Revisiting Vicinal Risk Minimization for Partially Supervised Multi-Label Classification Under Data Scarcity

Nanqing Dong*
Department of Computer Science
University of Oxford
nanqing.dong@cs.ox.ac.uk

Jiayi Wang
Mathematical Institute
University of Oxford

Irina Voiculescu
Department of Computer Science
University of Oxford
irina.voiculescu@cs.ox.ac.uk

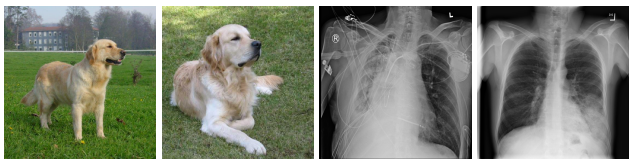
Abstract

Due to the high human cost of annotation, it is non-trivial to curate a large-scale medical dataset that is fully labeled for all classes of interest. Instead, it would be convenient to collect multiple small partially labeled datasets from different matching sources, where the medical images may have only been annotated for a subset of classes of interest. This paper offers an empirical understanding of an under-explored problem, namely partially supervised multi-label classification (PSMLC), where a multi-label classifier is trained with only partially labeled medical images. In contrast to the fully supervised counterpart, the partial supervision caused by medical data scarcity has non-trivial negative impacts on the model performance. A potential remedy could be augmenting the partial labels. Though vicinal risk minimization (VRM) has been a promising solution to improve the generalization ability of the model, its application to PSMLC remains an open question. To bridge the methodological gap, we provide the first VRM-based solution to PSMLC. The empirical results also provide insights into future research directions on partially supervised learning under data scarcity.

1. Introduction

Fueled by the joint development of theories [5, 16, 28] and hardware, deep learning has led to a significant leap in computer-aided diagnosis, reaching or even outperforming human-level performance [27]. As a data-driven method, DL models tend to require large-scale fully labeled images for supervised training. However, this is largely infeasible for many medical vision tasks due to high annotation costs, which gives rise to emerging research interests on partially supervised learning (PSL) [9, 12, 14, 15, 25, 29, 34, 37, 38].

*Corresponding Author



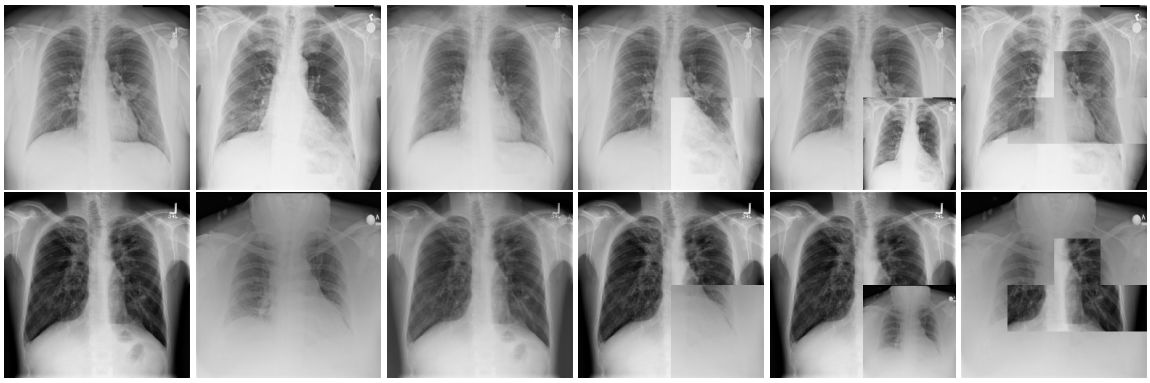
(a) (b) (c) (d)

Figure 1. (a) and (b) are two golden retrievers from ImageNet [7]. (c) and (d) are two chest X-ray (CXR) images from ChestX-ray14 [33]. In contrast to object-centric images, CXR images have multiple objects of semantic interests (*e.g.* organs).

Given a set of classes of interest, it is challenging to prepare a large dataset with all classes of interested annotated. Instead, it is more practical to source multiple relevant but partially labeled datasets, where each dataset is only annotated for a *true* subset of classes of interests. This can be interpreted from the perspective of multi-task learning (MTL) [2], where the task of interest can be decomposed into multiple sub-tasks.

Existing PSL studies [12, 14, 15, 25, 29, 34, 37, 38] tend to assume that large-scale partially labeled or even fully labeled data are available when designing the algorithms. However, this is infeasible in many specific scenarios, especially in the medical domain, where *data scarcity* has been a topic of active research. Dong *et al*. [9] first proposed to use data augmentation to mitigate the data scarcity in partially supervised semantic segmentation, where vicinal risk minimization (VRM) [3] is adopted to generate *vicinal* fully labeled image-label pairs with only partially labeled data. Though how to address the data scarcity issue of partially supervised multi-label classification (PSMLC) remains an open question, inspired by [9], we make a concrete first step towards it with VRM.

Various VRM-based data augmentation techniques [19, 20, 26, 31, 35, 36] have been designed to improve the generalization ability of a standard multi-class classifier trained



(a) Input 1 (b) Input 2 (c) MixUp [36] (d) CutMix [35] (e) ResizeMix [26] (f) PuzzleMix [20]

Figure 2. Illustration of state-of-the-art VRM methods on CXR images. First Row: Two CXR images are visually similar. Second Row: Two CXR images are visually different. The mixing ratio λ is 0.75 for all four methods displayed. Firstly, compared with MixUp [36], CutMix [35], ResizeMix [26], and PuzzleMix [20] generate less perceptually comfortable vicinal images (*i.e*. they do not look like real CXR images). Secondly, CutMix, ResizeMix, and PuzzleMix ignore the human structure similarity. Last but not least, in an MLC problem scenario, the vicinal images generated by CutMix, ResizeMix, and PuzzleMix might discard potential region of interests. For example, if the diseased region for the first input lies in the bottom right part of the CXR image, this might not be reflected in the vicinal images. In this work, we will focus on MixUp.

on general object-centric images (*e.g*. Fig. 1). The first challenge is to generate vicinal images. Though many of these methods [20, 26, 35] have reported state-of-the-art performance in multi-class classification tasks on general images, they tend to generate less meaningful vicinal images than MixUp [36] in the medical domain. This phenomenon is due to the fact that these methods are designed for object-centric images where the mixing process can potentially keep the semantic information of interest. As illustrated in Fig. 2, the vicinal images generated by CutMix [35], ResizeMix [26], and PuzzleMix [20] can not utilize *human structure similarity* [8–10] or preserve the region of interests (*e.g*. the infectious regions) for medical images with multiple objects. As a comparison, MixUp shows robust visual performance over the other variants. Thus, we use MixUp to generate vicinal images in this work. The second challenge is to generate vicinal labels. The existing VRM methods are only designed for fully labeled data. Due to partial supervision, there are missing labels. That is to say, while we can randomly sample two images to generate a vicinal image with MixUp, we can not define the corresponding vicinal label with missing labels, which is deemed as the major bottleneck of applying VRM to PSMLC.

While previous efforts [12] have been made to understand PSMLC with large-scale benchmark datasets from the perspective of label propagation [39], the problem formulation of data scarcity in this study differentiates our contributions from [12]. According to [9], VRM has shown unparalleled robustness in partially supervised semantic segmentation with only small-scale data. Motivated by this, we aim to leverage VRM to tackle PSMLC with data scarcity. Inspired by the principle of maximum entropy (PME) [18], we propose a simple yet robust MixUp-based method for PSMLC, which can efficiently improve the model performance with only access to partial labels. We evaluate the proposed method via a set of controllable experiments on ChestX-ray14 [33], a public multi-label CXR dataset of thoracic conditions. In addition to providing initial empirical insights into PSMLC, we also validate that self-supervised pre-training [4] can further improve the model performance together with the proposed method.

Our main contributions can be summarized as follows:

1. This is the first study of partially supervised multi-label classification (PSMLC) under data scarcity in the medical domain.

2. We adapt VRM to PSMLC and propose a simple yet robust PME-based technique to improve the model performance with only scarce partially labeled medical images.

3. The experimental results show initial empirical insights into future research directions on PSL under data scarcity.

2. Related Work

Partially Supervised Learning. Apart from [12], which provides an empirical understanding of PSMLC with large-scale general images (*e.g*. PASCAL VOC [13] and MS COCO [22], the major breakthroughs in PSL lie in the field of multi-organ or multi-structure segmentation on medical images [9, 14, 15, 25, 29, 34, 37, 38]. Gonzalez *et al*. [15] first

showed that, by only backpropagating the cross-entropy from the labeled part, the performance of multi-structure segmentation with sufficiently large-scale partially labeled data could be competitive with the performance of the fully supervised counterpart. However, in practice, it is unlikely to collect or curate a large amount of partially labeled data, *i.e*. only small-scale partially labeled datasets are available. Zhou *et al*. [38] proposed to utilize a fully labeled set to learn image priors. However, the fully labeled set might be difficult to acquire in practice. Shi *et al*. [29] proposed an *exclusion loss* for mutually exclusive classes, which can not be used in MLC. Closely related to our work, Dong *et al*. [9] proposed to mitigate the data scarcity based on VRM. In contrast to this work, [9] requires that the classes of interest to be mutually exclusive, while our work is the first study on PSMLC.

Vicinal Risk Minimization. Zhang *et al*. [36] first proposed MixUp, a VRM-based data augmentation method in the input space by mixing two random images and the corresponding labels by convex interpolation. This idea has been further extended by many variants. For example, CutMix [35] (Fig. 2d) replaces a random patch of the first image with a patch from the second image cropped at the same location. ResizeMix [26] (Fig. 2e) takes one step further by replacing a random patch of the first image with the resized second image. PuzzleMix [20] (Fig. 2f) could be viewed as a generalization of CutMix, where the locations of the random patches are determined by saliency. As discussed in Sec. 1, these variants of MixUp are not suitable for MLC or PSMLC tasks.

3. Preliminaries

MixUp. Let (x_i, y_i) and (x_j, y_j) be two image-label pairs randomly sampled from the training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{n}$. Here, $x \in \mathbb{R}^{H\times W\times C}$ is an image and $y \in \mathbb{R}^{K}$ is considered as a one-hot encoded binary vector, where there are K (mutually exclusive) classes of interest. The vicinal image-label pair $(\tilde{x}, \tilde{y})$ is defined as

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda)x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j \end{aligned} \tag{1}$$

where $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$ for $\alpha \in (0, \infty)$. From the perspective of MTL, a K-class MLC task could be decomposed into K binary classification tasks, where MixUp can be applied.

Weighted Loss. To combat class imbalance, the weighted binary cross-entropy (BCE) loss is commonly adopted in MLC. Given an image-label pair (x, y), we use y^i to denote the i^{th} entry of y. We have

$$\begin{aligned} \mathcal{L}(x, y^i) = &- w_+ y^i \log p(y^i = 1|x) \\ &- w_-(1-y^i)\log p(y^i = 0|x), \end{aligned} \tag{2}$$

where $w_+ = \frac{n_-}{n_+ + n_-}$ and $w_- = \frac{n_+}{n_+ + n_-}$ with n_+ and n_- the number of positive and negative cases for the i^{th} class respectively.

4. Partially Supervised Multi-Label Classification

4.1. Problem Formulation

Without loss of generality, given a *small-scale* training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{n}$, we define that, for each image-label pair (x, y), at least one entry of y is missing, *i.e*. x is partially labeled, and for each class of interest, there are both positive and negative cases labeled. Given a model f_θ, the goal of PSMLC is to maximize the prediction accuracy of f_θ with only *limited* partially labeled data.

4.2. MixUp with Partial Labels

Now, we will explore how to adapt MixUp to PSMLC. For simplicity, we illustrate with an example of K=2. Again, (x_i, y_i) and (x_j, y_j) are two image-label pairs. where y_i and y_j are partial labels. There are two cases.

4.2.1 Locally Full Supervision

In the first case, y_i and y_j have partial labels for the same class, *e.g*. $y_i = [?, y_i^2]$ and $y_j = [?, y_j^2]$, where ? denotes the missing label for the 1^{st} class and $y_i^2 \in \{0, 1\}, y_j^2 \in \{0, 1\}$. Obviously, Eq. (1) still holds if we only consider the 2^{nd} class. Under the interpretation of MTL, we have *locally* full supervision over the 2^{nd} class. However, to leverage MixUp in this case, there will be additional computational cost in batch-wise sampling, which could be a non-trivial overhead in practice. A trivial solution is to decompose a PSMLC problem into multiple binary classification problems. However, when K is large, this strategy will be inefficient as it does not utilize MTL.

4.2.2 Globally Partial Supervision

We are more interested in the second case, where y_i and y_j have partial labels for different classes, *e.g*. $y_i = [?, y_i^2]$ and $y_j = [y_j^1, ?]$. In fact, in a partially labeled dataset, the majority of randomly sampled pairs will fall in this case, which is also the major bottleneck of PSMLC. Similar to [9], we aim to transform PSMLC into fully supervised MLC. However, [9] defines the *vicinity distribution* [3] by utilizing human structure similarity, which is infeasible for PSMLC. Instead, we regularize the vicinity distribution by a simple probability trick.

The principle of maximum entropy (PME) [18] was first proposed in 1950s. With limited prior knowledge over the unknown true distribution, PME defines the vicinity distribution with the largest entropy. Concretely, for the k^{th}

Figure 3. Illustration of MixUp based on PME with two CXR images with 3 classes of interest. For full labels, Eq. (1) can be directly applied. For partial labels, Eq. (3) is only applied for the classes where the first image has no missing labels.

class, assume only x_i has the partial label (*e.g.* $y_i^k = 1$) and x_j has missing label ($y_j^k \leftarrow ?$), without any prior knowledge about x_j, we define $\hat{y}_j^k = 0.5$, *i.e.* considering y_j^k as an independent system with two possible states, discrete uniform distribution leads to the largest entropy. We define the vicinal label as

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda) x_j \\ \tilde{y}^k &= \lambda y_i^k + (1-\lambda)\hat{y}_j^k \end{aligned} \tag{3}$$

where $\lambda \sim \text{Uniform}(\alpha_k, 1)$ and $0.5 \leq \alpha_k < 1$ is a hyperparameter. Note, we use α_k instead of α as the choice of α_k can be dependent on the k^{th} class. This is different from MixUp-based methods [19, 20, 26, 31, 35, 36] that are designed for multi-class classification tasks. We will give more details of this design in Sec. 5.2. Though Eq. (3) has a similar format to Eq. (1), we want to highlight a few difference: (a) x_i should have locally full supervision for the k^{th} class (*e.g.* y_i^k=1); (b) λ is required to be larger than 0.5 for the known example (x_i in this case). As we are certain about y_i^k, there are only two possible states for y_j^k, *i.e.* y_j^k=1 or y_j^k=0. It can be inferred that if y_i^k=1, we have $0.75 < \tilde{y}^k = 0.5 + 0.5\lambda < 1$; if $y_i^k = 0$, we have $\tilde{y}^k = 0.5(1-\lambda) < 0.25$. Clauses (a) and (b) ensure that the generated vicinal label $\tilde{y}^k$ has a logically reasonable label distribution. An illustrative comparison between MixUp and the proposed method is presented in Fig. 3.

Relationship with Noisy Labels Learning from noisy labels [23] has been a seminal strategy in utilizing unlabeled data over the past decade [39]. Here, we define the noisy labels as the pseudo labels that are automatically generated by the algorithms for the unlabeled data. This strategy has shown state-of-the-art performance in semi-supervised learning benchmark tasks [30]. The problem of interest shares a similar problem formulation with semi-supervised learning, where unlabeled data exist. However, the proposed method differs from semi-supervised learning in three aspects. First, instead of learning the pseudo labels, we generate the pseudo labels by PME. That means, the proposed method is computationally efficient. Second, essentially, the proposed method is a data augmentation technique, which simultaneously augments the base image (x_i in Eq. (3)) and adds noise to the base label (y_i^k in Eq. (3)). In fact, adding noise to the ground truth labels has been shown as an effective method to improve the model robustness and generalization ability [32]. Last but not least, the generated pseudo labels ($\tilde{y}^k$) are logically reliable. As shown in [9], semi-supervised learning-based noisy labels might not be reliable if there are not enough labeled data to learn the model. In the problem of formulation, we focus on a data scarcity situation. Thus, a semi-supervised learning approach is not feasible.

4.3. Training Strategy

For fully labeled images, the linear combination in Eq. (1) can be performed efficiently with batch-wise processing, as shown in [36]. However, it requires additional computational cost to sample two image-label pairs. First, we have to decompose the task into K binary classification sub-tasks, which gives up the formulation of MTL and increases the number of forward and backward passes in the optimization process. Second, when sampling a pair, we need to make sure that two images are partially labeled for the same class of interest. Third, when K is large, two image-label pairs can have both locally full supervision and globally partial supervision. To fully utilize the advantages of random sampling in stochastic gradient descent, we need to efficiently implement Eq. (1) and Eq. (3) simultaneously in the batch processing.

To solve the above issues, we present a computation-efficient implementation for batch-wise training with only 8 lines of PyTorch code, shown in Listing 1.[1]

```
# x1, x2: batch of images, [B, 3, H, W]
# y1, y2: batch partial labels where the missing
#         entries are filled with 0.5, [B, K]
# alpha: hyperparameter
# model: neural network
# criterion: loss function with reduction='none'
lam = numpy.random.uniform(alpha, 1)
mask = y1 != 0.5
x = lam * x1 + (1. - lam) * x2
y = lam * y1 + (1. - lam) * y2
loss = (criterion(model(x), y) * mask).mean()
optimizer.zero_grad()
loss.backward()
optimizer.step()
```

Listing 1. Batch-wise training in PyTorch.

We also provide a concrete example for a better illustration of the mechanism behind our implementation. Let

[1]For simplicity, we use the same value of `alpha` for different classes to illustrate the main concept of the proposed method.

$y_1 = [1, ?, 0, ?]$ and $y_2 = [1, 1, ?, ?]$ be two label vectors with missing labels (*i.e.* $K = 4$). Given an arbitrary $\lambda \sim \text{Uniform}(0.5, 1)$ (say $\lambda = 0.75$), a step-by-step demonstration of Listing 1 is provided in Tab. 1.

Step	y_1	y_2	mask
–	[1, ?, 0, ?]	[1, 1, ?, ?]	–
Fill ? with 0.5	[1, 0.5, 0, 0.5]	[1, 1, 0.5, 0.5]	–
Get mask	[1, 0.5, 0, 0.5]	[1, 1, 0.5, 0.5]	[1, 0, 1, 0]
Get $\tilde{y}$	[1, 0.625, 0.25, 0.5]		[1, 0, 1, 0]

Table 1. A step-by-step demonstration of Listing 1. After getting $\tilde{y}$ with Eq. (1) (or Eq. (3)) with $\lambda = 0.75$, the BCE losses for $K = 4$ classes are computed following Eq. (2). However, only the BCE losses of the first and the third classes will be back-propagated as the BCE losses of the second and the fourth classes are zeroed out.

5. Experiments

5.1. Experimental Setup

Baselines. We consider four models in our experiments, where the four models share the same network backbone. The first one is a standard MLC model where the missing labels are ignored in the backpropagation [15]. We denote this model as *vanilla*. For the second model, we apply MixUp in a locally full supervision fashion to train an MLC model, where Eq. (1) is applied as described in Sec. 4.2.1. Following [36], we set $\alpha = 1$. We use the default valThe second baseline is denoted as MixUp. The third model is the proposed PME-based MixUp variant. Note, the third model is a unification of locally full supervision and globally partial supervision, where Eq. (1) and Eq. (3) are efficiently integrated (as shown in Sec. 4.3). For simplicity, we denote the proposed method as MixUp-PME. To understand the impact of PSMLC under data scarcity, we present the fourth model, which is the same MLC model trained with full labels. We denote this model as *Oracle*.

Data. We use the ChestX-ray14 [33] public multi-label dataset of thoracic conditions,[2] and adopt its default batch splits to ensure reproducibility. We use the first 1000 CXR images of the first batch as the training set and the second 1000 CXR images of the first batch as the test set. For simplicity, we illustrate the proposed method with a simple case that each image is only labeled for one class. We generate the partially labeled datasets by choosing 4 most common diseases among 14 identified conditions (*i.e.* $K = 4$), which are *infiltration*, *effusion*, *atelectasis*, and *nodule*.

[2] https://nihcc.app.box.com/v/ChestXray-NIHCC

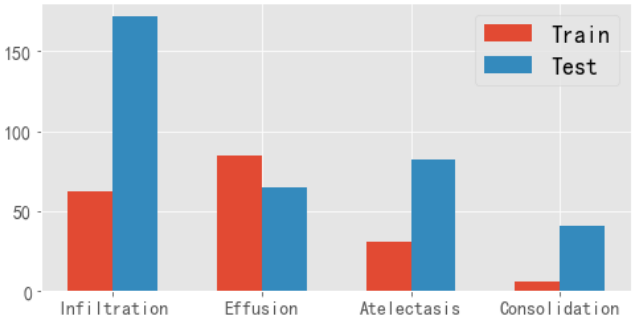


Figure 4. Label statistics of positive cases for the partially labeled training set and fully labeled test set.

Implementation. Following the setup of [27], we use DenseNet121 [17] as the network backbone. We minimize the weighted loss (Eq. (2)) by using a standard Adam optimizer [21] with fixed learning rate 10^{-3} and batch size 64. In the inference phase, we use 0.5 as the default threshold for the predicted probability score. We train all the baselines for 30 epochs and report the best mean F1-score for each baseline, where

$$F_1 = 2\,\frac{precision * recall}{precision + recall}, \tag{4}$$

i.e. the harmonic mean of precision and recall. We run each experiments for three times with different random seeds. All experiments are conducted in PyTorch [24] on an NVIDIA Tesla V100. Note, the traditional data augmentation that manipulates the image space does not directly solve the partial supervision problem [9]. Thus, for a fair comparison, we do not involve any traditional data augmentation. All CXR images are resized to a fixed size of 224×224. As a pre-processing step, instance normalization [6] is performed on each CXR image:

$$\hat{x}^{ij} = \frac{x^{ij} - \mu(x)}{\sigma(x)}, \tag{5}$$

where x is an image, $\hat{x}$ is the normalized image, (i, j) is the position of the pixel, and μ and σ are the mean and standard deviation of the pixels of x.

5.2. Empirical Analysis

To provide a comprehensive understanding of the problem of interest, we consider a simple situation in the first experiment. The training set is equally split into 4 subsets, where each subset only contains labels for one class. The label distributions of the partially labeled training set and the fully labeled test set are summarized in Fig. 4. In this experiment, we use the same value of α for different classes. The numerical results for four models are presented in Tab. 2, where we report the mean F1-score for four classes. For

Model	*Infiltration*	*Effusion*	*Atelectasis*	*Nodule*	Average
vanilla	0.3031	0.1807	0.1807	0.0704	0.1837
MixUp	0.2959	0.1701	0.1682	0.0707	0.1762
MixUp-PME	0.2949	0.1327	0.1584	**0.2151**	**0.2002**
AMP	0.3034	**0.1894**	**0.1889**	0.0788	0.1901
Oracle	**0.3547**	0.1710	0.1739	0.0699	0.1924

Table 2. Performance comparison for PSMLC. In the training set, each CXR image is only partially labeled for one class.

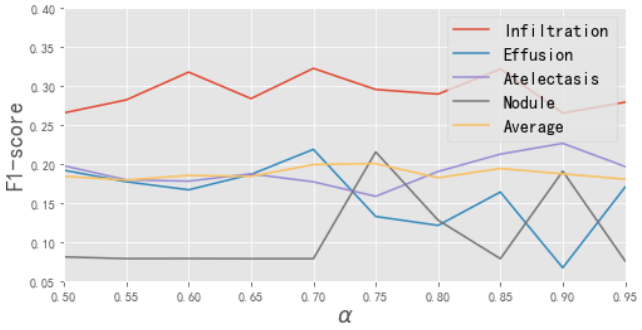


Figure 5. Sensitivity of MixUp-PME to α for different classes.

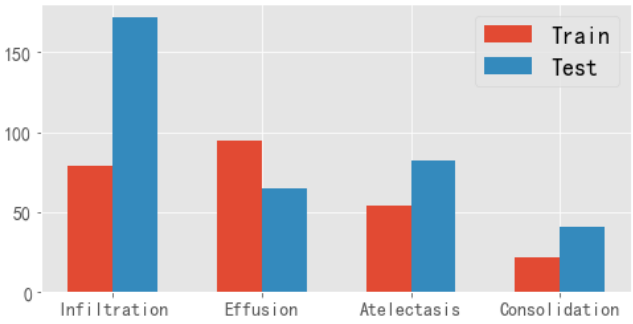


Figure 6. Label statistics of positive cases for the partially labeled training set and fully labeled test set.

MixUp-PME, we report the performance of $\alpha = 0.75$ in Tab. 2, which gives the highest mean F1-score. The complete results of MixUp-PME under different values of α are depicted in Fig. 5. Based on Tab. 2 and Fig. 5, there are four empirical findings. First, an MLC model trained with full labels might not outperform the one trained with partial labels. Second, MixUp might not improve the performance of PSMLC. Third, MixUp-PME can significantly improve the performance of the class(es) under extreme class imbalance. Fourth, the performance of MixUp-PME is sensitive to the value of α.

Adaptive MixUp-PME. We notice an interesting phenomenon in Fig. 5: while effusion achieves higher performance with smaller α, atelectasis and nodule tend to achieve higher performance with larger α. Intuitively, a MLC task can be decomposed into K different sub-tasks. It can be seen in Fig. 5 that the highest F1-score that can be achieved by MixUp-PME for each class is higher than the counterpart achieved by *vanilla*.[3] This means that, depending on the class imbalance situation and difficulty of the sub-task, α could be adaptive to different classes to improve the overall MTL performance. Thus, a reasonable hypothesis is *each sub-task should have an independent* α. To validate this hypothesis, we repeat the experiment of MixUp-PME with the suitable values of $\{\alpha_k\}_{k=1}^{K}$ inferred from Fig. 5. We denote this adaptive design as Adaptive MixUp-PME (AMP). The results are shown in Tab. 2, where AMP outperforms *vanilla* on all four classes this time. Surprisingly, AMP even outperforms *Oracle* on three classes. It is worth mentioning that, in this work, we use the mean F1-score as the major performance measurement. In practice, AMP might be preferred than MixUp-PME if the performance of individual class is more important than the average performance.

Robustness under MTL One limitation of the experiment in Tab. 2 is that only single class is partially labeled for each image. Under the setup of the first experiment, MixUp-PME might not be able to fully leverage the advantage of MTL, which is one of advantage of the proposed method. In the second experiment, we consider a challenging situation that each CXR image can be labeled for more than one class. To simulate this situation, for each class of each image, we randomly generate a binary number (0 or 1) from a Bernoulli distribution with equal possibilities (*i.e.* Bernoulli(0.5)). The label distributions of the simulated partially labeled training set and the fully labeled test set are summarized in Fig. 6. Note, in the experiment, MixUp can not be applied anymore as it is difficult to find two image-label pairs with the same set of labeled classes. Moreover, under this simulation, CXR images could also be unlabeled or fully labeled. We set $\alpha = 0.75$ for MixUp-PME following the first experiment. The quantitative comparison between MixUp-PME and *vanilla* is presented in

[3]The highest F1-score achieved by MixUp-PME under different values of α *vs.* the F1-score achieved by *vanilla*: 0.3219 *vs.* 0.3031 (infiltration), 0.2184 *vs.* 0.1807 (effusion), 0.2261 *vs.* 0.1807 (atelectasis), 0.2002 *vs.* 0.0704 (nodule).

Model	*Infiltration*	*Effusion*	*Atelectasis*	*Nodule*	Average
vanilla	0.3096	0.1341	0.1514	0.0553	0.1626
MixUp-PME	0.3139	**0.1717**	**0.2024**	**0.0783**	**0.1916**
Oracle	**0.3547**	0.1710	0.1739	0.0699	0.1924

Table 3. Performance comparison for PSMLC. In the training set, each CXR image could be labeled for multiple classes. MixUp can not be applied under this situation.

		Infiltration	*Effusion*	*Atelectasis*	*Nodule*	Average
Exp 1	*vanilla*	0.3031	0.1807	0.1807	0.0704	0.1837
	w/o SSL	0.2949	0.1327	0.1584	**0.2151**	**0.2002**
	w/ SSL	**0.3293**	**0.1953**	**0.1948**	0.0772	0.1991
Exp 2	*vanilla*	0.3096	0.1341	0.1514	0.0553	0.1626
	w/o SSL	0.3139	**0.1717**	0.2024	0.0783	0.1916
	w/ SSL	**0.3478**	0.1277	**0.2141**	**0.0912**	**0.1952**
Oracle		**0.3547**	0.1710	0.1739	0.0699	0.1924

Table 4. Impact of self-supervised pre-training on MixUp-PME. "*w/o* SSL" denotes that the model is not pre-trained. "*w/* SSL" denotes that the model is pre-trained.

Tab. 3. Compared with Tab. 2, the performance of *vanilla* is negatively influenced by this challenging experimental setup. On the contrary, MixUp-PME benefits from MTL with a huge performance gain. MixUp shows robust performance by outperforming *vanilla* on all four classes by a large margin and outperforming *Oracle* on three classes.

Impact of Unsupervised Pre-Training. Learning transferable representations from unlabeled data then fine-tuning with limited labels has been shown as a label-efficient learning paradigm. We leverage a state-of-the-art self-supervised learning (SSL) framework SimSiam [4] to pre-train the network backbone and repeat the first experiment above. Here, we assume additional large-scale unlabeled data are available. The pre-training is performed on 4 batches of ChestX-ray14 for 200 epochs.[4] We repeat the first and the second experiments, while this time, the models are initialized with the pre-trained weights. The results are shown in Tab. 4. With self-supervised pre-training, MixUp-PME further improves its performance over several classes and outperforms *vanilla* by a large margin. We conclude that SSL can be utilized to boost model performance under data scarcity.

6. Limitations and Future Directions

Data Decentralization. A fundamental assumption of this study is that the training partially labeled datasets can be collected and stored in a centralized fashion. However, in practice, especially in the medical domain, the decentralized datasets are stored in different hospitals [11]. Under the data regulations, it is not possible to apply MixUp-PME or AMP without exchanging users' data. An emerging research direction is federated PSMLC.

Domain Shift. In the experiments, we only consider the data scarcity and class imbalance. In addition, *domain shift* [1] could be a practical problem when collecting datasets from different sources [8]. The discussion on domain shift is left for future work.

Hyperparameters for AMP. In our second experiment, we choose the set of $\{\alpha_k\}_{k=1}^K$ based on posterior knowledge on the first experiment. In practice, a similar trick can be applied to find the suitable values of $\{\alpha_k\}_{k=1}^K$ on a small validation set. However, when K is large, this process could be troublesome. There is a trade-off between the performance and computational cost when applying AMP.

7. Conclusion

We present the first study of partially supervised multi-label classification (PSMLC) under data scarcity, an unexplored but practical problem in PSL. We propose a novel VRM method that is based the principle of maximum entropy. The experimental results show that the proposed method can be used to mitigate the data scarcity issue of PSMLC. In the future work, we will explore PSMLC under more data challenges.

Acknowledgements. The authors would like to thank Huawei Technologies Co., Ltd. for providing GPU computing service for this study.

[4] We use the second, the third, the fourth, and the fifth batches as the pre-training dataset, which contains 10^4 CXR images in total.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*, pages 137–144, 2007. 7

[2] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. 1

[3] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In *NIPS*, pages 416–422, 2001. 1, 3

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 2, 7

[5] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989. 1

[6] Wei Dai, Nanqing Dong, Zeya Wang, Xiaodan Liang, Hao Zhang, and Eric P Xing. Scan: Structure correcting adversarial network for organ segmentation in chest x-rays. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 263–273. Springer, 2018. 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 1

[8] Nanqing Dong, Michael Kampffmeyer, Xiaodan Liang, Zeya Wang, Wei Dai, and Eric Xing. Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In *MICCAI*, pages 544–552. Springer, 2018. 2, 7

[9] Nanqing Dong, Michael Kampffmeyer, Xiaodan Liang, Min Xu, Irina Voiculescu, and Eric Xing. Towards robust partially supervised multi-structure medical image segmentation on small-scale data. *Applied Soft Computing*, page 108074, 2022. 1, 2, 3, 4, 5

[10] Nanqing Dong, Michael Kampffmeyer, and Irina Voiculescu. Self-supervised multi-task representation learning for sequential medical images. In *ECML*, pages 779–794. Springer, 2021. 2

[11] Nanqing Dong and Irina Voiculescu. Federated contrastive learning for decentralized unlabeled medical images. In *MICCAI*, pages 378–387. Springer, 2021. 7

[12] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *CVPR*, pages 647–657, 2019. 1, 2

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 2

[14] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE TMI*, 2020. 1, 2

[15] Germán González, George R Washko, and Raúl San José Estépar. Multi-structure segmentation from partially labeled datasets. application to body composition measurements on ct scans. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 215–224. Springer, 2018. 1, 2, 5

[16] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. 1

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 5

[18] Edwin T Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620, 1957. 2, 3

[19] JangHyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. In *ICLR*, 2021. 1, 4

[20] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *ICML*, pages 5275–5285. PMLR, 2020. 1, 2, 3, 4

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2

[23] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, pages 1196–1204, 2013. 4

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, volume 32, 2019. 5

[25] Olivier Petit, Nicolas Thome, Arnaud Charnoz, Alexandre Hostettler, and Luc Soler. Handling missing annotations for semantic segmentation with deep convnets. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 20–28. Springer, 2018. 1, 2

[26] Jie Qin, Jiemin Fang, Qian Zhang, Wenyu Liu, Xingang Wang, and Xinggang Wang. Resizemix: Mixing data with preserved object information and true labels. *arXiv preprint arXiv:2012.11101*, 2020. 1, 2, 3, 4

[27] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1, 5

[28] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. 1

[29] Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, page 101979, 2021. 1, 2, 3

[30] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NIPS*, volume 33, pages 596–608, 2020. 4

[31] AFM Shahab Uddin, Mst Sirazam Monira, Wheemyung Shin, TaeChoong Chung, and Sung-Ho Bae. Saliencymix: A saliency guided data augmentation strategy for better regularization. In *ICLR*, 2021. 1, 4

[32] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *NIPS*, volume 30, pages 5601–5610, 2017. 4

[33] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pages 2097–2106, 2017. 1, 2, 5

[34] Yanyu Xu, Xinxing Xu, Lei Jin, Shenghua Gao, Rick Siow Mong Goh, Daniel SW Ting, and Yong Liu. Partially-supervised learning for vessel segmentation in ocular images. In *MICCAI*, pages 271–281. Springer, 2021. 1, 2

[35] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019. 1, 2, 3, 4

[36] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1, 2, 3, 4, 5

[37] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *CVPR*, pages 1195–1204, 2021. 1, 2

[38] Yuyin Zhou, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, and Alan L Yuille. Prior-aware neural network for partially-supervised multi-organ segmentation. In *ICCV*, pages 10672–10681, 2019. 1, 2, 3

[39] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002. 2, 4