Weighted Stratification in Multi-Label Contrastive Learning for Long-Tailed Medical Image Classification

Ying-Chih Lin and Yong-Sheng Chen(✉)

Institute of Computer Science and Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
{xup6yj.cs11, yschen}@nycu.edu.tw

Abstract. Multi-label classification (MLC) in medical image analysis presents significant challenges due to long-tailed class distribution and disease co-occurrence. While contrastive learning (CL) has emerged as a promising solution, recent studies primarily focus on defining positive samples, overlooking the low gradient problem associated with single-disease representation and the impact of co-occurring diseases. To address these issues, we propose ws-MulSupCon, a novel weighted stratification method in CL for MLC. Our gradient analysis indicates that separating the single-disease cases can amplify their gradient contributions. Accordingly, we stratify training samples into single- and multi-disease cases to enhance the representation learning of each disease. Moreover, we design a weighted loss function based on class frequency and disease comorbidity, mitigating the dominance of prevalent diseases and improving rare disease detection. To further discriminate between the healthy and diseased samples, a dedicated CL for healthy cases is introduced, improving overall classification performance and preventing false positives. Extensive experiments on NIH ChestXRay14 and MIMIC-CXR demonstrate that ws-MulSupCon outperforms SoTA methods across nearly all disease classes, showing its superiority and the effectiveness of learning long-tailed distribution in multi-label medical image classification. The code is available at `https://github.com/xup6YJ/ws-MulSupCon`.

Keywords: Multi-label classification · Contrastive learning · Long-tailed distribution.

1 Introduction

Multi-label classification (MLC) is a critical challenge in medical image [2,8,18,22] and computer vision [4,13,15,24] domains. MLC is particularly prevalent in medical diagnostics in modalities such as chest X-rays (CXR) and ophthalmoscopy, where a single examination image often captures multiple co-occurring diseases. Additionally, MLC poses greater challenges than single-label classification due to the intricate relationships between diseases, the similarities in disease characteristics on medical images, and the long-tail distribution of the collected data. The

scarcity of rare disease cases results in an imbalanced data distribution, leading to strong performance on prevalent diseases (head classes) while yielding poor results for rare conditions (tail classes) [21]. These issues highlight the inherent complexities of MLC in medical image analysis.

In recent years, contrastive learning (CL) has demonstrated remarkable performance in feature representation learning [3,6,11]. The core of CL lies in defining positive and negative samples to learn and distinguish discriminative features. However, self-supervised CL treats each instance as a unique class, leading to the class collision problem [23] and misclassification of semantically similar samples as negatives. To overcome this problem, SupCon [11] extends CL to supervised settings by utilizing label information for more effective positive/negative sample selection. Nonetheless, determining positive samples in MLC remains a significant challenge compared to single-label classification. While prior efforts have advanced the integration of supervised CL into MLC [1,5,14,20], they primarily focus on the relationship between a sample and its corresponding label set [19], neglecting the complex and diverse data distribution in multi-label scenarios which might hinder the generalization ability of the model.

Recent studies explore the relationship between the labels of different samples. MulSupCon [19] introduces two scenarios in MLC and defines them as **ANY**, where a sample shares at least one label with the anchor, and **ALL** for exactly matching the label set of the anchor. Through gradient analysis, they treat each anchor label independently and construct multiple positive sets for a single anchor sample. Huang *et al.* define the positive samples in **ANY** scenario and design the Similarity-Dissimilarity Loss to consider the various relations between samples and anchors [9]. SoftCon [17] assigns soft similarity scores to each sample pair, degrading the multi-label task into a single-label task. However, these methods still have several limitations in medical image scenarios. First, neglecting the distinction of feature representation between single- and multi-disease cases. Single-disease cases exhibit simpler feature representations, while multi-disease cases present more complex features due to co-occurring diseases. Computing contrastive loss without considering this distinction might limit the ability of the model to learn individual disease representations effectively, thereby limiting its capability to learn multi-disease cases. Second, the characteristics of comorbidities have not been sufficiently explored and incorporated into existing multi-label CL methods. Third, the problem posed by the long-tailed nature of medical image datasets remains largely unaddressed. These issues limit the learning efficacy in multi-label medical image classification.

To conquer these challenges, we proposed a novel CL method for MLC named ws-MulSupCon for seamless integration into convolution-based backbones. The main contributions of this work are as follows: (1) We conduct a gradient analysis and highlight the issue of low gradients in the contrast between anchors and single-disease samples, addressing a critical limitation in existing methods. (2) We propose a novel multi-label CL method that stratifies samples into single- and multi-disease cases, enabling the model to capture precise representations of each disease effectively. (3) To the best of our knowledge, this is the first multi-

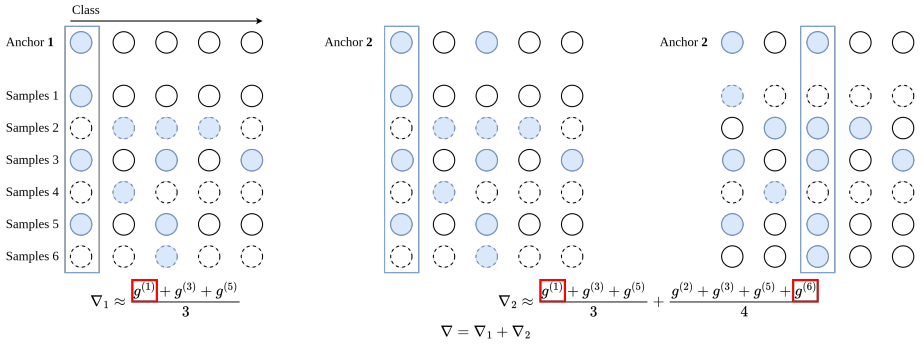

Fig. 1. Gradient illustration for MulSupCon [19]. MulSupCon considers each label separately and forms multiple positive sets for each anchor (blue boxes). Blue dots indicate the classes to which each sample belongs. Rows with circles outlined by dotted lines signify samples categorized in the negative set, while all other samples are assigned to the positive set. The notation g represents the gradient with the corresponding positive sample. The red boxes highlight single-disease cases where the gradient is weakened due to averaging with multi-disease cases.

label CL approach that simultaneously considers the challenges of long-tailed data distribution and inter-disease relationships. (4) Experimental results on two benchmark datasets validate the effectiveness of ws-MulSupCon in learning from long-tailed distribution, significantly reducing false positives compared to the loss-based method and achieving state-of-the-art (SoTA) performance.

2 Methodology

2.1 Preliminaries

For a batch of N samples and their corresponding labels, denoted as $\mathcal{B} = \{(\mathbf{x}^i, \mathbf{y}^i) \mid i = 1, 2, \ldots, N\}$, each sample $\mathbf{x}^i$ is associated with a multi-label set $\mathbf{y}^i = \{y^i_j \mid j = 1, \ldots, L\}$, where $y^i_j \in \{0, 1\}$ denotes the j-th label among a total of L classes. Following MulSupCon [19], which builds upon MoCo [6], $\mathbf{z}_q$ and $\mathbf{z}_k$ denote the query and key representations generated from a gradient descent-updated encoder and a momentum-updated encoder, respectively. Additionally, a queue $\mathcal{Q}$ is maintained to store $\mathbf{z}_k$ from previous batches as proposed in MoCo, facilitating efficient contrastive learning.

2.2 Multi-label contrastive learning

Gradient analysis. The loss function of MulSupCon is defined as:

$$\mathcal{L}_{\text{MulSupCon}} = \frac{1}{\sum_i |\mathbf{y}^i|} \sum_i \sum_j \frac{-1}{|\mathcal{P}^i_j|} \sum_{p \in \mathcal{P}^i_j} \log \frac{e^{s^i_p/\tau}}{\sum_a e^{s^i_a/\tau}}, \tag{1}$$

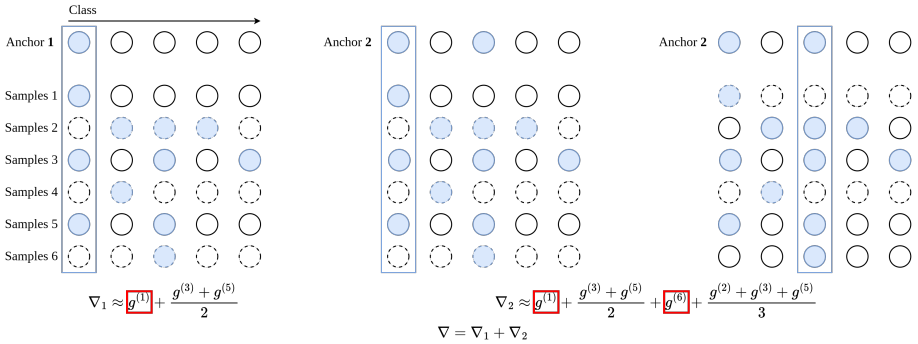

Fig. 2. Gradient illustration for single- and multi-disease cases stratification. Single-disease cases, representing simpler disease features, can contribute larger gradients, thereby enabling the model to learn more precise and distinct disease characteristics.

where $\mathcal{P}_j^i = \{a \mid y_j^i = y_j^a = 1,\ a = 1, ..., |\mathcal{A}|\}$ is a separate positive set of each disease j. Here, $\mathcal{A} = f_{\theta_m}(\mathcal{B}) \cup \mathcal{Q}$ represents all the features involved in contrastive loss computation, where f_{θ_m} denotes the momentum-updated encoder and $|\mathcal{A}| = |\mathcal{B}| + |\mathcal{Q}|$. The notation $s_n^i = \mathbf{z}_q^i \cdot \mathbf{z}_k^n$ denotes the inner product between the query and key representations. As illustrated in Fig. 1 with an example of two anchors, we can observe that the gradient in the contrast between anchors and single-disease cases is diminished by averaging with multi-disease cases. Since multi-disease cases exhibit more complex feature representations, we argue that single-disease cases should contribute larger gradients to better capture distinct and unambiguous disease characteristics. However, MulSupCon computes the loss for both cases jointly, diluting the gradients in single-disease samples.

Sample Stratification. To solve the aforementioned limitation, we stratify all the positive samples into single- and multi-disease cases to do CL while decoupling the loss into $\mathcal{L}_\mathrm{s}$ and $\mathcal{L}_\mathrm{m}$, which can be defined as:

$$\mathcal{L}_\mathrm{s} = \frac{1}{\sum_{i_s} |\mathbf{y}^{i_s}|} \Big(\sum_{i_s} \frac{-1}{|\mathcal{P}_s^{i_s}|} \sum_{p \in \mathcal{P}_s^{i_s}} \mathcal{S}_s + \sum_{i_s} \frac{-1}{|\mathcal{P}_m^{i_s}|} \sum_{p \in \mathcal{P}_m^{i_s}} \mathcal{S}_s \Big), \tag{2}$$

$$\mathcal{L}_\mathrm{m} = \frac{1}{\sum_{i_m} |\mathbf{y}^{i_m}|} \Big(\sum_{i_m} \sum_{j} \frac{-1}{|\mathcal{P}_s^{i_m}|} \sum_{p \in \mathcal{P}_s^{i_m}} \mathcal{S}_m + \sum_{i_m} \sum_{j} \frac{-1}{|\mathcal{P}_m^{i_m}|} \sum_{p \in \mathcal{P}_m^{i_m}} \mathcal{S}_m \Big), \tag{3}$$

where $\mathcal{S}_s = \log \frac{e^{s_p^{i_s}/\tau}}{\sum_a e^{s_a^{i_s}/\tau}}$ and $\mathcal{S}_m = \log \frac{e^{s_p^{i_m}/\tau}}{\sum_a e^{s_a^{i_m}/\tau}}$. Here, i_s, i_m represent the indices of single- and multi-disease anchors, respectively, while $\mathcal{P}_s$ and $\mathcal{P}_m$ correspond to the positive sample set with single- and multi-disease. Note that we omit the notation j in $\mathcal{P}_s$ and $\mathcal{P}_m$ for simplicity. The gradient of our proposed stratification method is illustrated in Fig. 2. However, a potential problem in multi-disease-related contrasts is that the gradients of $g^{(3)}$ and $g^{(5)}$ in ∇_1 are identical, despite sample 3 exhibiting a more intricate representation due to

its higher disease complexity. To further differentiate disease relationships, we weight the loss functions by the intersection over union, $\mathcal{W}_p^i = \frac{|\mathbf{y}^i \cap \mathbf{y}^p|}{|\mathbf{y}^i \cup \mathbf{y}^p|}$ between the anchor and sample. Therefore, the gradients in Fig. 2 will finally become:

$$\nabla_1 \approx g^{(1)} + \frac{\frac{1}{3}g^{(3)} + \frac{1}{2}g^{(5)}}{2}\,. \tag{4}$$

$$\nabla_2 \approx \frac{1}{2}g^{(1)} + \frac{\frac{2}{3}g^{(3)} + g^{(5)}}{2} + \frac{1}{2}g^{(6)} + \frac{\frac{1}{4}g^{(2)} + \frac{2}{3}g^{(3)} + g^{(5)}}{3}\,. \tag{5}$$

This design prioritizes the contrasts between the anchor and the sample with similar disease characteristics by assigning them larger gradients, thereby enhancing the ability of model to capture complex multi-disease representations.

Weighted Stratification. To conquer the challenge posed by long-tailed distribution, we weight $\mathcal{L}_{\text{single}}$ with a parameter $\bar{D}$, which can be defined as follows:

$$\mathcal{L}_{\text{ws}} = \frac{1}{\sum_{i_s} |\mathbf{y}^{i_s}|}\Big(\sum_{i_s} \frac{-1}{|\mathcal{P}_s^{i_s}|} \sum_{p \in \mathcal{P}_s^{i_s}} \bar{D}_j \mathcal{W}_p^{i_s} \mathcal{S}_s + \sum_{i_s} \frac{-1}{|\mathcal{P}_m^{i_s}|} \sum_{p \in \mathcal{P}_m^{i_s}} \bar{D}_j \mathcal{W}_p^{i_s} \mathcal{S}_s\Big)\,, \tag{6}$$

where $\bar{D}_j = \frac{D - d_j}{D}$, d_j denotes the number of cases with a certain disease j in the training data, and $D = \sum_j d_j$ is the sum of all the disease cases. By leveraging the disease distribution from the training dataset, this strategy effectively accounts for rare classes by assigning them higher weights, ensuring that their representations are adequately emphasized during training. On the other hand, to further consider the relationships between diseases, we weight $\mathcal{L}_{\text{multi}}$ with the reciprocal of the mean comorbidity score of each disease, which is defined as:

$$\begin{aligned}\mathcal{L}_{\text{wm}} = \frac{1}{\sum_{i_m} |\mathbf{y}^{i_m}|}\Big(&\sum_{i_m}\sum_{j} \frac{-1}{|\mathcal{P}_s^{i_m}|} \sum_{p \in \mathcal{P}_s^{i_m}} \bar{C}_j \mathcal{W}_p^{i_m} \mathcal{S}_m + \\ &\sum_{i_m}\sum_{j} \frac{-1}{|\mathcal{P}_m^{i_m}|} \sum_{p \in \mathcal{P}_m^{i_m}} \bar{C}_j \mathcal{W}_p^{i_m} \mathcal{S}_m\Big)\end{aligned}\,, \tag{7}$$

where $\bar{C}_j = \frac{1}{C_j / M_j}$, C_j denotes the comorbidity score, indicating the total count of co-occurring diseases with j-th disease across all multi-disease cases, M_j denotes the number of multi-disease cases containing the j-th disease. A higher mean comorbidity score for a disease class signifies that its features are more frequently encountered by the model due to its frequent co-occurrence with other diseases in multi-disease cases. Additionally, given that the classification performance of a disease is strongly related to its co-occurrence with other diseases [8], we incorporate the reciprocal of the mean comorbidity score to amplify the disease representations that are less frequently observed by the model in multi-disease CL. As the loss function establishes multiple positive sample sets for each class of the anchor, it can be weighted individually using $\bar{C}_j$ to better reflect the comorbidity situation of each disease class.

Healthy case contrastive learning. In $\mathcal{L}_{\text{single}}$ and $\mathcal{L}_{\text{multi}}$, we exclude healthy samples from the positive pairs to emphasize disease samples and prevent healthy samples from dominating the gradient during positive pair construction. Nevertheless, healthy samples are included in the negative pairs to ensure effective discrimination between disease and healthy samples. To further enhance the separation between disease and healthy cases in the embedding space, we introduce dedicated CL for healthy cases, enabling the model to distinguish between healthy and diseased cases more effectively. The loss function can be defined as:

$$\mathcal{L}_{\text{h}} = \frac{1}{|H|} \sum_{i}^{|H|} \frac{-1}{|\mathcal{P}_{\text{h}}|} \sum_{p \in \mathcal{P}_{\text{h}}} \log \frac{e^{s_p^i/\tau}}{\sum_a e^{s_a^i/\tau}}\,, \tag{8}$$

where H denotes the set of healthy anchor samples and the positive set $\mathcal{P}_{\text{h}} = \left\{a \mid |\mathbf{y}^a| = 0,\ a = 1, ..., |\mathcal{A}|\right\}$ contains the healthy samples only. Finally, the overall loss function can be defined as:

$$\mathcal{L}_{\text{all}} = (1-\lambda)(\mathcal{L}_{\text{ws}} + \mathcal{L}_{\text{wm}}) + \lambda \mathcal{L}_{\text{h}}\,, \tag{9}$$

where λ serves as a hyperparameter to balance the contributions of disease sample contrast and healthy sample contrast in the overall training process.

3 Experiments

Datasets and preprocessing. In this study, we utilize two large-scale CXR datasets, NIH ChestXRay14 (CXR-14) [16] and MIMIC-CXR (MIMIC) [10], to evaluate the performance of the proposed ws-MulSupCon method. The CXR-14 dataset comprises 112,120 frontal CXR images spanning 14 disease classes, while MIMIC includes 377,110 CXR images covering 13 disease classes. For MIMIC, we focus on the commonly used posterior-anterior (PA) direction images, retaining 96,155 CXR images for analysis. All images are resized into 224 $\times$ 224. Data augmentation for CL includes random horizontal flipping and random rotation in the range of -20 to 20 degrees to enhance robustness. Each dataset is randomly split at the patient level to ensure no sample overlap across the training, validation, and test sets, which comprise 70%, 10%, and 20% of the data, respectively.

Implementation details. In the pretraining phase, we train the model for 100 epochs using the Adam optimizer with an initial learning rate of 5×10^{-4}, managed by a cosine learning rate scheduler and a batch size of 64. Aligning to the same backbone model in MulSupCon, we utilize ResNet-50 [7] as our encoder. For the downstream task, the model is fine-tuned for 100 epochs using the Adam optimizer with the same initial learning rate of 5×10^{-4} and a batch size of 32. The learning rate is reduced by a factor of 0.1 when the validation loss plateaus. The model is trained from scratch using BCE loss. The optimal value of λ is determined via a comprehensive search over the interval [0,1], with stable performance within ± 0.5. Based on empirical results, λ is set to 0.7 for

Table 1. Results of performance comparison. The best scores are highlighted in red, and the second-best in blue.

Dataset	Type	Method	mAUC	mi-F1	ma-F1	mi-R	ma-R	mi-P	ma-P
CXR-14	Loss	TWML (CVPR'23) [12]	**80.62**	**32.94**	**25.84**	**64.84**	**45.56**	22.07	18.69
	CL	MulSupCon (AAAI'24) [19]	80.41	14.81	8.55	8.65	5.60	51.57	46.36
		Sim-Diss (arXiv'24) [9]	77.70	8.96	3.89	4.88	2.50	**54.51**	45.21
		SoftCon (GEOSCI'24) [17]	80.11	15.03	8.47	8.78	5.51	**52.02**	**48.27**
		ws-MulSupCon (Ours)	**81.99**	**20.44**	**12.95**	**12.82**	**8.73**	50.45	**47.79**
MIMIC	Loss	TWML (CVPR'23) [12]	**81.72**	**44.12**	**30.95**	**59.53**	**41.44**	35.05	27.23
	CL	MulSupCon (AAAI'24) [19]	81.04	31.18	16.89	20.67	12.91	63.45	**56.32**
		Sim-Diss (arXiv'24) [9]	80.64	29.56	14.85	19.20	11.51	**64.23**	55.39
		SoftCon (GEOSCI'24) [17]	81.70	32.49	18.14	21.75	13.69	**64.20**	52.35
		ws-MulSupCon (Ours)	**82.33**	**34.75**	**20.48**	**23.92**	**15.54**	63.48	**58.08**

Table 2. AUC comparison of all the classes on CXR-14.

Method	Inf.	Eff.	Ate.	Nod.	Mass	Pneumot.	Con.	Ple.	Car.	Emp.	Ede.	Fib.	Pne.	Her.
TWML	**70.61**	**87.49**	**79.14**	**70.81**	**82.20**	83.64	79.27	**76.82**	90.25	**85.76**	89.07	**75.57**	72.95	**85.09**
MulSupCon	70.25	87.17	78.54	71.62	81.68	**84.16**	**80.18**	75.94	**90.35**	84.64	88.71	74.12	**73.52**	84.88
Sim-Diss	69.59	85.74	77.55	67.39	75.29	81.87	79.15	72.89	87.48	79.66	88.30	70.64	72.31	79.97
SoftCon	70.22	87.07	78.70	70.72	81.05	83.57	79.43	75.98	89.88	84.05	**89.15**	74.99	73.25	83.42
Ours	**71.33**	**87.79**	**79.46**	**75.10**	**83.68**	**85.99**	**80.17**	**76.86**	**90.68**	**88.92**	**89.58**	**77.31**	**74.93**	**86.02**

CXR-14 and 0.75 for MIMIC. All implementations are conducted in PyTorch and executed on an NVIDIA 4090 GPU.

Performance comparison. We conduct a comprehensive comparison of ws-MulSupCon against several SoTA multi-label classification methods, as detailed in Table 1. The evaluated methods include the loss-based Two-Way Multi-Label Loss (TWML) [12] and CL-based studies such as MulSupCon [19], Sim-Diss [9], and SoftCon [17]. Performance is measured using seven key metrics: mAUC, micro/macro F1, micro/macro Recall, and micro/macro Precision. To ensure a fair comparison, all methods are trained under identical settings and backbone. Note that the loss-based method does not need to be pretrained. Source codes are obtained from official repositories where available, while Sim-Diss [9] and SoftCon [17] are implemented by us due to the absence of published codes.

As shown in Table 1, ws-MulSupCon achieves SoTA performance, attaining the highest mAUC of 81.99% on CXR-14 and 82.33% on MIMIC. Compared to the best results in other CL-based approaches, it improves mAUC by 1.58%, micro-F1 by 5.41%, macro-F1 by 4.4%, micro-recall by 4.04%, and macro-recall by 3.13% on CXR-14; and mAUC by 0.63%, micro-F1 by 2.26%, macro-F1 by 2.34%, micro-recall by 2.17%, and macro-recall by 1.85% on MIMIC, underscoring its superiority in disease detection. In contrast to the CL-based methods fine-tuned with BCE, the TWML attains high recall rates by heavily weighting positive disease classes in the loss design. However, this overemphasis results in elevated false positives, leading to decreased precision and a lower mAUC compared to the proposed method. Notably, ws-MulSupCon effectively balances recall and precision, ensuring robust diagnostic performance across both datasets.

Table 3. AUC comparison of all the classes on MIMIC.

Method	Opa.	Ple.	Ate.	Pne.	Car.	Ede.	Sup.	Les.	Enl.	Con.	Pneumo.	Fra.	other
TWML	76.52	94.02	**83.44**	76.93	**86.29**	**90.07**	**84.95**	71.56	77.46	82.85	**89.91**	**69.55**	**78.84**
MulSupCon	76.23	93.88	83.31	76.26	86.00	89.75	84.97	70.45	77.23	82.52	88.56	67.72	76.72
Sim-Diss	76.29	93.76	82.71	75.91	85.53	89.78	84.01	70.24	76.73	81.80	88.52	66.57	76.50
SoftCon	**76.97**	**94.06**	83.30	**77.11**	86.13	89.95	84.89	**71.72**	**77.85**	**83.02**	**89.91**	69.01	78.12
Ours	**78.04**	**94.41**	**84.27**	**78.42**	**86.31**	**90.45**	**84.99**	**72.91**	**77.91**	**83.76**	**90.45**	**69.62**	**78.78**

Table 4. Ablation studies for loss function design.

Dataset	Method	mAUC	mi-F1	ma-F1	mi-R	ma-R	mi-P	ma-P
CXR-14	Baseline	80.41	14.81	8.55	8.65	5.60	51.57	46.36
	Stratified	81.32	17.70	11.32	10.68	7.50	**51.76**	42.54
	Stratified+weighted	80.93	18.89	12.28	11.61	8.25	50.73	43.45
	ws-MulSupCon	**81.99**	**20.44**	**12.95**	**12.82**	**8.73**	50.45	**47.79**
MIMIC	Baseline	81.04	31.18	16.89	20.67	12.91	63.45	56.32
	Stratified	81.69	32.63	18.36	21.83	13.70	**64.61**	**58.82**
	Stratified+weighted	81.55	34.48	19.52	23.96	15.01	61.47	53.49
	ws-MulSupCon	**82.33**	**34.75**	**20.48**	**23.92**	**15.54**	63.48	58.08

Moreover, we assess AUC for all classes across both datasets to validate the effectiveness of ws-MulSupCon in learning long-tailed distribution, as detailed in Tables 2 and 3. Note that the diseases in the tables are sorted by sample count in descending order. Apart from a slight 0.01% decrease in AUC for the "consolidation" class on CXR-14 and a 0.06% reduction for the "pleural other" class on MIMIC, our proposed method consistently outperforms competing methods, achieving the highest AUC across nearly all disease classes. This underscores the strength of ws-MulSupCon in handling long-tailed data distribution.

Ablation study. To evaluate the impact of each design in ws-MulSupCon, we conduct a comprehensive ablation study on both datasets, as summarized in Table 4. Utilizing MulSupCon as the baseline, stratifying the samples into single- and multi-disease cases amplifies the gradient contrast between anchors and single-disease cases. This results in notable performance improvements, with increases in mAUC, micro-F1, macro-F1, micro-recall, and macro-recall by 0.91%, 2.89%, 2.77%, 2.03%, and 1.9% on CXR-14, and by 0.65%, 1.45%, 1.47%, 1.16%, and 0.79% on MIMIC, respectively. These results highlight the importance of explicitly modeling single-disease characteristics. Further enhancements with the weighting factors $\bar{D}$ and $\bar{C}$ yield additional gains of micro/macro-F1 and micro/macro-recall despite a slight decline in mAUC on both datasets. This suggests that incorporating both class frequency and comorbidity scores improves recall but introduces a higher false positive rate. Finally, the integration of healthy case CL further enhances the discrimination between disease and healthy feature representations, resulting in SoTA performance across all evaluation metrics. Notably, the consistent improvement in macro-level metrics demonstrates the effectiveness of ws-MulSupCon in tackling the challenge of rare disease classification, addressing a longstanding issue in medical image analysis.

4 Conclusion

In this paper, we propose ws-MulSupCon, a novel CL framework for multi-label medical image classification. Our method strategically stratifies the samples into single- and multi-disease cases, assigning higher gradient contributions to single-disease-related features to enhance representation learning. To address challenges associated with long-tailed distribution and disease co-occurrence, we design two weighted parameters to mitigate biases toward frequent classes. Additionally, we incorporate a healthy case CL to refine the ability of the model to distinguish between healthy and diseased cases. Comprehensive experiments validate the effectiveness of ws-MulSupCon in achieving SoTA performance.

Acknowledgments. This work was supported in part by the National Science and Technology Council, Taiwan, under Grants NSTC 113-2221-E-A49-137 and NSTC 113-2634-F-A49-003. We appreciate the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

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

References

1. Bai, J., Kong, S., Gomes, C.P.: Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In: international conference on machine learning. pp. 1383–1398. PMLR (2022)
2. Chen, B., Li, J., Lu, G., Yu, H., Zhang, D.: Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. IEEE journal of biomedical and health informatics **24**(8), 2292–2302 (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
4. Cole, E., Mac Aodha, O., Lorieul, T., Perona, P., Morris, D., Jojic, N.: Multi-label learning from single positive labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 933–942 (2021)
5. Gupta, R., Roy, A., Christensen, C., Kim, S., Gerard, S., Cincebeaux, M., Divakaran, A., Grindal, T., Shah, M.: Class prototypes based contrastive learning for classifying multi-label and fine-grained educational videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19923–19933 (2023)
6. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
8. Holste, G., Jiang, Z., Jaiswal, A., Hanna, M., Minkowitz, S., Legasto, A.C., Escalon, J.G., Steinberger, S., Bittman, M., Shen, T.C., et al.: How does pruning impact long-tailed multi-label medical image classifiers? In: International Conference

on Medical Image Computing and Computer-Assisted Intervention. pp. 663–673. Springer (2023)
9. Huang, G., Long, Y., Luo, C., Liu, S.: Similarity-dissimilarity loss with supervised contrastive learning for multi-label classification. arXiv preprint arXiv:2410.13439 (2024)
10. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
11. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
12. Kobayashi, T.: Two-way multi-label loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7476–7485 (2023)
13. Lanchantin, J., Wang, T., Ordonez, V., Qi, Y.: General multi-label image classification with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16478–16488 (2021)
14. Małkiński, M., Mańdziuk, J.: Multi-label contrastive learning for abstract visual reasoning. IEEE Transactions on Neural Networks and Learning Systems **35**(2), 1941–1953 (2022)
15. Pu, T., Sun, M., Wu, H., Chen, T., Tian, L., Lin, L.: Semantic representation and dependency learning for multi-label image recognition. Neurocomputing **526**, 121–130 (2023)
16. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106 (2017)
17. Wang, Y., Albrecht, C.M., Zhu, X.X.: Multi-label guided soft contrastive learning for efficient earth observation pretraining. IEEE Transactions on Geoscience and Remote Sensing (2024)
18. Zhang, K., Liang, W., Cao, P., Mao, Z., Yang, J., Zaiane, O.R.: Corlabelnet: a comprehensive framework for multi-label chest x-ray image classification with correlation guided discriminant feature learning and oversampling. Medical & Biological Engineering & Computing pp. 1–14 (2024)
19. Zhang, P., Wu, M.: Multi-label supervised contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 16786–16793 (2024)
20. Zhang, S., Xu, R., Xiong, C., Ramaiah, C.: Use all the labels: A hierarchical multi-label contrastive learning framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16660–16669 (2022)
21. 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)
22. Zhang, Y., Luo, L., Dou, Q., Heng, P.A.: Triplet attention and dual-pool contrastive learning for clinic-driven multi-label medical image classification. Medical image analysis **86**, 102772 (2023)
23. Zheng, M., Wang, F., You, S., Qian, C., Zhang, C., Wang, X., Xu, C.: Weakly supervised contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10042–10051 (2021)
24. Zhu, K., Fu, M., Wu, J.: Multi-label self-supervised learning with scene images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6694–6703 (2023)