Classification and Prediction of Age-Related Macular Degeneration Progression Using OCT Images and Multiple Instance Learning

Alberto J. Beltrán-Carrero[1(✉)], Javier Torresano-Rodríguez[2], Esther Santos-Vicente[2], María J. Aparicio Hernández-Lastras[2], Álvaro Caballero-Sastre[1], María J. Ledesma-Carbayo[1,3], and Juan J. Gómez-Valverde[1,3]

[1] Biomedical Image Technologies (BIT), ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain
aj.beltran.carrero@upm.es

[2] Ophthalmology Service of the Provincial Ophthalmic Institute, Hospital Universitario Gregorio Marañón, Madrid, Spain

[3] Centro de Investigación Biomédica en Red de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Madrid, Spain

Abstract. The study introduces a novel approach for classifying and predicting the progression of Age-related Macular Degeneration (AMD) using Optical Coherence Tomography (OCT) images and Multiple Instance Learning (MIL). AMD is a leading cause of vision impairment worldwide, making effective monitoring and treatment essential, particularly with anti-VEGF therapy. However, the increasing number of patients and the frequency of follow-up visits pose challenges for healthcare systems. This approach addresses two key tasks: (1) classifying changes between consecutive 2D OCT B-scans and (2) predicting disease progression within a 3-month period. For task 1, the model incorporates contextual information from adjacent B-scans and applies bidirectional cross-attention to learn time-dependent features. For task 2, a MIL-based architecture is used to identify the most significant slices within an OCT volume. The results demonstrated the effectiveness of the proposed methods. In task 1, the model achieved a mean score of 0.7488 across all evaluation metrics. For task 2, the mean score was 0.4478, reflecting the complexity of disease progression prediction. This approach offers improvements over baseline models and contributes to the development automated tools for AMD management, potentially easing the burden on ophthalmology services and improving personalized patient care.

Keywords: AMD · OCT · Anti-VEFG · MIL

G. Quellec et al. (Eds.): DIAMOND 2024/MARIO 2024, LNCS 15503, pp. 201–210, 2025.
https://doi.org/10.1007/978-3-031-86651-7_20

1 Introduction

Age-related macular degeneration (AMD) is a chronic, progressive retinal disease that affects the macula, the area of the retina responsible for the sharpest vision. As the leading cause of irreversible vision loss in the elderly population in developed countries, AMD is a significant public health concern, currently impacting approximately 196 million people worldwide [1]. In recent years, the introduction of anti-vascular endothelial growth factor (anti-VEGF) therapy for neovascular AMD has marked a major advancement in treatment, significantly improving outcomes [2]. However, the growing number of patients requiring treatment and frequent follow-ups has placed considerable strain on healthcare resources, resulting in substantial costs and the risk of overburdening ophthalmology services [3].

In this context, developing comprehensive solutions to optimize patient management workflows is crucial. By integrating clinical and imaging biomarkers, the diagnosis, treatment, and follow-up of AMD patients could become significantly more efficient [4,5]. A paradigm shift could be achieved through the application of deep learning algorithms, which can predict the need for anti-VEGF treatment and assist in monitoring disease progression as well as evaluating therapeutic efficacy [6]. The implementation of these technologies would not only reduce the burden on healthcare systems by enhancing resource allocation and reducing follow-up intervals, but also provide more personalized and timely interventions for patients [7].

Several studies have explored the application of machine learning and deep learning techniques to assist in the monitoring of AMD, most of these efforts focus on predicting the progression from early or intermediate non-exudative stages to advanced exudative stages [8,9], as well as forecasting treatment requirements [10–12]. Despite previous efforts, there remains an unmet need for accurate prediction of AMD progression in patients closely monitored on anti-VEGF treatment plans. The Monitoring Age-related Macular Degeneration Progression in Optical Coherence Tomography (MARIO) challenge at MICCAI 2024 aims to address this gap by evaluating both existing and novel algorithms for detecting progression of neovascular activity in OCT scans of patients with exudative AMD, with the ultimate goal of improving treatment planning.

The challenge is divided into two tasks. Task 1 focuses on pairs of 2D B-scans from two consecutive OCT scans, with the goal of classifying the changes between these slices, which are typically compared side by side by clinicians. Task 2 shifts the focus to predicting future disease progression within a 3-month period for patients undergoing anti-VEGF treatment based on 2D layers.

This work presents a novel approach that addresses both tasks, using two architectures inspired by Multiple Instance Learning (MIL) to effectively exploit the contextual information of successive B-scans.

2 Methods

2.1 Task 1: Classify Evolution Between Two Pairs of 2-D Slices from Two Consecutive 2D OCT Acquisitions

Dataset. Data for task 1 consisted in cases of 2 OCT B-scans and 2 fundus images associated with 2 consecutive visits. Along with image information, patient identifier, side eye, sex, age, visit number and the days between the two visits are also provided. Labels are given at B-scan level, that is, B-scans with the same fundus image associated (from the same acquisition) can have different labels. In this task, the following types of change in activity were defined: reduced (0), stable (1), worsened (2) and other (3).

Pre-processing. Original data samples consisted of a pair of individual OCT B-scans, the associated fundus, the clinical variables and the label assigned to that pair of B-scans. For the purposes of our method, we have only considered the B-scan pairs from t_i and t_{i+1}. First of all, we created the entire OCT volumes of each exam, by selecting all the B-scans that have associated the same fundus identifier, which is also associated with a single volume acquisition. Then, each sample consisted in a pair of OCT volumes and the labels associated to their B-scans pairs. To achieve this, we ensured that during volume creation all B-scans were correctly ordered, using the B-scan indexes provided in the original dataset. Therefore, we took the provided labels in the same order and created a list of labels, which were used as ground truth for each volume during training. Finally, we computed the mean and standard deviation of the volumes and performed intensity normalization using these values.

Method. We propose a model which aims to leverage the contextual information from adjacent B-scans within an OCT volume and over time. The overall architecture is illustrated in Fig. 1. We hypothesise that having contextual information might be potentially useful for the model, in order to better understand the activity of individual B-scans based on their neighbourhood. First, a batch of N 2-D slices extracted from an OCT volume was processed by a vision transformer image encoder. The initial weights of this encoder were taken from the RETFound OCT foundation model [13], which has shown remarkable performance in several retinal-related disease classification tasks and whose weights are publicly available. During training, we only freeze the patch-embedding layer of the encoder due to GPU memory limitations.

From each B-scan, we extracted the corresponding CLS token from the encoder and we stacked all the vectors from each time step, separately. Both vector matrices were then fed into a bidirectional cross-attention module, which produced a pair of attention matrices. This module was designed to process the information from all B-scans from both time steps together. In this way, we forced the model to learn the dependencies between two groups of consecutive slices over time. The module followed a similar mechanism to the cross-attention module of the transformer decoder [14]. It had three weight matrices for each

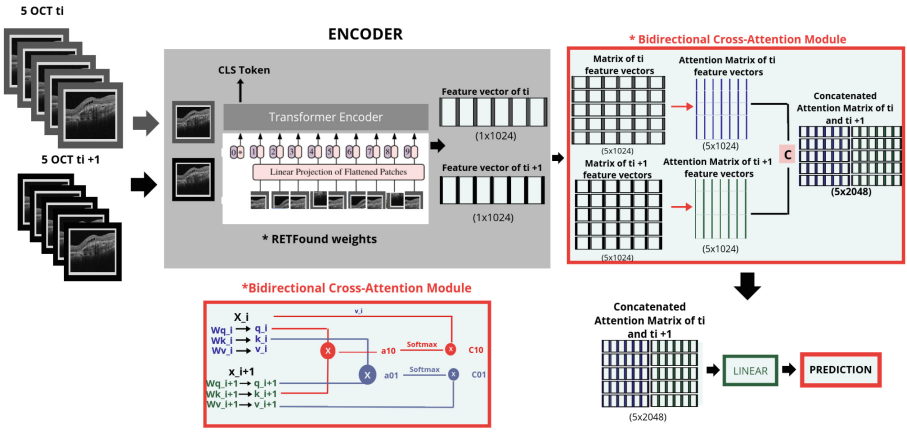

Fig. 1. Proposed architecture for classifying the type of change in activity between two consecutive OCT exams (task 1). The model is designed to leverage the context information from a window of 5 consecutive B-scans and the dependencies between B-scans from different time steps. Red arrows refer to the bidirectional cross-attention operation. (Color figure online)

time step: W_{qi}, W_{ki}, W_{vi}. W_{qi} referred to the matrix that produced the query for time step i, while W_{ki} and W_{vi} produced the keys and values, respectively. Then, two context matrices were computed according to Eq. 1. Since these matrices contain cross-time dependencies for each B-scan, they were concatenated along the feature axis and, finally, the resulting matrix was passed through a linear layer, which performed the classification for each individual B-scan. It is worth noting that the linear layer generated a prediction for each row of the matrix, i.e. for each B-scan, since we cannot assume that all slices from a close neighbourhood would have the same label.

$$C_{i+1,i} = Softmax\left(\frac{Q_i K_{i+1}^T}{\sqrt{d_k}}\right) V_i, \quad C_{i,i+1} = Softmax\left(\frac{Q_{i+1} K_i^T}{\sqrt{d_k}}\right) V_{i+1} \quad (1)$$

Given that the model is trained with random batches of N consecutive B-scans and that not all volumes have the same length, during inference time we had to implement a method to generate the predictions in batches of N B-scans. We simply divided the volumes into groups of N slices and, for those cases where the volume length was not divisible by N, the remaining slices were processed together at the end. This means that for some volumes, the predictions for the last B-scans were computed using a context window smaller than N.

2.2 Task 2: Prediction of Evolution Within 3 Months of AMD on OCT 2D Slices for Planning Treatment Anti-VEGF

Dataset. Data for task 2 consisted in cases of individual OCT B-scans and the corresponding fundus image, both associated with a single visit. Along with

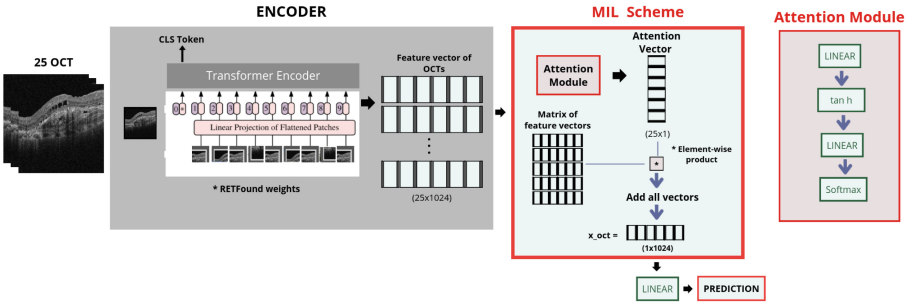

Fig. 2. Proposed architecture for predicting the type of change in activity of an OCT exam (task 2). The model is able to leverage the context information of N consecutive B-scans by following a Multiple Instance Learning (MIL) scheme.

image information, patient identifier, side eye, sex, age and visit number were also provided. For this task, while labels were given at the B-scan level, B-scans with the same fundus image associated (from the same exam) had the same label, i.e. we assumed that we had exam-level labels. In this task, the following types of change activity were defined: reduced (0), stable (1) and increased (2).

Pre-processing. The original samples consisted of a single B-scan, the associated fundus, the clinical variables and the assigned label. For the purposes of our method, we have only considered the OCT B-scans. We needed to create OCT volumes of the same size, by taking all B-scans with the same associated fundus identifier (unique for each volume acquisition). However, we noticed that the number of B-scans per volume varied across the dataset. While most of them were similar in size (19 and 25 B-scans/volume), there were volumes with smaller and larger size. To solve this issue without discarding samples, we decided to create the volumes with a constant size of 25 B-scans/volume. First, we selected the middle B-scan of the entire acquisition and we took the previous and consecutive 12 B-scans. For volumes with less than 25 B-scans, we repeated the extreme slices to complete the volume. In this way, we were able to create a dataset with volumes of the same size. Regarding the labels, we noticed that all the B-scans from the same volume were assigned the same label, thus, we can use it as a label for the whole volume. Finally, as in task 1, we computed the mean and standard deviation of all volumes, in order to perform intensity normalization.

Method. We propose a model architecture that integrates the information from N OCT B-scans from the same volume to produce a global label for all B-scans included. The architecture is illustrated in Fig. 2. First of all, a batch of N consecutive B-scans extracted from a volume was fed to the image encoder. As in task 1, the encoder consisted in a vision transformer whose initial weights were taken from the RETFound OCT model [13] and, this time, no layers were freeze

during training, so the encoder is completely fine-tuned. From each B-scan, a feature vector was extracted and all of them were then stacked. Then, the matrix containing all vectors from the OCT volume were passed through a MIL-oriented attention module. As shown in Fig. 2, this module followed the structure of a MLP with a *tanh* activation function, with the addition of a *Softmax* as final activation. The objective of this part of the model was to identify the B-scans with highest importance regarding the prediction of activity evolution. Since the B-scans of an entire volume were equally labelled, that is, the type of change in activity is associated to the whole exam, we assumed that there were regions of the volume more relevant to the identification of change activity, which aligned better with the nature of the disease activity, given that typical biomarkers, such as retinal fluids, are not uniformly distributed across all the volume. Therefore, the attention module learned to produce a vector that assigns an importance value to each slice. Once this vector is computed, the matrix of feature vectors is multiplied by the attention scores. Subsequently, the matrix is reduced to a single vector by adding of rows, which means adding values on the feature axis. This vector is then fed into the linear layer that produces the prediction. It should be noted that, in this case, the model generated an individual label for the entire OCT volume. Since not all volumes have the same number of slices, this label was extended to the corresponding original number of B-scans of the input volume.

2.3 Training

Training Data. After conducting the aforementioned pre-processing steps, a total of 573 OCT volumes for task 1 and 330 for task 2 comprised the final training dataset. All volumes underwent intensity scaling and normalization before training. Also, they were resized to [N, 224, 224] (N being the number of slices), in order to match the 2-D input size in which the image encoder was pre-trained. Then, we implemented a set of data augmentation operations using MONAI framework [15], which were conducted during model training. These operations included: random flip in all dimensions, random intensity scaling and random Gaussian noise addition. Regarding task 1, we ensured that the same augmentation operations were applied to paired volumes.

Moreover, for task 2, we implemented weighted random sampling during training. Weighted random sampling is a method which aims to balance the contribution of all classes to model training, by assigning the probability of taking a sample from one class based on the proportion of that class in the dataset, resulting in samples from majority classes being taken less often than those from minority classes. In our case, for each sample belonging to class c, its weight value corresponds to $w_c = N/N_c$, where N refers to the total number of samples in the dataset, and N_c refers to the number of samples in the dataset belonging to class c. This method helps the model to achieve better generalization to all classes without discarding samples. Nonetheless, this method can only be employed for task 2, since we have a single label per sample. In task 1, as commented before, a

Table 1. Evaluation metrics for task 1.

Model	F1-Score	Rk-corr. coef.	Specificity	Mean
RETFound [13]	0.7476	0.5113	0.8857	0.7149
Ours ($N = 5$)	**0.7849**	**0.5680**	**0.8936**	**0.7488**

sample of N B-scans taken from a volume can potentially have more than one label, so sample weighting becomes useless.

Optimization and Hardware. All experiments were conducted using the same optimization scheme. We employed the AdamW optimizer with a weight decay of 0.05 and $\beta_1 = 0.9, \beta_2 = 0.95$. We used a cosine learning rate scheduler with an initial learning rate of 0.0001 and 10 warm-up epochs. The total number of epochs was set to 100. We trained all our models in NVIDIA A100 40 GB GPUs provided by Magerit-3 cluster from the Centro de Supercomputación y Visualización de Madrid (CesViMa).

In order to select the best performing model from each experiment, we compute the mean of all evaluation metrics (presented in the following section) at the validation phase, which is conducted after every training epoch. Then, we saved the model with the global best mean metric.

2.4 Evaluation

The evaluation data comprised 7010 cases for task 1 and 3822 cases for task 2. As described before, for task 1 our model performed n-slice inference in order to generate a label for each B-scans pair. For task 2, we performed inference at volume level and the resulting label is assigned to all volume B-scans.

The performance metrics employed for model evaluation included: F1-score, Rk-correlation coefficient, Quadratic-weighted Kappa (for task 2 only) and Specificity. The implementation of these metrics was provided by the organisers.

3 Results

3.1 Task 1

Evaluation metrics and the confusion matrices for the experiments of task 1 are presented in Table 1 and Fig. 3. Our proposed model is compared with the performance of the RETFound model as a reference, which was also our best submission during the development phase. In this approach, we directly trained the vision transformer backbone to predict the label for a pair of individual B-scans.

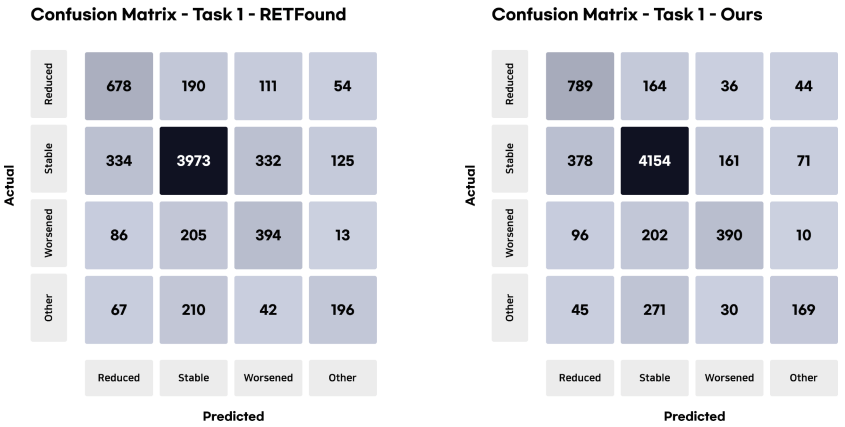

Fig. 3. Confusion matrices of best models for task 1. (Left) RETFound. (Right) Our model.

3.2 Task 2

Evaluation metrics and the confusion matrices for the experiments of task 2 are presented in Table 2 and Fig. 4. We have compared our model with respect to the RETFound and a ResNet 3D architecture. The RETFound was trained to predict the label of individual B-scans. The ResNet architecture is trained to predict the label of entire volumes. We also included the results of our proposed model trained using fixed-size volumes of 25 B-scans and using the complete original volumes.

Table 2. Evaluation metrics for task 2.

Model	F1-Score	Rk-corr. coef.	QW-Kappa	Specificity	Mean
RETFound [13]	0.7015	0.0262	0.0078	0.6721	0.3519
ResNet50-3D	0.7002	0.1330	**0.2422**	0.6925	0.4420
Ours (25 slices)	0.7310	**0.1736**	0.1874	**0.6990**	**0.4478**
Ours (all slices)	**0.7412**	0.1571	0.1520	0.6868	0.4343

4 Discussion

The main contribution of our methods was based on the assumption that the contextual information provided by adjacent B-scans would be useful in training a model for predicting the type of change in activity at the B-scan level. The results we obtained confirmed this assumption for both tasks showing that a model can effectively benefit from processing groups of consecutive B-scans and

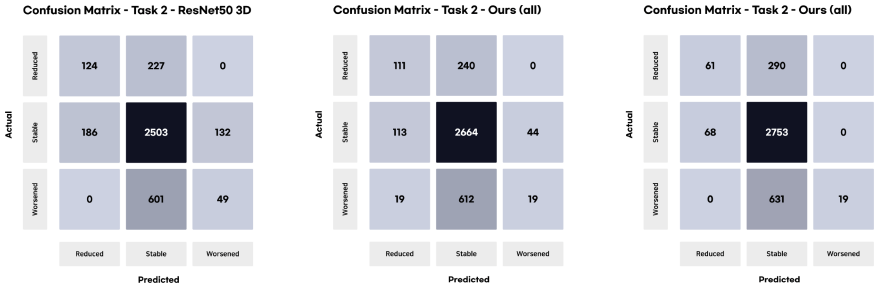

Fig. 4. Confusion matrices of best models for task 2. (Left) ResNet50 3D. (Center) Our model using volumes of 25 slices. (Right) Our model using all slices from original volumes.

improve its performance, compared to using only individual slices. In particular, we noticed such improvement for task 2, where models that leverages context information, by either using neighbouring slices or 3-D convolutions, clearly outperform a model which only uses individual B-scans for training. With respect to task 1, our bidirectional cross-attention module has shown that it is able to integrate information across B-Scans and time instants. Compared to other simpler approaches, for instance, recurrent neural networks, our module offers scalability and computing efficiency, as it is based on the Transformer's attention mechanism.

Our approach does not integrated other information (fundus and clinical variables) which can potentially help to improve the performance. Additionaly, we did not include any synthetic data augmentation method to handle the heavy class imbalance in the datasets, which is likely to provide better model generalization. Future lines of investigation should focus on integrating information from other sources and developing data effective augmentation techniques.

Code Repository: The code for this project is available on GitHub: https://github.com/BIT-UPM/mario_miccai_24_step_amd.

Acknowledgments. The authors acknowledge the support of Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, under grants TED2021-131951B-I00 and PID2022-141493OB-I00 (10.13039/501100011033/MCIN/AEI/ERDF, UE), cofinanced by European Regional Development Fund (ERDF), 'A way of making Europe' and the Next Generation EU funds. AJBC is supported by a FPI grant from the Ministerio de Ciencia e Innovación - PREP2022-000162. The authors gratefully acknowledge the Universidad Politécnica de Madrid (www.upm.es) for providing computing resources on Magerit Supercomputer.

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

References

1. Wong, W.L., et al.: Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob. Health **2**(2), e106–e116 (2014). https://doi.org/10.1016/S2214-109X(13)70145-1
2. Finger, R.P., et al.: Anti-vascular endothelial growth factor in neovascular age-related macular degeneration-a systematic review of the impact of anti-VEGF on patient outcomes and healthcare systems. BMC Ophthalmol. **20**, 1–14 (2020). https://doi.org/10.1186/s12886-020-01554-2
3. Ruiz-Moreno, J.M., Arias, L., Abraldes, M.J., Montero, J., Udaondo, P., The RAMDEBURS Study Group: Economic burden of age-related macular degeneration in routine clinical practice: the RAMDEBURS study. Int. Ophthalmol. **41**(10), 3427–3436 (2021). https://doi.org/10.1007/s10792-021-01906-x
4. Guymer, R., Wu, Z.: Age-related macular degeneration (AMD): more than meets the eye. The role of multimodal imaging in today's management of AMD-a review. Clin. Exp. Ophthalmol. **48**(7), 983–995 (2020). https://doi.org/10.1111/ceo.13837
5. Schmidt-Erfurth, U., Waldstein, S.M.: A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. Prog. Retin. Eye Res. **50**, 1–24 (2016). https://doi.org/10.1016/j.preteyeres.2015.07.007
6. Schmidt-Erfurth, U., et al.: Ai-based monitoring of retinal fluid in disease activity and under therapy. Prog. Retin. Eye Res. **86**, 100972 (2022). https://doi.org/10.1016/j.preteyeres.2021.100972
7. Tan, T.E., Wong, T.Y., Ting, D.: Artificial intelligence for prediction of anti-VEGF treatment burden in retinal diseases: towards precision medicine. Ophthalmol. Retina **5**(7), 601–603 (2021). https://doi.org/10.1016/j.oret.2021.05.001
8. Banerjee, I., et al.: A deep-learning approach for prognosis of age-related macular degeneration disease using SD-OCT imaging biomarkers. arXiv preprint arXiv:1902.10700 (2019). https://doi.org/10.48550/arXiv.1902.10700
9. Liu, Y., et al.: Prediction of OCT images of short-term response to anti-VEGF treatment for neovascular age-related macular degeneration using generative adversarial network. Br. J. Ophthalmol. **104**(12), 1735–1740 (2020). https://doi.org/10.1136/bjophthalmol-2019-315338
10. Bogunović, H., et al.: Prediction of anti-VEGF treatment requirements in neovascular AMD using a machine learning approach. Invest. Ophthalmol. Vis. Sci. **58**(7), 3240–3248 (2017). https://doi.org/10.1167/iovs.16-21053
11. Gallardo, M., et al.: Machine learning can predict anti-VEGF treatment demand in a treat-and-extend regimen for patients with neovascular AMD, DME, and RVO associated macular edema. Ophthalmol. Retina **5**(7), 604–624 (2021). https://doi.org/10.1016/j.oret.2021.05.002
12. Romo-Bucheli, D., Erfurth, U.S., Bogunović, H.: End-to-end deep learning model for predicting treatment requirements in neovascular AMD from longitudinal retinal oct imaging. IEEE J. Biomed. Health Inform. **24**(12), 3456–3465 (2020). https://doi.org/10.1109/JBHI.2020.3000136
13. Zhou, Y., Chia, M.A., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023). https://doi.org/10.1038/s41586-023-06555-x
14. Vaswani, A., et al.: Attention is all you need (2023). https://arxiv.org/abs/1706.03762
15. Jorge Cardoso, M., Li, W., et al.: Monai: an open-source framework for deep learning in healthcare (2022). https://arxiv.org/abs/2211.02701