Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study

Gregory Holste[1], Song Wang[1], Ziyu Jiang[2], Thomas C. Shen[3], George Shih[4], Ronald M. Summers[3], Yifan Peng[4⋆], and Zhangyang Wang[1⋆]

[1]The University of Texas at Austin, Austin, TX, USA
[2]Texas A&M University, College Station, TX, USA
[3]National Institutes of Health, Bethesda, MD, USA
[4]Weill Cornell Medicine, New York, NY, USA
yip4002@med.cornell.edu, atlaswang@utexas.edu

Abstract. Imaging exams, such as chest radiography, will yield a small set of common findings and a much larger set of uncommon findings. While a trained radiologist can learn the visual presentation of rare conditions by studying a few representative examples, teaching a machine to learn from such a "long-tailed" distribution is much more difficult, as standard methods would be easily biased toward the most frequent classes. In this paper, we present a comprehensive benchmark study of the long-tailed learning problem in the specific domain of thorax diseases on chest X-rays. We focus on learning from naturally distributed chest X-ray data, optimizing classification accuracy over not only the common "head" classes, but also the rare yet critical "tail" classes. To accomplish this, we introduce a challenging new long-tailed chest X-ray benchmark to facilitate research on developing long-tailed learning methods for medical image classification. The benchmark consists of two chest X-ray datasets for 19- and 20-way thorax disease classification, containing classes with as many as 53,000 and as few as 7 labeled training images. We evaluate both standard and state-of-the-art long-tailed learning methods on this new benchmark, analyzing which aspects of these methods are most beneficial for long-tailed medical image classification and summarizing insights for future algorithm design. The datasets, trained models, and code are available at https://github.com/VITA-Group/LongTailCXR.

Keywords: Long-Tailed Learning · Chest X-Ray · Class Imbalance.

1 Introduction

Like most diagnostic imaging exams, chest radiography produces a few very common findings, followed by many relatively rare findings [21,32]. Such a "long-tailed" (LT) distribution of outcomes can make it challenging to learn discriminative image features, as standard deep image classification methods will be

---

⋆ Denotes joint corresponding authorship.

biased toward the common "head" classes, sacrificing predictive performance on the infrequent "tail" classes [31]. In other settings and modalities, there are a select few examples of LT datasets, such as in dermatology [17] and gastrointestinal imaging [1]; however, the data from Liu et al [17] are not publicly available, and the *HyperKvasir* dataset [1] – while providing 23 unique class labels with several very rare conditions (<50 labeled examples) – only contains about 10,000 labeled images for classification. Additionally, while many studies offer techniques to combat class imbalance for medical image analysis problems [19,33,15,6], very few methods specifically address the challenges posed by an LT distribution, as there is no freely available benchmark for this purpose.

Only recently have studies begun to use the lens of "LT learning" to describe and improve medical image understanding solutions. For example, Galdran et al. [6] proposed Balanced-MixUp, an extension of the MixUp [29] regularization technique with class-balanced sampling, a common approach in the LT learning literature. Ju et al. [11] grouped rare classes into subsets based on prior knowledge (location, clinical presentation) and used knowledge distillation to train a "teacher" model to enforce the "student" to learn these groupings. Zhang et al. [30] combined a feature "memory" module, resampling of tail classes, and a re-weighted loss function to improve the LT classification of several medical datasets. More broadly, many relevant techniques have been developed in the related fields of imbalanced learning [33,19] and few-shot learning [24,14]. These medical image-specific techniques, plus the wealth of methods from the computer vision literature [3,8,28,16,4,2,25,12,31,9,20,13], provide a foundation from which the medical deep learning community can develop methods for medical LT classification.

Since no large-scale, publicly available dataset exists for LT medical image classification, we curate a large benchmark (>200,000 labeled images) of two thorax disease classification tasks on chest X-rays. Further, we evaluate state-of-the-art LT learning methods on this data, analyzing which components of existing methods are most applicable to the medical imaging domain.

Our contributions can be summarized as follows:

- We formally introduce the task of long-tailed classification of thorax disease on chest X-rays. The task provides a comprehensive and realistic evaluation of thorax disease classification in clinical practice settings.
- We curate a large-scale benchmark from the existing representative datasets NIH ChestXRay14 [28] and MIMIC-CXR [10]. The benchmark contains five new, more fine-grained pathologies, producing a challenging and severely imbalanced distribution of diseases. We describe the characteristics of this benchmark and will publicly release the labels.
- We find that the standard cross-entropy loss and augmentation methods such as MixUp fail to adequately classify the rarest "tail" classes. We observe that class-balanced re-weighting improves performance on infrequent classes, and "decoupling" via classifier re-training is the most effective approach for both datasets.
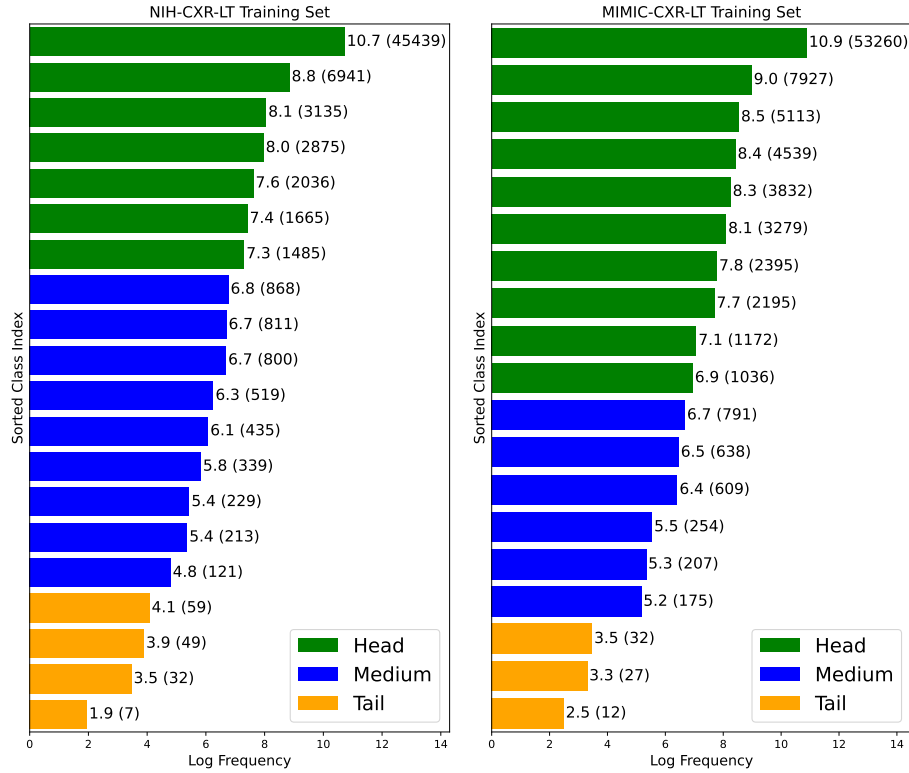
Fig. 1: Long-tailed distribution of thorax disease labels for the proposed NIH-CXR-LT (left) and MIMIC-CXR-LT (right) training datasets. Values by each bar represent log-frequency, while values in parentheses represent raw frequency. Textured bars represent newly added disease labels, which help create naturally long-tailed distributions without the need for artificial subsampling.

2   Long-Tailed Classification of Thorax Diseases

2.1   Task Definition

Disease patterns in chest X-rays are numerous, and their incidence exhibits a long-tailed distribution [32,21]: while a small number of common diseases have sufficient observed cases for large-scale analysis, most diseases are infrequent. Conventional computer vision methods may fail to correctly identify uncommon thorax disease classes due to the extremely imbalanced class distribution [21], introducing a new and clinically valuable LT classification task on chest X-rays. We formulate the LT classification task first by dividing thorax disease classes into "head" (many-shot: >1,000), "medium" (medium-shot: 100-1000, inclusive), and "tail" (few-shot: <100) categories according to their frequency in the training set.

2.2   Dataset Construction

We curate two long-tailed chest X-ray benchmarks, **NIH-CXR-LT** and **MIMIC-CXR-LT**, for NIH ChestXRay14 [26] and MIMIC-CXR [10], respectively. Each study of NIH ChestXRay14 and MIMIC-CXR usually contains one or more chest radiographs and one free-text radiology report. To generate a strongly long-tailed distribution without artificially subsampling, we introduce five new rare disease findings that are text-mined from radiology reports: Calcification of the Aorta, Subcutaneous Emphysema, Tortuous Aorta, Pneumomediastinum, and Pneumoperitoneum. We identify the presence or absence of new disease findings by parsing the text report associated with each study following the method detailed in RadText [23,26].

For this study, we only use frontal-view, single-label images, as most LT methods are developed specifically for multi-class (not multi-label) classification. Following the structure of previous LT benchmark datasets in computer vision, such as ImageNet-LT [18], we split NIH-CXR-LT and MIMIC-CXR-LT into *training*, *validation*, *test*, and *balanced test* sets. Since both datasets contain patients with multiple images, we split them at the patient level to prevent data leakage. Both validation and balanced test sets are small but perfectly balanced, where the balanced test set is a subset of the larger, imbalanced test set. This data split allows for evaluation consistent with the LT literature (via the balanced test set), as well as more traditional evaluation on a large naturally distributed set (via the test set). The resulting splits produce extreme class imbalance, with an *imbalance factor* – the cardinality of the most frequent training class divided by the cardinality of the least frequent training class – of 6,491 for NIH-CXR-LT and 4,438 for MIMIC-CXR-LT. Full detailed statistics and data split for NIH-CXR-LT and MIMIC-CXR-LT can be found in the Supplementary Materials.

NIH-CXR-LT. NIH ChestXRay14 contains over 100,000 chest X-rays labeled with 14 pathologies, plus a "No Findings" class. We construct a single-label, long-tailed version of the NIH ChestXRay14 dataset by introducing five new disease findings described above. The resulting NIH-CXR-LT dataset has 20 classes, including 7 head classes, 10 medium classes, and 3 tail classes. NIH-CXR-LT contains 88,637 images labeled with one of 19 thorax diseases, with 68,058 training and 20,279 test images. The validation and balanced test sets contain 15 and 30 images per class, respectively.

MIMIC-CXR-LT. We construct a single-label, long-tailed version of MIMIC-CXR in a similar manner. MIMIC-CXR is a multi-label classification dataset with over 200,000 chest X-rays labeled with 13 pathologies and a "No Findings" class. The resulting MIMIC-CXR-LT dataset contains 19 classes, of which 10 are head classes, 6 are medium classes, and 3 are tail classes. MIMIC-CXR-LT contains 111,792 images labeled with one of 18 diseases, with 87,493 training images and 23,550 test set images. The validation and balanced test sets contain 15 and 30 images per class, respectively.

Table 1: Long-tailed learning methods selected for benchmarking grouped by type of approach ("R" = Re-balancing, "A" = Augmentation, "O" = Other). "RW" = re-weighted with scikit-learn weights [22], "CB" = re-weighted with class-balanced weights [4].

Method	R	A	O	Method	R	A	O
Softmax (Baseline)				CB LDAM-DRW [2]	✓		
CB Softmax	✓			RW LDAM [2]	✓		
RW Softmax	✓			RW LDAM-DRW [2]	✓		
Focal Loss [16]	✓			MixUp [29]		✓	
CB Focal Loss [16]	✓			Balanced-MixUp [6]	✓	✓	
RW Focal Loss [16]	✓			Decoupling–cRT [12]	✓		✓
LDAM [2]	✓			Decoupling–τ-norm [12]	✓		✓
CB LDAM [2]	✓						

2.3   Methods for Benchmarking

In their survey, Zhang *et al.* group LT learning methods into three main categories: class re-balancing, information augmentation, and module improvement [31]. We simplify this categorization down to re-balancing, augmentation, and others, noting that some sophisticated methods can fall into more than one of these categories. We have summarized our selected methods for benchmarking with their corresponding categorizations in Table 1.

Class re-balancing, arguably the most common approach to LT learning, usually involves *resampling* the data such that it is effectively balanced during training or *re-weighting* a loss function to modulate the importance of classes based on their frequency. Resampling methods include SMOTE [3], which undersamples common classes and oversamples rare classes, and progressively-balanced sampling [12], which interpolates from instance- to class-balanced sampling; recent re-weighting strategies include Focal Loss [16], Label-Distribution-Aware Margin (LDAM) Loss [2], and Influence-Balanced Loss [20]. In addition to the baseline softmax cross-entropy loss function, we consider Focal Loss and LDAM, with optional deferred re-weighting (DRW). For re-weighting strategies, we select the "class-balanced" (CB) approach outlined in [4] and the re-weighting approach implemented by the scikit-learn library [22].

Approaches to "information augmentation" can include customized data augmentation, as well as transfer learning from related data domains. For this category, we choose MixUp [29] and Balanced-MixUp [6]. MixUp is an augmentation technique that linearly mixes pairs of input images and labels according to a Beta distribution, producing a strong regularizing effect. Balanced-MixUp, as explained earlier, is an extension of MixUp that linearly mixes pairs of images and labels, where one image is drawn from a batch of instance-balanced (naturally distributed) data and the other from class-balanced (resampled) data.

Lastly, other popular approaches to LT learning include ensembling, representation learning, classifier design, and decoupled training. For this category, we proceed with two straightforward decoupling methods: classifier re-training (cRT) and τ-normalization. Kang et al. [12] observed that they could achieve state-of-the-art results on several LT learning benchmarks by (1) learning representations from naturally distributed data, then (2) re-training or otherwise calibrating the classification head in order to better discriminate tail classes. After training a model on instance-balanced data, cRT freezes this trained backbone, then re-initializes and re-trains the classifier with class-balanced resampling. Directly using the model learned in step (1), τ-normalization scales each classifier's learned weights by their magnitude raised to the power τ.

2.4   Experiments and Evaluation

We evaluate the list of methods shown in Table 1 on NIH-CXR-LT and MIMIC-CXR-LT. To enable a fair comparison among all methods, we keep the entire training pipeline identical except for the method being applied. Specifically, we train a ResNet50 [7] pretrained on ImageNet [5], using the Adam optimizer with a learning rate of 1×10^{-4}. All models were trained for a maximum of 60 epochs with early-stopping based on overall validation accuracy. For full implementation details, refer to the Supplemental Materials and our code repository: https://github.com/VITA-Group/LongTailCXR.

We present results on both the balanced test set and imbalanced test set for each model and dataset. For the balanced test set, we report head, medium, and tail class accuracy. We additionally include the class-wise average ("overall") accuracy and the group-wise average ("avg") accuracy – namely, the mean of the head, medium, and tail accuracy; we use this metric since we seek a model that performs well across head, medium, and tail classes regardless of how many samples or classes belong to each group. For the imbalanced test set, we report the Macro-F1 score (the unweighted mean of class-wise F1 scores) and the balanced accuracy (the accuracy with samples weighted by inverse class frequency). We choose balanced accuracy since it is resistant to class imbalance, thus necessary since the test set follows the highly imbalanced real-world data distribution.

3   Results and Analysis

For the NIH-CXR-LT dataset, the baseline method fails to adequately classify tail classes, achieving 1.7% accuracy on those three rarest diseases (Table 2). The baseline of softmax cross-entropy loss achieves a group-wise average accuracy of 0.164, but improves to 0.309 and 0.288, respectively, when using class-balanced and scikit-learn weights. Furthermore, we see that re-weighting constantly improves performance, though it is inconsistent which re-weighting method provides more significant gains than others. We also see that DRW can additionally improve performance, as evidenced by the fact that both CB LDAM-DRW and RW LDAM-DRW outperform their counterparts without DRW. We find that

cRT decoupling achieves the best performance on both the balanced and imbalanced test sets, reaching 0.369 group-wise average accuracy on the balanced test set and 0.294 balanced accuracy on the test set. Classifier re-training is followed closely by RW LDAM-DRW, reaching 0.362 group-wise average accuracy and 0.294 balanced accuracy.

On MIMIC-CXR-LT, again, the baseline approach almost entirely fails to capture the tail classes, reaching 0.022 tail accuracy and 0.188 group-wise average accuracy (Table 3). Like with the NIH-CXR-LT results, re-weighting is always beneficial; for example, class-balanced re-weighting and scikit-learn re-weighting, respectively, improve focal loss performance from 0.181 to 0.278 and 0.299 group-wise average accuracy. Similarly, DRW brings even further gains to a re-weighted LDAM loss, improving group-wise accuracy by at least 0.05. Classifier re-training again achieves both the highest group-wise average accuracy on the balanced test set and the highest balanced accuracy on the test set by a considerable margin. For both the balanced and imbalanced test sets, the second-best method is a re-weighted LDAM loss with deferred re-weighting – CB LDAM-DRW for the balanced test set and RW LDAM-DRW for the test set.

Summary of Findings. Overall, we see that the standard approach of optimizing softmax cross-entropy with instance-balanced weights fails to adequately capture medium and tail classes for both NIH-CXR-LT and MIMIC-CXR-LT. In

Table 2: Results on NIH-CXR-LT. Accuracy is reported for the balanced test set ($N = 600$), where "Avg" accuracy is the mean of the head, medium, and tail accuracy. Macro-F1 score (mF1) and balanced accuracy (bAcc) are used to evaluate performance on the imbalanced test set ($N = 20,279$). The best and second-best results for a given metric are, respectively, bolded and underlined.

Method	Balanced Test Set					Test Set	
	Overall	Head	Medium	Tail	Avg	mF1	bAcc
Softmax	0.175	0.419	0.056	0.017	0.164	0.131	0.115
CB Softmax	0.333	0.295	0.415	0.217	0.309	0.177	0.269
RW Softmax	0.300	0.248	0.359	0.258	0.288	0.116	0.26
Focal Loss	0.160	0.362	0.056	0.042	0.153	0.142	0.122
CB Focal Loss	0.303	0.371	0.333	0.117	0.274	0.157	0.232
RW Focal Loss	0.255	0.286	0.293	0.117	0.232	0.090	0.197
LDAM	0.232	0.410	0.133	0.142	0.228	0.173	0.178
CB LDAM	0.295	0.357	0.285	0.208	0.284	0.161	0.235
CB LDAM-DRW	0.377	0.476	0.356	0.250	0.361	0.172	0.281
RW LDAM	0.353	0.305	0.419	0.292	0.338	0.111	0.279
RW LDAM-DRW	0.370	0.410	0.367	0.308	<u>0.362</u>	0.127	<u>0.289</u>
MixUp	0.170	0.419	0.044	0.017	0.160	0.132	0.118
Balanced-MixUp	0.213	0.443	0.081	0.108	0.211	0.167	0.155
Decoupling–cRT	0.380	0.433	0.374	0.300	**0.369**	0.138	**0.294**
Decoupling–τ-norm	0.280	0.457	0.230	0.083	0.257	0.144	0.214

Table 3: Results on MIMIC-CXR-LT. Accuracy is reported for the balanced test set ($N = 570$), where "Avg" accuracy is the mean of head, medium, and tail accuracy. Macro-F1 score (mF1) and balanced accuracy (bAcc) are used to evaluate performance on the imbalanced test set ($N = 23,550$). The best and second-best results for a given metric are, respectively, bolded and underlined.

Method	Balanced Test Set					Test Set	
	Overall	Head	Medium	Tail	Avg	mF1	bAcc
Softmax	0.281	0.503	0.039	0.022	0.188	0.183	0.169
CB Softmax	0.347	0.493	0.167	0.222	0.294	0.186	0.227
RW Softmax	0.314	0.473	0.139	0.133	0.249	0.163	0.211
Focal Loss	0.268	0.477	0.044	0.022	0.181	0.182	0.172
CB Focal Loss	0.288	0.373	0.117	0.344	0.278	0.136	0.191
RW Focal Loss	0.335	0.403	0.283	0.211	0.299	0.144	0.239
LDAM	0.261	0.497	0.000	0.000	0.166	0.172	0.165
CB LDAM	0.330	0.467	0.161	0.211	0.280	0.161	0.225
CB LDAM-DRW	0.379	0.520	0.156	0.356	<u>0.344</u>	0.197	0.267
RW LDAM	0.335	0.437	0.250	0.167	0.284	0.149	0.243
RW LDAM-DRW	0.365	0.447	0.256	0.311	0.338	0.177	<u>0.275</u>
MixUp	0.291	0.543	0.011	0.011	0.189	0.182	0.176
Balanced-MixUp	0.267	0.480	0.039	0.011	0.177	0.176	0.168
Decoupling–cRT	0.412	0.490	0.306	0.367	**0.387**	0.170	**0.296**
Decoupling–τ-norm	0.337	0.520	0.167	0.067	0.251	0.178	0.230

contrast to the empirical success of MixUp on many natural image-based problems and Balanced-MixUp on certain medical imaging tasks, we find that MixUp and Balanced-MixUp perform similarly to the baseline for these two tasks; perhaps linearly mixing radiographs destroys valuable high-contrast signal that is necessary for discriminating disease conditions. We see that re-weighting is always beneficial, though which re-weighting method provides larger gains appears to depend on its interaction with the loss function used. We also observe that DRW can provide additional gains to standard re-weighting when used with the LDAM loss. Finally, we see that cRT decoupling was the highest-performing method on both datasets, demonstrating that decoupled training can be a simple and powerful technique for long-tailed disease classification on chest X-rays. We note that reported metrics appear lower than prior work on the original NIH ChestXRay14 and MIMIC-CXR datasets since (1) we use different metrics (primarily accuracy, not area under the receiver operating characteristic curve), (2) we only use single-label images, and (3) the newly added classes are difficult and introduce confusion with the set of original diseases in each dataset.

4   Discussion and Conclusion

In summary, we have conducted the first comprehensive study of long-tailed learning methods for disease classification from chest X-rays. We publicly release all code, models, and data to encourage the development of long-tailed learning methods for medical image classification. While we adopted the standard practice of using ImageNet pretrained weights, this limited the list of candidate long-tailed learning methods we could use. For example, certain LT methods that use specialized architectures [25,27] or explore self-supervised learning [9,19] on other datasets are not compatible with ImageNet pretraining. Future work will explore various pretraining options, combating long-tailed data with a different weight initialization. Lastly, future work will also involve adapting multi-label long-tailed learning methods to these datasets, acknowledging the clinical reality that patients often present with multiple pathologies at once.

5   Acknowledgments

References

1. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data **7**(1), 1–14 (2020)
2. Cao, K., Wei, C., Gaidon, A., Aréchiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems, NeurIPS. pp. 1565–1576 (2019)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research (2002)
4. Cui, Y., Jia, M., Lin, T., Song, Y., Belongie, S.J.: Class-balanced loss based on effective number of samples. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 9268–9277 (2019)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
6. Galdran, A., Carneiro, G., Ballester, M.Á.G.: Balanced-MixUp for highly imbalanced medical image classification. In: Medical Image Computing and Computer Assisted Intervention, MICCAI. vol. 12905, pp. 323–333. Springer (2021)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 770–778 (2016)

8. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 5375–5384 (2016)
9. Jiang, Z., Chen, T., Mortazavi, B.J., Wang, Z.: Self-damaging contrastive learning. In: International Conference on Machine Learning, ICLR. pp. 4927–4939 (2021)
10. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data (2019)
11. Ju, L., Wang, X., Wang, L., Liu, T., Zhao, X., Drummond, T., Mahapatra, D., Ge, Z.: Relational subsets knowledge distillation for long-tailed retinal diseases recognition. In: Medical Image Computing and Computer Assisted Intervention, MICCAI. vol. 12908, pp. 3–12. Springer (2021)
12. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: International Conference on Learning Representations, ICLR (2020)
13. Kini, G.R., Paraskevas, O., Oymak, S., Thrampoulidis, C.: Label-imbalanced and group-sensitive classification under overparameterization. Advances in Neural Information Processing Systems, NeurIPS **34** (2021)
14. Li, X., Yu, L., Jin, Y., Fu, C., Xing, L., Heng, P.: Difficulty-aware meta-learning for rare disease diagnosis. In: Medical Image Computing and Computer Assisted Intervention, MICCAI. vol. 12261, pp. 357–366. Springer (2020)
15. Lin, C., Wu, H., Wen, Z., Qin, J.: Automated malaria cells detection from blood smears under severe class imbalance via importance-aware balanced group softmax. In: Medical Image Computing and Computer Assisted Intervention, MICCAI. vol. 12908, pp. 455–465. Springer (2021)
16. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, ICCV (2017)
17. Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., et al.: A deep learning system for differential diagnosis of skin diseases. Nature Medicine **26**(6), 900–908 (2020)
18. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2019)
19. Marrakchi, Y., Makansi, O., Brox, T.: Fighting class imbalance with contrastive learning. In: Medical Image Computing and Computer Assisted Intervention, MICCAI. Lecture Notes in Computer Science, vol. 12903, pp. 466–476. Springer (2021)
20. Park, S., Lim, J., Jeon, Y., Choi, J.Y.: Influence-balanced loss for imbalanced visual classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV. pp. 735–744 (2021)
21. Paul, A., Shen, T.C., Lee, S., Balachandar, N., Peng, Y., Lu, Z., Summers, R.M.: Generalized zero-shot chest x-ray diagnosis through trait-guided multi-view semantic embedding with self-training. IEEE Transactions on Medical Imaging (2021)
22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
23. Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R.M., Lu, Z.: Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits on Translational Science Proceedings **2018**, 188 – 196 (2018)
24. Quellec, G., Lamard, M., Conze, P., Massin, P., Cochener, B.: Automatic detection of rare pathologies in fundus photographs using few-shot learning. Medical Image Anal. **61**, 101660 (2020)

25. Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., Meng, D.: Meta-weight-net: Learning an explicit mapping for sample weighting. Advances in Neural Information Processing Systems, NeurIPS **32** (2019)
26. 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: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 3462–3471 (2017)
27. Wang, X., Lian, L., Miao, Z., Liu, Z., Yu, S.: Long-tailed recognition by routing diverse distribution-aware experts. In: International Conference on Learning Representations, ICLR (2020)
28. Wang, Y.X., Ramanan, D., Hebert, M.: Learning to model the tail. Advances in Neural Information Processing Systems, NeurIPS **30** (2017)
29. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
30. Zhang, R., E, H., Yuan, L., He, J., Zhang, H., Zhang, S., Wang, Y., Song, M., Wang, L.: MBNM: multi-branch network based on memory features for long-tailed medical image recognition. Comput. Methods Programs Biomed. **212**, 106448 (2021)
31. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. arXiv preprint arXiv:2110.04596 (2021)
32. Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE **109**(5), 820–838 (2021)
33. Zhuang, J., Cai, J., Wang, R., Zhang, J., Zheng, W.: CARE: class attention to regions of lesion for classification on imbalanced data. In: International Conference on Medical Imaging with Deep Learning, MIDL. Proceedings of Machine Learning Research, vol. 102, pp. 588–597. PMLR (2019)