

Hierarchy-Aware and Anatomy-Guided Learning for Lung Ultrasound Video Classification

Alya Almsouti¹[0009–0005–6349–7231], Lotfi Mecharbat¹, Noha Aboukhater²,
Yousef Alabrach³, Siddiq Anwar⁴, Andre Kumar⁵, Ibrahim Almakky¹, and
Mohammad Yaqub¹

¹ Mohamed bin Zayed University of Artificial Intelligence, UAE

alya.alsouti@mbzuai.ac.ae

² SEHA, UAE

³ Sheikh Khalifa Medical City, UAE

⁴ Khalifa University, UAE

⁵ Stanford University, USA

Abstract. Lung ultrasound (LUS) is a bedside tool for assessing pulmonary edema in patients at risk due to heart failure or impaired kidney function. However, automated LUS analysis remains challenging because of speckle noise, imaging artifacts, and operator-dependent acquisition variability. In this work, we present a deep learning framework for multi-class LUS video classification that explores two components: hierarchy-aware training, and anatomy-guided learning. Starting from a strong baseline, we introduce hierarchical training strategies and then introduce pleural line mask supervision to guide model attention toward anatomically relevant regions. We study four clinically relevant classes—healthy, B-lines, consolidations, and mixed B-lines with consolidations—using an open-access dataset of 1,886 videos from 219 patients, evaluated with patient-level five-fold cross-validation. Results show that hierarchy-aware training improves pathological separation relative to flat classification, while mask-guided attention supervision achieves the highest mean macro-F1 of 65.7% and produces more localized attention patterns. Transfer experiments on the external COVID-BLUEs dataset further show competitive and parameter-efficient adaptation while preserving pleural-focused attention behavior. These findings suggest that combining clinically structured objectives with anatomy-guided supervision is a practical approach to robust, interpretable LUS video analysis. Code and model implementations are available at <https://github.com/Alya-Almsouti/LUS-video-classification>.

Keywords: Lung ultrasound · Multi-class classification · Hierarchical learning · Attention supervision · Pulmonary edema

1 Introduction

Pulmonary edema is the accumulation of excess fluid in the lung interstitium and alveolar spaces, which impairs gas exchange. It is a frequent and clinically

significant complication in patients with heart failure, acute kidney injury (AKI), chronic kidney disease (CKD), and dialysis-dependent renal failure. Importantly, imaging signs of pulmonary congestion can appear before overt respiratory symptoms, making early detection essential for timely intervention and optimized fluid and treatment management [10, 15, 23]. In clinical settings, pulmonary edema is typically assessed using physical examination, lung auscultation, chest radiography, serum biomarkers, and bedside ultrasound; however, while these tools are complementary, ultrasound is the most sensitive modality for early extravascular lung water [4, 14, 17, 24].

Lung ultrasound (LUS), widely used within point-of-care ultrasound (POCUS), provides a bedside, radiation-free method for real-time assessment of pulmonary congestion, allowing for repeated evaluations as clinical conditions evolve [14, 4, 17]. This makes it particularly well-suited to both acute and longitudinal care settings, where dynamic fluid management decisions are required. Key findings in LUS are centered on pleural line appearance and artifact patterns. In a normally aerated lung, an intact pleural surface generates repeated horizontal reverberation artifacts (A-lines; Fig. 1a). As lung fluid increases and aeration decreases, vertical comet-tail artifacts (B-lines; Fig. 1b) emerge from the pleural line and extend to the bottom of the scan. With more advanced pathology, when fluid accumulates within the alveoli, subpleural tissue-like patterns appear (consolidations), often accompanied by pleural irregularity (Fig. 1c) [5, 25, 26]. Despite its clinical utility, LUS interpretation remains operator-dependent and is affected by ultrasound-specific noise, acquisition variability, and reader experience, making robust interpretation challenging for novice sonographers in busy workflows. This motivates automated LUS analysis as both a clinical decision-support tool and an educational aid. The main contributions of this work are:

- A LUS-specific framework that adapts hierarchy-aware objectives and mask-guided attention supervision for video classification.
- A SAM2-based semi-automatic video annotation pipeline that substantially reduced labeling time and enabled curation of a clinician-annotated pleural line segmentation dataset.
- Improved classification and interpretability: best multi-class performance, with attention maps more consistently focused on pleural and subpleural regions.
- Transfer experiments on the external COVID-BLueS severity and COVID detection tasks show competitive, parameter-efficient transfer while preserving pleural-focused attention behavior.

2 Related Work

Classical Machine Learning and Model-Based Methods. Early approaches for lung ultrasound (LUS) relied on handcrafted features and explicit geometric priors, with pleural line detection as a key task [5, 6, 25]. Representative pipelines used HMM-based modeling with Viterbi tracking, morphological and

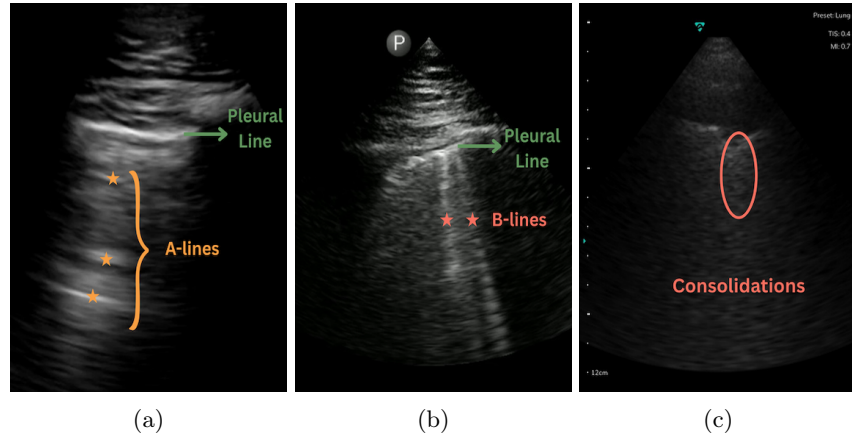


Fig. 1: Lung ultrasound artifacts: (a) A-lines and pleural line; (b) B-lines and pleural line; (c) Consolidations.

adaptive filtering for A-line enhancement, and Radon-transform-based line detection. While these approaches offered a degree of interpretability, their performance was often degraded by speckle noise, probe-angle variability, and cross-device domain shifts, ultimately limiting their generalization and scalability.

Supervised Deep Learning Methods. During the COVID-19 pandemic, LUS research shifted toward supervised deep learning for diagnosis, localization, and severity assessment. Early CNN-based approaches, such as those by Roy et al. and POCOVIDNet, leveraged class activation maps and confidence-based heuristics for post hoc interpretation, but offered limited spatial precision and weak anatomical grounding [4, 21]. Subsequent work moved beyond binary COVID-19 classification toward artifact localization, segmentation, and multi-class analysis. Detection-based models and U-Net variants were increasingly used to localize and segment B-lines and pleural patterns [14, 24]. However, many of these studies relied on private datasets, which constrained reproducibility [1, 12, 19, 27].

Representation Learning. More recent work has targeted the domain gap between natural and ultrasound images through ultrasound-specific representation learning. HiCo, USFM, spatio-temporal contrastive learning, and ultrasound foundation models improved transferability [7, 8, 13, 28]. However, key limitations remain for pulmonary edema assessment: many studies still emphasize frame-level predictions, ignoring important temporal information, and rely on private datasets with limited reproducibility. In addition, many end-to-end systems still provide limited anatomy-guided interpretability, with insufficient explicit supervision toward pleural line regions.

Attention Supervision and Hierarchical Learning. Mask- or localization-guided supervision has been used to encourage classifiers to attend to clinically relevant regions, including ViT-based attention supervision methods such as LGM-ViT and SPAN [3, 18]. Hierarchical classification has also been used to exploit clinically meaningful label structure, such as normal-vs-abnormal separation followed by subtype classification [22]. Our work adapts these ideas to video-level LUS classification, using pleural line masks as anatomical supervision for multi-class pulmonary edema assessment.

Rationale for Hierarchical and Anatomy-Guided Learning. The clinical decision process is naturally hierarchical: separating healthy from pathological findings is generally easier than distinguishing among pathological subtypes. This motivates hierarchy-aware objectives that encode the diagnostic structure during training. At the same time, clinicians visually anchor interpretation around the pleural line and adjacent subpleural region, where B-lines originate and consolidation patterns often emerge. This motivates anatomy-guided learning that explicitly guides model attention toward clinically relevant regions.

3 Methodology

Given a video consisting of N frames, we denote the input sequence as $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, where each frame $\mathbf{x}_t \in \mathbb{R}^{3 \times H \times W}$, $t = 1, \dots, N$ represents an RGB ultrasound image. Each frame is processed independently by a frame encoder $E(\cdot)$, which maps the input image to a d -dimensional latent representation $\mathbf{z}_t = E(\mathbf{x}_t)$, $\mathbf{z}_t \in \mathbb{R}^d$. The resulting sequence of frame-level embeddings is denoted as $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N) \in \mathbb{R}^{N \times d}$. To obtain one video-level representation, we apply a temporal aggregation function $T(\cdot)$ over the sequence $\mathbf{v} = T(\mathbf{Z})$.

Finally, the aggregated video-level representation is passed to a classifier head to produce the predicted logits $\hat{\mathbf{y}} = g(\mathbf{v})$. For classification, the training objective is defined as cross-entropy $\mathcal{L}_{\text{cls}}$ between predicted logits and the ground-truth label y .

3.1 Hierarchical Training

Because separating healthy from abnormal cases is clinically easier than distinguishing among abnormal subtypes, we encode this diagnostic structure directly in the training objective.

Hard Hierarchical Classification. In the hard hierarchical setting, separate binary models are trained for different sub-tasks and combined only at inference time. We evaluate two variants.

Two-stage hierarchical strategy. In the first strategy, two independent classifiers are trained: healthy vs. B-lines and healthy vs. consolidations. This yields

$$\hat{y}_{\text{HB}} = g_{\text{HB}}(T(E(\mathbf{X}))), \quad \hat{y}_{\text{HC}} = g_{\text{HC}}(T(E(\mathbf{X}))). \quad (1)$$

At inference, the final label is determined from the two outputs: both healthy $\rightarrow$ healthy; only one pathological output $\rightarrow$ that pathology; both pathological outputs $\rightarrow$ B-lines + consolidations.

One-versus-all strategy. The second strategy uses one-versus-all training with three classifiers: healthy-vs-not-healthy, B-lines-vs-not-B-lines, and consolidations-vs-not-consolidations, yielding $\hat{y}_{\text{H}}, \hat{y}_{\text{B}}, \hat{y}_{\text{C}}$. During inference, if $\hat{y}_{\text{H}}$ is healthy, the final label is healthy; otherwise, $\hat{y}_{\text{B}}$ and $\hat{y}_{\text{C}}$ determine the pathology label. If both pathological classifiers are negative, healthy is assigned as fallback.

Soft Hierarchical Training with Multiple Heads. In the soft hierarchical setting, a single shared model is trained with two task-specific heads. The temporal aggregator produces two video representations: $\mathbf{v}_{\text{healthy}}$ and $\mathbf{v}_{\text{path}}$, which are processed by $g_{\text{healthy}}(\cdot)$ and $g_{\text{path}}(\cdot)$, respectively. The first head predicts healthy vs. pathological and the second head predicts the pathology subtype:

$$\hat{\mathbf{y}}_{\text{healthy}} = g_{\text{healthy}}(\mathbf{v}_{\text{healthy}}), \quad \hat{\mathbf{y}}_{\text{path}} = g_{\text{path}}(\mathbf{v}_{\text{path}}) \quad (2)$$

where $\hat{\mathbf{y}}_{\text{healthy}} \in \mathbb{R}^2$ and $\hat{\mathbf{y}}_{\text{path}} \in \mathbb{R}^3$.

Training uses a hierarchical loss. Let $y_{\text{healthy}} \in \{0, 1\}$ denote healthy (0) vs. pathological (1). Additionally, let y_{path} be the ground-truth pathological class. The binary loss and subtypes loss are:

$$\mathcal{L}_{\text{healthy}} = \text{CE}(\hat{\mathbf{y}}_{\text{healthy}}, y_{\text{healthy}}), \quad \mathcal{L}_{\text{path}} = \text{CE}(\hat{\mathbf{y}}_{\text{path}}, y_{\text{path}}). \quad (3)$$

The total training loss is then defined as

$$\mathcal{L}(y) = \begin{cases} \lambda_{\text{healthy}} \mathcal{L}_{\text{healthy}}, & \text{if } y_{\text{healthy}} = 0, \\ \lambda_{\text{healthy}} \mathcal{L}_{\text{healthy}} + \lambda_{\text{path}} \mathcal{L}_{\text{path}}, & \text{if } y_{\text{healthy}} = 1, \end{cases} \quad (4)$$

where λ_{healthy} and λ_{path} are scalar weights that control the relative contribution of the hierarchical loss terms.

3.2 Mask-Guided Attention Supervision

Because clinicians primarily inspect the pleural line and adjacent subpleural region when interpreting LUS, we explicitly guide part of the model attention toward these anatomically relevant areas. This approach is illustrated in Fig. 2. We keep the same classification architecture and add only an attention supervision term to selected ViT heads of a certain block. Let $\mathbf{A} \in \mathbb{R}^{H \times N_{\text{tok}} \times N_{\text{tok}}}$ be the self-attention tensor of a chosen transformer block, with $N_{\text{tok}} = 1 + L$ (one

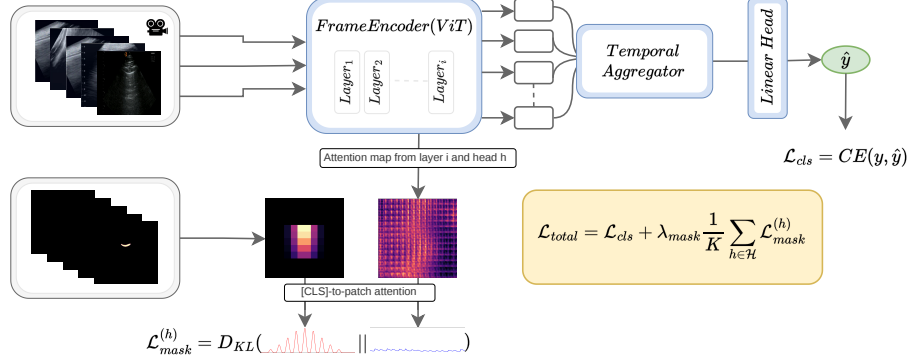


Fig. 2: **Mask-Guided Attention Supervision.** Video frames are first encoded by a ViT backbone and aggregated temporally for final classification, while pleural line masks are converted into target [CLS]-to-patch attention distributions to supervise selected attention heads toward anatomically relevant pleural regions.

[CLS] token and L patch tokens) and H is the number of heads. Supervision is applied to a head subset $\mathcal{H} \subset \{1, \dots, H\}$, where $|\mathcal{H}| = K$.

Given a binary pleural line mask $\mathbf{m} \in \{0, 1\}^{1 \times 224 \times 224}$, we downsample it to patch resolution $\mathbf{m}^p = \text{AvgPool}_{16}(\mathbf{m}) \in \mathbb{R}^{1 \times 14 \times 14}$. We then apply light smoothing, use progressive dilation across heads — where dilation is applied once for every two heads to encourage different heads to focus on neighborhoods of varying size around the pleural line — and convert each mask into a normalized patch-level target distribution constructing a target [CLS]-to-patch attention distribution:

$$\mathbf{q} = \text{Normalize}(\text{Flatten}(\mathbf{m}^p)), \quad \mathbf{q} \in [0, 1]^L. \quad (5)$$

For each supervised head $h \in \mathcal{H}$, we extract [CLS]-to-patch attention: $\mathbf{A}_{cls}^{(h)} = \mathbf{A}[h, 0, 1:] \in \mathbb{R}^L$. Mask supervision is imposed using KL divergence:

$$\mathcal{L}_{mask}^{(h)} = D_{KL}(\mathbf{A}_{cls}^{(h)} \parallel \mathbf{q}). \quad (6)$$

The final mask loss averages over supervised heads $\mathcal{L}_{mask} = \frac{1}{K} \sum_{h \in \mathcal{H}} \mathcal{L}_{mask}^{(h)}$. Finally, the overall objective augments classification loss:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{mask} \mathcal{L}_{mask}, \quad (7)$$

where λ_{mask} controls the strength of the attention supervision. We evaluated Block 12 as a late-stage supervision point, where patch features are expected to be more semantically mature, and Block 8 as an intermediate point to test whether earlier anatomical guidance can shape attention before the final layers. We compared 3 and 6 supervised heads to examine partial versus stronger attention constraints.

4 Experimental Design

4.1 Datasets and Segmentation Dataset Curation

We use the open-access Stanford LUS dataset [16], containing 1,886 videos from 219 patients (up to 12 videos per patient, depending on lung zones). Each video is labeled as one of four classes: healthy, B-lines, consolidations, or B-lines + consolidations. To further assess the transferability and generalizability of our method, we evaluated on the COVID-BLUeS dataset [27]. Because COVID-BLUeS does not provide the same four-class label space as the Stanford LUS dataset, we used the dataset-defined severity prediction and COVID detection tasks. The first task predicts severity of POCUS lung videos on a four-level scale (0–3), ranging from healthy (0) to severe (3), while the second task performs binary COVID detection (COVID vs. non-COVID).

Pleural line annotations were not provided in the open-access LUS dataset; therefore, a semi-automatic pipeline was developed and open-sourced [2] to obtain a segmentation dataset for mask-guided attention supervision. Initially, four medical doctors annotated the videos by splitting the dataset, such that each video was annotated once. Using the annotation platform, annotators placed point prompts on a selected frame from each video. Positive points were used to indicate pleural line regions, while negative points marked non-pleural areas. These prompts were provided to SAM2 [20] to generate segmentation masks, which were iteratively refined by the annotators based on the model output. Once satisfactory masks were obtained, they were aggregated across videos and manually cleaned to form an initial pleural line segmentation dataset. This produced 49,084 masks (15% of total frames) from 329 videos and 43 patients. Next, a U-Net model was trained on these curated masks and used for iterative auto-labeling and refinement (two cycles). The final segmentation set contained 147,342 frames from 941 videos and 200 patients (45% of total frames). Although no independent pleural line ground truth was available for Dice/IoU evaluation, the generated masks underwent manual quality control and cleaning to ensure accurate coverage of the pleural line. Since the masks are downsampled to (14×14) patch resolution before constructing the attention target, the supervision relies on accurate localization of the pleural line rather than pixel-perfect boundary delineation. This semi-automatic pipeline substantially reduced labeling time by replacing dense, frame-wise manual segmentation with sparse point prompting on selected frames.

4.2 Evaluation and Training Setup

Evaluation used patient-level five-fold cross-validation with precomputed splits. Each video was represented as a fixed-length sequence of $N = 150$ frames. For videos exceeding this length, frames were sampled according to a Gaussian distribution over the temporal axis centered at the midpoint, and videos with fewer than N frames were padded. All selected frames were then resized to 224×224 .

Additional details on the choice of fixed sequence length and temporal sampling strategy are provided in the supplementary material.

During training, we applied data augmentation to improve robustness, including random resized cropping (scale: 0.7–1.0), random horizontal flipping, and random rotations up to 10° . To better match ultrasound acquisition conditions, speckle noise was randomly injected into training samples. All images were normalized using the ImageNet mean and standard deviation. Validation data underwent only resizing and normalization.

Training ran for up to 100 epochs with a 3-epoch probing stage (frame encoder frozen), followed by end-to-end fine-tuning. AdamW optimizer (weight decay 0.01) was used with cosine annealing. Learning rates were 1×10^{-5} for non-encoder parameters and 5×10^{-6} for encoder parameters. Models were trained with cross-entropy, label smoothing (0.1), and fold-specific balanced class weights. Validation was monitored per epoch using macro-F1, weighted-F1, and balanced accuracy. Early stopping used macro-F1 (patience 20), and the best macro-F1 checkpoint per fold was reported. Unless noted otherwise, all results are validation-set metrics summarized over the five folds.

4.3 Baselines

In our baseline experiments, we varied both the frame encoder $E(\cdot)$ and the temporal aggregation function $T(\cdot)$ to identify which backbone and temporal aggregation strategy were most effective for the ultrasound modality and video-level representation learning. We compared four representative frame encoders: (i) an Ultrasound Foundation Model (USFM) [13], (ii) HiCo [28], (iii) ResNet-18 [11], and (iv) ViT-Small [9]. For all variants, only the backbone image encoder was used, and any task-specific classification heads were removed. USFM is a self-supervised ultrasound foundation model trained on a large dataset consisting of 2,187,915 images from different organs including the lung [13]. HiCo is a self-supervised ultrasound representation learning framework trained on 23,231 images extracted from lung and liver ultrasound videos [28].

Given frame-level embeddings $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)$, we evaluated four video-level aggregation strategies: (i) **Majority Voting (MV)**, where each frame was classified independently and the final video-level prediction corresponded to the most frequent class among frame-level predictions; during training, the video label was applied to each frame; (ii) **Mean Pooling (MP)**, where temporal information was aggregated as

$$\mathbf{v} = \frac{1}{N} \sum_{t=1}^N \mathbf{z}_t; \quad (8)$$

(iii) **Weighted Pooling (WP)**, where normalized attention weights $\{\alpha_t\}_{t=1}^N$ were computed from the frame embeddings to form

$$\mathbf{v} = \sum_{t=1}^N \alpha_t \mathbf{z}_t, \quad \text{with } \sum_{t=1}^N \alpha_t = 1; \quad (9)$$

(iv) **Transformer-Based Aggregation (T)**, where a temporal transformer was applied to $(\mathbf{z}_1, \dots, \mathbf{z}_N)$ and forms the video representation by concatenating the output [CLS] token with the mean-pooled representation of the remaining tokens for final video-level classification.

Table 1: Comparison of hierarchy-aware learning and anatomy-guided learning variants for four-class classification. Results are reported as macro-F1 (mean $\pm$ standard deviation) across folds. Class 0 indicates healthy class, Class 1: B-lines, Class 2: Consolidations, Class 3: B-lines + Consolidations.

Model- Configuration	F1-score				
	Class 0	Class 1	Class 2	Class 3	Macro
USFM-MV	79.8 $\pm$ 8.40	50.0 $\pm$ 10.1	25.6 $\pm$ 12.0	37.8 $\pm$ 9.00	49.8 $\pm$ 6.70
USFM-MP	76.3 $\pm$ 6.30	53.1 $\pm$ 6.60	7.7 $\pm$ 7.80	31.6 $\pm$ 9.50	43.5 $\pm$ 7.20
HiCo-MV	81.2 $\pm$ 6.00	57.5 $\pm$ 5.30	37.0 $\pm$ 7.70	41.6 $\pm$ 15.0	54.1 $\pm$ 2.10
ResNet18-T	89.2 $\pm$ 4.60	56.9 $\pm$ 3.60	35.9 $\pm$ 2.20	41.5 $\pm$ 14.1	56.1 $\pm$ 4.00
ViT-Small-T	91.6 $\pm$ 3.50	67.6 $\pm$ 5.50	51.8 $\pm$ 6.00	40.1 $\pm$ 9.50	62.8 $\pm$ 3.20
<i>Hierarchy-aware - Hard hierarchy:</i>					
Two-Stage	98.4$\pm$1.10	39.2 $\pm$ 8.10	10.1 $\pm$ 7.10	38.4 $\pm$ 6.20	46.5 $\pm$ 2.20
One-vs-all	<u>92.4 $\pm$ 3.10</u>	<u>71.3 $\pm$ 7.30</u>	40.0 $\pm$ 9.30	52.2$\pm$5.80	64.5 $\pm$ 1.60
<i>Hierarchy-aware - Soft hierarchy ($\lambda_{\text{path}} = 1$):</i>					
$\lambda_{\text{healthy}} = 1$	87.2 $\pm$ 7.50	63.3 $\pm$ 8.70	41.1 $\pm$ 3.30	46.0 $\pm$ 10.3	59.4 $\pm$ 3.00
$\lambda_{\text{healthy}} = 2$	89.5 $\pm$ 5.20	68.6 $\pm$ 6.00	51.0 $\pm$ 4.50	43.0 $\pm$ 9.10	63.4 $\pm$ 3.00
<i>Mask-guided attention supervision ($\lambda_{\text{mask}} = 2$):</i>					
Block 8, 3 Heads	92.1 $\pm$ 4.50	70.4 $\pm$ 7.90	53.0$\pm$7.2	45.5 $\pm$ 8.30	65.7$\pm$2.40
Block 12, 3 Heads	92.2 $\pm$ 3.80	71.4$\pm$6.60	48.3 $\pm$ 6.00	<u>49.0 $\pm$ 7.00</u>	<u>65.2 $\pm$ 1.10</u>
Block 12, 6 Heads	90.2 $\pm$ 6.00	68.2 $\pm$ 7.4	45.6 $\pm$ 6.20	47.1 $\pm$ 10.1	62.8 $\pm$ 2.70
<i>Hierarchy+ anatomy-guided</i>					
	90.4 $\pm$ 5.00	69.0 $\pm$ 7.40	51.0 $\pm$ 1.4	44.2 $\pm$ 12.9	64.0 $\pm$ 2.50

5 Results and Discussion

Results obtained using different frame encoders and temporal aggregation methods are presented in Table 1. To improve readability, the table reports the main configurations used for comparison, while the full evaluation of all backbones across all temporal aggregation strategies is provided in the supplementary material. Two key observations emerge. First, the choice of frame encoder had a greater impact on performance than the temporal aggregation method. USFM consistently performed the worst across all aggregation methods, followed by HiCo. ResNet18 showed variable performance, while ViT-Small consistently achieved the best results under all settings. This highlights that the quality of latent embeddings is a key factor in determining video classification model performance. Notably, although USFM and HiCo were pretrained on ultrasound scans (including lung), the ViT model initialized with ImageNet weights demon-

strated superior transferability for lung pathology discrimination. This may reflect a mismatch between ultrasound pretraining and the target task: USFM learns broad multi-organ ultrasound representations, while HiCo was pretrained on a smaller lung/liver dataset, whereas ViT-Small may provide more general visual features that adapt effectively during fine-tuning.

Second, the class-wise F1 scores were consistent across all backbone and aggregation configurations. The Healthy class (Class 0) achieved the highest performance, followed by B-lines (Class 1), while the lowest performance was observed for Consolidations and the mixed class (B-lines + Consolidations; Classes 2 and 3). This indicates that the lower performance for these classes may be due to the encoder learning representations of healthy patterns more effectively than those of pathological cases. In particular, extracting sufficiently discriminative latent embeddings for consolidation and mixed patterns appears to have been more challenging, possibly due in part to their lower representation across folds, as shown by the fold-level class distributions in the supplementary material. Moreover, the choice of backbone and temporal aggregator substantially influenced per-class performance. The most notable improvement was observed for Consolidations (Class 2), with an absolute gain of 44 % (from 7.7% to 51.8%). Overall, ViT-Small with transformer-based aggregation was selected as the default frame encoder and temporal aggregator for the rest of the experiments.

5.1 Effect of Hierarchy-Aware Learning

The hierarchical experiments in Table 1 (*Hierarchy-aware*) showed that the way hierarchy was enforced significantly affected performance. The two-stage strategy achieved the highest score for the healthy class (98.4%), likely because it relied on two independent classifiers and predicted healthy only when both agreed, resulting in a stricter and more confident decision boundary. In contrast, performance on B-lines and consolidations was lower, as the models were trained separately against the healthy class without explicitly learning to distinguish between pathologies. This lack of inter-pathology discrimination led to confusion between abnormal classes, reflecting the limitation of optimizing primarily for healthy vs. abnormal separation. On the other hand, the one-versus-all hierarchy achieved the strongest overall hierarchical performance (64.5%). This approach improved Class 1 and substantially boosted Class 3 performance by 12.1%, the highest gain observed across all methods, while maintaining strong healthy detection (92.4%). Compared to the two-stage design, this resulted in a more balanced class-wise performance. These findings suggest that explicitly training each classifier against all alternatives provided a stronger discriminative signal, leading to improved separation between pathological subtypes, particularly for the B-lines class. The observed performance for the soft hierarchy was as follows. With equal weighting ($\lambda_{\text{healthy}} = 1$, $\lambda_{\text{path}} = 1$), performance remained close to the baseline, with no substantial gains. Increasing the loss weight on the healthy class ($\lambda_{\text{healthy}} = 2$) led to an overall improvement, yielding modest gains over both the equal-weight setting and the baseline. Although the one-versus-all strategy achieved higher performance across hierarchical settings, it required

training three separate classifiers, whereas the soft hierarchy relied on a single model, highlighting a trade-off between performance and model complexity.

5.2 Effect of Mask-Guided Attention Supervision

The mask-guided attention supervision results (Table 1-*Mask-guided attention supervision*) showed that enforcing [CLS] attention toward anatomically relevant regions consistently improved performance across configurations. The most notable gains were observed when supervision was applied at Block 8 with 3 heads (heads 0, 1, 2), yielding the best overall performance (65.7%) and the highest improvement for Class 2, suggesting that guiding mid-level representations toward the pleural line was particularly effective for capturing pathology-specific patterns. Applying supervision at Block 12 also improved performance, with the 3-head setting achieving strong results for Class 1 and Class 3, indicating that later-layer refinement further enhanced class-specific discrimination. However, increasing the number of supervised heads to 6 (all heads) led to a drop in performance, suggesting that over-constraining attention may limit the model’s ability to capture complementary cues beyond the pleural line region.

Overall, both Block 8 and Block 12 supervision improved over the ViT-Small-T baseline, with Block 8 using 3 supervised heads achieving the highest mean macro-F1. The drop observed when supervising all 6 heads suggests that partial head supervision is preferable, likely because it guides attention toward pleural regions while preserving other heads for complementary visual cues. This analysis should be interpreted as a targeted ablation of selected supervision locations and head counts, rather than an exhaustive layer-wise and head-wise search.

Because fold-level variance was non-negligible, we further performed paired fold-level comparisons for the main configurations. The mask-guided model improved over the ViT-Small-T baseline on all five folds and achieved the highest mean macro-F1, but the difference did not reach conventional statistical significance with $n = 5$ folds (Wilcoxon signed-rank test, $p = 0.0625$). Therefore, we interpret the gain as a favorable performance trend rather than definitive statistical superiority. Additional confidence intervals, paired tests, and aggregated confusion matrices are provided in the supplementary material.

The effect of mask-guided supervision is illustrated in Fig. 3, which presents qualitative [CLS]-to-patch attention rollouts. The baseline model exhibits relatively diffuse attention, while mask-supervised models show clearer concentration near pleural structures. Taken together, anatomy-guided learning improved localization and yielded modest but consistent classification gains, whereas combining hierarchical learning and mask-guided attention supervision did not provide further improvement over the best anatomy-guided-only setting. One possible explanation is that the hierarchical objective and anatomy-guided loss impose partially overlapping constraints on the shared representation; therefore, when explicit pleural line attention supervision is already applied, the additional hierarchy loss may reduce optimization flexibility rather than provide complementary gains. An additional qualitative example is provided in the supplementary material.

Table 2: Transferability to the COVID-BLUEs severity task. Results are reported on the four-class severity setting (0–3), mean across the 5 folds.

Model	Accuracy	F1	Precision	Recall	Trainable Params
USFM-Majority Voting	31.5	25.3	31.5	23.0	86.14M
HiCo-Majority Voting	30.9	25.3	30.9	22.8	561.64M
ResNet-18-Transformer	45.5	43.7	45.5	46.0	15.73M
ViT-Small-Transformer	46.6	<u>46.0</u>	46.6	49.4	26.15M
<i>Transfer settings from pretrained model (previous task)</i>					
Linear head only (new head)	44.0	40.7	44.0	47.1	0.01M
Temporal aggregator + head fine-tuned	42.1	40.2	42.1	47.4	4.41M
Full fine-tuning	46.6	46.1	46.6	<u>48.2</u>	26.15M

Table 3: Transferability to the COVID-BLUEs COVID detection task. Results are mean across the 5 folds.

Model	Accuracy	F1	Precision	Recall	Trainable Params
ViT-Small-Transformer	61.4	72.7	94.3	59.2	26.15M
<i>Transfer settings from pretrained model (previous task)</i>					
Linear head only (new head)	66.1	70.2	74.2	68.0	0.01M
Temporal aggregator + head fine-tuned	66.2	72.9	83.3	65.6	4.41M
Full fine-tuning	62.0	71.7	88.9	62.0	26.15M

5.3 Transferability to External LUS Severity and COVID Tasks

To evaluate the transferability of the anatomy-guided model, we conducted experiments on an external dataset across two tasks.

For the severity prediction task (Table 2), all frame encoders were paired with their best-performing temporal aggregators identified from the baseline experiments. In addition, the best-performing anatomy-guided model was fine-tuned under different transfer settings to assess its adaptability. The results show that ultrasound-pretrained models (USFM and HiCo) performed poorly in this setting, despite their large capacity, suggesting limited task alignment. In contrast, models trained from scratch on the target dataset (ResNet-18 and ViT-Small) and full fine-tuning of the pretrained anatomy-guided model achieved competitive performance, indicating that leveraging the full model capacity and adapting all parameters was beneficial for this task. Notably, partial transfer strategies remained effective. Fine-tuning only the temporal aggregator or training a new classification head achieved reasonably competitive results while using significantly fewer trainable parameters. This highlights a clear trade-off between performance and efficiency, where lightweight adaptation offers a practical alternative when computational resources are constrained.

For COVID detection (Table 3), a different transfer pattern was observed. Temporal-aggregator+head fine-tuning achieved the best overall performance (F1 72.9, accuracy 66.2) while updating only 4.41M parameters, outperforming

both full fine-tuning (F1 71.7, accuracy 62.0, 26.15M parameters) and the directly trained ViT-Small baseline (F1 72.7, accuracy 61.4). Overall performance differences across methods were modest, but more parameter-efficient strategies consistently matched or exceeded full fine-tuning.

Our results on COVID-BLUEs are not directly comparable to those reported in the original study, as we could not identify the exact cross-validation split files from the released material. Therefore, we generated new patient-level splits, with all videos from a given patient kept within the same fold. For context, the original study reported best patient-level F1 scores of 81% for COVID detection and 51% for severity prediction. Thus, our experiments should be interpreted as an external transfer evaluation on COVID-BLUEs tasks rather than a strict reproduction of the original benchmark.

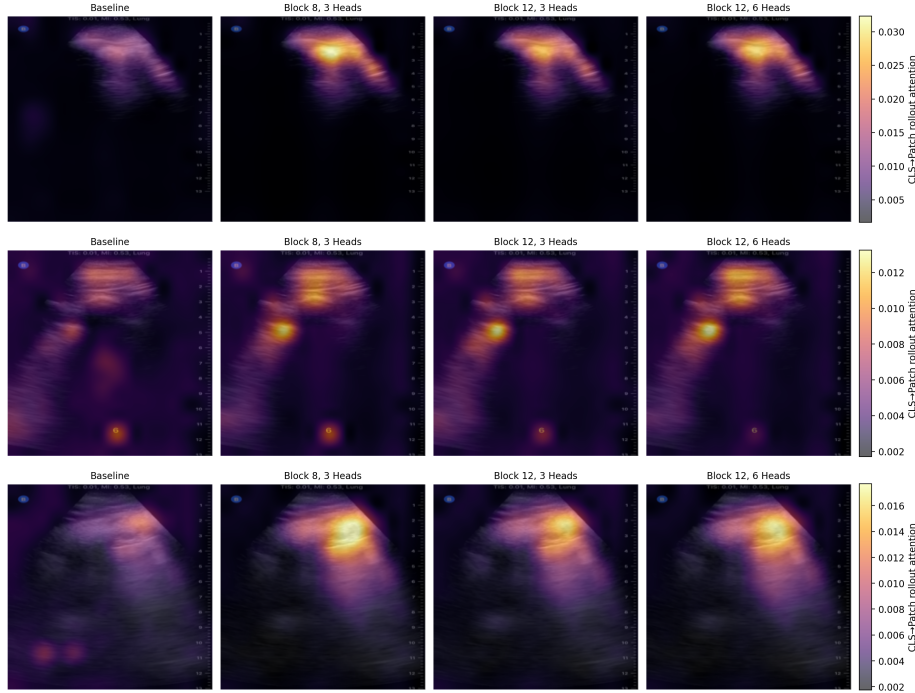


Fig. 3: Comparison of CLS-to-patch attention rollout maps across model configurations (as described in Section 5.2). Each column shows the spatial attention distribution over the input image, highlighting regions that contribute most strongly to the final frame embedding.

Qualitatively (Fig. 4), the directly trained ViT-Small Transformer baseline attended to regions outside clinically relevant areas, indicating less stable anatomical focus. In contrast, attention rollouts from transfer-based mod-

els (temporal-aggregator+head fine-tuning and full fine-tuning) remained consistently concentrated around the pleural line and adjacent subpleural region, even without mask-guided supervision during transfer. This suggests that the pretrained backbone preserved anatomically meaningful attention patterns and effectively generalized them to the external dataset.

6 Limitations

This study is limited by the moderate patient cohort size, class imbalance, and possible label ambiguity, particularly for consolidation and mixed-pathology cases, which were consistently harder to discriminate. During pleural line annotation, clinicians were allowed to flag label disagreements, and 34 of 329 videos (10.4%) were relabeled, suggesting that some cases contained ambiguous pathology patterns. Pleural line masks were generated semi-automatically and manually quality-checked; however, independent dense ground-truth masks and multi-annotator agreement were not available. Although attention supervision improved pleural-region localization, formal clinician evaluation of the attention maps remains future work.

7 Conclusion

This paper presents a LUS video classification framework built on clinically meaningful problem framing, hierarchy-aware training, and mask-guided attention supervision. Under patient-level evaluation on a public Stanford dataset, hierarchical objectives improve multi-class discrimination, and mask-guided attention supervision achieves the highest mean macro-F1 while improving anatomical localization. The results suggest that clinically structured objectives and explicit spatial guidance are practical directions for robust and interpretable pulmonary edema assessment from POCUS videos.

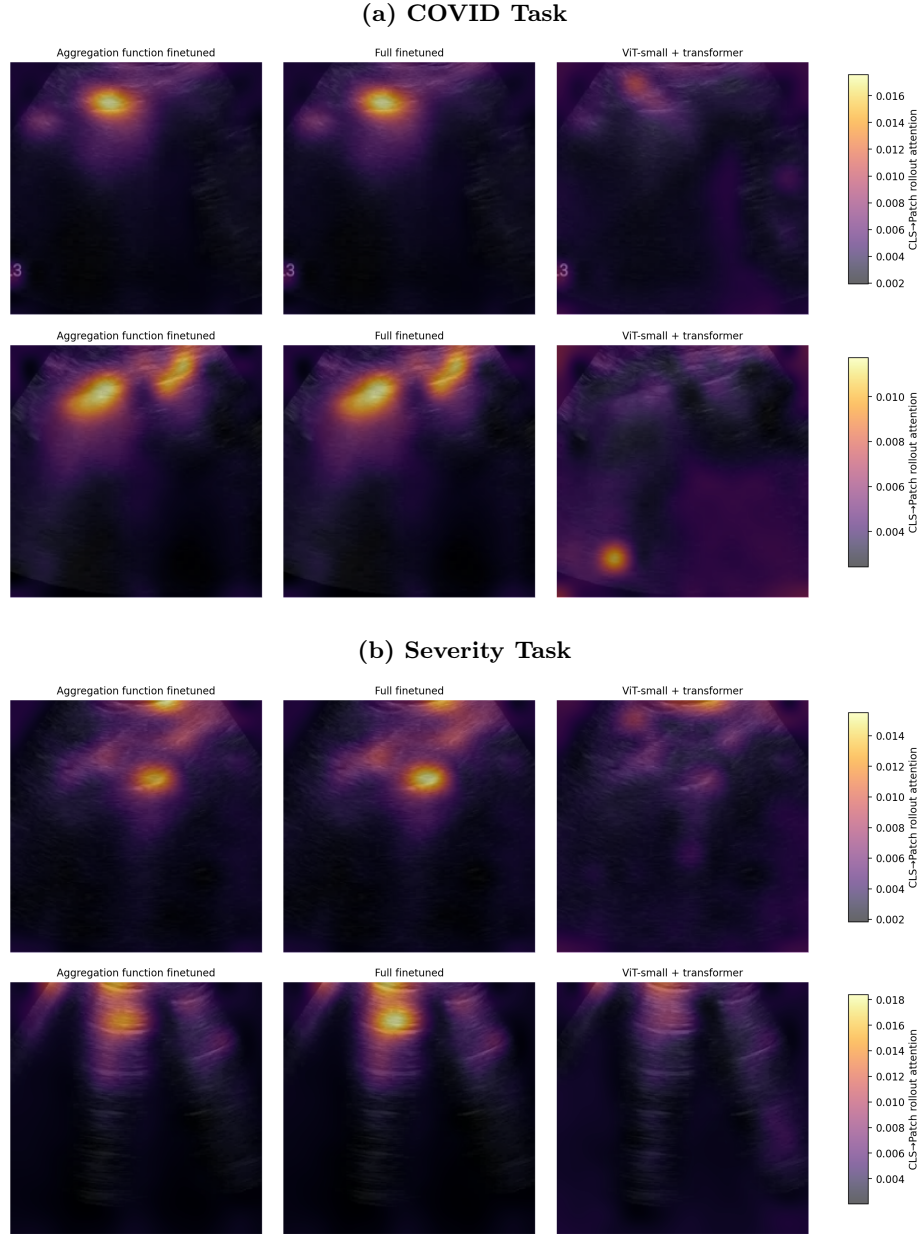


Fig. 4: Comparison of CLS-to-patch attention rollout maps across models finetuned on the COVID-BLUeS dataset. Each column visualizes the spatial attention distribution over the input image, highlighting regions that contribute most to the final frame embedding.

References

1. Abbasi, S., Wahd, A.S., Ghosh, S., Ezzelarab, M., Panicker, M., Chen, Y.T., Jaremko, J.L., Hareendranathan, A.: Improved a-line and b-line detection in lung ultrasound using deep learning with boundary-aware dice loss. *Bioengineering* **12**(3), 311 (2025)
2. Almsouti, A.: Video-segmentation-annotation-using-sam2. <https://github.com/Alya0/Video-Segmentation-Annotation-Using-SAM2> (2026)
3. Ben Itzhak, S., Kiryati, N., Portnoy, O., Mayer, A.: Localization-guided supervision for robust medical image classification by vision transformers. In: *Computer Vision – ECCV 2024 Workshops*. pp. 118–133. Springer (2025). https://doi.org/10.1007/978-3-031-92648-8_8
4. Born, J., Wiedemann, N., Cossio, M., Buhre, C., Brändle, G., Leidermann, K., Goulet, J., Aujayeb, A., Moor, M., Rieck, B., Borgwardt, K.: Accelerating detection of lung pathologies with explainable ultrasound image analysis. *Applied Sciences* **11**(2) (2021). <https://doi.org/10.3390/app11020672>, <https://www.mdpi.com/2076-3417/11/2/672>
5. Carrer, L., Donini, E., Marinelli, D., Zanetti, M., Mento, F., Torri, E., Smargiassi, A., Inchingolo, R., Soldati, G., Demi, L., Bovolo, F., Bruzzone, L.: Automatic pleural line extraction and covid-19 scoring from lung ultrasound data. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **67**(11), 2207–2217 (2020). <https://doi.org/10.1109/TUFFC.2020.3005512>
6. Chen, J., Li, J., He, C., Li, W., Li, Q.: Automated pleural line detection based on radon transform using ultrasound. *Ultrasonic Imaging* **43**(1), 19–28 (2021)
7. Chen, L., Rubin, J., Ouyang, J., Balaraju, N., Patil, S., Mehanian, C., Kulhare, S., Millin, R., Gregory, K.W., Gregory, C.R., et al.: Contrastive self-supervised learning for spatio-temporal analysis of lung ultrasound videos. In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. pp. 1–5. IEEE (2023)
8. Chen, Y., Zhang, C., Liu, L., Feng, C., Dong, C., Luo, Y., Wan, X.: Uscl: Pre-training deep ultrasound image diagnosis model through video contrastive representation learning. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 627–637. Springer (2021)
9. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations* (2021)
10. Faubel, S., Edelstein, C.: Mechanisms and mediators of lung injury after acute kidney injury. *Nature Reviews Nephrology* **12**, 48–60 (2016). <https://doi.org/10.1038/nrneph.2015.158>, <https://doi.org/10.1038/nrneph.2015.158>
11. 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)
12. Howell, L., Ingram, N., Lapham, R., Morrell, A., McLaughlan, J.R.: Deep learning for real-time multi-class segmentation of artefacts in lung ultrasound. *Ultrasonics* **140**, 107251 (2024)
13. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., et al.: Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical image analysis* **96**, 103202 (2024)

14. Joseph, J., Panicker, M.R., Chen, Y.T., Chandrasekharan, K., Mondy, V.C., Ayyappan, A., Valakkada, J., Narayan, K.V.: covecho resource constrained lung ultrasound image analysis tool for faster triaging and active learning. arXiv preprint arXiv:2206.10183 (2022)
15. Kharat, A., Tallaa, F., Lepage, M.A., Trinh, E., Suri, R.S., Mavrakanas, T.A.: Volume status assessment by lung ultrasound in end-stage kidney disease: A systematic review. *Canadian Journal of Kidney Health and Disease* **10**, 20543581231217853 (2023). <https://doi.org/10.1177/20543581231217853>, <https://doi.org/10.1177/20543581231217853>
16. Kumar, A., Nandakishore, P., Gordon, A.J., Baum, E., Madhok, J., Duanmu, Y., Kugler, J.: Creation of an open-access lung ultrasound image database for deep learning and neural network applications. medRxiv (2025). <https://doi.org/10.1101/2025.05.09.25327337>, <https://www.medrxiv.org/content/10.1101/2025.05.09.25327337v1>, preprint (not peer-reviewed)
17. Maw, A.M., Hassanin, A., Ho, P.M., McInnes, M.D., Moss, A., Juarez-Colunga, E., Soni, N.J., Miglioranza, M.H., Platz, E., DeSanto, K., et al.: Diagnostic accuracy of point-of-care lung ultrasonography and chest radiography in adults with symptoms suggestive of acute decompensated heart failure: a systematic review and meta-analysis. *JAMA network open* **2**(3), e190703–e190703 (2019)
18. Miao, K., Gokul, A., Singh, R., Petryk, S., Gonzalez, J.E., Keutzer, K., Darrell, T., Reed, C.J.: Prior knowledge-guided attention in self-supervised vision transformers. arXiv preprint arXiv:2209.03745 (2022)
19. Muñoz, M., Rubio, A., Cosarinsky, G., Cruza, J.F., Camacho, J.: Deep learning-based algorithms for real-time lung ultrasound assisted diagnosis. *Applied Sciences* **14**(24), 11930 (2024)
20. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
21. Roy, S., Menapace, W., Oei, S., Luijten, B., Fini, E., Saltori, C., Huijben, I., Chen-nakeshava, N., Mento, F., Sentelli, A., et al.: Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. *IEEE transactions on medical imaging* **39**(8), 2676–2687 (2020)
22. Salem, M., Abouzahir, S., Ghedira, H., El Saddik, A., Yaqub, M.: Plasmostage: A hierarchical deep learning framework for plasmodium parasite staging in malaria. In: 2025 IEEE Medical Measurements & Applications (MeMeA). pp. 1–6 (2025). <https://doi.org/10.1109/MeMeA65319.2025.11067957>
23. Stack, A.G., Casserly, L.F.: Pulmonary congestion in hemodialysis: an old chestnut worth screening for? *Clinical Journal of the American Society of Nephrology* **8**(8), 1279–1281 (aug 2013). <https://doi.org/10.2215/CJN.06210613>
24. Subramanyam, A., Sucharitha, M.: Multi-classification of lung diseases using lung ultrasound imaging. In: International Conference on Information Systems and Management Science. pp. 510–521. Springer (2023)
25. Susanti, H., Suprijanto: Image processing framework for pleural line (a-line) detection in video lung ultrasonography. In: 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES). pp. 99–102 (2021). <https://doi.org/10.1109/IECBES48179.2021.9398835>
26. Venkatesh, M., Deng, F., Gaillard, F., et al.: Shred sign (lungs). <https://doi.org/10.53347/rID-30808> (2026). <https://doi.org/10.53347/rID-30808>, radiopaedia.org, reference article. Accessed 25 Jan 2026

27. Wiedemann, N., de Korte-De Boer, D., Richter, M., van de Weijer, S., Buhre, C., Eggert, F.A., Aarnoudse, S., Grevendonk, L., Röber, S., Remie, C.M., et al.: Covid-blues-a prospective study on the value of ai in lung ultrasound analysis. *IEEE Journal of Biomedical and Health Informatics* (2025)
28. Zhang, C., Chen, Y., Liu, L., Liu, Q., Zhou, X.: Hico: hierarchical contrastive learning for ultrasound video model pretraining. In: *Proceedings of the Asian Conference on Computer Vision*. pp. 229–246 (2022)