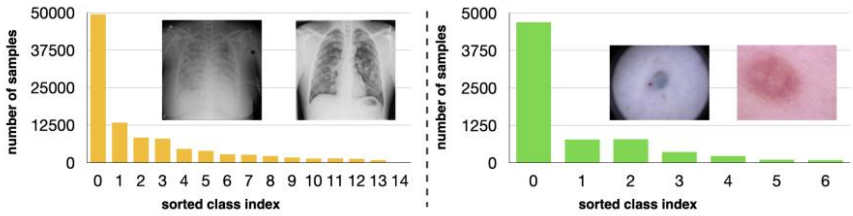
ACPL: Anti-curriculum Pseudo-labelling for Semi-supervised Medical Image Classification [CVPR 2022]

Problem:

Medical image analysis has a number of **multi-class** (e.g., a lesion image of a single class) and **multi-label** (e.g., an image from a patient can contain multiple diseases) problems, where both problems usually contain **severe class imbalances** because of the variable prevalence of diseases.



(b) Imbalanced distribution on multi-label Chest X-ray14 [39] (left) and multi-class ISIC2018 [36] (right)

Methods:

A small labelled training set $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}_L|}$.

A large unlabelled training set $\mathcal{D}_U = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}_U|}$

The pseudo-labelled set $\mathcal{D}_S$

An anchor set $\mathcal{D}_A$ contains informative pseudo-labelled samples

ACPL Optimisation

Algorithm 1 Anti-curriculum Pseudo-labelling Algorithm

require: Labelled set $\mathcal{D}_L$, unlabelled set $\mathcal{D}_U$, and number of training stages T
initialise $\mathcal{D}_A = \mathcal{D}_L$, and $t = 0$
warm-up train $p_{\theta_t}(\mathbf{x})$ **with**
$\theta_t = \arg\min_\theta \frac{1}{|\mathcal{D}_L|} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_L} \ell(\mathbf{y}_i, p_\theta(\mathbf{x}_i))$
while $t < T$ **or** $|\mathcal{D}_U| \neq 0$ **do**
  build pseudo-labelled dataset using CDSI from (2) and IM from (6):
  $\mathcal{D}_S = \{(\mathbf{x}, \tilde{\mathbf{y}}) | \mathbf{x} \in \mathcal{D}_U, h(f_{\theta_t}(\mathbf{x}), \mathcal{D}_A) = 1, \tilde{\mathbf{y}} = g(f_{\theta_t}(\mathbf{x}), \mathcal{D}_A)\}$
  update anchor set with ASP from (7):
  $\mathcal{D}_A = \mathcal{D}_A \bigcup (\mathbf{x}, \tilde{\mathbf{y}})$, where $(\mathbf{x}, \tilde{\mathbf{y}}) \in \mathcal{D}_S$, and $a(f_{\theta_t}(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A) = 1$
  $t \leftarrow t + 1$
  optimise (1) using $\mathcal{D}_L, \mathcal{D}_S$ **to obtain** $p_{\theta_t}(\mathbf{x})$
  update labelled and unlabelled sets:
  $\mathcal{D}_L \leftarrow \mathcal{D}_L \bigcup \mathcal{D}_S, \mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{D}_S$
end while
return $p_{\theta_t}(\mathbf{x})$

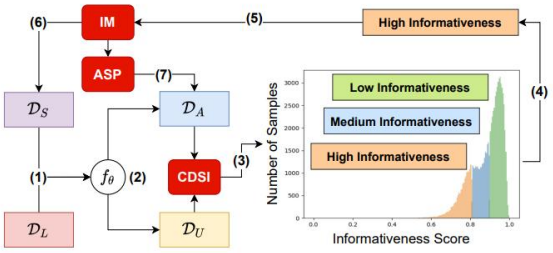


Figure 2. Anti-curriculum pseudo-labelling (ACPL) algorithm. The algorithm is divided into the following iterative steps: 1) train the model with $\mathcal{D}_S$ and $\mathcal{D}_L$; 2) extract the features from the anchor and unlabelled samples; 3) estimate information content of unlabelled samples with CDSI from (4) with anchor set $\mathcal{D}_A$; 4) partition the unlabelled samples into high, medium and low information content using (2); 5) assign a pseudo label to high information content unlabelled samples with IM from (6); 6) update $\mathcal{D}_S$ with new pseudo-labelled samples; and 7) update $\mathcal{D}_A$ with ASP in (7).

$$\ell_{ACPL}(\theta, \mathcal{D}_L, \mathcal{D}_S) = \frac{1}{|\mathcal{D}_L|} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_L} \ell(\mathbf{y}_i, p_\theta(\mathbf{x}_i)) + \frac{1}{|\mathcal{D}_S|} \sum_{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}_S} \ell(\tilde{\mathbf{y}}_i, p_\theta(\mathbf{x}_i))$$

Cross Distribution Sample Informativeness (CDSI)

First, we introduce a new approach to select the most informative unlabelled images (as far as possible from the distribution of labelled samples) to be pseudo-labelled.

The function that estimates if an unlabelled sample has high information content:

$$h(f_\theta(\mathbf{x}), \mathcal{D}_A) = \begin{cases} 1, & p_\gamma(\zeta = \text{high}|\mathbf{x}, \mathcal{D}_A) > \tau, \\ 0, & \text{otherwise}, \end{cases} \quad (2)$$

$\zeta \in \mathcal{Z} = \{\text{low}, \text{medium}, \text{high}\}$ represents the information content random variable

$\tau = \max\{p_\gamma(\zeta = \text{low}|\mathbf{x}, \mathcal{D}_A), p_\gamma(\zeta = \text{medium}|\mathbf{x}, \mathcal{D}_A)\}$

$p_\gamma(\zeta|\mathbf{x}, \mathcal{D}_A)$ can be decomposed into $p_\gamma(\mathbf{x}|\zeta, \mathcal{D}_A)p_\gamma(\zeta|\mathcal{D}_A)/p_\gamma(\mathbf{x}|\mathcal{D}_A)$

$p_\gamma(\zeta|\mathcal{D}_A) = \pi_\zeta$

$p_\gamma(\mathbf{x}|\zeta, \mathcal{D}_A) = n(d(f_\theta(\mathbf{x}), \mathcal{D}_A)|\mu_\zeta, \Sigma_\zeta), \quad (3)$

The probability in (3) is computed with the density of the unlabelled sample x with respect to the anchor set, as follows:

$$d(f_\theta(\mathbf{x}), \mathcal{D}_A) = \frac{1}{K} \sum_{\substack{(f_\theta(\mathbf{x}_A), \mathbf{y}_A) \in \\ \mathcal{N}(f_\theta(\mathbf{x}), \mathcal{D}_A)}} \frac{f_\theta(\mathbf{x})^\top f_\theta(\mathbf{x}_A)}{\|f_\theta(\mathbf{x})\|_2 \|f_\theta(\mathbf{x}_A)\|_2}, \quad (4)$$

Informative Mixup (IM)

Second, we introduce a new pseudo-labelling mechanism, called informative mixup, which combines the model classification with a K-nearest neighbor (KNN) classification guided by sample informativeness to improve prediction accuracy and mitigate confirmation bias.

After selecting informative unlabelled samples with (2), we aim to produce reliable pseudo labels for them.

$$\begin{aligned} \tilde{\mathbf{y}}_{\text{model}}(\mathbf{x}) &= p_\theta(\mathbf{x}), \\ \tilde{\mathbf{y}}_{\text{KNN}}(\mathbf{x}) &= \frac{1}{K} \sum_{(f_\theta(\mathbf{x}_A), \mathbf{y}_A) \in \mathcal{N}(f_\theta(\mathbf{x}), \mathcal{D}_A)} \mathbf{y}_A. \end{aligned} \quad (5)$$

$$\begin{aligned} \tilde{\mathbf{y}} = g(f_\theta(\mathbf{x}), \mathcal{D}_A) &= d(f_\theta(\mathbf{x}), \mathcal{D}_A) \times \tilde{\mathbf{y}}_{\text{model}}(\mathbf{x}) \\ &\quad + (1 - d(f_\theta(\mathbf{x}), \mathcal{D}_A)) \times \tilde{\mathbf{y}}_{\text{KNN}}(\mathbf{x}). \end{aligned} \quad (6)$$

Anchor Set Purification (ASP)

After estimating the pseudo label for informative unlabelled samples, we aim to update the anchor set with informative pseudo-labelled samples to maintain density score from (4) accurate in later training stages. We select the least connected pseudo-labelled samples to be inserted in the anchor set:

$$a(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A) = \begin{cases} 1, & c(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A) \leq \alpha, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where the pseudo-labelled samples with $a(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A) = 1$ and $\tilde{\mathbf{y}} = g(f_\theta(\mathbf{x}), \mathcal{D}_A)$ from (6) are inserted into the anchor set.

$c(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A)$ of a pseudo-labelled sample $f_\theta(\mathbf{x})$ in (7) is computed in three steps (see Fig. 3): 1) find the KNN samples $\mathcal{N}(f_\theta(\mathbf{x}), \mathcal{D}_A)$ from $f_\theta(\mathbf{x})$ to the anchor set $\mathcal{D}_A$; 2) for each of the K elements $(\mathbf{x}_A, \mathbf{y}_A) \in \mathcal{N}(f_\theta(\mathbf{x}), \mathcal{D}_A)$, find the KNN set $\mathcal{N}(f_\theta(\mathbf{x}_A), \mathcal{D}_U)$ from $f_\theta(\mathbf{x}_A)$ to the unlabelled set $\mathcal{D}_U$; and 3) $c(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A)$ is calculated to be the number of times that the pseudo-labelled sample $\mathbf{x}$ appears in the KNN sets $\mathcal{N}(f_\theta(\mathbf{x}_A), \mathcal{D}_U)$ for the K elements of set $\mathcal{N}(f_\theta(\mathbf{x}), \mathcal{D}_A)$. The threshold α in (7) is computed with $\alpha = \min_{\mathbf{x} \in \mathcal{D}_S} c(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A)$.

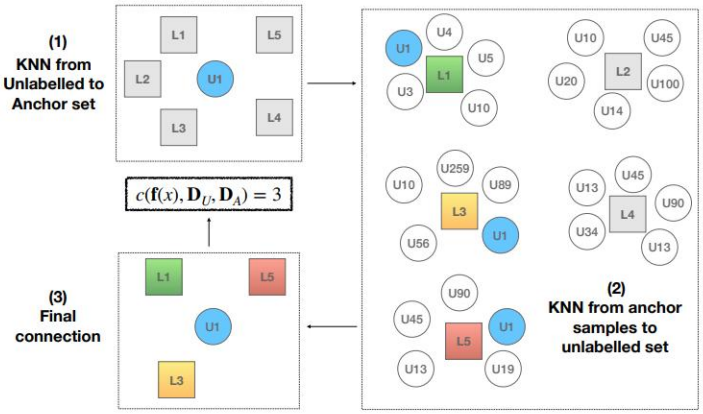


Figure 3. **ASP**: 1) find KNN samples from an informative unlabelled sample to the anchor set $\mathcal{D}_A$; 2) find KNN samples from each anchor sample of (1) to the unlabelled set $\mathcal{D}_U$; and 3) calculate the number of surviving nearest neighbours. Samples with the smallest values of $c(.)$ are selected to be inserted into $\mathcal{D}_A$.