CLOC: Contrastive Learning for Ordinal Classification with Multi-Margin N-pair Loss

Dileepa Pitawela[♦] Gustavo Carneiro[▲] Hsiang-Ting Chen[♦]
dileepa.pitawela@adelaide.edu.au g.carneiro@surrey.ac.uk tim.chen@adelaide.edu.au
[♦]University of Adelaide, Australia [▲] CVSSP, University of Surrey, UK

Abstract

In ordinal classification, misclassifying neighboring ranks is common, yet the consequences of these errors are not the same. For example, misclassifying benign tumor categories is less consequential, compared to an error at the pre-cancerous to cancerous threshold, which could profoundly influence treatment choices. Despite this, existing ordinal classification methods do not account for the varying importance of these margins, treating all neighboring classes as equally significant. To address this limitation, we propose CLOC, a new margin-based contrastive learning method for ordinal classification that learns an ordered representation based on the optimization of multiple margins with a novel multi-margin n-pair loss (MMNP). CLOC enables flexible decision boundaries across key adjacent categories, facilitating smooth transitions between classes and reducing the risk of overfitting to biases present in the training data. We provide empirical discussion regarding the properties of MMNP and show experimental results on five real-world image datasets (Adience, Historical Colour Image Dating, Knee Osteoarthritis, Indian Diabetic Retinopathy Image, and Breast Carcinoma Subtyping) and one synthetic dataset simulating clinical decision bias. Our results demonstrate that CLOC outperforms existing ordinal classification methods and show the interpretability and controllability of CLOC in learning meaningful, ordered representations that align with clinical and practical needs.

1. Introduction

Ordinal classification, the assignment of data to discrete, ordered categories, is essential across fields where ranking impacts key decisions. In healthcare, assessments such as TNM for cancer staging [22], APACHE for intensive care [28], and the Glasgow Coma Scale for brain injury [54] are essential for guiding treatment and predicting outcomes. This approach is also widely employed in psychology (e.g., Likert scales), education (grading proficiency levels), and finance (credit scoring), where maintaining ordinal relationships is key to deriving meaningful insights that inform actionable decisions [24].

Code is available at https://github.com/dpitawela/CLOC

Figure 1 illustrates the impact of various types of classification errors in ordinal contexts. Unlike standard classification problems, in ordinal classification or ranking problems, **not all label mistakes are equal** [8, 13, 58]; misclassifications with larger rank differences carry more significant consequences. For example, they often represent greater discrepancies in disease severity or treatment needs. Moreover, **not all adjacent label mistakes are equally significant**; for instance, an error between two benign tumor categories may have less impact compared to a misclassification at the boundary between pre-cancerous and cancerous stages, where the shift in diagnosis could significantly alter treatment and management [2, 52]. Similarly, a misclassification between minors and adults would have a larger impact than other adjacent ordinal categories [19], as it could lead to more substantial legal or social implications.

Existing methods have not been designed to address both of these challenges. *Regression* attempts to model ordered categories, but its single continuous output often loses the ordinal structure and lacks clear boundaries between ranks [14, 35, 62]. Additionally, its error penalties are less flexible, as typical regression or models, such as those minimizing mean squared error, do not inherently account for the ordinal relationships between classes nor allow modification. *Ordinal classification* methods [8, 40] are well-suited for tasks involving ordered categories, as they preserve the ranking structure and focus on maintaining the correct order, rather than merely predicting label distributions. However, most existing approaches focus on preserving order and minimizing label overlap in adjacent ranks, without considering the varying significance of adjacent label errors or examining the impact of margins between ranked labels.

We propose the new Contrastive Learning for Ordinal Classification (CLOC) method that is based on an optimization that minimizes the novel Multi-Margin N-Pair (MMNP) loss. The new MMNP loss is a central contribu-

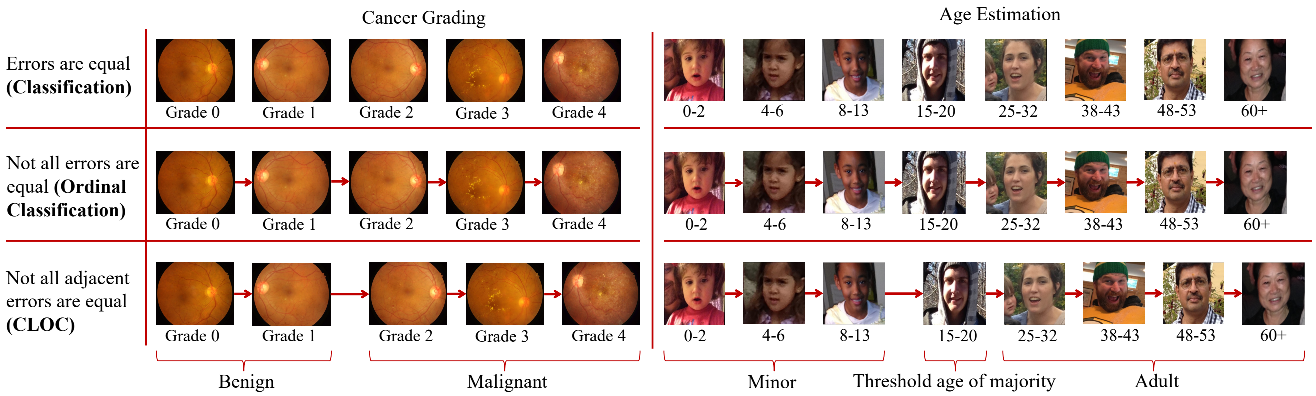

Figure 1. An illustration of the impact of classification errors in ordinal contexts. Standard classification treats all errors equally, while ordinal classification penalizes larger rank errors more heavily. CLOC further accounts for the varying impact of adjacent errors, particularly across critical boundaries like benign versus malignant categories or minor vs threshold age of majority. This figure shows samples from IDRID (cancer grading) [43] and Adience (age estimation) [16] datasets. Longer arrow indicates a greater impact from adjacent errors.

tion of this paper that introduces multiple learnable margins that allow flexible decision boundaries between adjacent labels, while its cumulative property preserves class ordering. These multiple learnable margins improve generalization by preventing overfitting to specific biases or minor variations in the training data, while also enabling more robust transitions between classes than previous ordinal classifiers that rely on fixed, hard decision boundaries. Additionally, CLOC enables nuanced ordinal classifications, making it particularly suited for challenging tasks like medical image analysis, where subtle distinctions between some categories are crucial. Furthermore, these learnable margins also offer interpretability of the importance between ordinal categories and provide an option for manual adjustment, allowing flexible control over critical decision boundaries. To summarize, our key contributions are:

- Novel Contrastive Learning for Ordinal Classification (CLOC) method that is based on an optimization that minimizes the novel Multi-Margin N-Pair (MMNP) loss, which provides learnable margins for flexible decision boundaries while preserving the ordinal structure.
- A novel method allowing users to control the model behavior through controlled margins, with experiments showing its effectiveness and robustness to biases.
- Enhanced interpretability via multi-margin learning, offering a new tool for interpretable AI and insights into class separability and decision boundaries in ordinal classification tasks.

We conducted experiments on five real-world image datasets (Adience, Historical Colour Image Dating, Knee Osteoarthritis, Indian Diabetic Retinopathy Image, and Breast Carcinoma Subtyping), as well as a synthetic dataset designed to simulate clinical decision biases. Our results show that CLOC outperforms existing ordinal classification methods. We also show empirical evidence of both CLOC's interpretability and controllability in learning meaningful, ordered representations that align closely with clinical and practical needs.

2. Related Work

Ordinal Classification: Early ordinal classification approaches [4, 23] relied on standard classification or regression, missing ordinal relationships and non-uniform label separations. More recent models like ORCNN [40] and RankingCNN [8] use multiple binary classifiers to predict rank order, while CNNPOR [37] introduced pairwise ordinal constraints to preserve structure. Methods like SORD [13] and POEs [34] use soft labels and probabilistic embeddings, respectively, to better capture ordinal information. Unlike these, which assume fixed class-pair error distributions that limit prioritization of critical boundaries, CLOC enables class-pair-specific distributions and allows direct, meaningful adjustments over critical decision boundaries.

Order learning approaches such as pairwise comparators [36] and Deep Repulsive Clustering (DRC) [30] compare instances directly to infer rank but are sensitive to reference quality. Moving Window Regression (MWR) [49] refines ranks iteratively with global and local regressors, though it also depends on good reference selection.

Vision-language models like OrdinalCLIP [35], Wang et al. [58], and NumCLIP [14] utilise CLIP for ordinal regression with rank-specific prompts and regularization, but these large models (190M+ parameters) are data-intensive. In contrast, our method efficiently captures ordinal relationships using a smaller model with contrastive learning, enabling scalable rank learning without heavy reliance on probabilistic distributions or large architectures.

Representation learning: Representation learning has been widely explored in classification, with contrastive learning becoming a leading approach, particularly in self-supervised contexts. The supervised variant, SupCon [27], has shown superior performance over traditional cross-entropy loss in tasks such as image recognition [27], long-tailed classification [57], out-of-domain detection [61], and anomaly detection [29]. While adaptations for ordinal tasks exist, such as in gaze estimation [59] and medical imaging [10, 15], these methods often fail to account for the ordinal nature of classes.

To address this, Xiao et al. [60] proposed using labeled distances to capture semantic similarities and maintain local structure. Li et al. [33] introduced rank-proportional distance scaling, while Suarez et al. [53] aligned embedding distances with rank differences. The work by [31] encoded order and metric relations using sample direction and distance. Further enhancements include Ordinal Log-Loss [6], which uses distance-aware weighting, and RankSim [21], which applies regularization for ordinal consistency. Rank-N-Contrast [62] introduced regression-aware embeddings to align distances with continuous targets.

In contrast, our method inherently models ordinal relationships through contrastive learning without relying on predefined metrics or additional regularization, leading to more flexible and robust feature representations.

Contrastive learning: CL focuses on maximizing similarity between positive pairs while separating negative ones, with temperature scaling and margin adjustments being critical parameters. Wang and Liu [56] shows that temperature tuning helps balance feature alignment and tolerance for semantically similar samples, while Shah et al. [48] introduces SVM-like margins in CL, though extending this to ordinal tasks is challenging. Rho et al. [45] explores single-margin adjustments for better generalization, but multi-margin strategies remain unexplored. Key CL frameworks like SimCLR [9] and DINO [5] focus on unsupervised learning through augmentations and self-distillation, respectively. SupCon [27] adapts CL for supervised tasks by clustering same-class samples, and Rank-N-Contrast (RNC) [62] introduces a rank-based loss for ordinal representation. Our approach, CLOC, builds on these advances with a novel multi-margin loss tailored for ordinal classification, showing superior performance and introducing better controllability and interpretability over decision boundaries across diverse datasets.

3. Methodology

Ordinal classification is a special case of multi-class classification where the labels have a natural ordering or ranking among them. Let the training dataset be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathcal{X}$ is a data sample and $y_i \in \mathcal{Y} = \{r_1, r_2, ..., r_C\}$ is ground truth label from a C-class problem that follow an ordered rank $r_1 \prec r_2 \prec ... \prec r_C$

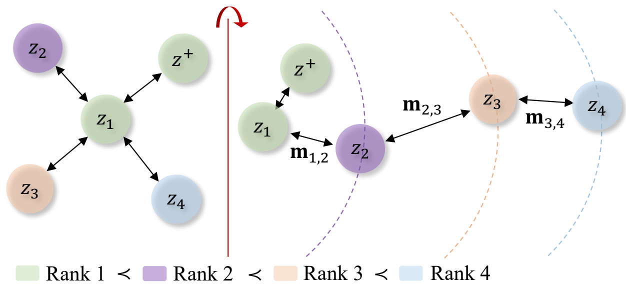

Figure 2. Set of data samples from different classes/ranks (indicated by colors), with their rank order relative to the anchor z_1 and its positive sample z^+, where rank $1 \prec$ rank $2 \prec$ rank $3 \prec$ rank 4. The MMNP loss pulls samples of the same rank closer to the anchor while pushing those with larger rank differences farther apart, resulting in wider margins in the representation space. Additionally, we can control specific margin values (e.g., by applying constraints) and use the learned margins to interpret class boundaries.

with C denoting the number of ranks, and $\prec$ indicating the ordering between different ranks. Similar to classification, ordinal classification seeks to predict $\hat{y}_i \in \mathcal{Y}$ from x_i. Data are passed through an encoder $e_\phi(\cdot)$ to generate feature embeddings, $e_\phi : \mathcal{X} \to \mathcal{Z}; \mathcal{Z} \in \mathbb{R}^d$ and followed by a classifier $c_\gamma(\cdot)$ to get output embeddings, $c_\gamma : \mathcal{Z} \to \mathcal{V}; \mathcal{V} \in \mathbb{R}^C$. The embeddings are subsequently converted to $\hat{y} = \mathsf{OneHot}(\mathsf{softmax}(\mathbf{v}))$ where $\mathbf{v} \in \mathcal{V}$, $\mathsf{softmax} : \mathbb{R}^C \to \Delta^{C-1}$ and $\mathsf{OneHot} : \Delta^{C-1} \to \mathcal{Y}$, returning a one-hot label representing the class with the largest prediction. In this context, the terms 'class', 'label' and 'rank' are used interchangeably to refer to the ordinal classification label.

3.1. Contrastive Learning for Ordinal Classification (CLOC)

CLOC addresses ordinal classification by learning a rank-aware representation through contrastive learning with multiple learnable classification margins. Unlike previous margin-based learning methods [51], which rely on a fixed, single margin, CLOC's multiple learnable margins allow the model to better capture more nuanced separations between ranks, offering added flexibility for various applications. This approach not only provides finer control over specific margins but also enhances interpretability of the learned boundaries, addressing limitations in earlier fixed-margin methods.

Formally, in a C-class classification setting, we define $C-1$ pairs of consecutive ranks indexing the margins, represented by the set $\mathcal{H} = \{(r_1, r_2), (r_2, r_3), ..., (r_{C-1}, r_C)\}$. The overall optimization of CLOC is based on minimizing the following objective function

$$\phi^*, \gamma^*, m^*_{\mathcal{H}} = \arg\min_{\phi,\gamma,m_{\mathcal{H}}} \frac{1}{N} \sum_{(x,y)\in\mathcal{D}} \ell_{CE}(y,\hat{y}) + \ell_{MM}(z,y,m_h)$$
$$\text{subject to } m_h > \rho, \forall h \in \mathcal{H} \tag{1}$$

where $m_{\mathcal{H}} = \{m_h\}_{h\in\mathcal{H}}$ with $m_h \in \mathbb{R}$, $z = e_\phi(x)$, $\hat{y} = c_\gamma(z)$, $\ell_{CE}(y,\hat{y}) = -y^\top \log(\hat{y})$ which is cross-entropy, and $\ell_{MM}(z,y)$ denotes the proposed MMNP loss defined below in Sec. 3.2. Overall, CLOC jointly optimizes ϕ, γ and $m_{\mathcal{H}}$ to learn a representation for ordinal classification that allows flexible decision boundaries between adjacent ranks and preserve the ordinal relationship across ranks.

3.2. Multi Margin N-Pair (MMNP) Loss

As shown in Fig. 2, the MMNP loss, denoted by $\ell_{MM}(\cdot)$ in Eq.(1), follows a contrastive learning setup, where the set of positive and negative samples for the anchor $(x,y) \in \mathcal{D}$, which produces the embedding $z = e_\phi(x)$, is defined respectively as $\mathcal{S}^+ = \{(z_j, y_j) \mid (x_j, y_j) \in \mathcal{D}, z_j = e_\phi(x_j), z_j \neq z, y_j = y\}$, and $\mathcal{S}^- = \{(z_k, y_k) \mid (x_k, y_k) \in \mathcal{D}, z_k = e_\phi(x_k), y_k \neq y\}$. In its simplest form, as an optimization over all pairs jointly, MMNP is given by,

$$\ell_{MM}(z,y,m_h) = \sum_{(z_j,y_j)\in\mathcal{S}^+} \sum_{(z_k,y_k)\in\mathcal{S}^-} \max(0,\ \mathsf{m}_{y,y_k} + \psi(z,z_k) - \psi(z,z_j)) \tag{2}$$

where $\psi(\cdot,\cdot)$ represents cosine similarity function, and $\mathsf{m}_{y,y_k} = m_{y,y+1} + ... + m_{y_{k-1},y_k}$ (inverse the index if $y_k \prec y$). That is, m_{y,y_k} is the cumulative sum of the margins between the anchor's rank y and the rank y_k of a negative sample. In practice, instead of forming the positive and negative sets from the whole dataset, we form them within the mini-batches, as explained in Sec. 3.3.

Properties of the MMNP Loss: The MMNP loss enhances contrastive learning by introducing cumulative multiple margins m_{y,y_k} in the embedding space. Previous contrastive loss based methods used for ordinal classification repels all negative samples in $\mathcal{S}^-$ from the anchor z equally, regardless of rank differences. In contrast, MMNP loss repels the negative samples from more distant ranks more strongly than those from closer ranks and thus preserve the ordinal relationship in the embedding space. More specifically, minimizing the MMNP loss imposes the condition $\psi(z,z_k) + \mathsf{m}_{y,y_k} \leq \psi(z,z_j)$. It means that given two negative samples (z_l, y_l) and (z_m, y_m) with ranks $y \prec y_l \prec y_m$, the similarity between the anchor z and z_m is constrained to be lower than that with z_l due to the cumulative margin $\mathsf{m}_{y,y_m} = \mathsf{m}_{y,y_l} + \mathsf{m}_{y_l,y_m}$. Additionally, the actual margin value between samples belonging to two consecutive ranked classes represents the learned distance to robustly separate samples from those two classes. Hence the learned margins can be used to interpret the difficulty in separating samples from consecutively ranked classes. Furthermore, note from Eq. (1), that it is possible to constrain m_h to a minimum value ρ, which is helpful when one wants to minimize mistakes between critical classes. Discussions about controlability and intepretability are in sections 4.4 and 4.5.

3.3. Training

As shown in Figure 3, CLOC employs a 2-phase training process [47]. In the first phase, the encoder, classifier and margins are jointly optimized to determine the margins that best separate classes and allow rank-aware representation, establishing a structured guide for representation learning. In the second phase, these margins are frozen, allowing the model to refine the representation while adhering to the learned margins. The two-phase training strategy is employed to prevent the margins from collapsing to a trivial solution during the co-optimization of the representation and multiple margins, which we discuss below in this section, with additional details in Appendix B.

Following the batch construction approach in [51], during both training phases, every mini-batch has, at least, samples from two distinct ranks, with a minimum of two samples taken from each rank. This setup ensures that each contrast includes at least one anchor, one positive sample, and two negative samples.

Phase One: Joint Learning of Margins and Representation

In this phase, we run the optimization in (1) to learn the margins $m^*_{\mathcal{H}}$, parameters of encoder ϕ^* and classifier γ^* for producing the embeddings. This phase seeks to optimize two objectives simultaneously: creating an ordered representation that is well-suited for classification, and learning margins that best separate ranks. However, this dual objective introduces a potential risk of margin collapse, where

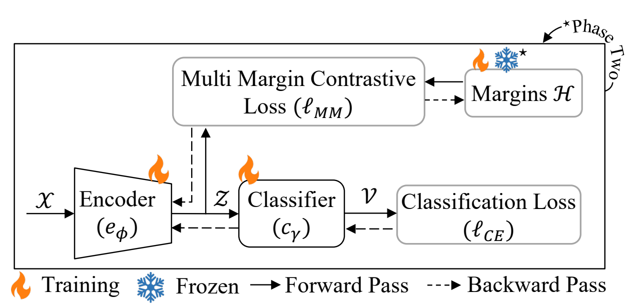

Figure 3. Two phase training for CLOC. In the first phase, the encoder, classifier and the margins are trained with the multi-margin contrastive loss and classification loss. In the second phase, the margins are frozen allowing the training of the encoder and classifier with the losses.

all margins go to zero. Margins may converge toward zero to reduce the overall loss in Eq. (1), as setting margins to zero simplifies the optimization problem where the feature extractor is no longer incentivized to produce a nuanced ranked representation, potentially undermining the goal of learning structured, rank-separated embeddings.

To reduce the rate of margin collapse and enable training for a large number of epochs, we implement following "precautionary" measures: 1) the learnable margins are activated using the softplus function, that provides a smooth lower bound and approaches to zero gradually; 2) the margins are initialized with random values drawn from a uniform distribution in the interval $[0.5, 1.0)$, ensuring a moderate start avoiding excessively small or large initial margins; and 3) early stopping is applied to prevent collapse.

Phase Two: Margin-guided Representation Learning

In the second phase, we optimize (1) only with respect to ϕ and γ with the learned margins being frozen. This second phase ensures that the frozen margins that learned in phase one guide the feature extractor while continuing to improve the classification accuracy.

The two-phase training allows integrating the precautionary mechanisms to phase one to address margin collapse. In phase two, with margins fixed, the model exclusively focuses on refining the feature representation for better convergence to an optimal solution.

4. Experiments

We conducted three experiments evaluating the controllability, interpretability, and overall performance of CLOC.

4.1. Datasets

The **Adience** dataset [16] serves as a benchmark for age prediction from facial images. It contains 26,580 colour images of 2,284 subjects, with each image categorized into one of eight age groups. We use the in-plane aligned version and five-folds originally released. Each fold contains an average of 7,372 images, yeilding 7,372 test images and 29,490 training images per run. Per fold image count along with the average class distribution, is in appendix A.1.

The **Historical Colour Image Dating (HID)** dataset [41] is a benchmark for evaluating algorithms that predict the decade a colour image was captured. The dataset spans five decades, from the 1930s to the 1970s, and consists of 1,325 images, with 265 images for each decade. Similar to [20, 41], we run standard five-fold cross-validation by randomly selecting 50 images for testing, with the remaining 215 images for training from each decade. Appendix A.2 visualizes samples from this dataset.

The **Knee Osteoarthritis (KOA)** dataset [7] aims to assess severity by the Kellgren-Lawrence (KL) grading system, where grade 0 represents a healthy knee and grade 4 indicates severe KOA. The dataset consists of 8,260 X-ray images, left and right joints combined. We train with 6,604 images and evaluate with 1,656 images. The class distribution of splits is in appendix A.3. The **Indian Diabetic Retinopathy Image Dataset (IDRiD)** [43] is designed to classify retinal fundus images into diabetic retinopathy grades defined by international clinical Diabetic Retinopathy (DR) scale. This scale identifies 5 grades, from 0 being no apparent DR to 4 being severe DR. The dataset contains 516 fundus images, split into 413 training images and 103 testing images, where each is assigned one of the 5 DR grades. The appendix A.4 details the class distribution of splits. The **BReAst Carcinoma Subtyping (BRACS)** dataset [3] facilitates classification of breast tumors from histological images. The grading system identifies 7 different subtypes of lesions ranging from grade 0 being normal to grade 6 being invasive carcinoma. The dataset includes 4,539 regions of interests from 189 patients split into 3,969 training and 570 testing samples. The class distribution of the dataset is in appendix A.5.

4.2. Experimental Setting

For Adience, the in-plane aligned version of the dataset is used, following the standard five-fold, subject-exclusive cross-validation protocol outlined in [16, 32]. The image size is set to 224x224 pixels, with random horizontal and vertical flips, affine transformations, and color jittering applied as data augmentations.

For HID, as described in [20, 41], standard five-fold cross-validation is performed by randomly selecting 50 images for testing and using the remaining 215 images for training from each decade. Training involves a random crop of 224x224 pixels, along with random horizontal and vertical flips, affine and perspective augmentations, and the AutoAugment protocol from [12] with a probability of 0.01. For testing, a center crop of 224x224 pixels is used.

For KOA, a center crop of the image at 120x200 pixels is applied, followed by histogram equalization. During training, random affine transformations, Gaussian blur, and perspective augmentations are used.

For both IDRID and BRACS, the image size is set to 224x224 pixels. During training, random horizontal and vertical flips, rotations, and Gaussian blur are applied. Additionally, random color jittering is used for IDRID, while random perspective augmentations are applied for BRACS.

For all experiments, an ImageNet pre-trained ResNet-50 [26] encoder is utilized, followed by two linear layers as the classifier. Both training phases use Adam optimizer with a learning rate of 0.001, with $\rho = 0$ (as per eq.1) and trains for 500 epochs. Early stopping is applied in Phase One when training accuracy reaches 95% and in Phase Two after 10 unimproved epochs. All experiments are performed on an

Method	Accuracy↑					Mean Absolute Error↓				
	Adiance	HID	KOA	IDRID	BRACS	Adiance	HID	KOA	IDRID	BRACS
Comparison with Ordinal Classification Methods										
ORCNN [40]	0.5670	0.3870	0.4597	0.5833	0.4677	0.5400	0.9500	0.8335	0.5972	0.8732
POE [34]	0.6050	0.5468	0.6410	0.5730	0.5160	0.4700	0.7600	0.4440	0.6770	0.8280
GOL [31]	0.6250	0.5620	0.6439	0.6311	0.5754	0.4300	**0.5500**	0.4330	0.6890	0.7060
MWR [49]	0.6260	0.5780	0.6380	0.6408	0.5649	0.4500	0.5800	0.5800	0.6752	0.8121
RnC † [62]	0.4679	0.5664	0.5886	0.5825	0.4632	0.8600	0.7288	0.5396	0.6152	0.8491
Comparison with Contrastive Learning Methods										
SimCLR [9]	0.3053	0.3988	0.5286	0.4563	0.3544	1.5303	1.0880	0.6964	1.0679	1.4859
SupCon [27]	0.5774	0.4812	0.5955	0.5243	0.5667	0.5131	0.8040	0.4896	0.7476	0.7877
DINO [5]	0.4691	0.5004	0.5069	0.5437	0.5263	0.8230	0.8564	0.7555	0.7476	1.0719
RnC † [62]	0.4679	0.5664	0.5886	0.5825	0.4632	0.8600	0.7288	0.5396	0.6152	0.8491
CLOC (Ours)	**0.6302**	**0.6208**	**0.6673**	**0.7379**	**0.6035**	**0.4100**	0.5534	**0.4167**	**0.4078**	**0.7054**

Table 1. Comparison of test set accuracy and mean absolute error of CLOC with contemporary ordinal classification methods (above) and contrastive learning methods (below) across five benchmark datasets. (†RnC is both an ordinal classification and a contrastive method.)

CLOC Margins	Error rate between class boundaries ↓							Overall Acc.↑
	C1 - C2	C2 - C3	C3 - C4	C4 - C5	C5 - C6	C6 - C7	C7 - C8	
IDRID								
Learned	0.1282	0.0270	0.0980	0.0937				0.7379
Controlled	0.1025	0.0000	0.1176	0.1562				0.6990
KOA								
Learned	0.2220	0.1700	0.0985	0.0620				0.6673
Controlled	0.2480	0.0900	0.1160	0.0620				0.6431
HID								
Learned	0.0300	0.1300	0.1500	0.2500				0.6280
Controlled	0.0200	0.1700	0.0800	0.2700				0.5640
Bracs								
Learned	0.1750	0.0931	0.0545	0.0555	0.1950	0.0602		0.6035
Controlled	0.2560	0.0740	0.0420	0.0490	0.0850	0.0180		0.4281
Adience								
Learned	0.8410	0.0955	0.1096	0.1899	0.2991	0.1522	0.2968	0.6134
Controlled	0.8600	0.1210	0.0640	0.2382	0.3247	0.1294	0.1718	0.5591

Table 2. Comparison of the overall accuracy and error rates for each pair of adjacent ranked classes between the original CLOC (Learned), where all margins are learned, and a modified version with fixed margins at critical decision boundaries (Controlled). Critical decision boundaries are highlighted in table 2.

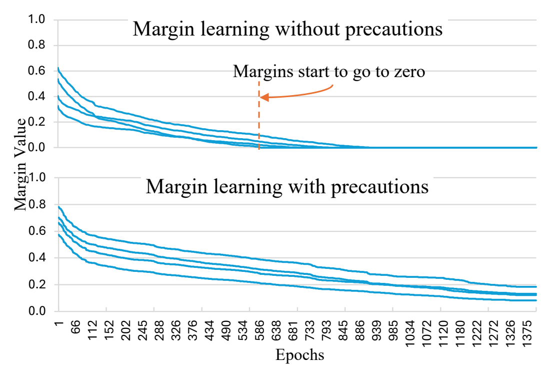

Figure 4. Training phase one with and without margin collapse "precaution" measures. Notice that training continues for larger number of epochs without margin collapse when "precaution" measures are applied.

RTX 4090 GPU.

We quantitatively assess performance by measuring accuracy and mean absolute error (MAE) on samples belonging to the test set of the datasets introduced above.

4.3. Comparison with Related Methods

Our experimental results, summarized in Table 1, compare CLOC against both ordinal classification and contrastive learning baselines across five datasets. For ordinal classification, we include ORCNN [40], POE [34], MWR [49], GOL [31], RnC [62] and for contrastive learning methods, we cover SimCLR [9], DINO [5], SupCon [27], and RNC [62]. Each contrastive learning method is evaluated using a ResNet-50 backbone pre-trained on ImageNet, with the full model re-trained on each respective training set and assessed on the corresponding test set. The results show that CLOC outperforms all other methods across every dataset. In particular, notice that for some datasets, such as HID and IDRID, CLOC is substantially better than competing methods, with accuracy 0.6202 (compared to 0.5664 of the second best–RnC) on HID and 0.7379 (compared to 0.6408 of the second best–MWR) on IDRID.

In addition, we provide an experiment with large number of classes in Appendix C.1 and a computational time comparison with related methods in Appendix C.2.

4.4. Controllability and Robustness to Biases

As shown in Figure 1, there can be truly consequential classification boundaries in a dataset, such as between cancer and non-cancer classification in a diagnostic medical imaging dataset, or at the age of majority in a face dataset. As explained in Sec. 3.2, our method allows the use of constraints in the learning of such critical margins, which is empirically demonstrated in this section.

Controlling the learning of critical margins: This experiment examines the effectiveness of manually fixing the margin to reduce errors between adjacent classes that intersect a key decision boundary. For the IDRID and KOA datasets, the key decision boundaries that influence treatment choices lie between grades 1 and 2 [39, 55]. BRACS has two critical boundaries to separate benign grade 2 from atypical grade 3, and atypical grade 4 from malignant grade 5. For the Adi-

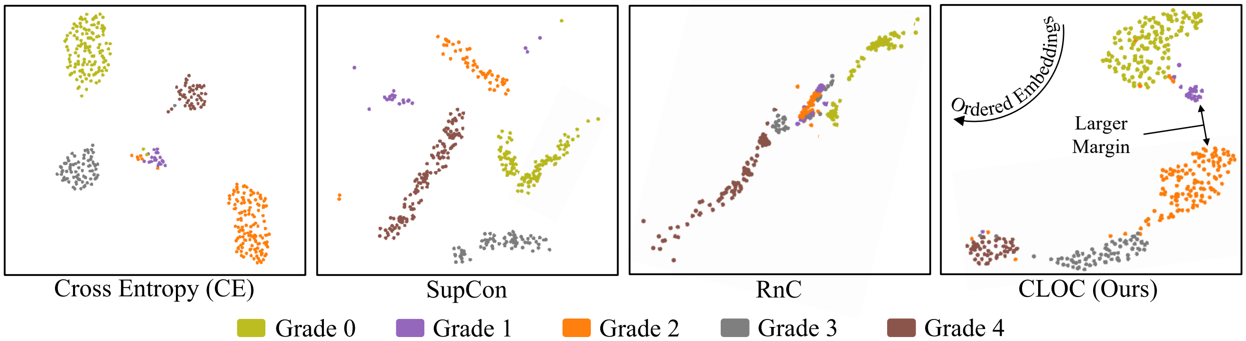

Figure 5. UMAP visualizations of learned representations for the IDRID dataset that focuses on cancer grade classification. Notice that proposed CLOC, effectively captures the inherent ordinal structure of cancer grades and maintains a larger margin at critical decision boundary, compared to other decision boundaries. Additional visualizations of GOL [31] and POE [34] are in Appendix C.4.

ence dataset, a practical critical decision boundary would be between age groups 8-13 and 15-20 to separate minors from threshold age of majority. For completeness, we assume a hypothetical critical boundary between the era of 50s and 60s in the HID dataset.

During the 2-phase optimization, we fix a larger value for the aforementioned critical decision margins and only optimize the remaining margins. The rest of the experiment settings were identical to Section 4.2.

Table 2 presents a comparison of overall accuracy and individual error rates between adjacent ranks for the learned margins approach (where all margins are optimized) and the controlled margin approach (where margin values are fixed at the critical decision boundary). Under the controlled margin condition, we observe a reduction in classification errors between classes near the critical boundary, albeit at the cost of lower overall accuracy.

Robustness to Biases: This experiment simulates a scenario of *diagnostic bias*. Studies in the medical domain have shown that practitioners have different subjective assessments [1, 42] and often lean towards positive diagnoses in high-stake cases [11, 46], a tendency explained by Error Management Theory (EMT) [25] and risk aversion. EMT suggests that when errors have unequal costs, people tend to err on the side of the less costly mistake; in medicine, a false negative (missed diagnosis) is typically viewed as more severe than a false positive. This bias, combined with doctors' inherent risk aversion to potentially missing a critical diagnosis, drives them to favor caution, increasing the likelihood of positive diagnoses even at the expense of false positives.

The experiment simulates diagnostic bias in the IDRID training set at a critical decision boundary between grades 1 and 2 by randomly relabeling 60% of the grade 1 samples as grade 2 and 30% of the grade 2 samples as grade 1. We report both the results of using learnable margins, where the model learns all margins automatically, and the results of a manually set margin that enforces a fixed, larger margin between grades 1 and 2. Tab. 3 presents a comparison with related methods on simulated biased training data. The results demonstrate that CLOC with a fixed critical decision margin is significantly more robust to label noise compared to CLOC with fully learnable margins. Compared to related methods, CLOC achieves accuracy gains exceeding 7.9% over the second best–GOL (Tab. 3).

4.5. Interpretability

The multi-margin property of CLOC offers a novel way to interpret data by the representation margins between ordinal classes. First, users can quantitatively examine the margins to infer the relative separation between ranks. For instance, in the IDRID dataset, the margins between consecutive grades are 0.35, 0.41, 0.23, and 0.30, where the largest margin (0.41) highlights the critical decision boundary—indicating the most challenging region for separation. A similar trend is observed in the Adience dataset, with margins of 0.11, 0.08, 0.20, 0.12, 0.13, 0.09, and 0.12, where the largest margin (0.20) occurs between age groups 8–13 and 15–20, again marking a critical boundary.

In addition, we further assess the interpretability of CLOC by comparing its UMAP [38] visualizations to those of nominal classification (CE), SupCon, and RnC on the IDRID dataset. As shown in Fig. 5, CLOC effectively preserves the ordinal structure of the IDRID grades and establishes a wider margin between grades 1 and 2, the dataset's most critical decision boundary. In contrast, RnC maintains the correct grade order but lacks inter-class separability. CE and SupCon form distinct clusters but fail to capture the ordinal relationships, showing the advantage of CLOC in maintaining order while improving class separation.

Method	Accuracy↑	MAE↓
SimCLR [9]	0.4661	1.0583
SupCon [27]	0.4660	0.9029
DINO [5]	0.3883	1.1262
ORCNN [40]	0.3786	1.0970
POE [34]	0.5240	0.7210
GOL [31]	0.6117	0.6700
MWR [49]	0.4039	0.9420
RnC [62]	0.5437	0.8544
CLOC (Learned)	0.6602	0.5340
CLOC (Fixed)	0.7087	0.4272

Table 3. Comparison of CLOC (learned or fixed margin) vs. related methods on biased training sets at critical boundaries.

At end of Training	Accuracy ↑					Mean Absolute Error ↓				
	Adiance	HID	KOA	IDRID	BRACS	Adiance	HID	KOA	IDRID	BRACS
Ablation study 1: With $C-1$ margins each set to 1										
phase 2	**0.6306**	**0.6492**	0.6612	0.7184	0.5912	**0.4000**	**0.5432**	0.4257	0.4272	0.8228
Ablation study 2: With a single learnable value for $C-1$ margins										
phase 1	0.5803	0.5916	0.6413	0.6893	0.5544	0.5500	0.6436	0.4644	0.5437	0.8649
phase 2	0.5939	0.5944	0.6594	0.7184	0.5860	0.5100	0.6396	0.4457	**0.4078**	0.7860
Ablation study 3: With $C-1$ learnable values for $C-1$ margins (CLOC)										
phase 1	0.5852	0.6028	0.6600	0.7282	0.5930	0.5500	0.6292	0.4426	**0.4078**	0.7333
phase 2	0.6302	0.6208	**0.6673**	**0.7379**	**0.6035**	0.4100	0.5534	**0.4167**	**0.4078**	**0.7054**

Table 4. Ablation study: 1) fixing all margins to one, 2) learning a single value for $C-1$ margins, and 3) learning $C-1$ margins (CLOC).

4.6. Ablation Studies

We experiment with three variations of CLOC, by manipulating margin parameters in different ways: 1) setting a fixed value of 1 to all margins, 2) learning a single value for all $C-1$ margins, and 3) learning every $C-1$ margin. All experiments follow the same setup described in Sec. 4.2.

Table 4 reports results for these ablation studies. It is evident that learning all $C-1$ margins results in higher accuracy and lower MAE compared to learning a single value for all margins. When all margins are set to 1, two out of the five datasets perform better, with a significant improvement only in the HID dataset. Overall, the results indicate that learnable multiple margins improves performance, indicating that fixed, uniform margins between consecutive classes may not effectively capture the nuanced differences among them. This finding highlights the advantage of learnable, class-specific margins to better model these distinctions.

We also observe that introducing a second training phase focused on learning representation using margins learned in phase one consistently improves accuracy compared to jointly optimising representation and margins.

Figure 4 shows margin learning in Phase One on IDRID with and without precautions introduced in Sec. 3.3, demonstrating that with "precautionary" measures, training can continue for more epochs without margin collapse. Appendix C.3 shows CLOC's accuracy increases as the number of parameters in different base models increases.

5. Discussion

Beyond overall accuracy: our result highlights a crucial insight: overall accuracy, while commonly used, may not adequately reflect the true importance of different outcomes in practical applications. By allowing flexible error prioritization, CLOC proves especially valuable in contexts where critical decision thresholds influence high-stakes outcomes, such as in medical applications. This finding opens a promising research direction focused on developing more refined metrics for classification, tailored to reflect the varying weights of errors in real-world contexts where some mistakes carry more serious consequences than others.

Connection between margins and decision boundaries: The coincidence of the largest learned margins with critical decision boundaries is intriguing (Section 4.5). One possible explanation is that the initial decision guidelines were defined based on distinct visual differences, which aligned the decisions with the most prominent visual features, making it easier to achieve separation in the representation space. Additionally, expert annotators may exercise greater caution near key boundaries, leading to more precise labeling and clearer representations. It is important to emphasize that the margin value alone does not offer a definitive explanation, as the confounding factors like visual distinctiveness and annotator caution are difficult to disentangle. Instead, much like other interpretable AI methods [18, 44], the margin value acts as an additional tool for understanding model behavior, providing valuable insights, particularly in high-stakes applications where decision boundaries are crucial.

Limitation: CLOC's performance decreases on datasets with a large number of classes, potentially due to the complexity of optimizing numerous margins (Appendix C.1). However, CLOC excels in ordinal classification, especially in medical applications, which typically feature moderate class numbers and decision boundaries.

6. Conclusion

We introduced CLOC, a new contrastive learning method for ordinal classification that learns an ordered representation based on the optimization of multiple margins with a novel multi-margin n-pair loss. CLOC learns flexible margins that adapt to the varying importance of adjacent label errors while maintaining ordinal relationships. Experiments on five real-world and one synthetic dataset show that our approach outperforms others. Beyond improved overall accuracy, CLOC offers enhanced controllability and interpretability, enabling practitioners to prioritize error reduction between specific ranks and set decision boundaries based on application needs, making it effective for complex tasks like medical image analysis.

Acknowledgments

G.C. acknowledges support from the Engineering and Physical Sciences Research Council (EPSRC) through grant EP/Y018036/1.

References

[1] Jodie C Avery, Alison Deslandes, Shay M Freger, Mathew Leonardi, Glen Lo, Gustavo Carneiro, G Condous, Mary Louise Hull, Louise Hull, Jodie Avery, et al. Noninvasive diagnostic imaging for endometriosis part 1: a systematic review of recent developments in ultrasound, combination imaging, and artificial intelligence. *Fertility and sterility*, 121(2):164–188, 2024. 7

[2] Jonathan Birch, Kathleen A Creel, Abhinav K Jha, and Anya Plutynski. Clinical decisions using ai must consider patient values. *Nature medicine*, 28(2):229–232, 2022. 1

[3] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093, 2022. 5

[4] Refik Can Malli, Mehmet Aygun, and Hazim Kemal Ekenel. Apparent age estimation using ensemble of deep learning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 9–16, 2016. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3, 6, 8

[6] François Castagnos, Martin Mihelich, and Charles Dognin. A simple log-based loss function for ordinal text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4604–4609, 2022. 3

[7] Pingjun Chen, Linlin Gao, Xiaoshuang Shi, Kyle Allen, and Yang Lin. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics*, 75:84–92, 2019. 5

[8] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5183–5192, 2017. 1, 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 6, 8

[10] Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C Thai, Kathleen Moore, Robert S Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical image analysis*, 79:102444, 2022. 3

[11] Marie Crowe, Maree Inder, and Richard Porter. Conducting qualitative research in mental health: Thematic and content analyses. *Australian & New Zealand Journal of Psychiatry*, 49(7):616–623, 2015. 7

[12] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123, 2019. 5

[13] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4738–4747, 2019. 1, 2

[14] Yao Du, Qiang Zhai, Weihang Dai, and Xiaomeng Li. Teach clip to develop a number sense for ordinal regression. *arXiv preprint arXiv:2408.03574*, 2024. 1, 2

[15] Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michele Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm Mcdonald, Camille Marie Piguet, et al. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 58–68. Springer, 2021. 3

[16] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014. 2, 5

[17] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzàlez, Hugo J. Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 243–251, 2015. 2

[18] Jordan D Fuhrman, Naveena Gorre, Qiyuan Hu, Hui Li, Issam El Naqa, and Maryellen L Giger. A review of explainable and interpretable ai with applications in covid-19 imaging. *Medical Physics*, 49(1):1–14, 2022. 8

[19] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2401–2412, 2013. 1

[20] Chi Keong Goh, Yanzhu Liu, and Adams Wai Kin Kong. A constrained deep neural network for ordinal regression. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 831–839, 2018. 5

[21] Yu Gong, Greg Mori, and Frederick Tung. Ranksim: Ranking similarity regularization for deep imbalanced regression. *arXiv preprint arXiv:2205.15236*, 2022. 3

[22] Frederick L Greene, Charles M Balch, Irvin D Fleming, April Fritz, Daniel G Haller, Monica Morrow, and David L Page. *AJCC cancer staging handbook: TNM classification of malignant tumors*. Springer Science & Business Media, 2002. 1

[23] Guodong Guo, Yun Fu, Charles R Dyer, and Thomas S Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188, 2008. 2

[24] Pedro Antonio Gutiérrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2015. 1

[25] Martie G Haselton and David M Buss. Error management theory: a new perspective on biases in cross-sex mind reading. *Journal of personality and social psychology*, 78(1):81, 2000. 7

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[27] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3, 6, 8

[28] William A Knaus, Jack E Zimmerman, Douglas P Wagner, Elizabeth A Draper, and Diane E Lawrence. Apache—acute physiology and chronic health evaluation: a physiologically based classification system. *Critical care medicine*, 9(8): 591–597, 1981. 1

[29] Okan Kopuklu, Jiapeng Zheng, Hang Xu, and Gerhard Rigoll. Driver anomaly detection: A dataset and contrastive learning approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 91–100, 2021. 3

[30] Seon-Ho Lee and Chang-Su Kim. Deep repulsive clustering of ordered data based on order-identity decomposition. In *International Conference on Learning Representations*, 2020. 2

[31] Seon-Ho Lee, Nyeong Ho Shin, and Chang-Su Kim. Geometric order learning for rank estimation. In *Advances in Neural Information Processing Systems*, 2022. 3, 6, 7, 8, 2

[32] Gil Levi and Tal Hassncer. Age and gender classification using convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 34–42, 2015. 5

[33] Changsheng Li, Qingshan Liu, Jing Liu, and Hanqing Lu. Learning distance metric regression for facial age estimation. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 2327–2330, 2012. 3

[34] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13896–13905, 2021. 2, 6, 7, 8, 3

[35] Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. Ordinalclip: Learning rank prompts for language-guided ordinal regression. *Advances in Neural Information Processing Systems*, 35:35313–35325, 2022. 1, 2

[36] Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *International Conference on Learning Representations*, 2019. 2

[37] Yanzhu Liu, Adams Wai Kin Kong, and Chi Keong Goh. A constrained deep neural network for ordinal regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 831–839, 2018. 2

[38] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 7

[39] Abdul Sami Mohammed, Ahmed Abul Hasanaath, Ghazanfar Latif, and Abul Bashar. Knee osteoarthritis detection and severity classification using residual neural networks on preprocessed x-ray images. *Diagnostics (Basel)*, 13(8):1380, 2023. 6

[40] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 6, 8

[41] Frank Palermo, James Hays, and Alexei A. Efros. Dating historical color images. In *ECCV (6)*, pages 499–512, 2012. 5

[42] Daniel Petashvili, Hu Wang, Alison Deslandes, Jodie Avery, George Condous, Gustavo Carneiro, Louise Hull, and Hsiang-Ting Chen. Learning subjective image quality assessment for transvaginal ultrasound scans from multi-annotator labels. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 7

[43] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *Data*, 3(3), 2018. 2, 5

[44] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3):e190043, 2020. 8

[45] Daniel Rho, Taesoo Kim, Sooill Park, Jaehyun Park, and Jaehan Park. Understanding contrastive learning through the lens of margins. *arXiv [cs.LG]*, 2023. 3

[46] Sophie Roborgh. From risk aversion to risk management. *The Companion to Peace and Conflict Fieldwork*, pages 115–129, 2021. 7

[47] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4393–4402. PMLR, 2018. 4

[48] Anshul Shah, Suvrit Sra, Rama Chellappa, and Anoop Cherian. Max-margin contrastive learning. *Proc. Conf. AAAI Artif. Intell.*, 36(8):8220–8230, 2022. 3

[49] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18760–18769, 2022. 2, 6, 8

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In

3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. 2

[51] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 3, 4

[52] Harold C Sox, Michael C Higgins, Douglas K Owens, and Gillian Sanders Schmidler. *Medical decision making*. John Wiley & Sons, 2024. 1

[53] Juan Luis Suárez, Salvador García, and Francisco Herrera. Ordinal regression with explainable distance metric learning based on ordered sequences. *Mach. Learn.*, 110(10): 2729–2762, 2021. 3

[54] Graham Teasdale, Andrew Maas, Fiona Lecky, Geoffrey Manley, Nino Stocchetti, and Gordon Murray. The glasgow coma scale at 40 years: standing the test of time. *The Lancet Neurology*, 13(8):844–854, 2014. 1

[55] Nikos Tsiknakis, Dimitris Theodoropoulos, Georgios Manikis, Emmanouil Ktistakis, Ourania Boutsora, Alexa Berto, Fabio Scarpa, Alberto Scarpa, Dimitrios I. Fotiadis, and Kostas Marias. Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. *Computers in Biology and Medicine*, 135:104599, 2021. 6

[56] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504. IEEE, 2021. 3

[57] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 943–952, 2021. 3

[58] Rui Wang, Peipei Li, Huaibo Huang, Chunshui Cao, Ran He, and Zhaofeng He. Learning-to-rank meets language: Boosting language-driven ordering alignment for ordinal classification. *Advances in Neural Information Processing Systems*, 36, 2023. 1, 2

[59] Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19376–19385, 2022. 3

[60] Bo Xiao, Xiaokang Yang, Yi Xu, and Hongyuan Zha. Learning distance metric for regression by semidefinite programming with application to human age estimation. In *Proceedings of the 17th ACM International Conference on Multimedia*, page 451–460, New York, NY, USA, 2009. Association for Computing Machinery. 3

[61] Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Zijun Liu, Yanan Wu, Hong Xu, Huixing Jiang, and Weiran Xu. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning. *arXiv preprint arXiv:2105.14289*, 2021. 3

[62] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 3, 6, 8

[63] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 2