IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 November 2023, accepted 25 November 2023, date of publication 28 November 2023,
date of current version 6 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3337398

RESEARCH ARTICLE

Self-Supervised Pretraining Improves Performance and Inference Efficiency in Multiple Lung Ultrasound Interpretation Tasks

BLAKE VANBERLO[1,2], (Graduate Student Member, IEEE), BRIAN LI[2,3], JESSE HOEY[1], (Member, IEEE), AND ALEXANDER WONG[3], (Senior Member, IEEE)

[1]Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[2]Deep Breathe, London, ON N6A 3S9, Canada
[3]Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada

Corresponding author: Blake Vanberlo (bvanberl@uwaterloo.ca)

The work of Blake Vanberlo was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through Vanier Scholar under Grant FRN 186945.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT In this study, we investigated whether self-supervised pretraining could produce a neural network feature extractor applicable to multiple classification tasks in B-mode lung ultrasound analysis. When fine-tuning on three lung ultrasound tasks, pretrained models resulted in an improvement of the average across-task area under the receiver operating characteristic curve (AUC) by 0.032 and 0.061 on local and external test sets respectively. Compact nonlinear classifiers trained on features outputted by a single pretrained model did not improve performance across all tasks; however, they reduced inference time by 49% compared to the serial execution of separate fine-tuned models. When training using 1% of the available labels, pretrained models consistently outperformed fully supervised models, with a maximum observed test AUC increase of 0.396 for the task of view classification. Overall, the results indicate that self-supervised pretraining is a useful strategy for producing initial weights for lung ultrasound classifiers.

INDEX TERMS Multi-task, self-supervised learning, ultrasound.

I. INTRODUCTION

Lung ultrasound (LUS) is a point-of-care ultrasound (POCUS) examination that is performed in acute care settings to rapidly narrow down differential diagnoses for patients in respiratory distress. In addition to its enhanced portability, safety, and affordability, LUS has exhibited diagnostic accuracy for a variety of respiratory conditions that is comparable to traditional modalities, such as chest radiography [1], [2], [3], [4] and computed tomography [2], [5], [6], [7]. Despite mounting evidence for its efficacy, there are barriers to the widespread adoption of POCUS, including reduced availability of training and lack of access to devices [8], [9], [10]. Multiple studies have proposed machine learning solutions for routine tasks in LUS interpretation, citing automation as a means to improve access to LUS [11], [12], [13]. A major barrier to the development of machine learning models for tasks in POCUS is the lack of access to curated, labeled datasets [14]. In addition to the sparsity of LUS expertise, the expense of soliciting experts to manually label retrospectively acquired ultrasound videos is prohibitive. As a result, there is remarkable value in discovering techniques that can reduce the amount of labeling required for retrospectively collected datasets.

Recent years have witnessed a surge of interest in self-supervised learning (SSL) as a strategy for representation learning in computer vision. Hailed as a means to

The associate editor coordinating the review of this manuscript and approving it for publication was Aysegul Ucar.

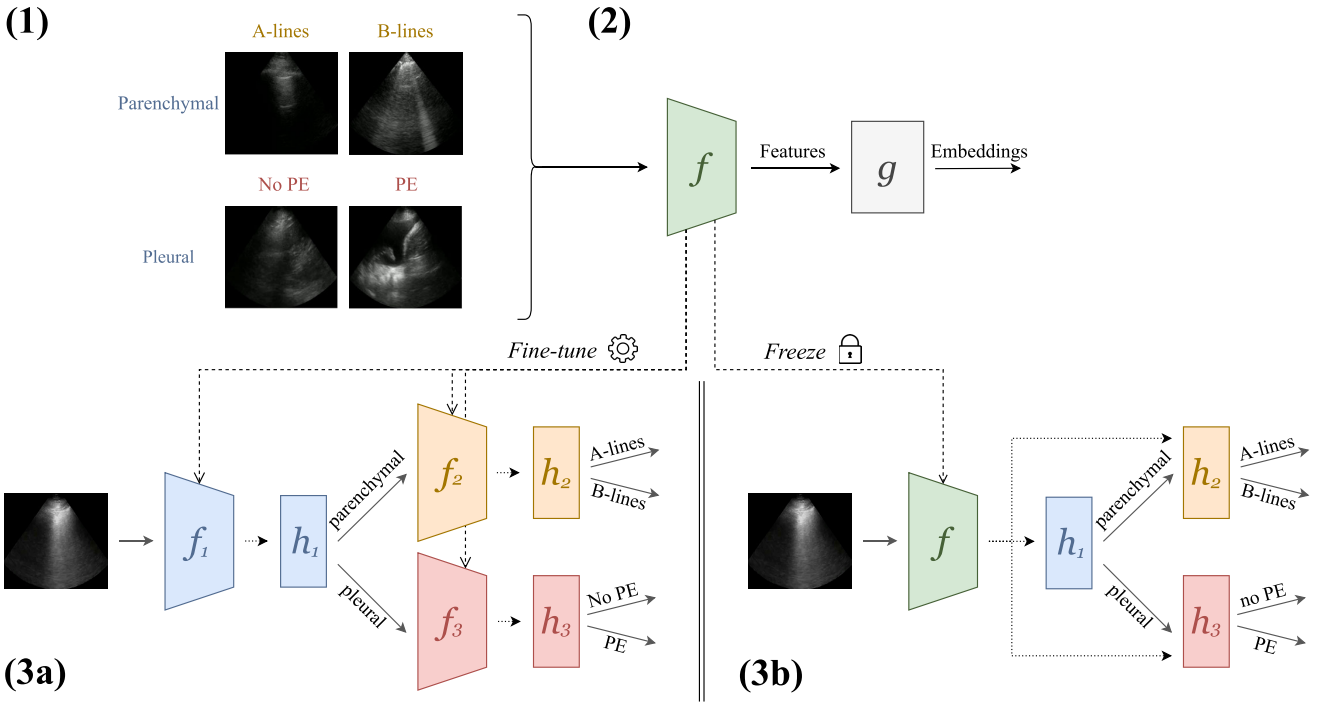


FIGURE 1. An overview of the methods described in this work. (1) Three tasks were identified for lung ultrasound (LUS) image classification: parenchymal versus pleural views, A-lines versus B-lines (applicable to parenchymal views), and pleural effusion (PE) versus no pleural effusion (applicable to pleural views). (2) A convolutional feature extractor f was pretrained to minimize a self-supervised objective, using unlabeled and labeled LUS images as input and trainable projector g. (3a) Task-specific models were defined by appending linear classifier or multilayer perceptron h_i to copies of pretrained f. The models were trained end-to-end for each task using labeled data. (3b) An alternative framework in which f's weights were not fine-tuned. Instead, task-specific models h_i were trained that each received f's feature representations as input.

productively leverage unlabeled data when labels are scarce, self-supervised pretraining produces a feature extractor that may be used to initialize the weights of a model in a supervised learning setting. It has been shown to improve performance on several supervised learning tasks in multiple domains of medical imaging, such as radiography [15], [16], computed tomography [16], [17], magnetic resonance imaging [16], [17], ultrasound [18], [19], and dermatology [15]. Studies have indicated that models pretrained with SSL perform comparably to fully supervised models even when fine-tuned with significantly less labeled data [15], [17]. Given the widespread paucity and expense of labeled medical images, it is therefore unsurprising that SSL has risen as a reasonable strategy to leverage unlabeled data. The primary objective of this study was to determine if contemporary self-supervised learning methods are a viable solution for improving performance in LUS classification tasks, particularly when the number of labels in a retrospectively acquired LUS dataset is low. We considered multiple tasks to ensure that any observed benefits of SSL were not confined to one particular LUS task.

The secondary objective of this work was to explore whether SSL methods can produce a single set of feature representations that are useful across different LUS tasks. More specifically, we wished to determine if such representations could be useful for the hierarchical arrangement of LUS interpretation tasks. LUS interpretation involves the recognition of multiple artifacts that narrow differential diagnoses in emergency and critical care scenarios, hereafter referred to as *multi-task LUS interpretation*. As is common across other medical imaging modalities, LUS interpretation can be conducted in a hierarchical manner. Interpreters engage in the predictive process of a decision tree, beginning with the root node and traversing down a single path, guided by decisions at each node. Examples of hierarchical interpretation from other modalities include the distinction of malignant pulmonary nodules on CT [20] and the identification of lipomatous tumors on MRI [21]. Past work in machine learning-based hierarchical medical imaging classification has resorted to training entirely separate classifiers for each node in the tree [22]. Our study sought to determine if a single feature extractor could produce meaningful representations for multi-task LUS interpretation. We hypothesized that self-supervised pretraining is suited for the task of developing a feature extractor that is useful for multiple classification tasks. The weights of the feature extractor could be fine-tuned for individual subsequent tasks (Fig. 1, 3a). Alternatively, holding the weights constant would facilitate the addition of new tasks to the multi-task LUS interpretion system by training classifiers on top of the features (Fig. 1, 3b).

The core contributions of this work are thus as follows: (1) an investigation of the suitability of self-supervised

feature extractors for multi-task interpretation of B-mode LUS, and (2) a tree-based classification strategy in which the inputs to the root node are obtained from a feature extractor pretrained with SSL. Section II provides a focused background of the SSL methods investigated in this study, an overview of the evidence regarding the impact of SSL in improving automatic ultrasound interpretation, and a summary of prior approaches to multi-task medical image interpretation. Section III describes the LUS interpretation tasks, datasets, and SSL pretraining protocol employed in this study. Section IV provides an evaluation of the performance. It also gives runtime statistics for each task and compares the fine-tuning of end-to-end models for each task against the training of multilayer perceptrons (MLP) on features outputted by a single pretrained extractor. Lastly, conclusions and recommendations for practitioners are given in Section V.

II. BACKGROUND

A. JOINT EMBEDDING SELF-SUPERVISED LEARNING

Broadly, self-supervised learning (SSL) is a form of unsupervised representation learning that is employed to pretrain a feature extractor for transfer learning. The feature extractor is trained to solve a *pretext task*, which is a supervised learning problem using labels that are computed from unlabeled data. The weights of the pretrained feature extractor can then be used to initialize a new model trained to solve a supervised learning task using a labeled dataset. In the joint embedding framework of SSL, the pretext task is designed to reduce the differences between representations of semantically related images that satisfy a pairwise relationship. Semantically related *positive pairs* of images are customarily passed through the feature extractor, with the output being sent through a projection head (typically a MLP), producing embeddings. A common approach for defining positive pairs is to produce two distinct distortions of the same image, ensuring that the distorted views do not alter the semantic content of the image. Joint embedding methods aim to maximize the similarity of the embeddings of positive pairs, thereby encouraging the feature extractor to learn to produce feature representations that are invariant to meaningless transformations. Refer to Fig. 2 for a visual depiction of a prototypical joint embedding method.

There are two cardinal categories of joint embedding SSL methods: contrastive and non-constrastive. In contrastive learning tasks, the SSL objective is designed to minimize the difference between embeddings of pairs of images satisfying the pairwise relationship (i.e., *positive pairs*) and maximize the difference between pairs that do not satisfy the relationship (i.e., *negative pairs*). SimCLR is a popular contrastive learning method where positive pairs are produced by transforming each image twice, where the parameters of the transformations are sampled from a distribution [23]. The model is trained to optimize the InfoNCE objective [24] to attract positive pairs and repel negative pairs in the embedding space.

Non-contrastive learning emerged as a strategy to address shortcomings in contrastive learning, such as its reliance on large batch sizes. Non-contrastive methods do not require negative pairs and focuses solely on minimizing the distance between embeddings of positive pairs. However, non-contrastive methods are vulnerable to information collapse – a degenerate solution where embeddings are consistently predicted as null vectors. To combat this adverse scenario, objectives have been proposed that included weighted terms to promote embedding decorrelation [25], [26] and nonzero variance in a batch [26].

B. JOINT EMBEDDING METHODS IN B-MODE ULTRASOUND

Multiple studies have assessed the impact of joint embedding self-supervised pretraining on the performance of machine learning solutions in diagnostic B-mode US tasks, particularly when labels are scarce. For example, contrastive and non-contrastive methods have been applied to breast tumor classification and left ventricle segmentation with mixed results [18], [27], [28]. Contrastive pretraining has been observed to improve the performance of models trained to distinguish benign and malignant breast tumors [18]. However, a similar study on a separate dataset found that a non-contrastive method did not outperform fully supervised models initialized with ImageNet-pretrained weights [27]. Similarly, contrastive and non-contrastive learning were reported to respectively improve and degrade on left ventricle segmentation in echocardiograms [28]. Anand et al. [29] performed a comprehensive evaluation of several joint embedding methods for the task of echocardiogram view classification, observing consistent improvement over full supervision.

Focusing on LUS applications, Chen et al. [30] proposed a custom contrastive learning objective with interpolated intra-video positive pairs, outperforming both fully supervised and SimCLR-pretrained models on the public POCUS dataset [31]. Adopting a curriculum learning approach, Basu et al. [19] achieved even better performance on POCUS with their contrastive learning method that employed progressively harder intra-video positive pairs. Both studies evaluated their ultrasound-specific contrastive learning approaches on small public LUS datasets. The authors pretrained feature extractors on a 22-video public LUS dataset acquired with devices manufactured by Butterfly Network. They fine-tuned their models for the task of COVID-19 pneumonia classification using the public POCUS [31] dataset. The current study adds to the previous literature exploring SSL for LUS tasks by (1) investigating the utility of the same pretrained feature extractor for multiple LUS tasks and (2) including experimentation with non-contrastive methods.

C. MULTI-TASK MEDICAL IMAGE INTERPRETATION

Several studies have addressed multi-task learning for multi-task medical imaging interpretation. For instance,

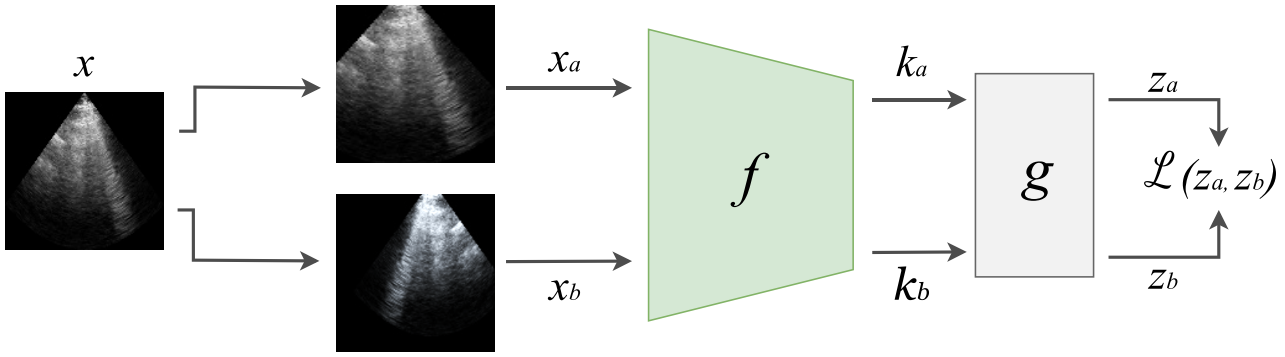


FIGURE 2. A depiction of the forward pass for a batch of images in a standard joint embedding SSL task. Batches of images are subjected to stochastic data transformations twice, producing distorted views x_a and x_b, which are passed through the feature extractor f to yield feature representations k_a and k_b. The projector g transforms k_a and k_b into embeddings z_a and z_b respectively. Typically, the objective function $\mathcal{L}(z_a, z_b)$ is optimized to maximize the similarity of z_a and z_b.

Zhang et al. [32] trained a single neural network with dedicated output layers for the classification of carotid plaques and estimation of the degree of stenosis on CT angiography imaging. Xu et al. [33] proposed a single convolutional neural network (CNN) architecture for adbominal US view classification and landmark localization using features from intermediate residual blocks as input for both tasks. Focusing on hierarchical interpretation, Fu et al. [34] proposed a system for medical image classification consisting of a convolutional neural network (CNN) followed by a decision tree in which each node is a linear classifier [34]. Decision trees with neural network nodes have also been proposed [22].

In this study, we showed that a single CNN pretrained with self supervision provides sufficient feature representations for multiple tasks, including tasks arranged hierarchically. Note that the methodology in this work is distinct from multi-task learning in that it explores the feasability of reusing a single self-supervised pretrained feature extractor for the development of multiple LUS classifiers.

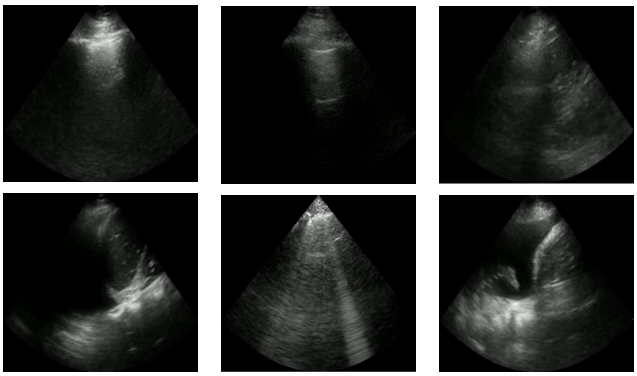
III. METHODS

A. LUS CLASSIFICATION TASKS

The LUS interpretative workflow addressed in this work has been described as a decision tree [35]. After determining the view, the interpreter traverses down the tree to look for increasingly specific artifacts that reduce a possible differential diagnosis. We focused on three binary classification tasks for LUS image interpretation: view classification (`View`), A-line versus B-line classification (`A/B`), and pleural effusion detection (`PE`). The `A/B` task is applicable to parenchymal LUS views, and the `PE` task is applicable to pleural LUS views. Table 1 summarizes these tasks, and Fig. 1 displays emblematic examples for each class.

B. DATA

Datasets from one local and one external healthcare institution were extracted from a private repository of LUS videos. Access to the data was permitted via ethics approval granted by Western University (REB 116838). The dataset



(a) `View`: Parenchymal (top), pleural (bottom)

(b) `AB`: A-lines (top), B-lines (bottom)

(c) `PE`: No PE (top), PE (bottom)

FIGURE 3. Examples of each class for each LUS binary classification task: `View` **(a),** `AB` **(b), and** `PE` **(c).**

had been previously labeled for the `View`, `AB`, and `PE` tasks by competent LUS interpreters during prior work [11], [36]. The labeled portion of the local dataset was split by patient identifier into a training set (70%), validation set (15%), and test set (15%), and the external dataset was reserved for testing only. Local videos with no labels were used only during self-supervised pretraining. Table 2 details the cardinalities and class distribution of these datasets. Regions peripheral to the US beam were expunged of extraneous visual artifacts, and the images were cropped to the boundaries of the beam. All images were downsampled to 128 × 128 pixels.

We also evaluated the effectiveness of SimCLR-pretrained weights on the public COVIDxUS dataset [37], splitting it by video identifier into a training, validation, and test set. Although patient identifiers were not available for every video, we ensured that multiple videos from the same patient identifier were contained in the same set. COVIDxUS contains 243 LUS videos (29 651 images) originating from a variety of manufacturers and clinical sources. Each example belongs to one of four classes: normal lung, COVID-19,

TABLE 1. A summary of the LUS tasks addressed in this study.

Task	Description	Negative Class	Positive Class
View	Distinguishing the type of LUS view	Parenchymal	Pleural
A/B	Normal versus abnormal lung tissue, as indicated by the presence of A-lines and B-lines respectively. Visualized in parenchymal views.	A-lines	B-lines
PE	Absence or presence of pleural effusion (PE). Visualized in pleural views.	No effusion	Effusion

TABLE 2. Breakdown of the institutional US datasets used in this study. For each LUS binary classification task, *x* / *y* indicates the number of negative and positive examples respectively.

	Local				External
	Unlabeled	Train	Validation	Test	Test
Videos	9993	3545	753	805	374
Patients	4919	975	210	210	53
Images	2 192 361	757 492	161 302	157 797	45317
View	-	520 128 / 237 364	113 436 / 47 866	109 684 / 48 113	33 483 / 11 834
A/B	-	195 000 / 88 361	43 980 / 17 269	43 378 / 17 719	12 173 / 12 078
PE	-	110 165 / 40 119	27 624 / 8236	21 081 / 9838	6830 / 2641

non-COVID-19 pneumonia pneumonia, and other pathologies. The dataset was sourced from openly available LUS examinations acquired at a variety of institutions with an assortment of ultrasound devices.

C. SELF-SUPERVISED PRETRAINING

Three joint embedding SSL methods were trialed to produce pretrained models for each LUS task: SimCLR (with $\tau = 0.1$) [23], Barlow Twins (with $\lambda = 0.005$) [25], and VICReg (with $\lambda = 25$, $\mu = 25$, $\nu = 1$) [26]. As was done in the original studies, positive pairs were produced by distorting images by applying stochastic data augmentations sampled from a family of transformations. Fig. 4 provides examples of augmented views of B-mode images from the local dataset. Below is the list of transformations, where P indicates the probability of a transformation being applied:

1) Random crop of $c \sim \mathcal{U}(0.5, 1.0)$ of the image's area. ($P = 0.8$).
2) Horizontal flip. ($P = 0.5$)
3) Multiplicative Gaussian noise, with SD $\sigma \sim \mathcal{U}(0.0, 0.1)$. ($P = 0.5$).
4) Brightness adjustment by $c \sim \mathcal{U}(0.5, 1.5)$. ($P = 0.7$).
5) Contrast adjustment by $c \sim \mathcal{U}(0.6, 1.0)$. ($P = 0.7$). With probability 0.5, this occurs before brightness adjustment.

Feature extractors were pretrained for 15 epochs using the union of the unlabeled and training images. The MobileNetV3 [38] architecture, initialized with ImageNet-pretrained weights, was employed as the feature extractor for all pretraining. The output of this architecture is a 576-dimensional feature representation vector.

D. EVALUATION PROTOCOL

We compared pretrained models with fully supervised models initialized with ImageNet-pretrained weights. The following

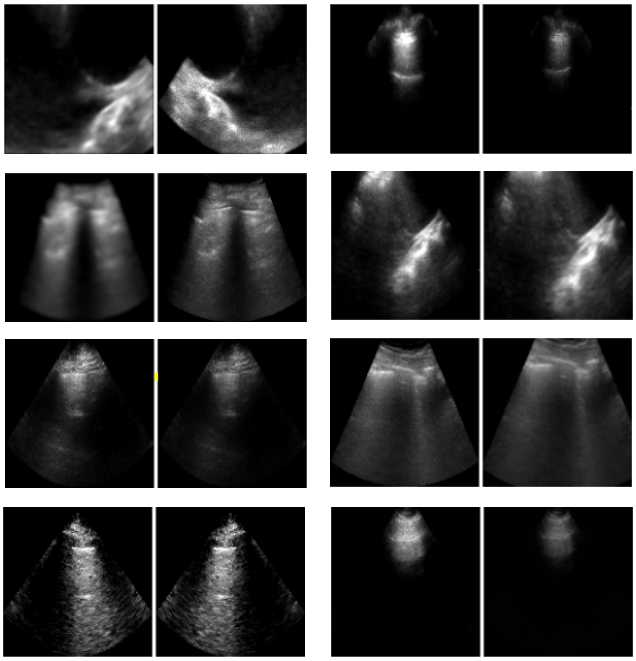


FIGURE 4. Augmented views of B-mode images, comprising positive pairs for self-supervised pretraining.

experiments were conducted to determine the pretrained models' effectiveness at learning the LUS tasks.

- **Linear classification** (LC): The weights of the feature extractor were held constant, and a linear classifier was trained using its outputted feature representations.
- **Fine-tuning** (FT): The weights of both the feature extractor and a linear head were trained.
- **Nonlinear classification** (NC): The weights of the feature extractor were held constant and a nonlinear head was trained on the features. The head consisted of a MLP with a single hidden layer of 32 nodes with rectified linear unit activation.

TABLE 3. AUC evaluated on the local and external test sets for the linear classification (LC), fine-tuning (FT), and nonlinear classification (NC) experiments. Results are presented for each of the `View`, `AB`, and `PE` tasks. The bottom row gives the geometric mean across tasks, with bold typeface indicating the best-performing pretraining strategy.

Task	Pretraining	Local test set			External test set		
		LC	FT	NC	LC	FT	NC
`View`	SimCLR	0.966	0.982	0.976	0.887	0.922	0.910
	Barlow Twins	0.959	0.978	0.973	0.889	0.920	0.908
	VICReg	0.960	0.980	0.975	0.894	0.917	0.909
	None	0.951	0.978	0.976	0.908	0.925	0.903
`AB`	SimCLR	0.949	0.976	0.969	0.890	0.896	0.898
	Barlow Twins	0.940	0.958	0.948	0.903	0.886	0.894
	VICReg	0.934	0.957	0.940	0.880	0.904	0.902
	None	0.923	0.945	0.939	0.820	0.833	0.856
`PE`	SimCLR	0.832	0.925	0.894	0.929	0.941	0.928
	Barlow Twins	0.855	0.906	0.893	0.845	0.940	0.923
	VICReg	0.849	0.935	0.871	0.878	0.936	0.921
	None	0.857	0.867	0.946	0.854	0.821	0.972
Mean	SimCLR	0.914	**0.961**	0.946	**0.902**	**0.919**	**0.912**
	Barlow Twins	**0.917**	0.947	0.937	0.879	0.915	0.908
	VICReg	0.913	0.957	0.928	0.884	0.919	0.911
	None	0.909	0.929	**0.954**	0.860	0.858	0.909

Fig. 1 (images 3a & 3b) illustrates how FT and NC each implement hierarchical LUS interpretation for the tasks of interest. In all trials, the initial learning rates for the feature extractor and head were 1×10^{-5} and 1×10^{-4} , respectively. The learning rates were multiplied by a factor of $e^{-0.02}$ each epoch. Models were trained for 10 epochs to minimize the binary cross-entropy loss function. The weights resulting in the lowest validation loss were retained. We assessed model performance by determining the area under the receiver operating characteristic curve (AUC) on the local and external test sets. All experiments were conducted using a system with an Intel i9-10900K CPU at 3.7 GHz and a Nvidia GeForce RTX 3090 GPU.

IV. RESULTS

A. LINEAR EVALUATION (LC)

Feature extractors were pretrained using SimCLR [23], Barlow Twins [25], and VICReg [26]. To evaluate the separability of the resulting representations with respect to the three LUS tasks, linear classifiers were trained on the output of the feature extractors. In each trial, the feature extractor's weights were held constant, and a perceptron was fitted for each binary LUS task, using the pretrained representations as input. Table 3 provides the performance of classifiers trained for the LC experiment on the local and external test sets. AUC was designated as the primary evaluation metric, but additional classification metrics are reported in the Appendix.

We compared the test AUC exhibited by classifiers initialized with self-supervised pretrained weights against classifiers initialized with ImageNet-pretrained weights (i.e., no self-supervision). Hereafter, we refer to the models initialized without self-supervised pretraining as *fully supervised*. In the case of linear evaluation, self-supervised pretraining resulted in greater performance on local test data for `View` and `AB`, but not for `PE`. On the external test set, the linear classifiers for pretrained models outperformed fully supervised models on `AB`, but not for `View`. The lack of difference for `View` may be the result of the greatly increased number of training examples labeled for `View` compared to the other tasks. SimCLR- and VICReg-pretrained models performed better on the `PE` task with external test data than fully supervised models, but Barlow Twins-pretrained models did not – a finding that was unique to this experiment.

B. FINE-TUNING EVALUATION (FT)

As in the LC experiment, a single-node fully connected layer was appended to the pretrained feature extractors. The entire network was then fine-tuned, facilitating task specialization by the feature extractor. The learning rate for the feature extractor was set to a tenth of that of the output layer (specifics are provided in Section III-D). Table 3 provides the AUC evaluated on the local and external test sets for the FT experiment. Each pretraining method achieved greater AUC on the local test set across all tasks. On the external test set, self-supervised pretrained models resulted in a minimum AUC improvement across pretraining methods of 0.053 and 0.115 for `AB` and `PE`, respectively. Improved external test set performance suggested improved generalizability of the pretrained classifiers. On the external `View` test set, fully supervised methods achieved the greatest AUC by a very small margin. In fact, the performance of fully supervised and pretrained models was close for `View` on both local and external test data – a finding that may be due to the substantially larger number of labels available for the `View` task compared to the others. Section IV-E explores how these results differ when considerably fewer labels are available for each task.

The general result that fine-tuned models initialized with self-supervised pretrained feature extractors outperform fully

TABLE 4. Mean class-wise AUC on the COVIDxUS test set for FT and NC.

Experiment	SimCLR [local]	SimCLR [COVIDxUS]	Supervised [ImageNet]
FT	0.627	**0.709**	0.623
NC	0.621	**0.751**	0.685

supervised baselines is consistent with previous studies that investigated the utility of ultrasound-specific contrastive pretraining objectives for COVID-19 classification [19], [30]. A caveat is that the experiments by Chen et al. [30] comparing pretrained and fully supervised models were conducted in a semi-supervised setting, as opposed to self-supervision. The present study expands on previous results, showing that (1) self-supervised pretraining improved performance of fine-tuned classifiers on three additional LUS interpretation tasks, and (2) that non-contrastive self-supervised approaches also improved LUS classifiers.

C. NONLINEAR CLASSIFICATION (NC)

The last experiment type performed on local and external test was nonlinear classification. Similar to the LC experiment, the objective was to evaluate the ability of unmodified pretrained feature representations to serve as inputs to a simple classifier Unlike LC, NC experiments make use of a 2-layer MLP, facilitating nonlinear decision boundaries. It was expected that such models would not perform as well as fine-tuning, since the feature extractor was barred from specializing in the task. However, should the performance gap with lightweight nonlinear classifiers be acceptably small, a hierarchical multi-task interpretation system could be constructed that resuses representations computed with one pass of the pretrained model. Here we detail the performance of lightweight classifiers on test data and compare them to fine-tuned models.

As shown in Table 3, pretraining did not consistently result in clear improvement on local test data. Similar to FT, there was little difference in performance on `View`. On the `AB` task, multiple pretrained models outperformed their fully supervised counterparts on local and external test data. Most notably, a fully supervised MLP achieved higher AUC on the `PE` task on the local and external test sets than any of the pretrained models. While we are not able to offer a direct explanation for this finding, we speculate that the performance gap was related to the fact that there may have been considerably fewer examples of pleural effusions in the pretraining data. With 27.0% of the images with a `PE` label containing a pleural effusion and 31.3% of the images in the entire training set labeled as being a pleural view, we estimate that the approximate prevalence of pleural effusions across the dataset was 8.3%. It is possible that the pretrained models learned to produce stronger representation for parenchymal views and non-PE pleural views.

D. PUBLIC DATASET EVALUATION

To promote experimental replicability, we investigated the effect of self-supervised pretraining with SimCLR on COVIDxUS, a public LUS dataset [37]. As shown in Table 4, models pretrained with SimCLR on the COVIDxUS training set achieved better mean class-wise test AUC than fully supervised models. These results were consistent with the improvement in COVID-19 pneumonia classification observed by previous works employing contrastive self-supervised pretraining [19], [30]. To explore the transferability of pretrained weights, we conducted a separate training run using weights pretrained with SimCLR on the local LUS dataset. Although COVIDxUS contained less than a tenth of the number of videos in the local training set alone, it was amalgamated from a variety of institutions and device manufacturers. Despite having been pretrained on several more videos, the models pretrained on local data greatly underperformed those pretrained on COVIDxUS alone, while performing comparably to fully supervised models. The results highlighted the importance of pretraining on a data distribution that is similar to the LUS interpretation task of interest.

E. LABEL EFFICIENCY

FT and NC were repeated for ImageNet-pretrained and SimCLR initialization using 1%, 10%, and 50% of the training set to evaluate the label efficiency of self-supervised pretrained models. As depicted in Fig. 5, self-supervision improved performance on the local test set in most cases. Moreover, the performance gain realized with SimCLR pretraining was largest when training with 1% of the labels. Recalling from Sections IV-A, IV-B, and IV-C that pretrained models did not clearly outperform fully supervised models for the `View` task, the label efficiency experiments highlighted the utility of self-supervised pretraining for `View` when fewer labeled training examples were available.

A notable finding was the leading performance of the fully supervised MLPs for `PE` (the NC experiment), which we suspect was related to the smaller size of the local test set. However, fully supervised MLPs trained on only 1% of the available `PE` labels greatly underperformed those with feature representations from the SimCLR-pretrained model.

F. QUALITATIVE EVALUATION OF REPRESENTATIONS

Seeking to better understand the results, we visualized two-dimensional t-SNE [39] projections of the representations outputted by both an ImageNet-pretrained and a SimCLR-pretrained feature extractor. As can be seen in Fig. 6, the projections for `PE` were not well-separated, even after pretraining with SimCLR, offering insight into the comparatively diminished improvement imbued by pretraining. In contrast, the projections suggested that self-supervised pretraining improved the separability of the data for the `AB` task, which was reflected in the decidedly stronger

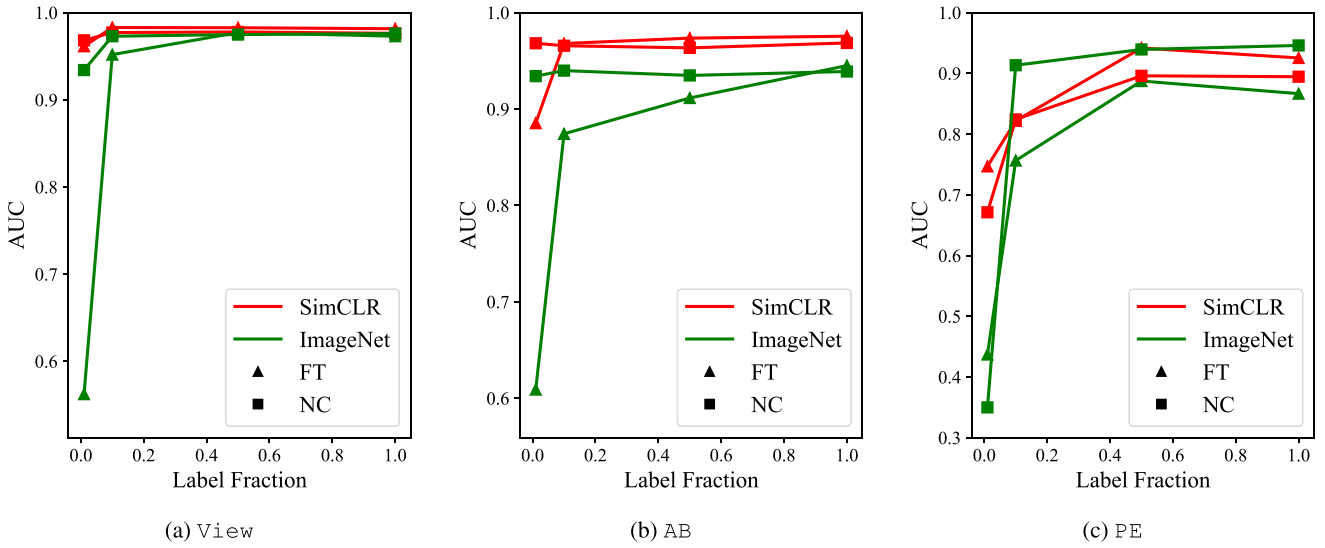


FIGURE 5. Local test AUC for supervised models initialized with ImageNet-pretrained weights and SimCLR-pretrained weights. Results are provided for the fine-tuning (FT) and nonlinear classification (NC) experiments training on various fractions of the labeled dataset.

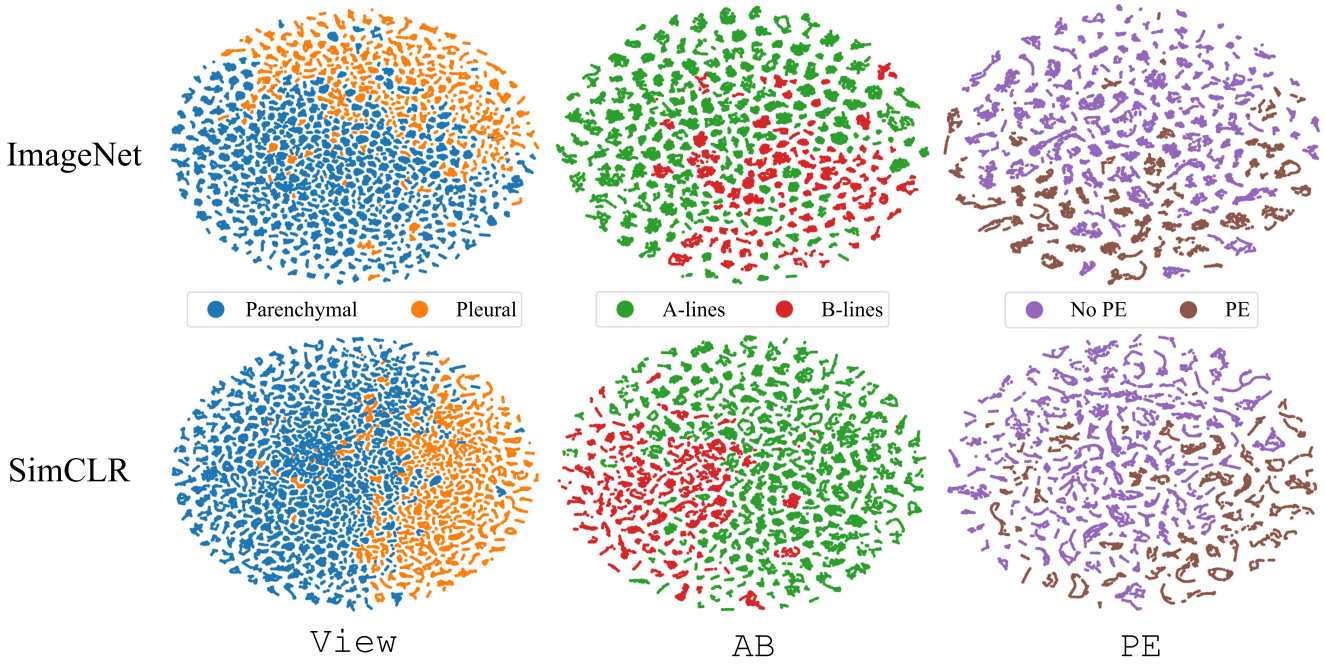


FIGURE 6. A comparison of t-SNE projections of features for the examples in the local test set outputted by a feature extractor initialized with ImageNet-pretrained weights before SimCLR pretraining (top) and after SimCLR pretraining (bottom).

performance of the SimCLR-pretrained model. The difference in performance after self-supervised pretraining was less clear for View, which may have occurred because there were significantly more labeled examples available for View (see Table 2). Moreover, the t-SNE projections for View exhibited separability before and after SimCLR pretraining.

G. INFERENCE EFFICIENCY

Recall that LUS interpretation consists of a hierarchically arranged set of tasks. Instead of training several task-specific models, the representations outputted by one self-supervised LUS feature extractor could be useful for multiple such tasks. Since speed is one of the essential qualities of POCUS, the design of automated LUS diagnostic assistive software

should prioritize runtime minimization. Real-time device inference could be accomplished by using copies of the output of a single feature extractor as input to multiple lightweight MLP classifiers. Furthermore, as LUS is a rapidly evolving diagnostic tool, classifiers for novel tasks could be integrated by training lightweight classifiers on the pretrained model's representations.

Although the results indicated that task-specific fine-tuning yielded the greatest per-task performance, any future work that improves nonlinear classification with pretrained weights would improve the execution time of a hierarchically arranged end-to-end LUS interpretation workflow. To quantify the potential runtime gains of using a single feature extractor for multi-task LUS classification, we compared the prediction time of two serially arranged fine-tuned CNNs (Fig. 1, 3a) against one feature extractor and two subsequent MLP classifiers (Fig. 1, 3b). Both configurations reflect the decision tree connecting the `View`, `AB`, and `PE` tasks in the simplified LUS interpretation workflow. After conducting 1000 serial predictions, the former took an average of 0.116 s (SD 0.003 s), while the latter took an average of 0.059 s (SD 0.001 s), underlining the runtime advantage of multi-task inference with a shared feature extractor. With each feature extractor and MLP requiring 3.7×10^7 and 3.7×10^4 floating point operations respectively, reusing the output of a single feature extractor as input to multiple task-specific MLPs would save considerable computational resources. The LUS diagnostic tree depicted in Fig. 1 would require approximately half the floating point operations if each node was a lightweight MLP instead of an entire CNN. Future work should focus on improving the applicability of unaltered feature representations from self-supervised pretrained models for multiple LUS classification tasks.

V. DISCUSSION AND CONCLUSION

In this study, joint embedding SSL methods were observed to improve the performance of classifiers on a variety of LUS tasks, particularly when only a small fraction of labels were available. Fine-tuning self-supervised pretrained models for each task consistently yielded the greatest performance gains for each task, with SimCLR-pretrained models improving across-tasks average AUC improvement of 0.032 and 0.061 on local and external test sets, respectively. When holding the weights of pretrained feature extractors constant, linear classifiers trained on representations from self-supervised models consistently achieved greater across-task average AUC on local and external test data. MLP classifiers trained on features outputted by self-supervised pretrained models did not outperform fully supervised models on all tasks. Nevertheless, low-dimensional projections of feature representations provided qualitative evidence that the test examples were well-separated with respect to two of the three tasks studied.

Based on the results of this study, practitioners working on developing automated LUS interpretation software aided by machine learning should strongly consider pretraining a feature extractor using any of the image-based contrastive or non-contrastive SSL methods investigated in the experiments, particularly when the majority of the available images are unlabeled. Given that pretraining improved the performance of fine-tuned models for multiple LUS tasks, the results support the conclusion that self-supervised pretraining is a viable method to boost LUS classifier performance. Moreover, practitioners may benefit from reduced dependence on expensive labeling expertise, achieving high classifier performance even without an abundance of labels. Despite the benefits of fine-tuning, it was observed that training two-layer MLP networks on the feature representations outputted by pretrained models did not consistently improve performance across tasks. As such, practitioners should consider fine-tuning for specific LUS tasks instead of training with a frozen pretrained feature extractor's representations.

There are several directions for future works that could improve the usefulness of feature representations for multiple LUS tasks. Given the greatly reduced inference time for multi-task LUS interpretation when reusing features from a single pretrained feature extractor, there would be great merit in future work that improves the quality of pretrained feature extractors and the separability of their outputs with respect to multiple tasks. Any comprehensive LUS interpretation software should be capable of distinguishing multiple cardinal artifacts in LUS images (e.g., A-lines, B-lines, pleural effusions, consolidations). Separability of feature representations with respect to multiple classification tasks would facilitate the training of multiple lightweight MLPs after a single forward pass of the feature extractor. To discover how feature extractors may be improved to output representations useful for multiple LUS tasks, future studies could systematically ascertain the effect of LUS-specific data augmentations in joint embedding methods. Alternative definitions of a positive pair for LUS videos could be explored as well. Instead of distorting the same image twice, one could explore the effect of intra-video positive pairs for multiple LUS tasks, along with sample weights for SSL objectives that exploit temporal proximity in B-mode videos. Future work could also examine the effect of fine-tuning a subset of the deeper layers in the feature extractor, effectively implementing a tradeoff between runtime efficiency for multi-task classification and full fine-tuning.

In summary, this study demonstrated that joint embedding self-supervised pretraining is a practical strategy for improving performance on LUS classification tasks when a fraction of the available data is labeled. More broadly, the findings imply that access to immense clinical labeling resources is not necessary to develop proficient LUS classifiers; rather, unlabeled data can be adapted via self-supervised learning to achieve improved performance.

APPENDIX
CLASSIFICATION METRICS

Tables 5 and 6 provide additional classification metrics from the evaluation on the local and external test sets, respectively

TABLE 5. Classification metrics calculated based on predictions for the local test set for the LC, FT, and NC experiments.

Task	Pretraining	Precision			Recall			Specificity		
		LC	FT	NC	LC	FT	NC	LC	FT	NC
View	SimCLR	0.922	0.930	0.932	0.855	0.921	0.946	0.968	0.969	0.972
	Barlow Twins	0.908	0.928	0.919	0.838	0.909	0.887	0.963	0.969	0.966
	VICReg	0.896	0.929	0.877	0.231	0.903	0.877	0.958	0.970	0.964
	None	0.913	0.912	0.916	0.799	0.906	0.879	0.967	0.961	0.965
AB	SimCLR	0.893	0.934	0.902	0.795	0.839	0.815	0.961	0.976	0.964
	Barlow Twins	0.812	0.872	0.857	0.799	0.818	0.793	0.925	0.951	0.946
	VICReg	0.827	0.859	0.825	0.792	0.819	0.774	0.932	0.945	0.933
	None	0.852	0.778	0.885	0.677	0.861	0.760	0.952	0.899	0.960
PE	SimCLR	0.824	0.908	0.848	0.593	0.680	0.662	0.941	0.968	0.945
	Barlow Twins	0.801	0.899	0.841	0.556	0.699	0.650	0.935	0.963	0.943
	VICReg	0.838	0.928	0.817	0.526	0.717	0.602	0.953	0.974	0.937
	None	0.925	0.996	0.940	0.468	0.371	0.667	0.982	0.999	0.980

TABLE 6. Classification metrics calculated based on predictions for the external test set for the LC, FT, and NC experiments.

Task	Pretraining	Precision			Recall			Specificity		
		LC	FT	NC	LC	FT	NC	LC	FT	NC
View	SimCLR	0.875	0.877	0.889	0.737	0.791	0.762	0.963	0.961	0.966
	Barlow Twins	0.892	0.931	0.930	0.725	0.769	0.724	0.969	0.980	0.981
	VICReg	0.946	0.918	0.926	0.672	0.771	0.720	0.986	0.976	0.980
	None	0.783	0.846	0.811	0.664	0.769	0.737	0.935	0.951	0.939
AB	SimCLR	0.906	0.866	0.876	0.676	0.736	0.729	0.929	0.885	0.896
	Barlow Twins	0.931	0.904	0.901	0.704	0.742	0.726	0.948	0.921	0.920
	VICReg	0.935	0.921	0.928	0.684	0.726	0.698	0.952	0.937	0.945
	None	0.810	0.730	0.847	0.612	0.830	0.665	0.855	0.691	0.880
PE	SimCLR	0.979	0.995	0.967	0.663	0.643	0.627	0.993	0.998	0.989
	Barlow Twins	0.936	0.996	0.984	0.588	0.605	0.615	0.979	0.999	0.995
	VICReg	0.876	0.998	0.982	0.618	0.602	0.649	0.954	0.999	0.994
	None	0.739	1.00	0.999	0.482	0.235	0.597	0.910	1.00	1.00

(detailed in Sections IV-A, IV-B, and IV-C). To supplement the AUC reported in Table 3, we provide precision, recall (i.e., sensitivity), and specificity. All available training labels were employed for these experiments.

Table 5 provides a more detailed picture of the performance on the local test set. Notably, metrics are similar for View, with most pretrained models exhibiting slightly greater performance in the fine-tuning setting for each metric. For the AB task, pretrained models exhibited decidedly greater precision and specificity, but lower recall. The fully supervised models predicted very few false positives for PE on the local test set, as evidenced by their high precision and specificity. However, the recall of linear and fine-tuned PE classifiers was abhorrently low, suggesting that a plenitude of false negative predictions were made. The fully supervised MLP trained for PE performed particularly well compared to pretrained models.

External test set metrics are reported in Table 6. In contrast to the local test set, the performance of pretrained models was nearly consistently greater than that of fully supervised models across all three tasks. Most notably, linear classifiers always achieved greater precision, recall, and specificity on external test data. The findings indicated that feature representations from pretrained models may be more generalizable than fully supervised models.

REFERENCES

[1] D. Lichtenstein, I. Goldstein, E. Mourgeon, P. Cluzel, P. Grenier, and J.-J. Rouby, "Comparative diagnostic performances of auscultation, chest radiography, and lung ultrasonography in acute respiratory distress syndrome," *J. Amer. Soc. Anesthesiol.*, vol. 100, no. 1, pp. 9–15, 2004.

[2] K. Nagarsheth and S. Kurek, "Ultrasound detection of pneumothorax compared with chest X-ray and computed tomography scan," *Amer. Surgeon*, vol. 77, no. 4, pp. 480–483, Apr. 2011.

[3] N. Xirouchaki, E. Magkanas, K. Vaporidi, E. Kondili, M. Plataki, A. Patrianakos, E. Akoumianaki, and D. Georgopoulos, "Lung ultrasound in critically ill patients: Comparison with bedside chest radiography," *Intensive Care Med.*, vol. 37, no. 9, pp. 1488–1493, Sep. 2011.

[4] K. Alrajhi, M. Y. Woo, and C. Vaillancourt, "Test characteristics of ultrasonography for the detection of pneumothorax," *Chest*, vol. 141, no. 3, pp. 703–708, Mar. 2012.

[5] P. Nazerian, G. Volpicelli, S. Vanni, C. Gigli, L. Betti, M. Bartolucci, M. Zanobetti, F. R. Ermini, C. Iannello, and S. Grifoni, "Accuracy of lung ultrasound for the diagnosis of consolidations when compared to chest computed tomography," *Amer. J. Emergency Med.*, vol. 33, no. 5, pp. 620–625, May 2015.

[6] D. Chiumello, M. Umbrello, G. F. S. Papa, A. Angileri, M. Gurgitano, P. Formenti, S. Coppola, S. Froio, A. Cammaroto, and G. Carrafiello, "Global and regional diagnostic accuracy of lung ultrasound compared to CT in patients with acute respiratory distress syndrome," *Crit. Care Med.*, vol. 47, no. 11, pp. 1599–1606, 2019.

[7] M. Wang, X. Luo, L. Wang, J. Estill, M. Lv, Y. Zhu, Q. Wang, X. Xiao, Y. Song, M. S. Lee, H. S. Ahn, J. Lei, and J. Tian, "A comparison of lung ultrasound and computed tomography in the diagnosis of patients with COVID-19: A systematic review and meta-analysis," *Diagnostics*, vol. 11, no. 8, p. 1351, Jul. 2021.

[8] A. K. Brady, C. R. Spitzer, D. Kelm, S. B. Brosnahan, M. Latifi, and K. M. Burkart, "Pulmonary critical care fellows' use of and self-reported barriers to learning bedside ultrasound during training," *Chest*, vol. 160, no. 1, pp. 231–237, Jul. 2021.

[9] Y. Y. Greenstein and K. Guevarra, "Point-of-care ultrasound in the intensive care unit: Applications, limitations, and the evolution of clinical practice," *Clinics Chest Med.*, vol. 43, no. 3, pp. 373–384, 2022.

[10] A. S. Ginsburg, Z. Liddy, P. T. Khazaneh, S. May, and F. Pervaiz, "A survey of barriers and facilitators to ultrasound use in low- and middle-income countries," *Sci. Rep.*, vol. 13, no. 1, p. 3322, Feb. 2023.

[11] R. Arntfield, D. Wu, J. Tschirhart, B. VanBerlo, A. Ford, J. Ho, J. McCauley, B. Wu, J. Deglint, R. Chaudhary, C. Dave, B. VanBerlo, J. Basmaji, and S. Millington, "Automation of lung ultrasound interpretation via deep learning for the classification of normal versus abnormal lung parenchyma: A multicenter study," *Diagnostics*, vol. 11, no. 11, p. 2049, Nov. 2021.

[12] N. Durrani, D. Vukovic, J. van der Burgt, M. Antico, R. J. G. van Sloun, D. Canty, M. Steffens, A. Wang, A. Royse, C. Royse, K. Haji, J. Dowling, G. Chetty, and D. Fontanarosa, "Automatic deep learning-based consolidation/collapse classification in lung ultrasound images for COVID-19 induced pneumonia," *Sci. Rep.*, vol. 12, no. 1, p. 17581, Oct. 2022.

[13] G. F. L. Tan, T. Du, J. S. Liu, C. C. Chai, C. M. Nyein, and A. Y. L. Liu, "Automated lung ultrasound image assessment using artificial intelligence to identify fluid overload in dialysis patients," *BMC Nephrol.*, vol. 23, no. 1, p. 410, Dec. 2022.

[14] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang, "Deep learning in medical ultrasound analysis: A review," *Engineering*, vol. 5, no. 2, pp. 261–275, 2019.

[15] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3458–3468.

[16] F. Haghighi, M. R. H. Taher, M. B. Gotway, and J. Liang, "DiRA: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20792–20802.

[17] Z. Zhou, V. Sodha, J. Pang, M. Gotway, and J. Liang, "Models genesis," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101840.

[18] S. Perek, M. Amit, and E. Hexter, "Self supervised contrastive learning on multiple breast modalities boosts classification performance," in *Proc. Int. Workshop Predictive Intell. Med.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 12928, 2021, pp. 117–127.

[19] S. Basu, S. Singla, M. Gupta, P. Rana, P. Gupta, and C. Arora, "Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 13434, 2022, pp. 423–433.

[20] H. Ma, Y. Guo, Q. Wang, M. Liu, Y. Qiang, X. Guo, Y. Guo, and Q. Chen, "Classification decision tree in CT imaging: Application to the differential diagnosis of solitary pulmonary nodules," *Chin. J. Radiol.*, pp. 50–55, 2008.

[21] E. J. Shim, M. A. Yoon, H. J. Yoo, C. G. Chee, M. H. Lee, S. H. Lee, H. W. Chung, and M. J. Shin, "An MRI-based decision tree to distinguish lipomas and lipoma variants from well-differentiated liposarcoma of the extremity and superficial trunk: Classification and regression tree (CART) analysis," *Eur. J. Radiol.*, vol. 127, Jun. 2020, Art. no. 109012.

[22] S. H. Yoo, H. Geng, T. L. Chiu, S. K. Yu, D. C. Cho, J. Heo, M. S. Choi, I. H. Choi, C. Cung Van, N. V. Nhung, B. J. Min, and H. Lee, "Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging," *Frontiers Med.*, vol. 7, p. 427, Jul. 2020.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[24] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[25] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.

[26] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[27] N.-Q. Nguyen and T.-S. Le, "A semi-supervised learning method to remedy the lack of labeled data," in *Proc. 15th Int. Conf. Adv. Comput. Appl. (ACOMP)*, Nov. 2021, pp. 78–84.

[28] M. Saeed, R. Muhtaseb, and M. Yaqub, "Contrastive pretraining for echocardiography segmentation with limited data," in *Proc. Annu. Conf. Med. Image Understand. Anal.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 13413, 2022, pp. 680–691.

[29] D. Anand, P. Annangi, and P. Sudhakar, "Benchmarking self-supervised representation learning from a million cardiac ultrasound images," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 529–532.

[30] Y. Chen, C. Zhang, L. Liu, C. Feng, C. Dong, Y. Luo, and X. Wan, "USCL: Pretraining deep ultrasound image diagnosis model through video contrastive representation learning," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer, 2021, pp. 627–637.

[31] J. Born, G. Brändle, M. Cossio, M. Disdier, J. Goulet, J. Roulin, and N. Wiedemann, "POCOVID-Net: Automatic detection of COVID-19 from a new lung ultrasound imaging dataset (POCUS)," 2020, *arXiv:2004.12084*.

[32] W. Zhang, G. Yang, N. Zhang, L. Xu, X. Wang, Y. Zhang, H. Zhang, J. Del Ser, and V. H. C. de Albuquerque, "Multi-task learning with multi-view weighted fusion attention for artery-specific calcification analysis," *Inf. Fusion*, vol. 71, pp. 64–76, Jul. 2021.

[33] Z. Xu, Y. Huo, J. Park, B. Landman, A. Milkowski, S. Grbic, and S. Zhou, "Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 711–719.

[34] G. FU, R. Wang, J. Li, M. Vakalopoulou, and V. Kalogeiton, "Me-NDT: Neural-backed decision tree for visual explainability of deep medical models," in *Medical Imaging With Deep Learning*, 2021.

[35] N. J. Soni, R. Arntfield, and P. Kory, *Point of Care Ultrasound E-Book*. Amsterdam, The Netherlands: Elsevier, 2019.

[36] B. VanBerlo, D. Smith, J. Tschirhart, B. VanBerlo, D. Wu, A. Ford, J. McCauley, B. Wu, R. Chaudhary, C. Dave, J. Ho, J. Deglint, B. Li, and R. Arntfield, "Enhancing annotation efficiency with machine learning: Automated partitioning of a lung ultrasound dataset by view," *Diagnostics*, vol. 12, no. 10, p. 2351, Sep. 2022.

[37] A. Ebadi, P. Xi, A. MacLean, S. Tremblay, S. Kohli, and A. Wong, "COVIDx-U.S.—An open-access benchmark dataset of ultrasound imaging data for AI-driven COVID-19 analytics," 2021, *arXiv:2103.10003*.

[38] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[39] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

BLAKE VANBERLO (Graduate Student Member, IEEE) received the B.E.Sc. degree in software engineering from Western University, London, ON, Canada, in 2017. He is currently pursuing the Ph.D. degree with the Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada. Since, he has been a Summer Research Assistant with the Robarts Research Institute and a Machine Learning Developer Intern with Unity Technologies. He was also a Freelance Artificial Intelligence Consultant, primarily focusing on development and deployment of municipal applications. During the Ph.D. degree, he was a Sessional Instructor with the University of Waterloo and the Director of Machine Learning with Deep Breathe. He is an Alumnus of the Schulich Leader Scholarship Program. He is also a recipient of the Vanier Canada Graduate Scholarship.

BRIAN LI is currently pursuing the bachelor's degree in biomedical engineering with the University of Waterloo, Waterloo, ON, Canada. Previously, he was a Machine Learning Engineering Intern with DarwinAI and a Data Science Intern with Capital One. He is also a Machine Learning Engineer with Deep Breathe while pursuing the bachelor's degree.

JESSE HOEY (Member, IEEE) received the Ph.D. degree in computer science from The University of British Columbia, in 2004. He is currently a Professor with the David R. Cheriton School of Computer Science, University of Waterloo, where he leads the Computational Health Informatics Laboratory (CHIL). He is also a Faculty Affiliate with the Vector Institute, and an Affiliate Scientist with KITE/TRI, Toronto. He has published over 100 peer-reviewed scientific articles. He is the Editor-in-Chief of IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and an Area Chair for the International Joint Conferences on Artificial Intelligence (IJCAI).

ALEXANDER WONG (Senior Member, IEEE) received the B.A.Sc. degree in computer engineering, the M.A.Sc. degree in electrical and computer engineering, and the Ph.D. degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 2005, 2007, and 2010, respectively. He is currently the Canada Research Chair of Artificial Intelligence and Medical Imaging, the Co-Director of the Vision and Image Processing Research Group, and a Professor with the Department of Systems Design Engineering, University of Waterloo. He is a fellow of the Royal Society of Public Health, Institution of Engineering and Technology, Institute of Physics, and International Society for Design and Development in Education. He has authored over 600 refereed journals and conference papers and patents, in various fields, such as computational imaging, artificial intelligence, computer vision, graphics, image processing, and multimedia systems. His research interests include integrative biomedical imaging systems design, operational artificial intelligence, and scalable and explainable deep learning. He is a member of the College of the Royal Society of Canada. He has received a number of awards, including two outstanding performance awards, the Distinguished Performance Award, the Engineering Research Excellence Award, the Sandford Fleming Teaching Excellence Award, the Early Researcher Award from the Ministry of Economic Development and Innovation, two magna cum laude awards and the Cum Laude Award from the Annual Meeting of the Imaging Network of Ontario, the Alumni Gold Medal, the Outstanding Paper Award at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges, in 2021, the Best Paper Award at the Conference on Neural Information Processing Systems (NIPS) Workshop on Transparent and Interpretable Machine Learning in 2017, the AquaHacking Challenge First Prize, in 2017, the Best Student Paper at the Ottawa Hockey Analytics Conference, in 2017, the Best Paper Award at the NIPS Workshop on Efficient Methods for Deep Neural Networks, in 2016, the Synaptive Best Medical Imaging Paper Award, in 2016, the Distinguished Paper Award by the Society of Information Display, in 2015, the Best Paper Award at the Conference of Computer Vision and Imaging Systems, in 2015 and 2017, and the Best Paper Award by the Canadian Image Processing and Pattern Recognition Society, in 2009 and 2014.