Exploring the Utility of Self-Supervised Pretraining Strategies for the Detection of Absent Lung Sliding in M-Mode Lung Ultrasound

Blake VanBerlo[1,3*], Brian Li[1,3], Alexander Wong[1], Jesse Hoey[1], Robert Arntfield[2,3]
[1]University of Waterloo, [2]Western University, [3]Deep Breathe
bvanberl@uwaterloo.ca

Abstract

Self-supervised pretraining has been observed to improve performance in supervised learning tasks in medical imaging. This study investigates the utility of self-supervised pretraining prior to conducting supervised fine-tuning for the downstream task of lung sliding classification in M-mode lung ultrasound images. We propose a novel pairwise relationship that couples M-mode images constructed from the same B-mode image and investigate the utility of data augmentation procedure specific to M-mode lung ultrasound. The results indicate that self-supervised pretraining yields better performance than full supervision, most notably for feature extractors not initialized with ImageNet-pretrained weights. Moreover, we observe that including a vast volume of unlabelled data results in improved performance on external validation datasets, underscoring the value of self-supervision for improving generalizability in automatic ultrasound interpretation. To the authors' best knowledge, this study is the first to characterize the influence of self-supervised pretraining for M-mode ultrasound.

1. Introduction

Pneumothorax (PTX) is a potentially life-threatening acute condition in which air occupies the space between the pleura of the lungs, resulting in collapse of the lung. Rapid identification of PTX is crucial in emergency, critical, and acute care settings to expedite medical intervention. Point-of-care lung ultrasound (LUS) is a quick, inexpensive, portable, imaging examination that does not expose patients to radiation. Despite its low prevalence compared to chest radiographs, LUS has been shown to exhibit superior diagnostic performance for the diagnosis of PTX [1, 18]. The lung sliding artefact, caused by the normal motion of the pleura, has been described as a means to rule out PTX [16]. Notably, the presence of lung sliding excludes a diagnosis of PTX within the purview of the ultrasound probe [16]. Conversely, PTX is likely present when lung sliding is absent.

Previous studies have demonstrated that deep convolutional neural networks (CNN) can be trained to distinguish between the presence and absence of lung sliding on motion mode (M-mode) ultrasound images [13,25]. Prior studies were limited by the amount of labelled data available for training and evaluation. Furthermore, previous studies initialized their networks using weights pretrained on the ImageNet dataset [9]. Despite the fact that M-mode images are profoundly distinct from the natural images present in ImageNet, it is common for medical imaging studies to leverage ImageNet-pretrained weights. They are publicly available for several common architectures and are able to extract low-level features present in medical images. Unfortunately, there are no publicly available equivalents for M-mode images, let alone LUS.

Self-supervised learning (SSL) is a representation learning strategy applicable in the absence of labelled data. CNNs pretrained using SSL have exhibited superior performance and label efficiency compared to fully supervised counterparts [5, 7, 11, 26]. Broadly, SSL pretraining consists of training a deep neural network to solve a *pretext task*, whose solution can be computed from unlabelled examples. The pretrained weights may be fine-tuned to solve a *downstream task* for which labels are present. This study explores the impact of self-supervised pretraining for the downstream task of detecting absent lung sliding in M-mode LUS, varying the choice of SSL method, weight initialization, data augmentation strategy, and inclusion of unlabelled data. Crucially, we demonstrate that incorporating large volumes of unlabelled M-mode images during the pretraining phase improves the performance of a fine-tuned classifier on external datasets. More specifically, our major contributions are as follows:

- A pairwise relationship for contrastive and non-contrastive learning that is specific to M-mode images
- A data augmentation pipeline specific to M-mode images in the context of pretraining

- A comprehensive investigation of factors that influence the utility of SSL pretraining for the downstream task of absent lung sliding detection, such as label efficiency, ImageNet initialization, and data augmentation
- Evidence that the inclusion of unlabelled data results in improved generalization to external datasets for absent lung sliding detection

Fig. 1 summarizes our methods. To the best of our knowledge, no study has investigated the efficacy of SSL for M-mode ultrasound tasks.

2. Related Work

2.1. Lung Sliding Classification

Multiple studies have explored the use of CNNs for automatically identifying absent lung sliding in LUS M-modes. Jaščur *et al*. [13] were the first to attempt this task. Using a dataset of 48 videos acquired from 48 post-thoracic surgery patients with a single ultrasound probe, their model achieved a sensitivity of 0.82 and specificity of 0.92. VanBerlo *et al*. [25] developed a binary classifier using a dataset of 2535 examples acquired from hundreds of patients with variable probes. When evaluated on a test set of 540 examples, their model attained a sensitivity and specificity of 0.935 and 0.873 respectively.

2.2. Self-supervised Learning in Computer Vision

Self-supervised learning methods in computer vision can be categorized in various manners, based on the pretext task. Generative methods, such as image colourization [27] and inpainting [20], often involve reconstructing a corrupted image. Predictive tasks, on the other hand, consist of learning to predict or undo a transformation applied to an image. For example, the jigsaw pretext task is characterized by unscrambling randomly permuted rectangular patches of an image [19].

Several contemporary approaches adopt the joint-embedding architecture, in which representations of paired samples are compared. In a contrastive learning task, objectives are designed to minimize the distance between representations of paired examples (i.e., *positive pairs*) and maximize the distance between those of examples from different pairs (i.e., *negative pairs*) [7, 12]. Non-contrastive methods aim to minimize the difference between positive pairs, disregarding negative pairs [5, 8, 11, 26]. Recent methods have added regularizers to mitigate a degenerate solution in non-contrastive learning in which representations for all examples trend toward zero vectors [5, 26].

In the context of joint-embedding methods, the *pairwise relationship* enforces constraints between examples that qualify them as a positive pair. Typically, a positive pair consists of two perturbations of a single example, where each perturbation is sampled from a distribution over image transformations. In their paper demonstrating how SimCLR improves performance and label efficiency in chest X-ray and dermatological image classification, Azizi *et al*. [3] suggested an alternative, situation-specific pairwise relationship that reflects a stronger inductive bias in the pretrained network – they considered any two distinct images of the same pathology to be a positive pair. For example, posteroanterior and lateral chest X-rays from the same patient encounter are a positive pair. A recent theoretical analysis by Balestriero and LeCun lends credence to the idea of context-specific positive pairs, providing justification that pretraining using SimCLR [7], Barlow Twins [26], or VICReg [5] will improve performance on a downstream supervised learning task, as long as the pairwise relationship aligns with the labels for that task [4]. The authors' results underline the importance of medical knowledge in designing pretext tasks, motivating the definition of the M-mode pairwise relationship presented in this study.

2.3. SSL for Medical Ultrasound

SSL for medical ultrasound is underexplored compared to other medical image modalities, but some studies have investigated its utility for brightness mode (B-mode) ultrasound images and videos. Jiao *et al*. [14] observed an improvement in performance on the downstream task of fetal plane detection after pretraining a CNN to both reorder the images of a shuffled fetal ultrasound video and predict a transformation that was applied to it. Basu *et al*. [6] explored the benefit of defining negative pairs within the same ultrasound video in addition to those across videos, constructing a contrastive learning procedure in which intra-video pairs are introduced to the model after inter-video pairs. Self-supervised pretraining has also been effective for multiple echocardiography tasks, including atrial fibrillation detection [10], left ventricle segmentation [21], and view identification [2].

3. Methods

3.1. Dataset

The datasets used for all training experiments in this study originated from a large, private B-mode LUS database collected from two hospitals within an academic healthcare institution, the use of which is permitted by ethics approval granted by Western University (REB 116838) A portion of this database was previously labelled for the presence or absence of lung sliding by a critical care physician possessing expertise in LUS (hereafter referred to as the *LUS expert*). All LUS videos were divided into 3 s segments, with excess frames discarded. The resulting dataset, hereafter referred to as $\mathcal{D}_{\text{lab}}$, contained 4793 videos. $\mathcal{D}_{\text{lab}}$ was then randomly split by patient into three partitions: a training set of 3254

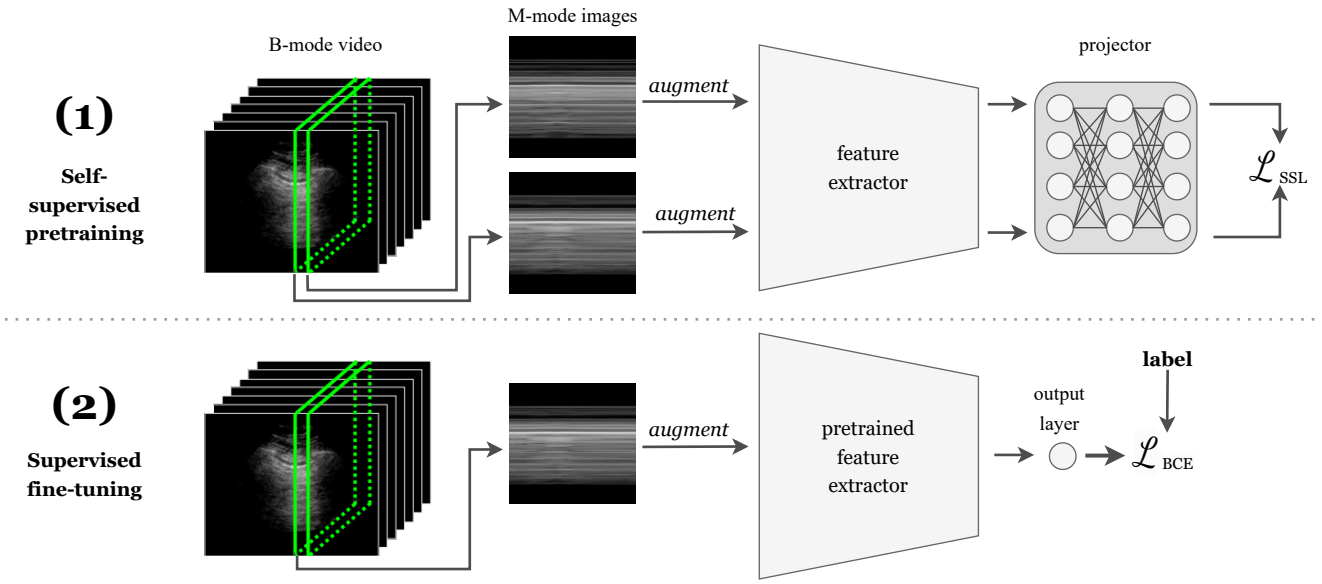


Figure 1. An overview of the methods in this paper. (1) Self-supervised pretraining using pairs of M-mode images extracted from the same B-mode LUS video. Both M-mode images are passed through a joint-embedding SSL architecture, consisting of a CNN feature extractor and multilayer perceptron projector in series. The objective, $\mathcal{L}_{SSL}$, is computed using the embeddings outputted by the projector. The pretrained CNN feature extractor is retained and the multilayer perceptron projector is discarded. (2) A single-node output layer is appended to the pre-trained feature extractor. The resulting model is fine-tuned to solve the downstream task of absent lung sliding detection by minimizing the binary cross entropy loss with respect to the labelled training set.

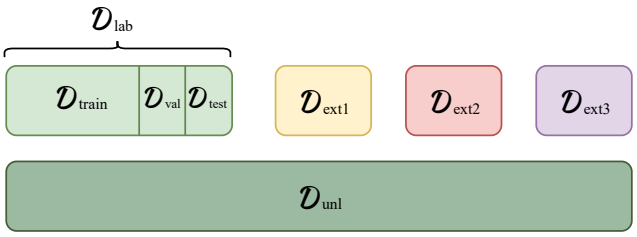


Figure 2. The datasets used in this study. $\mathcal{D}_{\text{lab}}$ is labelled for the downstream lung sliding classification task. It is split randomly by $70\%/15\%/15\%$ into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$, such that patients do not appear in multiple partitions.

videos ($\mathcal{D}_{\text{train}}$), validation set of 743 videos ($\mathcal{D}_{\text{val}}$), and test set of 796 videos ($\mathcal{D}_{\text{test}}$). The database was queried for additional videos containing the A-line artefact but that were missing labels for lung sliding. This additional tranche of 14249 unlabelled videos is referred to as $\mathcal{D}_{\text{unl}}$. Fig. 2 illustrates the dataset split.

To investigate model generalizability, we evaluate on three additional datasets from external healthcare institutions labelled for lung sliding: $\mathcal{D}_{\text{ext1}}$, $\mathcal{D}_{\text{ext2}}$, and $\mathcal{D}_{\text{ext3}}$. Tab. 1 provides details on the label and patient decomposition of all labelled datasets.

3.1.1 M-modes from B-modes

We follow a similar method outlined in [25] to extract M-mode images from greyscale B-mode videos. For the binary classification task, the vertical slice of the B-mode video with the maximum total pixel intensity is selected from all possible vertical slices within the horizontal bounds of the pleural line. This heuristic is consistent with the clinical notion that, in the vast majority of cases, the pleural line is the brightest artefact in a LUS image of the upper and middle lobes of the lung. All M-mode images were resized to 128×128 pixels prior to pretraining. When pretraining, we retain the top 50% of each video's M-modes, ordered by total pixel intensity. As will be discussed in Sec. 3.2.2, the pairwise relationship for M-modes requires distinct M-mode images from the same video.

3.2. Self-Supervised Learning

3.2.1 Pretext Tasks

We investigate three commonly employed joint-embedding self-supervised pretraining techniques from the literature. SimCLR [7] is a contrastive learning method that employs a temperature-scaled cross entropy objective. We set the temperature to $\tau = 0.1$. Barlow Twins [26] is a non-contrastive learning method that minimizes distance between embeddings of pairs and includes an embedding decorrelation reg-

		$\mathcal{D}_{\text{train}}$	$\mathcal{D}_{\text{val}}$	$\mathcal{D}_{\text{test}}$	$\mathcal{D}_{\text{unl}}$	$\mathcal{D}_{\text{ext1}}$	$\mathcal{D}_{\text{ext2}}$	$\mathcal{D}_{\text{ext3}}$
Lung sliding Present	Videos	2509	564	607	-	84	121	424
	Patients	366	76	72	-	25	55	155
Lung sliding Absent	Videos	745	179	189	-	26	30	77
	Patients	199	49	48	-	4	11	40
Total	Videos	3254	743	796	14 249	88	151	501
	Patients	565	125	120	3113	29	66	195

Table 1. Class decomposition

ularizer. We employ Barlow Twins with $\lambda = 0.005$ for the weight of the decorrelation regularizer. Inspired in part by Barlow Twins, VICReg [5] is a non-contrastive method that includes a regularizer that minimizes variance across the embedding dimension. We conduct pretraining trials with each of these methods. We set the weights of VICReg's three objective components to $\lambda = 25$, $\mu = 25$, and $\nu = 1$.

We pretrain for 100 epochs and use a batch size of 128 for all experiments. As in [3], we also investigate the effect of initializing the feature extractors with ImageNet-pretrained weights prior to self-supervised pretraining. After all pretraining experiments, weights of the projector are discarded and the feature extractor are preserved for initialization in the downstream task. We adopt the EfficientNetB0 [23] architecture as the feature extractor for all experiments, discarding the final block to reduce model capacity. In each experiment, the projector is a multilayer perceptron with 3 layers of 128 nodes, with the rectified linear unit activation applied to each hidden layer.

3.2.2 Pairwise Relationship

As outlined in Sec. 3.1.1, the brightest 50% of the M-mode images within the bounds of the pleural line are utilized for pretraining. We consider any such M-mode images from the same video to be a positive pair. Qualitatively, different M-mode images produced from the same B-mode video appear very similar. Crucially, they would have the same lung sliding label, fulfilling the alignment condition outlined in [4]. Fig. 3 displays examples of positive pairs of M-modes. We fix the pairwise relationship to focus on evaluating data augmentation transformations and the effects of pretraining under different settings, relegating an ablation study for the M-mode pairwise relationship to future work.

3.2.3 Data Augmentation Strategy

A pretraining data augmentation strategy was devised to improve the invariance against inconsequential transformations and noise inherent to M-mode images. Transformations were identified that simulate natural variations present in M-mode LUS but that do not impact the patterns exhibited by absent or present lung sliding, such as speckle noise and variable depth, gain, and contrast. The following series of stochastic transformations is applied to each M-mode image prior to subjecting it to the forward pass in pretraining:

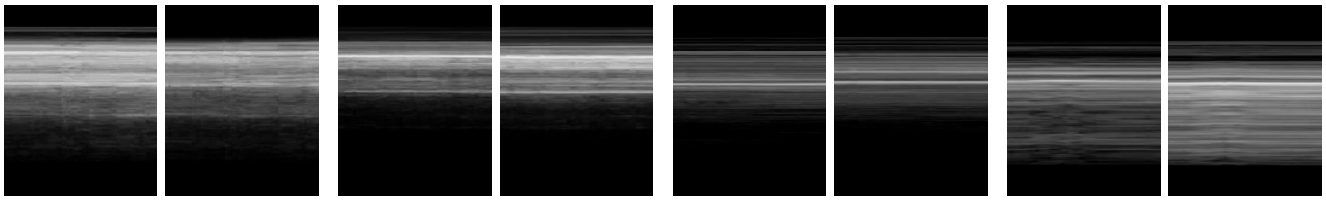
1. With probability 0.8, we random crop of $c \sim \mathcal{U}(0.08, 1.0)$ of the image's area, which is resized to its original dimensions. To ensure the pleural line was retained, the top of the crop was always within the upper half of the image.
2. With probability 0.5, horizontal flip
3. With probability 0.5, Gaussian blur with a horizontal kernel of 10 pixels and $\sigma \sim \mathcal{U}(0.1, 2.0)$
4. With probability 0.5, random additive Gaussian noise is added to each pixel, sampled from $\mathcal{N}(\mu, \sigma^2)$, where $\mu \sim \mathcal{U}(-10, 10)$ and $\sigma \sim \mathcal{U}(0, 25)$
5. With probability 0.5, application of speckle noise, simulated using multiplicative Gaussian noise, sampled from $\mathcal{N}(1, \sigma^2)$, with $\sigma \sim \mathcal{U}(0, 0.1)$
6. With probability 0.8, brightness adjustment by $c \sim \mathcal{U}(-0.4, 0.4)$
7. With probability 0.8, contrast adjustment by $c \sim \mathcal{U}(-0.4, 0.4)$. With probability 0.5, contrast adjustment occurs before brightness adjustment.

In summary, a positive pair consists of two M-mode images taken from a 3 s segment of the same original B-mode LUS video that are then transformed via data augmentation. See Fig. 4 for some examples of positive pairs.

3.3. Evaluation Protocol

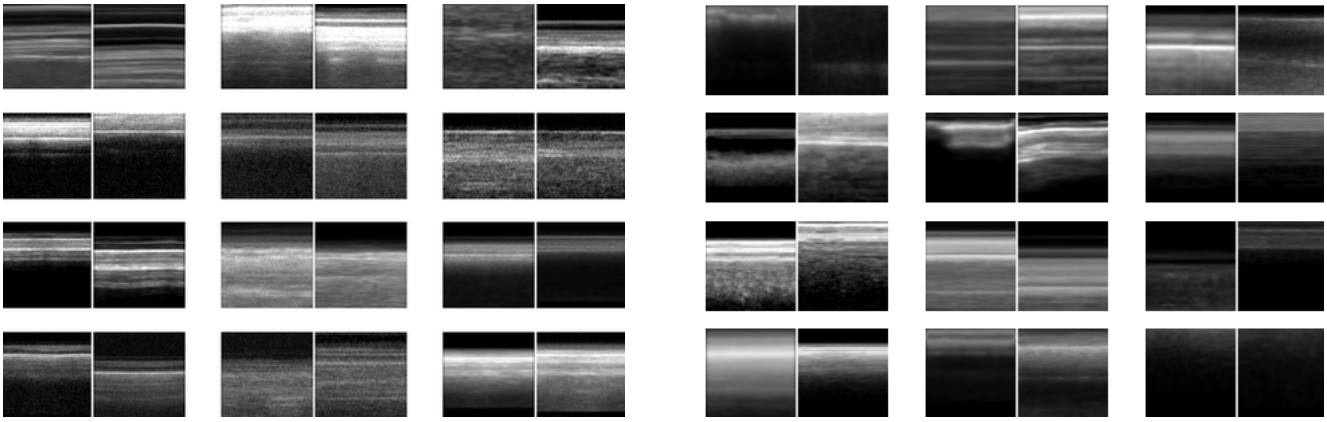
3.3.1 Lung Sliding Classification

The downstream supervised learning task is the identification of absent lung sliding in M-mode images, which is a binary classification task. M-mode images of the upper and middle lobes of the lung are well suited for this problem, as there exist established visual patterns employed by clinicians to distinsuish between present and absent lung sliding. The *seashore sign* is indicative of lung sliding [15], whereas the *barcode sign* signals absent lung sliding [17]. We consider absent lung sliding to be the positive class.



(a) Lung sliding, seashore sign. (b) Lung sliding, seashore sign. (c) Absent lung sliding, barcode sign. (d) Absent lung sliding, barcode sign

Figure 3. Examples of M-mode images satisfying the pairwise relationship of belonging to the same original B-mode video and intersecting the pleural line.



(a) Custom M-mode augmentation pipeline (b) BYOL augmentation pipeline [11]

Figure 4. A selection of M-mode positive pairs that have been subjected to the M-mode-specific data augmentation pipeline.

3.3.2 Data Augmentation Pipeline Evaluation

We first ascertain the merit of the custom M-mode LUS data augmentation pipeline for the pretext task (described in Sec. 3.2.3), as opposed to the ubiquitously cited augmentation pipeline for the "Bring Your Own Latent" (BYOL) SSL pretraining method proposed by Grill *et al*. [11], which we accordingly refer to as the *BYOL augmentation pipeline*. Unlike the original BYOL study, we do not apply hue changes because LUS M-mode images are greyscale. For each of the three studied pretraining methods, two CNN feature extractors are trained using $\mathcal{D}_{\text{train}}$ on the pretext task using the custom and BYOL augmentation pipelines respectively, producing a total of six feature extractors. A linear classifier is then appended and trained on the downstream task for each of the feature extractors. The best-performing augmentation pipeline across such experiments is then used for all subsequently performed pretraining trials.

3.3.3 Linear and Fine-tuning Evaluation

To investigate the effect of self-supervised pretraining on the performance in the downstream task, we adopt an evaluation protocol similar to those conducted by contemporary SSL publications that evaluate on natural images. We append a single-node fully connected layer to the feature extractor with sigmoid activation.

We consider two methods to evaluate the utility of weights pretrained with SSL – linear modelling and fine-tuning. In the linear evaluation, the weights of the pretrained model are fixed and a linear classifier is trained using the feature representations from the extractor. In the case of fine-tuning, all model weights past the first block of the network are subject to updates. In both cases, we train for 40 epochs with a learning rate of 1×10^{-4}, which is decayed by 0.03 every epoch after the 15^{th} epoch. There is a heavy class imbalance in $\mathcal{D}_{\text{lab}}$ that favours present lung sliding. As a result, we oversample the minority class, taking the second, third, and fourth brightest M-modes from each absent lung sliding video. Models are trained using the binary cross entropy loss function. Data augmentation is applied to training images, with random contrast reduction by $c \sim \mathcal{U}(0, 0.3)$, random brightness adjustment by $b \sim \mathcal{U}(-0.1, 0.1)$, additive Gaussian noise sampled from $\mathcal{N}(0, 5)$, and random horizontal flip. The model weights resulting in the lowest loss on $\mathcal{D}_{\text{val}}$ were saved for evaluation.

The performance of self-supervised pretrained models are compared against two fully supervised baselines, where weights are initialized with either ImageNet-pretrained weights or random weights.

	Random init.		ImageNet init.	
	BYOL	M-mode	BYOL	M-mode
SimCLR	0.585	0.658	0.864	0.827
Barlow Twins	0.554	0.578	0.826	0.818
VICReg	0.6208	0.6377	0.826	0.798

Table 2. AUC on $\mathcal{D}_{\text{val}}$ of linear classifier trained using feature representations from various self-supervised networks.

3.3.4 Label Efficiency

To assess the effects of SSL pretraining in its entirety, it is essential to consider its potency with respect to both performance differences and the ability to leverage otherwise unusable unlabelled datasets. Accordingly, we carry out a series of experiments that compare downstream performance under different levels of labelled data availability. We pretrain on $\mathcal{D}_{\text{train}}$ and fine-tune using progressively larger subsets of $\mathcal{D}_{\text{train}}$. Fine-tuning is repeated using ImageNet-pretrained weights and randomly initialized weights, facilitating comparisons in low-label scenarios. Lastly, we pretrain on a large dataset consisting of $\mathcal{D}_{\text{lab}}$ and $\mathcal{D}_{\text{unl}}$.

3.3.5 Explainability

Saliency maps (also referred to as "heatmaps") are a type of explanation for CNN predictions that consist of the orginal image superimposed onto a colour map indicating the regions that were most contributory to the prediction. We generate saliency maps using Grad-CAM [22] for selected predictions using models pretrained with SSL, initialized with ImageNet weights, and randomly initialized. The LUS expert compares the appropriateness of the saliency maps produced for pretrained and fully supervised models.

4. Results

4.1. Data Augmentation

We investigate the effect of applying our custom M-mode data augmentation pipeline in the pretraining phase. Three models were pretrained using the examined SSL methods. We execute linear evaluation trials and compare the AUC on $\mathcal{D}_{\text{val}}$. As demonstrated in Tab. 2, the custom M-mode augmentation pipeline results in higher performance in linear evaluation when the pretrained models are initialized randomly. Interestingly, the BYOL augmentation pipeline exhibits a marked improvement compared to the M-mode-specific augmentations in the case of ImageNet initialization. As a result, all further experiments pretrained from scratch and from ImageNet-pretrained weights use the M-mode augmentations and BYOL augmentations respectively.

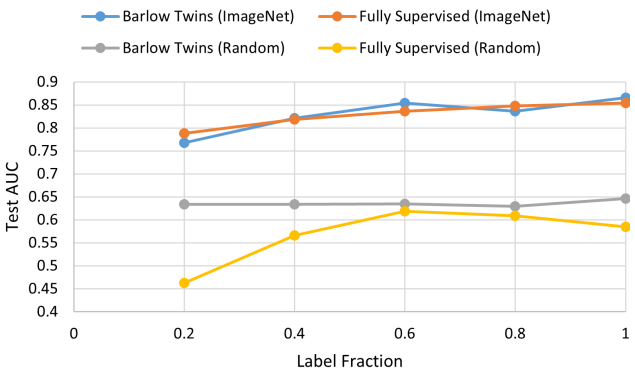


Figure 5. AUC on $\mathcal{D}_{\text{test}}$ for fine-tuned models initialized from weights pretrained using Barlow Twins and fully supervised models initialized randomly or with ImageNet-pretrained weights.

4.2. Comparison with Supervised Baselines

To assess the quality of the pretrained representations, we conduct linear evaluation and fine-tuning trials using pretrained feature extractors. The results on $\mathcal{D}_{\text{test}}$ are compared with a linear classifier trained atop feature extractors initialized randomly and with ImageNet-pretrained weights. Tab. 3 summarizes the performance on $\mathcal{D}_{\text{test}}$. Immediately apparent is the utility of initializing all feature extractors with ImageNet-pretrained weights, including those pretrained with SSL. Among the fine-tuning trials initialized with ImageNet, self-supervised pretrained models exhibit greater test performance. The results are less clear for linear evaluation using ImageNet weight initialization, as linear models using a frozen ImageNet-pretrained feature extractor achieve the greatest test AUC and sensitivity. We additionally find that in all cases where weights are initialized randomly, self-supervised pretrained models outperform the fully supervised baselines.

4.3. Label Efficiency

One of the major benefits of SSL is its ability to leverage unlabelled examples. To measure the effect of varying proportions of labelled data, we drop fractions of the labels in $\mathcal{D}_{\text{train}}$ and conduct fine-tuning using networks pretrained using Barlow Twins on all examples in $\mathcal{D}_{\text{train}}$. Fig. 5 details the results. The feature extractor trained from scratch benefitted from self-supervised pretraining in the low-label setting. However, it appears that this benefit is greatly diminished when the pretrained feature extractor and fully supervised feature extractor are both initialized with weights pretrained on ImageNet.

To further elucidate the benefit of pretraining with greater volumes of unlabelled data, we pretrain on both $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{unl}}$ using Barlow Twins and compare the performance to pretraining with $\mathcal{D}_{\text{train}}$ alone. As shown in Tab. 4, more unlabelled data greatly improves the test performance

Pretraining Method	Initialization	$\mathcal{D}_{\text{test}}$ AUC		$\mathcal{D}_{\text{test}}$ Specificity		$\mathcal{D}_{\text{test}}$ Sensitivity		$\mathcal{D}_{\text{test}}$ Accuracy	
		Linear	Fine-tune	Linear	Fine-tune	Linear	Fine-tune	Linear	Fine-tune
SimCLR	ImageNet	0.742	0.861	**0.764**	**0.909**	0.497	0.545	**0.701**	**0.823**
	Random	0.645	*0.662*	0.582	0.755	0.619	0.476	0.590	0.688
Barlow Twins	ImageNet	0.707	**0.866**	0.705	0.845	0.598	**0.741**	0.680	0.820
	Random	*0.646*	0.634	0.644	0.718	0.534	0.460	0.618	0.657
VICReg	ImageNet	0.738	0.834	0.661	0.84	0.703	0.656	0.671	0.797
	Random	0.609	0.619	*0.926*	*0.860*	0.286	0.318	*0.774*	*0.731*
None	ImageNet	**0.768**	0.854	0.638	0.834	**0.751**	0.714	0.665	0.805
	Random	0.500	0.585	0.000	0.563	*1.000*	*0.540*	0.237	0.558

Table 3. Downstream performance on $\mathcal{D}_{\text{test}}$, trained $\mathcal{D}_{\text{train}}$. *Typescript* entries correspond to the best performance when using randomly initialized weights and **boldface** entries identify the best performance for trials initialized with ImageNet-pretrained weights.

Initialization	Pretraining data	$\mathcal{D}_{\text{test}}$ AUC
Random	$\mathcal{D}_{\text{train}}$	0.634
	$\mathcal{D}_{\text{train}} + \mathcal{D}_{\text{unl}}$	0.799
ImageNet	$\mathcal{D}_{\text{train}}$	0.866
	$\mathcal{D}_{\text{train}} + \mathcal{D}_{\text{unl}}$	0.838

Table 4. Downstream classification performance of models pretrained using Barlow Net using labelled and unlabelled datasets.

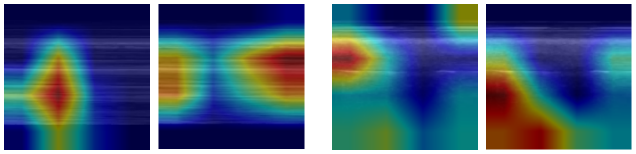
of the feature extractors pretrained using random initialization, while performance drops for those initialized with ImageNet weights.

4.4. Evaluation on External Datasets

We evaluate all pretrained and fully supervised fine-tuned models initialized with ImageNet-pretrained weights on $\mathcal{D}_{\text{ext1}}$, $\mathcal{D}_{\text{ext2}}$, and $\mathcal{D}_{\text{ext3}}$. The results, presented in Tab. 5, do not indicate that the self-supervised models pretrained on $\mathcal{D}_{\text{train}}$ outperform the fully supervised baseline when evaluated on datasets originating from other centres. However, with the exception of sensitivity, we find that performance on external datasets distinctly increases when models are pretrained on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{unl}}$ combined, highlighting a potential benefit of leveraging unlabelled data when labels are partially available.

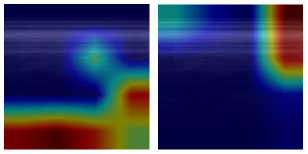
4.5. Explainability

To explore the patterns learned by the fine-tuned models, we select the pretrained network with the greatest $\mathcal{D}_{\text{test}}$ AUC (Barlow Twins) and produce saliency maps for 4 M-mode images in $\mathcal{D}_{\text{test}}$, using the pretrained network and a fully supervised network initialized with ImageNet-pretrained weights (see Fig. 6). The LUS expert reviewed the saliency maps without knowing which method was used to produce them and rated the appropriateness of the heatmaps using a 4-point Likert scale from 0 to 3, based on whether the highlighted regions correspond to the areas of clinical interest. For instance, despite the false positive prediction by the self-supervised model shown in Fig. 6a, the far field (bottom of the image) is dark, and the saliency map indicates activation by the poignant straight lines in the near field, producing a prediction of absent lung sliding. The saliency maps generated for the pretrained model and fully supervised model scored averages of 2.5 and 1.5 respectively.

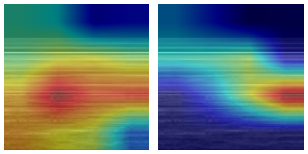


(a) False positive for both models.

(b) True negative for both models.



(c) False negative for the fully supervised model; true positive for the pretrained model.



(d) True positive for both models.

Figure 6. A selection of saliency maps for predictions produced by both a fully supervised model (left in each subfigure) and a model pretrained with Barlow Twins.

5. Discussion

The results in this study characterize the utility of self-supervised pretraining for M-mode LUS examinations. First, it is clear that fine-tuning a self-supervised feature extractor provides improved initial feature representations compared to randomly initialized weights. The gap in performance on unseen examples strikingly narrows when pretrained models and their fully supervised counterparts are initialized using the omnipresent ImageNet-pretrained weights. These results indicate that, while there may be performance gains when fine-tuning SSL-pretrained

	Pretraining dataset(s)	AUC	Specificity	Sensitivity	Accuracy
SimCLR	$\mathcal{D}_{\text{train}}$	0.765 [0.063]	0.765 [0.043]	0.532 [0.042]	0.729 [0.037]
	$\mathcal{D}_{\text{train}} + \mathcal{D}_{\text{unl}}$	0.778 [0.053]	**0.862** [**0.046**]	0.429 [0.103]	**0.834** [**0.056**]
Barlow Twins	$\mathcal{D}_{\text{train}}$	0.802 [0.071]	0.728 [0.043]	0.727 [0.138]	0.731 [0.048]
	$\mathcal{D}_{\text{train}} + \mathcal{D}_{\text{unl}}$	**0.833** [**0.059**]	0.761 [0.042]	0.723 [0.076]	0.756 [0.036]
VICReg	$\mathcal{D}_{\text{train}}$	0.807 [0.048]	0.727 [0.073]	**0.785** [**0.021**]	0.740 [0.059]
	$\mathcal{D}_{\text{train}} + \mathcal{D}_{\text{unl}}$	0.817 [0.053]	0.705 [0.076]	0.760 [0.114]	0.720 [0.054]
Fully Supervised	-	0.804 [0.076]	0.682 [0.097]	0.761 [0.075]	0.699 [0.071]

Table 5. Mean [std] performance across the three external datasets for pretrained and fully supervised models. Each was initialized with ImageNet weights and (pre)trained using $\mathcal{D}_{\text{train}}$. Models pretrained using $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{unl}}$ outperformed those pretrained using $\mathcal{D}_{\text{train}}$ alone.

models, the state-of-the-art constrastive and nonconstrative methods alone are not sufficient to reap substantial gains when trained on the labelled dataset and initialized with ImageNet-pretrained weights. However, when training from scratch, practitioners may benefit from pretraining.

A major insight derived from the results is that SSL-pretrained models improve performance on datasets from external centres for the task of lung sliding classification when using large volumes of unlabelled data that do not appear in the training set. Generalizability is a paramount concern for any organization deploying machine learning systems in healthcare settings. This finding is consistent with the results of studies showing that SSL pretraining improves performance on external data in other medical imaging domains, such as chest X-ray classification [3, 24], dermatologic image classification [3], and pathology slide classification [28]. Practitioners seeking to promote generalizability are therefore encouraged to employ self-supervised pretraining using any available unlabelled data.

Another noteworthy finding is that ImageNet-initialized feature extractors pretrained with the BYOL augmentation pipeline yield better performance for the downstream classification task. Again, the duality between ImageNet-pretrained and randomly initialized feature extractors manifests itself, as the M-mode augmentation pipeline produces better feature representations when pretrained models were initialized randomly. We conjecture that the BYOL augmentation pipeline's efficacy is related to the fact that the pretrained models were initialized with ImageNet-pretrained weights, since the original BYOL paper (and subsequent SSL publications, such as [5, 26]) focused on improving downstream tasks with ImageNet [11].

The present work has notable limitations. The training set consisted of LUS examinations collected within a single healthcare institution, thereby limiting heterogeneity with respect to sources of variance such as device manufacturer, practitioner skill sets, and patient populations. Secondly, SimCLR performs best when larger batch sizes are employed during pretraining [7]; however, due to material limitations, we utilize a considerably small batch size. In keeping with the authors' findings, we train for a large number of epochs to mitigate the impact of a small batch size. Lastly, this study did not meet or exceed the performance metrics reported in the most recent publication regarding automatic lung sliding classification [25]. However, unlike [25], we employ different datasets, we use standard binary cross entropy loss, we do not apply any techniques to mitigate overfitting, and we do not add fully connected layers between the feature extractor and the output layer. Rather than aiming to maximize the performance of the classifier, our focus is to study the effect of different SSL strategies and data augmentation distributions on the quality of representations.

There are multiple avenues for future work. First, a subsequent investigation could undertake a comprehensive inquiry into the data augmentation transformations for joint-embedding SSL methods applied to M-mode LUS. secondly, the lack of consensus regarding the augmentation pipeline motivates future work concerning the underlying reasons and possible discovery of novel M-mode ultrasound data transformations. Lastly, further research could explore alternative pretext tasks. In this study, we propose a pairwise relationship for M-mode images consisting of images that were taken from the same B-mode video, adopting common contrastive and non-contrastive learning pretext tasks; however, a novel pretext task could be formulated to better suit the M-mode ultrasound domain.

6. Conclusion

A selection of contemporary contrastive and non-contrastive SSL pretraining methods were investigated for LUS M-mode data, using a pairwise relationship specific to M-mode LUS. When evaluated on the downstream binary classification task of absent lung sliding detection, fine-tuned feature extractors initialized with self-supervised pretrained weights generally exhibited greater performance than fully supervised counterparts. Pretraining with larger unlabelled datasets resulted in improved metrics on evaluation datasets from external institutions. The results spur multiple directions for future work, such as the refinement of M-mode ultrasound data augmentation pipelines for SSL and the evaluation of alternative or novel pretext tasks.

References

[1] Khaled Alrajhi, Michael Y Woo, and Christian Vaillancourt. Test characteristics of ultrasonography for the detection of pneumothorax: a systematic review and meta-analysis. *Chest*, 141(3):703–708, 2012. 1

[2] Deepa Anand, Pavan Annangi, and Prasad Sudhakar. Benchmarking self-supervised representation learning from a million cardiac ultrasound images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 529–532. IEEE, 2022. 2

[3] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi. Big Self-Supervised Models Advance Medical Image Classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3458–3468. Institute of Electrical and Electronics Engineers Inc., 2021. ISSN: 15505499. 2, 4, 8

[4] Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *arXiv preprint arXiv:2205.11508*, 2022. 2, 4

[5] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. 1, 2, 4, 8

[6] Soumen Basu, Somanshu Singla, Mayank Gupta, Pratyaksha Rana, Pankaj Gupta, and Chetan Arora. Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*, pages 423–433. Springer, 2022. 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2, 3, 8

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[10] Fatemeh Taheri Dezaki, Tom Ginsberg, Christina Luong, Hooman Vaseli, Robert Rohling, Ken Gin, Purang Abolmaesumi, and Teresa Tsang. Echo-rhythm net: Semi-supervised learning for automatic detection of atrial fibrillation in echocardiography. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 110–113. IEEE, 2021. 2

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1, 2, 5, 8

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[13] Miroslav Jaščur, Marek Bundzel, Marek Malík, Anton Dzian, Norbert Ferenčík, and František Babič. Detecting the absence of lung sliding in lung ultrasounds using deep learning. *Applied Sciences*, 11(15):6976, 2021. 1, 2

[14] J. Jiao, R. Droste, L. Drukker, A.T. Papageorghiou, and J.A. Noble. Self-Supervised Representation Learning for Ultrasound Video. In *Proceedings - International Symposium on Biomedical Imaging*, volume 2020-April, pages 1847–1850. IEEE Computer Society, 2020. ISSN: 19457928. 2

[15] Daniel A Lichtenstein. *Whole body ultrasonography in the critically ill*. Springer Science & Business Media, 2010. 4

[16] Daniel A Lichtenstein and Yves Menu. A bedside ultrasound sign ruling out pneumothorax in the critically iii: lung sliding. *Chest*, 108(5):1345–1348, 1995. 1

[17] Daniel A Lichtenstein, Gilbert Mezière, Nathalie Lascols, Philippe Biderman, Jean-Paul Courret, Agnès Gepner, Ivan Goldstein, and Marc Tenoudji-Cohen. Ultrasound diagnosis of occult pneumothorax. *Critical care medicine*, 33(6):1231–1238, 2005. 4

[18] Khanjan Nagarsheth and Stanley Kurek. Ultrasound detection of pneumothorax compared with chest x-ray and computed tomography scan. *The American surgeon*, 77(4):480–483, 2011. 1

[19] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 69–84. Springer, 2016. 2

[20] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[21] Mohamed Saeed, Rand Muhtaseb, and Mohammad Yaqub. Contrastive pretraining for echocardiography segmentation with limited data. In *Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings*, pages 680–691. Springer, 2022. 2

[22] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6

[23] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 4

[24] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection

of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, pages 1–8, 2022. 8

[25] Blake VanBerlo, Derek Wu, Brian Li, Marwan A Rahman, Gregory Hogg, Bennett VanBerlo, Jared Tschirhart, Alex Ford, Jordan Ho, Joseph McCauley, et al. Accurate assessment of the lung sliding artefact on lung ultrasonography using a deep learning approach. *Computers in Biology and Medicine*, 148:105953, 2022. 1, 2, 3, 8

[26] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 1, 2, 3, 8

[27] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 2

[28] Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022. 8