Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Informatics in Medicine Unlocked 25 (2021) 100687

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Automated detection of pneumonia in lung ultrasound using deep video classification for COVID-19

Salehe Erfanian Ebadi [a,b,c], Deepa Krishnaswamy [a,c,*], Seyed Ehsan Seyed Bolouri [b], Dornoosh Zonoobi [b], Russell Greiner [d], Nathaniel Meuser-Herr [f], Jacob L. Jaremko [a,b], Jeevesh Kapur [b,e], Michelle Noga [a,c], Kumaradevan Punithakumar [a,c,d]

[a] Department of Radiology & Diagnostic Imaging, University of Alberta, Edmonton, Canada
[b] MEDO.ai, Edmonton, Canada
[c] Servier Virtual Cardiac Centre, Mazankowski Alberta Heart Institute, Edmonton, Canada
[d] Department of Computing Science, University of Alberta, Edmonton, Canada
[e] National University Hospital, Singapore
[f] Upstate Medical University, Syracuse, NY, USA

ARTICLE INFO

Keywords:
COVID-19
Lung ultrasound
Video classification
Convolutional neural networks

ABSTRACT

There is a crucial need for quick testing and diagnosis of patients during the COVID-19 pandemic. Lung ultrasound is an imaging modality that is cost-effective, widely accessible, and can be used to diagnose acute respiratory distress syndrome in patients with COVID-19. It can be used to find important characteristics in the images, including A-lines, B-lines, consolidation, and pleural effusion, which all inform the clinician in monitoring and diagnosing the disease. With the use of portable ultrasound transducers, lung ultrasound images can be easily acquired, however, the images are often of poor quality. They often require an expert clinician interpretation, which may be time-consuming and is highly subjective. We propose a method for fast and reliable interpretation of lung ultrasound images by use of deep learning, based on the Kinetics-I3D network. Our learned model can classify an entire lung ultrasound scan obtained at point-of-care, without requiring the use of preprocessing or a frame-by-frame analysis. We compare our video classifier against ground truth classification annotations provided by a set of expert radiologists and clinicians, which include A-lines, B-lines, consolidation, and pleural effusion. Our classification method achieves an accuracy of 90% and an average precision score of 95% with the use of 5-fold cross-validation. The results indicate the potential use of automated analysis of portable lung ultrasound images to assist clinicians in screening and diagnosing patients.

1. Introduction

The coronavirus disease 2019 (COVID-19) pandemic, caused by the severe acute respiratory syndrome coronavirus2 (SARS-CoV-2) has affected individuals around the world, causing over three million deaths to date (World HealthOrganization, 2020). The ability to quickly examine and diagnose patients with COVID-19 related pneumonia is crucial as the number of patients rises. Several diagnostic measures are used, for instance, the nasopharyngeal swab, but it may produce many false negatives in an affected person [1]. The reference standard used is the viral nucleic acid with reverse transcription polymerase chain reaction (RT-PCR) [2]. Delays may occur acquiring the results of the test, and this along with other issues such as inaccessibility proves the need for alternate methods of quick and reliable diagnosis.

Over the years lung ultrasound (LUS) imaging has proven effective at screening patients for a variety of disorders in the emergency care setting [3], and has been used to screen and diagnose patients with COVID-19 [4]. There are several advantages to employing ultrasound imaging (US) instead of traditional computed tomography (CT) or X-ray imaging: (1) US is a low-cost alternative compared to other imaging modalities, especially crucial as it may be beneficial in countries where affordability is a key issue (2) US does not expose the patient to harmful radiation produced by CT imaging, which is especially problematic as patients may be screened multiple times (3) One of the most significant advantages is that US scanners are highly portable and the transducers are easily sanitized, resulting in the effective scanning of affected patients in isolation (4) Importantly, LUS imaging can screen patients for acute respiratory distress syndrome (ARDS) [5]. As the lungs normally contain air, changes to the tissue or fluid and their corresponding ratios affect the US images, resulting in various artifacts [6]. A patient

* Corresponding author at: Department of Radiology & Diagnostic Imaging, University of Alberta, Edmonton, Canada.
E-mail address: deepa@ualberta.ca (D. Krishnaswamy).

https://doi.org/10.1016/j.imu.2021.100687
Received 11 June 2021; Received in revised form 28 July 2021; Accepted 29 July 2021
Available online 4 August 2021

with COVID-19 may present several symptoms that may manifest in the combination of various image artifacts in the US image. These abnormalities may range as follows, according to the Bedside Lung Ultrasound in Emergency (BLUE) protocol [7]:

1. **Artifact analysis (A-lines)**: Reflections occur between the transducer and the surface of the lung. An US image of a healthy lung might include A-lines, or horizontal artifacts, which occur when multiple reverberations are present below the pleural line [6,8]. The presence or absence of A-lines in different portions of the lung is an important factor in determining the severity of the ARDS.
2. **Artifact analysis (B-lines)**: B-lines are long vertical artifacts that usually begin from the pleural line [5,7]. B-lines often have a comet-tail appearance that is well-defined, which cover the A-lines. These occur when instead of air, other media are present in the lung (e.g. water or blood) [5].
3. **Alveolar consolidation and/or pleural effusion:** Alveolar consolidation is when the air spaces, the alveoli, are filled with a fluid such as water or blood [7]. Pleural effusion yields a distinct pattern in the US image [7] and is caused by excessive amounts of fluid between layers of the pleura.
4. **Lung sliding:** Normally, the lung slides in rhythm with the respiration of the patient at the pleural line. In the case of abnormalities, this does not occur and could indicate either pneumothorax if there are multiple A-lines and no B-lines, or pneumonia if there are only B-lines present [3,7].

Examples of A-line, B-lines, consolidation, and pleural effusion echo patterns are shown in Fig. 1. These artifacts may be difficult to observe in US images, making it challenging to interpret especially if the observer is not experienced. Therefore, the use of fast, automated, and most importantly, accurate, methods for the analysis of LUS images is of crucial importance.

1.1. Related work

There have been a number of techniques developed for the classification of various features in LUS images and videos, where many are models learned by deep learning using video or frame-based data as input. One classification method that uses a low-cost approach has been developed using the concept of subspace representations [9]. The authors proposed classifying LUS images into A-lines, B-lines, and consolidations using a multi-layer network termed the Subspace Approximation with Adjusted Bias (Saab) network. Another set of authors proposed an approach to classify COVID-19 severity on a single frame and video-level grading basis [10]. A scoring system of four levels was employed: (0) if pleural lines were present and included A-lines (1) if abnormalities are present (2) if there are small or large consolidations and (3) if there is a wide hyperechogenic area below the pleural surface. For frame-based score prediction, the authors developed a method termed Regularized Spatial Transformer Network (Reg-STN) to perform the classification into one of the four classes. For scoring videos, the authors made use of a parameterized aggregation layer, which combined the predictions from each frame.

Other researchers have developed techniques for multiple diseases that involve the presence of B-lines. In one approach, the authors developed a method to distinguish between three diseases associated with the presence of B-lines, namely, hydrostatic pulmonary edema (HPE), non-COVID-19 ARDS, and COVID-19 ARDS [11]. The authors performed frame-by-frame classification of US images using the network architecture of [12]. In another approach, the authors performed detection and classified pneumonia into eight clinical feature classes: normal subjects, B lines of a certain number, pleural effusion, and depth of hepatization [13]. Multiple features were manually grouped together into three classes, four classes, and the full eight classes. Three different neural network architectures were evaluated, namely VGG-19 [14], Resnet-101 [15] and EfficientNet-B5 [16].

Other methods have been developed in order to differentiate among patients with COVID-19, bacterial pneumonia, non-COVID-19 viral pneumonia as well as controls [17]. The authors have compiled a detailed dataset of both individual frames and videos from affected patients for classification. For frame-based classification, multiple network architectures were employed, including VGG-16 [14] and NasNET Mobile [18]. For the video-based classification, the VGG-16 network [14] along with Models Genesis [19] is used. The authors also evaluated their method on an independent test set from [10] in order to predict the severity score from 0 to 3. Instead of using an aggregation layer [10], the authors developed an approach that used an estimate of the confidence and ignored frames with low confidence values.

1.2. Our contribution

The purpose of the study was to diagnose and monitor the presence or absence of ARDS and pneumonia in the lungs. Our proposal aims to bridge the gap in the current diagnostic procedure of LUS by circumventing the need for the time-consuming training of human experts, as US patterns are difficult to discern [20] and also reduce the time that medical staff invests to diagnose a patient, as show-cased by [21]. We propose a framework for automatic detection of ARDS features seen in LUS that are present in pneumonia and COVID-19 patients. Our contributions can be summarized as follows:

1. We present a technique for classifying LUS video scans acquired at point-of-care without requiring any further processing or the need for operator intervention. The neural network can be easily re-trained with new data or new categories to adapt the model to the needs of specific applications involving LUS.
2. Our model obtains a classification accuracy of 90% and an average precision score of 95% micro-averaged over all classes denoting the disease stages, over a large dataset comprised of 1530 LUS scans with 287,549 frames. Furthermore, our model can be deployed on both the GPU and CPU with a classification speed of 220 milliseconds per video with an average of 240 frames (1090 frames per second).
3. Our model does not need expensive, arduous, and time-consuming manual data annotation and labeling; a task which is often a prerequisite for most supervised deep learning methods, making it suitable for quick adaptation and transfer learning to new tasks that involve spatio-temporal learning of signals.

2. Materials and methods

2.1. LUS dataset for ARDS feature classification

A major objective in our research was to obtain and create high-quality labeled data for LUS collected from multiple centers. We have a large dataset of 300 patient studies, with 100 studies per label (A-lines, B-lines, and consolidation, and/or pleural effusion). Each study has between 3–10 scans taken from different regions of the lung. Expert radiologists were then asked to label this dataset assessing three criteria relevant to ARDS: (1) presence of A-line artifacts, (2) presence of B-line artifacts, and (3) presence of consolidation and/or pleural effusion. The classification of 100 A-lines, 100 B-lines, and 100 consolidations is based on the ARDS criteria and is from the medical diagnostic record.

The resulting labeled dataset is comprised of 475 videos presenting A-lines, 491 videos presenting B-lines, and 564 videos presenting consolidation and/or pleural effusion, totaling 1530 videos. Each video has between 28-449 frames sampled at 30 Hz, and the total number of frames in the dataset is 287,549. On average, each video has 240 frames. Our models were trained and tested with 5-fold cross-validation that creates training and testing sets with 80% (1225 videos) and 20% (306 videos), respectively, for each fold.

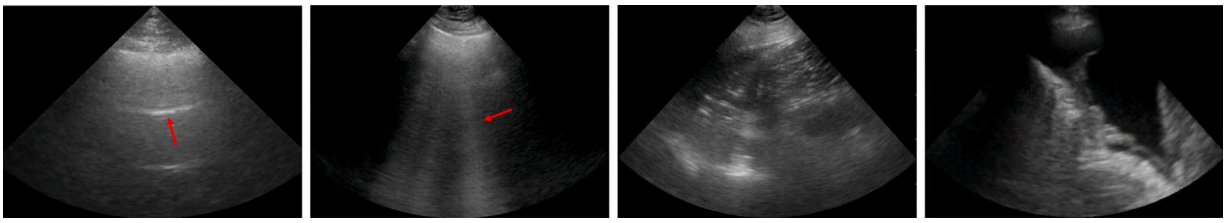

Fig. 1. Lung ultrasound profiles from left to right: A-lines, B-lines, consolidation, and pleural effusion.

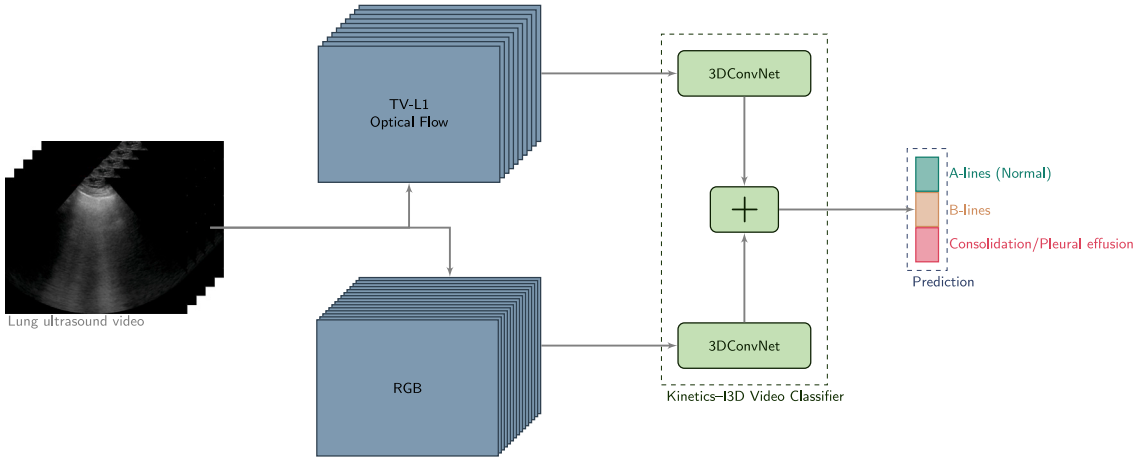

Fig. 2. The proposed two-stream inflated 3D ConvNet approach for the classification of pathology from lung ultrasound videos.

2.2. Video classification with Kinetics-I3D

We propose a Two-Stream Inflated 3D ConvNet (I3D) [22] for the task of classifying video sequences obtained by LUS devices at point-of-care; see Fig. 2. I3D has been originally developed for human action recognition from videos. It inflates the filters and pooling kernels in a 2D ConvNet that had previously been trained on the ImageNet dataset. The total number of parameters for the original model is 25M [22]. Here we demonstrate that I3D has the potential for transfer learning from biomedical imaging tasks that involve learning spatio-temporal signals. LUS is an excellent candidate since diagnosis is based on features that are seen in individual 2D images from the patients' videos and the temporal changes observed in each sequence. The training, inference, and validation were performed using the python programming language version 3.7 and TensorFlow version 1.15.

We train our network with the entire length of our videos, as certain features of interest (such as B-lines) could appear only during a single frame due to the operators' capture. Therefore, we cannot sub-sample or trim our video temporally to match the I3D input, which accepts 64 frames per video. This means that we had to re-implement both the training and testing pipeline in order to use videos of varying frame lengths. This is challenging with TensorFlow, as it does not provide a straightforward way of reading batches where each element in the batch has a different size. We, therefore, chose to train with a batch size of 1. The videos obtained from the LUS are between 28-449 frames long, sampled at 30 Hz. We manually crop the videos to contain only the scan region, then resized the videos to 224×224 pixels as an input to the network.

I3D has two streams of networks, where one accepts the RGB input video, and the other the optical flow fields extracted from the RGB video. The two networks are identical, and the classification result is obtained by fusing the output of the RGB network with the output of the flow network. Although I3D originally reported that using two streams results in better performance, which we also observed with our data, I3D can be used with the RGB stream alone as well, without much compromise on the classification scores.

We extract optical flow sequences with the TV-L1 algorithm [23]. Since US data has a high noise floor by nature and there are erratic temporal changes from one frame to the next, we slightly smooth out the extracted flow fields along the temporal channel using a 1D Gaussian filter with a smoothing factor of 1.15. We found this smoothing factor to be the best in preserving the temporal information while reducing high-frequency noise, yielding better flow fields with more easily interpreted visual features. Two examples of the original RGB images of two consecutive frames and the resulting optical flow magnitude are shown in Fig. 3. The values of the pixels resulting from optical flow are clipped to the range of $[-20, 20]$, and subsequently rescaled between -1 and 1. As the flow consists of two channels (x and y direction), a third channel of zeros is added. The value of 1 is added to all three channels and then divided by 2, resulting in values between 0 and 1 for all three channels [22].

A large portion of our dataset has been captured by keeping the probe fixed on pre-specified regions of the patient's anterior, lateral, and posterior chest locations. Naturally, any temporal motion seen in the videos resulted from the patient's respiration. Often, in the case of healthy scans exhibiting A-lines, these motions become very small, resulting in a very weak optical flow field that conveys little to no information and reduces the ability of the network to learn from the flow stream. As such, we boosted the extracted optical flow fields by an order of magnitude. Furthermore, with our dataset, down-weighting the x-component of the optical flow by half boosts the performance. This is intuitive as most of the sequences are captured without moving the probe, and thus little motion along the x-component is seen. Visually, the A-lines demonstrate themselves if motion along the y-direction is recorded. Down-scaling the x-component helps the model learn very subtle motions along the y-direction more effectively. Nevertheless, some of the videos in our dataset are captured with a sweeping motion that shows multiple regions of the lung. We elected to apply the same processing to these videos as well.

We warm-started our model training with the weights of the action recognition network, which itself was warm-started by training to recognize 1000 object categories in the ImageNet database. The ImageNet

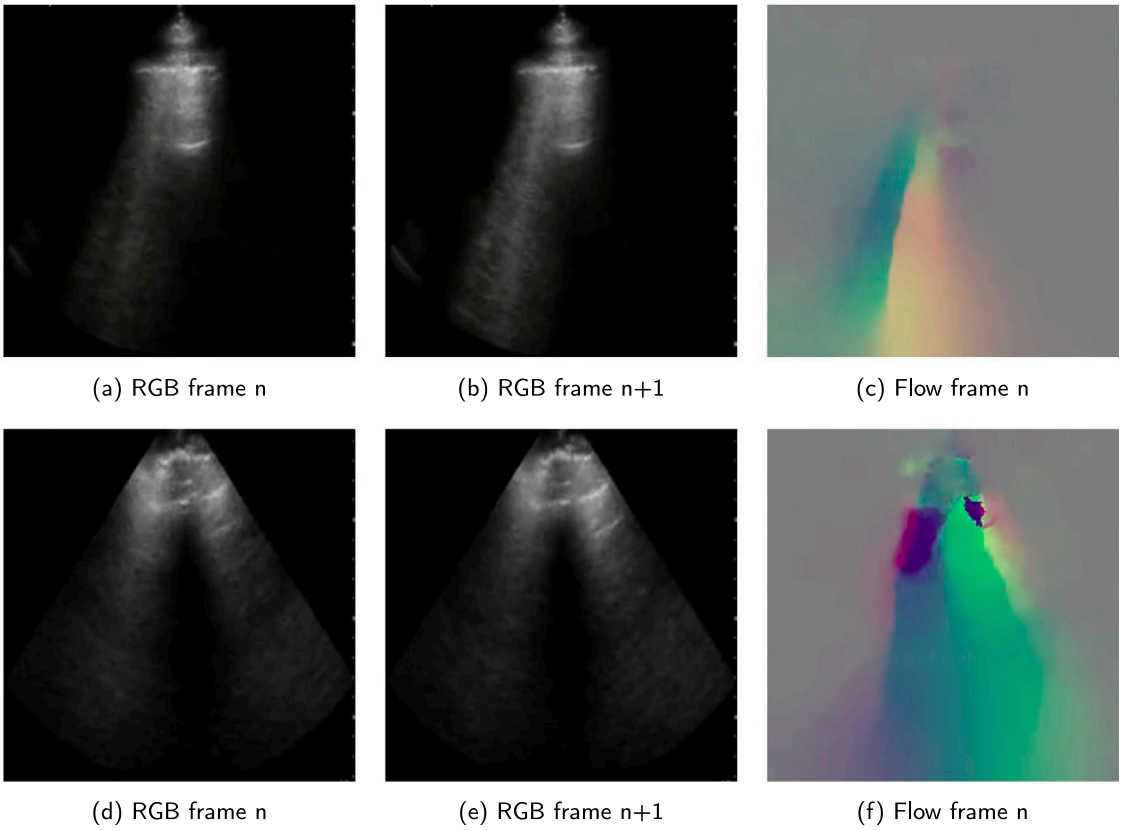

Fig. 3. Two examples (a) to (c) and (d) to (f) demonstrating the optical flow magnitude between two consecutive frames: (a) RGB frame n (b) RGB frame n+1 (c) Flow frame n (d) RGB frame n (e) RGB frame n+1 (f) Flow frame n.

training enables the I3D model to learn a rich representation for spatial features occurring in natural images. Leveraging these features with the temporal features of the kinetic action recognition dataset enables the model to learn not only scene-specific actions with an emphasis on the scene-setting but also temporal actions with an emphasis on the changes in the scene. Both of these properties are desirable for our LUS video classification; certain features are objective, such as the presence of A-lines, confluent B-lines, or consolidations, while other features are manifested temporally, such as appearing and disappearing B-lines with every respiration cycle, or compression atelectasis, or pleural sliding or its lack thereof. We validated this by testing the model in various initialization settings, namely, training from scratch, training from I3D weights without warm-starting, and training from I3D weights with warm-starting by ImageNet weights. We found that the models trained by warm-starting the ImageNet trained weights performed superior to all the other settings, and hence we report those results only.

Furthermore, our model can certainly be trained and validated using only the RGB stream of the I3D; as discussed earlier, the flow field stream tends to capture important temporal information for classification, which helps the model achieve better performance.

Training on videos, we used Adam optimizer with a learning rate of 1×10^{-4}, and momentum set to 0.9, with 1 NVIDIA RTX 2080 Ti. We trained all of our models for 3300 steps. Geometric and photometric data augmentations were not used, as we did not observe any improvement with them. The geometric augmentations fail to generalize the learning process when applied independently to each frame of the 3D video input. This indicates that the context is more important in a video, and for all the input frames that the network observes, temporal consistency must be strictly maintained. Photometric augmentations also added little to no improvement to the performance because our dataset is more or less homogeneous because of consistency during capture.

Since we use a batch size of 1, we also tested our model by eliminating batch normalization layers from our implementation of I3D but found no improvement with this setting. We report our results with the batch normalization layers. Intuitively when batch normalization is used, the model becomes more robust to illumination or gain changes that occur during a scan.

2.3. Handling data imbalance

Our dataset is imbalanced with respect to the class that shows healthy to moderately diseased features. The severe disease cases with consolidation and pleural effusion classes combined have only roughly 18% more cases than the healthy class. Therefore, the model trained on this data will be biased towards the abundant class. We, therefore, trained our model with Focal Loss [24] as an attention guiding mechanism, which penalizes the easily classified examples coming from the abundant class by modifying the standard cross-entropy loss such that it down-weights the loss assigned to the well-classified examples. A modulating factor is added to the cross-entropy loss function in order to improve the training on the difficult cases, where the hyperparameter $\gamma \geq 0$ is introduced. The equation for Focal loss is given in (1), where the value of p is the estimated probability of the class and is defined by $p \in [0, 1]$.

$$FL(p_t) = -(1 - p_t)^{\gamma} log(p_t) \tag{1}$$

We also tried over-sampling from the scarce class or under-sampling from the abundant class during training, both of which had a very small impact on the model's generalization because the focal loss is already effective enough. A K-fold cross-validation approach with $k = 5$ was used throughout our experiments. In the Discussion section, we showcase the attention capabilities of our model by visualizing its activation maps.

Table 1
Results are displayed from the 5-fold cross-validation classification of lung ultrasound videos into A-lines, B-lines, pleural effusion/consolidations.

Features	Precision	Recall	F1-score	Avg. precision micro Avg.	Number of scans
A-lines	0.87	0.91	0.88	0.94	475
B-lines	0.88	0.86	0.87	0.93	491
Cons./Pl.Eff.	0.96	0.92	0.94	0.98	564

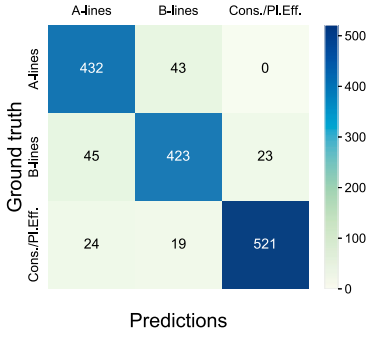

Fig. 4. Resulting confusion matrix from the multi-class classifier (with 5-fold cross-validation) evaluated on lung ultrasound videos.

3. Results

Our method learns to classify videos in a large dataset into multiple ARDS features (A-lines, B-lines, consolidation, and/or pleural effusion). The model also provides its confidence score for each classification outcome. The probability values for the presence of these features can be provided and ranked, suggesting a strong correlation with the actual ARDS features that are present in the video, which is useful for further statistical analysis or heuristics-based decision making.

As mentioned previously, we train and test our model with 5-fold cross-validation, and we ensured that the number of samples per class is similar in all folds. The classification performance of the model is shown in Table 1. The model is able to produce a classification of an 8 seconds long ultrasound video with 240 frames in 220 ms. The model learns to classify the severe disease cases (consolidation and/or pleural effusion) with a high F1-score. Notably, the 5-fold cross-validated models achieve a combined accuracy score of 0.90 and a balanced accuracy score of 0.90 that indicates a strong overall performance among all the classes. The models achieve a combined micro-averaged, average precision score of 0.95 overall. Furthermore, 1376 cases out of 1530 cases were correctly classified, resulting in a sensitivity of 90% on a large dataset.

Fig. 4 shows the confusion matrices for the classification on 5-fold cross-validated models. From these predictions, no healthy cases (A-lines) are confused with severe cases. In fact, severe disease cases can be correctly classified with a precision of 96% and a sensitivity of 93%. For healthy cases, the classification can be performed with a precision of 87% and sensitivity of 91%. The confusion matrices for individual models for each fold are shown in Fig. 5. From these results, it is evident that all the models achieve a low false negative rate and an even lower false positive rate for healthy vs. severe cases.

Nevertheless, there is room for improving the presented results, as certain cases caused the model confusion. We believe this is because our data had high variability in terms of its capture quality and device. Also, in-between cases, for example, the presence of A-lines and B-lines together, are commonly seen in LUS scans. As the patient's lung transitions from healthy to slightly de-aerated, the B-lines begin appearing and eventually obliterate the A-lines. The B-lines can also be seen together with consolidations in more advanced stages of the disease. In cases where consolidations and A-lines are seen together or on the same patient's LUS, it could signify that some regions of the patient's lung are transitioning from disease to healthy or are not yet affected adversely. As such, the model's performance could be improved significantly by acquiring more data for these transitioning cases or trimming the videos of the scans such that each scan only shows one part of the lung, as opposed to sweeps that show multiple regions. We elected to report our results with these videos included providing a better understanding of our model's capabilities.

Fig. 6 shows the Precision–Recall curves per class. On average, the area under the curve ranges from 93% to 98%. The Receiver Operating Characteristic curves with the area under the curve (ROC-AUC) are shown in Fig. 7. For all the classes, we obtained mean AUC scores ranging from 91% to 96%.

4. Discussion

4.1. Visualizing activation maps

Human expert radiologists tend to focus on specific critical regions of the LUS scans in order to form a diagnosis. Since our model can achieve a high classification performance, it would be interesting to visualize what would happen to the network's weights when the trained model is presented with a new scan to classify. This would also enable us to interpret the reliability of the results of the classification. A straightforward visualization technique is to show the activations of the convolutional neural network during the forward pass [25]. With this type of visualization, the activations usually start looking relatively blobby and dense, but as the training progresses, the activations usually become more sparse and localized. We, therefore, visualize the activations of the trained model for which we reported the classification results.

In order to visualize the activations of our network, we modify the I3D network structure by introducing convolution transpose layers for upsampling the activations of the last 3D convolutional layer before the pooling step (*Mixed_4f* layer in I3D) and discarding the last average pooling layer. The resulting upsampled activations have the same dimensions as the video height and width but with a single channel. Fig. 8 shows the upsampled activations of the network during a forward pass of a given input video. For ease of interpretation, we overlaid the activations on top of the original input image. The regions with high activation appear brighter green, and the rest of the regions containing little to no activations are shown in magenta.

From these visualizations, it can be seen that our classifier learns to focus on the regions of the lung where A-lines, B-lines, consolidation, or pleural effusion are present, all the while classifying these input videos correctly. With this insight, it is safe to assume that a robust classifier automatically discovers meaningful region of interest (ROI) detectors, representative of the learned feature categories [26]. With ROI detectors emerging as a result of learning to classify videos, it might be possible to utilize the same network to perform both video classification, and ROI localization in a single forward-pass without ever having been explicitly taught the notion of ROIs. Furthermore, the activation visualizations demonstrate that the network focuses its attention on the correct regions of the images to come up with a classification decision. This network attention is highly desirable in segmentation tasks, and therefore can be used to perform semi-supervised segmentation of these features without the need for training the model with pairs of images and their corresponding segmentation labels.

Curiously enough, although our model was not provided with lung sliding feature labels, the model tends to focus a great deal of its attention on the region where lung sliding may be present, which is surrounding the pleural line. Lung sliding is a key element for radiologists to interpret LUS scans but is otherwise a difficult feature to detect and localize for a computer vision model because it can take on a variety of appearances. It is therefore pleasantly surprising that the model has learned that the pleural region contains some important

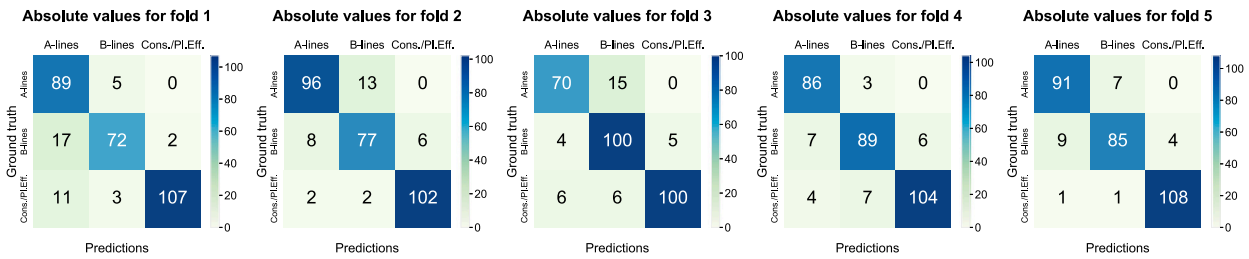

Fig. 5. Confusion matrices of our multi-class classifier on lung ultrasound videos for each individual fold.

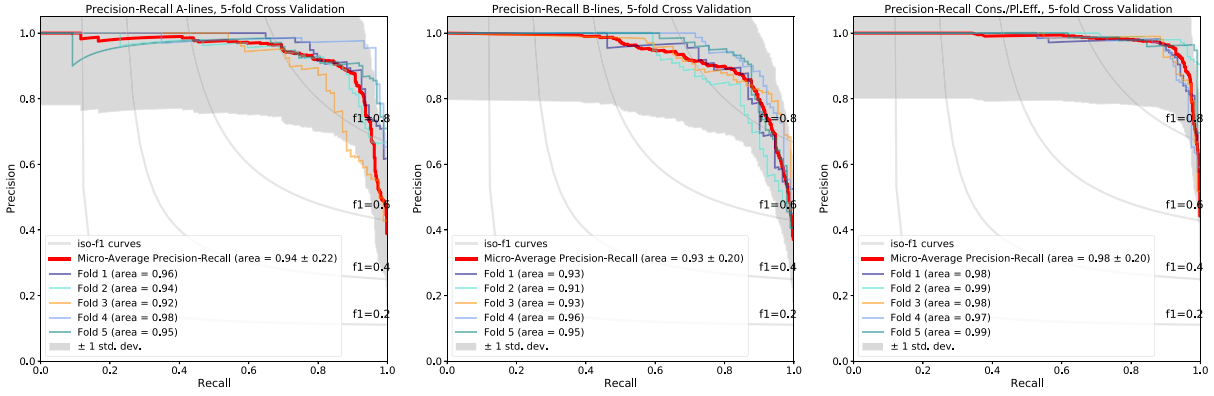

Fig. 6. Precision–Recall curve with 5-fold cross-validation. **From left to right:** Precision–Recall curves for A-lines, B-lines, and consolidation and/or pleural effusion classes.

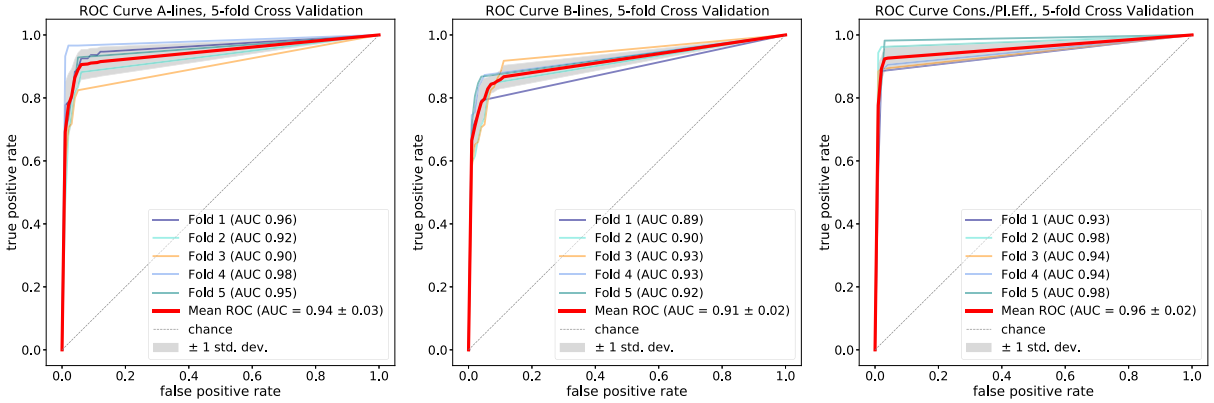

Fig. 7. Receiver operating characteristic curve with 5-fold cross-validation. **From left to right:** ROC curves for A-lines, B-lines, and consolidation and/or pleural effusion classes.

information that helps it to classify the videos. Intuitively for the model, any irregularity in the pleural line or lack of pleural sliding is an indicator of abnormalities in the lung.

In the top row of Fig. 8, we show three examples for videos labeled as A-line presence by our experts. It can be seen that the regions containing the pleural line and its reverberation artifacts down the scan highly activate the network, whereas the rest of the regions tend to be ignored. Similarly, for the second and third rows of Fig. 8, we show three examples of videos with B-line presence, and consolidation and/or pleural effusion presence, respectively. For the B-line category, the model is highly activated by the regions containing B-lines and the pleural line; conversely, for the consolidation and/or pleural effusion category, the model tends to be more sensitive to the artifacts caused directly by the presence of consolidations (air-bronchograms, or subpleural consolidations) or compression atelectasis features seen in pleural effusion.

4.2. Future work

We demonstrated the ability of the I3D video classifier in LUS video classification on a large dataset of videos. However, for certain videos, our model failed to produce the correct classification. We believe this is due to the fact that our data had high variability in terms of its capture quality and device. Also, in-between cases, for example, the presence of A-lines and B-lines together, are commonly seen in LUS scans. An interesting area to explore for future work is to handle these in-between cases. We were unable to do so since the in-between and transitional cases were scarce in our dataset, and therefore, would not provide statistically significant results.

An interesting approach to solving this problem might be to train expert binary classifiers for each category and then to combine their decision with ensemble modeling by heuristics-based analysis or bootstrap aggregating powered by the decision confidence the networks provide. It must be noted that our model is proposed as a decision

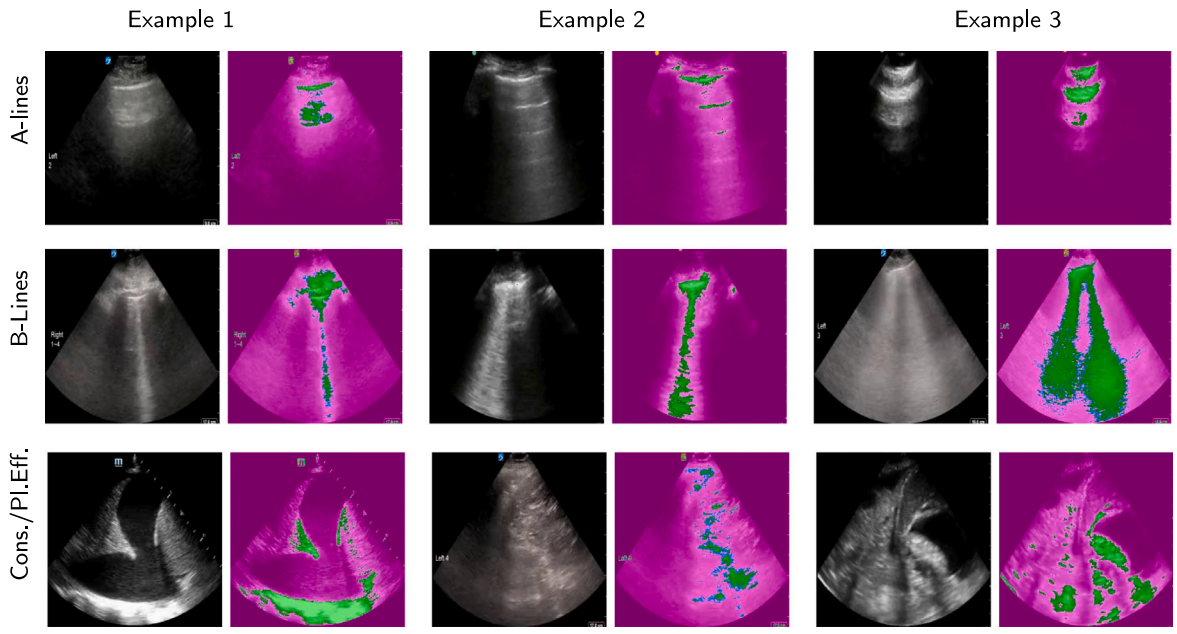

Fig. 8. Visualization of the classification network activation maps. **From top row to bottom:** three example videos and their respective activation maps for A-Lines, B-Lines, and consolidation and/or pleural effusion. For each frame the highlighted activations are shown by overlaying them on the frame. Regions of high activation appear brighter green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

support system and by no means intended to replace the already in place protocols and processes used to detect COVID-19 pneumonia. Therefore, a combination of our model and heuristics-based analysis would make the decision-making process more reliable in a clinical setting.

As with any supervised approach, our model is data-hungry, and as such more data collection efforts could help improve the results of our method significantly. More importantly, ensuring that there are enough samples for edge cases, which are rare but vitally important cases for the model to learn would help the model learn to generalize better.

Our model also has the potential for detecting and segmenting features in lung ultrasound scans. As an exercise, we up-sampled and visualized the activation maps of our video classifier at different stages in the hidden space and discovered that the model correctly focuses on regions of the frames that show the ARDS features, without any attention-guiding mechanism. This means that the hidden layers of our model can discover semantically meaningful concepts; for example, it discovers which visual feature in the video makes that video have the B-line label. This provides a great potential for fast semi-supervised segmentation, which does not require expensive manual frame-by-frame labeling that state-of-the-art segmentation methods do require. In prior literature, it was shown that object detectors could emerge from training ConvNets to perform scene classification [26]. With object detectors emerging as a result of learning to categorize scenes, the same network can perform both scene classification and object localization in a single forward-pass, without ever having been explicitly taught the notion of objects. We believe that extending our video classifier to the task of simultaneous video object classification and segmentation is an interesting subject for future research.

5. Conclusion

In this manuscript, we presented a novel technique for the classification of LUS videos obtained at point-of-care, based on a Two-Stream Inflated 3D ConvNet (I3D), originally developed for human action recognition from videos. The method presented in this manuscript can be used for fast and reliable interpretation of LUS using AI, which requires minimal labeling or data pre-processing effort, making it suitable for hospital use and fast deployment. Our technique can effectively categorize the main imaging features seen in LUS scans, such as A-lines, B-lines, consolidation, and pleural effusion, which unveil the degree to which the lungs have been affected by the infection. The interpretation of LUS scans is based on classifying the entire LUS scans (videos) obtained at point-of-care without the need for a frame-by-frame analysis or labor-intensive data labeling and pre-processing, which is usually a prerequisite for many AI-based techniques. We trained our video classifier using transfer learning by warm-starting the training from a network trained on the Kinetics-I3D, originally trained on the ImageNet dataset. The I3D network has two separate streams for RGB videos and optical flow field inputs that are fused together to provide the classification decision. The optical flow stream helps to retain certain important temporal information in the videos that improve the classification performance of the model.

We compared our results against ground truth assessments provided by expert radiologists. Our results indicate the effectiveness of this method in classifying between the main imaging features of LUS scans with an accuracy of 90%, a balanced accuracy of 90%, and an average precision score of 95%, with 5-fold cross-validated models. Further statistical analysis was provided to corroborate the effectiveness of the proposed methodology. Our model can provide an accurate classification of a given ultrasound scan that is 8 seconds long with 240 frames, in 220 milliseconds, as our system processes the entire scan with a single forward pass into the network, as opposed to the traditionally used frame-by-frame analysis which is time-consuming and could provide less reliable results, simply because temporal correlations between frames are lost.

One limitation of our method is the absence of patient information, as the data received was anonymized. Follow-up information concerning the patients is also not available for use. Future studies will ensure that more information is readily available, including patient characteristics and inclusion criteria.

We provided some insight as to the interpretation of the inner workings of our model by visualizing the network activations. The activations reveal that during a forward pass of a given video, the network correctly focuses on those regions of the image that contain the visual features that are important for diagnosis. Although we did not utilize any attention-guiding mechanisms, our network has implicitly learned to focus its attention on the same regions as a human

radiologist would. We believe that automated AI analysis of portable US imaging can help triage patients presenting to emergency with flu-like or breathing difficulty symptoms, determine who needs to be hospitalized, and immediately identify those patients who require ICU admission. We envision that our method could serve as an essential screening tool to allocate scarce hospital resources and help save lives in the COVID-19 pandemic.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Mitacs Accelerate, NSERC Alliance, and CIFAR AI & COVID-19 Catalyst Grants. The authors wish to thank Masood Dehghan for his valuable assistance in preparing this manuscript.

References

[1] Woloshi S, Patel N, Kesselheim A. False negative tests for SARS-CoV-2 infection—challenges and implications. N Engl J Med 2020;383(6):e38.
[2] Kuzan T, Altıntoprak K, Çiftçi H, Ergül U, Özdemir N, Bulut Mea. A comparison of clinical, laboratory and chest CT findings of laboratory-confirmed and clinically diagnosed COVID-19 patients at first admission. Diagnostic Intervent Radiol 2021;27(3):336–43.
[3] Volpicelli G, Elbarbary M, Blaivas M, Lichtenstein D, Mathis G, Kirkpatrick A, Melniker L, Gargani L, Noble V, Via G, Dean A. International evidence-based recommendations for point-of-care lung ultrasound. Intensive Care Med 2012;38(4):577–91.
[4] Smith M, Hayward S, Innes S, Miller A. Point-of-care lung ultrasound in patients with COVID-19–a narrative review. Anaesthesia 2020;75(8):1096–104.
[5] Soldati G, Smargiassi A, Inchingolo R, Buonsenso D, Perrone T, Briganti DF, Perlini S, Torri E, Mariani A, Mossolani EE, et al. Proposal for international standardization of the use of lung ultrasound for patients with COVID-19: A simple, quantitative, reproducible method. J Ultrasound Med 2020.
[6] Soldati G, Smargiassi A, Inchingolo R, Buonsenso D, Perrone T, Briganti DF, Perlini S, Torri E, Mariani A, Mossolani EE, et al. Is there a role for lung ultrasound during the COVID-19 pandemic? J Ultrasound Med 2020.
[7] Lichtenstein DA, Meziere GA. Relevance of lung ultrasound in the diagnosis of acute respiratory failure*: the BLUE protocol. Chest 2008;134(1):117–25.
[8] Carrer L, Donini E, Marinelli D, Zanetti M, Mento F, Torri E, Smargiassi A, Inchingolo R, Soldati G, Demi L, Bovolo F. Automatic pleural line extraction and COVID-19 scoring from lung ultrasound data. IEEE Trans Ultrason Ferroelectr Freq Control 2020;67(11):2207–17.
[9] Hou D, Hou R, Hou J. Interpretable saab subspace network for COVID-19 lung ultrasound screening. In: 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). 2020, p. 393–8.
[10] Roy S, Menapace W, Oei S, Luijten B, Fini E, Saltori C, Huijben I, Chennakeshava N, Mento F, Sentelli A, Peschiera E. Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. IEEE Trans Med Imaging 2020;39(8):2676–87.
[11] Arntfield R, VanBerlo B, Alaifan T, Phelps N, White M, Chaudhary R, Ho J, Wu D. 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 2021;11(3):e045120.
[12] Chollet F. Xception: Deep learning with depthwise separable convolutions, In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017:.
[13] Zhang J, Chng C, Chen X, Wu C, Zhang M, Xue Y, Jiang J, Chui C. Detection and classification of pneumonia from lung ultrasound images. In: 2020 5th International Conference on Communication, Image and Signal Processing (CCISP). 2020, p. 294–8.
[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv preprint arXiv:1409.1556.
[15] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016; p. 770–778.
[16] Tan M, Le Q. Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. 2019, p. 6105–14.
[17] Born J, Wiedemann N, Cossio M, Buhre C, Brändle G, Leidermann K, Aujayeb A, Moor M, Rieck B, Borgwardt K. Accelerating detection of lung pathologies with explainable ultrasound image analysis. Appl Sci 2021;11(2):672–95.
[18] Zoph B, Vasudevan V, Shlens J, Le Q. Learning transferable architectures for scalable image recognition, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018; p. 8697–8710.
[19] Zhou Z, Sodha V, Siddiquee M, Feng R, Tajbakhsh N, Gotway M, Liang J. Models genesis: Generic autodidactic models for 3d medical image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 2019, p. 384–93.
[20] Ng M-Y, Lee EY, Yang J, Yang F, Li X, Wang H, Lui MM-s, Lo CS-Y, Leung B, Khong P-L, et al. Imaging profile of the COVID-19 infection: radiologic findings and literature review. Radiol Cardiothoracic Imaging 2020;2(1):e200034.
[21] Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, Xue Z, Shi Y. Lung infection quantification of COVID-19 in CT images with deep learning. 2020, arXiv preprint arXiv:2003.04655.
[22] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017; p. 6299–6308.
[23] Zach C, Pock T, Bischof H. A duality based approach for realtime tv-l 1 optical flow. In: Joint Pattern Recognition Symposium. 2007, p. 214–23.
[24] Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection, In: Proceedings of the IEEE International Conference on Computer Vision, 2017; p. 2980–2988.
[25] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Springer; 2014, p. 818–33.
[26] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Object detectors emerge in deep scene CNNs. 2014, arXiv preprint arXiv:1412.6856.