Deep Clustering for Abdominal Organ Classification in Ultrasound Imaging

Hind Dadoun[a,*], **Hervé Delingette**[a], **Anne-Laure Rousseau**[b], **Eric de Kerviler**[c], **Nicholas Ayache**[a]
[a]Inria Epione Team, Sophia Antipolis, Université Côte d'Azur, France
[b]Hôpital Européen Georges Pompidou, NHance, Paris, France
[c]Saint Louis Hospital, AP-HP, Paris, France

Abstract. The use of ultrasound (US) imaging has developed considerably in several medical specialties recently. In particular, abdominal pain accounts for a significant part of medical consultations. In this context, ultrasound is the only non-invasive and non-ionizing imaging modality that allows real-time medical exploration of a specific body part. However, acquiring and interpreting US images remains a difficult and examiner-dependent task, with a limited number of trained operators. For abdominal organs, ultrasound images are even more difficult to interpret because some of the organs of interest are located deep inside the body and patient-related factors, such as the presence of fatty tissue, can hinder the reading. In this work, we present a simple framework for abdominal organ clustering using unlabeled ultrasound images. This method can serve as a tool to preprocess large uncurated databases, reducing the need for annotation in abdominal ultrasound studies. When few labeled examples are available, we explore how unlabeled data can be leveraged to improve the performance of multi-label classification as opposed to the traditional transfer learning approach. In particular, we show that for supervised fine-tuning, deep clustering is an effective pre-training method, with performance matching that of ImageNet pre-training using five times less labeled data. Finally, we combine this pre-training method with semi-supervised learning and report the performances.

Keywords: ultrasound imaging, representation learning, deep clustering, semi-supervised learning..

***Hind Dadoun, hind.dadoun@inria.fr**

1 Introduction

Several studies have sought to classify abdominal organs on ultrasound images. Due to the lack of freely available databases, most of them use in-house data-sets of very different sizes ranging from 4094[1] to 187,219[2] labeled images. They all share a common approach: the use of transfer learning[1] to initialize their classification models, combined with supervised learning methods. Transfer learning allows the reuse of deep networks trained on large scale annotated databases such as ImageNet, to learn specific tasks for which fewer labeled examples are available, while leveraging the learned general-purpose features of the original network. This applies to ultrasound images where labeled examples are hard to obtain due to the limited availability of expert annotations.

[1]either by re-training the complete model, or by tuning the last layers of the model only.

In a study, the authors evaluate transfer learning (and more specifically fine-tuning) with deep CNNs for the classification of abdominal ultrasound images.[1] Classes were chosen based on 11 categories specified on the images, which would correspond to standard plane views acquired during an abdominal examination. Images that fell outside these standard plane views were excluded, in addition to color or spectral Doppler, and images with very limited or unrecognizable anatomy. The study shows that using a large VGGNet network trained on 4094 images yields 77.9% accuracy on a test set of 1423 images. On a slightly different setting, the authors propose a multi-task learning framework for both the classification of views (including a class for other views), and a landmark detection for each relevant view.[2] A total of 187,219 ultrasound images from 706 patients were collected, and 20% of the dataset was used for testing. For the classification task, the study reports a 4.07% improvement compared to single-task learning, suggesting that the classification task benefits from sharing the low level features with the landmark detection task. Finally, Li *et al*[3] use a public dataset of 360 ultrasound images to propose a classification method that combines the deep learning techniques and k-Nearest-Neighbor (k-NN) classification for the multi-label classification of six anatomical structures (bladder, bowel, gallbladder, kidney, liver, and spleen), making again the assumption that each image contains a single organ. In particular they show that for various classification models (ResNet-101, ResNet-152, DenseNet-121, DenseNet-169, and DenseNet-201) the use of a k-NN classification on top of fixed features outperforms the fine-tuning of a fully connected layer. However they also show that this does not apply when using a Resnet50 model, which is the most commonly used model in this setting.

In summary, the available literature on the subject focuses on supervised methods using transfer learning to overcome the lack of labeled data. These methods all use a multi-class classification models, either considering one organ per image, which is rarely valid in practice, or considering

classes as standard views (containing several organs), which forces the model to classify only known standard views.

In this study we consider a different angle, and explore how unlabeled data can be used for the task of abdominal organ classification in ultrasound images in the multi-label setting (non-mutually-exclusive classes), as multiple organs may be visible simultaneously on the same image in abdominal ultrasound. This is achieved by considering three different approaches to exploit unlabelled data: i) self-supervised followed by supervised learning, ii) semi-supervised learning or iii) combining self-supervised pre-training followed by semi-supervised learning. Fig. 1 provides a schematic overview of these three different learning methods, in addition to the transfer learning approach.

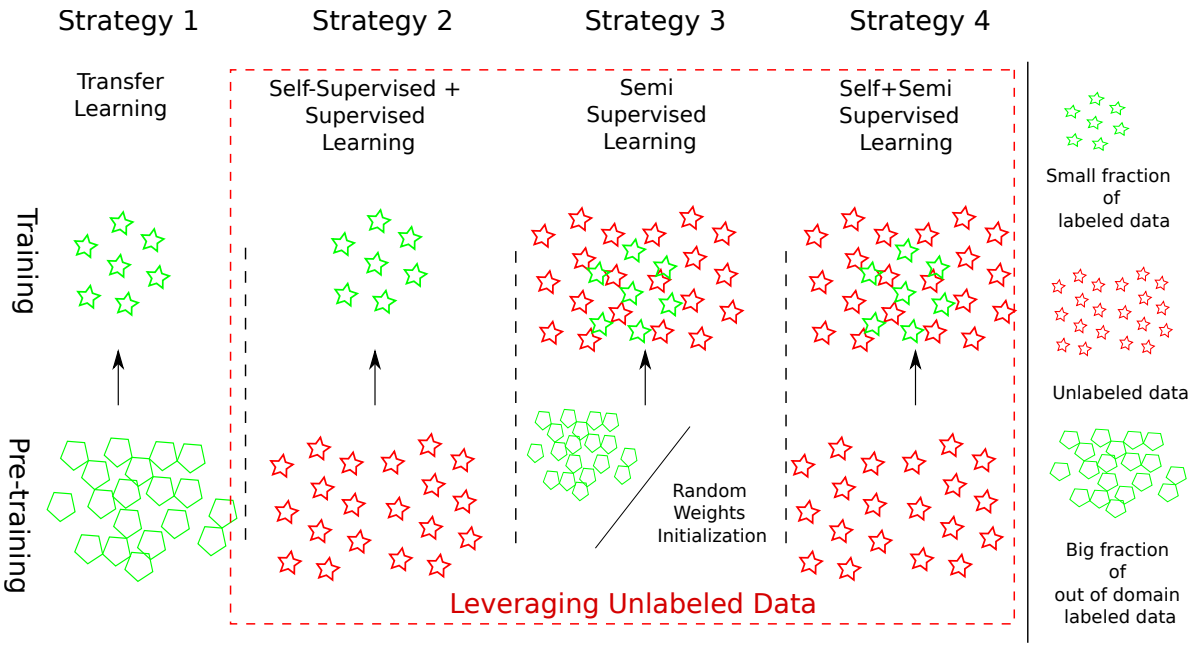

Fig 1 Overview of three different scenarios where unlabeled data (represented by red stars) can be leveraged with few labeled examples (represented by green stars): during pre-training in a self-supervised manner (Strategy 2), during training in a semi-supervised manner (Strategy 3), and during both stages (Strategy 4). Transfer learning (Strategy 1) is presented as a baseline method that does not require unlabeled data but rather a large amount of *out-of-domain* labeled data (represented by green polygons).

More specifically, we explore different questions in this context:

1. Are the features learned in a self-supervised manner more useful for downstream tasks on the same domain compared to features learned on a different domain database such as ImageNet (Fig. 1: Strategy 2 vs 1)?

2. Is there any advantage to leverage labeled and unlabeled data simultaneously during training through semi-supervised learning rather than consecutively as proposed by self-supervised methods (Fig. 1: Strategy 3 vs 2)?

3. Is there any relevance in combining the two paradigms (Fig. 1: Strategy 4)?

1.1 Self-supervised learning

Self-supervised learning methods are used to train networks on large scale unlabeled *in-domain* data using surrogate tasks before transferring those learned representations to more specific tasks on few labeled data. Following the taxonomy presented in Jing *et al*,[4] the surrogate tasks can be summarized in four categories:

- Generation based methods (e.g image inpainting[5]).
- Semantic-free label-based methods (e.g contour detection[6]).
- Cross modal-based methods (e.g audio and video correspondence[7–9]).
- Context based methods (e.g clustering[10]).

All of the above tasks share the same hypothesis, (1) Visual features are needed to solve the task and (2) the Convolutional Neural Networks (CNNs) can capture those features by solving the surrogate task. Previous works explored the use of such method in ultrasound imaging. For instance,

Jiao *et al.*[11] used cross-modal contrastive learning in multi-modal fetal ultrasound video and audio to learn strong representations and transfer them to a supervised task of fetal standard plane detection, demonstrating higher performance than with features whose weights were initialized with ImageNet. These results were obtained using 90 scans of 55,000 frames each for training, which in practice represents a large amount of labeled data.

1.2 Semi-supervised learning

Semi-supervised learning methods, on the other hand, utilize both labeled and unlabeled data during training to learn feature representations specific to the learning task. In this case, the networks are trained with an objective function composed of two terms: a supervised loss applied to labeled data and an unsupervised loss applied to unlabeled data. The latter is derived either using consistency regularization[12,13] or pseudo-labeling.[14,15] Most of these methods are evaluated on curated data sets where the data distribution is close to uniform and unlabeled data contains no novel class. Unfortunately, the strong performance of such methods does not always translate to non-curated data-sets[16] that violate assumptions implicit to the semi-supervised learning approaches. These assumptions[17] include smoothness, ("If two points x_1,x_2 in a high-density region are close, then so should be the corresponding outputs y_1,y_2"), and cluster assumption, ("If points are in the same cluster, they are likely to be of the same class"). In fetal 2D ultrasound imaging, a study[18] explored the use of consistency regularization with unsupervised data augmentation for anatomy classification to test whether the reported results in semi-supervised learning translate to non-ideal data. In particular, the authors show that the inclusion of challenging classes in the unlabeled set can harm the performance, compared to a supervised setting.

1.3 Deep Clustering

As described previously, several surrogate tasks can be considered to analyse the self-supervised learning framework. One of which is deep clustering, which we propose to use because it has the advantage of being closely related to the classification task. Furthermore, using the cluster assignments on a small development set allows to validate the cluster assumption needed for the semi-supervised methods. Traditionally, clustering methods such as K-Means aim to learn cluster assignments based on fixed feature representations. For high dimensional imaging data, in particular, such fixed features are generally not sufficiently discriminating and need to be learned during clustering. In a recent effort to address this shortcoming, Caron *et al.*,[10] proposed a deep clustering framework for the unsupervised learning of CNNs trained in an end-to-end fashion. The proposed method iterates between clustering the learned CNN features using K-Means (i.e clustering) and updating the CNN weights by predicting the cluster assignments as pseudo-labels (i.e representation learning). However, the process of alternating between these two learning objectives may be prone to error-propagation, which motivated other studies to propose simultaneous learning methods.[19,20] More recently, on a related application domain, Kart *et al*[21] adapted the framework of DeepCluster[10] to categorize uncurated large-scale cardiac MRI images, and used normalized mutual information (NMI) and cluster purity (CP) to evaluate the clustering quality which gives little information on the relevance and usefulness of the grouped semantic categories.

1.4 Contributions

Based on these observations, and keeping the above-mentioned questions in mind we propose a framework for the multi-label classification of abdominal organs in ultrasound images based on a large database of 6951 ultrasound examinations (89 830 images) from 5788 patients with very few

labeled examples. Our contributions can be listed as follows:

1. We propose to adapt two state-of-the-art multi-class methods to the multi-label classification setting: deep clustering with PICA[19] (Fig. 1: Strategy 2), and semi-supervised learning with FixMatch[22] (Fig. 1: Strategy 3).

2. We evaluate the use of deep clustering in self-supervised learning, and show that the learned features transfer better to the classification task, with performance higher than that of ImageNet initialization (Fig. 1: Strategy 2 vs 1).

3. We show that combining deep clustering pre-training with semi-supervised learning (Fig. 1: Strategy 4) yields robust results, even when the number of labelled examples is extremely limited (less than 275 images).

2 Methodology

2.1 Problem definition

We wish to assign to each ultrasound image a set of *non-mutually-exclusive* target labels corresponding to the C organs of interest: *liver, kidney, gallbladder, pancreas, spleen* and *bladder* or *other* if none of these organs are present in the image. This corresponds to the setting of multi-label classification, where the target label is a binary vector representing the absence or presence of each organ: $\mathbf{v} \in \{0,1\}^{C+1}$. We assume that we have access to a large amount of unlabeled data and very little labeled data. In this case, unlabeled data can be leveraged to improve classification performance either during pre-training and/or directly during training as presented in Fig. 1. In the remainder, we refer to the sets of unlabeled and labeled images respectively sampled during training at each iteration as $\mathbf{U} = \{I_1, ..., I_{N_u}\}$ and $\mathbf{S} = \{I_1, ..., I_{N_s}\}$ for the sake of clarity. We also

introduce a development set $\mathbf{D} = [(I_1, v_1), ..., (I_d, v_d)]$ whose objective is to choose the network settings that achieve the best performance on images the network has never seen. All the models presented hereafter share a common architecture composed of two parts:

1. A feature extractor $f(.)$ that maps an image $\mathbf{I}$ into a vector representation $\mathbf{x} = f(\mathbf{I})$.

2. A single and/or a multi-label classification head: $g(.)$ that assigns each feature representation $\mathbf{x}$ with a class membership distribution.

2.2 Deep clustering

Starting from a well-established deep clustering method (PICA[19]), we add a loss term to take into account the variability of the cluster size distribution at each iteration. To compare performance of both methods, we propose to use a small fraction of labeled data to assign each cluster to a relevant semantic multi-label category. The feature extractor of the best performing method can then be used to fine tune a supervised multi-label classification model.

2.2.1 PICA

Our method builds upon the framework of PICA presented in Fig. 2, where two different augmentation schemes $(\tilde{.})$ and $(\hat{.})$ are applied to all input images before passing through the CNN composed of a feature extractor $f(.)$ and a classification head $g_\tau(\mathbf{x}) = p = \{p_1, ..., p_K\}$ where K is the number of clusters. We denote by $\mathbf{P} \in \mathbb{R}^{N_u, K}$ the cluster prediction matrix of N_u images in $\mathbf{U}$. Then a Partition Uncertainty Index (PUI) is introduced as the cosine similarity set of all the cluster pairs:

$$\mathcal{M}_{PUI}(j_1, j_2) = cos(\hat{q}_{j_1}, \tilde{q}_{j_2}) \;\; \forall (j_1, j_2) \in [0, ..., K] \tag{1}$$

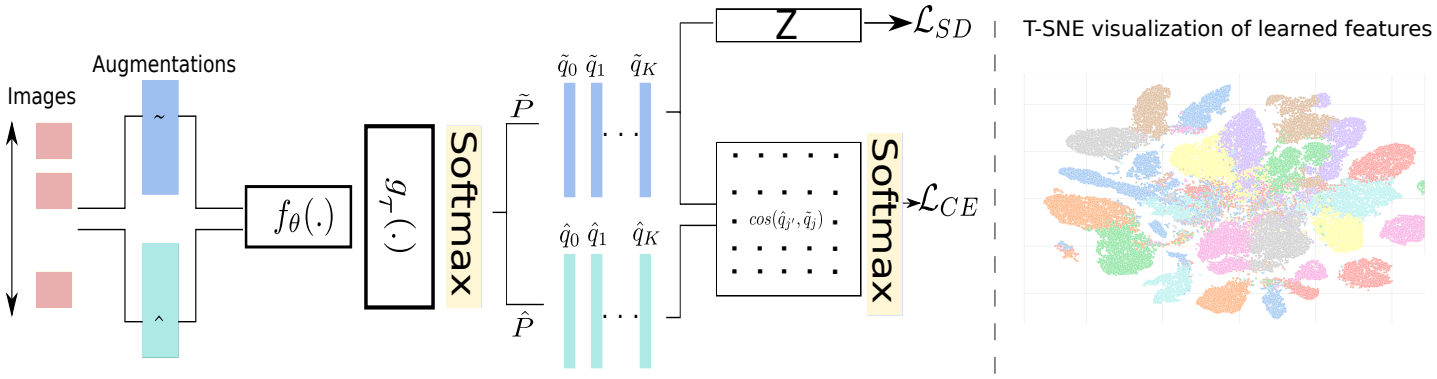

Fig 2 Overview of the deep clustering framework.Two different augmentation schemes $(\tilde{.})$ and $(\hat{.})$ are applied to all input images before passing through the CNN composed of a feature extractor $f(.)$ and a classification head $g_\tau(\mathbf{x}) = p = \{p_1, ..., p_K\}$. This framework can serve as an unsupervised classification model when no labeled examples are available. $\mathbf{P}$ is the cluster prediction matrix of all images in $\mathbf{U}$ and Z is the cluster size distribution. A T-SNE visualization of the learned features and their assigned clusters represented by different colors is shown on the right.

Where $\tilde{q}_j$ and $\hat{q}_j$ are the j-th rows of $\tilde{\mathbf{P}}$ and $\hat{\mathbf{P}}$ respectively.

To obtain the most confident predictions for each cluster without using any particular distance metric between samples, the authors propose to :

1. Force $\tilde{q}_{j_1}$ and $\hat{q}_{j_2}$ to be orthogonal when $j_1 \neq j_2$, which also means that their cosine similarity is equal to zero.

2. Force $\tilde{q}_j$ and $\hat{q}_j$ to be equal, meaning that predictions on two augmented views of the same image should be equal in which case the cosine similarity is equal to one.

To derive an objective function, the authors propose to apply a softmax operation as self-attention to each cluster j:

$$m_{j,j'} = \frac{\exp \mathcal{M}_{PUI}(j, j')}{\sum_{k=0}^{K} \exp \mathcal{M}_{PUI}(j, k)} \quad \forall j' \in [0, ..., K] \tag{2}$$

Yielding the following differentiable objective function:

$$\mathcal{L}_{CE} = \frac{1}{K} \sum_{j=0}^{K} -\log(m_{j,j}) \tag{3}$$

In addition, a constraint on the cluster size distribution Z is introduced to avoid trivial solutions:

$$\mathcal{L}_{NE} = \log(K) - H(Z) \tag{4}$$

with $z_i = \frac{\tilde{q}_i}{\sum_{k=0}^{K} \tilde{q}_k}$ and H is the entropy function.

The final objective function of PICA is formulated as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{NE}$$

with λ a hyper parameter.

2.2.2 Cluster size distribution

By minimising the negative entropy of the cluster size distribution $\mathcal{L}_{NE}$ introduced in (4), PICA forces $\mathbf{Z}$ to follow a uniform distribution at each iteration, meaning that all clusters have approximately the same size at each iteration. However, if the whole target data space is unbalanced, as it is the case in ultrasound imaging, then at each iteration the cluster size distribution needs to be also unbalanced. As we do not have any prior information to favor one cluster over any other, we set $\mathbf{Z}$ to follow a Symmetric Dirichlet distribution:

$$f\left(z_1, \ldots, z_K; \boldsymbol{\alpha}\right) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} z_i^{\alpha-1} \tag{5}$$

with B the Beta function, $\sum_{i=1}^{K} z_i = 1$, $z_i \geq 0$ for all $i \in \{1, \ldots, K\}$, and $\boldsymbol{\alpha} = [\alpha, ..., \alpha] \in \mathbb{R}^K$ such that $\alpha \geq 0$.

Doing so, the variance of the Dirichlet is captured by the magnitude of the chosen value (α) for the parameters:

- when $\alpha \gg 1$: the entropy of a Dirichlet draw is high, upper bounded by $\log(K)$ if $\alpha \to \infty$
- when $\alpha \ll 1$: the entropy of a Dirichlet draw approaches a delta peak on a random entry, which would correspond to an entropy equal to one.

In other words, (1) setting a large value for α would correspond to the case where at each iteration we force a uniform distribution (similar to using the negative entropy loss $\mathcal{L}_{NE}$), and (2) setting a small value for α would correspond to the trivial solution where at each iteration, all images are assigned to a single cluster. By taking α slightly higher than 1, we obtain a trade-off between both scenarios.

Therefore, we maximize the likelihood of $\mathbf{z}$ being a Symmetric Dirichlet distribution with $\alpha = 1 + \epsilon$:

$$\mathcal{L}_{SD} = -\frac{1}{log(K)} \sum_{k=0}^{K} (\alpha - 1) log(z_k) \tag{6}$$

The final objective function becomes:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{SD}(\alpha) \tag{7}$$

2.2.3 Matching clusters with multi-label targets

The objective of deep clustering is to separate the set of images into mutually exclusive clusters. Each cluster can then be matched with the dominant class and all images assigned to this cluster will have the same label. However, in our application, an image can contain several labels. As

such, we consider each different set of labels that exist in the multi-label dataset as a single label. By utilizing label priors, the number of plausible and most common combinations is set to $K < 2^C + 1$, where C corresponds to the number of classes. Doing so, the target label becomes a one hot encoding label ($\mathbf{w} \in \{0,1\}^K$) where the dimension K corresponds to the number of label combinations, leading to:

$$g_\tau(\mathbf{x}) = p = \{p_1, ..., p_K\} \tag{8}$$

To evaluate the relevance of the grouped semantic categories in the context of multi-label classification, instead of using directly the output probabilities p, we allocate to each cluster k a soft assignment. This is done by averaging the ground-truth multi-class labels $\mathbf{v} \in \{0,1\}^{C+1}$ of all images $\mathbf{I}$ in the development set $\mathbf{D}$ assigned to cluster k as shown in Algorithm 1: $\hat{v}_k = \frac{1}{|k|} \sum_{i \in k} v_i$.

Algorithm 1: Matching clusters with multi-label targets

Input

- Development set, $\mathbf{D} = [(I_1, v_1), ..., (I_d, v_d)]$.
- The cluster prediction matrix $\mathbf{P}$ of all images in $\mathbf{D}$.
- The target cluster number, K.
- The target class number, C.

Initialisation ;
$\hat{v}_k = [0, ..., 0] \in \mathbb{R}^{C+1}$
for $i \in \mathbf{D}$ ***:*** **do**
  $\hat{v}_{\arg\max(p_i)} += v_i$
end
for $k \in K$ ***:*** **do**
  $\hat{v}_k = \frac{1}{|k|} \times \hat{v}_k$
end
Output Cluster matching with dominating target class

2.2.4 *Fine-tuning for supervised learning*

The weights learned during the deep clustering phase can be used to initialize a multi-label supervised classification model. This implies using the same feature extractor f_θ but discarding entirely the single-label clustering head $g_\tau(\mathbf{x}) = p = \{p_1, ..., p_K\}$ and replacing it by a multi-label classification head $g_\Phi(\mathbf{x}) = p = \{p_0, ..., p_C\}$. During training, we use the supervised objective function defined in (9), where the target $\mathbf{v} \in \{0, 1\}^{C+1}$ is a vector representing the absence or presence of each organ. Model weights θ (feature extractor) and Φ (classification head) are learned jointly.

$$\mathcal{L}_{BCE} = \sum_{i \in \mathbf{S}} \frac{1}{C+1} \sum_{c=0}^{C} \mathbf{v}_c^{(i)} \log p_c^{(i)} + (1 - \mathbf{v}_c^{(i)})(1 - \log p_c^{(i)}) \tag{9}$$

2.3 *Semi-Supervised Classification*

Another common way of using unlabeled data is semi-supervised learning where unlabeled data is utilized simultaneously with labeled data during training using an unsupervised objective function for unlabeled examples. In the following, we present a state-of-the-art semi-supervised model for single-label classification. We then show how this model can be adapted to the multi-label classification setting. The feature extractor of this model can also be initialized with the deep clustering model, allowing to combine both methods.

2.3.1 FixMatch

The objective function $\mathcal{L}_u$ on the unlabeled set combines two lines of work:

- Pseudo-labeling: The objective is to use the model's prediction as pseudo-labels when the corresponding class probabilities fall above a certain threshold.

$$\mathcal{L}_u = \sum_{i \in \mathbf{U}} \mathbb{1}_{max(\tilde{p}^{(i)}) \gg \tau} \cdot \mathbf{H}(\tilde{y}^{(i)}, \tilde{p}^{(i)}) \tag{10}$$

  Where $\mathbf{H}$ is the cross-entropy and $\tilde{y}_c^{(i)} = \mathbb{1}_{\tilde{p}_c^{(i)} \geq \max_c(\tilde{p}_c^{(i)})}$ in the single-label setting.

- Consistency regularization: The objective is to force the model to output similar predictions when fed with perturbed versions of the same image. We denote by $\hat{(.)}$ and $\tilde{(.)}$ the strong and weak augmentations applied to the input images, the loss function in this case is:

$$\mathcal{L}_u = \sum_{i \in \mathbf{U}} \left\| \tilde{p}^{(i)} - \hat{p}^{(i)} \right\|^2 \tag{11}$$

Both objective functions can be combined, as described in FixMatch[22] by enforcing the pseudo-label $\tilde{y}^{(i)}$ obtained from the weak augmented version of the

image against the model's output probabilities for the strongly augmented version of the image $\hat{p}^{(i)}$ as follows:

$$\mathcal{L}_u = \sum_{i \in \mathbf{U}} \mathbb{1}_{max(\tilde{p}^{(i)}) \gg \tau} \cdot \mathbf{H}(\tilde{y}^{(i)}, \hat{p}^{(i)}) \tag{12}$$

Yielding the following global loss function in the semi-supervised setting:

$$\mathcal{L} = \sum_{j \in \mathbf{S}} \mathbf{H}(\mathbf{v}^{(j)}, p^{(j)}) + \sum_{i \in \mathbf{U}} \lambda \mathcal{L}_u(\tilde{p}^{(i)}, \hat{p}^{(i)}) \tag{13}$$

2.3.2 FixMatch for multi-label classification

The objective function derived in FixMatch supposes a single-label classification problem, where each pseudo-label is a one-hot encoding vector. Fig. 3 describes two ways of using FixMatch objective function in multi-label classification: i) using a multi-label objective function for both pseudo-labels and true labels by redefining the pseudo-label term (One-Head model) or ii) Using a single-label objective function for pseudo-labels and a multi-label objective function for true labels through the use of an additional classification head (Two-Head model).

1. One-Head model: One way to use the setting of FixMatch directly is to consider a different formulation for the pseudo-label. In this case, a Sigmoid

activation function is used instead of a Softmax function to ensure that the probability of each class is considered independently. Instead of taking the maximum probability of a class as the target label, we can threshold each probability class to 0.5:[23]

$$\tilde{y}_c^{(i)} = \mathbb{1}_{\tilde{p}_c^{(i)} \geq 0.5}$$

If the maximum probability of at least one class is greater than a threshold τ, then the derived pseudo-label is considered as a target label during training.

2. Two-Head model: Another way is to use a network with two classifiers $g_\phi(\mathbf{x}) = \{p_1, ..., p_K\}$ and $g_\kappa(\mathbf{x}) = \{\bar{p}_1, ..., \bar{p}_C\}$. Doing so we can compute $\mathcal{L}$ on the outputs of g_ϕ and add a supervised loss (Eq 9) on the output of g_κ. The semi-supervised objective function is slightly modified as follows:

$$\mathcal{L} = \sum_{j \in \mathbf{S}} \mathbf{H}(\mathbf{w}^{(j)}, p^{(j)}) + \mathcal{L}_{BCE}(\mathbf{v}^{(j)}, \bar{p}^{(j))}) + \lambda \sum_{i \in \mathbf{U}} \mathcal{L}_u(\tilde{p}^{(i)}, \hat{p}^{(i)})$$

2.3.3 Self and Semi-Supervised Learning

The feature extractor $f_\theta(\boldsymbol{x})$ of the semi-supervised learning model can be initialized either randomly (without pre-training) or from a pre-trained model (usually trained on a different database with a large size of labeled examples). It can also be

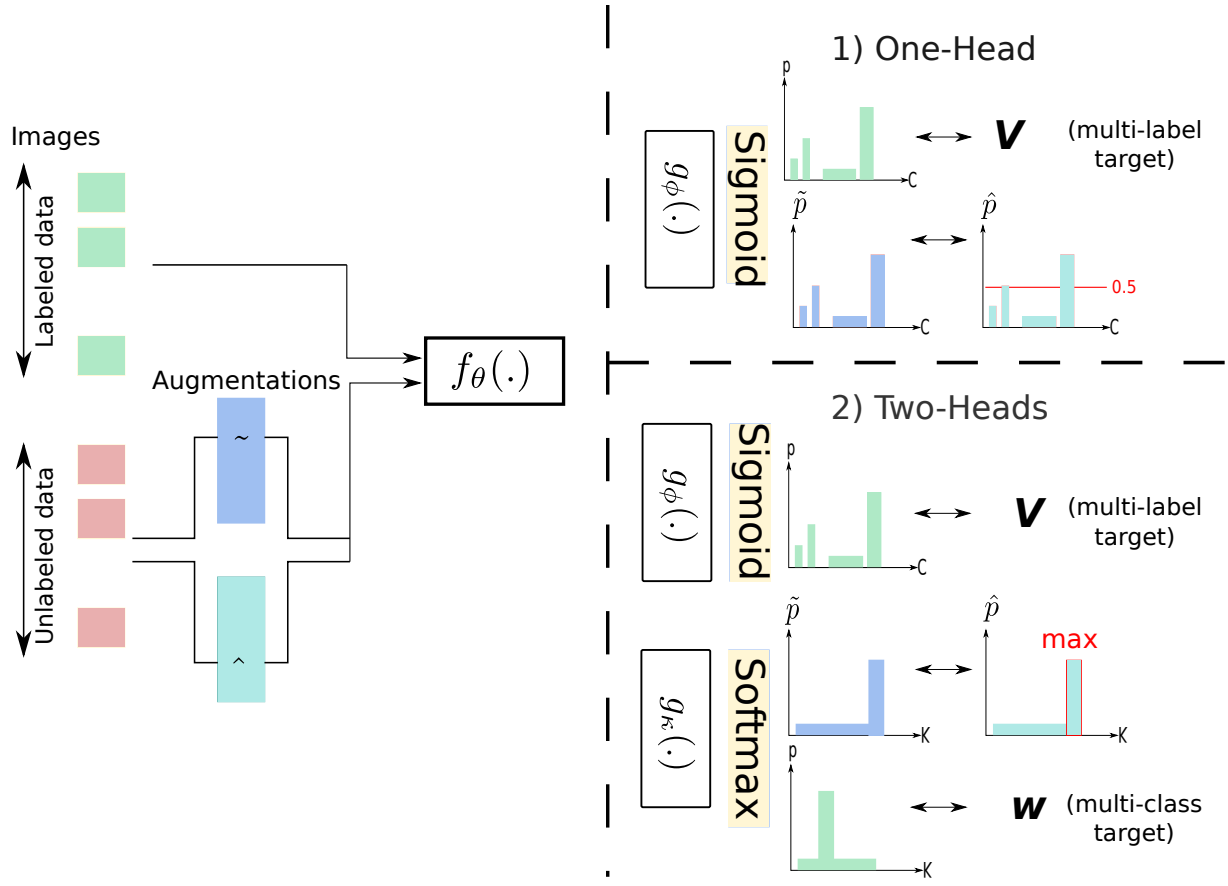

Fig 3 Semi-supervised learning: Two approaches to using FixMatch objective function in multi-label classification: i) Using a multi-label objective function for both pseudo-labels and true labels or ii) Using a single-label objective function for pseudo-labels and a multi-label objective function for true labels.

initialized by the weights learned during the deep clustering phase, thus unifying these two approaches. The key distinctions with the fine-tuning method presented in 2.2.4 are the classification head(s) and the objective function, which includes the unlabeled data during training as well.

Table 1 Abdominal organ classification results for the unsupervised model using deep clustering trained on 84967 unlabeled images. Performance is obtained on a test set of 101 abdominal exams ($n_{images} = 1242, n_{labels} = 1566$) with 5 trials:the average and best results are reported separately.

		Liver		Gallbladder		Other		Kidney		Pancreas		Spleen		Bladder		weighted avg	
		Best	Mean	Best	Mean	Best	Mean	Best	Mean	Best	Mean	Best	Mean	Best	Mean	Best	Mean
F1-score	LSD	72.08	**70.71**	50.0	47.44	65.23	**60.41**	71.5	**68.94**	65.08	**54.08**	69.8	59.35	74.75	**68.09**	66.95	**64.73**
	PICA	69.94	67.36	59.9	**50.05**	58.33	49.98	73.59	64.47	58.21	48.23	68.75	**62.74**	64.66	57.9	65.99	60.38
Precision	LSD	72.8	**67.66**	79.25	61.68	63.89	**60.05**	78.71	**64.04**	76.09	**58.03**	74.17	56.21	91.43	**75.42**	70.29	**63.82**
	PICA	70.51	60.47	77.78	**61.74**	56.76	51.18	69.27	59.04	62.9	40.49	72.73	**59.82**	90.62	69.97	63.82	57.78
Recall	LSD	81.8	74.7	54.78	**42.96**	71.43	**61.84**	86.08	**76.2**	68.06	54.72	71.11	65.19	82.76	**66.55**	73.24	**68.05**
	PICA	91.89	**77.77**	53.91	42.78	62.22	51.24	83.23	72.59	76.39	**66.39**	71.11	**66.67**	74.14	53.1	70.11	66.42
n_{labels}		555		115		315		316		72		135		58		1566	

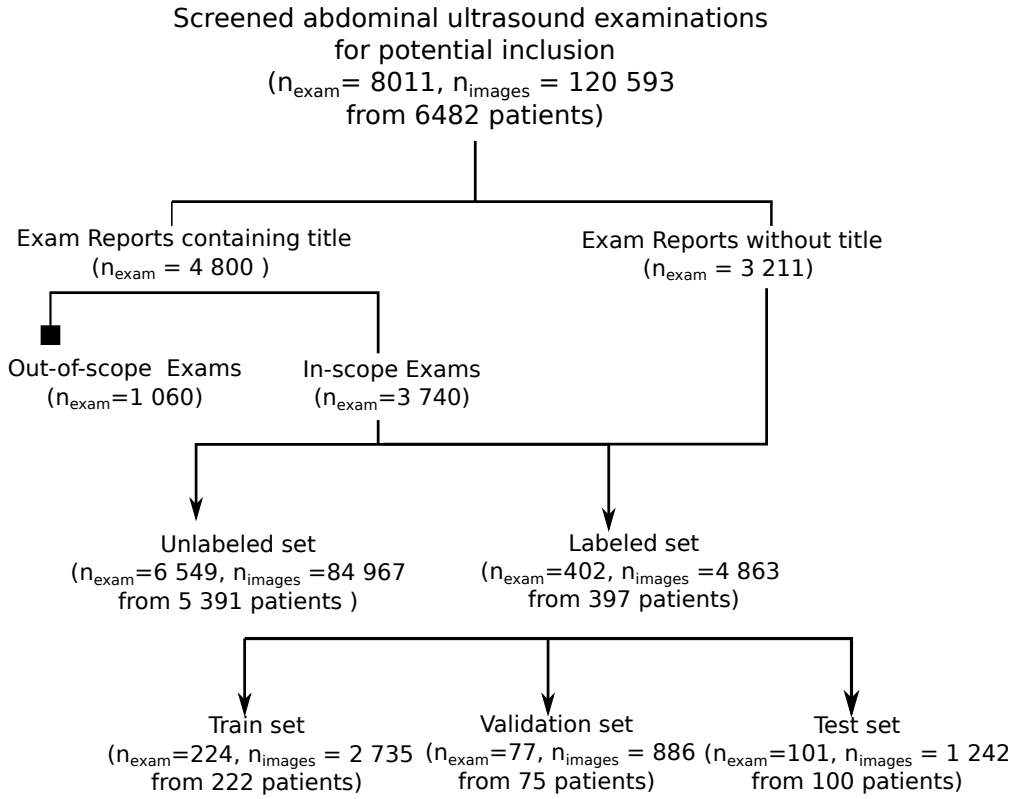

Fig 4 Flowchart describing the distribution of US examinations.

3 Experiments

3.1 Data set

International Review Board approval was obtained for this retrospective study, in collaboration with the Clinical Data Warehouse registered under the number

(no. IRB00011591). Multi-vendor data collected directly from the picture archiving and communication system (PACS) consisted of RGB freeze frames captured during ultrasound examination tagged as ”abdominal” along with a textual report written by a physician. In total, 8011 abdominal ultrasound examinations (120 593 images) from 6482 patients were extracted. Abdominal exams are performed to look for abnormalities in the abdomen, but also to determine if blood is circulating at a normal rate and level (i.e., Color Doppler Mode), or provide guidance during a biopsy procedure. In some cases, sus-pubic and endo-vaginal ultrasound examinations are performed in addition and combined with the abdominal examination. However, in this study, we are only interested in gray-scale ultrasound images (i.e., Brightness Mode) obtained with traditional ultrasound systems and focused on six anatomical structures, which would correspond to the following examination report titles: *abdominal, renal, bladder, liver*, or *urinary tract* ultrasound. For 60% of the reports, a title was available, which allowed us to exclude 1060 out-of-scope examinations that did not fit the five aforementioned titles (conforming to the medical experts' recommendations). Because this information was not always available, we decided to create an additional class *other* when none of the organs of interest were present in the image. The final database consisted of 6951 ultra-

sound examinations (89 830 images) from 5788 patients. 224 examinations (2735 images) from 222 patients were randomly selected for the labeled training set and 6549 examinations (84967 images) from 5391 patients for the unlabeled training set. 77 examinations (886 images) from 75 patients were randomly selected for the validation set and 101 examinations (1242 images) from 100 patients for the test set. The sets were constructed to ensure that there is no overlap of patients between sets. Data partition is summarized in Fig. 4. An adjudication panel was used as an external standard of reference. The panel consisted of four physicians, either radiologists or holders of a French national diploma in US imaging, and four sonographers holders of a French national diploma in US imaging from six different health institutions with more than three years of experience. The annotators, who worked on a tailor-made annotation platform, were asked to activate the tags corresponding to the organs present in each image. Each image was annotated once by one of the experts, while images for the test set were annotated twice.

3.2 Experimental settings

Training was conducted with four GeForce GTX 1080 Ti GPUs using Pytorch with a computation time ranging from few hours to 24 hours depending on the experiment. During training, and for all experiments, at each iteration we randomly

sample a subset t of exams, instead of sampling a set of images, to make sure that almost all the classes are represented in a batch since standard abdominal exams include views of all organs of interest. For each comparison, we use the same network architecture and training protocol, including the hyper parameters, optimizer, learning rate, number of epochs, data augmentation and data preprocessing. In particular, for all experiments a Resnet18 backbone is used, with a fixed learning rate of 0.0001, Adam optimizer, and unlabeled batch size of 16 examinations. For experiments with unsupervised training, fixed number of epochs (n = 100) is used and model weights of the last epoch are selected for evaluation. For experiments with supervised training, early stopping is performed if no improvement is observed on the validation set for seven epochs to avoid over-fitting. For experiments with semi-supervised training, a fixed amount of iteration steps (5000) is used. Finally for both supervised and semi-supervised training, model weights are selected at the epoch where the model performed best on the validation set. Performance is reported in terms of precision, recall and F1-score for each class and for the weighted average of all classes on the test set. To report these metrics, a class-specific threshold was set to select the probabilities output by the networks; the threshold was defined as the value that maximizes the F1 score of each class

during the validation step. Finally, to reflect the performance stability, each experiment was iterated four times, by initializing the random number generator (i.e seed value) differently.

3.3 Results

We present the results of the deep clustering method when no labeled data is available during training, and analyze how the added loss term (LSD in (6)) impacts the performance. In addition, we present the results of using unlabeled data in the presence of labeled data ranging from 275 to 2742 labeled examples.

3.3.1 Training with unlabeled data

Table. 1 compares the classification performance of the deep clustering models for all classes on the test set. Multi-labels were assigned to the clusters after training as presented in Algorithm 1. Without using any labeled data during training, deep clustering yields reasonable results for all organs. In particular, the proposed constraint on the cluster size distribution (LSD) outperforms on average the PICA baseline for almost all classes with an F1-score Weighted Average of 64.83% and 60.7% respectively with the default hyperparameter $\lambda = 2$ used in the original

paper to weight both terms of the objective function. Moreover, LSD seems to be more robust to the change of seeds, in fact the difference between the mean and the best results is almost insignificant in LSD, whereas in PICA, the seed has a greater impact on the performance. We further evaluate the sensitivity of both models to choices of λ by testing different values: $0.5, 1, 2, 5,$ and 10 with four different seeds. Fig. 5 displays the box plot of the F1-score values. We can see that PICA is very sensitive to the hyper-parameter λ, while LSD's performance is more stable, with a higher median.

3.3.2 Training with both labeled and unlabeled data

Performance is evaluated for several scenarios presented in Fig. 1, namely using unlabeled data simultaneously with labeled data during training in a semi-supervised manner, and/or during pre-training with deep clustering as a self-supervised objective function. All these scenarios are compared against to tradiotional transfer learning approach.

Semi-supervised learning:

Fig. 6 compares performances in terms of mean and 95% confidence intervals of F1-score weighted average for both semi-supervised methods presented in Fig.

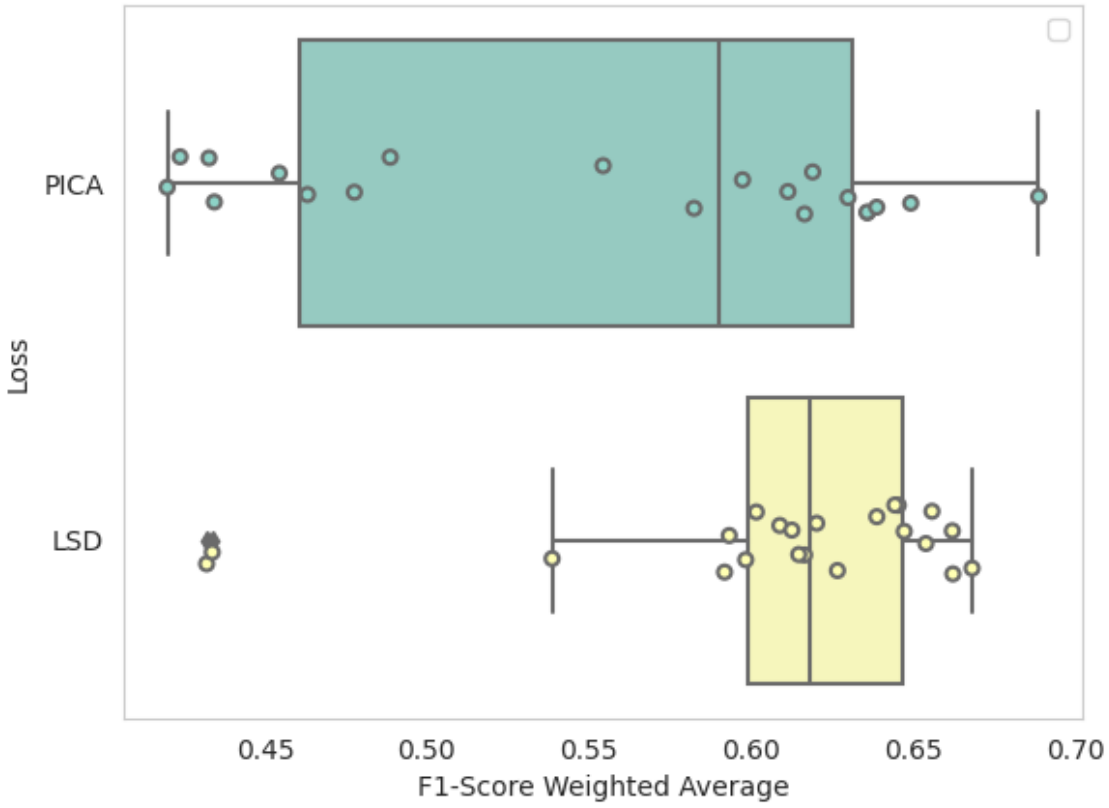

Fig 5 A box-plot showing F1-score weighted average values for each loss using five different values of λ over four experiments with different seeds.

3. We find that using the Two-Head model, rather than the One-Head model, yields better results regardless of the amount of annotated data used. One possible explanation is that by having only a multi-label objective function, classes with probability below 0.5 are automatically considered negative pseudo-labels, which propagates errors during training.

Deep Clustering as pre-training model:

We select the weights of the best performing deep clustering model (Section 2.2)

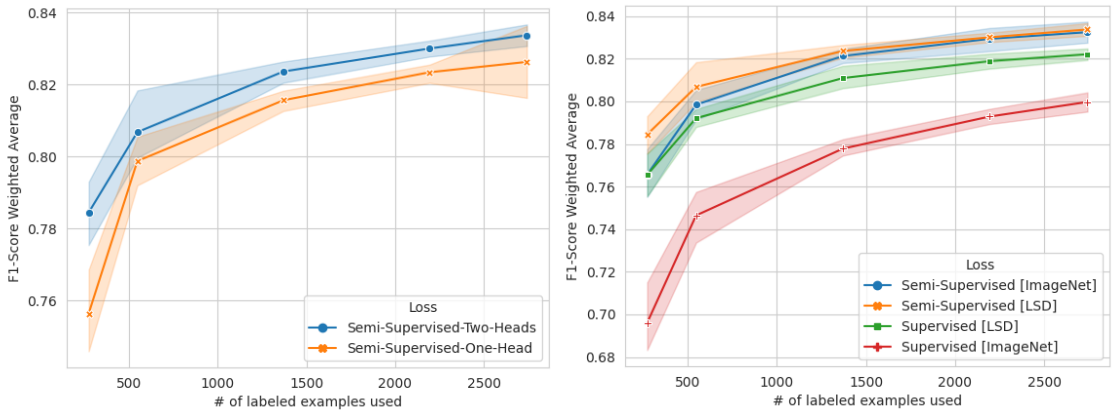

Fig 6 Mean and 95% confidence intervals of F1-score weighted average using all unlabeled images ($n_u = 84967$) with 10%, 20%, 50% and 100% of labeled images ($n_s = 2742$). On the left: One-Head vs Two Head semi-supervised learning models presented in Fig. 3. On the right: Semi-Supervised vs Supervised Learning models with different pretraining methods.

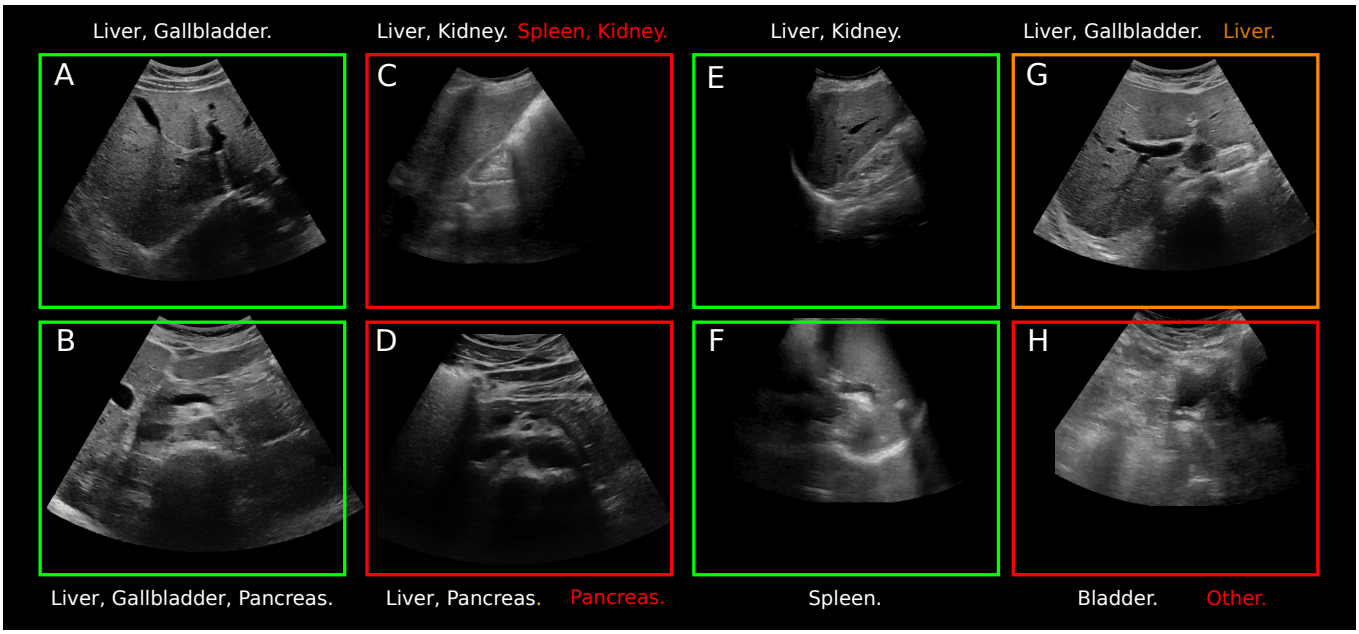

Fig 7 The images highlighted by a green and a red square represent successful and unsuccessful cases respectively. Labels in white represent the true classes assigned to the image, and yellow labels refer to the predicted classes when the classification is incomplete or (partially) incorrect. The image highlighted by an orange square represents a case where the "true" class *Gallbladder* assigned to the image is likely to be inaccurate.

to initialize our supervised (Section 2.2.4) and semi-supervised (Section 2.3.3) frameworks and compare the performance with ImageNet weight's initialization.

Fig. 6 shows these performances in terms of mean and 95% confidence intervals of F1-score weighted average. We observe that deep clustering pre-training outperforms ImageNet pre-training in all settings. An even more interesting result, is the performance difference using only 275 labeled example images (10% of the labeled set) with 69.2% vs 76.9% F1-score weighted average and 76.5% vs 79.3% in the supervised setting and semi-supervised setting respectively. Best performance is obtained with semi-supervised learning and 2742 labeled example images with an F1-score weighted average of 84.1% as shown in Table. 2. Fig. 7 showcases images classified by the best performing model with four examples of successful cases and four other cases where at least one label was misclassified or not detected. In particular, Picture (A) shows a *liver* and a *gallbladder*, (B) a *liver*, a *gallbladder* and a *pancreas*, (E) a *liver* and a *kidney*, and (F) a *spleen*. The image highlighted by an orange square (G) represents a case where the class *Gallbladder* assigned to the image is likely to be inaccurate after consulting with several experts, and where only the *liver* is visible. Images highlighted by a red square represent cases where at least one label was misclassified or not detected. Picture (D) shows a *liver* and a *pancreas*, (C) shows a *kidney* and *liver*, here the model correctly classified the *kidney* but confused the *liver* with the *spleen*, and

Table 2 Abdominal organ classification results for the best performing model:A two-head semi-supervised learning model pre-trained with deep clustering and trained using 2742 labelled examples images. Performance is obtained on a test set of 101 abdominal exams ($n_{images} = 1242, n_{labels} = 1566$).

	Liver	**Gallbladder**	**Other**	**Kidney**	**Pancreas**	**Spleen**	**Bladder**	**weighted avg**
F1-score	0.89	0.78	0.83	0.87	0.66	0.75	0.89	0.84
Precision	0.87	0.90	0.86	0.82	0.55	0.76	0.98	0.84
Recall	0.90	0.70	0.80	0.93	0.81	0.74	0.81	0.85
n_{labels}	555	115	315	316	72	135	58	1566

finally (H) the model classified the image as *other* instead of *bladder*.

4 Discussion

In this work, we analyze the benefits of using unlabeled ultrasound data when the available labeled dataset is limited, in contrast to related studies using transfer learning where the model is retrained, after being initially trained to classify color photographs on ImageNet. We observed that deep clustering can be used as an unsupervised model for the classification of abdominal organs with reasonable performance. We further improved the performance by taking into account the possible imbalance of classes in a given batch, using a symmetric Dirichlet prior, which incidentally makes the method less sensitive to the choice of the hyperparameter λ. In addition, an extensive study was conducted to analyze how large amounts of unlabeled data could be used to improve the performance of a model trained on few labeled data and to determine the extent to which the size of the labeled data set impacts performance.

Of the three questions stated in the introduction, we addressed the first one (Fig. 1: Strategy 2) by using the weights of the deep clustering model as initialization to a supervised multi-label classification model, and showed that the features learned in this self-supervised manner were more useful for downstream tasks on the same domain compared to features learned on ImageNet, regardless of the amount of labeled data used. Regarding the second question (Fig. 1: Strategy 3 vs 2), we showed that the performance obtained by leveraging labeled and unlabeled data simultaneously during training was slightly better than that obtained by using them consecutively as proposed by the self-supervised methods. As for the third question (Fig. 1: Strategy 4) we showed that there is indeed a relevance in combining the two paradigms through deep clustering when the amount of labeled data is extremely limited. In other words, the gap between deep clustering and ImageNet pre-training in the semi-supervised setting decreases as the amount of labeled data increases, suggesting that depending on the amount of labeled data available, ImageNet pre-training in the semi-supervised setting may suffice.

Our study has several limitations: first to be able to use the deep clustering method, multi-label classification was transformed to a single-label classification. Doing so, the method cannot benefit from the potential relationships between labels as

they are considered independent. Second, by adding a Dirichlet prior on the cluster size distribution, an additional hyperparameter is introduced which in our opinion should remain fixed but further investigation on the values of this hyperparameter is needed. The effect of self-supervised and semi-supervised learning for the task of organ classification in abdominal ultrasound images was evaluated on a single data set using only two methods. Thus exploration of other methods on multiple data sets is warranted. Finally, an in-depth investigation on the effect of class imbalance for semi-supervised learning is needed, the proposed method could benefit from taking into account the imbalance in the pseudo-labels generated during the training.

5 Conclusion

In summary, we provided a framework for the classification of abdominal organs in a large-scale multi-vendor ultrasound database. Specifically, and in contrast to the aforementioned related studies, we use an unrefined database with a very limited number of labeled examples. Both self-supervised and semi-supervised learning methods were explored to leverage the unlabeled data. In addition, we adapted these methods to a multi-label framework that, to our knowledge, has not been

addressed before. Classification of abdominal organs in large ultrasound databases is an important step for future work on US based diagnosis. Indeed, several steps are necessary to build and pre-process the database before training a model for such a task.[24] First the abdominal exams are selected, then amongst the images of the exam, the ones containing the organ(s) of interest are manually selected by a trained operator. All (or part) of the selected images are further annotated by a panel of experts according to the predefined task (e.g., detection of abnormalities in the kidney), and finally used to train a machine learning model. Our study can help to automate some of these processes at minimal cost. In general, US studies have the potential to speed up the democratization of access to medical imaging in developing countries where healthcare providers consider lack of training to be the main limitation to the use of US.[25]

Disclosures

Conflicts of interest should be declared under a separate header. If the authors have no relevant financial interests in the manuscript and no other potential conflicts of interest to disclose, a statement to this effect should also be included in the manuscript.

Acknowledgments

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur. We thank Deepomatic for making available an annotation platform to the NHance project. We thank Remi Rousseau, MS, Guillaume Oules, MS, and Alexandre Dubreucq, MS, both from Ecole Polytechnique, for their help in the design of the NHance project, and Salma Eloualydy, MS in the help provided to construct our database. The authors are grateful to the radiology team from Necker Hospital, and particularly Jean-Michel Correas MD PhD, Anne-Marie Tissier MD, Sylvain Bodard MD. We thank the French Health Data Hub for providing resources and support. The authors are grateful for the help provided by the team of the radiology department at Saint Louis Hospital: Claudine Singh, Fanny Joujou, Kemel Khezzane, Constance De Margerie MD, PhD. We thank Olivier Lucidarme MD, PhD and Charles Doulin from the radiology department at Pitié-Salpêtrière Hospital. The authors greatly appreciated the help that Mariama Bah, MD, and Gregory Khelifi, MD, Albane De Kerautem, MD, provided. This work was supported by the clinical research

unit of Saint Louis Hospital: Jerome Lambert, MD, PhD, and Claire Montlahuc, MD, PhD. Ultrasound Images selected for the figures were processed one by one to remove any potentially re-identifying information.

Data, Materials, and Code Availability

The code will be made available upon publication.

References

1 P. M. Cheng and H. S. Malhi, "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," *Journal of digital imaging* **30**(2), 234–243 (2017).

2 Z. Xu, Y. Huo, J. Park, *et al.*, "Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 711–719, Springer (2018).

3 K. Li, Y. Xu, Z. Zhao, *et al.*, "Automatic recognition of abdominal organs in ultrasound images based on deep neural networks and k-nearest-neighbor classification," in *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 1980–1985, IEEE (2021).

4 L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence* **43**(11), 4037–4058 (2020).

5 D. Pathak, P. Krahenbuhl, J. Donahue, *et al.*, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536–2544 (2016).

6 Z. Ren and Y. J. Lee, "Cross-domain self-supervised multi-task feature learning using synthetic imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 762–771 (2018).

7 B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *Advances in Neural Information Processing Systems* **31** (2018).

8 E. Bonmati, Y. Hu, A. Grimwood, *et al.*, "Voice-assisted image labelling for endoscopic ultrasound classification using neural networks.," *IEEE Transactions on Medical Imaging* (2021).

9 A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906* (2021).

10 M. Caron, P. Bojanowski, A. Joulin, *et al.*, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, 132–149 (2018).

11 J. Jiao, Y. Cai, M. Alsharid, *et al.*, "Self-supervised contrastive video-speech representation learning for ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 534–543, Springer (2020).

12 M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *Advances in neural information processing systems* **29** (2016).

13 S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242* (2016).

14 G. J. McLachlan, "Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis," *Journal of the American Statistical Association* **70**(350), 365–369 (1975).

15 D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, **3**(2), 896 (2013).

16 J.-C. Su, Z. Cheng, and S. Maji, "A realistic evaluation of semi-supervised learning for fine-grained classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12966–12975 (2021).

17 O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks* **20**(3), 542–542 (2009).

18 J. Tan, A. Au, Q. Meng, *et al.*, "Semi-supervised learning of fetal anatomy from ultrasound," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, 157–164, Springer (2019).

19 J. Huang, S. Gong, and X. Zhu, "Deep semantic clustering by partition confidence maximisation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8849–8858 (2020).

20 W. Van Gansbeke, S. Vandenhende, S. Georgoulis, *et al.*, "Scan: Learning to classify images without labels," in *European Conference on Computer Vision*, 268–285, Springer (2020).

21 T. Kart, W. Bai, B. Glocker, *et al.*, "Deepmcat: Large-scale deep clustering

for medical image categorization," in *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*, 259–267, Springer (2021).

22 K. Sohn, D. Berthelot, N. Carlini, *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems* **33**, 596–608 (2020).

23 M. N. Rizve, K. Duarte, Y. S. Rawat, *et al.*, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," *arXiv preprint arXiv:2101.06329* (2021).

24 H. Dadoun, A.-L. Rousseau, E. de Kerviler, *et al.*, "Detection, localization, and characterization of focal liver lesions in abdominal us with deep learning," *Radiology: Artificial Intelligence* (2022).

25 S. Shah, B. A. Bellows, A. A. Adedipe, *et al.*, "Perceived barriers in the use of ultrasound in developing countries," *Critical ultrasound journal* **7**(1), 1–5 (2015).

List of Figures

3 Semi-supervised learning: Two approaches to using FixMatch objective function in multi-label classification: i) Using a multi-label objective function for both pseudo-labels and true labels or ii) Using a single-label objective function for pseudo-labels and a multi-label objective function for true labels.

4 Flowchart describing the distribution of US examinations.

5 A box-plot showing F1-score weighted average values for each loss using five different values of λ over four experiments with different seeds.

6 Mean and 95% confidence intervals of F1-score weighted average using all unlabeled images ($n_u = 84967$) with 10%, 20%, 50% and 100% of labeled images ($n_s = 2742$). On the left: One-Head vs Two Head semi-supervised learning models presented in Fig. 3. On the right: Semi-Supervised vs Supervised Learning models with different pretraining methods.

List of Tables

2 Abdominal organ classification results for the best performing model:A two-head semi-supervised learning model pre-trained with deep clustering and trained using 2742 labelled examples images. Performance is obtained on a test set of 101 abdominal exams ($n_{images} = 1242, n_{labels} = 1566$).