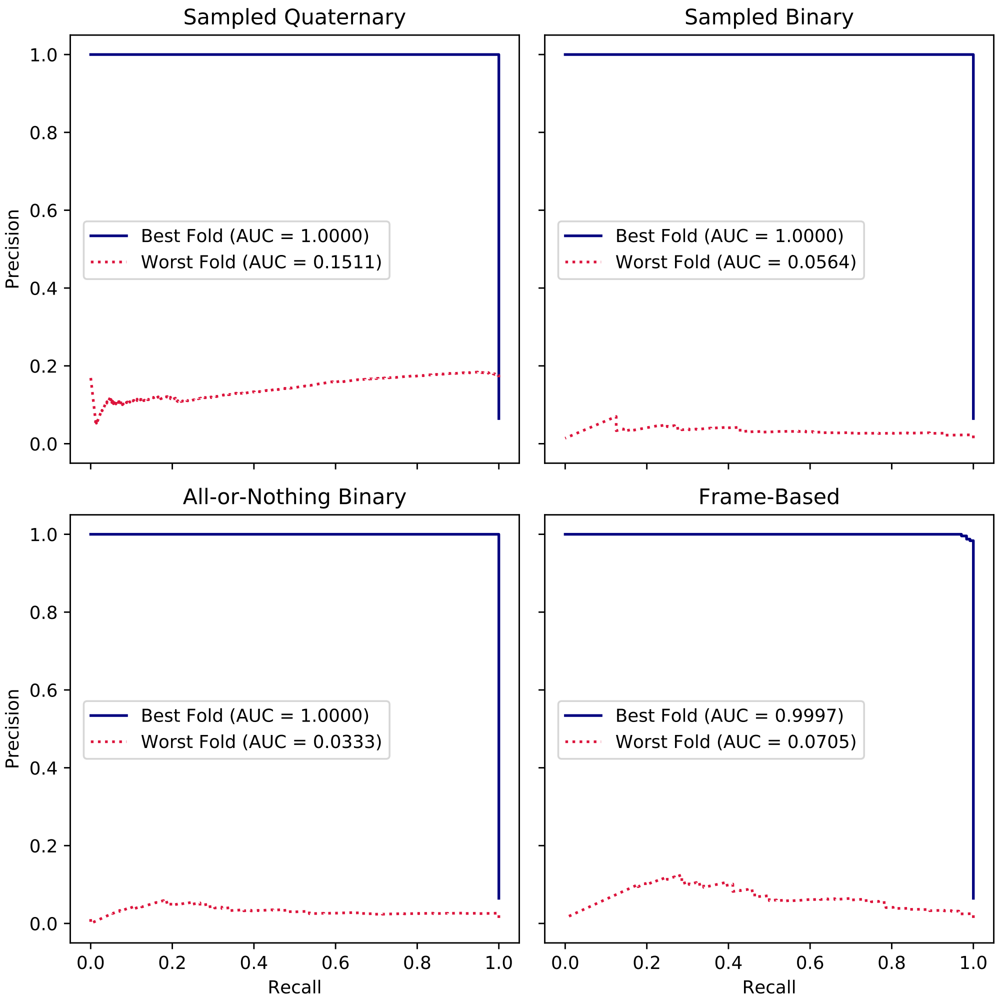


Fig. 6. The best and worst fold test set precision-recall curves across the 10 folds for which videos labelled N (Section II.A2) were excluded from the training, and for which videos labelled Y* were excluded from the test set. These are given with their associated average precisions given by the Area Under Curve (AUC) scores, for models trained using the various labelling methods: the sampled quaternary and binary video-based methods, the all-or-nothing binary video-based method, and the frame-based method. One of the 10 folds, which contained only healthy frames, was excluded from the results.

the 9 p-values corresponding to 9 folds. In all cases, the percentages of folds for which the null hypothesis was rejected were significantly greater than the 5% rate expected due to chance if the scores corresponding to a pair of labelling methods were insignificantly different. More concretely, the sampled quaternary method was found to be significantly different compared to the sampled binary method (mean/std p: 0.12 ± 0.29) for 50% of folds, the all-or-nothing binary method (mean/std p: 0.17 ± 0.29) for 50% of folds, and the frame-based method (mean/std p: 0.17 ± 0.3) for 60% of folds. Additionally, the sampled binary method was found to be significantly different compared to the all-or-nothing binary method (mean/std p: 0.24 ± 0.26) for 30% of folds, and the frame-based method (mean/std p: 0.16 ± 0.27) for 60% of folds. Finally, the frame-based method was found to be significantly different from the all-or-nothing binary-based method (mean/std p: 0.19 ± 0.36) for 70% of folds.

The metrics for the analysis of inter-observer agreement is given by Table VI. In the columns under the heading *data after criteria*, the inter-observer metrics (viz. Cohen kappa, % agreement) are used to show the agreement between two experts in labelling frames associated with imaging patterns from the consolidation/collapse pathology. These metrics are calculated before and after the clinical criteria had been applied to demonstrate to what degree the experts agree on consolidation/collapse frames labelled Y (high image quality, conclusively diagnosed pathology, clear anatomical markers), labelled Y* (conclusively diagnosed pathology, exactly one of: unclear anatomical markers or lower image quality), and labelled N (e.g. lower image quality, inconclusively diagnosed pathology, unclear anatomical markers). The *expert 1 / algorithm* and *expert 2 / algorithm* rows show the percent agreements of each expert's labels, and the classifier's predictions for the test set, which includes a mixture of healthy and unhealthy patients for each of the labelling methods. Since the training of the algorithm excluded the frames that did not satisfy the clinical criteria and were therefore labelled N, the percent agreement and Cohen kappa scores are absent from these rows.

The quaternary method performs the best in terms of percent agreement with the worst performing method being the all-or-nothing video-based method (Table VI). This metric along with the other metrics (e.g. PR-AUC, accuracy) show that the performance of the algorithm is at least on par if not at certain times slightly better than the trained experts after the application of the data exclusion criteria and the clinical criteria.

TABLE IV

THE MEAN/STD TEST-SET CONFUSION MATRIX FOR THE 10-FOLD CROSS VALIDATION TEST-SET RESULTS WITH VIDEOS LABELLED N (SECTION II.A2) EXCLUDED. POSITIVES CORRESPOND TO FRAMES FROM LUS VIDEOS THAT CONTAIN SIGNATURES OF CONSOLIDATIONS (AND ASSOCIATED IMAGING PATTERNS) WHILE NEGATIVES CORRESPOND TO FRAMES THAT DO NOT.

	Mean / std (%) with Y* frames excluded from the test set			
Method	**TP**	**FP**	**FN**	**TN**
Frame-based	5.68 ± 4.66	5.93 ± 7.18	4.08 ± 5.52	84.31 ± 28.82
All-or-nothing binary video based	6.75 ± 3.90	9.74 ± 13.80	3.01 ± 3.46	80.50 ± 31.80
Sampled binary video based	6.73 ± 3.98	9.51 ± 11.43	3.03 ± 3.52	80.73 ± 30.94
Sampled quaternary video based	7.59 ± 4.08	8.16 ± 11.72	2.17 ± 3.42	82.08 ± 33.69

One of the 10 folds, which contained only healthy frames, was excluded from the results. Here Y frames are frames Y* videos (Section II.A2).*

TABLE V

THE MEAN/STD TEST-SET METRICS FOR THE 10-FOLD CROSS VALIDATION, FOR WHICH VIDEOS LABELLED N (SECTION II.A2) WERE EXCLUDED.

	Mean / std (%) with Y* frames excluded from the test set				
Method	**PR-AUC**	**Recall**	**Precision**	**F1-Score**	**Accuracy**
Frame-based	60.08 ± 39.38	63.28 ± 36.62	53.01 ± 37.44	54.69 ± 38.02	90.18 ± 9.35
All-or-nothing binary video based	64.37 ± 39.32	69.18 ± 28.12	56.27 ± 37.86	59.10 ± 34.96	87.21 ± 17.07
Sampled binary video based	62.39 ± 40.09	69.22 ± 28.30	52.75 ± 36.37	55.92 ± 32.31	87.41 ± 14.97
Sampled quaternary video based	73.34 ± 30.37	83.67 ± 23.62	59.26 ± 28.14	66.78 ± 25.13	88.73 ± 15.87

One of the 10 folds, which contained only healthy frames, was excluded from the results. Here Y frames are frames Y* videos (Section II.A2).*

TABLE VI

METRICS FOR THE INTER-OBSERVER AGREEMENT ANALYSIS.

	Cohen kappa score / % agreement			
Comparison	**Data after criteria**			
Expert 1 / Expert 2	0.537 [Y/Y*/N] (0.805) Frame-based	0.956 [Y] (0.99) All-or-nothing video	0.552 [Y*] (0.91) Sampled binary video	0.149 [N] (0.58) Sampled quaternary video

Expert 1 / Algorithm	{Y/Y*}
	(90.088)
	{Y}
	(91.106)
		{Y/Y*}
		(87.626)
		{Y}
		(89.154)
			{Y/Y*}
			(87.823)
			{Y}
			(89.240)
				{Y/Y*}
				(89.791)
				{Y}
				(90.994)

Expert 2 / Algorithm	{Y/Y*}
	(90.144)
	{Y}
	(91.602)
		{Y/Y*}
		(87.767)
		{Y}
		(89.186)
			{Y/Y*}
			(87.835)
			{Y}
			(89.202)
				{Y/Y*}
				(89.956)
				{Y}
				(91.026)

Key: (% agreement), [video quality], {clinical criteria (unhealthy/healthy dataset)}

IV. DISCUSSION

In this paper, we utilised an automated deep learning-based approach that identifies consolidation/collapses in LUS images to aid in the diagnosis of late stages of COVID-19 induced pneumonia. Here we extend our previous work on pleural effusion pathology classification [23] by proposing an improvement to its video-based method, namely, through our sampled quaternary video-based method. We have evaluated this quaternary method by comparing it with the frame-based method and video-based method of our previous pleural effusion work and with the sampled binary video-based approach.

There is a trade-off between the quantity and quality of training data: By increasing the quality of the dataset by excluding videos, we reduce our dataset size. This trade-off was managed by excluding N but not Y* video frames from the training set, preferring instead of excluding only Y* from the test set. This exclusion of Y* from the test set was done because, in order to evaluate how well the compared classifiers fare against the challenge of label noise, the test sets ground truths must be virtually free of it. For images that do not fall under the Y criteria (Section A2), the inter-observer agreement varies greatly as the diagnosis of a given frame with consolidation/collapse gets more inconclusive as the contributing factors begin to accumulate These contributing factors range from varying image quality, location of key anatomical markers, artefacts hindering or obstructing the associated imaging pattern found with consolidation/collapse, and improper or poor placement of US probe resulting in unusable or highly questionable image.

While it is hard to gauge the performance of the model with respect to false positives and false negatives, since those metrics are with respect to a specific classification threshold of 0.5, domain knowledge may be leveraged. In the case of false positives, for which the algorithm predicted the frame as containing consolidation/collapse when there was none, the algorithm sometimes labels possible artefacts that resemble consolidation/collapse or liver like features if located in the RPL scanning region. A possible approach to addressing this issue is by providing the anatomical information of the liver and accounting for that during the trainings.In the case of false negatives, for which the algorithm predicted the frame as being free from consolidation/collapse when it was instead present, the misclassified frames were sometimes drawn from LUS videos for which the consolidation/collapse was hidden beneath an inflated lung and was only visible when the patient exhaled. This limitation could be addressed similarly to the false positive case, by accounting for patient inhales/exhales during training. Alternately, patient breathing rhythm could be accounted for automatically by an algorithm that accounts for the temporal relationship between frames.

To evaluate our classifiers in a classification-threshold independent manner, the PR-AUC score was used in place of the more common ROC-AUC score. This is because the dataset employed had an imbalanced class distribution [35]. The imbalanced class problem was exacerbated by the fact that in the case of the quaternary labelling method, there are three decision thresholds to be calibrated as opposed to one. While this is a limitation in our case, it may be argued that in cases where decision threshold calibration is feasible, the extra degrees of freedom of the quaternary method allows a more fine-tuned threshold calibration compared to the binary methods.

The fact that the sampled quaternary method outperformed the sampled binary method with respect to the PR-AUC score is expected, and in fact this quaternary method performed best overall in this respect. This is because the quaternary approach performs a smoother transition between healthy and non-healthy classifications by providing 4 classes rather than 2. This increase in class granularity may be limiting the maximum ascent in training loss per training iteration due to the inherent error of video-based labelling in comparison to frame-based labelling. More concretely, the fact that the overall loss function used incorporated a SORD cross-entropy loss function as one of its terms allows long-distance errors to be penalised harsher than low-distance errors. This is beneficial because, just as some predicted severity scores are closer to the true severity score in [22], some predicted sampled quaternary video-based method labels are more representative of the true number of frames with signatures of consolidation/collapse than others. This contrasts with cases where the classes are independent of each other.

In our findings, the inaccurately supervised learning of the video-based methods outperformed the supervised learning of the frame-based method with respect to PR-AUC score. This may be explained in terms of the bias-variance trade off, with the video-based methods shifting the trade off towards bias and the frame-based method shifting it towards variance. Indeed, for the video-based methods, a single label is being assigned to all frames of the video, so that the degrees of freedom the model has to overfit the noisy labelling data is reduced. Instead, a more uniform labelling noise is introduced across the frames of a video, which may be reducing the over-confidence of the video-based classifier has on any single frame. The quaternary method may be facilitating this reduction of over-confidence by preventing an overly high degree of noise from being injected uniformly across the labels of a video, as may be the case for the binary approaches. In this sense, our quaternary method is similar to label smoothing, a regularisation method that reduces over-confidence by smoothing labels and label noise, which been shown to be effective for datasets with incorrect labels present. Note, however, that the frame-based method still performed best in terms of accuracy, which may be due to the loss function used optimising error as opposed to a metric similar to PR-AUC score or F1 score that is more suitable for class imbalance scenarios.

While the sampled quaternary method outperformed the all-or-nothing binary method, there is a trade-off in terms of labelling effort. If medical reports are readily available, then the all-or-nothing method would not require additional labelling effort or clinical expertise. Hence, if the reduction in labelling time is of a higher priority than classification accuracy, then the all-or-nothing method may be preferred. The labelling effort of the sampled quaternary and sampled binary methods, on the other hand, are identical: this is because the labelling effort depends only on the percentage of frames sampled (10% in our case). Indeed, the sampled binary and quaternary methods are flexible in the sense that the trade-off between labelling effort and classification accuracy is easy to adjust: the higher the percentage of frames sampled, the higher the labelling effort.

The key limitation of our work was the limited size of high-quality training examples. Future work could address this limitation through transfer learning, a more sophisticated approach to data augmentation than the online data augmentation that was used, or by applying our classifier to a larger and higher quality dataset. Each of these approaches mitigate overfitting, while the latter approach reduces the dataset label noise. Since overfitting and label noise are posited as the reason why our video-based approaches outperform the frame-based method, the suggested future work could provide evidence that our video-based approaches indeed mitigate label noise or overfitting in a manner similar to label smoothing. Additionally, due to the large number of healthy training examples and limitations on expert time, the Section A2 Y/Y*/N categorisation was not applied to the healthy examples. Future work could apply this categorisation to both the healthy and unhealthy videos.

V. CONCLUSION

Our work provides a tool for automatic consolidation/collapse identification of LUS video frames during point-of-care testing. Out of the labelling methods considered, the video-based methods were intended to reduce labelling effort while minimising the resulting loss of accuracy.

The video-based methods outperformed the frame-based method. This may be a result of overfitting due to variance due to our small dataset size or label noise. Specifically, our video-based methods may be more robust label noise and variance than the frame-based method. That is, in the bias-variance trade-off the frame-based method shifts the trade-off towards variance while the video-based method shifts the trade-off towards bias. It is expected that if the classifier was run on a larger dataset of higher quality, the frame-based method would outperform the video-based methods. However, this must be confirmed through future work.

The best performing method was the sampled quaternary method, which employed a novel training approach using four classes, and performed better than the medical report based all-or-nothing method of [23]. However, if medical reports corresponding to LUS videos are readily available, then the all-or-nothing method may be preferred for scenarios where a reduction in labelling time is prioritised higher than classification accuracy.

References

[1] S. Kulkarni, B. Down, and S. Jha, ”Point-of-care lung ultrasound in intensive care during the COVID-19 pandemic.”

[2] G. S. Shrestha, D. Weeratunga, and K. Baker, ”Point-of-Care Lung Ultrasound in Critically ill Patients.”

[3] G. Soldati, R. Smargiassi A Auid- Orcid: — Fau - Inchingolo, R. A.-O. Inchingolo, D. A.-O. Buonsenso, D. F. Perrone T Fau - Briganti, D. F. Briganti, et al., ”Proposal for International Standardization of the Use of Lung Ultrasound for Patients With COVID-19: A Simple, Quantitative, Reproducible Method.”

[4] M. Allinovi, A. Parise, M. Giacalone, A. Amerio, M. Delsante, A. Odone, et al., ”Lung Ultrasound May Support Diagnosis and Monitoring of COVID-19 Pneumonia,” Ultrasound Med Biol, vol. 46, pp. 2908-2917, 2020.

[5] T. A.-O. Perrone, G. Soldati, L. Padovini, A. Fiengo, G. Lettieri, U. Sabatini, et al., ”A New Lung Ultrasound Protocol Able to Predict Worsening in Patients Affected by Severe Acute Respiratory Syndrome Coronavirus 2 Pneumonia.”

[6] N. Bonadia, A. Carnicelli, A. Piano, D. Buonsenso, E. Gilardi, C. Kadhim, et al., ”Lung Ultrasound Findings Are Associated with Mortality and Need for Intensive Care Admission in COVID-19 Patients Evaluated in the Emergency Department,” Ultrasound Med Biol, vol. 46, pp. 2927-2937, 2020.

[7] K. Yasukawa and T. Minami, ”Point-of-Care Lung Ultrasound Findings in Patients with COVID-19 Pneumonia.”

[8] e. a. David Canty, ”FUSE Lung Ultrasound course (Lung Ultrasound Diagnosis Tutorial), Focused Lung Ultrasound Image Quality Scoring (LUQS) Tool, Focused Lung Ultrasound Interpretation Score (LUIS),” U. o. M. Royal Melbourne Hospital, Ed., ed, 2020.

[9] V. A.-O. Manivel, A. Lesnewski, S. Shamim, G. Carbonatto, and T. Govindan, ”CLUE: COVID-19 lung ultrasound in emergency department.”

[10] P. A.-O. Pietersen, K. R. Madsen, O. Graumann, L. Konge, B. U. Nielsen, and C. B. Laursen, ”Lung ultrasound training: a systematic review of published literature in clinical lung ultrasound training.”

[11] A. Y. Denault, S. Delisle, D. Canty, A. Royse, C. Royse, X. C. Serra, et al., ”A proposed lung ultrasound and phenotypic algorithm for the care of COVID-19 patients with acute respiratory failure,” Can J Anaesth, vol. 67, pp. 1393-1404, 2020.

[12] F. Sadik, A. G. Dastider, and S. A. Fattah, ”SpecMEn-DL: spectral mask enhancement with deep learning models to predict COVID-19 from lung ultrasound videos,” Health Information Science and Systems, vol. 9, p. 28, 2021/07/09 2021.

[13] G. Muhammad and M. Shamim Hossain, ”COVID-19 and Non-COVID-19 Classification using Multi-layers Fusion From Lung Ultrasound Images,” Information Fusion, vol. 72, pp. 80-88, 2021/08/01/ 2021.

[14] W. Xue, C. Cao, J. Liu, Y. Duan, H. Cao, J. Wang, et al., ”Modality alignment contrastive learning for severity assessment of COVID-19 from lung ultrasound and clinical information,” Medical Image Analysis, vol. 69, p. 101975, 2021/04/01/ 2021.

[15] J. Chen, C. He, J. Yin, J. Li, X. Duan, Y. Cao, et al., ”Quantitative Analysis and Automated Lung Ultrasound Scoring for Evaluating COVID-19 Pneumonia With Neural Networks,” IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 68, pp. 2507-2515, 2021.

[16] M. La Salvia, G. Secco, E. Torti, G. Florimbi, L. Guido, P. Lago, et al., ”Deep learning and lung ultrasound for Covid-19 pneumonia detection and severity classification,” Computers in Biology and Medicine, vol. 136, p. 104742, 2021/09/01/ 2021.

[17] R. Arntfield, B. VanBerlo, T. Alaifan, N. Phelps, M. White, R. Chaudhary, et al., ”Development of a convolutional neural network to differentiate among the etiology of similar appearing pathological B lines on lung ultrasound: a deep learning study,” BMJ Open, vol. 11, p. e045120, 2021.

[18] G. Baloescu C Fau - Toporek, S. Toporek G Fau - Kim, K. Kim S Fau - McNamara, R. McNamara K Fau - Liu, M. M. Liu R Fau - Shaw, R. L. Shaw Mm Fau - McNamara, et al., ”Automated Lung Ultrasound B-Line Assessment Using a Deep Learning Algorithm.”

[19] L. Carrer, E. Donini, D. Marinelli, M. Zanetti, F. Mento, E. Torri, et al., ”Automatic Pleural Line Extraction and COVID-19 Scoring From Lung Ultrasound Data,” IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 67, pp. 2207-2217, 2020.

[20] R. J. G. v. Sloun and L. Demi, ”Localizing B-Lines in Lung Ultrasonography by Weakly Supervised Deep Learning, ¡italic¿In-Vivo¡/italic¿ Results,” IEEE Journal of Biomedical and Health Informatics, vol. 24, pp. 957-964, 2020.

[21] J. Born, G. Brändle, M. Cossio, M. Disdier, J. Goulet, J. e. e. Roulin, et al., ”POCOVID-Net: Automatic Detection of COVID-19 From a New Lung Ultrasound Imaging Dataset (POCUS),” ArXiv, vol. abs/2004.12084, 2020.

[22] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, et al., ”Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound,” IEEE Transactions on Medical Imaging, vol. 39, pp. 2676-2687, 2020.

[23] C.-H. Tsai, J. van der Burgt, D. Vukovic, N. Kaur, L. Demi, D. Canty, et al., ”Automatic deep learning-based pleural effusion classification in lung ultrasound images for respiratory pathology diagnosis,” Physica Medica, vol. 83, pp. 38-45, 2021/03/01/ 2021.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ”Rethinking the Inception Architecture for Computer Vision,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826, 2016.

[25] M. Lukasik, S. Bhojanapalli, A. K. Menon, and S. Kumar, ”Does label smoothing mitigate label noise?,” ArXiv, vol. abs/2003.02819, 2020.

[26] X. Cid, D. Canty, A. Royse, A. B. Maier, D. Johnson, D. El-Ansary, et al., ”Impact of point-of-care ultrasound on the hospital length of stay for internal medicine inpatients with cardiopulmonary diagnosis at admission: study protocol of a randomized controlled trial—the IMFCU-1 (Internal Medicine Focused Clinical Ultrasound) study,” Trials, vol. 21, p. 53, 2020/01/08 2020.

[27] D. Canty, K. Haji, A. Denault, and A. Royse, ”Lung Ultrasound in Anaesthesia and Critical Care Medicine,” in Perioperative Medicine – Current Controversies, K. Stuart-Smith, Ed., ed Cham: Springer International Publishing, 2016, pp. 345-389.

[28] ”A randomised trial of focused cardiac, lung, and femoral and popliteal vein ultrasound on the length of stay in internal medicine admissions with a cardiopulmonary diagnosis,” Melbourne (NSW): NHMRC Clinical Trials Centre, Royal Melbourne Hospital (Australia); 2019 - Identifier ACTRN12618001442291. http://www.ANZCTR.org.au/ACTRN12618001442291.aspx2019.

[29] Z.-H. Zhou, ”A brief introduction to weakly supervised learning,” National Science Review, vol. 5, pp. 44-53, 2018.

[30] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, ”Spatial Transformer Networks,” in NIPS, 2015.

[31] C. Winship and R. D. Mare, ”Regression Models with Ordinal Variables,” American Sociological Review, vol. 49, pp. 512-525, 1984.

[32] R. Díaz and A. Marathe, ”Soft Labels for Ordinal Regression,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4733-4742.

[33] P. S. Efraimidis and P. G. Spirakis, ”Weighted random sampling with a reservoir,” Information Processing Letters, vol. 97, pp. 181-185, 2006/03/16/ 2006.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., ”PyTorch: An Imperative Style, High-Performance Deep Learning Library,” ed, 2019.

[35] T. Saito and M. Rehmsmeier, ”The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” PLoS One, vol. 10, pp. e0118432-e0118432, 2015.

[36] K. Boyd, K. H. Eng, and C. D. Page, ”Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals,” Berlin, Heidelberg, 2013, pp. 451-466.

[37] H. Abdi, ”Bonferroni and Šidák corrections for multiple comparisons,” Encyclopedia of measurement and statistics, vol. 3, pp. 103-107, 2007.

[38] W. Haynes, ”Bonferroni Correction,” in Encyclopedia of Systems Biology, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds., ed New York, NY: Springer New York, 2013, pp. 154-154.

[39] M. L. McHugh, ”Interrater reliability: the kappa statistic,” Biochem Med (Zagreb), vol. 22, pp. 276-282, 2012.

[40] F. J. Provost, ”Machine Learning from Imbalanced Data Sets 101,” 2008.