Test-Time Adaptation with Calibration of Medical Image Classification Nets for Label Distribution Shift

Wenao Ma[1], Cheng Chen[1], Shuang Zheng[2,3], Jing Qin[4], Huimao Zhang[2,3(✉)], and Qi Dou[1(✉)]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong
qdou@cse.cuhk.edu.hk
[2] Department of Radiology, The First Hospital of Jilin University, Changchun, China
huimao@jlu.edu.cn
[3] Jilin Provincial Key Laboratory of Medical Imaging & Big Data, Changchun, China
[4] Centre for Smart Health, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Abstract. Class distribution plays an important role in learning deep classifiers. When the proportion of each class in the test set differs from the training set, the performance of classification nets usually degrades. Such a label distribution shift problem is common in medical diagnosis since the prevalence of disease vary over location and time. In this paper, we propose the first method to tackle label shift for medical image classification, which effectively adapt the model learned from a single training label distribution to arbitrary unknown test label distribution. Our approach innovates distribution calibration to learn multiple representative classifiers, which are capable of handling different one-dominating-class distributions. When given a test image, the diverse classifiers are dynamically aggregated via the consistency-driven test-time adaptation, to deal with the unknown test label distribution. We validate our method on two important medical image classification tasks including liver fibrosis staging and COVID-19 severity prediction. Our experiments clearly show the decreased model performance under label shift. With our method, model performance significantly improves on all the test datasets with different label shifts for both medical image diagnosis tasks. Code is available at https://github.com/med-air/TTADC.

Keywords: Test-time adaptation · Label distribution shift · Medical image classification

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-16437-8_30.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
L. Wang et al. (Eds.): MICCAI 2022, LNCS 13433, pp. 313–323, 2022.
https://doi.org/10.1007/978-3-031-16437-8_30

1 Introduction

Intelligent medical image diagnosis has witnessed great success on accurate predictions for various tasks such as disease staging [14,23], lesion diagnosis [9,15], and severity prediction [11,25]. However, real-world use of classification models is challenged by the inevitable shift in class distributions on test data at deployment [1,27,32,33]. Usually, the proportion of samples belonging to each class is associated with patient demographics and region-related prevalence of disease, which differs from one hospital to another. This issue is called *label distribution shift*, which means that the label distribution can change across training and test datasets. As label distribution plays a vital role in classification tasks [7,16], such shift can make the learned classifier become suboptimal on unseen datasets, thus suffering from performance degradation in testing.

Label distribution shifts are very common in medical diagnosis as the disease distributions vary across location and time. For example, the prevalence of liver diseases significantly differs among regions due to the difference in vaccination coverage [31]. Such label shifts often degrade the performance of a learned classifier on test data, leading to erroneous predictions as observed in prior works [3,4,6]. For example, Davis et al. find that the prediction accuracy of their machine learning models decreases due to the declining incidence of acute kidney injury over time [6]. Since the proportion of normal and disease cases differs between the screening and diagnostic scenarios, an accurate model for screening purpose could perform poorly for diagnosis purpose, even for the same disease [3]. Park et al. [20] show in three disease classification models that dataset shifts including the label shift can lead to unreasonable predictions. Despite being observed in many real applications, the problem of label distribution shift has not yet been tackled for medical image diagnosis, severely hindering the large-scale deployment of deep models in clinical practice.

To generalize model under label shift, if the label distribution of test data can be known, such as the uniform distribution assumption made in [24,30], the label shift can be alleviated by re-sampling training data or adjusting the prediction probability in the softmax loss [22,24] accordingly. In practical scenarios, however, it is unlikely to anticipate the label distribution of test data, which is usually unknown and arbitrary, and may even continuously change. In this regard, we aim to mitigate label shift in a highly practical yet challenging setting, where the test label distribution is unknown and the trained model itself must accommodate label shift by utilizing the test data only. To tackle this problem, we consider two key ingredients. Firstly, since the test label distribution can be arbitrary, it is important to enlarge the capacity of models for an extensive label distribution space. The difficulty lies in how to establish such a representative space during model learning from the training set with a fixed label distribution. Secondly, motivated by the recent test-time learning works [28,29], although the knowledge of test dataset is unknown during model training, it can be explored from the test data at inference time.

In this paper, to our best knowledge, we present the first work to effectively tackle the label distribution shift in medical image classification. Our method

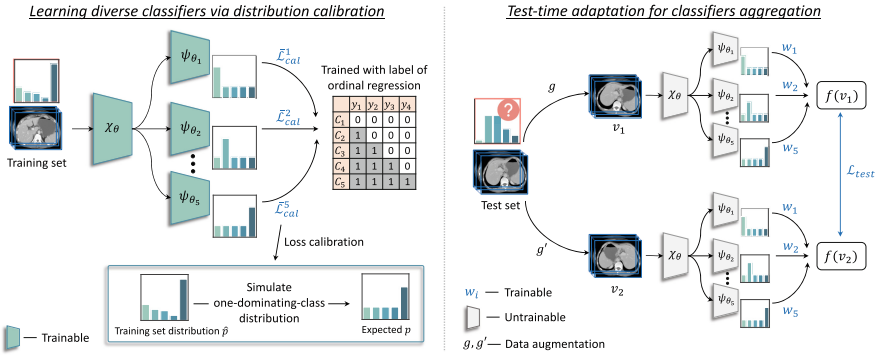

Fig. 1. Overview of our proposed method for test-time adaptation by calibration of medical image classification networks for label distribution shift.

learns representative classifiers with distribution calibration, by extending the concept of balanced softmax loss [24,34] to simulate multiple distributions that one class dominates other classes. Compared with [34], our method can be more flexible and be more targeted for ordinal classification, as our one-dominating-class distributions can represent more diverse label distributions and we use ordinal encoding instead of one-hot encoding to train the model. Then, at model deployment to new test data, we dynamically combine the representative classifiers by adapting their outputs to the label distribution of test data. The test-time adaptation is driven by a consistency regularization loss to adjust the weights of different classifier. We evaluate our method on two important medical applications of liver fibrosis staging and COVID-19 severity prediction. With our proposed method, the label shift can be largely mitigated with consistent performance improvement.

2 Method

2.1 Problem Formulation of Label Distribution Shift

For disease diagnosis, consider a classification task that aims to train a model to predict the disease class y correctly given an input image x. Let $\hat{p}(x, y)$ and $\tilde{p}(x, y)$ denote the training and test set distributions respectively. In practice, a deployed model often suffers from label distribution shift, which means the label distribution of training set $\hat{p}(y)$ is different from that of test set $\tilde{p}(y)$, i.e., $\hat{p}(y) \neq \tilde{p}(y)$, but the conditional distributions are consistent, i.e., $\hat{p}(x|y) = \tilde{p}(x|y)$. This phenomenon is especially common in medical image classification, where the disease label y is often the causal variable and the image data x can be regarded as the manifestations of a disease [26,32]. According to the Bayesian inference $\hat{p}(y|x) = \frac{\hat{p}(x|y)\hat{p}(y)}{\hat{p}(x)}$, the model prediction $\hat{p}(y|x)$ is strongly coupled with the label distribution $\hat{p}(y)$, thus the shift in $\hat{p}(y)$ can cause erroneous prediction of $\hat{p}(y|x)$.

Regarding this problem, our goal is to adapt a classifier that is learned from the training set to perform well on any unseen test set with label distribution shift.

2.2 Learning Diverse Classifiers via Distribution Calibration

Since the test label distribution can be arbitrary, to generalize models under label shift, we consider it is important to enlarge the capacity of classifiers to a broad range of label distributions. However, during training, the model is only presented to a fixed training label distribution thus has limited capacity. Inspired by balanced softmax [24] which calibrates skewed label distribution to be uniform by adding a compensating term to the softmax loss, we propose to learn diverse classifiers via dedicated distribution calibration. As shown in Fig. 1, our insight is to simulate representative one-dominating-class distributions so that the proper combination of learned classifiers can handle arbitrary test label distribution.

Before introducing how to achieve distribution calibration, we first clarify the ordinal encoding in our classification task. To encourage classification network to learn the commonness of all classes and the distinctions between different classes, we use ordinal encoding [18] instead of one-hot encoding for the ordinal classes in our liver fibrosis staging and COVID-19 severity prediction tasks. This ordinal encoding performs multiple binary classifications with sigmoid function and combines the multiple binary outputs by taking the highest class that is predicted as 1 as the final prediction. Furthermore, for distribution calibration in our ordinal regression, we extend the balanced softmax to the sigmoid function and derive the corresponding compensating term. Let $p(y_i = 1|x)$ be the desired conditional probability for the expected label distribution, and $\hat{p}(y_i = 1|x)$ be the desired conditional probability of the training set, and assume $p(y_i = 1|x)$ is expressed by the standard sigmoid function of the network output ϕ_i in i-th ordinal vector: $p(y_i = 1|x) = \frac{e^{\phi_i}}{1+e^{\phi_i}}$, then the $\hat{p}(y_i = 1|x)$ with the same output ϕ_i can be expressed as:

$$\hat{p}(y_i = 1|x) = \frac{e^{\phi_i - \log\left(\frac{r_i'}{1-r_i'} \cdot \frac{1-r_i}{r_i}\right)}}{1 + e^{\phi_i - \log\left(\frac{r_i'}{1-r_i'} \cdot \frac{1-r_i}{r_i}\right)}}, \tag{1}$$

where r_i' and r_i are the positive label proportion in the i-th ordinal vector for the expected label distribution p and factual label distribution respectively $\hat{p}$, and the term $\log\left(\frac{r_i'}{1-r_i'} \cdot \frac{1-r_i}{r_i}\right)$ is the compensating term. The proof of Eq. (1) is provided in the supplementary material.

In this way, the calibrated loss function is:

$$\bar{\mathcal{L}}_{cal} = -\sum_{i=1}^{K-1} \left(y_i \log \hat{p}(y_i = 1|x) + (1 - y_i) \log\left(1 - \hat{p}(y_i = 1|x)\right)\right), \tag{2}$$

where K denotes the total number of classes. This calibrated loss function enables the model learned on the training label distribution to generate the prediction for the expected label distribution.

Moreover, we aim to properly construct different r_i' to simulate K one-dominating-class distributions for K classifiers. Assume the proportion of dominating class j is λ times other classes, then the value of r_i' can be calculated as:

$$r_i' = 1 - \frac{i - \mathbf{1}_{i \geq j} \cdot (1 - \lambda)}{\lambda + K - 1}, \tag{3}$$

where $\mathbf{1}_{i \geq j}$ is the indicator function. The derivation of Eq. (3) can be found in the supplementary material. Notably, our distribution-calibrated networks use independent parameters only at the last stages and fully-connected layer of networks, while share the parameters at other layers (see the shared network χ_θ and the independent networks ψ_θ in Fig. 1). This is motivated by the observation that decoupling the representation learning and classification gives more generalizable representations [10]. In this way, we obtain diverse classifiers to handle different label distributions, but adding only minimal computational cost.

2.3 Test-Time Adaptation for Dynamic Classifier Aggregation

After obtaining diverse distribution-calibrated classifiers during training phase, then at test time, the key is how to aggregate these classifiers to handle the unknown test label distribution with the given inference samples. To build the connection between the obtained classifiers and the test data, we aggregate the outputs of all classifiers with learnable weights, which are dynamically adapted using information implicitly provided by the test data. It's worth to mention that a set of test data, which can reflect the label distribution in the test center, should be accessible simultaneously during this phase.

Specifically, the aggregated output is defined as $\hat{p}_{\text{agg}} = \sum_{k=1}^{K} w_k \hat{p}_k$, where $\sum_{k=1}^{K} w_k = 1$ and $\hat{p}_k$ is the output of k-th classifiers with the form of ordinal vector. As different combination of $\{w_1, w_2, ..., w_K\}$ can enable the model to deal with different test label distributions, the aim of our test-time adaptation is to find the optimal combination for a given test set. Our assumption is that if the aggregated model has adapted to a particular test label distribution, for the test images generated from such a label distribution, the model should give similar predictions to perturbed versions of the same image. Based on this assumption, we design a consistency regularization mechanism to drive the test-time learning. Given an input x, we generate two augmented views $g(x) = v_1$ and $g'(x) = v_2$ using the data augmentation approaches, including rotating, flipping, and shifting the images, and adding Gaussian noise to the images. The two views are then forwarded to the trained model $f(\cdot)$ respectively, yielding the ordinal encoded output $f(v_1) = \hat{p}_{\text{agg}} = w_1 \cdot \hat{p}_1 + w_2 \cdot \hat{p}_2 + \cdots + w_K \cdot \hat{p}_K$ and $f(v_2) = \hat{p}'_{\text{agg}} = w_1 \cdot \hat{p}'_1 + w_2 \cdot \hat{p}'_2 + \cdots + w_K \cdot \hat{p}'_K$. The consistency regularization for the outputs of the two views is imposed with a cosine similarity loss:

$$\mathcal{L}_{\text{test}} = -cos(f(v_1), f(v_2)) = -\frac{f(v_1) \cdot f(v_2)}{\|f(v_1)\|_2 \times \|f(v_2)\|_2}, \tag{4}$$

The loss $\mathcal{L}_{\text{test}}$ drives the updates of the weights set $\{w_1, w_2, ..., w_K\}$ with the implicit knowledge of label distribution on the test set, while the other network

parameters of $f(\cdot)$ are frozen. This implicit knowledge is reflected by the consistency that measures whether the aggregated model has adapted to the test label distribution successfully. Each weight of $\{w_1, w_2, ..., w_K\}$ is initialized to $\frac{1}{K}$ and we use softmax function to maintain the sum of them equals to one after each iteration. As a result, the test results can be obtained after the test-time adaptation given the optimized weights set.

3 Experiment

3.1 Dataset and Experimental Setup

Datasets. We have validated our proposed method on two tasks: 1) liver fibrosis staging with an in-house abdominal CT dataset, and 2) COVID-19 severity prediction with a public chest CT dataset (iCTCF [17]). The liver CT dataset consists of three centers with different label distributions, including 823 cases from our center, 99 cases from external center A and 50 cases from external center B. The ground truths of the liver fibrosis staging come from the pathology results of liver biopsy. The liver fibrosis disease is divided into 5 stages, including no fibrosis (F0), portal fibrosis without septa (F1), portal fibrosis with few septa (F2), numerous septa without cirrhosis (F3) and cirrhosis (F4). Segmentation of the liver is pre-computed with an out-of-the-box tool in a related clinical study, so we adopt it in our paper as the region of interest for classification. The slice thickness of the CT images is 5 mm and the in-plane resolution is 512×512. For the COVID-19 dataset, it contains 969 cases from HUST-Union Hospital for training and 370 cases from HUST-Liyuan Hospital for test. The severity of COVID-19 is divided to 6 levels: control (S0), suspected (S1), mild (S2), regular (S3), severe (S4) and critically (S5). The preprocessing and automatic lung segmentation process are the same as a recent work [2] on this dataset.

Experimental Setting. For liver fibrosis staging, we take 630 cases from our center as the training set, 193 cases from our center as evaluation set and the data from two external centers as two different test sets. For COVID-19 severity prediction, we use the data from HUST-Union Hospital for training and data from HUST-Liyuan Hospital for test. Label distribution statistics of different centers for both datasets are provided in supplementary.

Evaluation Metrics. For both tasks, the diagnosis performance is evaluated with accuracy, area under the receiver operating characteristic curve (AUC) and Obuchowski index (OI) [19], as reported in related works [2,5,21]. Considering the AUC is defined for binary classification while ours are multi-class classification tasks, we combine the classes and convert the multi-class classification to several binary classifications. Specifically, we calculate the AUC of F0 vs F1-4, F0-1 vs F2-4, F0-2 vs F3-4 and F0-3 vs F4 for the liver fibrosis staging, and the AUC of S0 vs S1-5, S0-1 vs S2-5, S0-2 vs S3-5, S0-3 vs S4-5 and S0-4 vs S5 for COVID-19 severity prediction. We report the average of all the AUC values as overall performance. The Obuchowski index (OI) is a metric which is proved to

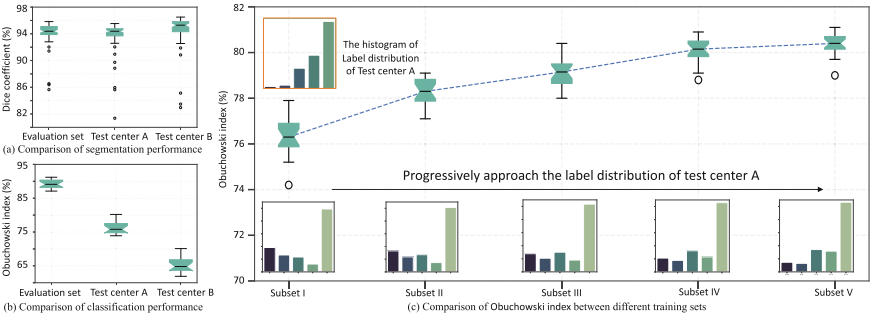

Fig. 2. Analysis of model performance with label distribution shift.

have no bias when label distributions are different between training and test sets [12].

Implementation Details. Considering model efficiency while still capturing 3D information in CT scans, we use ResNet-50 to get a vector of spatial features and then forward the features of adjacent slices to a LSTM module and a fully connected layer for classification. We train the models using Adam with an initial learning rate of $1e-5$, a weight decay of $1e-4$ and batch size of 4. Our models are implemented using a workstation with four NVIDIA TITAN Xp GPUs.

3.2 Experimental Results

Observation of Label Distribution Shift. Label distribution shift and data distribution shift are two types of dataset shift, as introduced in previous work [27]. We first clearly show in the multi-center liver CT datasets that under label distribution shift, the performance of classification model would degrade. Figure 2(a) and (b) present that the segmentation performance for region-of-interest liver extraction is consistent between the evaluation set and test sets, while the final classification performance of Obuchowski index largely decreases by 12.9% and 27.5% at test set A and B. It worth to mention that the label distribution of evaluation set is consistent with the training set while the test sets are not. In Fig. 2(c), we progressively adjust the class distribution of training set to approach the label distribution of test set A, by random sampling a certain proportion of images belonging to each class. We can see that the classification performance increases when the class distribution of the training set becomes closer to the test set. These experiments clearly demonstrate it is indeed the label shift causes the performance drop of classification model in our datasets.

Comparison with State-of-the-Art Methods. We here compare our method with state-of-the-art approaches for label shift in natural images as strong competitors, including **BALMS** [24], which calibrates the training label distribution to be uniform, **LADE** [8], which disentangles the training label

Table 1. Quantitative comparison of different methods on the test sets of the two tasks. Results are reported with average and standard deviation over three independent runs.

Methods	Task 1: Liver fibrosis staging						Task 2: COVID-19 Severity prediction		
	Test center A			Test center B					
	AUC	Accuracy	OI	AUC	Accuracy	OI	AUC	Accuracy	OI
Baseline	77.7 ± 0.7	52.5 ± 0.8	76.3 ± 0.5	68.8 ± 0.6	40.7 ± 0.9	66.3 ± 0.5	68.4 ± 0.8	36.2 ± 1.0	65.2 ± 0.6
BALMS [24]	80.3 ± 0.5	54.9 ± 0.5	78.3 ± 0.5	70.1 ± 0.7	44.0 ± 1.6	67.0 ± 0.4	69.5 ± 0.6	36.2 ± 1.0	66.4 ± 0.5
LADE [8]	80.6 ± 0.5	57.6 ± 0.8	78.5 ± 0.5	69.3 ± 0.6	46.0 ± 1.6	67.9 ± 0.5	68.3 ± 0.6	37.2 ± 0.9	66.2 ± 0.5
TADE [34]	80.9 ± 0.6	59.9 ± 1.9	79.2 ± 0.5	70.2 ± 0.8	47.3 ± 0.9	68.5 ± 0.7	69.6 ± 0.8	38.3 ± 1.6	68.4 ± 0.6
TENT [29]	78.9 ± 0.8	53.2 ± 0.5	77.0 ± 0.7	69.9 ± 0.5	42.7 ± 0.9	67.1 ± 0.5	69.1 ± 0.8	36.4 ± 0.8	65.6 ± 0.7
Focal Loss [13]	80.2 ± 0.6	53.2 ± 0.5	78.0 ± 0.5	69.1 ± 0.7	43.3 ± 0.9	67.7 ± 0.6	69.5 ± 0.6	36.5 ± 1.0	66.5 ± 0.5
TTADC (ours)	$\mathbf{82.3 \pm 0.4}$	$\mathbf{61.0 \pm 1.0}$	$\mathbf{80.2 \pm 0.4}$	$\mathbf{72.4 \pm 0.6}$	$\mathbf{50.7 \pm 0.9}$	$\mathbf{69.6 \pm 0.4}$	$\mathbf{71.1 \pm 0.6}$	$\mathbf{40.2 \pm 1.1}$	$\mathbf{69.8 \pm 0.5}$

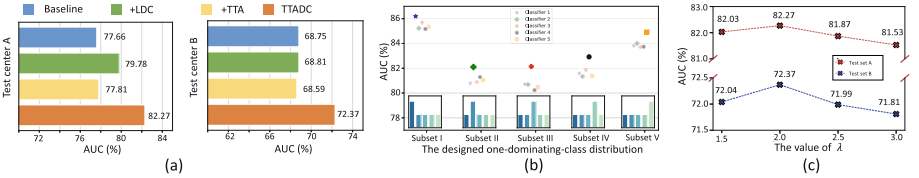

Fig. 3. Ablation analysis of our method on liver CT dataset. (a) Contribution of LDC and TTA in our method; (b) Performance of our learned diverse classifiers on different one-dominating-class distributions; (c) Effect of the value of λ on model performance.

distribution from the model prediction, and **TADE** [34], which also proposes to train multiple networks with different expertise but their networks are less representative than ours. Note that BALMS, LADE, and TADE need to use our derived compensating term in Eq. 1 to be applied in our classification tasks with ordinal regression. We also compare our method with **TENT** [29], which is a general test-time adaptation approach for domain shift problem, and **Focal Loss** [13], which can alleviate class imbalance by increasing the focus on hard samples.

Table 1 presents the comparison results on the test centers of both liver fibrosis staging and COVID-19 severity prediction. Our TTADC significantly improves the model performance over baseline on all test sets, with 4.6%, 3.6%, 2.7% increase in AUC, 8.5%, 10.0%, 4.0% increase in Accuracy, and 3.9%, 3.3%, 4.6% increase in OI respectively, outperforming all the comparison methods. The results validate the effectiveness of our distribution calibration and test-time adaptation on addressing arbitrary label shift. Our method clearly outperforms the domain adaptation method TENT, showing the necessity of designing approach specifically for label shift. Although not significant, Focal loss can also generally improve over baseline, indicating the alleviation of class imbalance may help reduce the effect of label shift. The other methods on tackling label shift generally outperform TENT and Focal loss. Our method and TADE which learn multiple classifiers obtain better performance than BALMS and LADE which use uniform distribution assumption, showing the importance of enlarg-

ing the model capacity for proper test-time adaptation. Our method also clearly outperforms TADE, demonstrating the combination of our one-dominating-class distributions can represent more diverse test label distributions.

Ablation Analysis. Comprehensive ablation studies have been conducted with the liver CT dataset to analyze the key ingredients regarding our TTADC. As shown in Fig. 3(a), adding only the learning distribution-calibrated classifier (LDC) or test-time adaptation (TTA) over baseline is not able to improve over baseline. This is as expected since the two key components are strongly coupled, i.e., the diverse classifiers need to be properly aggregated at test time for the unknown label distribution. In Fig. 3(b), we manually sample a few images from the training center to construct the test subsets with different one-dominating-class distributions. We can see that given the test subset with k-th class dominating other classes, the best performance comes from the k-th classifier, demonstrating that our proposed distribution calibration successfully generate classifiers that have expertise on different one-dominating-class distributions. Moreover, Fig. 3(c) compares the model performance trained with different λ in Eq. 3. The results show that the optimal choice of λ is 2.

4 Conclusion

We present, to our best knowledge, the first method to generalize deep classifiers to unknown test label distributions for medical image classification. Our methods innovates distribution calibration to learn multiple representative classifiers during training, which are then dynamically aggregated via test-time adaptation to deal with arbitrary label shift. Our method is general and experiments on two important medical diagnosis tasks demonstrate the effectiveness of our method.

Acknowledgement. This work was supported in part by the Hong Kong Innovation and Technology Fund (Project No. ITS/238/21), in part by the CUHK Shun Hing Institute of Advanced Engineering (project MMT-p5-20), in part by the Shenzhen-HK Collaborative Development Zone, in part by Jilin Provincial Key Laboratory of Medical Imaging & Big Data (20200601003JC), Radiology, and in part by Technology Innovation Center of Jilin Province (20190902016TC).

References

1. Azizzadenesheli, K., Liu, A., Yang, F., Anandkumar, A.: Regularized learning for domain adaptation under label shifts. In: International Conference on Learning Representations (2019)
2. Bao, G., et al.: COVID-MTL: multitask learning with Shift3D and random-weighted loss for COVID-19 diagnosis and severity assessment. Pattern Recogn. **124**, 108499 (2022)
3. Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., Tsaneva-Atanasova, K.: Artificial intelligence, bias and clinical safety. BMJ Qual. Saf. **28**(3), 231–237 (2019)

4. Chen, I.Y., Joshi, S., Ghassemi, M., Ranganath, R.: Probabilistic machine learning for healthcare. Annu. Rev. Biomed. Data Sci. **4**, 393–415 (2021)
5. Choi, K.J., et al.: Development and validation of a deep learning system for staging liver fibrosis by using contrast agent-enhanced CT images in the liver. Radiology **289**(3), 688–697 (2018)
6. Davis, S.E., Lasko, T.A., Chen, G., Siew, E.D., Matheny, M.E.: Calibration drift in regression and machine learning models for acute kidney injury. J. Am. Med. Inform. Assoc. **24**(6), 1052–1061 (2017)
7. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **42**(4), 463–484 (2011)
8. Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., Chang, B.: Disentangling label distribution for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6626–6636 (2021)
9. Hussein, S., Kandel, P., Bolan, C.W., Wallace, M.B., Bagci, U.: Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches. IEEE Trans. Med. Imaging **38**(8), 1777–1787 (2019)
10. Kang, B., et al.: Decoupling representation and classifier for long-tailed recognition. In: International Conference on Learning Representations (2020)
11. Konwer, A., et al.: Attention-based multi-scale gated recurrent encoder with novel correlation loss for COVID-19 progression prediction. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 824–833. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_79
12. Lambert, J., Halfon, P., Penaranda, G., Bedossa, P., Cacoub, P., Carrat, F.: How to measure the diagnostic accuracy of noninvasive liver fibrosis indices: the area under the ROC curve revisited. Clin. Chem. **54**(8), 1372–1378 (2008)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
14. Liu, M., Zhang, D., Shen, D.: Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment. IEEE Trans. Med. Imaging **35**(6), 1463–1474 (2016)
15. Mesejo, P., et al.: Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE Trans. Med. Imaging **35**(9), 2051–2063 (2016)
16. Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., et al.: A unifying view on dataset shift in classification. Pattern Recogn. **45**(1), 521–530 (2012)
17. Ning, W., et al.: Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. Nat. Biomed. Eng. **4**(12), 1197–1207 (2020)
18. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output CNN for age estimation. In: CVPR, pp. 4920–4928 (2016)
19. Obuchowski, N.A., Goske, M.J., Applegate, K.E.: Assessing physicians' accuracy in diagnosing paediatric patients with acute abdominal pain: measuring accuracy for multiple diseases. Stat. Med. **20**(21), 3261–3278 (2001)
20. Park, C., Awadalla, A., Kohno, T., Patel, S.: Reliable and trustworthy machine learning for health using dataset shift detection. In: NeurIPS, vol. 34 (2021)
21. Park, H.J., et al.: Radiomics analysis of gadoxetic acid-enhanced MRI for staging liver fibrosis. Radiology **290**(2), 380–387 (2019)

22. Peng, J., Bu, X., Sun, M., Zhang, Z., Tan, T., Yan, J.: Large-scale object detection in the wild from imbalanced multi-labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9709–9718 (2020)
23. Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X.: Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 201–209. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_23
24. Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al.: Balanced meta-softmax for long-tailed visual recognition. Adv. Neural. Inf. Process. Syst. **33**, 4175–4186 (2020)
25. Roy, S., et al.: Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. IEEE Trans. Med. Imaging **39**(8), 2676–2687 (2020)
26. Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., Mooij, J.: On causal and anticausal learning. In: ICML (2012)
27. Subbaswamy, A., Saria, S.: From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics **21**(2), 345–352 (2020)
28. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning, pp. 9229–9248. PMLR (2020)
29. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations ICLR (2021)
30. Wang, X., Lian, L., Miao, Z., Liu, Z., Yu, S.X.: Long-tailed recognition by routing diverse distribution-aware experts. In: International Conference on Learning Representations (2021)
31. Williams, R.: Global challenges in liver disease. Hepatology **44**(3), 521–526 (2006)
32. Wu, R., Guo, C., Su, Y., Weinberger, K.Q.: Online adaptation to label distribution shift. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
33. Zhang, K., Schölkopf, B., Muandet, K., Wang, Z.: Domain adaptation under target and conditional shift. In: ICML, pp. 819–827. PMLR (2013)
34. Zhang, Y., Hooi, B., Hong, L., Feng, J.: Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. arXiv preprint arXiv:2107.09249 (2021)