# Distribution-Aware Multi-Label FixMatch for Semi-Supervised Learning on CheXpert.

Sontje Ihler[1], Felix Kuhnke, Timo Kuhlgatz[1], Thomas Seel[1]
[1]Institute of Mechatronic Systems, Leibniz Universität Hannover
✉ sontje.ihler@imes.uni-hannover.de

## Abstract

*Semi-supervised learning (SSL) has achieved remarkable success for multiclass classification in recent years, yielding a promising solution for medical image classification where labeled data is scarce but unlabeled images are accessible. In the context of multi-label problems however, SSL is still under-explored. In this work we adapt FixMatch to the multi-label scenario, specifically focusing on CheXpert, a multi-label chest X-ray classification dataset which is imbalanced and only partially labeled. Leveraging distribution alignment, our proposed method, ML-FixMatch+DA, achieves solid performance gains in SSL tasks (AUC: +2.6%) and in a missing label scenario (AUC: +1.9%). In contrast to previous work we achieve a performance gain on CheXpert using FixMatch. We show that in contrast to multiclass FixMatch, where distribution alignment is optional, it is essential for multi-label FixMatch to handle class imbalance and generate reliable (positive and negative) pseudo-labels. Our pseudo-label selection is based on a single threshold for all classes and handles imbalance with no prior knowledge on label distributions. Our adaptation keeps the simplicity of the original multiclass FixMatch with no added hyperparameters (even for imbalanced data) and demonstrates the feasibility of simple SSL for multi-label problems, filling a crucial gap in the literature.*

## 1. Introduction

Automating medical image diagnosis holds the promise of transforming healthcare by streamlining the diagnostic process, making it faster and more efficient. At the heart of this innovation is the use of neural networks, which require extensive data for training. This need for data ensures that the algorithms learn accurately from a vast array of examples, covering a wide range of conditions and scenarios.

Manual labeling of large medical image datasets is not feasible at large scale as it requires medical experts and is therefore time-consuming and expensive. At the same time medical images are captured at large scale in everyday clinical practice making unlabeled images much easier accessible than labeled ones. The goal is to exploit this data with minimal expert annotation (efficient labeling). There are two promising solutions to solve this: (1) automatic labeling using natural language processing like done for the CheXpert dataset [8] and (2) semi-supervised learning (SSL) which combines supervised learning from labeled data and self-supervised learning from unlabeled data [10, 11, 16, 24, 27].

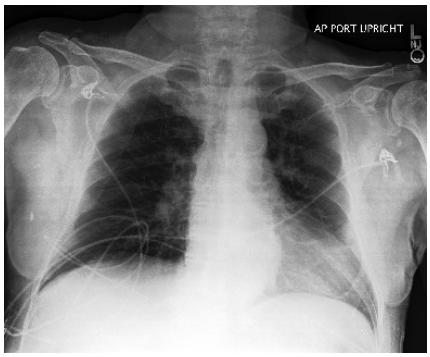
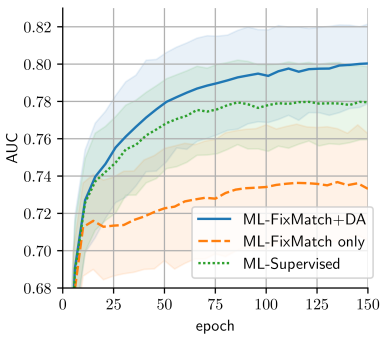
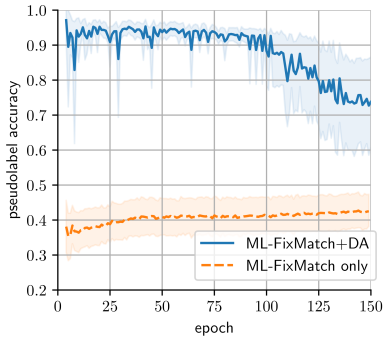



Figure 1. Example image from X-ray classification dataset CheXpert [8]. The dataset is challenging for semi-supervised learning as the dataset has **multi-label annotations**, the label distribution is **imbalanced**, and because the dataset was automatically labeled from patient files each image is **only partially labeled**.

One of the most popular approaches for SSL is FixMatch [16]. Its popularity derives from being highly effective while being simple at the same time. FixMatch combines consistency regularization (ensuring consistent predictions across augmentations) and pseudo-labeling to leverage unlabeled data effectively. Even though FixMatch is so popular for SSL, it was designed only for multiclass classification, *i.e.* each image sample represents exactly one class. Unfortunately, many medical image diagnosis tasks are actually multi-label problems *i.e.* an image can show more than one pathology class or none at all. Compared to SSL




(a) AUC for our proposed ML-FixMatch+DA vs. supervised baseline.




(b) Distribution alignment (DA) boosts pseudo-label accuracy

Figure 2. We propose distribution-aware multi-label FixMatch (ML-FixMatch+DA) for semi-supervised multi-label learning. (a) ML-FixMatch+DA is able to exploit the unlabeled data to improve model performance (AUC) compared to the supervised baseline. ML-FixMatch+DA has no added hyper-parameters compared to multiclass FixMatch and shares its simplicity. Positive and negative pseudo-labels are both obtained using a single threshold, the same threshold is also used for all classes. This is achieved by incorporating distribution alignment (DA) into the FixMatch training to handle imbalance. (b) DA boosts multi-label pseudo-label accuracy on CheXpert from 40% to over 90%, making it possible to perform semi-supervised learning on CheXpert using a simple, straight-forward multi-label FixMatch adaptation with no added hyperparameters. In the later training stages the pseudo label accuracy decreases but the model performance for ML-FixMatch+DA is still improving.

for multiclass classification problems, which has made great progress over the last years (heavily based on FixMatch), SSL for multi-label classification is still an understudied problem [22].

One example of a medical multi-label dataset is the previously mentioned CheXpert dataset [8], see Fig. 1. CheXpert is a dataset for automated chest X-ray image interpretation, which features automatically generated labels from patient reports. One might think this solves the limited label issue, however, patient files are not designed to extract pathological information by algorithms. They generally do not provide a full diagnosis for each image but often only describe one or few relevant pathologies and/or the difference to a previous diagnostic stage [18]. In conclusion, the training set of the CheXpert dataset is not labeled for pathologies not mentioned in the corresponding patient file of an image *i.e.* it is only partially labeled. This indicates that automatic labeling itself is not enough to solve the efficient labeling issue, even with the now drastic improvement in language models. However, it can actually form a great basis for SSL.

Zenk *et al*. [25] recently applied FixMatch to the CheXpert dataset and found that it does not lead to improved model performance. However, it seems they did not adapt FixMatch to the multi-label scenario of CheXpert.

In this work we show that FixMatch can in fact improve model performance on the CheXpert dataset if adapted to the multi-label scenario and the class distribution is taken into account. Due to its real-world nature, the label distribution of the CheXpert dataset is imbalanced. We show that our adaptation distribution-aware multi-label FixMatch (ML-FixMatch+DA) not only works for a standard SSL scenario but also when learning from incomplete, *i.e.* missing labels. Our adaptation starts by changing the multiclass to multi-label losses and by proposing a simple strategy to generate negative pseudo-labels for the multi-label losses. We do this without adding new hyperparameters or complexity to FixMatch. Tuning hyperparameters is a challenging task with limited labels and it is best if it can be avoided. We achieve this by employing distribution alignment (DA) [1] for pseudo-label generation which is a an established add-on for multiclass FixMatch [1, 20, 25].

Our contributions are manifold: (I) We are the first to provide an adaptation of FixMatch to a multi-label classification task (ML-FixMatch) which is direct and straightforward without adding hyperparameters. (II) By incorporating (parameter-free) distribution alignment, we are able to use only a single threshold value to obtain highly accurate pseudo-labels. This means all pseudo-labels of all classes, indifferent of being positive or negative pseudo-labels, can be masked with a single parameter. (III) In contrast to previous work we are able to increase model performance on CheXpert using FixMatch in 1) an SSL task and 2) learning from incomplete labels. (IV) While DA is optional for multiclass FixMatch to increase model performance, our findings demonstrate that DA is actually critical for multi-label FixMatch to work on imbalanced class distributions.

## 2. Multiclass FixMatch

For easier understanding of our adaptations we will first revisit the concept and math behind multiclass FixMatch [16] in this section. We will then build upon the provided equa-

tions in the following section to help highlight the difference between multiclass and multi-label FixMatch.

### 2.1. Multiclass Problem

FixMatch was designed for multiclass SSL. Multiclass classification describes the problem where each data sample is affiliated with exactly one class. We use the following annotations to describe the multiclass problem. Let $x \in X$ be a training sample with true class distribution $p \in \{0,1\}^C$ with $|p| = 1$, where $C$ is the number of classes (or pathologies in our case). For a given model $f$ the model's prediction for $x$ is $f : x \to q \in [0,1]^C$ and $|q| = 1$ with a predicted probability between 0 and 1 for each class and $p_c$ being the probability for class $c$. $|q| = 1$ is enforced by a Softmax activation after the last model layer.

### 2.2. Multiclass FixMatch

FixMatch is a multiclass SSL strategy that leverages both labeled and unlabeled data using pseudo-labels. It combines a supervised loss $L_s$ and a self-supervised pseudo-label loss $L_{\text{PL}}$ with weighting factor $\alpha$:

$$L = L_{\text{s}} + \alpha \cdot L_{\text{PL}}. \tag{1}$$

The supervised loss function $L_s$ is computed from the labeled data using cross entropy loss:

$$L_{\text{s}} = \frac{1}{N} \sum_{i}^{N} p_i \log(q_i^w). \tag{2}$$

For easier reading we simplified $\sum_i^C p_{i,c} \log(q_{i,c})$ to $p_i \log(q_i)$ in the above and all following equations. FixMatch uses weakly augmented training samples in its supervised loss. We provide loss computation per batch. The number of labeled samples per batch is $N$.

To train on the unlabeled data, FixMatch uses augmentation anchoring to create pseudo-labels. FixMatch creates two different augmentations of a training sample with the idea that both augmentations share the same class distribution $p$ as they both derive from the same image sample. FixMatch incorporates weak and strong image augmentations, $aug_w$ and $aug_s$ respectively. We refer to predictions from weak augmentations to $q^w = f(aug_w(x))$ and from strong augmentations to $q^s = f(aug_s(x))$. Positive pseudo-labels $\tilde{p}^+ \in [0,1]^C$ with $|\tilde{p}^+| = 1$ are created from the weakly augmented images by creating hard labels from $q^w$. FixMatch then filters these pseudo-labels by confidence so that only high-confidence predictions are retained for loss computation to avoid noisy training.

A prediction is considered confident if the predicted probability is higher than a threshold $t$ resulting in the following pseudo-label loss:

$$\tilde{p}^+ = \mathbb{1}(q^w > t), \tag{3}$$

$$L_{\text{PL}} = \frac{1}{M} \sum_{i}^{M} \mathbb{1}_{q_i^w > t} \log(q_i^s). \tag{4}$$

where $M$ is the number of unlabeled samples in a batch. If the model is not confident about a prediction, it is discarded. The pseudo-label loss is naturally small in the beginning of the training process. Only $m = \sum_i^M \mathbb{1}(\tilde{p}_i^+ = 1)$ are selected for pseudo-labels but the loss is still normalized with $M$ and $M >> m$. The loss increases over time with increasing prediction confidence.

## 3. Multi-Label FixMatch

This section explains how we adapt multiclass FixMatch to the multi-label setting and incorporate distribution alignment to handle imbalance and label masking to handle missing labels in CheXpert. While distribution alignment is optional for multiclass FixMatch to boost performance, it is critical for single-threshold, multi-label FixMatch adaptation to actually work on class-imbalanced multi-label data.

### 3.1. Multi-label Problem

Multi-label classification describes the problem where each data sample can be affiliated to an arbitrary number of classes and therefore have more than one label associated with it. In applications the number of possible classes is limited to the number of observed classes. It is also possible that there is no class affiliation.

To underline the similarities between multiclass and multi-label FixMatch we adapt the previous notation to our multi-label problem. Again let $x \in X$ be a training sample but now with a true multi-label class distribution $p \in \{0,1\}^C$, where $C$ again is the number of classes. Again for a given model $f$ the model's prediction for $x$ is $f : x \to q \in [0,1]^C$ with a predicted probability between 0 and 1 for each class. In the multi-label case this is achieved by applying Sigmoid activation to each class after the final model layer. (In the multi-label case the elements of $p$ (nor $q$) must **not** sum up to 1.)

### 3.2. Multi-label Adaptation for FixMatch

The overall concept of ML-FixMatch is identical to multiclass FixMatch which results in the same composition of the loss function $L'$ from a supervised loss $L'_s$ and a pseudo-label loss $L'_{\text{PL}}$ with weighting factor $\alpha$:

$$L' = L'_{\text{s}} + \alpha L'_{\text{PL}}. \tag{5}$$

We first adapt the supervised loss from cross entropy to binary cross entropy to optimize for multi-label classification. Binary cross entropy is identical to cross entropy for

positive labels with an added term for negative labels (see Sec. 2.2):

$$L'_{\mathrm{s}} = \frac{1}{N}\sum_{i}^{N} p_i \log(q_i) + (1-p_i)\log(1-q_i). \quad (6)$$

In the second step the pseudo-label loss is also reformulated to binary cross entropy. FixMatch only generates positive pseudo-labels. We propose a simple strategy to generate negative pseudo-labels. To maintain FixMatch' simplicity we mirror the generation of positive pseudo-labels and generate negative pseudo-labels by filtering predictions with very low probabilities using the confidence threshold $(1-t)$. We refer to a negative pseudo-label as $\tilde{p}^-$.

$$\tilde{p}^- = \mathbb{1}(q < (1-t)), \quad (7)$$

$$L'_{\mathrm{PL}} = \frac{1}{M}\sum_{i}^{M} \mathbb{1}_{(q_i^w > t)} \log(q_i^s) + \mathbb{1}_{(q_i^w < (1-t))} \log(1-q_i^s). \quad (8)$$

Using the same value $t$ for the upper and lower confidence threshold will however fail in the presence of class imbalance. To solve this we align the model predictions for the pseudo-labels to a uniform distribution (for each class) using distribution alignment.

### 3.3. Distribution Alignment for ML-FixMatch

The concept of ML-FixMatch+DA is based on Distribution Alignment (DA). DA was introduced by ReMixMatch to reduce confirmation bias [1]. Distribution alignment aims to mitigate the impact of class imbalance pseudo-label generation during SSL. It ensures that the learned model's predictions align with the underlying data distribution, even when the distribution is skewed. This is achieved by computing the expected probability values for each class *i.e.* the mean probability $\bar{q}_{\mathrm{c}}$ of all model predictions for each class $c$. Each model prediction $q$ is then aligned to $q^*$ by normalizing each class prediction with $\bar{q}_{\mathrm{c}}$ :

$$q^*_{\mathrm{c}} = q_{\mathrm{c}} \cdot \frac{1}{\bar{q}_{\mathrm{c}}}, \quad (9)$$

$$\bar{q}_{\mathrm{c}} = \frac{1}{M}\sum_{i}^{M} q_{i,\mathrm{c}}. \quad (10)$$

This results in the aligned pseudo loss function:

$$L^*_{\mathrm{PL}} = \frac{1}{M}\sum_{i}^{M} \mathbb{1}_{(q_i^{*,w} > t)} \log(q_i^s) + \mathbb{1}_{(q_i^{*,w} < (1-t))} \log(1-q_i^s). \quad (11)$$

The computation of the expected values $\bar{q}_{\mathrm{c}}$ for each class from probabilities $q$ are the same for multiclass and multi-label. The difference lies in the activation function that was used to obtain $q$. While multiclass DA computes $q$ from Softmax, multi-label DA requires Sigmoid activation.

DA has the advantage that it predicts $\bar{q}_{\mathrm{c}}$ solely based on the model's predictions and does not pose any assumptions on the class distributions based on the labeled data (like *e.g.* logit adjustment [12]). This is a necessity for generating distribution-aware pseudo-labels for CheXpert as the class distribution between labeled samples and unlabeled samples differ drastically.

For ML-FixMatch+DA $L^*_{PL}$ replaces $L'_{PL}$ in Eq. (5).

### 3.4. CheXpert Masking for Missing Labels

CheXpert is only partially labeled. We address this by masking $L'_s$ similarly to $L'_{\mathrm{PL}}$ in Eq. (8) and only compute the supervised loss from existing labels 1 and 0. To counteract fluctuation in our supervised loss due to fluctuating amount of labels in a batch we divide by the number of labels $n$ in a batch rather than the labeled batch size $N$:

$$L^*_{\mathrm{s}} = \frac{1}{n}\sum_{i}^{N} \mathbb{1}_{p_i=1} \log(q_i) + \mathbb{1}_{p_i=0} \log(1-q_i), \quad (12)$$

$$n = \sum_{i}^{N} \mathbb{1}_{p_i=1} + \mathbb{1}_{p_i=0}. \quad (13)$$

For learning on CheXpert $L^*_s$ replaces $L'_s$ in Eq. (5).

## 4. Experiments

In this section, we describe our experimental setup to show the effectiveness of our proposed FixMatch adaptation ML-FixMatch+DA on the CheXpert dataset [8].

### 4.1. Dataset and Datasplit

**CheXpert** We perform our experiments on the multi-label chest X-ray dataset CheXpert [8]. The training set was automatically labeled from patient files, while the validation and test set were manually labeled by radiologists. The training set contains approx. 225k images and has labels for 15 pathologies. We follow common protocol to only use five of the 15 pathologies for model optimization and validation: atelectasis, cardiomegaly, consolidation, edema and pleural effusion [7, 8, 25]. The automatically labeled training set is special as the labels are incomplete ie. not fully labeled and it contains uncertain labels. There are four label categories:

1. pathology present according to patient file (1)
2. pathology explicitly not present according to patient file (0)
3. pathology mentioned in patient file but algorithm is unsure if patient has pathology (*u* for uncertain)
4. not mentioned in patient file (*not labeled*)

**CheXpert5000** SSL experiments require labeled and unlabeled samples. To ensure repeatability we use the public

Table 1. Label distribution of one of the five CheXpert5k training sets for five pathologies [7]. The dataset is only partially labeled with confident labels 1 and 0. It also contains uncertain labels *u* and missing labels *not labeled*. Because CheXpert is a multi-label dataset each image can contain 5 labels, one for each pathology. If this training set was fully labeled it would have 25k labels. Due to it being only partially labeled this set only has approx. 7k confident labels. The imratio describes imbalance of the label distribution and is computed from the ratio of positive to all confident labels for each class. It cannot be assumed that the class distribution of the labeled samples and the unlabeled samples is the same or similar. The unlabeled samples contain more negative samples than the labeled samples.

| **Label** | **Atelectasis** | **Cardiomegaly** | **Consolidation** | **Edema** | **Pleural Effusion** |
|---|---|---|---|---|---|
| 1 | 813 | 637 | 363 | 1443 | 2086 |
| 0 | 16 | 146 | 368 | 367 | 485 |
| *u* | 810 | 178 | 666 | 332 | 271 |
| *not labeled* | 3361 | 4039 | 3603 | 2858 | 2158 |
| # confident labels | 829 | 783 | 731 | 1810 | 2571 |
| imratio | 98.1 | 81.4 | 49.3 | 79.7 | 81.3 |

datasplit from the CheXpert5000 (CheXpert5k) study [7] as our labeled data to mimic the limited label scenario. It splits the full CheXpert training set into a new training (124,664 samples), validation (16,989 samples) and test set (25,205 samples) by uniform sampling *i.e*. all datasets have about the same distribution as the full dataset. From the new training set the authors again sampled five subsets of 5000 samples each intended as a new dataset for limited data learning. We provide the class distribution and the imratio, a metric for label imbalance, in Tab. 1. Only the training subsets are used in this study.

**Evaluation set** For validation we follow common practice and evaluate our experiments on the official validation set of 234 fully labeled chest X-rays which were fully manually annotated by three board certified radiologists.

### 4.2. Experimental Setup

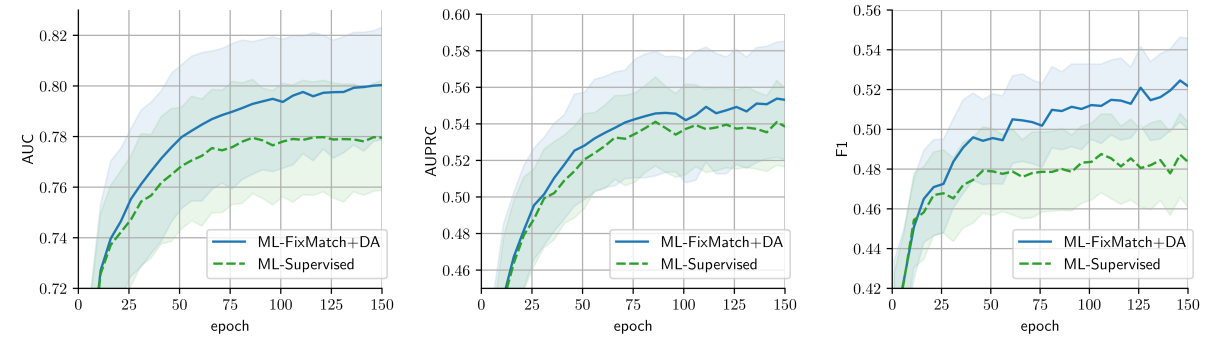
On our experiments we validate the feasibility of our approach in two settings and perform an ablation study for distribution alignment.

**Feasibility Study** For our feasibility study, we perform standard SSL (study I) on two hyperparameter configurations, as well as a study solely focusing on the missing labels (study II). An overview of our experimental settings is presented in Tab. 2. To assess the effectiveness of our method in a standard SSL setting (study I), we conduct experiments using the CheXpert5k training subsets described previously as labeled data, hence using a total of 5000 labeled samples. As we rely on the confident labels (1 and 0) only for supervised learning this results in which results in approx. 7000 labels (see Table 1). We perform our standard SSL experiments on two different set of hyperparameters to test the robustness of our approach. As unlabeled data we use the remaining full CheXpert5k training set with 120k samples and hence 600k potential pseudo-labels (5 for each unlabeled image sample). In a second experiment (study II), we test if the model performance can also be improved using only the uncertain and non-labeled image samples (label *u* and *not labeled* in Table 1) from the partially labeled CheXpert5k datasets *i.e*. if we only use those labels as potential pseudo-labels. This is especially interesting as these were skipped by the automatic labeler and might therefore be more difficult than additional unlabeled samples (which the automatic labeler was able to annotate with confident labels). For this study of exploiting the missing labels we therefore reuse the labeled CheXpert5k subsets and use the missing labels (uncertain and not labeled) of this training subset for unlabeled data. Because the CheXpert5k subsets are only partially labeled (with 7k labels) they have approximately 18k missing labels which can be exploited by ML-FixMatch+DA to generate pseudo-labels. We compare our ML-FixMatch+DA with a supervised ML-Baseline which was trained solely supervised with binary cross entropy.

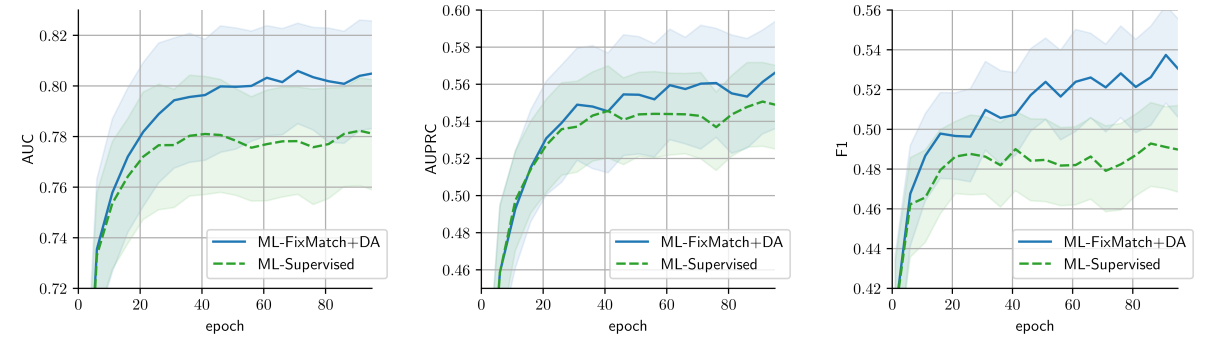
Table 2. Feasibility study design. (I) We validate our approach for two different hyperparameter configurations for the standard SSL setup with a labeled and an unlabeled dataset. (II) We validate our approach only on the missing labels (uncertain and not labeled) of the CheXpert5k training set without adding additional unlabeled data. The CheXpert5k datasets have approximately 18k missing labels which can be exploited by ML-FixMatch+DA to generate pseudo-labels.

| **Study** | **# labeled images** | **# labels** | **# unlab. images** | **# potential PL*** | **N** | **M** |
|---|---|---|---|---|---|---|
| I | 5k | ∼7k | 120k | 600k | 32 | 64 |
| I | 5k | ∼7k | 120k | 600k | 12 | 84 |
| II | 5k | ∼7k | (5k) | 18k | 32 | 64 |

**Ablation study for DA** To validate the necessity of DA in our approach we perform an ablation study where we compare ML-FixMatch+DA to ML-FixMatch without DA.




(a) Reproducing CheXpert5000 hyperparameters: labeled batch size 32, learning rate 3e-3, unlabeled batch size 64




(b) Reproducing 1:7 FixMatch ratio: labeled batch size 12, learning rate 1e-3, unlabeled batch size 84

Figure 3. Results for the feasibility study for standard SSL setting. ML-FixMatch+DA improves model performance on all performance metrics *i.e.* area under the curve (AUC), area under the precision recall curve (AUPRC) and the F1 score. We show the results over five runs. While the supervised baseline is already converged ML-FixMatch+DA is still improving.

### 4.3. Implementation Details

Following the CheXpert5k study we employ BitM-50x1 [9] a ResNet-50 [6] variant trained on ImageNet-21k [14] for improved generalization performance .

For our standard SSL experiment (study I), we validate two hyperparameter configurations (1) We adapt the hyperparameters from the CheXpert5k study. We use SGD optimizer, learning rate of 3e-3 and batch size 32 for our labeled data. We add batch size of 64 for unlabeled data which maxes out VRAM (24GB) for one batch. (2) FixMatch recommends a ratio of 1:7 for labeled to unlabeled training samples. We therefore employ a labeled batch size 12 with an unlabeled batch size 84. Due to the decreased labeled batch size we decrease the learning rate to 1e-1.

For our missing label setting (study II) and ablation study we adapt the first hyperparameter configuration. Because we don't want to use the large CheXpert5k validation set, as it is not representative for limited learning and small validation sets have a risk of being unreliable, we do not use a validation set to estimate optimal training time instead we train our models for a fixed amount of epochs. These were determined by the convergence of the supervised baseline. For configuration (1) we used 150 epochs and configuration (2) we used 100 epochs. We use a constant learning rate for both configurations. It is common practice to map the uncertain labels and missing labels to either 0 or 1 to increase the amount of training labels for supervised learning, however, this leads to noise in the training labels. To keep noise to a minimum, we only use confident labels 1 and 0 and ignore all other labels.

For weak and strong augmentation we directly adapt FixMatch augmentations (but no Cutout[3]) [16]. We skip Cutout to avoid label flips in the pseudo-labels. Following FixMatch we set $\alpha$ to 0.5 and $t$ to 0.95. We use Sigmoid activation in the final layer. We do not use temperature scaling [4] for two reasons. One, temperature scaling assumes a Softmax probability distribution and is not directly applicable for our multi-label scenario. Two, the default hyperparameter for temperature scaling in multiclass FixMatch is set to 1 anyway which is equivalent to no scaling. We further follow [9] and [7] and use CheXpert at resolution 320x320 which we downscale to 224x224 for model training and validation. The pseudo label distribution $\bar{q}$ is estimated from the previous $256 \times M$ predictions of unlabeled samples. $M$ is the number of unlabeled samples per batch. We use PyTorch [13] and the timm libray [21].

## 5. Results

### 5.1. Feasibility

**Study I: Standard SSL** To validate the feasibility for standard SSL we compare ML-FixMatch+DA to a supervised baseline computed from the confident CheXpert5k labels with binary cross entropy. We provide area under the curve (AUC), area under the precision recall curve (AUPRC) and the F1 score for both hyperparameter configurations. We follow common binary classification protocol and set the decision threshold for the F1 score at 0.5. The F1 score is a combined metric of recall and precision which rewards high precision and recall. It rewards the balance of the two. We provide the mean over five runs with different labeled data splits (see Sec. 4.1).

Our proposed method ML-FixMatch+DA improves model performance on all metrics and for both hyperparameter configurations showing the feasibility of our approach. Our results are shown in Fig. 3. This proves that the model is able to exploit the unlabeled data and learn from the pseudo-labels. Our approach can handle CheXpert's imbalance with a single threshold for pseudo-label generation. The improved F1 score is especially interesting: a higher F1 score results from a more balanced recall to precision ratio which means that ML-FixMatch+DA pushes the model to have an optimal decision threshold at 0.5. This is beneficial as this is a sign of a non-biased classifier. Please note that a FixMatch epoch contains more images than a supervised epoch, however, the supervised baselines were fully converged while ML-FixMatch+DA was still improving.

We tested our approach with two sets of standard hyperparameter configurations (no hyperparameter tuning) to get an idea of the robustness. We see robust improvements for both configurations. These are promising findings towards robustness. Robustness to changing hyperparameters is a big win in limited data learning.

**Study II: Exploiting missing labels** In our second feasibility experiment we test if the model performance can also be improved using only the uncertain and non-labeled image samples from the partially labeled CheXpert5k datasets. We did this by again employing the confident labels from the limited 5k datasets for supervised training and instead of providing the rest of the full CheXpert dataset we provided the limited dataset again but without labels. We can see from Fig. 4 and Tab. 3 that we can achieve a similar performance gain to the SSL setting only by using only the uncertain or unlabeled labels in the 5k training samples. This means that ML-FixMatch+DA is able to exploit the uncertain and unlabeled data that was not labeled by the automatic labeler. ML-FixMatch+DA can therefore also be beneficial for partially labeled data and missing labels which are common occurrences for medical datasets [23], especially if they are created from automatic annotations.

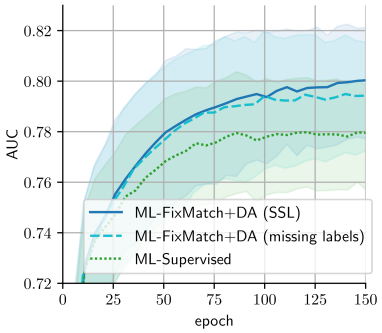



Figure 4. Results for feasibility of exploiting missing labels. We can see that only using the missing labels as unlabeled data improves the model performance significantly.

### 5.2. Ablation study for DA

In a final experiment, we show that DA is crucial in our approach. We can see that in Fig. 2 FixMatch with no DA performs poorer than the supervised baseline *i.e.* the model degenerates during training with pseudo-labels. This is due to very poor pseudo-label accuracy. In Fig. 2b we see how DA boosts pseudo-label accuracy from 40% to above 90%.

## 6. Related Work

In this section we provide a short review on current work on multi-label semi-supervised learning. Generally multi-label semi-supervised learning is a very understudied problem with few publications [5, 10, 15, 22, 24], with imbalance even more understudied [10, 22]. Most of these are in the medical context [5, 10, 24].

Apart from FixMatch [16] closest to our work is class-distribution-aware thresholding (CAT) [22]. It is, to our knowledge, the first and only work to explicitly address imbalanced, multi-label semi-supervised learning. Their approach is very similar to ours and mainly differs in the confidence-thresholding of the pseudo-labels where our approach is much simpler. While we rely on a single threshold for the whole algorithm, CAT requires two thresholds for each class, one for positive and one for negative pseudo-labels. Furthermore, CAT relies on the labeled data class distribution to approximate the class distribution for unlabeled data. In our case the labeled and unlabeled data distribution differ drastically so CAT will not be able to estimate reliable thresholds.

The most recent works on multi-label SSL for medical image classification are ACPL [10] and PEFAT [24]. They both perform experiments on the multiclass dataset ISIC 2018 [2, 17] and the multi-label dataset chestX-ray14 [19]. ChestX-ray14 is a slightly older dataset for x-ray classification than CheXpert. ACPL selects pseudo-labels

Table 3. Results for our feasibility study for standard SSL setting and exploiting missing labels. Metrics are computed from five runs. $N$ is the number of labeled samples per batch with a total of 84 samples per batch. Results for $N = 32$ were taken at epoch 150. Results for $N = 12$ were taken at epoch 100. Our proposed ML-FixMatch+DA always improves the supervised baseline.

| Experiment | Method | N | AUC | AUPRC | F1 |
|---|---|---|---|---|---|
| Baseline | ML-Supervised | 32 | $0.7794 \pm 0.0220$ | $0.5351 \pm 0.0300$ | $0.4866 \pm 0.0193$ |
| Standard SSL (I) | ML-FixMatch +DA | 32 | $\mathbf{0.8004 \pm 0.0177}$ | $\mathbf{0.5530 \pm 0.0226}$ | $\mathbf{0.5210 \pm 0.0158}$ |
| Missing labels (II) | ML-FixMatch +DA | 32 | $0.7942 \pm 0.0175$ | $0.5473 \pm 0.0185$ | $0.5044 \pm 0.0226$ |
| Baseline | ML-Supervised | 12 | $0.7829 \pm 0.0164$ | $0.5460 \pm 0.0323$ | $0.4888 \pm 0.0071$ |
| Standard SSL (I) | ML-FixMatch +DA | 12 | $\mathbf{0.8026 \pm 0.0121}$ | $\mathbf{0.5611 \pm 0.0163}$ | $\mathbf{0.5290 \pm 0.0071}$ |

based on the distance in feature space. They claim that unlabeled samples with a larger distance to labeled samples are more informative and are more likely to belong to the minority class. This would be beneficial in imbalanced SSL. PEFAT selects pseudo-labels based on consistency but instead of selecting high confidence predictions they select low loss predictions (self-supervised loss based on consistency). Both ACPL and PEFAT compute selection thresholds based on Gaussian mixture models which is great as thresholds generically adapt to datasets or tasks at hand. However, it is noticeable that both ACPL and PEFAT were designed with a focus on multiclass classification and do not fully embrace the multi-label problem. ACPL creates pseudo-labels based on a graph-based nearest neighbor approach which to our understanding leads to single-label pseudo-labels as does PEFAT's argmax method. Allowing only a single label (multiclass label) for each image can be a problem for multi-label learning. Single-label pseudo-labels from argmax will favor pseudo-labels for the easiest pathology in an image (easy classes) and therefore an over-selection of pseudo-labels for these classes. We don't know the effect of single-label pseudo-labels selected by ACPL. It would be interesting if these are robust to the just mentioned effect. Neither methods employs negative pseudo-labels.

Recent works for semi-supervised learning on CheXpert are Gyawali *et al*. [5] and Zenk *et al*. [25]. Zenk *et al*. use FixMatch on CheXpert but it seems like did not adapt FixMatch to a multi-label setting. Gyawali *et al*. use global latent mixing and mixup [26]. The supervised baseline seems to have been trained using multiclass cross entropy instead of multi-label binary cross entropy[1] This again bears the risk to favor easy classes during training and ignore the rest if easy classes are present in an image sample.

To our best knowledge the first to introduce negative pseudo-labels into semi-supervised learning were Rizve *et al*. when they proposed uncertainty-aware pseudo-labeling (UPS) [15] for multiclass and multi-label semi-supervised learning. Instead of relying on augmentation anchoring and confidence like FixMatch and our adaptation, they combine confidence and uncertainty to select reliable pseudo-labels. The selection requires several parameters. The uncertainty estimation adds additional complexity compared to our approach. UPS theoretically addresses our problem, however it was recently outperformed by ACPL [10] which was then outperformed by PEFAT [24] on chestX-ray14.

Every year there is a large number of variations built upon the FixMatch concept for multiclass classification. We believe that our adaptation ML-FixMatch+DA holds the same potential for multi-label SSL.

[1]Information taken from the paper. However, the authors provide code with binary cross entropy, so we are not sure.

## 7. Conclusion

In conclusion, this work successfully achieved its objective of adapting FixMatch to the semi-supervised multi-label learning (SSMLL) scenario while maintaining simplicity and avoiding the introduction of complexity or additional hyperparameters. By providing a straightforward adaptation of FixMatch, our approach ensures accessibility and usability. This simplicity mirrors the popular characteristics that have established FixMatch as a cornerstone in the field of semi-supervised learning

The significance of this adaptation is underscored by the relatively understudied nature of SSMLL compared to its single-label counterpart. While FixMatch has garnered widespread acclaim for its efficacy in simplifying multiclass SSL, our work extends this simplicity to the realm of multi-label SSL, filling a crucial gap in the literature.

Furthermore, our approach does not require prior knowledge about label distribution, making it particularly applicable to automatically labeled medical datasets where assumptions about class distributions cannot be derived from labeled data. This aspect is pivotal for ensuring robustness and generalizability in medical image diagnosis tasks.

Overall, our adaptation of FixMatch to SSMLL represents a significant advancement in the field, offering a practical and effective solution for enhancing the efficiency and accuracy of medical image diagnosis.

## References


[1] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *International Conference on Learning Representations*, 2020. 2, 4

[2] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 7

[3] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *ArXiv*, 2017. 6

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *International Conference on Machine Learning*, 34, 2017. 6

[5] Prashnna Kumar Gyawali, Sandesh Ghimire, Pradeep Bajracharya, Zhiyuan Li, and Linwei Wang. Semi-supervised medical image classification with global latent mixing. *Medical Image Computing and Computer Assisted Intervention*, 2020. 7, 8

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEE Conference on Computer Vision and Pattern Recognition*, 2016. 6

[7] Sontje Ihler, Felix Kuhnke, and Svenja Spindeldreier. A comprehensive study of modern architectures and regularization approaches on chexpert5000. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 654–663. Springer, 2022. 4, 5, 6

[8] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 1, 2, 4

[9] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 6

[10] Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20706, 2022. 1, 7, 8

[11] Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE transactions on medical imaging*, 39(11):3429–3440, 2020. 1

[12] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 4

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *Conference on Neural Information Processing Systems*, 33, 2019. 6

[14] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 6

[15] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 7, 8

[16] 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. *Advances in neural information processing systems*, 33:596–608, 2020. 1, 2, 6, 7

[17] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 7

[18] Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagnostic Pathology*, 19, 2024. 2

[19] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 7

[20] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10857–10866, 2021. 2

[21] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *Conference on Neural Information Processing Systems*, 35, 2021. 6

[22] Ming-Kun Xie, Jia-Hao Xiao, Hao-Zhe Liu, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Class-distribution-aware pseudo labeling for semi-supervised multi-label learning. *Neural Information Processing Systems*, 37, 2023. 2, 7

[23] Xuanang Xu, Hannah H. Deng, Jaime Gateno, and Pingkun Yan. Federated multi-organ segmentation with inconsistent

labels. *IEEE Transaction on Medical Imaging*, 42(10):2948–2960, 2023. 7

[24] Qingjie Zeng, Yutong Xie, Zilin Lu, and Yong Xia. Pefat: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15671–15680, 2023. 1, 7, 8

[25] Maximilian Zenk, David Zimmerer, Fabian Isensee, Paul F Jäger, Jakob Wasserthal, and Klaus Maier-Hein. Realistic evaluation of fixmatch on imbalanced medical image classification tasks. In *Bildverarbeitung für die Medizin 2022: Proceedings, German Workshop on Medical Image Computing, Heidelberg, June 26-28, 2022*, pages 291–296. Springer, 2022. 2, 4, 8

[26] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 8

[27] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14471–14481, 2022. 1