ConPro: Learning Severity Representation for Medical Images using Contrastive Learning and Preference Optimization

Hong Nguyen[1] Hoang Nguyen[2] Melinda Chang[1] Hieu Pham[2]
Shrikanth Narayanan[1] Michael Pazzani[1]

[1]University of Southern California, Los Angeles, United States
[2]Vinuni-Illinois Smart Health Center, Hanoi, Vietnam

{hongn, shri, mpazzani}@usc.edu, mchang@chls.usc.edu, hieu.ph@vinuni.edu.vn

Abstract

*Understanding the severity of conditions shown in images in medical diagnosis is crucial, serving as a key guide for clinical assessment, treatment, as well as evaluating longitudinal progression. This paper proposes **ConPrO**: a novel representation learning method for severity assessment in medical images using **Co**ntrastive learning-integrated **Pr**eference **O**ptimization. Different from conventional contrastive learning methods that maximize the distance between classes, ConPrO injects into the latent vector the distance preference knowledge between various severity classes and the normal class. We systematically examine the key components of our framework to illuminate how contrastive prediction tasks acquire valuable representations. We show that our representation learning framework offers valuable severity ordering in the feature space while outperforming previous state-of-the-art methods on classification tasks. We achieve a 6% and 20% relative improvement compared to a supervised and a self-supervised baseline, respectively. In addition, we derived discussions on severity indicators and related applications of preference comparison in the medical domain.*

1. Introduction

Recent advances in supervised [12, 27, 29, 35] and self-supervised contrastive learning [5, 9, 38, 41] offer a strong foundation for image understanding and interpretation, including in medical applications. Crucially, latent vectors are acquired from data to capture increasing amounts of contextual information within an image and across contextual classes. Self-supervised contrastive learning attempts to exploit domain knowledge by bringing 'positive' samples closer together in the embedding space while pushing 'negative' samples apart. A positive pair often consists of augmented versions of the same sample and negative pairs are created using the anchor and randomly selected samples from the data batch. On the other hand, supervised contrastive learning (SupCon) studies cross-class relations by grouping embeddings from the same class to the same cluster and pushing different clusters far from each other.

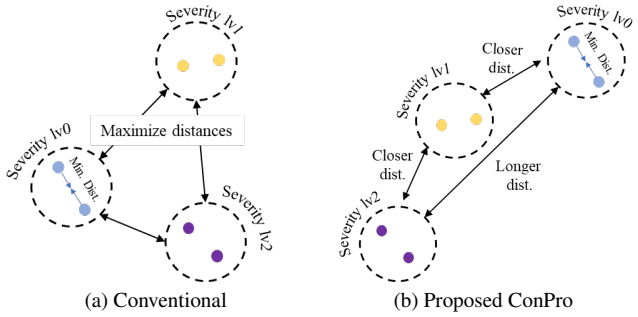
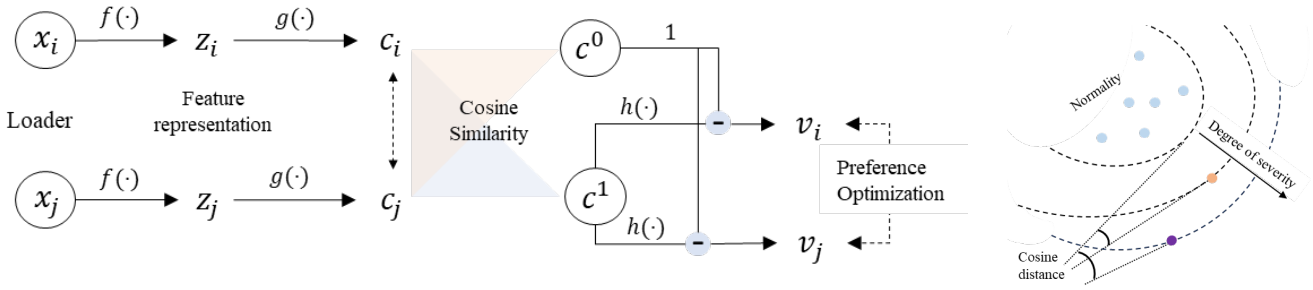


Figure 1. Conventional (a) and target (b) representation for severity modeling in latent space. A darker color represents a higher severity level, and '0' represents normality. Our proposed method targets to embed distance relation to severity classes in representation space

Conventionally, supervised contrastive learning treats all classes equally and maximizes inter-class distance, as shown in Fig. 1a. However, this approach ignores the different level of similarity between classes, some classes should be further away than others. For instance, class "dog" should have a closer relation to "wolf" than "table." Similarly, in the medical domain (Fig. 1b), the conditions of similar severity should have a smaller distance than those of large severity differences. Furthermore, bridging the gap from the non-medical images to the medical domain presents a distinct challenge since medical images are burdened by label sharpness [14], experts' annotation biases [22, 39], weak labels [13, 36], and noise from various imaging modalities [17]. These challenges may degrade the



(a) Learning Framework

(b) Desired latent space

Figure 2. ConPro learning framework (a) includes contrastive learning and preference optimization to get the desired latent space (b)

learned knowledge within the latent representation of medical images.

Despite playing an essential role in clinical practice, severity disparity in medical images has not been investigated well in the computer vision literature. A common approach is to explore image-derived clinical severity as a multi-class classification problem [4, 6, 7, 24, 32, 43], where each class corresponds to a different (quantized) severity level. The classifier then learns to distinguish attributes/traits that are probabilistically different between severity classes. However, a higher severity level may still share some traits or attributes with lower severity levels. Thus, another proposed approach was to rank images with respect to severity scores. In [30], the authors introduced the concept of severity ranking for fundus photography. Following this idea, [11, 18, 24, 42, 43] improved that work by incorporating multi-scoring, multitasking learning, and experts' agreement. However, an undesirable outcome of current severity ranking methods is that all outputs can "collapse" to a constant value. Thus, a single score for the severity of the entire image is a significant loss of information and interpretability. Our work addresses this challenge by introducing ranking loss in the representation space.

In this paper, we inject severity information into a latent space vector using contrastive learning-integrated preference optimization (ConPrO). Not only does ConPrO show a reliable representation arrangement, but it also outperforms state-of-the-art (SOTA) algorithms on classification tasks. The contributions of this paper are summarized below:

- We propose ConPrO: a novel representation learning method that incorporates class severity information within the latent space. ConPrO performs better in the F1 score, by a relative 20% compared to SimCLR (self-supervised) and 6% compared to SupCon (supervised).
- We introduce an evaluation metric, Mean Absolute Exponential Error (MAEE), for a specific problem (severity classification). MAEE penalizes incorrect prediction at higher severity classes in case of data imbalance.
- We show that increasing the number of reference vectors helps reduce MAEE and offer discussion on the potential application of preference comparison in the medical domain.

2. Related Works

Visual Contrastive Learning. The fundamental concept of contrastive learning is to push the latent vectors of different classes far apart from each other while pulling latent vectors within the same class closer. This method was introduced in representation learning as a self-supervised [5, 9, 38, 41] or supervised way [12, 27, 29, 35] to inject relative information to embeddings in the latent space. Shekoofeh et al. [1] show that contrastive learning improves the robustness and data efficiency of medical imaging tasks. Furthermore, the authors suggest that visual representation learning is a key component for building large (foundation) vision models. Moreover, recent works [8, 9, 33] have made efforts to relate the success of contrastive learning from the perspectives of mutual information, choices of feature encoder, and loss function.

Preference comparison. Originating from learning to rank problems [3, 37] in recommendation systems, several works [18, 30] have used preference comparison for ranking disease severity. Yu [40] proposed Relative Distance Ranking Loss, which measures the similarity between image patches and their reference image. Recently, preference comparison was re-introduced in Large Language Models (LLMs) as a way to optimize preference of generative pairs of answers given an input. RLHF [20] learns the reward function from pairwise comparisons of output text and optimizes it via reinforcement learning. More recently, Direct Preference Optimization (DPO) [26] simplifies RLHF by optimizing a language model directly to align with human preferences without relying on explicit reward modeling or reinforcement learning. DPO updates parameters by maximize preferences likelihood between pair of samples, which can be a potential approach toward severity ranking in latent spaces.

We note that none of the previous works investigate medical, image-driven preference with respect to severity, nor

relative distance between classes. As such, severity-level dependency (Figure 1b) brings up a new problem in representation learning. We also note that although the individual components of our framework have been presented in prior research, the innovation of our framework is its combination to solve severity ranking problems with particular choices of reward/loss function.

3. Method

The framework, as shown in Fig. 2a, contains two main consecutive phases including binary contrastive learning ('Con' step) and preference optimization ('PrO' step). In the 'Con' step, we maximize the latent distance between the normal and the abnormal class. In the 'Pro' phase, we re-arrange the relative distance of severity levels within abnormal classes with respect to reference vectors in the normal class. The detailed implementation and loss function of both steps are presented in the following.

3.1. Contrastive Learning

We group severity classes into a single abnormal class and perform binary contrastive learning between normal[1] and abnormal samples. The motivation for this phase is to group the positive samples in a cluster that is well-separated from the abnormal cluster in the latent space. Subsequently, the normal cluster is used as an anchor for preference comparison. The framework contains a feature extractor $f(\cdot)$ which is a convolutional neural network that encodes images to latent vectors. The contrastive head $g(\cdot)$ maps those vectors to the contrastive space.

Supervised Contrastive Objective. For simplicity, we use margin contrastive loss (although there are multiple attractive options such as NT-Xent, XT-Logistic [5])

$$\mathcal{L}_{Con}(c_i, c_j, y_{ij}) = \mathbb{E}[y_{ij}\, d_{\cos}(c_i, c_j) + \dots \\ (1 - y_{ij})\, \mathbf{max}(0, m - d_{\cos}(c_i, c_j))] \tag{1}$$

where m is maximum margin, ranging from 0 to 2 (here we choose $m = 2$) and $d_{\cos}$ denotes the cosine distance

$$d_{\cos}(c_i, c_j) = \frac{c_i^\top c_j}{\|c_i\|\,\|c_j\|} \tag{2}$$

between a pair of vectors (c_i, c_j).

3.2. Preference Optimization over Latent Space

Inspired by current advances in learning to rank algorithms as well as preference optimization algorithms in Natural Language Processing (NLP) such as RLHF [20] and DPO [26], our objective is to present a simple approach for severity comparison over the representation space.

[1]The term "normal" in this paper refers to non-abnormal cases; e.g., images categorized as "normal" with respect to a specific pathology may not necessarily indicate healthy condition

Algorithm 1 ConPro Pseudocode

Require: Pre-defined f, g, h, contrastive loader $\mathcal{C}$, preference loader $\mathcal{P}$
for (x_i, x_j, y_{ij}) in $\mathcal{C}$ **do** ▷ *'Con' step*
  $z_i, z_j \leftarrow f(x_i), f(x_j)$
  $c_i, c_j \leftarrow g(z_i), g(z_j)$
  Calculate $\mathcal{L}_{Con}(c_i, c_j, y_{ij})$ ▷ Eq. (1)
  Update network f and g to minimize $\mathcal{L}_{Con}$
end for
for $(c_i, c_j, \pi_{\mathbf{0}}, y_{ij})$ in $\mathcal{P}$ **do** ▷ *'PrO' step*
  $\nu_i, \nu_j \leftarrow h(c_i), h(c_j)$
  Draw π_0 from $\mathcal{P}$
  Calculate $\mathcal{L}_{PrO}(\nu_i, \nu_j, \pi_0, y_{ij})$ ▷ Eq. (4)
  Update network f, g and h to minimize $\mathcal{L}_{PrO}$
end for
return encoder network f and **discard** g, h

Preference Comparison Objective. Both RLHF and DPO use Bradley-Terry preference model [2] to construct the loss function. Given some prior knowledge π_0, the Bradley-Terry model calculates the preference likelihood over a pair of samples with respect to labelers' severity measurement, denoted as $\nu_i > \nu_j|\pi_0$, where ν_i and ν_j are the preferred and dispreferred completion amongst (ν_i, ν_i) respectively. The preferences are assumed to be generated by some predefined reward model r^*. In this work, we choose the reward function $r^* = d_{\cos}$ as the cosine distance from severity to normality. Intuitively, we try to pull the less severe latent vectors closer to the "normality" anchor while pushing more severe cases far apart. We define the "normality" anchor as a vector or set of vectors belonging to the normal class. Under the Bradley-Terry model, we derive a simplified probability measure for pairwise severity comparison:

$$p^*(\nu_i > \nu_j|\pi_{\mathbf{0}}) = \sigma(r^*(\nu_i, \pi_0) - r^*(\nu_j, \pi_0)) \\ = \frac{1}{1 + \exp[d_{\cos}(\nu_i, \pi_0) - d_{\cos}(\nu_j, \pi_0)]} \tag{3}$$

where $x_{\mathbf{0}}$ represented normality and $\mathrm{d}_{\cos}$ is the cosine distance. We can then formulate the problem in hand as binary classification and use the negative log-likelihood loss to re-parameterize the feature space:

$$\mathcal{L}_{PrO}(\nu_i, \nu_j, \pi_0, y_{ij}|r^* = d_{\cos}) \\ = \mathbb{E}[\log(\sigma(r^*(\nu_i, \pi_0) - r^*(\nu_j, \pi_0)))] \tag{4}$$

4. Experiments

Datasets and Preprocessing. We study two real-world datasets that contain severity labels: Papilledema and VinDr-Mammogram. The first two datasets contain discrete class labels while VinDr measures symptom severity on a

Table 1. Multiclass classification results. SupCon-n denote n-classes supervised contrastive learning and SupCon-2 is the first stage of ConPro. ImageNet denotes a pre-trained model on ImageNet dataset. For MAEE score, the lower, the better

	Papilledema			VinDr-Mammo		
Methods	Macro F1	Recall	MAEE	Macro F1	Recall	MAEE
Multiclass classification task:						
ImageNet	38.7 ± 4.2	39.8 ± 3.1	6.4 ± 0.40	17.9 ± 2.0	20.6 ± 0.9	2.9 ± 0.06
SupCon-2	46.3 ± 5.5	47.4 ± 4.7	5.2 ± 0.30	34.9 ± 0.7	33.6 ± 1.1	2.4 ± 0.03
SupCon-n	45.5 ± 4.1	46.9 ± 3.6	5.0 ± 0.26	20.8 ± 1.5	23.1 ± 1.2	2.9 ± 0.03
SimCLR	40.3 ± 3.9	43.8 ± 3.0	4.8 ± 0.26	25.1 ± 1.4	25.7 ± 1.2	2.7 ± 0.05
ConPrO (ours)	$\mathbf{48.5 \pm 3.8}$	$\mathbf{49.4 \pm 4.1}$	$\mathbf{4.8 \pm 0.25}$	$\mathbf{35.6 \pm 0.8}$	$\mathbf{34.7 \pm 1.0}$	$\mathbf{2.4 \pm 0.03}$

continuous scale. Details of the statistics and preprocessing of each dataset are described below.

- *Papilledema* is a controlled dataset comprising 331 pediatric fundus images obtained clinically from 105 subjects from 2011 to 2021. The dataset contains a five-level severity rating for Papilledema. De-identified clinical datasets were uploaded to the HIPAA-compliant Research Electronic Data Capture (REDCap) database.
- *VinDr-Mammo* [19] is a public Vietnamese collection of full-field digital mammography comprising 5,000 four-view examinations with breast-level evaluations and annotated findings between 2018 and 2020. These examinations underwent independent double readings, with any disagreements resolved through third-party radiologist arbitration. The authors state that there are no ethical concerns. Approval was granted by the Institutional Review Boards of Hanoi Medical University Hospital and Hospital 108 to release de-identified data. The VinDr-Mammo dataset assesses the Breast Imaging-Reporting and Data System (BI-RADS) for breast level. It has 7 categories from 0 to 6 and be used as a risk evaluation and quality assurance tool. The datasets only contain the mammograms with BI-RADS from 1 to 5. In this work, we used this assessment to estimate the severity of the targeted breast.

Evaluation. We evaluate the final image representation using standard protocols [5, 38]. This involves training a linear classifier on the frozen-weight feature encoder, using the validation F1 score to choose the best model. That model is then used to compute the final scores on the test set. We evaluate the representation vector via classification tasks and use several evaluation metrics including Top-1 F1 scores (macro F1), Recall and Mean Absolute Exponential Error (MAEE). We proposed to used MAEE as a variant of Mean Absolute Error (MAE) for severity classification problems. MAEE is computed as

$$\text{MAEE} = \frac{1}{n}\sum_{i}^{n} \mathbf{e}^{|y_i - \hat{y}_i|} \tag{5}$$

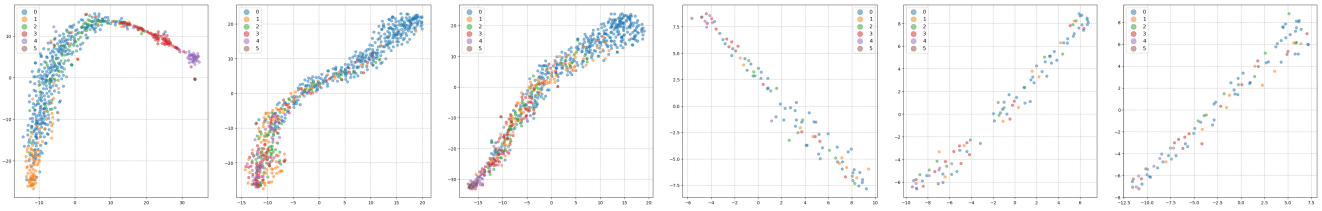
Both MAE and MAEE measure the error between predictions and true levels of severity. However, while MAE evaluates regression problems with a linear penalty, MAEE assigns exponential penalties for incorrect severity level predictions.

Experimental Setup. We trained all datasets on a GTX 3090 with a batch size of 16. Our choice of feature encoder is Resnet-50. If not stated otherwise, our explorations utilize the following settings.

- *Data splitting:* Since each subject may have multiple visits or images have multiple views, we split train/val/test by $70/15/15$ for Papilledema dataset and $72/8/20$ for VinDr-Mammo by subject IDs to avoid data leakage. We chose pairs for preference optimization by randomly selecting 10^5 pairs with replacements for training and 10^4 for evaluation.
- *Optimizer:* We use stochastic gradient descent (SGD) with momentum 0.9. We update the ResNet-50 encoder with a learning rate of 10^{-3} and the projection head with a learning rate of 0.01 for both Supervised Contrastive Learning (Con) and Preference Comparison (PrO).
- *Projection Head:* Resnet-50 outputs 2048-d vectors. The projection head $g(\cdot)$ is a fully-connected (FC) layer with a 256-d output. In the 'PrO' step, the projection head $h(\cdot)$ of preference comparison is an FC layer that keeps the preference vectors on the same dimension with normality.
- *Prediction Head:* We use a 2-layer MLP with a hidden dimension of 256. The activation function is ReLU, and we use 10% dropout. For fine-tuning, since both datasets are unbalanced, we use Cross Entropy with estimated class weights as the loss function.

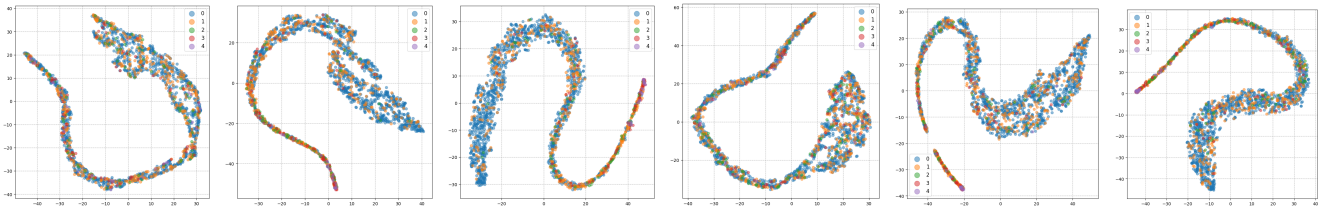
5. Results

Classification task performance. The baseline is Resnet-50 pre-trained on the ImageNet dataset. Other SOTA methods include self-supervised (SimCLR) and supervised (SupCon-n) models. For all baselines we freeze the Resnet pre-trained weights and only fine-tune the prediction head on our target task. Table 1 shows that ConPro outperforms the ImageNet baseline, SimCLR, and SupCon-n (SupCon-2 is the "Con" step of our method) on all metrics in both datasets. ConPrO reaches the highest macro F1 score of



(a) SupCon-5 Train-set (b) SupCon-2 Train-set (c) ConPrO Train-set (d) SupCon-5 Test-set (e) SupCon-2 Test-set (f) ConPrO Test-set

Figure 3. T-SNE visualization of representation vectors of (a) training and (b) test set after supervised contrastive learning (c) training (d) test set after preference optimization. The plots are samples from the Papilledema dataset. All figures use cosine distance. Label '0' denotes normality, and "1-5" denotes increasing level of severity.



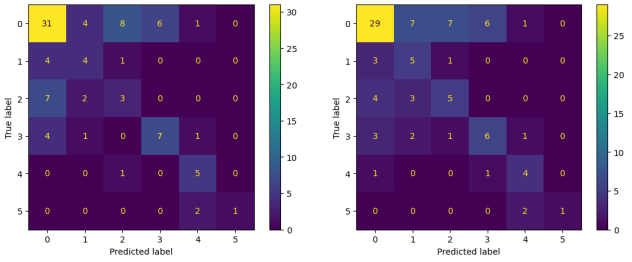
(a) SupCon-5 Train-set (b) SupCon-2 Train-set (c) ConPrO Train-set (d) SupCon-5 Test-set (e) SupCon-2 Test-set (f) ConPrO Test-set

Figure 4. T-SNE visualization of representation vectors of (a) training and (b) test set after supervised contrastive learning (c) training (d) test set after preference optimization. The plots are 2000 random samples from the VinDr-Mammo dataset. All figures use cosine distance. Label '0' denotes normality, and "1-4" denotes increasing level of severity.

48.55% and 35.6%, respectively, in the Papilledema dataset on the 6-class classification task and the VinDr-Mammo dataset on the 5-class classification task. Compared to the "Con" step, the "PrO" step injects useful information from the "Con" step to latent space, boosting the performance by 4.8% and 2% on both datasets.

ConPrO better represents severity in feature space. As shown in Fig. 3 and 4, we qualitatively evaluate the feature representations of our method on both datasets. Comparing the 'PrO' (Fig. 3b and 4b) step with the 'Con' step (Fig. 3c and 4c), we show that the preference optimization successfully re-arranges the abnormal samples with respect to severity classes. The same behavior can be seen in the test set. Moreover, SupCon-5 (Fig. 3a) shows discrimination of severity classes, but the positions of the embeddings are not relative to the severity scores, which is not ideal for severity interpretation.

MAE versus MAEE. Different from the F1 score that captures exact classification prediction, MAE and MAEE take prediction error into account. Fig. 5 represents two confusion matrices having the same F1 score. While Fig. 5b shows a better MAE score, Fig. 5a presents a greater value of MAEE. MAEE shows more sensitivity to incorrect predictions that deviate significantly from the ground truth, such as misclassifications between severity labels '3' and '4' as '0'. This study opts to utilize MAEE as it helps us identify potential serious incorrect predictions in severity classification since there is no distinct boundary between



(a) MAE = 0.88 and MAEE = 5.05 (b) MAE = 0.86 and MAEE = 5.24

Figure 5. Confusion matrices of same setting on two independent run. Two matrix have the same F1 score but different in MAE and MAEE

severity classes. For instance, ophthalmologists usually group classes as mild {1,2}, moderate {3}, and severe {4,5} since the finer annotation scale by experts also tends to have uncertainty. For mammogram examination, the BI-RADS score is also grouped into normal {1}, benign {2,3}, and malignant {4,5}.

Choices of normality indicator. To calculate the reward function in equation 4, we need to compute the distance from each vector in pairs with respect to the normal class. Thus, "normality" is abstract and represented by a cluster of vectors. We randomly chose n vectors in the normal class to get the mean of these vectors to get a single anchor representation for representing "normality" (with respect to lack of severity). Table 2 shows how F1 scores and

MAEE change when varying the number of referenced vectors. Interestingly, by increasing the number of reference vectors, the framework lowers MAEE error across classes. As a trade-off, the macro F1 score will fall. The intuition is that by updating model parameters in the "Pro" step, we also update the normality anchor. By choosing the mean of multiple normal vectors as an anchor, we attempt to empirically lower the effect of model updating (no theoretical proof is derived in this work and is left for future work).

Table 2. F1 Scores and MAEE versus the number of referenced vectors per pairs in the "PrO" step. For MAEE score, the lower, the better

	Number of reference vectors		
Metric	1	10	20
Macro F1	**46.3** $\pm$ **5.7**	45.2 $\pm$ 5.1	43.7 $\pm$ 6.3
MAEE	5.4 $\pm$ 0.52	5.4 $\pm$ 0.42	**5.2** $\pm$ **0.39**

6. Discussion

Explainable AI for Severity Ranking. Definitive reasons for the success of contrastive learning still remain incomplete in published literature. On the theoretical front, some [33] argue that the success is attributed to maximizing mutual information, while others emphasize the importance of the loss function [8, 9]. However, in terms of explainability, it remains an open question what attributes contribute to positive pairs and what distinguishes negative pairs. The majority of literature [10, 15, 16, 23] uses contrastive (counterpart) methods to explain Deep Learning models. Yet, there are limited works [31, 34] that focus on interpreting representations in contrastive learning. In this paper, we inject clinical condition severity knowledge into medical imaging representations, but we do not know what knowledge the latent space learned in deciding which sample is more severe. Thus this problem remains a topic for future investigation.

Subject matter expert-centric explanations, such as clinical judgments, may differ from what an AI model learns [21]. Performance of image-driven explainable AI (XAI) in clinical settings tends to degrade under three major pathological characteristics [28]: multiple instances (pathology has multiple possible instantiations of interest and it is ranked variably by the preference of experts), size variety (instance size may vary between subjects, heterogeneity of clinical presentation, variability in severity between patients, and longitudinal changes in clinical manifestations that may be more important for diagnostic consideration than the severity of pathology at presentation), and pathology shape complexity. Reconciling user-centric and expert-centric explanations is a yet to be fully solved research problem wherein preference optimization can be advantageous.

Severity Indicator in Medical Domain. In interpreting ophthalmologic images, physicians frequently use comparisons in making diagnoses. For instance, in evaluating for glaucoma, asymmetry in the cup-to-disc ratio between two eyes of the same patient has predictive value for glaucoma diagnosis [25]. Thus, it is easier for ophthalmologists to compare two fundus images to decide which one has more asymmetry, rather than assign a class label for each image. For that reason, preference comparison may play an important part in improving diagnosis. This also bring up multiple challenge including

- *Severity indicators on multiple perspective of diagnosis:* In glaucoma diagnosis, assessments often rely on either fundus photos or results from visual field tests. It's crucial to recognize that the severity reflect from fundus photo different from severity of visual field test although there may be a strong correlation. There is often discrepancy between metrics of structure (photos) and function (visual field tests) in ophthalmology (and other fields of medicine). Some patients with glaucoma have normal visual fields (pre-perimetric glaucoma).
- *Severity preference on multiple pathologies:* One image may endure multiple conditions (e.g. VinDr-Mammo dataset represent 15 types of pathologies and each image may contain more than one type). The challenge lies in comparing and prioritizing the severity of multiple pathologies within the same image.

7. Conclusion

This paper presents a representation learning method to inject severity information in the latent representation space. We meticulously examine the components of the suggested framework and show that (1) ConPrO not only demonstrates a dependable representation via TSNE visualization but also surpasses state-of-the-art algorithms in classification tasks by at least 6% in F1 score, (2) proposed MAEE metric penalizes serious incorrect prediction which fit well to the severity classification problem, (3) choosing good "normality" anchor can help reduce MAEE score. Finally, we discuss several problems of preference comparison and explainable AI as potential directions for future work.

References

[1] Shekoofeh Azizi, Laura Culp, Jana von Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, Seyedeh Sara Mahdavi, Ellery Wulczyn, Boris Babenko, Megan Wilson, Aaron Loh, Po-Hsuan Cameron Chen, Yuan Liu, Pinal Bavishi, Scott Mayer McKinney, Jim Winkens, Abhijit Guha Roy, Zach Beaver, Fiona Ryan, Justin D. Krogue, Mozziyar Etemadi, Umesh Telang, Yun Liu, Lily H. Peng, Greg S Corrado, Dale R. Webster, David J. Fleet, Geoffrey E. Hinton, Neil Houlsby, Alan Karthikesalingam, Mohammad Norouzi, and Vivek

Natarajan. Robust and efficient medical imaging with self-supervision. *ArXiv*, abs/2205.09723, 2022. 2

[2] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 3

[3] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, page 89–96, New York, NY, USA, 2005. Association for Computing Machinery. 2

[4] Tej Bahadur Chandra, Bikesh Singh, and Deepak Jain. Disease localization and severity assessment in chest x-ray images using multi-stage superpixels classification. *Computer Methods and Programs in Biomedicine*, page 106947, 2022. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org, 2020. 1, 2, 3, 4

[6] Matthew D. Li et al. Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *npj Digital Medicine*, 3, 2020. 2

[7] Zongyun Gu, Yan Li, Zijian Wang, Junling Kan, Jianhua Shu, Qing Wang, and Yugen Yi. Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention. *Intell. Neuroscience*, 2023, 2023. 2

[8] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron Van Den Oord. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org, 2020. 2, 6

[9] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ArXiv*, abs/1808.06670, 2018. 1, 2, 6

[10] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Contrastive explanations for model interpretability. *CoRR*, abs/2103.01378, 2021. 6

[11] Jayashree Kalpathy-Cramer, J Peter Campbell, Deniz Erdogmus, Peng Tian, Dharanish Kedarisetti, Chace Moleta, James D Reynolds, Kelly Hutcheson, Michael J Shapiro, Michael X Repka, Philip Ferrone, Kimberly Drenser, Jason Horowitz, Kemal Sonmez, Ryan Swan, Susan Ostmo, Karyn E Jonas, R V Paul Chan, Michael F Chiang, and Imaging and Informatics in Retinopathy of Prematurity Research Consortium. Plus disease in retinopathy of prematurity: Improving diagnosis by ranking disease severity and using quantitative image analysis. *Ophthalmology*, 123(11): 2345–2351, 2016. 2

[12] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, pages 18661–18673. Curran Associates, Inc., 2020. 1, 2

[13] Kiran Kokilepersaud, Mohit Prabhushankar, Ghassan AlRegib, Stephanie Trejo Corona, and Charles Wykoff. Gradient-based severity labeling for biomarker classification in oct. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3416–3420, 2022. 1

[14] Nicholas Konz and Maciej A Mazurowski. The effect of intrinsic dataset properties on generalization: Unraveling learning differences between natural and medical images. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[15] Chris Lin, Hugh Chen, Chanwoo Kim, and Su-In Lee. Contrastive corpus attribution for explaining representations. In *The Eleventh International Conference on Learning Representations*, 2023. 6

[16] Tim Miller. Contrastive explanation: a structural-model approach. *The Knowledge Engineering Review*, 36:e14, 2021. 6

[17] L. Morra, Luca Piano, Fabrizio Lamberti, and Tatiana Tommasi. Bridging the gap between natural and medical images through deep colorization. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 835–842, 2020. 1

[18] Hong Nguyen, Cuong V. Nguyen, Shrikanth Narayanan, Benjamin Y. Xu, and Michael Pazzani. Explainable severity ranking via pairwise n-hidden comparison: a case study of glaucoma, 2023. 2

[19] Hieu T. Nguyen, Ha Q. Nguyen, Hieu H. Pham, Khanh Lam, Linh T. Le, Minh Dao, and Van Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *medRxiv*, 2022. 4

[20] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, pages 27730–27744. Curran Associates, Inc., 2022. 2, 3

[21] Michael Pazzani, Severine Soltani, Robert Kaufman, Samson Qian, and Albert Hsiao. Expert-informed, user-centric explanations for machine learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12280–12286, 2022. 6

[22] Steffen Erhard Petersen, Mohammed Yunus Khanji, Sven Plein, Patrizio Lancellotti, and Chiara Bucciarelli-Ducci. European association of cardiovascular imaging expert consensus paper: a comprehensive review of cardiovascular magnetic resonance normal values of cardiac chamber size and aortic root in adults and recommendations for grading severity. *European heart journal cardiovascular Imaging*, 2019. 1

[23] Vipin Pillai, Soroush Abbasi Koohpayegani, Ashley Ouligian, Dennis Fong, and Hamed Pirsiavash. Consistent explanations by contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10213–10222, 2022. 6

[24] Yazan Qiblawey, Anas M. Tahir, Muhammad Enamul Hoque Chowdhury, Amith Abdullah Khandakar, Serkan Kiranyaz, Tawsifur Rahman, Nabil Ibtehaz, Sakib Mahmud, Somaya Al-Madeed, and Farayi Musharavati. Detection and severity classification of covid-19 in ct images using deep learning. *Diagnostics*, 11, 2021. 2

[25] Mary Qiu, Michael V. Boland, and Pradeep Y. Ramulu. Cup-to-disc ratio asymmetry in u.s. adults: Prevalence and association with glaucoma in the 2005–2008 national health and nutrition examination survey. *Ophthalmology*, 124(8):1229–1236, 2017. 6

[26] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3

[27] Ruslan Salakhutdinov and Geoffrey E. Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In *International Conference on Artificial Intelligence and Statistics*, 2007. 1, 2

[28] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven Truong, Chanh Nguyen, Van Doan Ngo, Jayne, DO Seekins, Francis Blankenberg, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4:867 – 878, 2022. 6

[29] Anshul Shah, Suvrit Sra, Rama Chellappa, and Anoop Cherian. Max-margin contrastive learning. *Proceedings of AAAI Conference (2022)*. 1, 2

[30] Peng Tian, Yuan Guo, Jayashree Kalpathy-Cramer, Susan Ostmo, John Peter Campbell, Michael F. Chiang, Jennifer Dy, Deniz Erdogmus, and Stratis Ioannidis. A severity score for retinopathy of prematurity. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 1809–1819, New York, NY, USA, 2019. Association for Computing Machinery. 2

[31] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, pages 6827–6839. Curran Associates, Inc., 2020. 6

[32] Sam B. Tran, Huyen T. X. Nguyen, Chi Phan, Ha Q. Nguyen, and Hieu H. Pham. A novel transparency strategy-based data augmentation approach for bi-rads classification of mammograms. In *2023 IEEE Statistical Signal Processing Workshop (SSP)*, pages 681–685, 2023. 2

[33] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 2, 6

[34] Lei Wang, Ee-Peng Lim, Zhiwei Liu, and Tianxiang Zhao. Explanation guided contrastive learning for sequential recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, page 2017–2027, New York, NY, USA, 2022. Association for Computing Machinery. 6

[35] Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*. MIT Press, 2005. 1, 2

[36] Dufan Wu, Kuang Gong, Chiara Daniela Arru, Fatemeh Homayounieh, Bernardo Bizzo, Varun Buch, Hui Ren, Kyungsang Kim, Nir Neumark, Pengcheng Xu, Zhiyuan Liu, Wei Fang, Nuobei Xie, Won Young Tak, Soo Young Park, Yu Rim Lee, Min Kyu Kang, Jung Gil Park, Alessandro Carriero, Luca Saba, Mahsa Masjedi, Hamidreza Talari, Rosa Babaei, Hadi Karimi Mobin, Shadi Ebrahimian, Ittai Dayan, Mannudeep K. Kalra, and Quanzheng Li. Severity and consolidation quantification of covid-19 from ct images using deep learning based on hybrid weak labels. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3529–3538, 2020. 1

[37] Tian Xia, Shaodan Zhai, and Shaojun Wang. Plackett-luce model for learning-to-rank task. *ArXiv*, abs/1909.06722, 2019. 2

[38] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. 1, 2, 4

[39] Qi Yang, Qiang Liu, Haibo Xu, Hong Lu, Shiyuan Liu, and Hongjun Li. Imaging of coronavirus disease 2019: A chinese expert consensus statement. *European Journal of Radiology*, 127:109008, 2020. 1

[40] Xin Yu, Yurun Tian, Fatih Porikli, Richard Hartley, Hongdong Li, Huub Heijnen, and Vassileios Balntas. Unsupervised extraction of local image descriptors via relative distance ranking loss. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2893–2902, 2019. 2

[41] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1476–1485, 2019. 1, 2

[42] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2074–2083, 2019. 2

[43] İlkay Yıldız, Peng Tian, Jennifer Dy, Deniz Erdoğmuş, James Brown, Jayashree Kalpathy-Cramer, Susan Ostmo, J. Peter Campbell, Michael F. Chiang, and Stratis Ioannidis. Classification and comparison via neural networks. *Neural Networks*, 118:65–80, 2019. 2