Mitigating Grade Boundary Conflicts for Disease Severity Grading with Ambiguous Labels

Zehui Liao[1,2], Shishuai Hu[1,2], and Yong Xia[1,2(✉)]

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[2] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
yxia@nwpu.edu.cn

Abstract. Accurate disease severity grading is essential in medical image analysis, but label ambiguity due to inter-annotator variability and unclear grade boundaries often reduces model performance. While ordinal prior-based methods address this by incorporating ordering information into model training, they still fail to effectively resolve the underlying conflicts at grade boundaries, which arise from inconsistent grading standards across different annotators. To address this challenge, we propose the Label **H**armonization via **D**econflicted Grade Boundaries and **U**nimodal Projection (**HDU**) framework. HDU introduces two key modules: the Grade Boundaries Harmonizing (GBH) module, which resolves conflicts at grade boundaries by recalibrating labels through prototype-guided re-estimation, and the Unimodal Projection (UniProj) module, which projects calibrated labels onto a unimodal distribution to preserve the ordinal nature of grading tasks. Additionally, an early learning regularization strategy mitigates overfitting to ambiguous labels. We validate HDU through experiments on two disease grading datasets, LIDC-IDRI and Kaggle DR, both of which exhibit label ambiguity. The results show that HDU outperforms existing methods, excelling in handling ambiguous labels while effectively harmonizing grade boundaries and incorporating ordinal constraints.

Keywords: Disease severity grading · Ambiguous labels · Grade boundary conflicts.

1 Introduction

Recent advances in deep neural networks (DNNs) have significantly improved disease severity grading, such as malignancy grading of lung nodules and diabetic retinopathy [17,3,7]. These models excel in performance when trained on large-scale, high-quality annotated datasets. However, lesion annotation requires expert input, which is time-consuming and inherently prone to errors due to ambiguous boundaries between adjacent grades and inter-annotator variability [19,21,20]. Such ambiguity in the assigned labels degrades data quality [16] and consequently undermines model performance and generalization [14,13].

Consequently, enhancing model robustness in the presence of ambiguous labels is crucial for achieving accurate and reliable disease grading.

In disease severity grading, label ambiguity typically manifests as samples being associated with candidate labels that are subjective and mutually inconsistent [12,10,16]. A common approach is to generate ***proxy labels*** via averaging, taking the median, majority voting, or label sampling [12,24,25,9]. While these methods help resolve annotation discrepancies, they often fail to fully exploit the information contained in multiple candidate labels and may even introduce additional errors, thereby degrading model performance. Beyond proxy label generation, some studies adopt ***data selection strategies*** to reduce the impact of ambiguous labels by emphasizing reliable samples during training [16,10]. Representative frameworks leverage annotation consistency [16] or uncertainty estimation [10] to select reliable data and adopt divide-and-conquer training schemes [16] or curriculum learning [10] to guide the model toward more stable and discriminative representations. ***Ordinal prior-based methods*** further incorporate the inherent ordinal relationships between adjacent grades [2] into model training by constraining predicted label distributions to be unimodal or by generating soft labels based on Gaussian assumptions over ordinal categories [1,22]. However, existing methods primarily focus on improving robustness through reliable data selection or ordinal priors constraints, without directly addressing the underlying conflict at grade boundaries caused by inconsistent grading standards among different annotators. Therefore, we advocate addressing this issue from the perspective of harmonizing grade boundaries in multi-annotator disease severity grading data.

In this paper, we propose the Label **H**armonization via **D**econflicted Grade Boundaries and **U**nimodal Projection (**HDU**) framework to address label ambiguity in disease severity grading. Within HDU, we introduce two key modules: the Grade Boundaries Harmonizing (GBH) module and the Unimodal Projection (UniProj) module. The GBH module captures representative prototype features for each grade in every training batch. By computing the similarity between sample features and prototypes, GBH re-estimates probabilistic labels, harmonizing grade boundaries in a data-driven manner. To ensure reliable prototypes at the outset, the network undergoes a warm-up phase before training. The original labels are then calibrated using the prototype-guided labels. To enforce the ordinal nature of disease severity grading, the UniProj module projects the calibrated labels onto a unimodal distribution. These projected labels are then used to supervise model training. Additionally, to mitigate overfitting to ambiguous labels in later training stages, we incorporate an early learning regularization strategy [18]. We conducted experiments on two grading datasets, LIDC-IDRI and Kaggle DR, both of which involve ambiguous label issues. The results show that HDU outperforms existing methods in addressing ambiguous labels, underscoring the importance of resolving grade boundary conflicts and accounting for the unique characteristics of grading tasks. Furthermore, the effectiveness of each key component in the HDU framework is also validated.

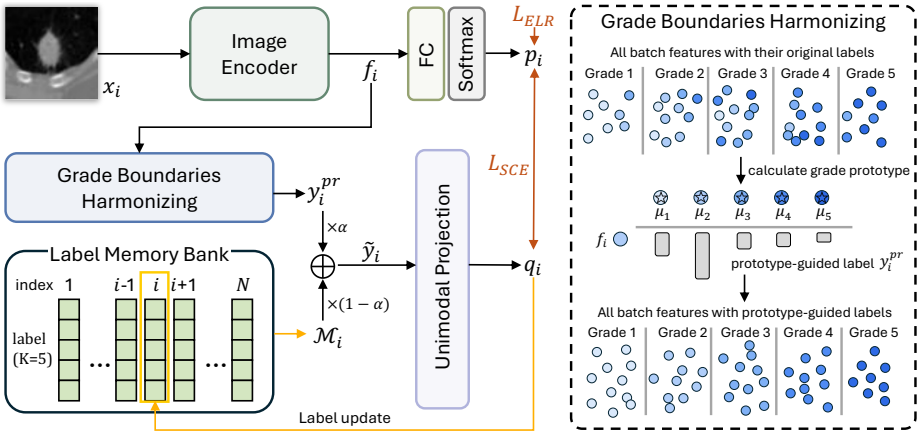

Fig. 1. Overview of the proposed HDU framework, which includes a backbone network, a grade boundaries harmonizing module, and a unimodal projection module. Given an input image x_i, its label is calibrated and projected, then used to supervise model training, alongside early learning regularization.

Our contribution are three-fold: (1) We propose the HDU framework, introducing the Grade Boundaries Harmonizing module to resolve conflicts at grade boundaries and enhance label consistency through prototype-guided re-estimation. (2) We introduce the Unimodal Projection module, which projects calibrated labels onto a unimodal distribution, effectively incorporating the ordinal nature of grading tasks and outperforming prediction-level constraints used in other methods. (3) Extensive experiments demonstrate that HDU outperforms existing methods in handling ambiguous labels, providing superior performance by harmonizing grade boundaries and ensuring reliable predictions.

2 Method

2.1 Problem Formalization and Method Overview

For a K-grade disease severity task, the training set $D = \{(\boldsymbol{x}_i, Y_i)\}_{i=1}^{N}$ consists of N image-label pairs, with $\boldsymbol{x}_i \in \mathcal{R}^{C\times H\times W}$ representing the i-th image of C channels and spatial dimensions $H \times W$. The candidate labels for $\boldsymbol{x}_i$ are $Y_i = \{y_i^j\}_{j=1}^{N_i}$, where $y_i^j \in [K]$ and N_i is the number of annotators, ranging from one to multiple. The goal is to train a grading model that is robust to label ambiguity, using D with ambiguous labels.

As shown in Fig. 1, our HDU framework consists of a backbone network, the GBH module, the UniProj module. Each training iteration follows three main steps: First, the GBH module recalibrates the original labels by generating soft labels based on the similarity to class prototypes. Second, the corrected labels are processed by the UniProj module, which projects them onto a unimodal

distribution to preserve the ordinal nature of the grading task. Finally, the grading loss is computed using the projected labels, along with the early learning regularization term to mitigate overfitting.

2.2 Backbone Network

The backbone network consists of an encoder $\mathcal{F}_E(\Theta_E)$ and fully connected (FC) layers $\mathcal{F}_{FC}(\Theta_{FC})$, where Θ_E and Θ_{FC} represent their respective parameters. The encoder is composed of several convolutional layers followed by an average pooling layer. Given an input image $\boldsymbol{x}_i$, the encoder extracts its feature $\boldsymbol{f}_i = \mathcal{F}_E(\boldsymbol{x}_i; \Theta_E)$, which is then passed through the FC layers and a softmax function $\mathcal{S}$ to compute the probability output for each grade $\boldsymbol{p}_i = \mathcal{S}(\mathcal{F}_{FC}(\boldsymbol{f}_i; \Theta_{FC}))$.

During training, we maintain a ***label memory bank*** $\mathcal{M} = \{\mathcal{M}_i\}_{i=1}^N$, where each $\mathcal{M}_i \in [0,1]^K$ is initialized with the normalized frequency of each grade label in the candidate label set Y_i, such that the sum of its elements equals 1. The values in $\mathcal{M}$ are updated with the calibrated probabilistic labels during each training iteration [11]. In each training round, grade prototypes are computed from the feature of training samples. To initialize the prototypes reliably in the first round, we perform a ***warm-up phase*** on the backbone network. During this phase, sample labels are derived from the initial values in $\mathcal{M}$, and the loss is computed using soft cross-entropy (SCE) with probabilistic labels.

2.3 Grade Boundaries Harmonizing

To resolve grade boundary conflicts during training, we construct grade prototypes [26] in each batch and generate pseudo-labels for all samples. Let the batch size be B and feature dimension d. For sample x_i, we perform L_2 normalization on its feature $\boldsymbol{f}_i$, resulting in $\hat{\boldsymbol{f}}_i$. Given the probabilistic label $\mathcal{M}_i$, we assign each sample $\boldsymbol{x}_i$ to the grade with the highest probability $\hat{c}_i = \arg\max_{k \in \{1,\ldots,K\}} \mathcal{M}_{i,k}$. For each grade, we aggregate features within it to obtain its prototype:

$$\mu_k = \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \hat{\boldsymbol{f}}_i, \quad \mathcal{B}_k = \{i \mid \hat{c}_i = k\}, \quad k = 1, \ldots, K, \tag{1}$$

where $|\mathcal{B}_k|$ is the number of samples in grade k. We use class-distribution-based stratified sampling to ensure all grades are represented in each batch, stabilizing prototype computation. For $\boldsymbol{x}_i$, we compute its similarity to each class prototype:

$$s_{i,k} = \hat{\boldsymbol{f}}_i^\top \mu_k, \quad k = 1, \ldots, K, \tag{2}$$

and convert these similarities into prototype-guided pseudo labels using softmax $\mathbf{y}_i^{pr} = \mathcal{S}\left(\frac{\mathbf{s}_i}{\tau}\right)$, where τ is the temperature coefficient, set to 0.1 in this study. To calibrate the original label $\mathcal{M}_i$, we fuse it with $\mathbf{y}_i^{pr}$ via weighted summarizing:

$$\tilde{\mathbf{y}}_i = (1 - \alpha) \times \mathcal{M}_i + \alpha \times \mathbf{y}_i^{pr}, \tag{3}$$

where $\alpha \in [0,1]$ serves as a weight factor.

2.4 Unimodal Projection

Although the prototype-guided label $\tilde{\mathbf{y}}_i$ alleviates boundary conflicts, it may still violate the unimodal property required by ordinal grading tasks. To enforce ordinal consistency, we introduce a unimodal projection module that aligns $\tilde{\mathbf{y}}_i$ with the semantic ordering of grades. Unlike existing approaches that impose unimodality constraints on model predictions [1,2], UniProj projects the soft labels themselves onto the unimodal space, thereby providing more direct and explicit ordinal supervision during training. Given $\tilde{\mathbf{y}}_i \in [0,1]^K$ (denoted as $\mathbf{y}$ in this subsection for simplicity), we solve:

$$\mathbf{q} = \arg\min_{\mathbf{q}\in\mathcal{U}} \|\mathbf{q}-\mathbf{y}\|_2^2, \quad \text{s.t. } \mathbf{q} \geq 0,\ \sum_{k=1}^{K} q_k = 1, \tag{4}$$

where $\mathcal{U}$ denotes the set of unimodal probability distributions that follow a "first increasing, then decreasing" pattern. To enforce unimodality, we adopt the ***piecewise isotonic regression*** strategy. For each possible peak position $k \in \{1,\ldots,K\}$, we constrain $\{y_1,\ldots,y_k\}$ to be monotonically increasing and $\{y_k,\ldots,y_K\}$ to be monotonically decreasing. The two segments are merged at the peak and normalized to ensure a valid probability distribution. We evaluate all peak candidates and select the solution with the smallest Euclidean distance to the original distribution. This projection enforces strict unimodality while preserving the overall shape of the original distribution, effectively eliminating multimodal artifacts caused by annotator disagreements. In implementation, we use the classic ***pool adjacent violators*** algorithm for monotonic regression, and the projected label of $\tilde{\mathbf{y}}_i$ is denoted as $\boldsymbol{q}_i$.

2.5 Loss Function and Label Memory Bank Update

The loss function consists of a primary grading loss $\mathcal{L}_{\text{SCE}}$ and a early learning regularization term $\mathcal{L}_{\text{ELR}}$. For sample x_i, the loss is calculated as:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{SCE}}(\boldsymbol{p}_i, \boldsymbol{q}_i) \;+\; \lambda\, \mathcal{L}_{\text{ELR}}(i, \boldsymbol{p}_i)\,, \tag{5}$$

where λ is the regularization weight coefficient. While the $\mathcal{L}_{\text{SCE}}$ is the SCE loss between the prediction $\boldsymbol{p}_i$ and label $\boldsymbol{q}_i$, $\mathcal{L}_{\text{ELR}}$ maintains the stability between current prediction $\boldsymbol{p}_i$ and historically stable prediction $\boldsymbol{t}_i$, mitigating the overfitting caused by noisy gradients. Specifically, $\boldsymbol{t}_i$ is calculated in each epoch using a momentum-based approach:

$$\mathbf{t}_i \;\leftarrow\; \beta\,\mathbf{t}_i + (1-\beta)\,\frac{\boldsymbol{p}_i}{\sum_{k=1}^{K} p_{i,k}}, \tag{6}$$

where $\beta \in [0,1)$ controls the retention of historical information. The ELR term is then defined as:

$$\mathcal{L}_{\text{ELR}} \;=\; \log\Bigl(1 - \boldsymbol{t}_i^{\top}\boldsymbol{p}_i\Bigr). \tag{7}$$

Label Memory Bank Update. After each forward pass, we update the label memory bank $\mathcal{M}$ with the latest projected labels $\mathbf{q}_i$ via $\mathcal{M}_i \leftarrow \mathbf{q}_i$. This update enables iterative refinement of sample-level pseudo-labels, allowing them to progressively approach the underlying true distribution during training.

3 Experiments and Analysis

3.1 Experimental Setup

Datasets. Our HDU is evaluated on the LIDC-IDRI and Kaggle DR datasets. The **LIDC-IDRI** dataset [4] includes 1,018 chest CT scans with 2,568 lung nodules, each assigned a malignancy score (1–5) by 1 to 4 radiologists. Nodules are categorized into unreliable and reliable subsets based on annotation consistency. The unreliable subset contains 1,048 nodules with inconsistent annotations and 766 annotated by a single radiologist, while the reliable subset consists of 394 nodules with consistent diagnoses from multiple radiologists. We randomly select 256 reliable nodules as the test set. The remaining 138 reliable nodules, together with all unreliable nodules, form the training set, from which 10% is further used for validation. The **Kaggle Diabetic Retinopathy (DR)** dataset [6] contains 88,702 fundus images from 44,351 patients, with left and right eye images provided for each patient. The images are categorized into five severity levels: Normal, NPDR I, NPDR II, NPDR III, and PDR. Following Ju et al. [10], 57,213 images were re-labeled, among which 917 expert-verified images serve as the gold-standard test set. The remaining 56,296 images retain their original noisy labels and are used for training. Due to class imbalance (42,185 Normal images), we randomly subsample 8,406 Normal images, resulting in a training set of 22,517 samples, from which 10% is used for validation.

Implementation Details and Evaluation Metrics. All images are resized to 224×224. We adopt EfficientNet-B0 [23] as the backbone for LIDC-IDRI and ResNet34 [8] for Kaggle DR, both pretrained on ImageNet [5]. The models are trained for 80 epochs (LIDC-IDRI) and 100 epochs (Kaggle DR) with batch sizes of 128 and 256, respectively. The initial learning rate is set to 1×10^{-4} and decayed exponentially with a rate of 0.95. We use the Adam optimizer with a weight decay of 1×10^{-4}. The hyperparameters of HDU are selected via grid search, with $\lambda = 1.5$ and $\beta = 0.2$. All experiments are conducted on an NVIDIA GeForce RTX 3080 Ti GPU. Results are averaged over three random runs. Performance is evaluated using accuracy, AUC, recall, and F1-score.

3.2 Experimental Results

Comparative Experiments. We compare HDU with several representative methods for handling ambiguous labels in disease severity grading. First, we include four simple baselines that construct proxy labels from multiple annotations using statistical strategies: *Mean Voting*, *Majority Voting*, *Label Sampling*, and *Soft Label*. These methods directly aggregate candidate labels and train a

Table 1. Performance (%) of HDU, four baselines, and five representative methods on the LIDC-IDRI dataset. The best results are in **bold**.

Methods	Accuracy	AUC	F1-Score	Recall
Mean Voting	50.65 ± 1.87	82.78 ± 1.22	46.80 ± 2.68	54.76 ± 3.16
Majority Voting	65.23 ± 0.96	84.96 ± 1.24	52.23 ± 3.54	51.62 ± 3.93
Soft Label	65.36 ± 0.66	88.22 ± 0.30	53.39 ± 1.72	51.47 ± 2.32
Label Sampling	67.19 ± 1.94	88.21 ± 1.02	54.40 ± 1.33	53.61 ± 2.50
ELR+LS	72.92 ± 0.66	88.23 ± 1.26	55.30 ± 1.29	54.02 ± 2.09
ILDE+LS	69.92 ± 1.99	88.64 ± 0.39	57.02 ± 1.48	58.04 ± 1.64
DAR	72.00 ± 1.57	88.85 ± 0.09	56.72 ± 3.37	55.50 ± 3.80
SRJT	73.08 ± 1.03	88.24 ± 0.18	57.56 ± 2.95	56.07 ± 2.68
URL	73.83 ± 1.94	87.41 ± 0.18	58.19 ± 3.66	56.78 ± 3.12
Ours	**74.48** ± 0.66	**91.57** ± 0.94	**62.22** ± 1.30	**62.29** ± 1.15

Table 2. Performance (%) of HDU, a baseline, and four representative methods on the Kaggle DR dataset. The best results are in **bold**.

Methods	Accuracy	AUC	F1-Score	Recall
Baseline	56.49 ± 1.51	80.43 ± 0.77	47.63 ± 1.53	57.75 ± 2.18
ELR	59.47 ± 1.12	80.25 ± 1.19	46.84 ± 2.56	56.19 ± 5.50
ILDE	61.18 ± 1.74	73.23 ± 2.85	44.94 ± 1.95	53.12 ± 7.54
SRJT	58.14 ± 0.87	79.29 ± 1.28	46.25 ± 0.50	58.28 ± 2.33
URL	56.74 ± 1.73	80.86 ± 1.21	47.21 ± 1.69	59.43 ± 1.24
Ours	**62.38** ± 0.44	**83.07** ± 1.71	**48.99** ± 0.83	**62.24** ± 1.04

standard model using cross-entropy loss. Second, we consider two representative methods designed for label noise in classification: *ILDE* [15] and *ELR* [18], which mitigate label errors through noise transition matrix estimation or regularization. Third, we include three methods specifically designed for ambiguous labels in grading tasks: *DAR* [16], *SRJT* [22], and *URL* [1]. DAR emphasizes reliable samples during training, while SRJT and URL further incorporate ordinal priors to constrain model predictions. Compared with these methods, HDU explicitly resolves grade boundary conflicts through prototype-based Grade Boundaries Harmonizing and enforces ordinal consistency via Unimodal Projection. By jointly addressing boundary inconsistency and ordinal structure at the label level, HDU learns more stable and semantically coherent grading boundaries under multi-annotator ambiguity.

Comparative Results. Tables 1 and 2 present the comparison results on LIDC-IDRI and Kaggle DR, both of which contain substantial label ambiguity. Across the two datasets, simple proxy-label baselines and probabilistic aggregation strategies show limited performance, while label-noise-robust methods (ELR, ILDE) and grading-specific approaches (DAR, SRJT, URL) provide moderate improvements. Nevertheless, HDU consistently achieves the best results across all evaluation metrics on both datasets. On LIDC-IDRI, HDU improves AUC and F1-score by 2.93% and 4.03%, respectively, compared to the best competi-

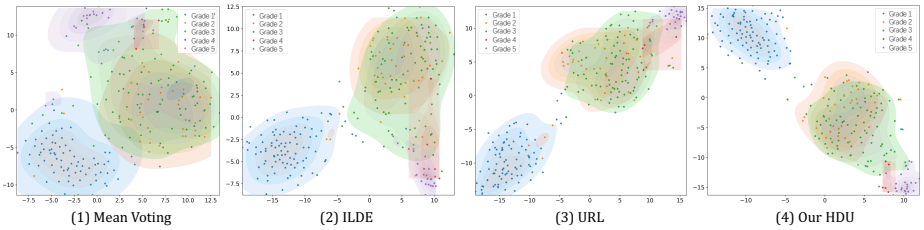

Fig. 2. t-SNE visualization of performance on the LIDC-IDRI test set for the baseline method Mean Voting, comparison methods ILDE and URL, and our HDU.

Table 3. Performance (%) of HDU and its three variants on the LIDC-IDRI dataset. The best results are in **bold**.

HDU				Accuracy	F1-Score
Soft Label	UniProj	ELR	GBH		
√				65.36 ± 0.66	53.39 ± 1.72
√	√			66.02 ± 2.92	54.96 ± 4.19
√	√	√		72.00 ± 1.57	56.72 ± 3.37
√	√	√	√	**74.48** ± 0.66	**62.22** ± 1.30

tor. On Kaggle DR, HDU also achieves consistent gains, improving AUC by 2.21% and F1-score by 1.36% over the strongest baseline. These results demonstrate that explicitly resolving grade boundary conflicts and enforcing ordinal consistency leads to more robust learning under ambiguous multi-annotator supervision. The t-SNE visualization in Fig. 2 shows that the HDU framework clusters samples more compactly for each grade, showing clear advantages over baseline and comparison methods. Compared with URL, HDU achieves comparable recall on most grades while improving Grade 1 recall from 86.59% to 88.65% and Grade 5 recall from 63.15% to 78.94%, showing stronger discrimination of non-severe and severe cases. However, distinguishing between adjacent grades, particularly Grade 1 and Grade 2, remains challenging.

Ablation Study. To evaluate the contribution of each component, we conduct ablation experiments on LIDC-IDRI by progressively adding modules to the Soft Label baseline. Specifically, we compare: (1) Soft Label only; (2) Soft Label + UniProj; (3) Soft Label + UniProj + ELR; and (4) the full HDU framework with GBH. As shown in Table 3, introducing UniProj brings consistent improvements, indicating that enforcing unimodality effectively incorporates ordinal constraints at the label level. Adding ELR further improves performance, demonstrating its ability to mitigate overfitting to ambiguous labels. Finally, incorporating GBH yields the largest performance gain, confirming that resolving grade boundary conflicts is critical for robust grading. The full HDU framework achieves the best results across all metrics, validating the complementary effects of boundary harmonization, ordinal projection, and noise-robust regularization.

4 Conclusion

This paper addresses the issue of ambiguous labels in disease severity grading, focusing on grade boundary conflicts arising from multi-annotator inconsistency. We propose the HDU framework, which harmonizes grade boundaries through prototype-based label calibration and enforces ordinal consistency via unimodal projection. This approach enhances stability and semantic coherence under ambiguous supervision. Extensive experiments on LIDC-IDRI and Kaggle DR datasets show that HDU consistently outperforms existing methods, underscoring the importance of explicitly modeling grade boundary consistency for reliable disease severity grading.

Acknowledgments. This work was supported in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang, China, under Grant 2025C01201(SD2), in part by the National Natural Science Foundation of China under Grant 92470101, in part by the Project of National Key Clinical Specialty (Department of Medical Imaging, No. 2024017), in part by the Ningbo Key Laboratory of Digital Imaging and Medical-Engineering Interdisciplinarity (No. 2024031).

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

References

1. Ai, X., Liao, Z., Xia, Y.: Url: Combating label noise for lung nodule malignancy grading. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 1–11. Springer (2023)
2. Cardoso, J.S., Cruz, R.P., Albuquerque, T.: Unimodal distributions for ordinal regression. IEEE Transactions on Artificial Intelligence **6**(9), 2498–2509 (2025)
3. Chen, L., Gu, D., Chen, Y., Shao, Y., Cao, X., Liu, G., Gao, Y., Wang, Q., Shen, D.: An artificial-intelligence lung imaging analysis system (alias) for population-based nodule computing in ct scans. Computerized Medical Imaging and Graphics **89**, 101899 (2021)
4. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013)
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. Ieee (2009)
6. Dugas, E., Jared, Jorge, Cukierski, W.: Diabetic retinopathy detection (2015), `https://kaggle.com/competitions/diabetic-retinopathy-detection`
7. Gu, D., Liu, G., Xue, Z.: On the performance of lung nodule detection, segmentation and classification. Computerized Medical Imaging and Graphics **89**, 101886 (2021)
8. 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. pp. 770–778 (2016)

9. Jensen, M.H., Jørgensen, D.R., Jalaboi, R., Hansen, M.E., Olsen, M.A.: Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 540–548. Springer (2019)
10. Ju, L., Wang, X., Wang, L., Mahapatra, D., Zhao, X., Zhou, Q., Liu, T., Ge, Z.: Improving medical images classification with label noise using dual-uncertainty estimation. IEEE Transactions on Medical Imaging **41**(6), 1533–1546 (2022)
11. Kurian, N.C., Varsha, S., Bajpai, A., Patel, S., Sethi, A.: Improved histology image classification under label noise via feature aggregating memory banks. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2022)
12. 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)
13. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Transformer-based annotation bias-aware medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 24–34. Springer (2023)
14. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Modeling annotator preference and stochastic annotation error for medical image segmentation. Medical Image Analysis **92**, 103028 (2024)
15. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Instance-dependent label distribution estimation for learning with label noise. International Journal of Computer Vision **133**(5), 2568–2580 (2025)
16. Liao, Z., Xie, Y., Hu, S., Xia, Y.: Learning from ambiguous labels for lung nodule malignancy prediction. IEEE Transactions on Medical Imaging **41**(7), 1874–1884 (2022)
17. Liu, L., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Multi-task deep model with margin ranking loss for lung nodule analysis. IEEE Transactions on Medical Imaging **39**(3), 718–728 (2019)
18. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. Advances in neural information processing systems **33**, 20331–20342 (2020)
19. Ma, Y., Hou, J., Zhang, C., Zhou, Y., Ge, Z., Xie, H., Ju, L.: Benchmarking real-world medical image classification with noisy labels: Challenges, practice, and outlook. arXiv preprint arXiv:2512.09315 (2025)
20. Riotto, E., Tsai, W.S., Khalid, H., Lamanna, F., Roch, L., Manoj, M., Sivaprasad, S.: Intergrader agreement on qualitative and quantitative assessment of diabetic retinopathy severity using ultra-widefield imaging: Inspired study report 1. Diagnostics **15**(14), 1831 (2025)
21. Srinivasan, S., Suresh, S., Chendilnathan, C., Prakash V, J., Sivaprasad, S., Rajalakshmi, R., Anjana, R.M., Malik, R.A., Kulothungan, V., Raman, R., et al.: Inter-observer agreement in grading severity of diabetic retinopathy in wide-field fundus photographs. Eye **37**(6), 1231–1235 (2023)
22. Takezaki, S., Tanaka, K., Uchida, S.: Self-relaxed joint training: Sample selection for severity estimation with ordinal noisy labels. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 368–377. IEEE (2025)
23. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)

24. Wu, B., Sun, X., Hu, L., Wang, Y.: Learning with unsure data for medical image diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10590–10599 (2019)
25. Xu, X., Wang, C., Guo, J., Gan, Y., Wang, J., Bai, H., Zhang, L., Li, W., Yi, Z.: Mscs-deepln: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks. Medical Image Analysis **65**, 101772 (2020)
26. Yang, W., Zhang, R., Chen, J., Wang, L., Kim, J.: Prototype-guided pseudo labeling for semi-supervised text classification. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 16369–16382 (2023)