TRANSFORMER-BASED MULTI-PROTOTYPE APPROACH FOR DIABETIC MACULAR EDEMA ANALYSIS IN OCT IMAGES

Plácido L. Vidal [⋆†] *Joaquim de Moura* [⋆†] *Jorge Novo*[⋆†] *Marcos Ortega*[⋆†] *Jaime S. Cardoso* [‡λ]

[⋆] Centro de Investigación CITIC, Universidade da Coruña, 15071, A Coruña, Spain
[†] Grupo VARPA, Instituto de Investigación Biomédica de A Coruña (INIBIC), 15006, A Coruña, Spain
[‡] Centre for Telecommunications and Multimedia, INESC TEC, 4200-465, Porto, Portugal
[λ] Faculdade de Engenharia, Universidade do Porto, 4099-002, Porto, Portugal

ABSTRACT

Optical Coherence Tomography (OCT) is the major diagnostic tool for the leading cause of blindness in developed countries: Diabetic Macular Edema (DME). Depending on the type of fluid accumulations, different treatments are needed. In particular, Cystoid Macular Edemas (CMEs) represent the most severe scenario, while Diffuse Retinal Thickening (DRT) is an early indicator of the disease but a challenging scenario to detect. While methodologies exist, their explanatory power is limited to the input sample itself. However, due to the complexity of these accumulations, this may not be enough for a clinician to assess the validity of the classification. Thus, in this work, we propose a novel approach based on multi-prototype networks with vision transformers to obtain an example-based explainable classification. Our proposal achieved robust results in two representative OCT devices, with a mean accuracy of 0.9099 ± 0.0083 and 0.8582 ± 0.0126 for CME and DRT-type fluid accumulations, respectively.

***Index Terms*—** Multi-Prototype, Transformers, Optical Coherence Tomography, Diabetic Macular Edema, Explainable Artificial Intelligence

1. INTRODUCTION

Optical Coherence Tomography (OCT) is a imaging technique that generates a cross-sectional representation of the studied tissues (as well as any pathological structures in them) [1]. By means of OCT, we can study the presence of edemas in the retinal layers. These fluid leakages, caused by the degeneration of the delicate vessels in the retina, are amongst the main causes of blindness in developed countries. This degeneration leads to swelling in the central part of the retina (the macula), causing the appearance of Diabetic Macular Edemas (DME). An early diagnosis is critical for this pathology as, if left undiagnosed (or misdiagnosed), it can lead to irreversible damage to the main retinal structures [2].

For this reason, computer-aided diagnosis methodologies try to reduce the influence of the subjectivity of the expert in the monitoring and treatment of this pathology. However, most of these methodologies still represent complex black boxes, meaning their internal logic and inner workings are hidden and even experts cannot fully understand the rationale behind their predictions [3]. In the ophthalmological domain, this decision-making process is critical, as it might severely impact the quality of life of the patient while being based on misleading information. For this reason, methodologies that aim to unveil the underlying processes of these black box methodologies were published. For example, Vidal *et al.* [4, 5] with a proposal based on an exhaustive voting strategy to generate a map of the regions recognised as pathological, Wang *et al.* [6] approached the explainability on the domain by studying the gradient-based class-activation maps, and Reza *et al.* [7] through the LIME framework [8].

However, these explainable methodologies offer a representation only focused on what is already present in the image, not further extending the relation to prior knowledge. To address this relevant issue, and based on the classic learning paradigm of "bag-of-visual-words" Chen *et al.* [9] proposed an architecture that integrates the concept of prototypical parts into the "reasoning" process of the network. This architecture uses a regular convolutional neural network and, progressively, determines a set of prototypes per considered class that represent archetypal aspects of the classified ideal (which demonstrated its performance in other signal processing domains [10]). To perform the classification, the latent feature vector of an in-

This research was funded by Instituto de Salud Carlos III, Government of Spain, PI17/00940 research project; Ministerio de Ciencia e Innovación y Universidades, Government of Spain, RTI2018-095894-B-I00 research project, Ayudas para la formación de profesorado universitario (FPU), grant ref. FPU18/02271 and Ayudas complementarias de movilidad destinadas a beneficiarios del programa de Formación del Profesorado Universitario (FPU) EST22/00696; Ministerio de Ciencia e Innovación, Government of Spain through the research project with reference PID2019-108435RB-I00; Consellería de Cultura, Educación e Universidade, Xunta de Galicia, Grupos de Referencia Competitiva, grant ref. ED431C 2020/24; Axencia Galega de Innovación (GAIN), Xunta de Galicia, grant ref. IN845D 2020/38;CITIC, Centro de Investigación de Galicia ref. ED431G 2019/01, receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%).

put image patch is compared against all the prototypes, and assigned a similarity score. These values are weighted and aggregated to obtain the per-class final score.

This strategy has been successfully adapted to other medical imaging domains such as chest radiography in the works of Kim *et al.*[11], where they convert the original multi-class approach into a multi-label. In the same imaging domain, but for the diagnosis of COVID-19 and pneumonia, Singh and Yow [12] proposed a close variant of the original proposal of Chen *et al.* [9], but introducing the concept of negative-positive prototypes. This means that not only considers the positive connection between similarity scores, but takes also into account negative relationship between other classes to calculate the overall per-class score. Nonetheless, the attempts to translate this approach to the DME in OCT domain were unsuccessful, as convergence would be dispersed due to the similarity of the pathological structures with other typical patterns that are present in the normal retinal layers.

However, Song *et al.* [13] recently introduced the concept of global and local prototype branches in the architecture with mutual correction, aiming to fix the problems that arose from the usage of Visual Transformers as the backbone of prototypical part networks. Their proposal added a parallel global branch to the prototypical core of the architecture. This branch, which provides guidance for the local branch (and also contributes to the final mutual joint decision), aids preventing the background dispersion common in transformer architectures. This has the collateral advantage that also contributes compensating the issue presented by the classical ProtoPNet in scenarios with homogeneity between class representative patterns. Moreover, thanks to the masking of the global branch, the final response maps product of the prototype usage results in sharp attention maps.

Thus, in this work we propose a transformer-based multi-prototype approach for DME analysis in OCT images. To do so, we perform an in-depth analysis of the behavior of this prototypical part architecture in two scenarios: CME vs. all (which includes DRT and normal retinal patterns) and DRT vs. all (which includes CME and normal retinal patterns). These two types represent the most significant types of DME retinal fluid. Moreover, we validate the proposed approach in a multi-device manner, in two of the most representative devices of the domain with a heterogeneous set of configurations.

2. MATERIALS AND METHODS

2.1. Dataset

The original expert-labeled image bank is composed of a multivendor dataset of 356 OCT images. 177 of the images were obtained from a CIRRUS™ HD-OCT 500 Carl Zeiss Meditec OCT device ranging in resolution from 682×446 pixels to $1,680 \times 1,050$ pixels and 179 from an HRA+OCT SPECTRALIS® from Heidelberg Engineering, Inc. ranging in resolution from 714×291 pixels to $1,535 \times 496$ pixels. To properly establish an scenario where the prototypes are representative of the pathology, the dataset contains images from both left and right eyes, including healthy and patients afflicted by different severity levels of DME. The protocols followed during the development of this project were conducted in accordance with the Declaration of Helsinki, approved by the Ethics Committee of Investigation from A Coruña/Ferrol (2014/437).

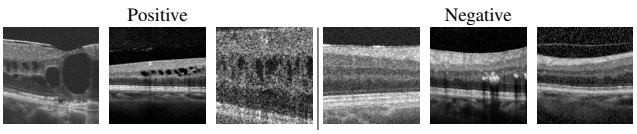

Fig. 1. Examples of three CME positive and negative samples.

Given that the classes are unbalanced in the bank of OCT images (and to further take advantage of the features present in each image), we extract different non-overlapping samples from the images to generate the dataset. This way, we are able to break down the independent regions in the images and balance the information. Thus, we can fully characterize all classes and study the suitability of the approach in a context with balanced data. To do so, we extract a window of 224×224 pixels centered column-wise relative to the image and row-wise relative to the retinal region in that column. Then, we repeat this process in both directions recursively with a distance of 224 horizontal pixels between sample centers. This process is done until the center column of the samples falls outside of the image. If only a part of the sample falls outside of the image, we mirror the borders of the image to fill the sample with the same patterns.

For both scenarios (CME vs. all and DRT vs. all) and OCT devices in this work, the negative cases also include samples with pathological patterns, albeit from other types of fluid accumulations. That is, the only criterion for a sample to be considered positive is the presence of the pathology, not also the absence of other types of fluid. For example, in the negative class, CME vs. all would also include DRT samples. This is shown in Fig. 1, where both positive and negative samples for CME include DRT patterns.

Finally, we randomly divide the multi-device dataset into 66.66% of the samples for training and 33.33% for validating the model. This way, the total dataset is divided into 1,486 samples for training and 495 for testing.

2.2. Methodology

In Fig. 2, we present a schematic representation of the transformer-based multi-prototype network. As we can see, the architecture considers two different types of prototype branches: global and local. While the local prototypes focus on particular image tokens, the class token aggregates information from all image tokens and produces a high-level abstraction. Moreover, the global branch serves as guidance

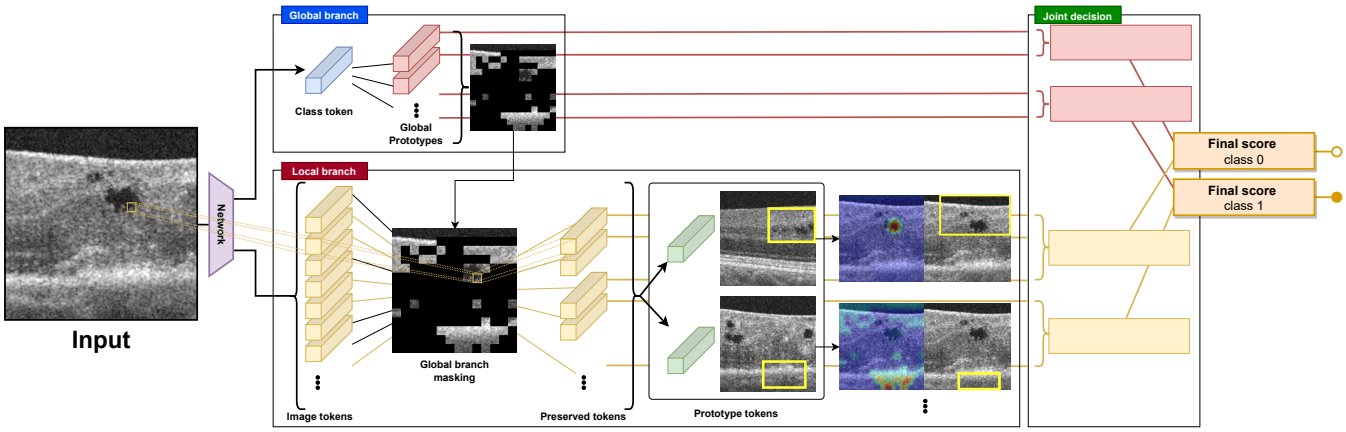

Fig. 2. Diagram of the transformer-based multi-prototype network. Note the global and local branches, where both converge towards the final weighted summed logits.

to the local branch, masking the background to focus the attention on foreground features. However, unlike the ProtoPFormer architecture [13], we do not include the prototypical part concentration loss that was designed to promote inter-prototype divergence and allow local prototypes to focus on diverse prototypical parts. In our scenario, this loss prevented the proper convergence of the models.

As the original proposal of the prototypical part networks with visual transformers backbone, we tested the Vision Transformer architectures Data-efficient image Transformers (DeiT) [14] and Class-attention image Transformers (CaiT) [15]. More specifically, we use the DeiT Tiny (DeiT-Ti), DeiT Small (DeiT-S), and CaiT XXS-24 (CaiT-XXS-24), similar to DeiT albeit replacing final self-attention modules with class-attention ones and freezing the patch embeddings in them. In this line, we first perform five warm-up epochs with a learning rate (LR) of 0.0001. After this warmup, the LR of the model are changed to 0.003 for both the add-on layer and prototype vectors, while 0.0004 for the visual transformer architectures. In our scenario, as texture patterns play a key role on the classification [16], we only used as on-line data augmentation strategy an horizontal flip with a 50% probability. This is done so because, due to the symmetric nature of the retinas in OCT images, any pattern present in the samples might appear in both horizontal orientations depending on the eye. The same way, batch size of 64 allowed for a better convergence during training.

All the models are trained for 200 epochs using cross-entropy loss as the learning process stagnates from this point onwards (and the generalisation capabilities of the model significantly diminish). We also perform five random repetitions per model and experiment to measure the robustness of the proposal. As optimizer, we use AdamW [17] with a cosine scheduler as the official DeiT implementations pretrained with ImageNet [18]. The epoch intervals to decay LR was set to five and the decay rate to 0.01 with a weight of 0.05.

The number of reserve tokens in the last layer, given that the samples tend to contain a lot of background information, was set to 64. At the same time, while in scenarios where structural information is more relevant (such as the classification of cars or birds like in the original proposal of ProtoPFormer), given that in our work the main discriminator is related to textural information, we set the dimension of the prototypes to 64. This reduced dimension helps to the filtering of unwanted noise and results in a better abstraction of textural patterns. Moreover, given that a bigger prototype dimension causes the introduction of noise, given the delicate balance between texture and noise information in this scenario (specially DRT), this lesser dimension size aided the performance of the models. Finally, the number of global prototypes per class, in our case, was empirically set to 10 and 24 for the local branch. A greater number of prototypes showed no improvements at the cost of generalisation and a lesser number began to deteriorate the performance of the model.

3. RESULTS AND DISCUSSION

In Tables 1 and 2, we present the results of both multi-device experiments. To validate the proposed approach, we measure the accuracy, F1 Score and Matthew's Correlation Coefficient (MCC). In the first scenario, CME vs. all, we can notice how the global branch of the network achieves better results than the local one. In fact, jointly, the final classification seems to be detrimental to the overall score. In particular, looking at the activations of the prototypes per sample, we noticed how the model tended to pay attention to the innermost and outermost layers of the retina instead of the actual fluid accumulations. In Fig. 3, we present the prototypical activations in a particular input sample of two characteristic global prototypes. As the reader might notice, the mask generated by reserved tokens completely overlaps the full pathological region.

This behavior is caused by the nature of the problem con-

Table 1. Mean and standard deviation for all the random repetitions for the CME vs. all experiments. The independent scores per branch and final results are presented.

Global branch	**Accuracy**	**F1 Score**	**MCC**
DeiT-Ti	0.9046 ± 0.0084	0.9191 ± 0.0083	0.8036 ± 0.0173
DeiT-S	0.9059 ± 0.0129	0.9209 ± 0.0097	0.8068 ± 0.0267
CaiT-XXS-24	0.9103 ± 0.0075	0.9247 ± 0.0056	0.8153 ± 0.0162
Local branch	**Accuracy**	**F1 Score**	**MCC**
DeiT-Ti	0.9030 ± 0.0051	0.9180 ± 0.0054	0.8004 ± 0.0109
DeiT-S	0.9026 ± 0.0118	0.9187 ± 0.0088	0.7998 ± 0.0238
CaiT-XXS-24	0.9095 ± 0.0039	0.9236 ± 0.0036	0.8137 ± 0.0080
Final	**Accuracy**	**F1 Score**	**MCC**
DeiT-Ti	0.9018 ± 0.0057	0.9168 ± 0.0061	0.7979 ± 0.0119
DeiT-S	0.9038 ± 0.0113	0.9194 ± 0.0087	0.8025 ± 0.0230
CaiT-XXS-24	0.9099 ± 0.0083	0.9243 ± 0.0064	0.8145 ± 0.0179

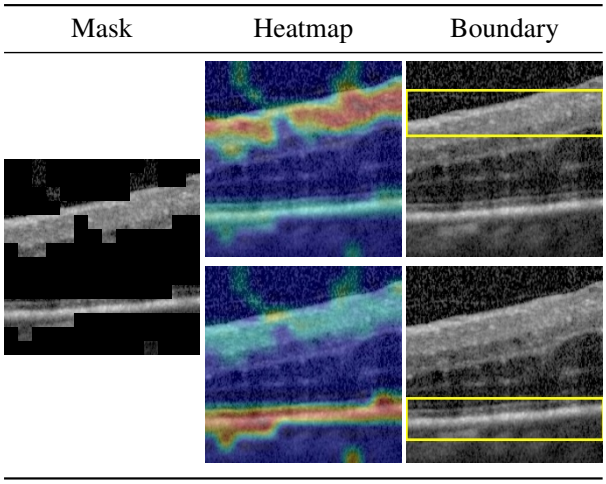

Fig. 3. Examples of two prototype activations centered in the innermost and outermost layers from the CME vs. all model.

sidered in this scenario. As we are evaluating CME vs. all, the negative class includes DRT patterns. As CME patterns are usually mixed with them, they tend to be obfuscated. Moreover, the fluid patterns between devices are very different, while normal retinal ones are similar. This way, it is indeed more versatile to consider as a prototype a normal retina instead of the fluid accumulations.

Regarding DRT, we can see the opposite scenario as with the previous experiment. Now the local branch attains the best score. In this case, as the positive class is represented by the DRT (which is usually intermingled with retinal tissues), the previous approach of paying attention to particular layers is not that useful. However, analyzing the prototypes for the positive class in Fig. 4, we can see how the global prototypes are actually sparse, not focused on the overall diffuse thickening of the retina that characterises DRT.

Finally, analyzing approaches with similar scenarios, we can assess that the classification power traded to explanatory

Table 2. Mean and standard deviation for all the random repetitions for the DRT vs. all experiments. The independent scores per branch and final results are presented.

Global branch	**Accuracy**	**F1 Score**	**MCC**
DeiT-Ti	0.8448 ± 0.0088	0.8615 ± 0.0087	0.6854 ± 0.0183
DeiT-S	0.8549 ± 0.0083	0.8713 ± 0.0100	0.7080 ± 0.0162
CaiT-XXS-24	0.8574 ± 0.0116	0.8750 ± 0.0115	0.7107 ± 0.0236
Local branch	**Accuracy**	**F1 Score**	**MCC**
DeiT-Ti	0.8481 ± 0.0099	0.8651 ± 0.0091	0.6907 ± 0.0199
DeiT-S	0.8558 ± 0.0103	0.8745 ± 0.0064	0.7091 ± 0.0210
CaiT-XXS-24	0.8598 ± 0.0139	0.8766 ± 0.0132	0.7147 ± 0.0281
Final	**Accuracy**	**F1 Score**	**MCC**
DeiT-Ti	0.8469 ± 0.0118	0.8640 ± 0.0119	0.6885 ± 0.0227
DeiT-S	0.8537 ± 0.0097	0.8716 ± 0.0091	0.7054 ± 0.0193
CaiT-XXS-24	0.8582 ± 0.0126	0.8737 ± 0.0111	0.7117 ± 0.0252

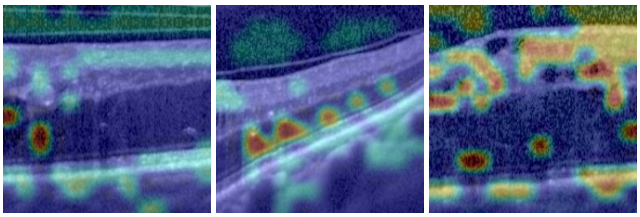

Fig. 4. Random examples of sparse activations of different prototypes from the DRT vs. all model.

power might not be significant. As reference, the work of Otero *et al.* [19] achieves similar accuracy for CME, around 0.9049. DRT achieves slightly better results, with an 0.0305 of improvement. Nonetheless, this comparison is unfair, as both works study different scenarios, datasets and methodologies.

4. CONCLUSIONS

In this work, we present a novel approach exploring all the potential of transformer-based multi-prototype networks for DME analysis in OCT images. In particular, we analyzed two of the most challenging scenarios from two representative devices of the domain. The usage of these prototypical networks allows for an explainability unparalleled respect to other explainable artificial intelligence techniques, as the models offer the classification in relation to previously known samples. Thus, experts would not only know the presence of pathological structures, but also the reason behind said decision. As future works, given the potential of this proposal, a multi-label approach would be interesting to explore the joint patterns and prototypes of other relevant pathologies and signal processing domains. Moreover, this architecture would significantly benefit from the shared information between classes to improve its performance. Similarly, a study on the reserve tokens for the local branch guidance would benefit from a per-class study, as well as adapting the prototypical part concentration loss.

5. REFERENCES

[1] Corinna E. Psomadakis, Nadeem Marghoob, Brady Bleicher, and Orit Markowitz, "Optical coherence tomography," *Clinics in Dermatology*, vol. 39, no. 4, pp. 624–634, July 2021.

[2] James H.B. Im, Ya-Ping Jin, Ronald Chow, and Peng Yan, "Prevalence of diabetic macular edema based on optical coherence tomography in people with diabetes: A systematic review and meta-analysis," *Survey of Ophthalmology*, vol. 67, no. 4, pp. 1244–1251, July 2022.

[3] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, 2019.

[4] Plácido L. Vidal, Joaquim de Moura, Macarena Díaz, Jorge Novo, and Marcos Ortega, "Diabetic macular edema characterization and visualization using optical coherence tomography images," *Applied Sciences*, vol. 10, no. 21, 2020.

[5] Placido L. Vidal, Joaquim de Moura, Jorge Novo, and Marcos Ortega, "Cystoid fluid color map generation in optical coherence tomography images using a densely connected convolutional neural network," in *2019 International Joint Conference on Neural Networks (IJCNN)*. July 2019, IEEE.

[6] Yiyang Wang, Mirtha Lucas, Jacob Furst, Amani A. Fawzi, and Daniela Raicu, "Explainable deep learning for biomarker classification of OCT images," in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. Oct. 2020, IEEE.

[7] Md Tanzim Reza, Farzad Ahmed, Shihab Sharar, and Annajiat Alim Rasel, "Interpretable retinal disease classification from oct images using deep neural network and explainable ai," in *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, 2021, pp. 1–4.

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[9] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su, "This looks like that: Deep learning for interpretable image recognition," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[10] Zhao Ren, Thanh Tam Nguyen, and Wolfgang Nejdl, "Prototype learning for interpretable respiratory sound analysis," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9087–9091.

[11] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon, "Xprotonet: Diagnosis in chest radiography with global and local explanations," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. 2021, pp. 15719–15728, Computer Vision Foundation / IEEE.

[12] Gurmail Singh and Kin-Choong Yow, "These do not look like those: An interpretable deep learning model for image recognition," *IEEE Access*, vol. 9, pp. 41482–41493, 2021.

[13] Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song, "Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition," 2022.

[14] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou, "Training data-efficient image transformers and; distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, Marina Meila and Tong Zhang, Eds. 18–24 Jul 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357, PMLR.

[15] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou, "Going deeper with image transformers," 2021.

[16] Joaquim de Moura, Plácido L. Vidal, Jorge Novo, José Rouco, Manuel G. Penedo, and Marcos Ortega, "Intraretinal fluid pattern characterization in optical coherence tomography images," *Sensors*, vol. 20, no. 7, 2020.

[17] Ilya Loshchilov and Frank Hutter, "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[19] Iago Otero, Plácido L. Vidal, Joaquim de Moura, Jorge Novo, and Marcos Ortega, "Computerized tool for identification and enhanced visualization of macular edema regions using OCT scans," in *27th European Symposium on Artificial Neural Networks, ESANN 2019, Bruges, Belgium, April 24-26, 2019*, 2019.