Balanced-MixUp for Highly Imbalanced Medical Image Classification

Adrian Galdran[1(✉)], Gustavo Carneiro[2], and Miguel A. González Ballester[3,4]

[1] Bournemouth University, Poole, UK
agaldran@bournemouth.ac.uk
[2] University of Adelaide, Adelaide, Australia
gustavo.carneiro@adelaide.edu
[3] BCN Medtech, Department of Information and Communication Technologies,
Universitat Pompeu Fabra, Barcelona, Spain
ma.gonzalez@upf.edu
[4] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

Abstract. Highly imbalanced datasets are ubiquitous in medical image classification problems. In such problems, it is often the case that rare classes associated to less prevalent diseases are severely under-represented in labeled databases, typically resulting in poor performance of machine learning algorithms due to overfitting in the learning process. In this paper, we propose a novel mechanism for sampling training data based on the popular MixUp regularization technique, which we refer to as Balanced-MixUp. In short, Balanced-MixUp simultaneously performs regular (*i.e.*, instance-based) and balanced (*i.e.*, class-based) sampling of the training data. The resulting two sets of samples are then mixed-up to create a more balanced training distribution from which a neural network can effectively learn without incurring in heavily under-fitting the minority classes. We experiment with a highly imbalanced dataset of retinal images (55K samples, 5 classes) and a long-tail dataset of gastro-intestinal video frames (10K images, 23 classes), using two CNNs of varying representation capabilities. Experimental results demonstrate that applying Balanced-MixUp outperforms other conventional sampling schemes and loss functions specifically designed to deal with imbalanced data. Code is released at https://github.com/agaldran/balanced_mixup

Keywords: Imbalanced learning · Long-tail image classification

1 Introduction

Backed by the emergence of increasingly powerful convolutional neural networks, medical image classification has made remarkable advances over the last years, reaching unprecedented levels of accuracy [20]. However, due to the inherent difficulty in collecting labeled examples of rare diseases or other unusual instances in a medical context, these models are often trained and tested on large datasets
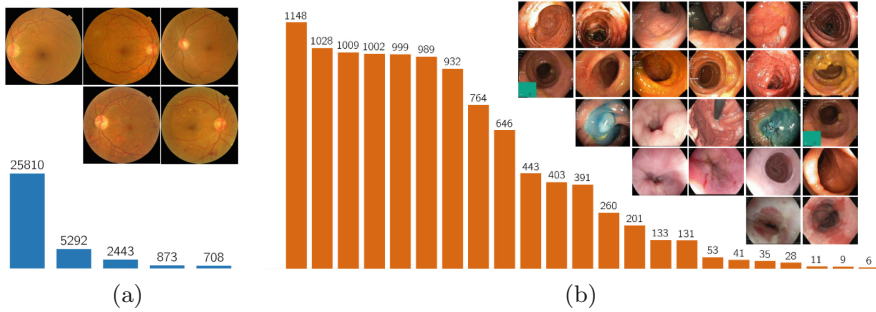
Fig. 1. The two different data imbalance scenarios considered in this paper. Left: Highly imbalanced problem (DR grading from retinal images, $K = 5$ [17]). Right: Long-tailed data (Gastro-intestinal image classification, $K = 23$ [2]).

containing a more balanced distribution with respect to image classes than the one found in real-life clinical scenarios, which typically has a long-tailed distribution. Under such severe data imbalance, over-represented (i.e., majority) classes tend to dominate the training process, resulting in a decrease of performance in under-represented (i.e., minority) classes [23,31]. Therefore, developing training methods that are adapted to strong data imbalance is essential for the advance of medical image classification.

Common solutions to address data imbalance involve data re-sampling to achieve a balanced class distribution [4], curriculum learning [13], adapted loss functions, *e.g.* cost-sensitive classification [9,30], or weighting the contribution of the different samples [7,19]. Another approach uses synthetic manipulation of data and/or labels to drive the learning process towards a more suitable solution, like label smoothing [8] or SMOTE [5]. Our proposed technique is a combination of modified training data sampling strategies with synthetic data manipulation via the well-known MixUp regularization method [28]. Therefore it is deeply connected with MixUp, with the fundamental difference that, while MixUp randomly combines training samples without taking into account their classes, we carefully mix examples from minority categories with examples from the other classes in order to create a more suitable training data distribution.

Some approaches have been recently explored involving MixUp and data imbalance. ReMix [6] mixes up by keeping the minority class label, instead of mixing up the labels. Similarly, MixBoost [14] attempts to combine active learning with MixUp to select which training samples to mix from each category, adding an extra complexity layer to the sampling process. Another popular technique that is related to our approach is SMOTE [5]. However, SMOTE generates convex combinations of input samples only between nearest neighbors of the same class. While this creates extra examples in regions of the space where there is less data, it does not provide the same regularizing benefits as MixUp.

In this paper, we propose Balanced-MixUp, a new imbalanced-robust training method that mixes up imbalanced (instance-based) and balanced (class-based) sampling of the data. Experiments on two different medical image classification tasks with highly imbalanced and long-tailed data (as shown in Fig. 1), using neural networks of different complexities, show that Balanced-MixUp is effective in creating a more evenly distributed training data and also regularizing over-parametrized models, resulting in better performance than other conventional and recent approaches to handle data imbalance in all the considered cases.

2   Methodology

In this paper we intend to combine the MixUp regularization technique with a modified sampling strategy for imbalanced datasets. This section introduces each of these concepts and then our proposed combined technique.

2.1   MixUp Regularization

The MixUp technique was initially introduced in [28] as a simple regularization technique to decrease overfitting in deep neural networks. If we denote (x_i, y_i) a training example composed of an image x_i and its associated one-hot encoded label y_i, with f_θ being a neural network to be trained to approximate the mapping $f(x_i) = y_i \ \forall i$, then MixUp creates synthetic examples and labels as follows:

$$\hat{x} = \lambda x_i + (1 - \lambda)x_j, \quad \hat{y} = \lambda y_i + (1 - \lambda)y_j \tag{1}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$, with $\alpha > 0$. Since $\lambda \in [0, 1]$, $\hat{x}$ and $\hat{y}$ are random convex combinations of data and label inputs. Typical values of α vary in $[0.1, 0.4]$, which means that in practice $\hat{x}$ will likely be relatively close to either x_i or x_j. Despite its simplicity, it has been shown that optimizing f_θ on mixed-up data leads to better generalization and improves model calibration [26].

### 2.2   Training Data Sampling Strategies

When dealing with extremely imbalanced data, the training process is typically impacted by different per-class learning dynamics that result in underfitting of minority classes, which are rarely presented to the model and may end up being entirely ignored [15]. Modified sampling strategies can be utilized to mitigate this effect, like oversampling under-represented categories, although merely doing so often leads to counter-productive outcomes, *e.g.* repeatedly showing to the model the same training examples may lead to the overfitting of minority classes [5].

Before detailing how to combine oversampling of minority classes with MixUp regularization, we introduce some further notation to describe sampling strategies. Given a training set $\mathcal{D} = \{(x_i, y_i), \ i = 1, ..., N\}$ for a multi-class problem

with K classes, if each class k contains n_k examples, we have that $\sum_{k=1}^{K} n_k = N$. We can then describe common data sampling strategies as follows:

$$p_j = \frac{n_j^q}{\sum_{k=1}^{K} n_k^q},\tag{2}$$

being p_j the probability of sampling from class j during training. With this, selecting $q = 1$ amounts to picking examples with a probability equal to the frequency of their class in the training set (*instance-based sampling*), whereas choosing $q = 0$ leads to a uniform probability $p_j = 1/K$ of sampling from each class, this is, *class-based sampling* or oversampling of minority classes. Another popular choice is *square-root sampling*, which stems from selecting $q = 1/2$.
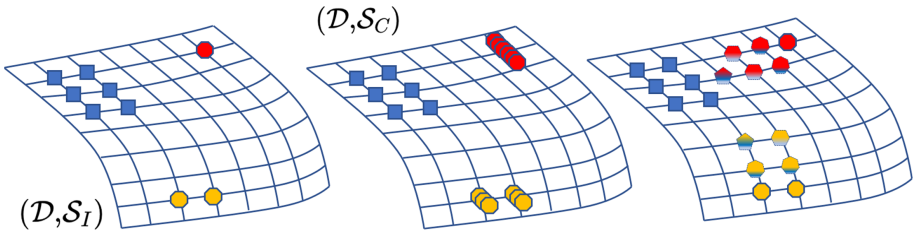
Fig. 2. Schematic illustration of Balanced-MixUp (right) filling underpopulated areas in the data space - linked to infrequent categories - with synthetic examples.
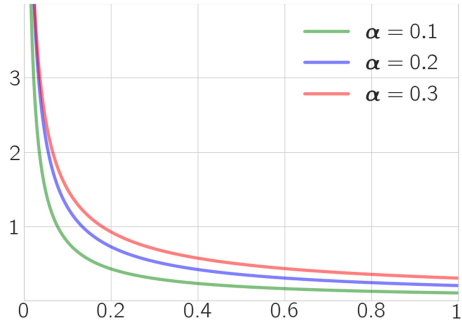
2.3 Balanced-MixUp

We propose to combine the two training enhancements described above into a single technique that is suitable for learning under strong data imbalance. We introduce an adaptation to this scenario of MixUp, referred to as Balanced-MixUp, in which two data points are not randomly sampled without considering their category, but rather sampling one following an instance-based strategy and the other following a class-based sampling.

Let us refer to a training dataset together with a sampling strategy as $(\mathcal{D}, \mathcal{S})$, with $\mathcal{S}_I$ and $\mathcal{S}_C$ referring to instance-based and class-based sampling respectively. With this notation, Balanced-MixUp can be described as follows:

$$\hat{x} = \lambda x_I + (1 - \lambda) x_C, \quad \hat{y} = \lambda y_I + (1 - \lambda) y_C,\tag{3}$$

for $(x_I, y_I) \in (\mathcal{D}, \mathcal{S}_I), (x_C, y_C) \in (\mathcal{D}, \mathcal{S}_C)$. This induces a more balanced distribution of training examples by creating synthetic data points around regions of the space where minority classes provide less data density, as illustrated in Fig. 2. At the same time, adding noise to the labels helps regularizing the learning process.

In our case, departing from the original MixUp formulation, we obtain the mixing coefficient as $\lambda \sim \text{Beta}(\alpha, 1)$. This results in an exponential-like distri bution as shown in Fig. 3, which leads to convex combinations in which examples from $(\mathcal{D}, \mathcal{S}_I)$ receive more weight, preventing overfitting in minority classes. In addition, it allows us to formulate our technique as depending on a single hyperparameter with an intuitive behavior: as α grows,

Fig. 3. $\text{Beta}(\alpha, 1)$ distribution for varying α

examples from minority classes are mixed up with greater weights, increasing class balance at the expense of risking some overfitting.

2.4 Training Details

All models in this paper are CNNs implemented in Pytorch [22] and optimized to reduce the Cross-Entropy loss using Stochastic Gradient Descent with a batch size of 8 and a learning rate of 0.01. During training, standard data augmentation operations are applied, and the learning rate is cyclically annealed towards 0. The metric of interest in each task (see next section) is used to monitor model's performance on the validation set and the best one is kept. After training, we generate predictions by averaging the result of the input image and its horizontally flipped version, which is common practice to regularize predictions [25].

3 Experimental Analysis

In this section we detail our experimental setup: considered datasets with relevant statistics, baseline techniques for comparison, evaluation metrics, and a discussion on the numerical differences in performance between methods.

3.1 Experimental Details

We consider two different image classification tasks: Diabetic Retinopathy (DR) grading from retinal fundus images and gastrointestinal image (GI) classification.

– For DR grading, we use the Eyepacs database[1] the largest publicly available dataset for this task. This database has images labeled as belonging to one out of five possible DR grades, ranging from DR0 (no DR) to DR4 (proliferative DR). Due to the nature of this disease, and the very subtle signs associated to the DR1 class (just a few microaneurysms present), the class distribution is highly imbalanced, as shown in Fig. 1. The database contains ∼35,000 training

---

[1] https://www.kaggle.com/c/diabetic-retinopathy-detection.

examples and $\sim$55,000 images assigned to a separate test set. We use 10% of the training set for validation and early-stopping purposes, and derive performance measures from the official test set. As an external database to examine model generalization we also use the 1748 retinal images in the Messidor-2 dataset[2], with the grades provided[3] in [17].

– For GI image classification, we use the Hyper-Kvasir dataset[4], recently made available [2]. This database contains 10,662 images labeled with anatomical landmarks as well as pathological and normal findings, and it represents a challenging classification problem due to the larger amount of classes (23) and the varying class frequencies, with minority classes very rarely represented, see Fig. 1. Please note that there is also an official test set [12], but we found this to contain only 11 different classes. Since our focus is on assessing model performance under a long-tail distribution, we prefer to discard this set, running all our experiments and computing performance for all models in a stratified 5-fold cross-validation manner.

For performance assessment in the DR case, we adopt as the main evaluation a measure based on the quadratic-weighted kappa score (quad-κ), which is the standard performance measure in grading problems, as it considers the correct ordering of predictions. We also report Matthews Correlation Coefficient (MCC), which has been shown to be meaningful under strong class imbalance [3], and the Kendall-τ coefficient. For the GI image classification task, we adopt as the main assessment measure MCC, as it was used for performance ranking in a recent challenge associated to this database [12]. Since we are dealing with a long-tail data distribution, we also report Balanced Accuracy (the mean of the diagonal in a normalized confusion matrix) and the macro-F1 score, which capture well performance in minority classes.

For comparison purposes, in both tasks we contrast Balanced-MixUp with other conventional sampling approaches: 1) Class-based sampling (oversampling of minority classes, as recommended in [4]), 2) Instance-based sampling, and 3) sampling based on the square root of class-frequencies [21]. We also include training with loss functions that are meant to perform better under severe class imbalance, namely 4) the popular Focal Loss [18], which modulates the contribution of each example according to the amount of loss it generates, and 5) a Class-Balanced loss recently introduced that weighs the loss of each class during training based on the *effective* number of samples, see [7].

In all cases, in order to observe the behavior of the different techniques with models of varying sizes, we train two different CNN baseline architectures: a MobileNet V2 [24] and more powerful ResNext50, modified as in [16]. For sensitivity analysis purposes, we measure the impact of varying the only hyperparameter α in Balanced-MixUp, with $\alpha \in \{0.1, 0.2, 0.3\}$ and report the corresponding results.

---

[2] https://www.adcis.net/en/third-party/messidor2/.
[3] https://www.kaggle.com/google-brain/messidor2-dr-grades.
[4] https://endotect.com/.

3.2 Discussion on the Numerical Results on DR Grading

Assessment measures for the task of DR grading from retinal images are displayed in Table 1. The leftmost three columns show performance of a MobileNet V2 CNN in terms of quad-κ, MCC and Kendall-τ. In this case, it can be seen that Balanced-MixUp with $\alpha = 0.1$ outperforms all other approaches in terms of the three considered figures of merit. Note that setting $\alpha = 0.2$ still outperforms all other techniques. It is also interesting to observe the differences when we switch to a larger model, ResNext50. In particular, we can see that: 1) the performance of almost all methods is increased, indicating that ResNext50 is a more suitable choice for this problem; and 2) Balanced-MixUp with $\alpha = 0.2$ is now the dominant approach, reaching the highest scores in all considered measures. This is consistent with the intuition that Balanced-MixUp can regularize the learning process, and that increasing its hyperparameter α leads to a stronger regularization Indeed, for this architecture the second best method is Balanced-MixUp with $\alpha = 0.3$. It is also worth stressing that, again, the three values of α translate in better results than any of the other considered alternatives.

Table 1. Performance comparison with two different CNN architectures for DR grading on the Eyepacs test set. Best results are marked bold.

	MobileNet V2			ResNeXt-50		
	quad-κ	MCC	Kendall-τ	quad-κ	MCC	Kendall-τ
Class-sampling	74.58	49.34	63.30	74.45	51.12	64.14
Instance-sampling	78.75	61.79	73.39	80.17	62.78	74.32
Sqrt-sampling [21]	79.32	59.66	72.22	79.38	58.77	72.22
Focal Loss [18]	78.59	60.74	72.70	79.73	62.66	73.37
CB Loss [7]	77.84	61.08	72.85	79.19	61.83	74.32
Bal-Mxp: $\alpha = 0.1$	**79.61**	**62.41**	**74.05**	80.35	63.31	74.74
Bal-Mxp: $\alpha = 0.2$	79.43	62.22	73.69	**80.78**	**63.78**	**75.05**
Bal-Mxp: $\alpha = 0.3$	78.34	61.65	73.25	80.62	63.51	74.89

It is important to stress that the results shown in Table 1 do not represent the state-of-the-art in this dataset. Recent methods are capable of reaching greater quad-κ values by means of specialized architectures with attention mechanisms [11], or semi-supervised learning with segmentation branches [29]. Our results are however aligned with other recently published techniques, as shown in Table 2, which also contains test results in the Messidor-2 dataset (without any re-training), demonstrating their competitiveness and generalization ability. Moreover we consider Balanced-MixUp to be a supplementary regularizing technique of the learning process from which other methods could also benefit.

3.3 Discussion on the Numerical Results on GI Image Classification

The second task we consider is Gastro-Intestinal image classification. Here the class imbalance is greater than in the DR problem, in addition to the presence of 23 categories, which turn this into a more challenging dataset. Table 3 shows the median performance in a stratified 5-fold cross-validation process, when considering exactly the same techniques and architectures as in the previous section.

Although in this task the performance measures appear to be more mixed than in the DR problem, we can still see how Balanced-MixUp appears to bring improvements over the other competing approaches. Again, the ResNext50 architecture delivers better performance in terms of MCC, although differences are not as high as before, and some methods show a degraded performance when applied with this larger architecture (compared to using MobileNet).

Table 2. Performance comparison in terms of quad-κ for methods trained on the Eyepacs dataset and tested (no retraining) on Eyepacs and Messidor-2.

	QWKL	DR|graduate	Cost-sensitive	Iter. Aug.	Bal-Mxp
	PRL'18 [27]	MedIA'20[1]	MICCAI'20 [9]	TMI'20 [10]	Rx50, $\alpha = 0.2$
Eyepacs	74.00	74.00	78.71	80.00	**80.78**
Messidor-2	–	71.00	79.79	–	**85.14**

Table 3. Stratified 5-Fold cross-validation (median) results for GI image classification on the Hyper-Kvasir dataset. Best results are marked bold.

	MobileNet V2			ResNeXt-50		
	MCC	B-ACC	Macro-F1	MCC	B-ACC	Macro-F1
Class-sampling	89.74	61.84	61.84	88.97	59.57	58.40
Instance-sampling	90.49	62.13	62.16	90.74	61.41	61.69
Sqrt-sampling [21]	90.11	63.01	62.66	90.29	62.84	62.85
Focal Loss [18]	90.30	62.09	61.78	90.23	62.12	62.36
CB Loss [7]	85.84	54.93	54.87	89.67	**63.84**	63.71
Bal-Mxp: $\alpha = 0.1$	**90.90**	63.49	62.77	91.05	62.55	62.92
Bal-Mxp: $\alpha = 0.2$	90.54	63.44	63.85	**91.15**	62.80	**64.00**
Bal-Mxp: $\alpha = 0.3$	90.39	**64.76**	**64.07**	90.84	62.34	62.35

Focusing first on the MCC score, we can appreciate again the regularizing effect of Balanced-MixUp: for both architectures, applying it leads to improved performance. Moreover, we can again see the same trend as above: the MobileNet model performs better with less regularization ($\alpha = 0.1$), and increasing α slowly deteriorates the MCC. As expected, when using the ResNeXt architecture, a

greater $\alpha = 0.2$ delivers better performance. It is worth noting that the MCC reached by the MobileNet model with $\alpha = 0.1$ is greater than in any of the considered methods and architectures, excluding models regularized with Balanced-MixUp and varying α.

Finally, it is relevant to observe that when we consider balanced accuracy and macro-F1, which can capture better performance in minority classes, for almost all methods it appears to be beneficial to use a smaller architecture. This appears to reduce to some extent the minority class overfitting due to oversampling, observed in the low performance of class-based sampling. Interestingly, even with this small architecture, Balanced MixUp increases the performance for all values of α, which further shows that our approach not only helps regularizing large models but also contributes to an improved learning on minority classes.

4   Conclusion

Imbalanced data is present in almost any medical image analysis task. Therefore, designing techniques for learning in such regimes is a challenge of great clinical significance. This paper introduces Balanced-MixUp to deal with heavy imbalanced data distributions by means of the combination of the popular MixUp regularization technique and modified training data sampling strategies. Balanced-MixUp is easy to implement and delivers consistent performance improvements when compared with other popular techniques. Its extension to other tasks beyond image classification represents a promising future research direction.

References

1. Araujo, T., et al.: DR|GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Med. Image Anal. **63**, 101715 (2020)
2. Borgli, H., et al.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data **7**(1), 283 (2020)
3. Boughorbel, S., Jarray, F., El-Anbari, M.: Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE **12**(6), 0177678 (2017)
4. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. **106**, 249–259 (2018)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**(1), 321–357 (2002)
6. Chou, H.-P., Chang, S.-C., Pan, J.-Y., Wei, W., Juan, D.-C.: Remix: rebalanced mixup. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020. LNCS, vol. 12540, pp. 95–110. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-65414-6_9
7. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

8. Galdran, A., et al.: Non-uniform label smoothing for diabetic retinopathy grading from retinal fundus images with deep neural networks. Trans. Vis. Sci. Technol. **9**(2), 34–34 (2020)

9. Galdran, A., Dolz, J., Chakor, H., Lombaert, H., Ben Ayed, I.: Cost-sensitive regularization for diabetic retinopathy grading from eye fundus images. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 665–674. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_64

10. González-Gonzalo, C., Liefers, B., Ginneken, B., Sánchez, C.I.: Iterative augmentation of visual evidence for weakly-supervised lesion localization in deep interpretability frameworks: application to color fundus images. IEEE Trans. Med. Imaging **39**(11), 3499–3511 (2020)

11. He, A., Li, T., Li, N., Wang, K., Fu, H.: CABNet: category attention block for imbalanced diabetic retinopathy grading. IEEE Trans. Med. Imaging **40**(1), 143–153 (2021)

12. Hicks, S., Jha, D., Thambawita, V., Halvorsen, P., Hammer, H.L., Riegler, M.: The EndoTect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In: 25th International Conference on Pattern Recognition (ICPR) (2020)

13. Jiménez-Sánchez, A., et al.: Medical-based deep curriculum learning for improved fracture classification. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 694–702. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_77

14. Kabra, A., et al.: MixBoost: synthetic oversampling with boosted mixup for handling extreme imbalance. arXiv arXiv: 2009.01571 (September 2020)

15. Kang, B., et al.: Decoupling representation and classifier for long-tailed recognition. In: ICLR (2020)

16. Kolesnikov, A., et al.: Big Transfer (BiT): general visual representation learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 491–507. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_29

17. Krause, J., et al.: Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology **125**(8), 1264–1272 (2018)

18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection, pp. 2980–2988 (2017)

19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 318–327 (2020)

20. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)

21. Mahajan, D., et al.: Exploring the limits of weakly supervised pretraining. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 185–201. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_12

22. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: NEURIPS 2019, pp. 8024–8035 (2019)

23. Quellec, G., Lamard, M., Conze, P.H., Massin, P., Cochener, B.: Automatic detection of rare pathologies in fundus photographs using few-shot learning. Med. Image Anal. **61**, 101660 (2020)

24. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (June 2018)

25. Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J.: When and why test-time augmentation works. arXiv arXiv:2011.11156 (November 2020)
26. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: improved calibration and predictive uncertainty for deep neural networks. In: Advances in Neural Information Processing Systems (2019)
27. de la Torre, J., Puig, D., Valls, A.: Weighted kappa loss function for multi-class classification of ordinal data in deep learning. Pattern Recogn. Lett. **105**, 144–154 (2018)
28. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
29. Zhou, Y., et al.: Collaborative learning of semi-supervised segmentation and classification for medical images. In: Conference on Computer Vision and Pattern Recognition (June 2019)
30. Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans. Knowl. Data Eng. **18**(1), 63–77 (2006)
31. 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, pp. 588–597. PMLR (May 2019)