Noisy Label Learning for Large-scale Medical Image Classification

Fengbei Liu[1] Yu Tian[1,4] Filipe R. Cordeiro[2] Vasileios Belagiannis[3]
Ian Reid[1] Gustavo Carneiro[1]

[1] Australian Institute for Machine Learning, University of Adelaide
[2] Universidade Federal Rural de Pernambuco, Brazil
[3] Universität Ulm, Germany
[4] South Australian Health and Medical Research Institute

Abstract. The classification accuracy of deep learning models depends not only on the size of their training sets, but also on the quality of their labels. In medical image classification, large-scale datasets are becoming abundant, but their labels will be noisy when they are automatically extracted from radiology reports using natural language processing tools. Given that deep learning models can easily overfit these noisy-label samples, it is important to study training approaches that can handle label noise. In this paper, we adapt a state-of-the-art (SOTA) noisy-label multi-class training approach to learn a multi-label classifier for the dataset Chest X-ray14, which is a large scale dataset known to contain label noise in the training set. Given that this dataset also has label noise in the testing set, we propose a new theoretically sound method to estimate the performance of the model on a hidden clean testing data, given the result on the noisy testing data. Using our clean data performance estimation, we notice that the majority of label noise on Chest X-ray14 is present in the class 'No Finding', which is intuitively correct because this is the most likely class to contain one or more of the 14 diseases due to labelling mistakes.

Keywords: Noisy label learning · Chest X-ray14.

1 Introduction

Recent progress on deep neural networks (DNN) has enabled considerable success in various tasks, including visual classification [8,10] and medical image analysis [13,17,20,24,30,32]. The successful performance of DNNs depends on a massive amount of carefully labelled data produced by human annotators, which is time-consuming and expensive to obtain. An alternative to this manual labelling is based on an automatic labelling produced by search engines [25,28]. Similar methods were proposed in medical image analysis with an automatic extraction of chest disease labels with the application of natural language processing (NLP) tools on the radiologists' reports [32]. Both alternative labelling processes are designed to mitigate the high cost needed to produce expertly annotated data. However, the use of these alternative labelling processes often produces unreliably labelled datasets because search engines can provide incorrect searching results and NLP extracted disease labels without verification by doctors may

contain incorrect labels [22,23]. Even when relying on human experts, labelling medical images is not straightforward because the task can be naturally difficult. For example, radiologists may not be 100% confident in a diagnosis and the label will reflect that issue. These unreliable labels are called *noisy labels* because they do not represent the ground truth labels required by a DNN in supervised learning.

The main issue of noisy labels is that the training of a DNN is known to overfit them, causing a decline in the model accuracy. For instance, Zhang et al. [37] have shown that even with completely random labels, DNNs can fit the entire training dataset with zero error rate, leading to low generalization ability on validation or testing sets. The use of popular regularization techniques does not solve the label noise issues because these techniques were designed to generalise well to a clean testing set, given the training on a clean training set. Hence, methods specifically designed to address label noise [15,17,18] are needed to enable the training of DNNs in such scenario.

The **first contribution** of this paper is to adapt the noisy label multi-class learning method 'early learning regularisation' (ELR) [18] to train a multi-label classifier for the dataset Chest X-ray14 [32], which is a large-scale dataset known to contain label noise in the training set [22,23]. Our work is the first to show the effectiveness of ELR in the noisy multi-label problem of Chest X-ray14 [32]. A limitation of Chest X-ray14 [32] is that not only the training set, but the testing set also has label noise. Therefore, **a second contribution** of this paper is a new theoretically sound method to estimate the classification accuracy on hidden clean data, given the accuracy on the noisy testing set. This method is new in machine learning and enables the use of noisy testing sets, which was not explored by previous noisy label learning papers [15]. Our proposed clean data performance estimation allows us to investigate how label noise affects Chest X-ray14 [32], and we noticed that the majority of noise is present in the class 'NoFinding'. This result has been noticed by other researchers [22,23] and is intuitive because the 'NoFinding' class is the most likely to contain one or more of the 14 diseases because of labelling mistakes.

2 Related Work

2.1 Chest X-ray Classification

Chest X-ray is the most widely used imaging modality for diagnosing chest diseases, and as a result, it attracts a great deal of attention from the medical imaging community. With several large-scale datasets proposed, training a DNN to automatically label diseases from chest X-ray images has become a popular classification task [9,16,21,32]. Wang et al. [32] proposed a DNN to automatically classify diseases from Chest X-ray. Guan et al. [6] used an attention mechanism to improve thorax classification. Other papers [16] use extra bounding box information to improve classification performance. However, some extensive studies [12,22,23] show that NLP labelled Chest X-ray datasets [32] contain a fair proportion of label noise because of NLP inaccuracies. This noise affects not only the training, but also the testing set. Hence, even if one uses a training algorithm that is robust to noisy labels, the results on the testing set do not represent a reliable estimation of the functionality of the model because of the lack of a clean ground truth for testing.

2.2 Noisy Label Learning

Noisy label learning focuses on maintaining model robustness to a large proportion of label noise in the training set. Different approaches can be characterized as: *sample selection* [7, 11, 15], *robust loss function* [4, 18, 33], and *transition matrix based* [5, 34] methods. Sample selection approaches have shown the best results in dealing with label noise. They work by automatically selecting clean samples from the noisy training set and retraining the model with those selected clean samples. The remaining samples are considered noisy, and are either excluded from training or relabeled by the training model. The critical challenge is how to design the mechanism for selecting clean samples. Arpit et al. [1] observed that DNNs are prone to fit clean samples first and overfit to noisy samples after a certain number of epochs. This observation inspired an important selection criterion called *small loss selection*. The criterion treats small training loss samples as clean samples. Han et al. [7] and Yu et al. [35] integrate the small loss idea for training two DNNs, where the updating of the parameters of one DNN depends on the small loss samples from the other DNN. The disagreements between the DNNs help with the discovery of label error and the estimation of robust prediction. Previous methods discard the noisy samples and only retrain with clean samples, but [15] rely on semi-supervised learning where noisy samples have their labels removed. This transforms noisy-label into a typical semi-supervised learning problem with clean samples as a labelled subset and noisy samples as an unlabeled set. However, small loss sample selection assumes a balanced training set, which is not present in multi-label learning problems (such as Chest X-ray14 [32]), where minority class samples can be treated as noisy samples because of their large losses. For this reason, all methods based on small loss selection cannot be directly translated to the multi-label learning. Methods based on robust loss functions and transition matrices can deal better with class imbalances, so we explore them for our multi-label classifier in Sec. 3.

3 Method

We assume that we have a clean label distribution, denoted by $D_{X,Y}$, where $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are random variables for the images and labels, respectively, with $\mathcal{X} \subset \mathbb{R}^{H \times W \times C}$ and $\mathcal{Y} \subset \{0, 1\}^{|\mathcal{Y}|}$. From $D_{X,Y}$, it is possible to sample training images and labels with $(\mathbf{x}, \mathbf{y}) \sim D_{X,Y}$, where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. We also have a noisy label distribution $\tilde{D}_{X,\tilde{Y}}$ (with $X, \tilde{Y} \in \mathcal{X} \times \mathcal{Y}$) that generates image samples and noisy labels with $(\mathbf{x}, \tilde{\mathbf{y}}) \sim \tilde{D}_{X,\tilde{Y}}$. The noisy training set is denoted by $\tilde{\mathcal{S}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^{|\tilde{\mathcal{S}}|}$ and the noisy testing set is represented by $\tilde{\mathcal{V}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^{|\tilde{\mathcal{V}}|}$, where in both sets, we have $(\mathbf{x}, \tilde{\mathbf{y}}) \sim \tilde{D}_{X,\tilde{Y}}$.

3.1 Estimating Classification on a Hidden Clean Distribution

In a noisy label problem, we typically have the noisy training and testing sets to model and test our classifier, denoted by $h : \mathcal{X} \rightarrow \mathcal{Y}$, where $h \in \mathcal{H}$ (i.e., $\mathcal{H}$ is the space of classifiers). However, an important question is what is the performance of $h(.)$ in

samples from the clean distribution $D_{X,Y}$? Chen et al. [3] recently proposed theoretical results to show that the maximisation of the training accuracy using $\tilde{S}$ produces an optimal classifier and that model selection using a noisy validation set is reliable. In this section, we use these results to derive a classification accuracy lower bound on a hidden clean distribution, given the classification accuracy on a noisy testing set.

Given that the results in [3] apply to multi-class problems (i.e. one label, multiple possible values for the label), we divide our multi-label problem into $|\mathcal{Y}|$ such problems, where for each problem, there are only two possible labels, 0 and 1. Similarly to [3], we assume that X is independent of $\tilde{Y}$ given Y, which implies that a noise transition matrix $\mathbf{T} \in [0,1]^{2\times 2}$ is used to corrupt Y into $\tilde{Y}$, with $\mathbf{T}_{kl} = P(\tilde{Y} = l|Y = k)$. The classification accuracy of $h \in \mathcal{H}$ on the clean distribution is defined by

$$A_D(h) = \mathbb{E}_{(X,Y)\sim D}[\mathbf{1}(h(X) = Y)], \tag{1}$$

where $\mathbf{1}(.)$ is the indicator function, and $A_{\tilde{D}}(h)$ is similarly defined. The classification accuracy of h on the noisy testing set $\tilde{\mathcal{V}}$ is defined as

$$A_{\tilde{V}}(h) = \frac{1}{|\tilde{V}|}\sum_{i=1}^{|\tilde{V}|}\mathbf{1}(h(\mathbf{x}_i) = \tilde{\mathbf{y}}_i)), \tag{2}$$

with the classification accuracy on the training set $A_{\tilde{S}}(h)$ similarly defined.

Theorem 1. *Considering training a classifier $h \in \mathcal{H}$ on noisy training set $\tilde{S}$, then with probability at least $1-\delta$ (for $\delta \in (0,1]$),*

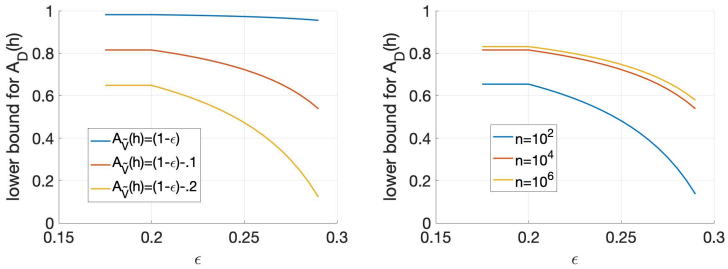
$$A_D(h) \geq 1 + \frac{1}{\tau}\left(-1 + \epsilon + A_{\tilde{V}}(h) - \sqrt{\frac{\ln(1/\delta)}{2|\tilde{\mathcal{V}}|}}\right), \tag{3}$$

where $\tau = \min_{k,l\in\{0,1\},k\neq l}(\mathbf{T}_{kk} - \mathbf{T}_{kl})$, and the noise rate defined by $\epsilon = 1 - \sum_{k=0}^{1} P(Y = k)\mathbf{T}_{kk}$, with $P(Y = k)$ denoting the prior of class k.

Proof. Please see proof in the supplementary material.

Theorem 1 shows that a lower bound to the classification accuracy on the clean distribution set $A_D(h)$ depends on the accuracy on the noisy test set $A_{\tilde{V}}(h)$, the arbitrary value of $\delta \in (0,1]$, the size of the testing set $|\tilde{\mathcal{V}}|$, the noise rate ϵ, and on the noise transition matrix $\mathbf{T}$. More specifically, from [3], we have $\left(A_{\tilde{V}}(h) - \sqrt{\frac{\ln(1/\delta)}{2|\tilde{\mathcal{V}}|}}\right) \leq (1-\epsilon)$, which means that $\left(-1 + \epsilon + A_{\tilde{V}}(h) - \sqrt{\frac{\ln(1/\delta)}{2|\tilde{\mathcal{V}}|}}\right) \leq 0$, and consequently, $A_D(h) \leq 1$. Hence, τ reduces when ϵ increases, which in turn will reduce the lower bound for $A_D(h)$ – this is shown in Fig. 1-(a). Also, increasing the test set size $|\tilde{\mathcal{V}}|$ will make $\left(A_{\tilde{V}}(h) - \sqrt{\frac{\ln(1/\delta)}{2|\tilde{\mathcal{V}}|}}\right) \rightarrow (1-\epsilon)$, and consequently increase the lower bound for $A_D(h)$, as shown in Fig. 1-(b).

Given that both ϵ and $\mathbf{T}$ are unavailable, we estimate them with the method to estimate a transition matrix proposed in [34]. We use the transition matrix estimation from [34] because it does not need a set of clean samples, which is normally assumed by previous transition matrix estimators [19].



(a) Varying $A_{\tilde{\mathcal{V}}}(h)$ (b) Varying testing set size $n = |\tilde{\mathcal{V}}|$

Fig. 1: Lower bound for $A_D(h)$, computed from (3), as a function of ϵ for (a) different values of $A_{\tilde{\mathcal{V}}}(h) \leq 1 - \epsilon$, assuming $|\tilde{\mathcal{V}}| = 10^4$; and (b) different testing set sizes $n = |\tilde{\mathcal{V}}|$, assuming $A_{\tilde{\mathcal{V}}}(h) = (1 - \epsilon) - 0.1$.

3.2 Early Learning Regularisation

We adapt the Early Learning Regularisation (ELR) algorithm by Liu et al. [18], which is based on a regularised cross-entropy loss that boosts the gradient of samples that are estimated to contain clean labels and dampens the gradient of samples estimated to contain noisy labels. Assuming that the classifier $h \in \mathcal{H}$ is represented by a DNN $f_\theta : \mathcal{X} \to [0,1]^{|\mathcal{Y}|}$, the ELR loss is defined by:

$$\ell_{ELR}(\theta, \tilde{\mathcal{S}}) = \frac{1}{|\tilde{\mathcal{S}}|} \sum_{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \tilde{\mathcal{S}}} \left(\ell_{CE}(\mathbf{p}_i, \tilde{\mathbf{y}}_i) + \lambda \log \left(1 - \mathbf{t}_i^\top \mathbf{p}_i\right)\right), \tag{4}$$

where $\ell_{CE}(\mathbf{p}_i, \tilde{\mathbf{y}}_i) = -\tilde{\mathbf{y}}_i^\top \log \mathbf{p}_i$, $\mathbf{p}_i = f_\theta(\mathbf{x}_i)$ (with $\mathbf{p}_i = [0,1]^{|\mathcal{Y}|}$) represents the model output, $\mathbf{t}_i = [0,1]^{|\mathcal{Y}|}$ denotes a target vector of probabilities of the presence of each label in $\mathcal{Y}$ for sample i, which is computed using a moving average from past model outputs. The training of the classifier $f_\theta(.)$ using the ELR loss in (4) is empirically shown to produce outstanding results on multi-class noisy label datasets. Our paper is the first to apply this loss to a multi-label problem, where the main differences compared to the original multi-class formulation [18] are: 1) both the model output $\mathbf{p}_i$ and moving target $\mathbf{t}_i$ do not need to sum to one – that is, they represent $|\mathcal{Y}|$ binary classification problems, so $\sum_{c=1}^{|\mathcal{Y}|} \mathbf{p}_i(c) \in [0, |\mathcal{Y}|]$ (similarly for $\mathbf{t}_i$); and 2) we modify the MixUp [38] operation in ELR (more specifically the ELR$_+$ [18]) to mix only the estimated target from early model prediction, but not the images, when calculating ELR regularization – this is done based on the observation in [31] that suggests that MixUp is not suitable for multi-label classification.

4 Experiment

4.1 Dataset

We evaluate our method on the Chest X-ray14 dataset [32] that contains 112,120 frontal-view chest X-ray images with 14 pathologies. Each X-ray can be labelled with

multiple diseases, making it a multi-label problem. For the train/val/test split, we follow official training and testing splits [32]. We further split 10% of the training images (at the patient level) for validation to estimate hyper-parameters and select models.

4.2 Implementation Details

The backbone model for our classifier $f_{\theta(.)}$ from Sec. 3.2 is the DenseNet121 [10]. We use Adam [14] optimizer with batch size 16. Images are resized from original 1024 $\times$ 1024 to 512$\times$512 and normalized with ImageNet [27] mean and standard deviation. For data augmentation, we employ random resized crop, 5-degree random rotation and random horizontal flipping. The network is initialized with ImageNet [27] pre-trained weights. For metric evaluation, we follow [32,36] and use AUC for each pathology and average AUC for all 14 labels. Because AUC varies dramatically on different dataset splits, we only consider methods that follow the official split [32]. To compute the bound results from (3), we consider 15 binary classification accuracies, where we use the labels from the 14 pathologies, and also the label 'NoFinding'. For computing this accuracy, we consider that the classifier labels a test image $\mathbf{x}$ with a particular disease $c \in \mathcal{Y}$ when $\mathbf{p}(c) \geq 0.5$ (with $\mathbf{p} = f_\theta(\mathbf{x})$), and if $\mathbf{p}(c) < 0.5$ for all $c \in \mathcal{Y}$, then the image is labelled with 'NoFinding'.

4.3 Experimental Setup

We compare our adaptation of the ELR training (in Sec. 3.2) using the DenseNet121 model, with the following methods. Wang et al. [32] presented the baseline methods, along with the original Chest X-ray dataset. Li et al. [16] combine local and global models to classify images into the Chest X-ray14 labels. CheXNet [24] is a weakly supervised DenseNet121 model to classify the 14 diseases. The category-wise residual attention learning (CRAL) [6] framework handles the multi-label problem using an attention module that explores the relationships between labels. Ma et al. [21] proposed a cross-attention network to deal with multi-label imbalanced weakly-supervised learning problems. The current SOTA for Chest X-ray14 is the model by Hermoza et al. [9] that is a weakly supervised disease classifier that combines region proposal and saliency detection. We also show results from DivideMix [15], which uses a noisy label learning algorithm based on small loss sample selection and semi-supervised learning – DivideMix has the SOTA results in noisy-label learning benchmarks. All methods above use the official train/val/test split [32]

4.4 Experimental Results

The comparison with the methods in [6,9,15,16,21,24,32] are presented in Table 1. Note that for [16] we take result tested on official test split with DenseNet121 [10]. Our method improves by 1.6% over the previous SOTA [9] in terms of the mean AUC. The result from DivideMix [15] is worse than most baselines because it relies on small-loss sample selection scheme to distinguish noisy from clean samples. However, a small loss on Chest X-ray14 can also represent a hard sample because it is a severely imbalanced dataset. Only a small fraction of samples includes Fibrosis, Pneumonia and

Method	Wang et al. [32]	Li et al. [16]	CheXNet [24]	CRAL [6]	Ma et al. [21]	Hermoza et al. [9]	DivideMix [15]	Ours
Atelectasis	70.00	72.90	75.50	78.10	77.70	77.50	72.00	**79.44**
Cardiomegaly	81.00	84.60	86.70	88.30	**89.40**	88.10	86.0	88.42
Effusion	75.90	78.10	81.50	83.10	82.90	83.10	80.12	**86.16**
Infiltration	66.10	67.30	69.40	69.70	69.60	69.50	61.93	**71.38**
Mass	69.30	74.30	80.20	83.00	83.80	82.60	75.30	**84.63**
Nodule	66.90	75.80	73.50	76.40	77.10	78.90	68.50	**81.36**
Pneumonia	65.80	63.30	69.80	72.50	72.20	74.10	57.60	**76.66**
Pneumothorax	79.90	79.30	82.80	86.60	86.20	87.90	76.70	**87.86**
Consolidation	70.30	72.00	72.20	75.80	750	74.70	63.70	**79.37**
Edema	80.50	71.00	83.50	85.30	84.60	84.60	80.40	**86.08**
Emphysema	83.30	75.10	85.60	91.10	90.80	93.60	87.20	**94.15**
Fibrosis	78.60	76.10	80.30	82.60	82.70	83.30	73.40	**84.70**
Pleural Thicken	68.40	73.00	74.90	780	77.90	79.30	73.20	**80.23**
Hernia	87.20	66.80	89.40	91.80	**93.40**	91.70	87.70	92.00
Mean	74.50	73.90	78.90	81.60	81.70	82.10	74.50	**83.74**

Table 1: Class-level AUC of our method and other supervised SOTA approaches.

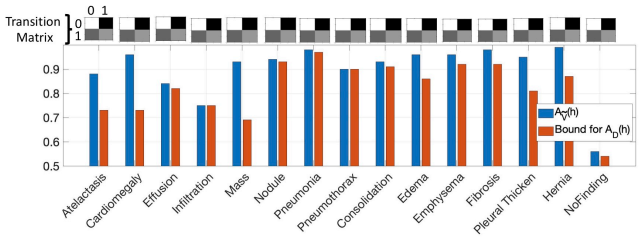


Fig. 2: Class-level accuracy on noisy test set ($A_{\tilde{V}}(h)$) and estimated lower bound accuracy on clean test set ($A_D(h)$), computed as described in (1). On top, we show the estimated transition matrix for each class.

these samples will be discard during DivideMix small loss selection process. In addition, DivideMix utilises the SOTA semi-supervised method MixMatch [2], but a recent work [26] claims that MixMatch does not perform better than Mean-Teacher [29] in multi-label setup. Liu et al.'s approach [18] uses Mean-Teacher for target estimation and does not rely on small-loss selection.

We also estimate the lower bound accuracy on a hidden clean distribution given the classification accuracy computed in the noisy testing set in Fig. 2. We first notice that the lower bound accuracy on a hidden clean distribution is usually lower than the accuracy computed from the noisy test set. According to the results in Sec. 3.1, this happens when the noisy testing set accuracy is smaller than $1-\epsilon$, where the gap can be explained by the term $\sqrt{\frac{\ln(1/\delta)}{2|\tilde{\mathcal{V}}|}}$. Another important point is the low value for the bound for class 'NoFinding', implying that this class is relatively more noisy than the other classes representing the pathologies. Oaken studied this dataset in [22,23] and mentioned that the 'NoFinding' label may include diseases, either from one of 14 pathologies or from additional diseases not present in the label set. This observation implies that 'NoFinding' is likely to be more noisy than the pathology labels, which explains the result in Fig. 2.

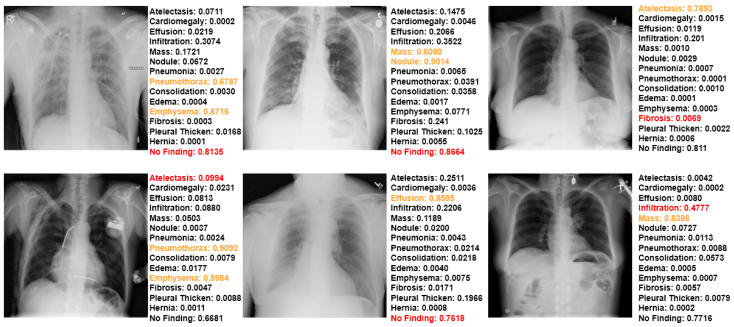


Fig. 3: Chest X-ray14 images with model prediction and original noisy ground truth. All images are extracted from [22,23] and confirmed to be noisy. Red is original noisy label and orange is model prediction.

In [22,23], Oakden identified several images that are certain to contain label noise, so we validate our model on these images to confirm that our model does not overfit the noisy label samples. In Fig. 3, we show the results from our model prediction on the noisy label images from [22,23]. These results show that our model can correctly detect noisy labels from the original annotation of the Chest X-ray14 [32].

5 Conclusion

In this paper, we argue that certain medical image classification problems, such as the classification on Chest X-ray14 [32], must use noisy-label learning methods because they contain noise in their labels due to mistakes in the automatic extraction of labels from NLP algorithms, or the difficulty of the classification task. We proposed an adaptation of the ELR noisy-label learning method [18] to work for the multi-label problem of Chest X-ray14 [32]. Given the lack of clean testing set to assess the performance of this method, we also proposed a theoretically sound bound for the classification accuracy on a hidden clean distribution, given the classification accuracy on the noisy testing set. Comparisons with the SOTA on Chest X-ray14 [32] show that our noisy-label learning is substantially more accurate, indicating the advantage of using noisy-label learning for this problem. The results related to the classification bound show that the optimal classifier result from the noisy testing set has a gap that can be explained by the theory in Sec. 3.1, and that the 'NoFinding' class is the one with the most noise in the dataset, confirming the result in [22,23].

References

1. Arpit, D., et al.: A closer look at memorization in deep networks. arXiv preprint arXiv:1706.05394 (2017)
2. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems. pp. 5049–5059 (2019)
3. Chen, P., Ye, J., Chen, G., Zhao, J., Heng, P.A.: Robustness of accuracy metric and its inspirations in learning with noisy labels. arXiv preprint arXiv:2012.04193 (2020)
4. Ghosh, A., Kumar, H., Sastry, P.: Robust loss functions under label noise for deep neural networks. arXiv preprint arXiv:1712.09482 (2017)
5. Goldberger, J., Ben-Reuven, E.: Training deep neural-networks using a noise adaptation layer (2016)
6. Guan, Q., Huang, Y.: Multi-label chest x-ray image classification via category-wise residual attention learning. Pattern Recognition Letters **130**, 259–266 (2020)
7. Han, B., et al.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Advances in neural information processing systems. pp. 8527–8537 (2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
9. Hermoza, R., et al.: Region proposals for saliency map refinement for weakly-supervised disease localisation and classification. arXiv preprint arXiv:2005.10550 (2020)
10. Huang, G., et al.: Densely connected convolutional networks. In: CVPR. pp. 4700–4708 (2017)
11. Huang, J., et al.: O2u-net: A simple noisy label detection approach for deep neural networks. In: ICCV. pp. 3326–3334 (2019)
12. Irvin, J., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI. vol. 33, pp. 590–597 (2019)
13. Jonmohamadi, Y., et al.: Automatic segmentation of multiple structures in knee arthroscopy using deep learning. IEEE Access **8**, 51853–51861 (2020)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394 (2020)
16. Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.J., Fei-Fei, L.: Thoracic disease identification and localization with limited supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8290–8299 (2018)
17. Liu, F., et al.: Self-supervised depth estimation to regularise semantic segmentation in knee arthroscopy. In: MICCAI. pp. 594–603. Springer (2020)
18. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. arXiv preprint arXiv:2007.00151 (2020)
19. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence **38**(3), 447–461 (2015)
20. Liu, Y., Tian, Y., Maicas, G., Pu, L.Z.C.T., Singh, R., Verjans, J.W., Carneiro, G.: Photoshopping colonoscopy video frames. In: ISBI. pp. 1–5. IEEE (2020)
21. Ma, C., et al.: Multi-label thoracic disease image classification with cross-attention networks. In: MICCAI. pp. 730–738. Springer (2019)
22. Oakden-Rayner, L.: Exploring the chestxray14 dataset: problems. Wordpress: Luke Oakden Rayner (2017)
23. Oakden-Rayner, L.: Exploring large-scale public medical image datasets. Academic radiology **27**(1), 106–112 (2020)

24. Rajpurkar, P., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
25. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. arXiv preprint arXiv:1803.09050 (2018)
26. Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M.: In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. arXiv preprint arXiv:2101.06329 (2021)
27. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
28. Shen, Y., Ji, R., Chen, Z., Hong, X., Zheng, F., Liu, J., Xu, M., Tian, Q.: Noise-aware fully webly supervised object detection. In: CVPR. pp. 11326–11335 (2020)
29. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems. pp. 1195–1204 (2017)
30. Tian, Y., Maicas, G., Pu, L.Z.C.T., Singh, R., Verjans, J.W., Carneiro, G.: Few-shot anomaly detection for polyp frames from colonoscopy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 274–284. Springer (2020)
31. Wang, Q., Jia, N., Breckon, T.P.: A baseline for multi-label image classification using an ensemble of deep convolutional neural networks. In: ICIP. pp. 644–648. IEEE (2019)
32. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR. pp. 2097–2106 (2017)
33. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: ICCV. pp. 322–330 (2019)
34. Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., Sugiyama, M.: Are anchor points really indispensable in label-noise learning? arXiv preprint arXiv:1906.00189 (2019)
35. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I.W., Sugiyama, M.: How does disagreement help generalization against label corruption? arXiv preprint arXiv:1901.04215 (2019)
36. Zhang, C., Chen, F., Chen, Y.Y.: Thoracic disease identification and localization using distance learning and region verification. arXiv preprint arXiv:2006.04203 (2020)
37. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530 (2016)
38. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)