This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Learning Disease State from Noisy Ordinal Disease Progression Labels

Gustav Schmidt[1], Holger Heidrich[2], Philipp Berens[1], and Sarah Müller[1]

[1] Hertie Institute for AI in Brain Health, University of Tübingen, Germany
[2] Department of Computer Science, University of Tübingen, Germany
sar.mueller@uni-tuebingen.de

Abstract. Learning from noisy ordinal labels is a key challenge in medical imaging. In this work, we ask whether ordinal disease progression labels (*better*, *worse*, or *stable*) can be used to learn a representation allowing to classify disease state. For neovascular age-related macular degeneration (nAMD), we cast the problem of modeling disease progression between medical visits as a classification task with ordinal ranks. To enhance generalization, we tailor our model to the problem setting by (1) independent image encoding, (2) antisymmetric logit space equivariance, and (3) ordinal scale awareness. In addition, we address label noise by learning an uncertainty estimate for loss re-weighting. Our approach learns an interpretable disease representation enabling strong few-shot performance for the related task of nAMD activity classification from single images, despite being trained only on image pairs with ordinal disease progression labels[1].

Keywords: Few-Shot Learning · Ordinal Labels · Label Noise · Age Related Macular Degeneration · Optical Coherence Tomography.

1 Introduction

Changes apparent in medical images can be informative about the progression of a disease, playing a critical role in guiding clinical decision making, particularly for conditions requiring timely interventions. One such example is the treatment of neovascular age-related macular degeneration (nAMD) with anti-vascular endothelial growth factor (anti-VEGF) therapy. Here, the treatment is guided by the presence and extent of exudative signs, such as intraretinal and subretinal fluid as relevant biomarkers [19]. These are best assessed with optical coherence tomography (OCT) imaging. Accurate prediction of disease progression in this context could help to optimize treatment schedules and improve patient outcomes. Different deep learning approaches have been proposed to analyze AMD based on individual OCT B-scans, including disease activity classification [1, 13], biomarker identification [11, 8], and disease progression modeling [18, 4].

Interestingly, judging disease progression between OCT B-Scan pairs is known to be easier and less biased for clinicians than assigning categorical severity scores

[1] https://github.com/berenslab/Learning-Disease-State

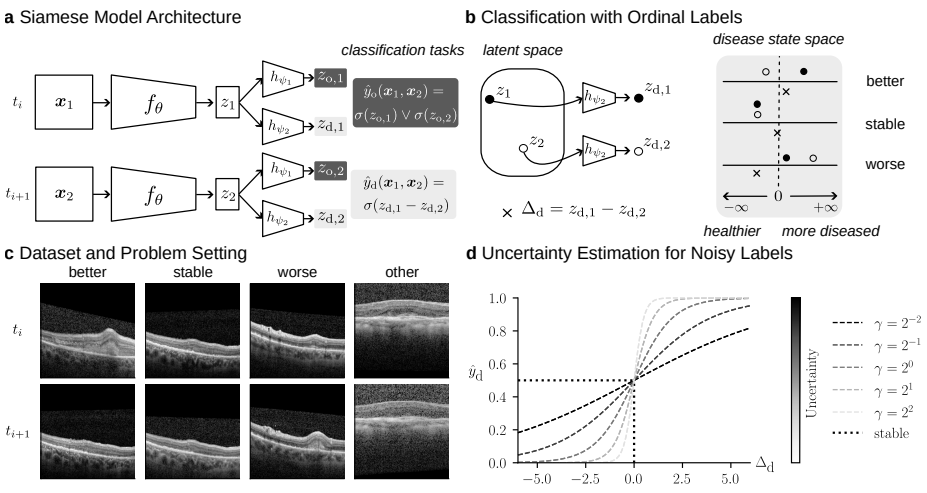

Fig. 1. Overview figure of the proposed approach. (a) Our Siamese model analyses two OCT B-scans (c), outputs the probability of being ungradable ($\hat{y}_\mathrm{o}$), and predicts the disease progression ($\hat{y}_\mathrm{d}$). (b) For disease progression prediction our model internally learns a disease state space (z_d) for each image. (d) Moreover, we fit an uncertainty estimate for each image pair to account for label noise. $\sigma(\cdot)$ is the sigmoid function.

to individual B-Scans used in the studies cited above, also because it is closer to the clinical task performed in a routine assessment [12]. Here, we ask whether we can use such coarse, ordinal information about whether nAMD has improved, worsened, or remained stable between two visits (Fig. 1 **c**) to learn about the underlying continuous disease state. While ordinal regression has been explored in machine learning [3, 21, 7] and medical imaging [2, 22, 6, 16, 10], conventional methods neither preserve the continuity of the label space by discretizing K labels into K-1 binary tasks nor do they take noisy labels into account. Here, we introduce a method that is able to learn a continuous and interpretable disease state from noisy ordinal disease progression labels. Importantly, unlike [14], where they require Euclidean distances to healthy anchors, our model learns a continuous disease state directly, without such constraints.

To this end, we used the MARIO challenge dataset [23] from MICCAI 2024, which consists of labeled image pairs indicating the disease progression between two patient visits. We frame disease progression between medical visits as a classification task with ordinal ranks (Fig. 1 **a,b**). To enhance generalization and interpretability, we propose the following model design choices:

1. **Independent image encoding**: We encode a scalar disease state for each image independently (Fig. 1 **a**) to ensure that the model, although trained on labeled image pairs, retains a meaningful disease state representation on image level (Fig. 1 **b**).
2. **Antisymmetric logit space**: Since disease progression labels capture differences between image pairs, we directly model progression as the difference

between two disease states Fig. 1 **b**). This enforces an antisymmetric equivariant logit space – derivable from a property of the sigmoid function.
3. **Ordinal scale awareness**: Unlike conventional ordinal regression, our method considers both order and known distances between labels. This results in a continuous, structured and interpretable representation of disease progression, capturing intraclass and interclass relationships.
4. **Uncertainty-aware loss re-weighting**: We mitigate the impact of label noise with an uncertainty estimation by a learnable slope parameter for the sigmoid function (Fig. 1 **d**) for loss re-weighting [20]. This accounts for different sources of label noise, allowing the model to better capture inherent ambiguities in clinical grading.

We then show that the learned representation leads to strong few-shot out-of-distribution performance on an in-house OCT dataset labeled for nAMD activity.

2 Methods

2.1 Dataset and Preprocessing

For training and evaluation, we used the MARIO challenge data [23], specifically the development set with 14,496 labeled OCT B-scan pairs from 68 patients, each up to 10 visits per patient. All OCT volumes were acquired with Heidelberg Spectralis and volumes were registered between consecutive visits using the Spectralis follow-up option. To standardize the input dimension, we padded all images to match the largest resolution occurring in the dataset (496×1024). Furthermore, we applied training-time data augmentation, ensuring that the same augmentations were applied for both images out of a pair. Augmentations included random resize cropping between 20%-100% of the original size, resized to 224×385, and random horizontal flipping. We split our dataset patient-wise in 85% for training (5-fold cross-validation) and 15% for testing.

2.2 Problem Setting

The B-scan image pairs were annotated by ophthalmologists into seven initial classes, which were later simplified into three disease progression classes *better*, *worse*, *stable*, and one *other* category for ungradable image pairs (Fig. 1 **c**). Therefore, the challenge framed the task as a 4-class classification problem, where the disease progression labels are on an ordinal scale with symmetric distances – the *stable* class is between *better* and *worse*.

2.3 Obtaining Disease State from Coarse Progression Labels

We used a Siamese neural network to process labeled B-scan pairs (Fig. 1 **a**). However, instead of training an off-the-shelf Siamese network on the 4-class task using the concatenation of high-dimensional latent representations from each branch, like in [4, 16], we tailored our model to the problem setting to enhance

interpretability and generalization. Because, the *other* class is a very different category than the disease progression categories, we separated the disease information and the *other* class with two independent heads ($\boldsymbol{z}_\mathrm{d}$ and $\boldsymbol{z}_\mathrm{o}$ in Fig. 1 **b**) and treated both heads as independent classification tasks. Moreover, we forced every branch to output a scalar and hence a more interpretable output on image-level before computing information on image-pair level for the classification tasks where we had labels. This has the advantage that our model predicts a disease state z_d for each B-scan individually, even though we only train it on labeled B-scan pairs. Then, we reasoned that disease progression prediction is mainly about the difference between the image pairs. In theory, we could subtract registered B-scan pairs from each other and operate on difference images, however, to be more robust to possible registration errors and other noise differences between visits, we computed the differences in the logit (unnormalized log probability) space. Therefore, we predicted the disease progression with

$$\hat{y}_\mathrm{d}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sigma(\Delta_\mathrm{d}) = \frac{1}{1 + e^{-\Delta_\mathrm{d}}},\ \Delta_\mathrm{d} = z_{\mathrm{d},1} - z_{\mathrm{d},2}. \tag{1}$$

Estimating disease progression is akin to a binary classification task with the two classes *worse* $\hat{=}$ 0 and *better* $\hat{=}$ 1. However, our setup extended binary classification with a third label *stable* $\hat{=}$ 0.5. By setting the labels in this way, they inherently follow the order and known distances between our labels by positioning the *stable* class between *better* and *worse* (Fig. 1 **d**). However, unlike conventional ordinal regression methods that discretize K labels into K-1 binary tasks, our approach preserved continuity of the disease progression space, highlighting intraclass and interclass relationships. This extension also leads to an antisymmetric equivariance with respect to the image order, so that $\hat{y}_\mathrm{d}(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1 - \hat{y}_\mathrm{d}(\boldsymbol{x}_2, \boldsymbol{x}_1)$, which assumes that matching pairs in both time directions (forward and backward) helps learning about disease progression. Even if this inductive bias is not strictly true, as e.g. fluids in the retina may leave lasting traces, we empirically found that the resulting model can be used to detect the presence of biomarkers like intra- and subretinal fluids.

For the *other* binary classification task, we merged the two *other* head predictions for the individual images with a logical OR operation – if at least one image of the pair is ungradable, e.g., due to noise, then the model should already predict the *other* class. We applied De Morgan's laws to reformulate the *other* prediction to the differential computation

$$\begin{aligned} \hat{y}_\mathrm{o}(\boldsymbol{x}_1, \boldsymbol{x}_2) &= p((\boldsymbol{x}_1 \text{ is } \textit{other}) \vee (\boldsymbol{x}_2 \text{ is } \textit{other})) && (2) \\ &= 1 - (1 - \sigma(z_{o,1})) \cdot (1 - \sigma(z_{o,2})) = 1 - \sigma(-z_{o,1}) \cdot \sigma(-z_{o,2}) && (3) \end{aligned}$$

where $\sigma(x)$ is the sigmoid function with the property $1 - \sigma(x) = \sigma(-x)$.

2.4 Modeling Label Noise

To model label noise and account for uncertain examples, we included a learnable slope parameter γ (Fig. 1 **d**) into the disease progression tasks sigmoid function

$$\hat{y}_{\mathrm{d}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sigma(\Delta_{\mathrm{d}}, \gamma) = \frac{1}{1 + e^{-\gamma \cdot \Delta_{\mathrm{d}}}} \tag{4}$$

to give our model the possibility to set this parameter for every B-scan pair at training time. Intuitively, we interpreted the γ values as an uncertainty estimate for each image pair, where lower and higher values than $\gamma = 1$ referred to higher and lower uncertainty, respectively (Fig. 1 **d**). However, γ could also be misused as a shortcut by the model for "hard" to classify image pairs by setting γ very low and hence achieve lower loss on those samples. Therefore, we introduced a regularizer, which regularizes γ to be close to its default value $\gamma = 1$. We defined $\gamma = 2^{\alpha}$ and add $|\alpha|$ as a regularizer to the final optimization problem

$$\theta^*, \psi_1^*, \psi_2^*, \alpha^* = \underset{\theta, \psi_1, \psi_2, \alpha}{\arg\min}\, \mathrm{BCE}(y_{\mathrm{o}}, \hat{y}_{\mathrm{o}}) + \mathrm{BCE}(y_{\mathrm{d}}, \hat{y}_{\mathrm{d}}) + \lambda|\alpha| \tag{5}$$

with BCE as the binary cross entropy loss. We balanced the dataset with a weighted random sampler and trained all models with 5-fold cross-validation, a ResNet50 [9] backbone and AdamW [15] optimizer ($\mathrm{lr} = 10^{-4}$) for 60 epochs and selected the model by the best validation loss. For our models with noise estimation, we set $\lambda = 0.15$, selected by grid search.

3 Results

3.1 Disease Progression Classification

We used the MARIO OCT B-scan dataset [23] to train our architecture suitable for handling ordinal disease progression labels (Fig. 1) and first evaluated it for disease progression classification on the metrics of the MARIO challenge [17]. We compared its performance to a naïve classifier (clf.) trained for a 4-class problem with categorical cross-entropy. Here, the naïve classifier was a Siamese model which concatenates high-dimensional image embeddings from each branch and processes them through learnable layers to the final classification output, without producing interpretable scalar per-image outputs. For our model, to assign each point in $\hat{y}_{\mathrm{d}}$ to a class, we optimized a symmetric decision boundary around 0.5

Table 1. Classification performance (in %) for disease progression.

Model	F1 Score	Rk-corr.	Specificity	Bal. Acc.	Precision	Recall
naïve clf.	70 ± 5	44 ± 4	87 ± 1	60 ± 3	57 ± 3	60 ± 3
ours	61 ± 7	36 ± 5	86 ± 1	59 ± 5	47 ± 3	59 ± 5
ours + noise estim.	60 ± 7	36 ± 6	86 ± 2	60 ± 4	47 ± 4	60 ± 4

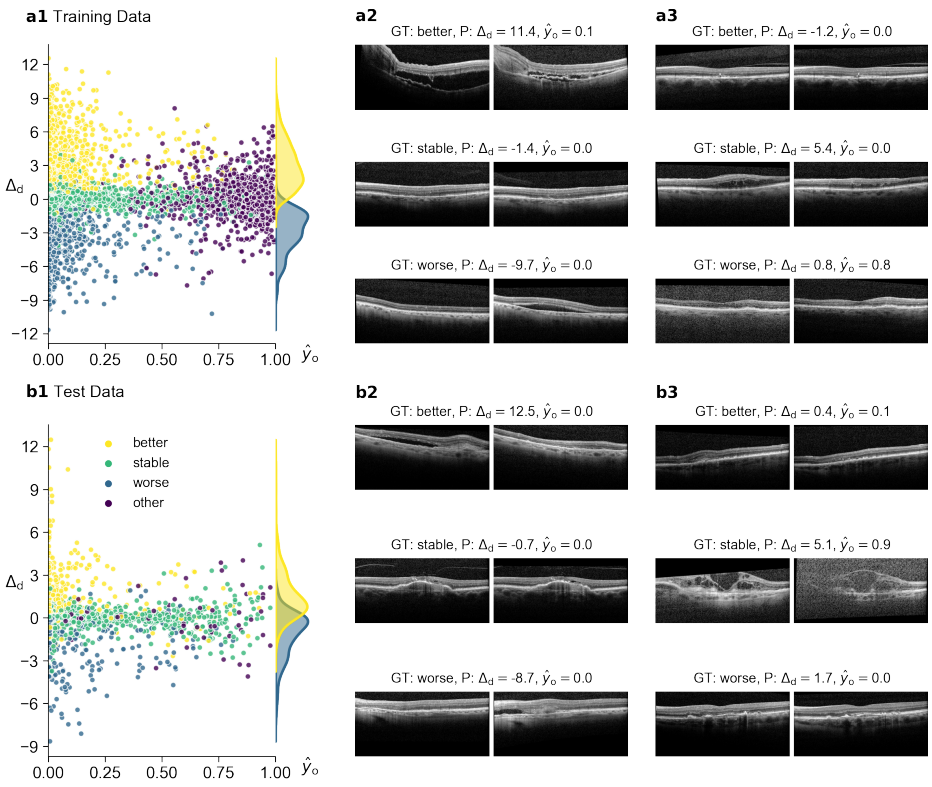

Fig. 2. Our model maps coarse ordinal labels to a continuous variable Δ_d that preserves ordinal ranks. (a1) Δ_d (averaged over folds) vs. the probability of belonging to the *other* class for training data. (a2) Examples with low probability of *other* illustrating the graded nature of disease progression judgments, with the magnitude of Δ_d corresponding to the magnitude of observed changes. (a3) Examples with "incorrect" predictions compared to the ground truth labels. Top: Labeled as *better*, but looks stable, as indicated by Δ_d. Middle: Labeled as *stable*, but looks improved, reflected in Δ_d. Bottom: Labeled *worse*, but looks stable and noisy captured by $\hat{y}_\mathrm{o}$. (b1-b3) as in (a1-a3) but for test data. (GT: ground truth, P: prediction).

based on the validation data. Our model performed comparable to the the naïve classifier (Table 1) for specificity, balanced accuracy and recall. However, in other metrics like the F1 score, our model showed reduced performance, mainly due to the performance for the *stable* class. While the focus of this work was not on optimizing performance for the disease progression classification task, our model would have been placed 17/21 on the MARIO leaderboard, when evaluated on 15% of the training data in the cross-validation setting (as we did not have access to the MARIO validation set which was used to rank the participants).

However, compared to the naïve classifier, our model internally learned a continuous disease progression representation Δ_d. We qualitatively analyzed what

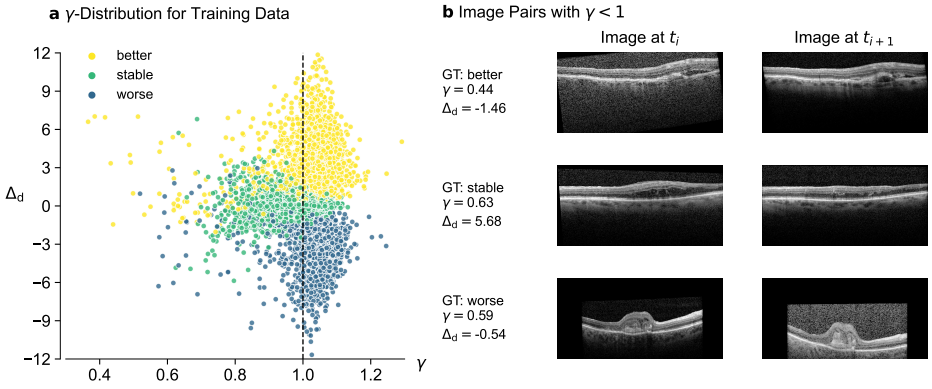

Fig. 3. Modeling label noise via uncertainty parameter. (a) Lower uncertainty parameter γ (averaged over folds) is observed for images with $\Delta_{\mathrm{d}} \approx 0$, indicating higher uncertainty. (b) Example images show that γ reflects various noise sources: acquisition noise, mislabeling and registration errors (top-bottom). (GT: ground truth).

additional information this representation could provide both for the training and test data (Fig. 2 **a1**,**b1**). When selecting image pairs with large values in Δ_{d}, large changes between the images were visible, indicating that the representation accurately reflected disease progression information (Fig. 2 **a2**,**b2**). Additionally, we probed the representation to understand why some image pairs were incorrectly predicted or had low confidence predictions (Fig. 2 **a3**,**b3**). We found that in many cases, the ground truth labels did not match well what was visible on the images, as the images did not show clear evidence for the labeled class but rather for one of the others, indicating that many ground truth labels were noisy. Also, the *other* class prediction helped to detect corrupted images with high success.

3.2 Learning an Uncertainty Parameter for Label Noise

Our observation that ground truth labels were unreliable and noisy in many cases was corroborated by the fact that adjacent B-scans in OCT volumes sometimes received different labels, despite showing similar structural patterns. Therefore, we extended our model with a mechanism to discount noisy labels during training, learning an uncertainty parameter γ for each B-scan pair (Fig. 3 **a**). In fact, the learned γ was smaller for image pairs close to the decision boundaries. We found that image pairs with low γ values corresponded to noise from acquisition, mislabeling or failed registration (Fig. 3 **b**, top to bottom). To verify this qualitative finding, we analyzed the γ values of examples adjacent in the OCT volume but labeled differently. We found that 24% of examples labeled *better* next to *worse* had γ values below 0.85, compared to 15% for both *better* to *stable* and *worse* to *stable* transitions, while only around 8% for cases remaining in the same state (e.g. *better* to *better*).

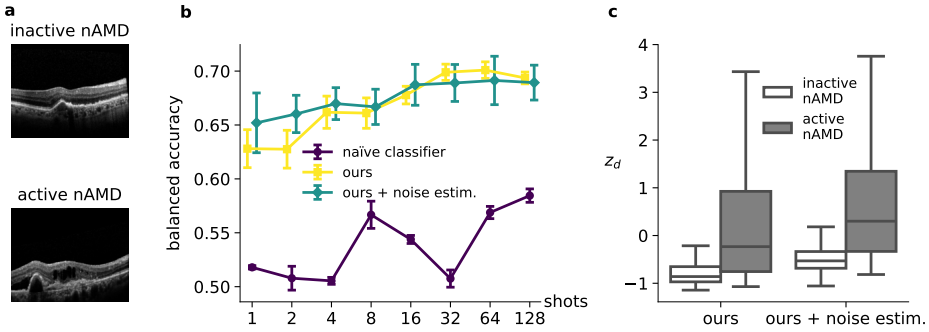

Fig. 4. Our model allows accurate out-of-distribution few-shot nAMD activity classification. (a) Example OCT B-scans from the in-house dataset. (b) Balanced accuracy as a function of the number of single B-scans per class for our method (yellow), our method with noise estimation (green) and a naïve classifier (purple). (c) Inactive vs. active nAMD B-scans lead to distributions in the z_d space (here shown for one fold).

3.3 Out-of-Distribution Few-Shot Disease Activity Classification

Because our model was able to extract a meaningful continuous disease progression representation from the coarse ordinal labels, we hypothesized that the representation learned by z_d could be used for performing a related task on an out-of-distribution (OOD) dataset. The in-house dataset consists of 2,886 B-scans of 50 patients acquired with Heidelberg Spectralis (Fig. 4 **a**) – 878 with nAMD activity and 2,008 without. The dataset is not publicly available but was first used in [1] and approved by the local institutional ethics committee. For evaluation, we mapped the B-scans from 27 patients into the z_d space. For training, we used the remaining 23 patients to optimize the decision boundary between inactive and active nAMD patients for the z_d logits on a small set of images (we call k-shots, where k is the number of B-scans per class) without retuning the representation in any way. Compared to a logistic regression classifier on the embeddings of the naïve model, our models showed better nAMD activity classification performance, even when the decision boundary was determined using very little data (Fig. 4 **b**). Furthermore, the model with noise estimation proved to be the most robust model in terms of balanced accuracy, especially for very few shots (Fig. 4 **b**,**c**).

4 Discussion

We introduced a novel approach to learn from coarse ordinal disease progression labels (*better*, *stable* or *worse*) by tailoring a deep learning model to the task, showing strong few-shot generalization performance on an in-house OOD OCT dataset on a related, but not identical task of nAMD activity classification. Our model provided enhanced interpretability by learning a continuous disease

progression space internally, compared to models directly targeting the classification task. Additionally, by explicitly modeling the noise in the ordinal labels, we were able to explore various sources of label noise and show more robust OOD performance. In future work, an enhanced backbone architecture could be used to improve disease progression classification on the MARIO challenge data. Furthermore, we observed that the *stable* class and our noise model appear to interfere, leading the model to assign lower gamma values to this class. Therefore, exploring alternative approaches for modeling label noise (e.g., [5]) could offer a promising direction for future work. Finally, the disease state space currently learned by the model is restricted to one-dimension – an insightful ablation experiment would be to investigate how the performance would change if we allow a higher-dimensional latent representation.

Acknowledgments. We thank Sebastian Damrich, Verena Jasmin Hallitschke, and Julius Gervelmeyer for helpful discussions and their valuable feedback. This project was supported by the Hertie Foundation and by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy with the Excellence Cluster 2064 "Machine Learning — New Perspectives for Science", project number 390727645. PB is a member of the Else Kröner Medical Scientist Kolleg "ClinbrAIn: Artificial Intelligence for Clinical Brain Research". The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting SM.

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

References

1. Ayhan, M.S., Faber, H., Kühlewein, L., Inhoffen, W., Aliyeva, G., Ziemssen, F., Berens, P.: Multitask Learning for Activity Detection in Neovascular Age-Related Macular Degeneration. Translational Vision Science & Technology **12**(4), 12–12 (04 2023). https://doi.org/10.1167/tvst.12.4.12
2. Baek, S., Sim, J., Wu, G., Kim, W.H.: OCL: Ordinal Contrastive Learning for Imputating Features with Progressive Labels. In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15002. Springer Nature Switzerland (October 2024)
3. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters **140**, 325–331 (2020). https://doi.org/https://doi.org/10.1016/j.patrec.2020.11.008
4. Emre, T., Chakravarty, A., Rivail, A., Riedl, S., Schmidt-Erfurth, U., Bogunović, H.: TINC: Temporally Informed Non-contrastive Learning for Disease Progression Modeling in Retinal OCT Volumes, p. 625–634. Springer Nature Switzerland (2022). https://doi.org/10.1007/978-3-031-16434-7_60
5. Englesson, E., Azizpour, H.: Robust Classification via Regression for Learning with Noisy Labels. In: The Twelfth International Conference on Learning Representations (2024), `https://openreview.net/forum?id=wfgZc3IMqo`
6. Gao, Z., Zhao, H., Wu, Z., Wang, Y., Lip, G.Y.H., Shantsila, A., Shantsila, E., Zheng, Y.: Coral-CVDs: A Consistent Ordinal Regression Model for Cardiovascular Diseases Grading. In: Ophthalmic Medical Image Analysis. Springer Nature Switzerland (2025)

7. Garg, B., Manwani, N.: Robust Deep Ordinal Regression Under Label Noise. In: Asian conference on machine learning. pp. 782–796. PMLR (2020)
8. Hanson, R.L., Airody, A., Sivaprasad, S., Gale, R.P.: Optical coherence tomography imaging biomarkers associated with neovascular age-related macular degeneration: a systematic review. Eye **37**(12), 2438–2453 (2023)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (2015)
10. Hoebel, K.V., Lemay, A., Campbell, J.P., Ostmo, S., Chiang, M.F., Bridge, C.P., Li, M.D., Singh, P., Coyner, A.S., Kalpathy-Cramer, J.: A generalized framework to predict continuous scores from medical ordinal labels (2023)
11. Holland, R., Kaye, R., Hagag, A.M., Leingang, O., Taylor, T.R., Bogunović, H., Schmidt-Erfurth, U., Scholl, H.P., Rueckert, D., Lotery, A.J., Sivaprasad, S., Menten, M.J.: Deep Learning-Based Clustering of OCT Images for Biomarker Discovery in Age-Related Macular Degeneration (PINNACLE Study Report 4). Ophthalmology Science **4**(6), 100543 (2024). https://doi.org/https://doi.org/10.1016/j.xops.2024.100543
12. Kalpathy-Cramer, J., Campbell, J.P., Erdogmus, D., Tian, P., Kedarisetti, D., Moleta, C., Reynolds, J.D., Hutcheson, K., Shapiro, M.J., Repka, M.X., et al.: Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. Ophthalmology **123**(11), 2345–2351 (2016)
13. Koseoglu, N.D., Grzybowski, A., Liu, T.Y.A.: Deep Learning Applications to Classification and Detection of Age-Related Macular Degeneration on Optical Coherence Tomography Imaging: A Review. Ophthalmology and Therapy **12**, 2347 – 2359 (2023)
14. Li, M.D., Chang, K., Bearce, B., Chang, C.Y., Huang, A.J., Campbell, J.P., Brown, J.M., Singh, P., Hoebel, K.V., Erdoğmuş, D., et al.: Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. NPJ digital medicine **3**(1), 48 (2020)
15. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (2019)
16. Polat, G., Çağlar, Ü.M., Temizel, A.: Class distance weighted cross entropy loss for classification of disease severity. Expert Systems with Applications **269**, 126372 (2025). https://doi.org/https://doi.org/10.1016/j.eswa.2024.126372
17. Quellec, G., El Habib Daho, M., Zeghlache, R. (eds.): Image-based prediction of retinal disease progression. Lecture notes in computer science, Springer International Publishing, Cham, Switzerland (Apr 2025)
18. Rivail, A., Schmidt-Erfurth, U., Vogl, W.D., Waldstein, S.M., Riedl, S., Grechenig, C., Wu, Z., Bogunović, H.: Modeling Disease Progression in Retinal OCTs with Longitudinal Self-supervised Learning (2019). https://doi.org/10.1007/978-3-030-32281-6_5
19. Schmidt-Erfurth, U., Waldstein, S.M.: A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. Progress in Retinal and Eye Research **50**, 1–24 (2016)
20. Shi, J., Zhang, K., Guo, C., Yang, Y., Xu, Y., Wu, J.: A survey of label-noise deep learning for medical image analysis. Medical Image Analysis **95** (2024). https://doi.org/https://doi.org/10.1016/j.media.2024.103166
21. Shi, X., Cao, W., Raschka, S.: Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. Pattern Analysis and Applications **26**(3), 941–955 (2023)

22. Tang, W., Yang, Z., Song, Y.: Disease-grading networks with ordinal regularization for medical imaging. Neurocomputing **545**, 126245 (2023). https://doi.org/https://doi.org/10.1016/j.neucom.2023.126245
23. Youven Zeghlache, R.: MARIO: Monitoring Age-related Macular Degeneration Progression in Optical Coherence Tomgraphy. `https://youvenz.github.io/MARIO_challenge.github.io/` (2024). https://doi.org/10.5281/zenodo.10992295