Appendix

A. Analysis of labeled severity scores

Figure A1 shows that the average severity score correlates weakly with respiratory rate and oxygen saturation.

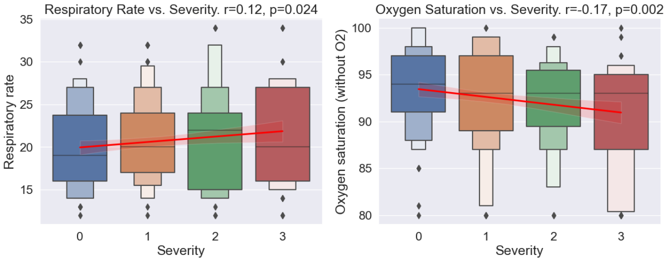


Fig. A1: Labeled severity vs Respiratory Rate (left) and Oxygen Saturation (right). The red regression line indicates the correlation between the two variables. Both effects are statistically significant.

B. Schemata provided with the CovEcho model

Joseph et al. [21] devise a scheme to score the frame quality based on the detected landmarks. 30 points are given if at least pleura is detected, 15 and 10 for rib and shadow respectively, and 45 points for either A-lines, B-lines, B patch or consolidations, yielding a maximum of 100.

Furthermore, they provide a scheme to derive a severity score from the detected patterns:

0: A-lines are present.
1: Single or multiple B-lines are detected.
2: Confluent appearance of B-lines (B patch)
3: Degraded by the appearance of consolidations due to the effusion in between two pleural surfaces.
4: Air bronchograms.

Figure A2 shows the confusion matrix when comparing our severity scores to the predicted severity scores derived with this scheme. The correlation is low (Spearman $\rho = 0.1$) and the CovEcho severity score underestimates the severity score.

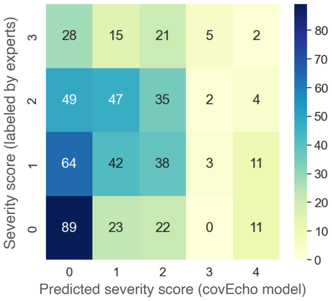


Fig. A2: Confusion matrix

C. Feature importance of clinical variables in RF

The RF (Random Forest) model trained to predict the PCR test result solely from clinical variables can be analyzed with feature importance techniques. Figure A3 visualises the top 25 most important features and reveals that the top four most important features are CBC data. The two most important by a notable margin are lactate dehydrogenase (LDH) and C-reactive protein (CRP) values. This confirms the results in section IV-C that show a substantial decrease in accuracy when witholding blood test data.

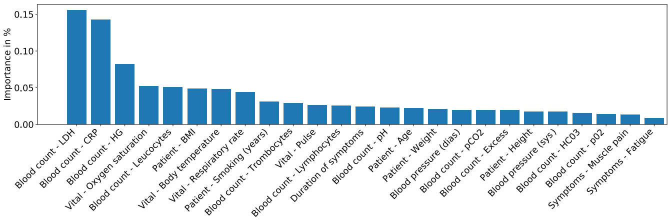


Fig. A3: Feature significance of the RF classifier for all features. Shown are only the top 25 most important features of the 66 total.

D. Training of convolutional neural networks

Following [7], the models are warm-started from Imagenet [18] and we only finetune the last six layers. We use Adam optimizer with a learning rate of 0.0001 and cross-entropy loss. The hyperparameters and layer sizes were tuned with the W&B [6]. The full network was trained for up to 50 epochs in 5-fold cross validation. Furthermore, we found it advantageous to opt for a "warm start" of the network by pretraining on the public POCUS dataset by [7] (only including healthy and COVID-19-infected patients); however, this is only possible when training only on images, since the POCUS dataset does not provide sufficient clinical variables. Here, we augment data with random rotations (20% in either direction), random contrast (factor of 0.2) and random zoom to prevent the network from finding a way to classify the images based on the US probe borders.

We further proposed to merge images with clinical variables as input. In this case, the clinical variables are encoded as a normalized vector and passed through one fully-connected layer before being concatenated with the output of the last convolutional layer. The clinical variables provided to the model comprise all variables collected at admission, including the variables listed in Table I, together with further variables that are not listed in Table I for the sake of simplicity, specifically (for the full table see Supplementary Material):

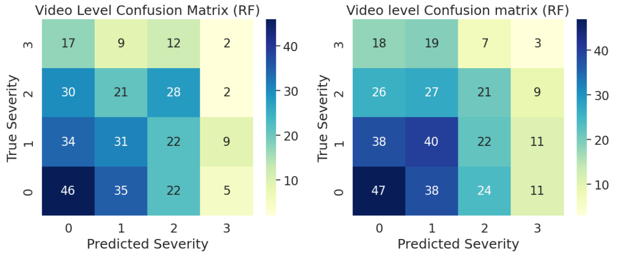
- Inclusion criteria: "Fever or chills","Cough","Difficulty breathing","Loss of taste", "Loss of smell", "Sore throat","Congestion","Runny nose"
- Additional collected symptoms: "Earache", "Wheezing", "Joint pain", "Dyspnea", "Decreased consciousness", "Confusion", "Abdominal pain", "Nausea", "Vomitting", "Diarrhea", "Skin rash", "Lymphadenopathy", "Ageusia/Dysgeusia", "Anosmia/Hyposmia", "Emergency case"
- Pulmonary diseases: "DPLD", "Cystic fibrosis", "Pneumothorax", "Tuberculosis", "Dyspnea", "Other"

E. Severity score prediction

Figure A4 presents the confusion matrices for predicting the disease severity from the CovEcho or ICLUS models. The severity tends to be underestimated.

Model	Images			Images and Features			Images and Features without Blood Test		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
POCOVID-Net	**73.4%**$_{\pm 5}$	**0.90**$_{\pm 0.13}$	0.55$_{\pm 0.24}$	78.4%$_{\pm 9}$	0.87$_{\pm 0.12}$	0.69$_{\pm 0.19}$	78.4%$_{\pm 8}$	0.77$_{\pm 0.14}$	0.80$_{\pm 0.12}$
ResNet50	63.2%$_{\pm 14}$	0.77$_{\pm 0.33}$	0.50$_{\pm 0.35}$	**80.3%**$_{\pm 13}$	0.67$_{\pm 0.30}$	**0.93**$_{\pm 0.08}$	**83.4%**$_{\pm 5}$	**0.84**$_{\pm 0.01}$	0.83$_{\pm 0.11}$
NASNetMobile	58.2%$_{\pm 7}$	0.61$_{\pm 0.25}$	**0.90**$_{\pm 0.13}$	78.5%$_{\pm 17}$	0.67$_{\pm 0.30}$	0.90$_{\pm 0.13}$	75.0%$_{\pm 17}$	0.60$_{\pm 0.38}$	**0.89**$_{\pm 0.09}$
MobileNetV2	68.4%$_{\pm 10}$	0.64$_{\pm 0.20}$	0.73$_{\pm 0.13}$	78.4%$_{\pm 9}$	**0.88**$_{\pm 0.15}$	0.69$_{\pm 0.22}$	78.3%$_{\pm 13}$	0.70$_{\pm 0.20}$	0.87$_{\pm 0.12}$
EfficientNetB7	71.7%$_{\pm 9}$	0.77$_{\pm 0.17}$	0.65$_{\pm 0.25}$	76.3%$_{\pm 10}$	0.64$_{\pm 0.20}$	0.89$_{\pm 0.09}$	76.8%$_{\pm 9}$	0.81$_{\pm 0.12}$	0.73$_{\pm 0.27}$

TABLE A1: Performance of different network architectures on the three configurations of the dataset together with the standard deviation over the five folds. The results of a late fusion between the best three networks can be seen in the last row. Highlighted in bold are the highest values per column.



(a) CovEcho-based **(b)** segmentation-based

Fig. A4: Confusion matrix of the best performing model for the video level severity score prediction using the CovEcho model class counts and the segmentation class areas respectively.

F. Results for image-based models for diagnosis

Table A1 and Table A2 list the same results as Table II but including standard deviations.

Model	Accuracy	Sensitivity	Specificity
LR	0.66$_{\pm 0.19}$	0.60$_{\pm 0.16}$	0.73$_{\pm 0.22}$
RF	0.54$_{\pm 0.04}$	0.66$_{\pm 0.13}$	0.42$_{\pm 0.17}$
SVM	0.66$_{\pm 0.19}$	0.60$_{\pm 0.16}$	0.73$_{\pm 0.22}$

(a) covEcho detection model

Model	Accuracy	Sensitivity	Specificity
LR	0.79$_{\pm 0.04}$	0.82$_{\pm 0.12}$	0.77$_{\pm 0.19}$
RF	0.74$_{\pm 0.04}$	0.72$_{\pm 0.08}$	0.77$_{\pm 0.09}$
SVM	0.79$_{\pm 0.04}$	0.82$_{\pm 0.12}$	0.77$_{\pm 0.19}$

(b) ICLUS segmentation model

TABLE A2: COVID-19 detection accuracy of models trained on segmentation class area (patient-level results).

G. Relation between segmented pixel count and COVID-19 diagnosis

Figure A5 shows the distribution of pixel count by class. A-lines are detected less frequently for COVID-19 positive cases, whereas B-lines and white lungs are more frequent.

H. Human diagnosis of COVID-19 from LUS

In Table A3 the results of human diagnosis are given for all tested aggregation schemes. In particular, we tested to average the severity score and to apply different thresholds, or to count the number of videos with B-lines occurring. The highest correspondence to the PCR test result is achieved when assuming that patients with B-lines in more than 3 videos are COVID-positive.

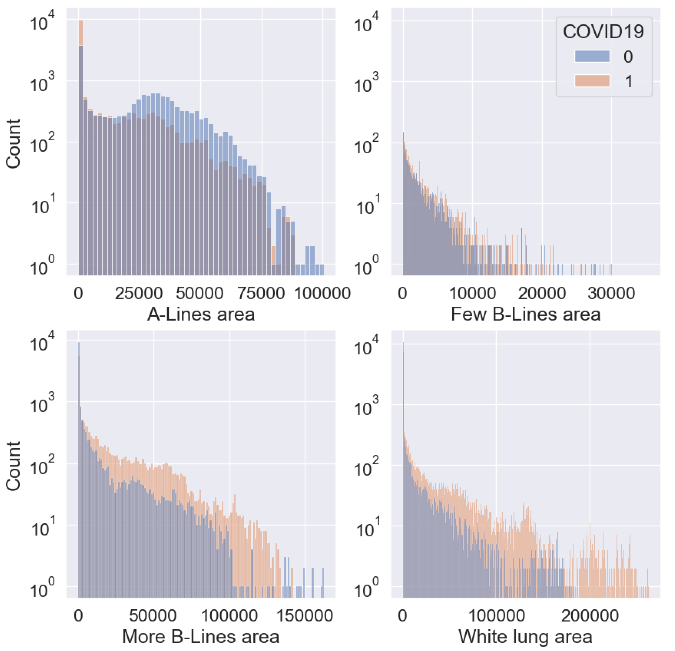


Fig. A5: One histogram for each class assigned by the ICLUS model summarizing the pixel counts in all frames. The plots use the logarithmic scale on the y-axis for better visibility. All the distributions are significantly different when split by the COVID-19 variable ($p < 0.001$).

Level	Rule for diagnosis	Accuracy	Sensitivity	Specificity
Patient	Average severity ≥ 0.5	0.49	0.85	0.10
	Average severity ≥ 1	0.54	0.58	0.50
	Average severity ≥ 2	0.49	0.18	0.83
	B-lines in > 1 videos	0.63	0.79	0.47
	B-lines in > 2 videos	0.63	0.73	0.53
	B-lines in > 3 videos	**0.65**	0.58	0.73
Video	B-lines visible	0.59	0.71	0.44
	Severity > 0	0.51	0.66	0.33
	Severity > 1	0.48	0.36	0.64

TABLE A3: Predicting the PCR test result via manual severity assessment from LUS videos. The best match is achieved with B-line counting