Learning Robust Classifier for Imbalanced Medical Image Dataset with Noisy Labels by Minimizing Invariant Risk

Jinpeng Li[1], Hanqun Cao[1], Jiaze Wang[1], Furui Liu[3], Qi Dou[1,2], Guangyong Chen[3(✉)], and Pheng-Ann Heng[1,2]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong

[2] Institute of Medical Intelligence and XR, The Chinese University of Hong Kong, Shatin, Hong Kong

[3] Zhejiang Lab, Hangzhou, China

gychen@zhejianglab.com

Abstract. In medical image analysis, imbalanced noisy dataset classification poses a long-standing and critical problem since clinical large-scale datasets often attain noisy labels and imbalanced distributions through annotation and collection. Current approaches addressing noisy labels and long-tailed distributions separately may negatively impact real-world practices. Additionally, the factor of class hardness hindering label noise removal remains undiscovered, causing a critical necessity for an approach to enhance the classification performance of noisy imbalanced medical datasets with various class hardness. To address this paradox, we propose a robust classifier that trains on a multi-stage noise removal framework, which jointly rectifies the adverse effects of label noise, imbalanced distribution, and class hardness. The proposed noise removal framework consists of multiple phases. Multi-Environment Risk Minimization (MER) strategy captures data-to-label causal features for noise identification, and the Rescaling Class-aware Gaussian Mixture Modeling (RCGM) learns class-invariant detection mappings for noise removal. Extensive experiments on two imbalanced noisy clinical datasets demonstrate the capability and potential of our method for boosting the performance of medical image classification.

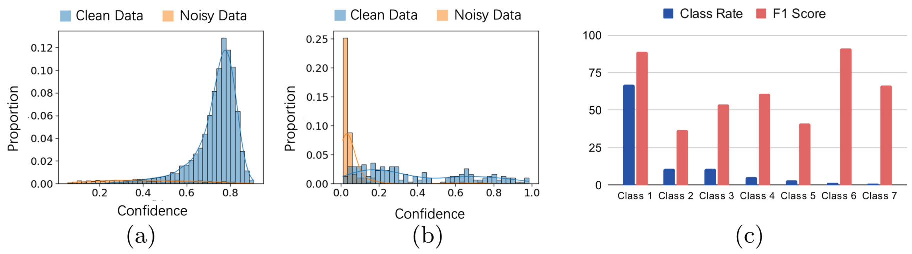
Keywords: Imbalanced Data · Noisy Labels · Medical Image Analysis

1 Introduction

Image classification is a significant challenge in medical image analysis. Although some classification methods achieve promising performance on balanced and clean medical datasets, balanced datasets with high-accuracy annotations are

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_30.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 306–316, 2023.
https://doi.org/10.1007/978-3-031-43987-2_30



(a) (b) (c)

Fig. 1. Analysis of confidence distributions and class hardness on imbalanced noisy HAM10000 dataset [22]. (a) and (b) are confidence distributions of clean and noisy data on the majority class and the minority class, respectively (c) is the relationship between class rate and F1 score among different classes.

time-consuming and expensive. Besides, pruning clean and balanced datasets require a large amount of crucial clinical data, which is insufficient for large-scale deep learning. Therefore, we focus on a more practical yet unexplored setting for handling imbalanced medical data with noisy labels, utilizing all available low-cost data with possible noisy annotations. Noisy imbalanced datasets arise due to the lack of high-quality annotations [11] and skewed data distributions [18] where the number of instances largely varies across different classes. Besides, the class hardness problem where classification difficulties vary for different categories presents another challenge in removing label noise. Due to differences in disease epidemicity and collection difficulty, rare anomalies or anatomical features render diseases with low epidemicity easier to detect. However, existing techniques [12,23,24] fail to jointly address these scenarios, leading to inadequate classification outcomes. Therefore, noisy-labeled, imbalanced datasets with various class hardness remain a persistent challenge in medical classification.

Existing approaches for non-ideal medical image classification can be summarized into noisy classification, imbalanced recognition, and noisy imbalanced identification. Noisy classification approaches [3,7,23] conduct noise-invariant learning depending on the big-loss hypothesis, where classifiers trained with clean data with lower empirical loss aid with de-noising identification. However, imbalanced data creates different confidence distributions of clean and noisy data in the majority class and minority class as shown in Fig. 1, which invalidates the big-loss assumption [3,4]. Imbalanced recognition approaches [9,15,21] utilize augmented embeddings and imbalance-invariant training loss to re-balance the long-tailed medical data artificially, but the disturbance from noisy labels leads to uncasual feature learning, impeding the recognition of tail classes. Noisy long-tailed identification technique [25] has achieved promising results by addressing noise and imbalance concerns sequentially. However, the class hardness problem leads to vague decision boundaries that hinders accurate‘ noise identification.

In this work, we propose a multi-stage noise removal framework to address these concerns jointly. The main contributions of our work include: 1) We decompose the negative effects in practical medical image classification, 2) We minimize the invariant risk to tackle noise identification influenced by multiple factors, enabling the classifier to learn causal features and be distribution-invariant, 3) A re-scaling class-aware Gaussian Mixture Modeling (CGMM) approach is proposed to distinguish noise labels under various class hardness, 4) We evaluate our method on two medical image datasets, and conduct thorough ablation studies to demonstrate our approach's effectiveness.

2 Method

2.1 Problem Formulation

In the noisy imbalanced classification setting, we denote a medical dataset as $\{(x_i, y_i)\}_{i=1}^{N}$ where y_i is the corresponding label of data x_i and N is the total amount of instances. Here y_i may be noisy. Further, we split the dataset according to class categories. Then, we have $\{\mathcal{D}_j\}_{j=1}^{M}$, where M is the number of classes; $\mathcal{D}_j$ denotes the subset for class j. In each subset containing N_j samples, the data pairs are expressed as $\{(x_i^j, y_i^j)\}_{i=1}^{N_j}$. Without loss of generality, we order the classes as $N_1 > N_2 > ... > N_{M-1} > N_M$. Further, we denote the backbone as $\mathcal{H}(\cdot;\theta), \mathcal{X} \rightarrow \mathcal{Z}$ mapping data manifold to the latent manifold, the classifier head as $\mathcal{G}(\cdot;\gamma), \mathcal{Z} \rightarrow \mathcal{C}$ linking latent space to the category logit space, and the identifier as $\mathcal{F}(\cdot;\phi), \mathcal{Z} \rightarrow \mathcal{C}$. We aim to train a robust medical image classification model composed of a representation backbone and a classifier head on label noise and imbalance distribution, resulting in a minimized loss on the testing dataset:

$$\min \sum_{i \in D} L(\mathcal{G}\left[\mathcal{H}(x_i^{test};\theta);\gamma\right], y_i^{test}) \tag{1}$$

2.2 Mapping Correction Decomposition

We decompose the non-linear mapping $p(y = c|x)$ as a product of two space mappings $p_{\mathcal{G}}(y = c|z) \cdot p_{\mathcal{H}}(z|x)$. Given that backbone mapping is independent of noisy imbalanced effects, we conduct further disentanglement by defining e as the negative effects and $\mathcal{P}$ as constant for fixed probability mappings:

$$\begin{aligned} p(y=c|x,e) &= p_{\mathcal{H}}(z|x) \cdot p_{\mathcal{G}}(y=c|z,e) \\ &= p_{\mathcal{H}}(z|x) \cdot \{p_{\mathcal{G}}(y=c|z) \cdot p_{\mathcal{G}}(y=c|e)\} \\ &= p_{\mathcal{H}}(z|x) \cdot p_{\mathcal{G}}(y=c|z) \cdot \{p_{\mathcal{G}}(y=c|[e_i, e_n, e_m])\} \\ &= \mathcal{P} \cdot \overbrace{p_{\mathcal{G}}(e_i|y=c)}^{\textbf{Imbalance}} \overbrace{p_{\mathcal{G}}(e_m|y=c,e_i)}^{\textbf{Hardness}} \overbrace{p_{\mathcal{G}}(e_n|y=c,e_i,e_m)}^{\textbf{Noise}} \end{aligned} \tag{2}$$

The induction derives from the assumption that the incorrect mapping $p_{\mathcal{G}}(y = c|z, e)$ conditions on both pure latent to logits mapping $p_{\mathcal{G}}(y = c|z)$ and adverse effects $p_{\mathcal{G}}(y = c|e)$. By Bayes theorem, we decompose the effect into imbalance,

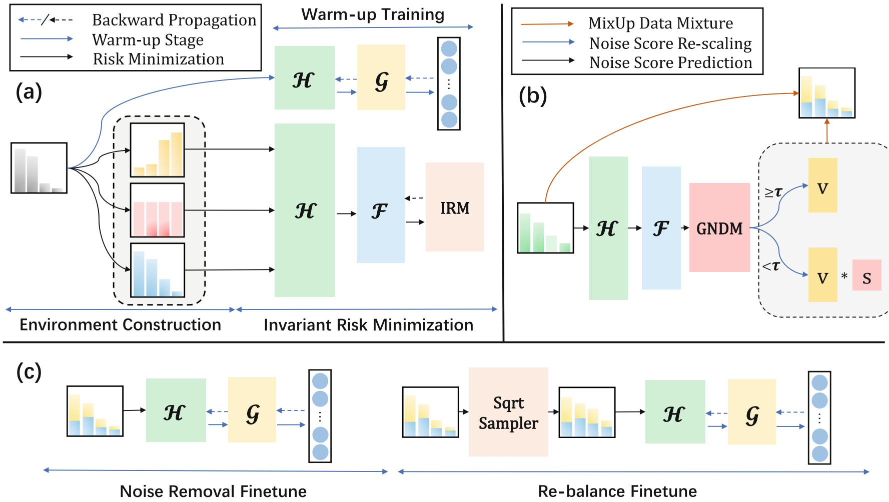


Fig. 2. Protocol for Noisy Long-tailed Recognition: (a) shows warm-up and MER schemes. Backbone $\mathcal{H}$ and classifier $\mathcal{G}$ are first trained in the warm-up phase. Noise identifier $\mathcal{F}$ is optimized across three constructed environments with $\mathcal{H}$ fixed during MER. (b) represents RCGM scheme for class-aware noise detection and score re-scaling. (c) displays final fine-tuning procedures including noise removal finetune and re-balanced finetune.

noise, and mode (hardness), where the noise effect depends on skew distribution and hardness effect; and the hardness effect is noise-invariant.

Currently, noise removal methods only address pure noise effects ($p_{\mathcal{G}}(e_n|y = c)$), while imbalance recognition methods can only resolve imbalanced distribution, which hinders the co-removal of adverse influences. Furthermore, the impact of hardness effects has not been considered in previous studies, which adds an extra dimension to noise removal. In essence, the fundamental idea of noisy classification involves utilizing clean data for classifier training, which determines the importance of noise identification and removal. To address these issues, we propose a mapping correction approach that combines independent noise detection and removal techniques to identify and remove noise effectively.

2.3 Minimizing Invariant Risk Across Multi-distributions

Traditional learning with noisy label methods mainly minimize empirical risk on training data. However, they fail to consider the influence of imbalanced distributions, which might cause a biased gradient direction on the optimization subspace. Following [25], we minimize the invariant risk [2] across multi-environment for independent detector learning. By assuming that the robust classifier performs well on every data distribution, we solve the optimizing object by finding the optima to reduce the averaged distance for gradient shift:

$$\begin{gathered}\min_{\substack{\mathcal{H}_\theta:\mathcal{X}\to\mathcal{Z}\\ \mathcal{F}_\phi:\mathcal{Z}\to\mathcal{Y}}}\sum_{\varepsilon\in\mathcal{E}_{tr}}\mathcal{L}(\mathcal{F}_\phi\circ\mathcal{H}_\theta)\\ \text{s.t. } \mathcal{F}_\phi\in\arg\min_{\bar{\mathcal{F}}_\phi:\mathcal{Z}\to\mathcal{Y}}\mathcal{L}(\bar{\mathcal{F}}_\phi\circ\mathcal{H}_\theta),\quad\forall\varepsilon\in\mathcal{E}_{tr},\end{gathered}\tag{3}$$

where ε represents an environment (distribution) for classifier $\mathcal{F}_\phi$ and backbone $\mathcal{H}_\theta$; and $\mathcal{L}$ denotes the empirical loss for classification. Since the incorrect mapping is not caused by feature representation, the backbone $\mathcal{H}_\theta$ is fixed during the optimization. By transferring the constraints into a penalty in the optimizing object, we solve this problem by learning the constraint scale ω [2]:

$$\min_{\mathcal{F}}\sum_{\epsilon}\mathcal{L}(\mathcal{F}\circ\mathcal{H})+\left\|\nabla_{w|w=1}\mathcal{L}(w\cdot\mathcal{F}\circ\mathcal{H})\right\|_2^2.\tag{4}$$

Ideally, the noise removal process is distribution-invariant if data is uniformly distributed w.r.t. classes. By the law of large numbers, all constructed distributions should be symmetric according to the balanced distribution to obtain a uniform expectation. To simplify this assumption, we construct three different data distributions [25] composed of one uniform distribution and two symmetric skewed distributions instead of theoretical settings. In practice, all environments are established from the training set with the same class categories.

2.4 Rescaling Class-Aware Gaussian Mixture

Existing noise labels learning methods [1,13] cluster all sample loss or confidence scores with Beta Mixture Model or Gaussian Mixture Model into noisy and clean distributions. From the perspective of clustering, definite and immense gaps between two congregate groups contribute to more accurate decisions. However, in medical image analysis, an overlooked mismatch exists between class hardness and difficulty in noise identification. This results in ineffectiveness of global cluster methods in detecting label noises across all categories. To resolve the challenge, we propose a novel method called rescaling class-aware Gaussian Mixture Modeling (RCGM) which clusters each category data independently by fitting confidence scores q_{ij} from ith class into two Gaussian distributions as $p_i^n(x^n|\mu^n,\Sigma^n)$ and $p_i^c(x^c|\mu^c,\Sigma^c)$. The mixed Gaussian $p_i^M(\cdot)$ is obtained by linear combinations α_{ik} for each distribution:

$$p_i^M(q_{ij}):=\sum_{k\in\{c,n\}}\alpha_{ik}p_i^k\left(q_{ij}\mid\mu_i^k,\Sigma_i^k\right),\tag{5}$$

which produces more accurate and independent measurements of label quality. Rather than relying on the assumption that confidence distributions of training samples depend solely on their label quality, RCGM solves the effect of class hardness in noisy detection by individually clustering the scores in each category. This overcomes the limitations of global clustering methods and significantly enhances the accuracy of noise identification even when class hardness varies.

Instead of assigning a hard label to the potential noisy data as [8] which also employs a class-specific GMM to cluster the uncertainty, we further re-scale the confidence score of class-wise noisy data. Let x_{ij} be the jth in class i, then its probability of having a clean label is:

$$\gamma_{ij} = \frac{\alpha_{ik} p_i^c \left(q_{ij} \mid \mu_i^c, \Sigma_i^c\right)}{p_i^M \left(q_{ij}\right)}, \tag{6}$$

which is then multiplied by a hyperparameter s if the instance is predicted as noise to reduce its weight in the finetuning. With a pre-defined noise selection threshold as τ, we have the final clean score as:

$$v(x_{ij}) := \begin{cases} \gamma_{ij} & \text{if } \gamma_{ij} \geq \tau \\ s \cdot \gamma_{ij} & \text{if } \gamma_{ij} < \tau \end{cases} \tag{7}$$

2.5 Overall Learning Framework for Imbalanced and Noisy Data

In contrast to two-stage noise removal and imbalance classification techniques, our approach applies a multi-stage protocol: warm-up phases, noise removal phases, and fine-tuning phases as shown in Fig. 2. In the warm-up stage, we train backbone $\mathcal{H}$ and classifier $\mathcal{G}$ a few epochs by assuming that $\mathcal{G}$ only remembers clean images with less empirical loss. In the noise removal phases, we learn class-invariant probability distributions of noisy-label effect with MER and remove class hardness impact with RCGM. Finally, in the fine-tuning phases, we apply MixUp technique [13,25,26] to rebuild a hybrid distribution from noisy pairs and clean pairs by:

$$\begin{aligned} \hat{x}_{kl} &:= \alpha_{kl} x_k + (1-\alpha_{kl}) x_l, \quad \forall x_k, x_l \in \mathcal{D} \\ \hat{y}_{kl} &:= \alpha_{kl} y_k + (1-\alpha_{kl}) y_l, \quad \forall y_k, y_l \in \mathcal{D} \end{aligned} \tag{8}$$

where $\alpha_{kl} := \frac{v(x_k)}{v(x_l)}$ denotes the balanced scale; and $\{(\hat{x}_{kl}, \hat{y}_{kl})\}$ are the mixed clean data for classifier fine-tuning. Sqrt sampler is applied to re-balance the data, and cross-stage KL [12] and CE loss are the fine-tuning loss functions.

3 Experiment

3.1 Dataset and Evaluation Metric

We evaluated our approach on two medical image datasets with imbalanced class distributions and noisy labels. The first dataset, HAM10000 [22], is a dermatoscopic image dataset for skin-lesion classification with 10,015 images divided into seven categories. It contains a training set with 7,007 images, a validation set with 1,003 images, and a testing set with 2,005 images. Following the previous noisy label settings [25], we add 20% noise to its training set by randomly flipping labels. The second dataset, CHAOYANG [29], is a histopathology image dataset manually annotated into four cancer categories by three pathological

Table 1. Quantitative comparisons with state-of-the-art methods on HAM10000 and CHAOYANG datasets. The second-best performances are underlined.

Method	HAM10000			CHAOYANG		
	Macro-F1	B-ACC	MCC	Macro-F1	B-ACC	MCC
Focal Loss [14]	66.16	65.21	59.86	70.84	69.10	70.27
Sqrt-RS [17]	67.02	66.28	55.09	70.39	68.56	69.71
PG-RS [10]	62.91	63.29	51.14	71.03	69.35	70.18
CB-Focal [5]	63.41	72.63	52.21	73.20	72.46	71.37
EQL [21]	60.94	66.18	55.53	71.09	70.53	70.77
EQL V2 [20]	58.33	54.70	52.01	69.24	68.35	67.78
CECE [5]	40.92	56.75	37.46	47.12	47.56	43.50
CLAS [28]	69.61	70.85	63.67	71.91	71.46	70.71
FCD [12]	71.08	72.85	66.58	71.82	70.07	71.76
DivideMix [13]	69.72	70.84	65.33	50.44	49.34	50.29
NL [16]	44.42	42.52	55.81	71.75	69.99	71.63
NL+Sqrt-RS	62.46	61.42	52.44	71.77	70.88	71.46
GCE [27]	50.47	48.91	63.63	21.04	28.12	4.13
GCE+Sqrt-RS	70.81	70.76	65.86	70.83	69.77	70.21
GCE+Focal	66.19	68.71	61.82	72.91	71.25	72.68
Co-Learning [19]	58.05	51.02	57.73	60.73	59.78	64.13
H2E [25]	69.69	69.11	63.48	69.36	67.89	68.59
Ours	**76.43**	**75.60**	**70.19**	**74.50**	**72.75**	**73.08**

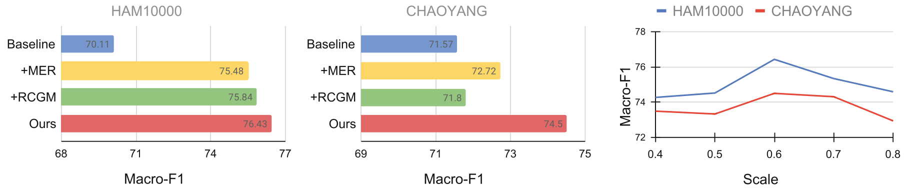
experts, with 40% of training samples having inconsistent annotations from the experts. To emulate imbalanced scenarios, we prune the class sizes of the training set into an imbalanced distribution as [5]. Consequently, CHAOYANG dataset consists of a training set with 2,181 images, a validation set with 713 images, and a testing set with 1,426 images, where the validation and testing sets have clean labels. The imbalanced ratios [12] of HAM10000 and CHAOYANG are 59 and 20, respectively. The evaluation metrics are Macro-F1, B-ACC, and MCC.

3.2 Implementation Details

We mainly follow the training settings of FCD [12]. ResNet-18 pretrained on the ImageNet is the backbone. The batch size is 48. Learning rates are 0.06, 0.001, 0.06 and 0.006 with the cosine schedule for four stages, respectively. We train our models by SGD optimizer with sharpness-aware term [6] for 90, 90, 90, and 20 epochs. The size of input image is 224×224. The scale and threshold in RCGM are 0.6 and 0.1, respectively.

3.3 Comparison with State-of-the-Art Methods

We compare our model with state-of-the-art methods which contain noisy methods (including DivideMix [13], NL [16], GCE [27], Co-Learning [19]), imbalance methods (including Focal Loss [14], Sqrt-RS [17], PG-RS [10], CB-Focal [5], EQL [21], EQL V2 [20], CECE [5], CLAS [28], FCD [12]), and noisy imbalanced classification methods (including H2E [25], NL+Sqrt-RS, GCE+Sqrt-RS, GCE+Focal). We train all approaches under the same data augmentations and network architecture. Table 1 exhibits the overall comparison of all approaches. We first conclude that noisy imbalanced setting does negatively affect learning with noise methods and imbalanced methods. In imbalanced methods, CECE only obtains 40.92 in Macro-F1 on HAM10000 and 47.12% Macro-F1 on CHAOYANG. In noisy methods, NL and GCE also suffer great performance declines. We mix these weakly-performed approaches with methods from the other category, observing the accuracy improvement. Compared to GCE, GCE+Sqrt-RS achieves +20.34% Macro-F1 on HAM10000 and +49.79% Macro-F1 on CHAOYANG. Similar increases happen in GCE & GCE+Focal and NL & NL+Sqrt-RS. Then, we compare our approach to state-of-the-art methods of the noisy (DivideMix), imbalanced (FCD), and noisy long-tailed (H2E) methods. Our framework achieves improvements in all metrics on both datasets, demonstrating the rationality of the assumption and the effectiveness of our framework.



(a) (b) (c)

Fig. 3. Ablation analysis. (a) and (b) Quantitative performance comparison of different components of our method on HAM10000 and CHAOYANG datasets, respectively. (c) Comparative results of our approach with different s values.

3.4 Ablation Studies

As shown in Fig. 3, we evaluate the effectiveness of the components in our method by decomposing them on extensive experiments. We choose the first stage of FCD [12] as our baseline. Figure 3a and 3b show that only using MER or RCGM achieves better performance than our strong baseline on both datasets. For example, MER achieves 5.37% and 1.15% improvements on HAM10000 and CHAOYANG, respectively, demonstrating the effectiveness of our noise removal techniques. Further, our multi-stage noise removal technique outperforms single MER and RCGM, revealing that the decomposition for noise effect and hardness effect works on noisy imbalanced datasets. We find that the combination of MER and RCGM improves more on CHAOYANG dataset. This is because CHAOYANG has more possible label noise than HAM10000 caused by the high annotating procedure. From Fig. 3c, we observe the accuracy trends are as the scale increases and achieve the peak around 0.6. It indicates the re-scaling process for noise weight deduction contributes to balancing the feature learning and classification boundary disturbance from the mixture of noisy and clean data. Furthermore, similar performance trends reveal the robustness of scale s.

4 Conclusion and Discussion

We propose a multi-step framework for noisy long-imbalanced medical image classification. We address three practical adverse effects including data noise, imbalanced distribution, and class hardness. To solve these difficulties, we conduct Multi-Environment Risk Minimization (MER) and rescaling class-aware Gaussian Mixture Modeling (RCGM) together for robust feature learning. Extensive results on two public medical image datasets have verified that our framework works on the noisy imbalanced classification problem. The main limitation of our work is the manually designed multi-stage training protocol which lacks simplicity compared to end-to-end training and warrants future simplification.

Acknowlegdement. This work described in this paper was supported in part by the Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089. The work was also partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Number: T45-401/22-N) and by a grant from the Hong Kong Innovation and Technology Fund (Project Number: GHP/080/20SZ). The work was also partially supported by a grant from the National Key R&D Program of China (2022YFE0200700), a grant from the National Natural Science Foundation of China (Project No. 62006219), and a grant from the Natural Science Foundation of Guangdong Province (2022A1515011579).

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Unsupervised label noise modeling and loss correction. In: Chaudhuri, K., Salakhutdinov, R. (eds.) ICML 2019 (2019)
2. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
3. Chen, P., Liao, B.B., Chen, G., Zhang, S.: Understanding and utilizing deep neural networks trained with noisy labels. In: ICML (2019)
4. Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In: ICCV (2015)

5. Cui, Y., Jia, M., Lin, T., Song, Y., Belongie, S.J.: Class-balanced loss based on effective number of samples. In: CVPR (2019)
6. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021. OpenReview.net (2021). https://openreview.net/forum?id=6Tm1mposlrM
7. Frénay, B., Verleysen, M.: Classification in the presence of label noise: a survey. IEEE TNNLS **25**(5), 845–869 (2013)
8. Huang, Y., Bai, B., Zhao, S., Bai, K., Wang, F.: Uncertainty-aware learning against label noise on imbalanced datasets. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, 22 February–1 March 2022, pp. 6960–6969. AAAI Press (2022). https://ojs.aaai.org/index.php/AAAI/article/view/20654
9. Kang, B., et al.: Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217 (2019)
10. Kang, B., et al.: Decoupling representation and classifier for long-tailed recognition. In: ICLR (2020)
11. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med. Image Anal. **65**, 101759 (2020)
12. Li, J., et al.: Flat-aware cross-stage distilled framework for imbalanced medical image classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13433, pp. 217–226. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_21
13. Li, J., Socher, R., Hoi, S.C.H.: Dividemix: learning with noisy labels as semi-supervised learning. In: ICLR 2020 (2020)
14. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
15. Liu, J., Sun, Y., Han, C., Dou, Z., Li, W.: Deep representation learning on long-tailed data: a learnable embedding augmentation perspective. In: CVPR (2020)
16. Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S.M., Bailey, J.: Normalized loss functions for deep learning with noisy labels. In: ICML 2020 (2020)
17. Mahajan, D., et al.: Exploring the limits of weakly supervised pretraining. In: ECCV (2018)
18. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: a survey. IEEE TNNLS (2022)
19. Tan, C., Xia, J., Wu, L., Li, S.Z.: Co-learning: learning from noisy labels with self-supervision. In: Shen, H.T., et al. (eds.) ACM 2021 (2021)
20. Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q.: Equalization loss V2: a new gradient balance approach for long-tailed object detection. In: CVPR (2021)
21. Tan, J., et al.: Equalization loss for long-tailed object recognition. In: CVPR 2020 (2020)
22. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**(1), 1–9 (2018)
23. Xue, C., Dou, Q., Shi, X., Chen, H., Heng, P.A.: Robust learning at noisy labeled medical images: applied to skin lesion classification. In: ISBI 2019 (2019)

24. Xue, C., Yu, L., Chen, P., Dou, Q., Heng, P.A.: Robust medical image classification from noisy labeled data with global and local representation guided co-training. IEEE TMI **41**(6), 1371–1382 (2022)
25. Yi, X., Tang, K., Hua, X.S., Lim, J.H., Zhang, H.: Identifying hard noise in long-tailed sample distribution. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13686, pp. 739–756. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19809-0_42
26. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
27. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) NIPS 2018 (2018)
28. Zhong, Z., Cui, J., Liu, S., Jia, J.: Improving calibration for long-tailed recognition. In: CVPR 2021 (2021)
29. Zhu, C., Chen, W., Peng, T., Wang, Y., Jin, M.: Hard sample aware noise robust learning for histopathology image classification. IEEE TMI **41**(4), 881–894 (2021)