Contents lists available at ScienceDirect

Physica Medica

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

Automatic deep learning-based pleural effusion classification in lung ultrasound images for respiratory pathology diagnosis

Chung-Han Tsai [a], Jeroen van der Burgt [a], Damjan Vukovic [a,b], Nancy Kaur [c], Libertario Demi [d], David Canty [e], Andrew Wang [e], Alistair Royse [e], Colin Royse [e,f], Kavi Haji [e], Jason Dowling [g], Girija Chetty [c], Davide Fontanarosa [a,b,*]

[a] *School of Clinical Sciences, Queensland University of Technology, Gardens Point Campus, 2 George St, Brisbane 4000, QLD, Australia*
[b] *Centre for Biomedical Technologies (CBT), Queensland University of Technology, Brisbane, Queensland, Australia*
[c] *School of IT & Systems, Faculty of Science and Technology, University of Canberra, 11 Kirinari Street, Bruce 2617, ACT, Australia*
[d] *Ultrasound Laboratory Trento, Department of Information Engineering and Computer Science, University of Trento, Italy*
[e] *Department of Surgery (Royal Melbourne Hospital), University of Melbourne, Royal Parade, Parkville 3050, VIC, Australia*
[f] *Outcomes Research Consortium, Cleveland Clinic, Cleveland, OH, USA*
[g] *CSIRO Health and Biosecurity, The Australian eHealth Research Centre, Australia*

ARTICLE INFO

Keywords:
Lung ultrasound
Machine learning
Weakly supervised deep learning
Pleural effusion diagnosis, COVID-19

ABSTRACT

Lung ultrasound (LUS) imaging as a point-of-care diagnostic tool for lung pathologies has been proven superior to X-ray and comparable to CT, enabling earlier and more accurate diagnosis in real-time at the patient's bedside. The main limitation to widespread use is its dependence on the operator training and experience. COVID-19 lung ultrasound findings predominantly reflect a pneumonitis pattern, with pleural effusion being infrequent. However, pleural effusion is easy to detect and to quantify, therefore it was selected as the subject of this study, which aims to develop an automated system for the interpretation of LUS of pleural effusion. A LUS dataset was collected at the Royal Melbourne Hospital which consisted of 623 videos containing 99,209 2D ultrasound images of 70 patients using a phased array transducer. A standardized protocol was followed that involved scanning six anatomical regions providing complete coverage of the lungs for diagnosis of respiratory pathology. This protocol combined with a deep learning algorithm using a Spatial Transformer Network provides a basis for automatic pathology classification on an image-based level. In this work, the deep learning model was trained using supervised and weakly supervised approaches which used frame- and video-based ground truth labels respectively. The reference was expert clinician image interpretation. Both approaches show comparable accuracy scores on the test set of 92.4% and 91.1%, respectively, not statistically significantly different. However, the video-based labelling approach requires significantly less effort from clinical experts for ground truth labelling.

1. Introduction:

Ultrasound imaging is routinely used for diagnostics and is particularly suitable for mass screening of a range of diseases. Amid the COVID-19 pandemic, lung ultrasound (LUS) has proven to be very useful in the diagnosis and management of patients presenting with suspected or confirmed COVID-19 respiratory infection [1–5]. The diagnostic accuracy of LUS in the diagnosis of respiratory pathology, including COVID-19 disease, in expert hands, surpasses chest X-ray and approaches that of computed tomography (CT) [6]; but, unlike CT, which is the current standard investigation for COVID-19, LUS does not require the patient's transfer to the radiology department due to portability and ease of access to the ultrasound machine. Before diagnosis at initial presentation to healthcare, in particular in resource poor areas with limited or no access to COVID-19 testing, LUS can substitute for other standard investigations as a rapid, affordable and non-invasive test that may also reduce infection risk. The identification and spatial localization within the lungs of several different pathologies, among which pleural effusion, atelectasis, consolidations and interstitial syndrome, are paramount for an accurate and reliable diagnosis. These may present differently depending on the stage of disease progression. For example, pleural thickening in early stages versus sub-pleural thickening in later stages of

[*] Corresponding author at: School of Clinical Sciences, Queensland University of Technology, Gardens Point Campus, 2 George St, Brisbane 4000, QLD, Australia.
E-mail address: d3.fontanarosa@qut.edu.au (D. Fontanarosa).

https://doi.org/10.1016/j.ejmp.2021.02.023
Received 3 December 2020; Received in revised form 9 February 2021; Accepted 22 February 2021
Available online 8 March 2021

COVID-19 disease. The current problem preventing widespread use of LUS for these applications is that image acquisition and image interpretation are complex and depend on the competency in LUS of the sonographers. The training necessary to reach adequate competency is lengthy and is currently limited by COVID-related physical distancing restrictions around the world. As a result, the availability of these professionals is insufficient to cover the growing demand so, despite its popularity and advantages, LUS has not become more extensively adopted.

In this work we focused on the automatic detection of one of these pathologies potentially involved in COVID-19 induced pneumonitis, pleural effusion. Pleural effusion typically manifests itself as an anechoic area in the intrapleural space (i.e., an area of ultrasound transmission without reflections within the space) [7]. Diagnosis of pleural effusion is important as drainage of fluid can relieve respiratory failure and microbiological culture can guide effective antibiotic treatment. In addition to improving accuracy of diagnosis or exclusion of effusion, LUS can reduce the risk of potentially lethal complications from blind insertion of percutaneous pleural catheters by visualisation of the anatomy of the chest wall, thorax and abdomen, and enabling real-time guidance of the sharp metal drain introducer [8].

The aim of this study was to develop an algorithm that interprets LUS datasets to identify pleural effusion, with comparable or improved accuracy compared to clinical standards in order to allow faster diagnosis and a more robust result irrespective of the competence of the sonographer performing the examination. Machine learning applications for ultrasound imaging interpretation, especially in diagnostics, have grown significantly in recent years [9]. The increasing interest in LUS during the COVID-19 pandemic has sparked the production of several new works on the automation of the diagnosis using deep learning [10] and/or machine learning [11]. But very sparse literature is in general available on automatic detection and quantification of pleural effusion. Notably, Kulhare et al. [12] explicitly mention the pathology, among the others investigated, in their work which is based on short videos of in-vivo swine models; the results they report, though, are based on a smaller dataset, and are significantly worse than the ones produced in our work.

The performances of two labelling methods, frame-based and scanning position-based, have been compared in an effort to reduce the workload on clinical experts that are required to perform the labelling.

2. Material and Methods:

2.1. Dataset

The dataset used for this work was obtained from a study at the Royal Melbourne Hospital (Melbourne Health Human Research Ethics Committee approval HREC/18/MH/269, trial registration: http://www.ANZCTR.org.au/ACTRN12618001442291.aspx) which includes LUS data from 70 patients with either image evidence of pleural effusion (39 patients) and normal lungs (31 patients). For each patient, LUS videos were acquired at six different anatomical zones (scanning positions), three per side, following a standardized LUS protocol. These include right anterior (RANT), left anterior (LANT), right posterior upper (RPU), left posterior upper (LPU), right posterior lower (RPL), and left posterior lower (LPL) zones [13], as shown in Fig. 1. Each patient received at least six ultrasound videos which corresponded to the three anatomical zones per side. In total 623 ultrasound videos were acquired using a Sonosite X-Porte ultrasound imaging system (Fujifilm, Bothell, WA, USA) with a 1–5 MHz phased array transducer probe (SonoSite X-Porte rP19xp) which resulted in 99,209 2D ultrasound images. A phased array probe was used as it has superior depth penetration to a microconvex probe [14], allowing complete visualisation of large effusions and hence estimation of volume, which will be developed in the algorithm. Although phased array probes have less resolution than microconvex and curvilinear probes, resolution is not as important for assessment of effusion than it is for pleural pathology such as interstitial syndrome or pneumothorax [14]. This study was restricted to the detection of anatomical findings, and not of lung ultrasound artifacts. All images were stored using commercial DICOM archiving software (Synapse Cardiovascular, FujiFilm Australia, Murarrie QLD 4172), and de-identified images were used for analysis.

A patient's lungs were considered normal (free of pleural effusion) if none of the associated ultrasound videos from any of the anatomical zones showed clinical signs of pleural effusion and categorized as "Normal". However, if at least one of the ultrasound videos demonstrated pleural effusion, the lungs were categorized as "Abnormal". From the total number of ultrasound images, 20,120 images (20%) showed signs of pleural effusion (see Table 1 and Table 4).

2.2. Pre-processing

An open-source DICOM processing package [15] was used to extract the original pixel data from the compressed DICOM format. The decompression of the pixel data involved the conversion of its photometric interpretation from 'YBR_FULL_422' colour space to RGB colour space [16]. In the ultrasound images a variety of overlays, including text (with no patient or institution identification being present), watermarks and trademarks from the ultrasound imaging system were present. All

Table 1
Ultrasound image dataset overview.

Condition	Number of patients	Number of videos	Number of images	Number of pleural effusion videos
Normal	31	252	44,211	n/a
Abnormal	39	371	54,998	141
Total	**70**	**623**	**99,209**	**141**

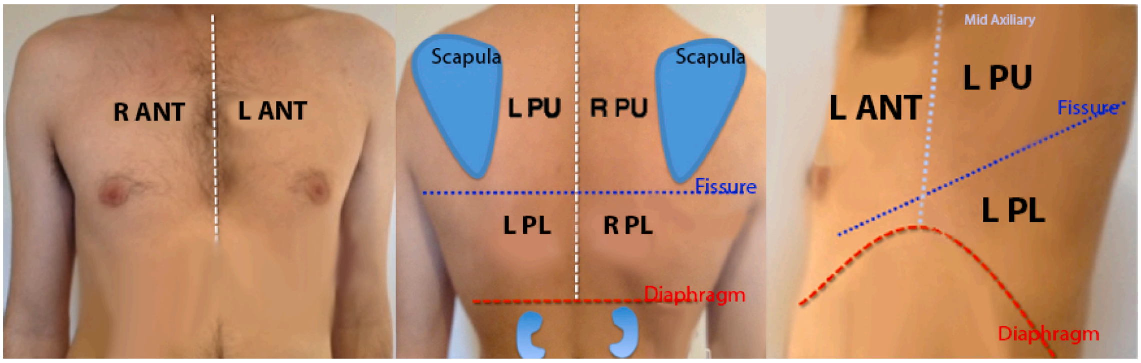

Fig. 1. (Left) Right anterior and left anterior; (Middle) LPU, LPL, RPU, and RPL; (Right) Lateral view for LANT, LPU, and LPL (Need to request for the authorization of the picture use).

overlays, both inside and outside the ultrasound sectors were replaced by black (background) pixels. The last pre-processing operation was to crop the images to the size of the minimal rectangle containing the ultrasound sector to reduce the image data size before feeding it into the deep learning model.

2.3. Labelling

To provide the deep learning model with ground truth labels for training purposes, a video- and a frame-based labelling approach were used following a weakly supervised and supervised approach respectively.

For each patient, an iLungScan™ report (developed by the Ultrasound Education Group at the University of Melbourne, https://ilungscan.com), validated by LUS experts and clinicians from the University of Melbourne, was available to indicate whether signs of pleural effusion were present in each of the six scanning positions. If present, all images in the videos belonging to that particular scanning position were assigned a Score V(video)1. Otherwise, the images were assigned a Score V0.

For the frame-based labelling, each ultrasound image has been assigned a binary label indicating if it contains clinical signs of pleural effusion (Score F(rame)1) or not (Score F0) by a clinical expert.

2.4. Data split

The dataset was split into a training and a test set by randomly putting 90% of all patients (63 patients) in the training set and the remaining 10% (7 patients) into the test set. Both the training and test set had a mix of normal cases and pleural effusion patients, but it should be noticed that this does not necessarily ensure balanced classes of normal and pleural effusion videos or frames in each set. For this reason, a weighted random sampler [17,18] from PyTorch was applied to avoid adverse effects of the imbalanced classes. To perform cross validation, 10 folds of training and test sets were created in which each patient appeared at least once in the test set. The details of data split for video- and frame-based trainings for each fold are shown in Tables 2 and 3 respectively.

The class distribution for video-based and frame-based labelling approaches over the entire dataset is presented in Table 4. The video-based class distribution of ultrasound images with Score V1 and Score V0 was approximately 20% vs. 80% in the training set, as shown in Table 5. The frame-based class distribution (Score F1 vs. Score F0) in the training set was approximately 16% vs. 84%, as shown in Table 6.

2.5. Deep learning architecture

The deep learning model introduced in [19] was designed to assign a severity score to ultrasound images of COVID-19 patients. This model made use of a Regularised Spatial Transformer Network (Reg-STN) architecture consisting of convolutional neural networks (CNN) [20] for feature extraction and a Spatial Transformer Network (STN) [21] to perform weakly supervised localization of COVID-19 pathological signs. In our work, instead, the model was used for binary classification of pleural effusion, the specific architecture as shown in the green rectangle of Fig. 1 of [19] is described in detail in section IV.C of the same paper. The approach was weakly supervised because the labelling was performed at the video level, i.e. the experts labelled the video as including or not including pleural effusion, but did not specifically identify in which frame the pathology is present. It is important to note that, while alone it is not sufficient for a diagnosis, pleural effusion is one of the pathologies potentially associated with COVID-19, and its proper evaluation is an essential step for a reliable automatic detection of the viral infection.

In agreement with the approach proposed in [19], the Adam optimizer was used and the model was trained with a batch size of 64 for 120 epochs. The training of the model was performed on a Linux workstation with a dual Nvidia Titan RTX GPU (128 GB memory, Intel i9-9820X CPU @ 3.30 GHz x20, 24 GB dedicated GPU memory) (Lambda Labs, San Francisco, CA, USA).

2.6. Evaluations

The performance of the deep learning model for both frame-based and video-based labelling approaches was evaluated by calculating classification accuracy, precision, recall and F-score metrics. In addition, confusion matrices were generated. It is important to note that the model trained by the video-based labelling approach ('the video-based model') generated predictions on the frame-based level (images) and was then evaluated on the frame-based ground truth labels, i.e., not on the video-based ground truth labels it was trained with. In other words, the video-based approach is weakly supervised.

3. Results

Overall, the video-based labelling approach reached 91.12% mean accuracy in the test set over the 10-folds, and the frame-based labelling approach reached 92.38%, which was 1.26% higher than the video-based labelling approach. The F1-score of the frame-based and the video-based approaches were used to determine which folds had the best and worst performance: for both approaches, the same folds resulted either the best or the worst. The F1 scores of the best fold are 87.71% and 90.47% for the video-based approach and the frame-based approach, respectively, as shown in Table 7.

A T-test was performed on the accuracies of the two labelling approaches and produced a p-value of 0.422. Using a common significance level of 0.05, this means there is no statistically significant difference between the performance of the video-based and the frame-based labelling approaches.

Based on the F1-scores, the best and the worst folds for the test set

Table 2
Data split overview for videos in the training set and the test set for each fold. The total number of videos was 623 for each fold.

Fold #	Training set: Total (% out of 623)	Training set: Normal Patient (% out of the training set total)	Training set: Pleural Effusion Patient (% out of the training set total)	Test set: Total (% out of 623)	Test set: Normal Patient (% out of the test set total)	Test set: Pleural Effusion Patient (% out of the test set total)
0	563 (90.37%)	218 (38.72%)	345 (61.28%)	60 (9.63%)	34 (56.67%)	26 (43.33%)
1	566 (90.85%)	221 (39.05%)	345 (60.95%)	57 (9.15%)	31 (54.39%)	26 (45.61%)
2	546 (87.64%)	235 (43.04%)	311 (56.96%)	77 (12.36%)	17 (22.08%)	60 (77.92%)
3	558 (89.57%)	226 (40.5%)	332 (59.5%)	65 (10.43%)	26 (40.0%)	39 (60.0%)
4	566 (90.85%)	223 (39.4%)	343 (60.6%)	57 (9.15%)	29 (50.88%)	28 (49.12%)
5	561 (90.05%)	221 (39.39%)	340 (60.61%)	62 (9.95%)	31 (50.0%)	31 (50.0%)
6	561 (90.05%)	235 (41.89%)	326 (58.11%)	62 (9.95%)	17 (27.42%)	45 (72.58%)
7	561 (90.05%)	239 (42.6%)	322 (57.4%)	62 (9.95%)	13 (20.97%)	49 (79.03%)
8	561 (90.05%)	216 (38.5%)	345 (61.5%)	62 (9.95%)	36 (58.06%)	26 (41.94%)
9	564 (90.53%)	234 (41.49%)	330 (58.51%)	59 (9.47%)	18 (30.51%)	41 (69.49%)

Table 3
Data split overview for images in the training set and the test set for each fold. The total number of images was 99,209 for each fold.

Fold #	Training set: Total (out of 99,209)	Training set: Normal Patient (out of the training set total)	Training set: Pleural Effusion Patient (out of the training set total)	Test set: Total (out of 99,209)	Test set: Normal Patient (out of the test set total)	Test set: Pleural Effusion Patient (out of the test set total)
0	89,934 (90.65%)	38,763 (43.1%)	51,171 (56.9%)	9275 (9.35%)	5448 (58.74%)	3827 (41.26%)
1	90,153 (90.87%)	39,199 (43.48%)	50,954 (56.52%)	9056 (9.13%)	5012 (55.34%)	4044 (44.66%)
2	88,394 (89.1%)	39,498 (44.68%)	48,896 (55.32%)	10,815 (10.9%)	4713 (43.58%)	6102 (56.42%)
3	87,012 (87.71%)	38,633 (44.4%)	48,379 (55.6%)	12,197 (12.29%)	5578 (45.73%)	6619 (54.27%)
4	91,312 (92.04%)	39,514 (43.27%)	51,798 (56.73%)	7897 (7.96%)	4697 (59.48%)	3200 (40.52%)
5	89,738 (90.45%)	39,989 (44.56%)	49,749 (55.44%)	9471 (9.55%)	4222 (44.58%)	5249 (55.42%)
6	87,893 (88.59%)	40,510 (46.09%)	47,383 (53.91%)	11,316 (11.41%)	3701 (32.71%)	7615 (67.29%)
7	86,542 (87.23%)	40,791 (47.13%)	45,751 (52.87%)	12,667 (12.77%)	3420 (27.0%)	9247 (73.0%)
8	90,859 (91.58%)	38,651 (42.54%)	52,208 (57.46%)	8350 (8.42%)	5560 (66.59%)	2790 (33.41%)
9	91,044 (91.77%)	42,351 (46.52%)	48,693 (53.48%)	8165 (8.23%)	1860 (22.78%)	6305 (77.22%)

Table 4
Class distributions for video-based and frame-based labelling approaches.

Dataset	Number of images for video-based labelling approach	Number of images for frame-based labelling approach
Overall	**99,209**	**99,209**
Normal class (Score 0)	79,089 (80%)	83,061 (84%)
Pleural effusion class (Score 1)	20,120 (20%)	16,148 (16%)

Table 5
Class distribution of the training and test set for video-based labelling: number of images for each fold.

Fold #	Training set: Normal class (Score V0)	Training set: Pleural effusion class (Score V1)	Test set: Normal class (Score V0)	Test set: Pleural effusion class (Score V1)
0	71,036 (78.99%)	18,898 (21.01%)	8053 (86.82%)	1222 (13.18%)
1	71,383 (79.18%)	18,770 (20.82%)	7706 (85.09%)	1350 (14.91%)
2	70,758 (80.05%)	17,636 (19.95%)	8331 (77.03%)	2484 (22.97%)
3	69,642 (80.04%)	17,370 (19.96%)	9447 (77.45%)	2750 (22.55%)
4	72,251 (79.13%)	19,061 (20.87%)	6838 (86.59%)	1059 (13.41%)
5	71,053 (79.18%)	18,685 (20.82%)	8036 (84.85%)	1435 (15.15%)
6	70,653 (80.39%)	17,240 (19.61%)	8436 (74.55%)	2880 (25.45%)
7	69,655 (80.49%)	16,887 (19.51%)	9434 (74.48%)	3233 (25.52%)
8	71,930 (79.17%)	18,929 (20.83%)	7159 (85.74%)	1191 (14.26%)
9	73,440 (80.66%)	17,604 (19.34%)	5649 (69.19%)	2516 (30.81%)

Table 6
Class distribution of the training and test set for frame-based labelling: number of images for each fold.

Fold #	Training set: Normal class (Score F0)	Training set: Pleural effusion class (Score F1)	Test set: Normal class (Score F0)	Test set: Pleural effusion class (Score F1)
0	74,733 (83.1%)	15,201 (16.9%)	8328 (89.79%)	947 (10.21%)
1	75,508 (83.76%)	14,645 (16.24%)	7553 (83.4%)	1503 (16.6%)
2	74,156 (83.89%)	14,238 (16.11%)	8905 (82.34%)	1910 (17.66%)
3	72,998 (83.89%)	14,014 (16.11%)	10,063 (82.5%)	2134 (17.5%)
4	76,223 (83.48%)	15,089 (16.52%)	6838 (86.59%)	1059 (13.41%)
5	74,770 (83.32%)	14,968 (16.68%)	8291 (87.54%)	1180 (12.46%)
6	74,093 (84.3%)	13,800 (15.7%)	8968 (79.25%)	2348 (20.75%)
7	73,332 (84.74%)	13,210 (15.26%)	9729 (76.81%)	2938 (23.19%)
8	75,822 (83.45%)	15,037 (16.55%)	7239 (86.69%)	1111 (13.31%)
9	75,914 (83.38%)	15,130 (16.62%)	7147 (87.53%)	1018 (12.47%)

were selected to present the confusion matrices in the following Tables 8–11. In the best fold, the video-based labelling approach had 1881 true positives (3 less true positives than the frame-based approach) while it had 9789 true negatives (127 less true negatives than the frame-based approach). In the worst fold, the video-based labelling approach had 420 true positives (119 more true positives than the frame-based approach) while it had 6486 true negatives (261 less true negatives than the frame-based approach).

For the training set, the video-based labelling approach reached 99.97% mean accuracy. The frame-based labelling approach reached 99.93%, 0.04% lower compared to the video-based labelling approach. The F1-scores are both around 99.98% for the video-based approach and the frame-based approach, respectively, as shown in Table 12.

In Tables 13 and 14 the confusion matrices for the video-based and the frame-based labelling approach, evaluated on the training set, are presented. The frame-based labelling approach had 43,806 true positives (536 more true positives than the video-based approach) and 44,026 true negatives (332 more true negatives than the video-based approach).

Fig. 2 shows an example of LUS images where the prediction is different for the frame- and the video-based approach. Fig. 3 shows LUS

Table 7
The comparison of mean accuracy, F1-score, precision and recall between the video-based and the frame-based approach, evaluated on the test set.

Metrics	Video-based labelling approach	Frame-based labelling approach
Mean accuracy	91.1179%	92.3785%
Standard deviation	3.3525	3.1525
Accuracy (the best fold)	95.68%	96.75%
F1-score (the best fold)	87.71%	90.47%
Precision (the best fold)	87.29%	92.76%
Recall (the best fold)	88.14%	88.28%
Accuracy (the worst fold)	84.58%	86.30%
F1-score (the worst fold)	40.02%	34.98%
Precision (the worst fold)	38.85%	42.82%
Recall (the worst fold)	41.26%	29.57%

Table 8
Confusion matrix of the best fold for the video-based labelling approach, evaluated on the test set.

Confusion Matrix	Score 1 (Actual)	Score 0 (Actual)
Score 1 (Predicted)	TP: 1,881	FP: 274
Score 0 (Predicted)	FN: 253	TN: 9,789

Table 9
Confusion matrix of the best fold for the frame-based labelling approach, evaluated on the test set.

Confusion Matrix	Score 1 (Actual)	Score 0 (Actual)
Score 1 (Predicted)	TP: 1,884	FP: 147
Score 0 (Predicted)	FN: 250	TN: 9,916

Table 10
Confusion matrix of the worst fold for the video-based labelling approach, evaluated on the test set.

Confusion Matrix	Score 1 (Actual)	Score 0 (Actual)
Score 1 (Predicted)	TP: 420	FP: 661
Score 0 (Predicted)	FN: 598	TN: 6,486

Table 11
Confusion matrix of the worst fold for the frame-based labelling approach, evaluated on the test set.

Confusion Matrix	Score 1 (Actual)	Score 0 (Actual)
Score 1 (Predicted)	TP: 301	FP: 402
Score 0 (Predicted)	FN: 717	TN: 6,745

images for which the prediction is the same for both the frame- and the video-based approach. Each subfigure of Figs. 2 and 3 corresponds to each of the combinations of confusion matrix elements of the video- and frame-based approach on the test set in the best fold (Fold #3), as shown in Table 15. The corresponding numbers of LUS images are also indicated in the table for each of the eight confusion matrix element combinations. Table 15 presents the numbers of LUS images for each of the combinations of confusion matrix elements of the video- and frame-based approach on the training set.

Table 12
The comparison of accuracy and F1-score between the video-based and the frame-based approach, evaluated on the training set.

Metric	Video-based approach with the frame-based ground truth labels	Frame-based approach
Mean Accuracy	99.97%	99.93%
Accuracy (the best fold)	99.99%	99.95%
Precision (the best fold)	99.98%	99.95%
Recall (the best fold)	99.99%	99.96%
F1-score (the best fold)	99.99%	99.95%

Table 13
Confusion matrix of the best fold for the video-based labelling approach, evaluated on the training set.

Confusion Matrix	Score 1 (Actual)	Score 0 (Actual)
Score 1 (Predicted)	TP: 43,270	FP: 7
Score 0 (Predicted)	FN: 5	TN: 43,694

Table 14
Confusion matrix of the best fold for the frame-based labelling approach, evaluated on the training set.

Confusion Matrix	Score 1 (Actual)	Score 0 (Actual)
Score 1 (Predicted)	TP: 43,806	FP: 23
Score 0 (Predicted)	FN: 17	TN: 44,026

4. Discussion

In this work we have demonstrated the first step to automating LUS image evaluation, in particular introducing automatic detection of pleural effusion, both using a video-based and a frame-based labelling approach. This is part of a larger project to provide potentially untrained operators with a reliable diagnostic tool for COVID-19 induced respiratory disease.

This is challenging because a combination of intercurrent pathologies including pleural effusion, atelectasis (collapsed lung), consolidation, interstitial syndrome and pneumothorax need to be identified as these can be alternative cause of a patient's respiratory symptoms. With our initial effort, we focused on pleural effusion because besides developing a tool to address the COVID-19 pneumonia diagnosis, we believe an accurate, timely and quantitatively reliable estimate of effusion is also pertinent for improved patient care. Moreover, pleural effusion is the easiest pathology to identify on LUS. In COVID-19 lung pathology, interstitial syndrome and bronchopneumonia are significantly more common but have more subtle ultrasound characteristics. Pleural effusion, by contrast, is far more easily distinguished by its harsh contrast and clear interface to adjacent soft tissues. As mentioned, LUS helps determine the location and severity of pleural effusion, which in turn have a pivotal role in deciding if further invasive management is appropriate. Our group is also working concurrently on the other lung pathologies, for each of which the approach is the same: the algorithms developed can be combined into a lung pathology diagnostic tool, or as separate specific identification and evaluation tools. Future work includes extension of the algorithm to automatically segment pleural effusion, which will enable estimation of the pleural effusion volume, an important parameter useful for treatment decisions.

Since it is now clear that the most impacted regions in the world are and will be the resource-poor ones, due to the limited access to good level healthcare, it is vital to propose diagnostic methods and systems that are cost-effective, deployable at the bedside, easy to use and

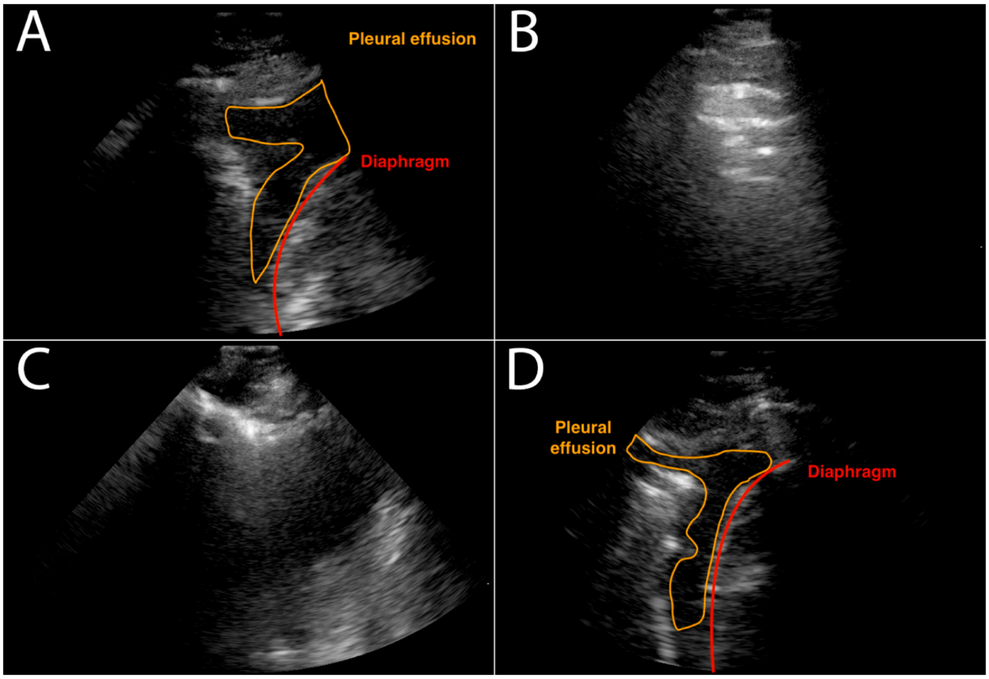

Fig. 2. Images for which the prediction is different between the video- and frame-based labelling approach. (A) True positive (frame-based) and false negative (video-based) This LUS image from the LPL scanning position shows signs of pleural effusion; (B) True negative (frame-based) and false positive (video-based) This LUS image from the RPL scanning position shows no pleural effusion signs; (C) False positive (frame-based) and true negative (video-based). This LUS image from the RANT scanning position shows no signs of pleural effusion; (D) false negative (frame-based) and true positive (video-based) This LUS image from the RPL scanning position shows signs of pleural effusion.

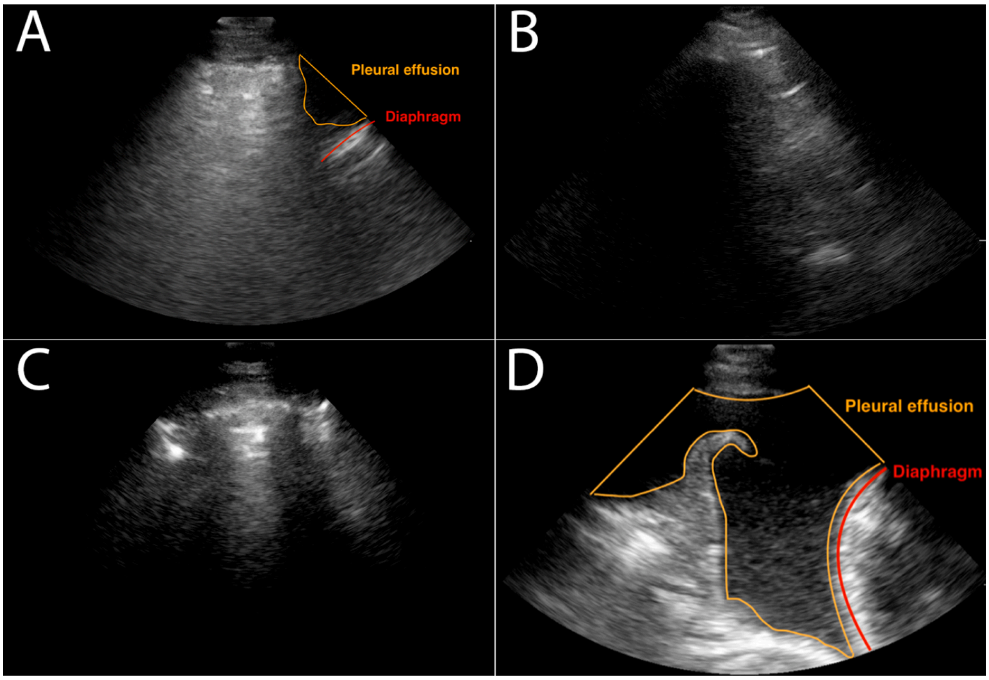

Fig. 3. Images for which the prediction is the same for the video- and frame-based labelling approach. (A) False negative. This LUS image from the RPL scanning position shows signs of pleural effusion; (B) False positive. This LUS image from the RPL scanning position shows no strong signs of pleural effusion. The small black rectangles at the bottom of the image are the replaced black pixels of the overlays inside the ultrasound sector; (C) True negative. This LUS image from the LPU scanning position shows no signs of pleural effusion; (D) True positive. This LUS image from the RPL scanning position shows the pathological signs of pleural effusion.

reliable. Ultrasound based systems hold the potential to fulfil all these requirements if a sufficient level of automatisation at the image acquisition and at the interpretation stages is available. Automation will also increase the reliability of LUS, reduce the risk of error in diagnosis, and inappropriate patient management. The LUS training burden may also be reduced in the presence of algorithms such as the one introduced here to assist the sonographer. It is estimated to take between 3 and 6 months of full-time training to reach proficiency under current conventional training approaches. Finally, the experience developed with this project will be transferable to other types of clinical ultrasound imaging which suffer from similar limitations such as transthoracic echocardiography, abdominal ultrasound and ultrasound screening for deep venous thrombosis. It is imperative to highlight that access to ultrasound machines is no longer a barrier as handheld ultrasound systems cost as low as US$500 in 2020. This makes US an economical imaging modality compared to others. Provided that sufficient automation is implemented, LUS is presently the only portable, cost-effective and real-time imaging technique for monitoring individuals at risk of subclinical deterioration in the outpatient setting. LUS is also an efficient risk stratification tool. It can be used to predict possible deterioration of subclinical pulmonary disease in hospital wards and indicate the need for early intubation and mechanical ventilation, which may avoid unnecessary deaths. In critical care settings, LUS helps predict which patients will benefit from prone or high positive end-expiratory pressure

Table 15
Confusion matrix comparison of the best fold (Fold #3) between the frame- and the video-based approach, evaluated on the test set (N/A: not applicable).

Cross-conditions	TP (frame-based)	TN (frame-based)	FP (frame-based)	FN (frame-based)
TP (video-based)	1,807 LUS images Fig. 3 (D)	N/A	N/A	74 LUS images Fig. 2 (D)
TN (video-based)	N/A	9,710 LUS images Fig. 3 (C)	79 LUS images Fig. 2 (C)	N/A
FP (video-based)	N/A	206 LUS images Fig. 2 (B)	68 LUS images Fig. 3 (B)	N/A
FN (video-based)	77 LUS images Fig. 2 (A)	N/A	N/A	176 LUS images Fig. 3 (A)

versus prone ventilation and which ones will not.

In this work, a frame-based and a video-based (weakly supervised) labelling method were implemented, where the latter is completely novel and showed a comparable accuracy score with significantly less labelling effort. It also has the potential to decrease the variability connected with labelling operator dependence, making the algorithm more robust. Furthermore, the video-based approach has a higher recall score for the worst fold which is important to not miss any cases. These results are based on [19] but, by contrast, we have performed binary classification instead of providing a 4 level severity score. The results show that using a video-based labelling approach leads to statistically equal performance to a frame-based labelling approach while reducing the amount of labelling effort significantly. For example, for the dataset used here, only 623 video labels were required compared to 99,209 frame labels. Although the accuracy is still high for the worst performing fold, the recall and the F1 score are significantly lower than the best fold. This means that the generalization of the algorithm could be improved which needs to be investigated further.

It is important to highlight that there is no known lung ultrasound feature that is pathognomonic for COVID-19 infection. Which means that training the algorithm adding the information about COVID diagnosis would not help identify COVID-19 patients from effusion evaluations. In particular, the prevalence of pleural effusions seems to be variable from available studies [22]. Effusion can be categorised into simple (anechoic), or complex (with septations or internal echogenicity). Simple effusions usually indicate transudates while complex effusions are generally associated with exudates. The process of obtaining pleural fluid specimens is considered an aerosol generation procedure and is therefore generally avoided where possible in patients suspected or confirmed for COVID-19 infection. The definitive characterisation of the pleural effusion requires chemical and biochemical analysis in the laboratory. To date, COVID-19 related pleural effusions were reported mostly to be exudative [23,24]. Based on these, it's more than likely that COVID-19 related pleural effusions are indistinguishable from effusions of other infective causes. It is the combination of clinical history, underlying epidemiological risk and exposures, together with the ultrasound findings that may lead to a reliable diagnosis of COVID-19 induced respiratory disease. So, accurately and precisely identifying all the possible pathologies that are part of the ultrasound findings is an essential step towards automatic COVID-19 diagnosis, and the work presented in our paper is the first step as mentioned in this direction. Effusion among these pathologies is particularly important because its presence may be predictive of a worse prognosis and can indicate bacterial superinfection in COVID-19, based on prior experience with MERS-CoV [22]. On the other side, as previously mentioned, identifying pleural effusion is also a valuable tool in general, besides COVID-19. In fact, it allows clinicians to consider alternative causes for patients' shortness of breath, fever etc. such as, for example, presence of a concurrent medical condition such as acute decompensated heart failure, that also requires prompt treatment.

The problem is inherently three-dimensional because, as most scanning protocols nowadays suggest, it is necessary to scan both lungs at several different locations to diagnose reliably this type of viral pneumonitis. As a matter of fact, it is actually four-dimensional, because the longitudinal evolution of some of the pathologies associated with the disease provides essential information to make appropriate treatment decisions. A limitation of the training we have performed in this paper is that it was performed using single 2D frames, instead of whole videos or 3D volumes and that no volumetric information (i.e., for example, using context information together with the frame) was added. One of the next steps in our project will be to implement spatial localization of each of the frames from the patient and produce a diagnosis based on the combination of the evaluations performed along all the scanning planes. Spatial localization is paramount especially in the COVID-19 induced disease framework where a single plane of view is not sufficient to formulate a reliable diagnosis. We plan to train our algorithms on other pathologies which are involved in these types of pneumonia such as, for example, atelectasis, consolidations, interstitial syndrome and pneumothorax. Ultimately, the ability to automatically and accurately identify all lung pathologies together with their correct spatial localization will allow for a clinically accurate diagnosis.

The training was performed with data from one single institution, part of future work is to include more institutions in the collaboration. Nevertheless, we are confident these results are representative of a general population since there is general international scientific consensus on the acquisition protocol followed by the hospital providing the data. Moreover, the data used contained only healthy cases or pleural effusion cases, therefore in principle it is not possible to establish at this stage whether the presence of other pathologies may affect these results. Another potential limitation in this study is that the ultrasound images were not verified with the current gold standard, CT of the chest, and it is possible that small effusions may have been missed. However, the accuracy of ultrasound is very close to CT, having a sensitivity and specificity at detecting pleural effusion between 90% and 100% [25]. In addition, it is likely that if ultrasound did miss a few effusions, the size of these effusions would be small and probably not clinically significant. It is also vital to highlight that the purpose of this study was to train an algorithm which could perform at the same level as expert clinicians who use LUS for pleural effusion diagnosis in normal clinical setups. As previously mentioned, these diagnoses are often performed without the support of CT scans, as LUS is currently considered a reference imaging modality for this task, since it is advantageous over CT for reasons including lower cost, absence of ionising radiation, and because it allows real-time scanning for procedural guidance. So, for the purpose of this work, the images used for training can be considered ground truths, since they have been reviewed by two clinicians trained in LUS, who independently confirmed presence of pleural effusion, one at point-of-care, and one at a later date for the purpose of this study.

5. Conclusions

In this work, it has been shown that it is possible to diagnose automatically, efficiently and reliably pleural effusion on clinical LUS images using deep learning (in particular based on the Reg-STN architecture). The algorithm proposed interprets LUS datasets to identify pleural effusion, with comparable or improved accuracy compared to clinical standards (for both the frame-based and scanning position-based labelling methods) and allows faster diagnosis and a more robust result irrespective of the competence of the sonographer performing the examination. We have also demonstrated that video-based labelling approach can achieve comparable results to a frame-based labelling approach for classifying pleural effusion in lung ultrasound images. This significantly reduces the input required from clinical experts to provide ground truth labels.

This is the first step towards full automation of LUS evaluation for lung pathologies, which will democratise effective access to diagnostic tools for diseases such as viral pneumonias, including COVID-19 induced respiratory inflammations.

Acknowledgments

The authors would like to thank Dr. Ximena Cid (Department of Surgery, University of Melbourne, Melbourne, Australia) for contribution of the LUS images; Prof. Gustavo Carneiro (University of Adelaide, Adelaide, Australia) and Dr. Maria Antico for the contribution in the initial discussions; and the support received by the authors of [19] with the initial setup of the code.

References

[1] Soldati G, Smargiassi A, Inchingolo R, Buonsenso D, Perrone T, Briganti DF, et al. Proposal for international standardization of the use of lung ultrasound for COVID-19 patients; a simple, quantitative, reproducible method. J Ultrasound Med 2020. https://doi.org/10.1002/jum.15285.

[2] Mento F, Perrone T, Macioce VN, Tursi F, Buonsenso D, Torri E, et al. On the impact of different lung ultrasound imaging protocols in the evaluation of patients affected by coronavirus disease 2019. J Ultrasound Med 2020. https://doi.org/10.1002/jum.15580.

[3] Perrone T, Soldati G, Padovini L, Fiengo A, Lettieri G, Sabatini U, et al. A new lung ultrasound protocol able to predict worsening in patients affected by severe acute respiratory syndrome coronavirus 2 pneumonia. J Ultrasound Med 2020. https://doi.org/10.1002/jum.15548.

[4] Gargani L, Soliman-Aboumarie H, Volpicelli G, Corradi F, Pastore MC, Cameli M. Why, when, and how to use lung ultrasound during the COVID-19 pandemic: Enthusiasm and caution. Eur Heart J Cardiovasc Imaging 2020. https://doi.org/10.1093/ehjci/jeaa163.

[5] Cid X, Canty D, Royse A, Maier AB, Johnson D, El-Ansary D, 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) s. Trials 2020. https://doi.org/10.1186/s13063-019-4003-2.

[6] Amatya Y, Rupp J, Russell FM, Saunders J, Bales B, House DR. Diagnostic use of lung ultrasound compared to chest radiograph for suspected pneumonia in a resource-limited setting. Int J Emerg Med 2018. https://doi.org/10.1186/s12245-018-0170-2.

[7] Volpicelli G, Elbarbary M, Blaivas M, Lichtenstein DA, Mathis G, Kirkpatrick AW, et al. International evidence-based recommendations for point-of-care lung ultrasound. Intensive Care Med 2012. https://doi.org/10.1007/s00134-012-2513-4.

[8] Havelock T, Teoh R, Laws D, Gleeson F. Pleural procedures and thoracic ultrasound: British Thoracic Society pleural disease guideline 2010. Thorax 2010;(2). https://doi.org/10.1136/thx.2010.137026.

[9] Huang Q, Zhang F, Li X. Machine learning in ultrasound computer-aided diagnostic systems: a survey. Biomed Res Int 2018. https://doi.org/10.1155/2018/5137904.

[10] Baloescu C, Toporek G, Kim S, McNamara K, Liu R, Shaw MM, et al. Automated lung ultrasound B-Line Assessment using a deep learning algorithm. IEEE Trans Ultrason Ferroelectr Freq Control 2020. https://doi.org/10.1109/TUFFC.2020.3002249.

[11] Carrer L, Donini E, Marinelli D, Zanetti M, Mento F, Torri E, et al. Automatic pleural line extraction and COVID-19 scoring from lung ultrasound data. IEEE Trans Ultrason Ferroelectr Freq Control 2020. https://doi.org/10.1109/TUFFC.2020.3005512.

[12] Kulhare S, Zheng X, Mehanian C, Gregory C, Zhu M, Gregory K, et al. Ultrasound-based detection of lung abnormalities using single shot detection convolutional neural networks BT - simulation, image processing, and ultrasound systems for assisted diagnosis and navigation. In: Stoyanov D, Taylor Z, Aylward S, Tavares JMRS, Xiao Y, Simpson A, et al., editors., Cham: Springer International Publishing; 2018, p. 65–73.

[13] Ford JW, Heiberg J, Brennan AP, Royse CF, Canty DJ, El-Ansary D, et al. A pilot assessment of 3 point-of-care strategies for diagnosis of perioperative lung pathology. Anesth Analg 2017. https://doi.org/10.1213/ANE.0000000000001726.

[14] Brogi E, Gargani L, Bignami E, Barbariol F, Marra A, Forfori F, et al. Thoracic ultrasound for pleural effusion in the intensive care unit: A narrative review from diagnosis to treatment. Crit Care 2017. https://doi.org/10.1186/s13054-017-1897-5.

[15] Mason D. SU-E-T-33: Pydicom: An Open Source DICOM Library. Med. Phys., 2011. https://doi.org/10.1118/1.3611983.

[16] Nema D. DICOM PS3.3 2016c -. Information Object Definitions. 2016.

[17] PyTorch. PyTorch documentation - PyTorch master documentation. PyTorch; 2019.

[18] Efraimidis PS, Spirakis PG. Weighted random sampling with a reservoir. Inf Process Lett 2006. https://doi.org/10.1016/j.ipl.2005.11.003.

[19] Roy S, Menapace W, Oei S, Luijten B, Fini E, Saltori C, et al. Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. IEEE Trans Med Imaging 2020. https://doi.org/10.1109/TMI.4210.1109/TMI.2020.2994459.

[20] Van Sloun RJG, Demi L. Localizing B-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results. IEEE J Biomed Heal Informatics 2020. https://doi.org/10.1109/JBHI.2019.2936151.

[21] Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial transformer networks. Adv. Neural Inf. Process. Syst. 2015.

[22] Aujayeb A. Clarification on pleural effusions in COVID-19. Radiol Cardiothorac Imaging 2020. https://doi.org/10.1148/ryct.2020200330.

[23] Chong WH, Huggins JT, Chopra A. Characteristics of pleural effusion in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pneumonia. Am J Med Sci 2021;361:281–4. https://doi.org/10.1016/j.amjms.2020.09.008.

[24] Hussein M, Haq IU, Hameed M, Thomas M, Elarabi A, Allingawi M, et al. Pleural effusion as an isolated finding in COVID-19 infection. Respir Med Case Rep 2020. https://doi.org/10.1016/j.rmcr.2020.101269.

[25] Soni NJ, Franco R, Velez MI, Schnobrich D, Dancel R, Restrepo MI, et al. Ultrasound in the diagnosis and management of pleural effusions. J Hosp Med 2015. https://doi.org/10.1002/jhm.2434.