Frequency-Aware Multi-Margin Loss for Imbalanced Medical Ordinal Classification

Haojie Yin[1], Tianyi Liu[2], Chengcheng Feng[1], and Kaizhu Huang[1⋆]

[1] Duke Kunshan University
haojie.yin@duke.edu, kaizhu.huang@dukekunshan.edu.cn
[2] Xi'an Jiaotong-Liverpool University

Abstract. Medical ordinal classification aims to predict ordered stages of disease progression. Existing ordinal metric learning methods enforce globally monotonic embeddings with fixed inter-stage margins to preserve stage ordering, implicitly assuming uniform stage transitions. However, real-world medical datasets are typically long-tailed, making severe stages rare yet clinically critical. Applying fixed ordinal constraints leads such imbalanced data to structural incompatibility: fixed margins cause boundary compression for sparse severe stage transitions, while data imbalance biases rank embeddings against rare stages. To address these issues, we propose the Frequency-aware Multi-Margin Loss (FMML). FMML introduces Boundary-Level Scaling to adapt inter-stage margins based on transition sparsity and Anchor-Level Scaling to strengthen supervision for rare stage anchors, effectively preserving ordinal consistency with long-tailed adaptation. Furthermore, since metrics like MAE fail to capture the systematic directinoal bias, we introduce Long-Tailed Rank Deviation (LTRD). This metric evaluates ordinal classification tasks while accounting for both rank errors and data distribution. Experiments on four imbalanced medical grading datasets demonstrate that the proposed method achieves superior performance. The source code is publicly available at https://github.com/yinpub/ltrank2026.

Keywords: Data imbalance · Ordinal classification · Contrastive learning · Distance metric learning · Long-tailed distribution.

1 Introduction

Disease progression is a gradual process, as seen in tumor staging, from early to late stages [14]. Unlike standard multi-class classification, which treats labels as independent categories, ordinal classification acknowledges the inherent rank order of disease stages. This approach is clinically relevant as it reflects progressive pathological shifts inherent in disease maturation. By preserving this latent ordinal structure, models can better capture biological progression and ensure predictions consistent with the disease trajectory [9].

⋆ Corresponding author

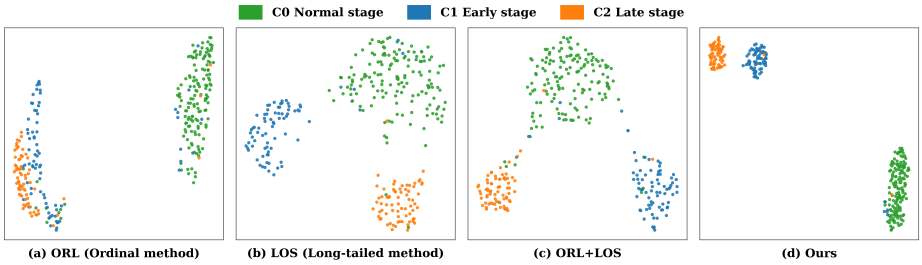

Fig. 1. UMAP visualization of the learned feature embeddings on the GAMMA dataset. Different colors present different classes. The number of classes 0 far exceeds that of 1 and 2. The data presents a long-tailed distribution.

To arrange samples according to their ranks in the embedding space, ordinal metric learning methods typically utilize a margin-based triplet loss with a fixed threshold between instances [7, 6]. This constraint ensures that samples from advanced disease stages are separated from healthy samples while remaining closer to adjacent earlier-stage samples. While this global ranking constraint works well under balanced conditions, it faces challenges when confronted with the long-tailed distributions typical of real-world medical datasets. Tail classes representing rare but critical severe stages lack sufficient sample support [3]. Consequently, as shown in Figure 1(a), their feature representations fail to form distinct clusters and are easily compressed by the overwhelmingly large head classes, leading to boundary collapse in the tail region.

Addressing this boundary collapse introduces a fundamental structural conflict. Standard long-tailed solutions, such as reweighting or margin expansion, treat each category independently [17]. They attempt to enlarge the decision space for rare classes to prevent mis-classification [4]. However, ordinal methods inherently couple adjacent classes together to maintain the progressive order. If a long-tailed method independently expands the margin of a specific tail class, it forces that class into the feature space reserved for its adjacent stages, seen in Figure 1(b). Therefore, as shown in Figure 1(c), directly combining long-tailed corrections with ordinal constraints either destroyed the progressive ordering or fails to overcome the imbalance, preventing accurate grading of severe stages.

To resolve this incompatibility, the inter-class margins need to dynamically adapt to the data distribution without breaking the global ordinal constraints. Along this line, we propose Frequency-aware Multi-Margin Loss (FMML). FMML incorporates two frequency-aware balancing strategies. Boundary-Level Scaling introduces the learnable multi-stage margins, adaptively enlarging inter-stage separation for sparse and severe disease transitions and thereby preventing boundary compression by dominant classes. Anchor-Level Scaling strengthens tail gradients at the embedding level. Unlike conventional imbalanced solutions applied mainly at the classifier layer, FMML balances head and tail through ordinal geometry formation, so that long-tailed correction and ordinal metric

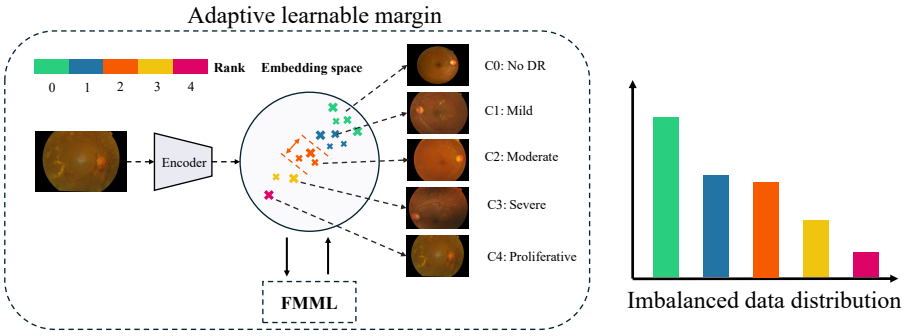

Fig. 2. Overview of the Proposed FMML Framework.

learning become complementary rather than interfering. It facilitates a smoother boundary formation process, ensuring that tail class structure while simultaneously preserving the disease progression ordinal structure, as shown in Fig. 1(d).

Furthermore, traditional metrics such as Accuracy and Mean Absolute Error (MAE) cannot capture systematic directional bias in imbalanced ordinal settings. For example, in tumor grading, MAE equally penalizes misclassifying a normal case (head class) as late stage (tail class) and a late stage case as normal. The latter is clearly more harmful. It is evident that a suitable metric should account for the fact that rank errors on tail classes are more harmful. We therefore introduce Long-Tailed Rank Deviation (LTRD), a frequency-aware metric explicitly designed to measure class-wise rank deviation.

Our contributions are: (1) We propose FMML, a novel frequency-aware multi-margin loss that stabilizes ordinal feature geometry for imbalanced medical classification. (2) We propose LTRD, an ordinal evaluation metric that captures class-wise directional bias under long-tailed distributions. (3) We demonstrate consistent performance gains across four diverse imbalanced medical grading datasets.

2 Method

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, each sample x_i is associated with an ordinal label $y_i \in \{0, 1, \ldots, K-1\}$ with natural ordering $0 \prec 1 \prec \cdots \prec K-1$. Let n_k denote the number of samples in class k. Our objective is to learn an embedding function $f : \mathcal{X} \to \mathbb{R}^d$ that preserves progressive ordinal transitions in the representation space under severe data imbalance.

2.1 Ordinal Metric Learning with Structured Margins

We adopt a margin-based metric learning framework [8, 11] to structure the embedding space for ordinal classification. Given normalized embeddings $z_i = f(x_i)/\|f(x_i)\|$, similarity is defined as $s_{ij} = z_i^\top z_j$. We consider consecutive class

boundaries as elementary separation units. Let $\tilde{m}_h > 0$ denote the learnable margin associated with the boundary between rank h and $h+1$. These margins represent local separation constraints for consecutive stages. Metric learning construct cumulative class positions $c_k = \sum_{h=0}^{k-1} \tilde{m}_h$, which guarantees a strictly ordered layout $c_0 < c_1 < \cdots < c_{K-1}$ along a one-dimensional ordinal axis. Under this construction, the required margin between any two classes becomes the sum of the margins of all consecutive boundaries between them. Specifically, the ordinal margin between samples i and j is defined as $M_{ij} = |c_{y_i} - c_{y_j}|$. Thus, if two samples are separated by multiple intermediate stages, the imposed separation equals the accumulation of all intervening boundary margins. The margins are reparameterized by softplus function to ensure that all values remain positive. This additive structure directly links global pairwise constraints to local consecutive boundaries. In this way, ordinal geometry is encoded directly through boundary-wise constraints that propagate to global pairwise relationships.

An anchor sample is encouraged to be more similar to positive samples than to negative samples by at least a predefined margin m, typically via a hinge constraint of the form

$$L_i = \max(0,\, s_{in} + m - s_{ip}).$$

This formulation enforces a fixed separation between classes in the embedding space. However, a uniform margin is insufficient for ordinal classification settings. A constant margin ignores the data distribution structure inherent in the label space. Tail classes participate less frequently during optimization. As a result, they fail to establish decision boundaries. Simply amplifying the gradients of tail classes at the decision layer disrupt boundary formation, since this strategy remains decoupled from metric learning at the embedding level. Motivated by this observation, we introduce a learnable multi-margin that explicitly facilitates the formation of decision boundaries. The metric objective then replaces the fixed margin m with the ordinal margin M_{ij}, yielding the per-anchor loss

$$\mathcal{L}_i = \frac{1}{|P(i)||N(i)|} \sum_{p \in P(i)} \sum_{n \in N(i)} \max(0,\, s_{in} + M_{in} - s_{ip}).$$

2.2 Frequency-Aware Modulation

To stabilize ordinal geometry under long-tailed distribution, we introduce two complementary mechanisms.

Boundary-Level Scaling (Geometric Stabilization). Even with balanced update strength, boundaries between rare consecutive stages remain vulnerable to compression due to limited data support. For boundary h between class h and $h+1$, we define boundary density $b_h = n_h + n_{h+1}$ and modulate the margin as $\hat{m}_h = \beta_h \tilde{m}_h$, where $\beta_h = (1/b_h)^\delta$ with $\delta \geq 0$. Sparse boundaries therefore receive enlarged effective margins. This explicitly enforces stronger geometric separation between rare adjacent stages, preventing shrinkage of their decision

regions. While anchor-level scaling adjusts optimization behavior, boundary-level scaling directly modifies embedding geometry. Together, they address both representation drift and boundary compression.

Anchor-Level Scaling (Optimization Stabilization). Minority stages participate in fewer positive-negative interactions within each mini-batch, leading to weaker corrective updates and increased risk of representation drift. We therefore scale each anchor's contribution by $\alpha(y_i) = (1/n_{y_i})^\gamma$ with $\gamma \geq 0$, where n_{y_i} denotes the number of samples in class y_i. This scaling increases the effective update strength for rare stages, counteracting optimization imbalance and reducing feature collapse. Importantly, anchor scaling regulates optimization dynamics without modifying the geometric margin structure.

2.3 Frequency-Aware Multi-Margin Loss

Using frequency-modulated cumulative positions $\hat{c}_k = \sum_{h=0}^{k-1} \hat{m}_h$, the ordinal margin between two samples is defined as $M_{ij} = |\hat{c}_{y_i} - \hat{c}_{y_j}|$. For each anchor i, with positive set $P(i)$ and negative set $N(i)$ within the mini-batch as

$$\mathcal{L}_i = \alpha(y_i) \frac{1}{|P(i)||N(i)|} \sum_{p \in P(i)} \sum_{n \in N(i)} \max\big(0,\, s_{in} + M_{in} - s_{ip}\big).$$

Minimizing this hinge term enforces $s_{in} + M_{in} \leq s_{ip}$, ensuring that negative samples are separated from the anchor by at least the ordinal margin. The overall objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_i$. This unified formulation jointly enforces intra-stage compactness and progressive inter-stage separation.

2.4 Long-Tailed Rank Deviation: A Frequency-Aware Ordinal Metric

Standard metrics such as accuracy or MAE are insufficient for long-tailed ordinal classification. In medical grading tasks, it is critical not only to measure ordinal deviation magnitude, but also to detect systematic bias under imbalance distribution. To address these limitations, we propose **Long-Tailed Rank Deviation**, a novel frequency-aware ordinal metric that explicitly captures class-wise rank deviation and directional distortion. For each class c with N_c samples, we define the class-wise rank error as

$$RE_c = \frac{1}{N_c} \sum_{i:y_i=c} |y_i - \hat{y}_i| + \lambda \left| \frac{1}{N_c} \sum_{i:y_i=c} (\hat{y}_i - y_i) \right|,$$

where the first term measures average ordinal deviation and the second term penalizes systematic over or under estimation. The recommended range is 0 to 1. We set λ to 0.8 in our experiments. The bias term becomes large when a minority class is consistently predicted toward dominant neighboring stages. To prevent head-class dominance, we assign weight $w_c = 1/\sqrt{p_c}$ with $p_c = N_c/N$.

The final metric is defined as

$$\text{LTRD} = \frac{\sum_c w_c \, RE_c}{\sum_c w_c}.$$

LTRD effectively evaluates ordinal deviation magnitude under imbalanced data, providing a more sensitive assessment for clinically critical grading tasks.

3 Experimental Results

3.1 Datasets and Implementation Details

We conduct experiments on four publicly imbalanced medical ordinal grading datasets: GAMMA [16], BUSI [1], APTOS 2019, and KOA [10]. GAMMA is a 3-class glaucoma grading dataset that includes two imaging modalities: fundus photography (FP) and optical coherence tomography (OCT). In our experiments, we use only the FP modality. BUSI is a 3-class breast ultrasound dataset (Normal, Benign, Malignant). APTOS 2019 is a 5-class diabetic retinopathy grading benchmark based on retinal fundus images, with labels ranging from No DR to Proliferative DR. KOA is a 5-class knee osteoarthritis grading dataset based on X-ray images (Grade 0 to Grade 4), representing increasing structural degeneration of the knee joint. All experimental results are reported using ACC, MAE, the proposed LTRD, and per-class F1-scores.

We adopt ResNet-50 as the backbone network for all experiments. All input images are resized to 224×224. During training, standard data augmentation strategies are applied, including random horizontal flipping, slight rotation, and color perturbation. We use the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 8. BUSI and GAMMA are trained for 100 epochs, while APTOS and KOA are trained for 80 epochs. We set $\delta = 0.4$ and $\gamma = 0.5$. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

Table 1. Performance comparison on BUSI and GAMMA datasets. The best results are in **bold**, and the second best are underlined.

Method	BUSI						GAMMA					
	Acc↑	MAE↓	LTRD↓	C0	C1	C2	Acc↑	MAE↓	LTRD↓	C0	C1	C2
WCE [5]	0.877	0.174	0.285	0.822	0.903	0.860	0.750	0.280	0.518	0.911	0.500	**0.655**
BAW [2]	0.877	0.185	0.351	0.806	0.900	0.878	0.770	0.270	0.637	0.933	0.596	0.583
BPaCo [3]	0.872	0.174	0.314	0.822	0.906	0.830	0.770	0.240	0.508	0.941	0.542	0.640
LOS [15]	0.872	0.185	0.347	0.800	0.897	0.868	0.780	0.250	0.561	0.933	0.628	0.591
PCOL [13]	0.862	0.195	0.478	0.781	0.892	0.846	0.700	0.310	0.766	0.922	0.592	0.148
ORL [6]	0.877	0.159	0.283	0.870	0.899	0.835	0.780	0.220	0.480	0.941	0.621	0.600
CLOC [12]	0.867	0.174	0.373	0.818	0.900	0.827	0.770	0.240	0.503	0.925	0.577	0.619
Ours	**0.913**	**0.118**	**0.205**	**0.882**	**0.930**	**0.897**	**0.810**	**0.190**	**0.460**	**0.970**	**0.642**	0.652

3.2 Main Results

The quantitative comparisons are reported in Table 1 and Table 2. As observed, our method consistently outperforms all existing approaches across four medical grading benchmarks, achieving the best overall performance in terms of accuracy, MAE, and LTRD.

Table 2. Performance comparison on APTOS and KOA datasets.

Method	APTOS								KOA		
	Acc↑	MAE↓	LTRD↓	C0	C1	C2	C3	C4	Acc↑	MAE↓	LTRD↓
WCE [5]	0.812	0.262	1.019	0.973	0.529	0.707	0.400	0.511	0.675	0.408	**0.474**
BAW [2]	0.822	0.246	1.037	0.973	0.563	0.743	0.267	0.577	0.653	0.451	0.661
BPaCo [3]	0.839	0.208	0.949	0.975	0.500	0.776	0.313	**0.667**	0.670	0.409	0.517
LOS [15]	0.847	0.202	0.918	**0.988**	0.633	0.764	0.375	0.600	0.678	0.406	0.600
PCOL [13]	0.828	0.240	1.139	0.985	0.633	0.781	0.313	0.205	0.596	0.434	0.893
ORL [6]	0.817	0.238	1.230	0.985	0.000	0.730	0.000	**0.667**	0.681	0.390	0.726
CLOC [12]	0.825	0.221	0.889	0.980	0.571	0.762	0.278	0.539	0.673	0.394	0.516
Ours	**0.858**	**0.191**	**0.851**	0.983	**0.667**	**0.804**	**0.412**	0.588	**0.702**	**0.375**	0.491

On the imbalanced subsets, our method achieves the strongest overall class-wise performance, demonstrating its ability to mitigate minority drift and boundary compression under long-tailed distributions. For head classes, our method remains on par with or superior to state-of-the-art baselines, indicating that the gains on tail stages do not compromise dominant class stability. Our approach not only surpasses competing methods in metrics but also achieves well balanced performance across all stages, confirming the effectiveness of jointly modeling ordinal structure and class frequency in long-tailed medical grading tasks. Some long-tailed methods improve tail class performance, but this improvement often comes at the expense of performance on other classes. The tradeoff mainly results from the fact that these methods disrupt the ordinal structure. This further highlights the balanced nature of our method.

3.3 Ablation Studies

We conduct ablation experiments on BUSI and GAMMA (Table 3) to evaluate the contribution of each frequency-aware component. Starting from the multi-margin ordinal contrast baseline, introducing anchor-level scaling consistently improves overall accuracy while reducing both MAE and LTRD, indicating enhanced optimization stability under imbalance. Further incorporating boundary-level scaling leads to additional reductions in ordinal deviation, suggesting improved stage discrimination and more stable inter-stage spacing. The

Table 3. Ablation study on BUSI and GAMMA.

Method	BUSI			GAMMA		
	Acc↑	MAE↓	LTRD↓	Acc↑	MAE↓	LTRD↓
Baseline	0.863	0.188	0.400	0.750	0.250	0.570
w/o Anchor-level Scaling	0.873	0.179	0.392	0.770	0.240	0.552
w/o Boundary-level Scaling	0.908	0.152	0.324	0.780	0.230	0.503
FMML	**0.913**	**0.118**	**0.205**	**0.810**	**0.190**	**0.460**

full FMML model achieves the best performance across all metrics on both datasets, demonstrating that anchor-level and boundary-level scaling provide complementary benefits for long-tailed ordinal learning.

3.4 Parameter Sensitivity of Scaling Factors γ and δ

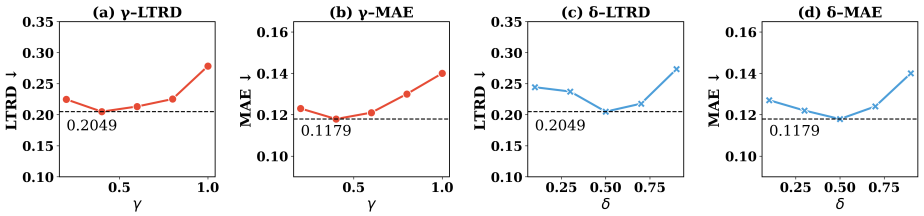

Fig. 3. (a)-(b) Effect of anchor-level scaling. (c)-(d) Effect of boundary-level scaling.

As shown in Fig. 3, both the anchor-level exponent γ and the boundary-level exponent δ exhibit clear non-monotonic effects on LTRD and MAE. Moderate values of γ stabilize minority updates and alleviate representation drift, while appropriate δ enlarges sparse inter-stage margins and mitigates boundary compression. However, excessively large scaling factors lead to over-amplified optimization or geometric distortion, disrupting global ordinal consistency and smooth stage transitions. These results demonstrate that a proper balance between optimization stabilization and geometric modulation is essential for maintaining stable ordinal structure.

4 Conclusion and Future Work

This work identifies a fundamental incompatibility between ordinal learning and long-tailed corrections, which leads to unstable feature geometry and consequently unreliable grading. To address this, we propose FMML, which adaptively

adjusts inter-stage margins while preserving global ordinal consistency. By reshaping ordinal geometry in a frequency-aware manner, FMML enables reliable representation learning under severe imbalance. We further introduce LTRD to quantify rank bias in long-tailed ordinal settings. FMML effectively addresses ordinal classification problems under long-tailed distributions.

Future Work. Distribution mismatch in long-tailed ordinal classification is an important yet underexplored problem. This issue is especially critical in medical grading tasks. For example, during outbreaks of emerging diseases, initial training data may exhibit an extremely long-tailed distribution, dominated by early stage cases. By the time of deployment, the distribution of severe cases may have already shifted. In addition, patients labeled as early stage during training can later progress to severe stages, creating an individual level stage shift. Such evolving distribution shifts impose additional robustness challenges. In the next stage, we will focus on addressing distribution mismatch to improve the robustness of ordinal models in dynamic clinical environments.

Acknowledgments. The work was partially supported by the following: National Natural Science Foundation of China under No. 92370119, and 62376113.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Cai, W., He, J., Guo, X., Sun, H., Tong, H., Wei, B., Wu, H., Yao, D., Guo, D.: Regression-assisted classification for ct-based portal hypertension diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 533–542. Springer (2025)
3. Cai, Z., Wei, T., Lin, L., Chen, H., Tang, X.: Bpaco: Balanced parametric contrastive learning for long-tailed medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 383–393. Springer (2024)
4. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems **32** (2019)
5. 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)
6. Jia, G., Ma, X., Bao, B.K.: Deep orientational representation learning for ordinal regression. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
7. Lee, S.H., Shin, N.H., Kim, C.S.: Geometric order learning for rank estimation. Advances in Neural Information Processing Systems **35**, 27–39 (2022)
8. Lee, S.H., Shin, N.H., Kim, C.S.: Order learning-an overview. APSIPA Transactions on Signal and Information Processing **12**(1), e34 (2023)

9. Lei, Y., Zhu, H., Zhang, J., Shan, H.: Meta ordinal regression forest for medical image classification with ordinal labels. IEEE/CAA Journal of Automatica Sinica **9**(7), 1233–1247 (2022)
10. Mohammed, A.S., Hasanaath, A.A., Latif, G., Bashar, A.: Knee osteoarthritis detection and severity classification using residual neural networks on preprocessed x-ray images. Diagnostics **13**(8), 1380 (2023)
11. Nguyen, B., Morell, C., De Baets, B.: Distance metric learning for ordinal classification based on triplet constraints. Knowledge-Based Systems **142**, 17–28 (2018)
12. Pitawela, D., Carneiro, G., Chen, H.T.: Cloc: Contrastive learning for ordinal classification with multi-margin n-pair loss. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15538–15548 (2025)
13. Saleem, A., Lewis, J.R., Gilani, S.Z.: A hybrid contrastive ordinal regression method for advancing disease severity assessment in imbalanced medical datasets. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–23. Springer (2025)
14. Shen, L., Margolies, L.R., Rothstein, J.H., Fluder, E., McBride, R., Sieh, W.: Deep learning to improve breast cancer detection on screening mammography. Scientific reports **9**(1), 12495 (2019)
15. Sun, S., Lu, H., Li, J., Xie, Y., Li, T., Yang, X., Zhang, L., Yan, J.: Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance. In: The Thirteenth International Conference on Learning Representations (2025)
16. Wu, J., Fang, H., Li, F., Fu, H., Lin, F., Li, J., Huang, Y., Yu, Q., Song, S., Xu, X., et al.: Gamma challenge: glaucoma grading from multi-modality images. Medical Image Analysis **90**, 102938 (2023)
17. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(9), 10795–10816 (2023)