2025 IEEE Symposium on Trustworthy, Explainable and Responsible Computational Intelligence (CITREx Companion) | 979-8-3315-1972-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/CITRExCompanion65208.2025.10981490

CNN Layer-wise explanations for the prediction of Diabetic Retinopathy from OCTA images

1st Shawn Azimov
Computer Science Dept., UPC
Barcelona, Spain
shawn.azimov@estudiantat.upc.edu

2nd Jordi Delgado
Computer Science Dept., UPC
Barcelona, Spain
jdelgado@cs.upc.edu

3rd Enrique Romero
Computer Science Dept., UPC
Barcelona, Spain
eromero@cs.upc.edu

4th Alfredo Vellido
Computer Science Dept., UPC
Barcelona, Spain
avellido@cs.upc.edu

5th Javier Zarranz-Ventura
Institut Clínic d'Oftalmología (ICOF)
Hospital Clínic de Barcelona
Barcelona, Spain
zarranz@clinic.cat

6th Caroline König
Computer Science Dept., UPC
Barcelona, Spain
ckonig@cs.upc.edu

***Abstract*—Optical Coherence Tomography Angiography (OCTA) is a promising new imaging technique in ophthalmology for diagnosing Diabetic Retinopathy (DR) in diabetic patients. Early detection and treatment of DR is crucial to prevent vision loss and blindness. Deep Learning-based approaches to image classification have proven to be very effective in image analysis tasks, but their adoption hinges on the ability of the medical professionals to understand how these networks make their decisions. This study focuses on exploring explainable artificial intelligence techniques to better understand how a Convolutional Neural Network (CNN) classifies OCTA images of diabetic patients with varying levels of DR. This exploration is based on Signature Activation, a technique that generates holistic and class-agnostic explanations of the CNN's decisions.**

***Index Terms*—diabetic retinopathy, explainable artificial intelligence, signature activation, convolutional neural network**

I. Introduction

Diabetic Retinopathy (DR) is a medical condition that can damage the patient's retina as a result of the underlying Diabetes Mellitus (DM) disease, leading to low vision or blindness if left untreated [1]. DR occurs when the high blood sugar caused by diabetes damages the blood vessels in the retina. Both Type 1 and Type 2 DM can cause DR, though it is more prevalent in Type 1 DM patients. Early detection and treatment of DR in diabetic patients is crucial, which can be assisted by the analysis of retinal images of different modalities [2] using Deep Learning (DL) methods [3]–[5].

DL methods in general can be considered to be "black boxes", meaning that it is very difficult to understand, and therefore justify, their inner workings, or ascertain how they make prediction or classification decisions based on the available inputs. The most common DL approach for image analysis is Convolutional Neural Networks (CNN). In recent years, various methods have been applied to CNNs to better understand how they make their predictions/classifications. They fall within the domain of explainable AI (XAI) [6] and, if pertaining to image data, saliency methods.

This study explores a particular saliency method called Signature Activation (SA), and compares it to another related method known as Gradient-weighted Class Activation Mapping (Grad-CAM). Both are used on CNN models, trained on a real clinical image dataset of Type 1 Diabetes patients for a prediction task involving the presence or absence of DR. The dataset comprises images acquired via Optical Coherence Tomography Angiography (OCTA). From the trained CNNs, saliency maps are generated through SA and Grad-CAM. These saliency maps are analyzed to study how such methods can be used to improve explainability, in particular when used in the medical setting under consideration.

II. Materials

The OCTA images dataset analyzed in this study was obtained in a clinical trial sponsored by Fundació Clínic per a la Recerca Biomèdica, Hospital Clínic de Barcelona and Fundació La Marató de TV3 (ClinicalTrials.gov NCT03422965). For each eye, 4 different OCTA images were collected, including 2 types of scanning protocols - 3 × 3 mm and 6 × 6 mm. The latter provides a wider view of the retinal area, while the former is more focused on the central region. The original dataset involves six classes: five of them include patients with Type 1 DM and either no-DR or DR in different stages of development, while a sixth consists of control patients. Our experiments concern the classification of Type 1 DM patients with no DR (label 0, 438 patients) *vs.* Type 1 DM patients with different levels of DR: label 1, comprising cases of mild NPDR (158), moderate NPDR (24), severe NPDR (2) and proliferative DR (1). This choice is made on the basis of medical relevance.

The dataset was shuffled and split into training (492 items) and validation (131) sets using an 80/20 stratified split. For this study, 3 × 3 mm superficial capillary plexus (SCP) images,

This study was funded by Spanish PID2022-143299OB-I00/AEI/10.13039/501100011033/FEDER, UE and by Fundació La Marató de TV3, La Marató 2015, Diabetis i Obesitat (grant 201633.10), and projects PI18/00518 and PI21/01384, funded by ISCIII and co-funded by the European Union.

which focus on the superficial vascular plexus, were used, following previous studies [7], [8]. Various types of data augmentation (including included random left-right flips, up-down flips, 90 degree rotations, brightness, contrast, central zoom and random zoom) were applied to compensate some of the underrepresented categories of the dataset, such as severe NDPR and proliferative DR.

III. Methods

As mentioned in the introduction, in this study we focus on the SA [9] and Grad-CAM [10] techniques. Signature Action is a class-agnostic method, meaning that it provides a holistic view of what a model's layer focuses on, regardless of which class is represented in the image. Such methods can be especially useful for medical imaging, where it is important to show the saliency maps of all of the relevant parts of the image contributing to a classification.

Grad-CAM is a popular explainability technique for CNNs [10], which improves upon previous interpretability methods such as CAM (Class Activation Mapping), by allowing it to work on a broad range of network architectures. It differs from SA in some aspects. While Grad-CAM is class-discriminative in multi-class scenarios, meaning that it generates separate visualizations for each class, SA is class-agnostic, generating a holistic activation map for all classes. Regarding the architecture of CNN, we used the VGG19 [11], [12] model in the experiments. This is a popular architecture belonging to the VGG (Visual Geometry Group) family of networks. It consists of 19 layers, 16 of which are convolutional, while the other 3 are fully-connected layers. The convolutional layers consist of 5 blocks with max-pooling layers in between of stride 2, gradually reducing the dimensions of the feature maps by half. The blocks contain 2, 2, 4, 4 and 4 convolutional layers each, respectively, and use 3×3 kernel filters and ReLU activation. The three fully connected layers have 4096, 4096, and 1000 channels each in the original design.

CNN models were trained using binary cross-entropy loss with an Adam optimizer. A checkpoint callback was used to save the best model based on the lowest achieved validation loss during training. Various performance assessment metrics were calculated, including accuracy (ACC), AUC score, F1 score per class, and Brier score per class [13].

SA explanations were generated for various images from the validation dataset. The method was applied to various layers of the base model, resulting in activation maps of varying degree of resolution prior to resizing (depending on the shape of the given layer), in such a way that how the data is propagated through the model's depth can be seen. The explanations of the predictions obtained by SA were post-processed to provide an easily readable representation from the original image.

The VGG19 model contains 5 blocks, with 16 total convolutional layers, and each of them is used to visualize the saliency maps. To reduce the number of layers, just the last activation layers of each block are used to visualize the propagation of signal through the model's depth.

IV. Experiments and Results

A. Batch size experiments

For experiments with the VGG19 model, varying batch sizes of the training set (16, 32 and 64) were tested. The results are shown in Table I. The model trained with a batch size of 32 quite consistently shows the best results across most metrics, and was therefore used in the remaining experiments.

TABLE I
Validation set results according to several metrics for the DR 2-class dataset, using VGG19 with varying batch sizes.

BATCH	LOSS	ACC	AUC	F1-0	F1-1	BR0	BR1
16	0.507	0.756	0.752	0.842	0.467	**0.083**	0.390
32	**0.492**	**0.778**	**0.777**	**0.851**	**0.567**	0.092	**0.340**
64	0.504	0.709	0.766	0.810	0.387	0.091	0.369

B. Dataset augmentation experiments

The results for the models trained with different data augmentation techniques and parameters are summarized in Table II. For the sake of clarity, the augmentation variants applied are assigned an identifier, as listed below:

- 1. Brightness (offset 0.1) + Contrast (offset 0.1)
- 2. Brightness (offset 0.1) + Contrast (offset 0.1) + Central Zoom (offset 0.1)
- 3. Brightness (offset 0.2) + Contrast (offset 0.2) + Central Zoom (offset 0.2)
- 4. Brightness (offset 0.1) + Contrast (offset 0.1) + Central Zoom (offset 0.2)
- 5. Central Zoom (offset 0.2)
- 6. Random Zoom (offset 0.2)
- 7. Random Zoom (offset 0.3)
- 8. Brightness (offset 0.1) + Contrast (offset 0.1) + Random Zoom (offset 0.2)

TABLE II
Validation set results for the DR 2-class dataset obtained with VGG19, with varying augmentation techniques (AUGM).

AUGM	LOSS	ACC	AUC	F1-0	F1-1	BR0	BR1
1	0.493	**0.779**	0.778	**0.851**	**0.567**	0.092	0.340
2	0.494	0.763	0.782	0.847	0.475	0.080	0.377
3	0.526	0.741	0.742	0.835	0.393	**0.071**	0.439
4	0.546	0.718	0.739	0.819	0.351	0.076	0.455
5	0.498	0.741	0.769	0.828	0.469	0.091	0.358
6	**0.478**	0.771	**0.792**	0.847	0.545	0.094	**0.323**
7	0.509	0.748	0.749	0.834	0.476	0.093	0.364
8	0.513	0.741	0.762	0.830	0.452	0.084	0.398

The results show that the best validation loss and AUC were achieved when using random zoom with an offset of 0.2. The model with brightness and contrast with an offset of 0.1 shows the best performance on the accuracy metric and the F1 scores for both class 0 and class 1. When combining these augmentations, however, the model performance worsens. Central zoom augmentation was also tested, but the results were worse than those corresponding to random zoom.

C. *Experiments with different classification heads*

For the next set of experiments, the custom VGG19 model was tested using three different types of classification heads, namely Global Average Pooling, Global Max Pooling, as well as a Flatten layer followed by a Dense layer with 128 neurons. Various hyperparameters were tested for each type of model and Table III summarizes the best results for each architecture.

TABLE III
VALIDATION SET METRICS FOR THE DR 2-CLASS DATASET USING VGG19 MODELS WITH DIFFERENT CLASSIFICATION HEADS.

MODEL	LOSS	ACC	AUC	F1-0	F1-1	BR0	BR1
AvgPool	**0.478**	**0.771**	**0.792**	**0.847**	**0.546**	0.094	0.323
MaxPool	0.512	0.748	0.756	0.833	0.492	**0.092**	0.364
Flatten	0.527	0.748	0.759	0.833	0.492	0.121	**0.311**

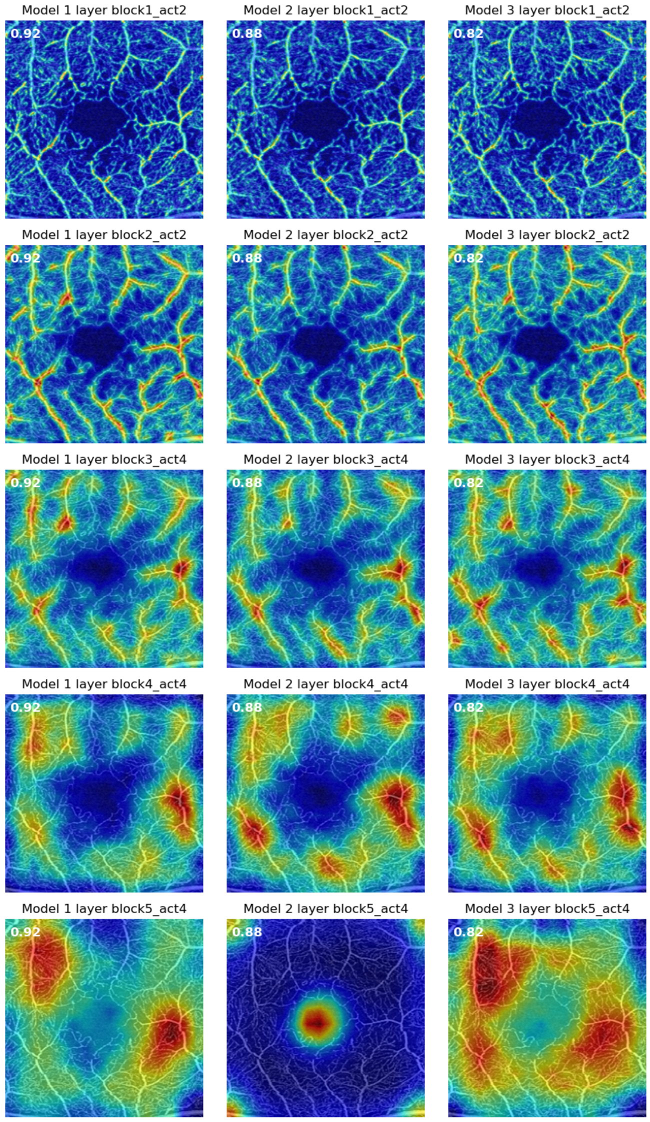
Model 1 layer block1_act2
Model 2 layer block1_act2
Model 3 layer block1_act2
0.92
0.88
0.82
Model 1 layer block2_act2
Model 2 layer block2_act2
Model 3 layer block2_act2
0.92
0.88
0.82
Model 1 layer block3_act4
Model 2 layer block3_act4
Model 3 layer block3_act4
0.92
0.88
0.82
Model 1 layer block4_act4
Model 2 layer block4_act4
Model 3 layer block4_act4
0.92
0.88
0.82
Model 1 layer block5_act4
Model 2 layer block5_act4
Model 3 layer block5_act4
0.92
0.88
0.82

Fig. 1. Signature Activations for three different VGG19 models with Global Average Pooling (left), Global Max Pooling (middle) and Flatten layer with a Dense layer (right) for correctly-predicted image 86, label DR. Values from 0 (dark blue) to 1 (dark red).

According to these results, the model with Global Average Pooling layer prior to classification performs best. The Global Max Pool layer and the Flatten layer models achieved the same accuracy, but the Global Max Pool model shows slightly better validation loss. Figure 1 shows the side by side comparison between the SA maps of three models tested for the illustrative example correctly-predicted *image 86* of class 1. They show that there is a very similar trend for earlier layers of all three models, where they consistently focus on the vessels up to block 4. However, they do greatly differ in the last layer, where the Global Max pooling model focuses on the foveal avascular zone and the other models (AvgPooling and Flatten) still focus in the vessels rather than the perifoveal region.

The model using Global Average Pooling layer prior to classification achieved the best metrics. Therefore, this model was used to generate the SA maps in the remaining experiments.

D. *Signature Activations*

Figure 2 shows a selection of the saliency maps from the activation layers from block 1 to 5 for a DR-type image.

The SA of the VGG19 model show a gradual focus on larger and larger features of the image, which is what you would expect from a multi-layer CNN, so this model seems a good fit for the task in terms of complexity. Interestingly, it focuses on the vessels even in the last layers (block 5), revealing that SA concentrates in the vessels and dismisses the perifoveal region.

E. *Signature Activation comparison to Grad-CAM*

Figure 3 shows the SA (left column) and Grad-CAM activation maps (right clumn) of 5 layers of the VGG19 model for *image 86* of type DR (class 1). A similar figure for *image 100* (class 0) is, for the sake of brevity, available in an *external repository*).

Comparing the two methods and their corresponding activation maps of the same image and layers, some clear differences can be found. In earlier layers of the model, Grad-CAM provides a more sparse visualization of the model's focus, highlighting many distinct points in the image, rather than the blood vessels distinctly, as SA does. This is likely to do with the fact that Grad-CAM utilizes gradients in the calculation of the activation maps, which can potentially add more noise to the resulting maps, while SA does not. The Grad-CAM activation maps of later layers of the model show a more holistic focus on the areas of the image in the case of *image 86* of type DR. The same analysis was done for images of type No-DR (*image 100*), where very few and localized areas are highlighted by the method.

Another interesting distinction between both explainability methods is that some activation maps show the model focusing on the center of the image, the foveal avascular zone (FAZ). The shape and size of the FAZ has been investigated as biomarkers of diabetic retinopathy [14], [15]. From these parameters, size has been the most widely accepted one, but none is universally accepted [16]. In any case, this supports

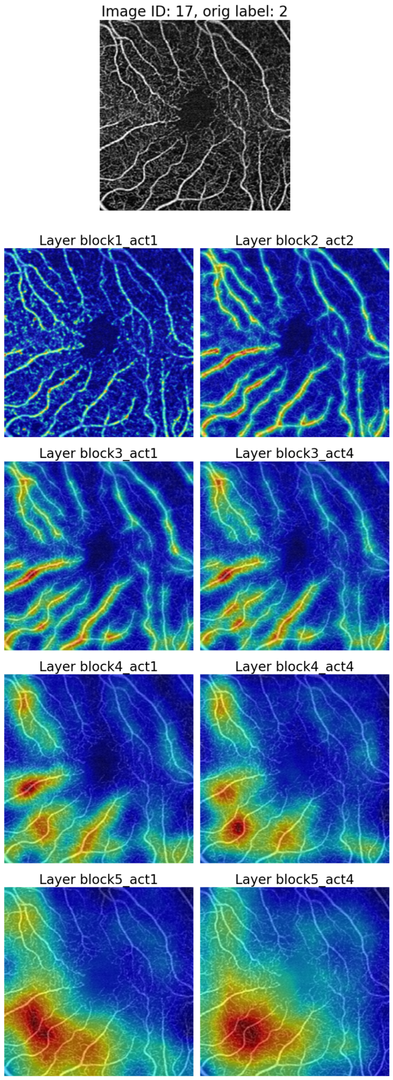

Fig. 2. SA of activation layers of the VGG19 model (Image 17)

the findings of the study, which shows that some models are focusing clearly in this region.

V. Discussion

The experimental results with different classification head architectures clearly show that there are important differences in their SA maps, in particular in the last activation layer as reported in section IV-C. They show that the choice of the pooling/flatten layer can significantly impact the resulting activation maps of the final activation layer. In fact, the differences begin to show not just in the activation maps of the final layer, but rather in the 4 layers of the final block of the model, becoming most pronounced in the last layer. However, equally important is the finding that the rest of the layers of the models are not significantly impacted by this architectural change according to their SA maps.

The VGG19 model with the best metrics was used to generate SA maps as well as Grad-CAM activation maps for several images. A comparison of both methods reveals that, in earlier layers of the model, Grad-CAM provides a more sparse visualization of the model's focus rather than highlighting the blood vessels distinctly as happens with the SA method. The Grad-CAM activation maps of later layers of the model show a more holistic focus on the areas of the image similar to the holistic view of the signature activation maps. The study reveals that the SA and Grad-CAM methods provide different insights into the inner workings of the model. Signature Activation is quite consistent in highlighting the vessels in the images, in stark contrast with Grad-CAM.

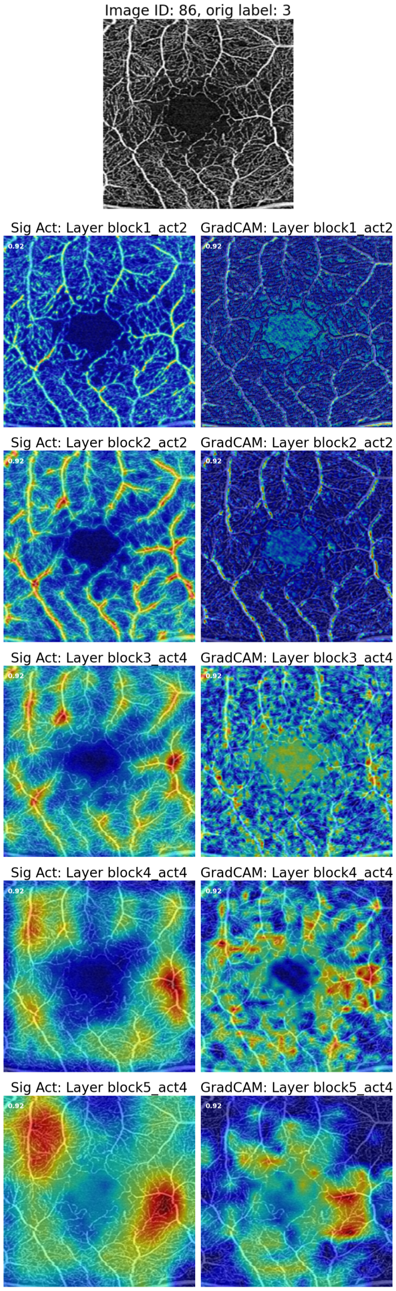
Image ID: 86, orig label: 3
Sig Act: Layer block1_act2
GradCAM: Layer block1_act2
Sig Act: Layer block2_act2
GradCAM: Layer block2_act2
Sig Act: Layer block3_act4
GradCAM: Layer block3_act4
Sig Act: Layer block4_act4
GradCAM: Layer block4_act4
Sig Act: Layer block5_act4
GradCAM: Layer block5_act4
0.92

Fig. 3. SA (left) and Grad-CAM (right) for image 86 (Label DR).

References

[1] Diabetic Retinopathy. https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy, 2024. Accessed: 2024-09-17.

[2] Beau J Fenner, Raymond LM Wong, Wai-Ching Lam, Gavin SW Tan, and Gemmy CM Cheung. Advances in retinal imaging and applications in diabetic retinopathy screening: a review. *Ophthalmology and therapy*, 7:333–346, 2018.

[3] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.

[4] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

[5] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[6] Paulo JG Lisboa, Sascha Saralajew, Alfredo Vellido, Ricardo Fernández-Domenech, and Thomas Villmann. The coming of age of interpretable and explainable machine learning models. *Neurocomputing*, 535:25–39, 2023.

[7] Laura Carrera Escale, Anass Benali Bendahmane, Ann Christin Rathert, Ruben Martín Pinardel, Carolina Bernal Morales, Anibal Alé Chilet, Marina Barraso Rodrigo, Sara Marín Martinez, Alfredo Vellido Alcacena, and Enrique Romero Merino. Radiomics-based assessment of optical coherence tomography angiography images for diabetic retinopathy diagnosis. *Ophtalmology science*, 3(2, article 100259), 2023.

[8] V Sudha and TR Ganeshbabu. A Convolutional Neural Network Classifier VGG-19 Architecture for Lesion Detection and Grading in Diabetic Retinopathy Based on Deep Learning. *Computers, Materials & Continua*, 66(1), 2021.

[9] Jose Roberto Tello Ayala, Akl C. Fahed, Weiwei Pan, Eugene V. Pomerantsev, Patrick T. Ellinor, Anthony Philippakis, and Finale Doshi-Velez. Signature Activation: A Sparse Signal View for Holistic Saliency. *arXiv:2309.11443*, 20 Sep 2023.

[10] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from Deep Networks via Gradient-based Localization. *arXiv:1610.02391*, 3 Dec 2019.

[11] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*, 4 Sep 2014.

[12] Sheldon Mascarenhas and Mukul Agarwal. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*, volume 1, pages 96–99. IEEE, 2021.

[13] Kaspar Rufibach. Use of Brier score to assess binary predictions. *Journal of clinical epidemiology*, 63(8):938–942, 2010.

[14] Brian D Krawitz, Shelley Mo, Lawrence S Geyman, Steven A Agemy, Nicole K Scripsema, Patricia M Garcia, Toco YP Chui, and Richard B Rosen. Acircularity index and axis ratio of the foveal avascular zone in diabetic eyes and healthy controls measured by optical coherence tomography angiography. *Vision research*, 139:177–186, 2017.

[15] Torben Guijarro, Javier Zarranz-Ventura, Enrique Romero, and Alfredo Vellido. Diabetic retinopathy prediction from OCTA-based vessel tortuosity metrics using machine learning. In *19^{th} conference on Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB 2024)*.

[16] Wasim A Samara, Abtin Shahlaee, Murtaza K Adam, M Ali Khan, Allen Chiang, Joseph I Maguire, Jason Hsu, and Allen C Ho. Quantification of diabetic macular ischemia using optical coherence tomography angiography and its relationship with visual acuity. *Ophthalmology*, 124(2):235–244, 2017.