Weakly Supervised Lesion Localization of Nascent Geographic Atrophy in Age-Related Macular Degeneration

Heming Yao[1], Adam Pely[1], Zhichao Wu[2,3], Simon S. Gao[1], Robyn H. Guymer[2,3], Hao Chen[1], Mohsen Hejrati[1], and Miao Zhang[1(✉)]

[1] Genentech, South San Francisco, CA, USA
zhang.miao@gene.com
[2] Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, East Melbourne, VIC, Australia
[3] Ophthalmology, Department of Surgery, The University of Melbourne, Melbourne, VIC, Australia

Abstract. The optical coherence tomography (OCT) signs of nascent geographic atrophy (nGA) are highly associated with GA onset. Automatically localizing nGA lesions can assist patient screening and endpoint evaluation in clinical trials. This task can be achieved with supervised object detection models, but they require laborious bounding box annotations. This study thus evaluated whether a weakly supervised method could localize nGA lesions based on the saliency map generated from a deep learning nGA classification model. This multi-instance deep learning model is based on 2D ResNet with late fusion and was trained to classify nGA on OCT volumes. The proposed method was cross-validated using a dataset consisting of 1884 volumes from 280 eyes of 140 subjects, which had volume-wise nGA labels and expert-graded slice-wise lesion bounding box annotations. The area under Precision-Recall curve (AUPRC) or correctly localized lesions was 0.72($\pm$0.08), compared to 0.77($\pm$0.07) from a fully supervised method with YOLO V3. No statistically significant difference is observed between the weakly supervised and fully supervised methods (Wilcoxon signed-rank test, $p = 1.0$).

Keywords: OCT · weakly supervised learning · object detection

1 Introduction

Nascent geographic atrophy (nGA), originally described by Wu et al. [1], describes features of photoreceptor degeneration seen on optical coherence tomography (OCT) imaging that are strongly associated with the development of geographic atrophy (GA), a late stage complication of age-related macular degeneration (AMD). A recent study reported that the development of nGA in individuals with intermediate AMD was associated with a 78-fold increased rate of GA development [2]. Thus, nGA could potentially

H. Yao and A. Pely—Contributed equally to this work.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 477–485, 2023.
https://doi.org/10.1007/978-3-031-43907-0_46

act as an earlier biomarker of AMD progression, or potentially as an earlier endpoint in intervention studies aiming to slow GA development [3]. Thus being able to easily identify eyes with nGA and localize nGA lesions is important in clinical trials and research.

However, identifying and grading the location of nGA lesions in OCT volume scans can be a laborious task, and would be an operationally expensive undertaking in clinical trials. Automation of this task would be invaluable when seeking to quantify the number of nGA lesions present, or when seeking to identify a smaller subset of B-scans for manual expert review (an "AI-assisted" approach). While the localization of nGA could be tackled by supervised object detection models – as demonstrated in other types of lesions [4–6] – it takes domain experts a large amount of time to provide sufficient number of lesion level annotations (e.g. with a bounding box). On the other hand, weakly supervised methods require only coarse annotations, and they have been popular in computer vision tasks where dense annotations are difficult to obtain [7].

In this work, we sought to develop a deep learning-based method to automate the localization of nGA lesions on OCT imaging, trained only on the information about the presence or absence of nGA at the volume level. A weakly supervised algorithm was developed that utilizes the saliency maps from Gradient Class Activation Maps (GradCAM) technique [8]. While existing literature has demonstrated the ability of the GradCAM in post-hoc model interpretation [9–12], it is unknown whether the saliency map can help further localize the class-related lesions or abnormalities that are often sparse anatomically. We thus explored the possibility of GradCAM in identifying the location of nGA-related abnormalities after training a model for classifying nGA in a 3D volume scan.

2 Methods

2.1 Dataset

This study included participants in the sham treatment arm of the Laser Intervention in the Early Stages of AMD study (LEAD, clinicaltrials.gov identifier, NCT01790802). The LEAD study was conducted according to the International Conference on Harmonization Guidelines for Good Clinical Practice and the tenets of the Declaration of Helsinki. Institutional review board approval was obtained at all sites and all participants provided written informed consent.

The participants of LEAD study were required to have bilateral large drusen and a best-corrected visual acuity of 20/40 or better in both eyes at baseline [13]. Participants were evaluated at the baseline and every 6-month follow up visits for up to 36-months. At each visit, OCT imaging was performed following pupillary dilation, by obtaining a 3D volume scan consisting of 49 B-scans (i.e. 2D slices along X-Z direction) covering a $20° \times 20° \times 1.9$ mm region of the macula, with $1024 \times 49 \times 496$ voxels anisotropically sampled along X, Y and Z d directions respectively.

Multimodal imaging was used to assess the development of late AMD as an endpoint in the LEAD study, which included nGA detected on OCT imaging. In order to evaluate the association between nGA and the subsequent development of GA as detected on color fundus photographs (CFP; the historical gold standard for atrophic AMD) in a

previous study, OCT imaging and CFP were independently re-graded for the presence of nGA and GA respectively [14]. In this sub-study, we included individuals who did not have nGA at baseline based on the above independent re-grading of OCT imaging, and who had at least one follow-up visit.

A total of 1,884 OCT volumes from 280 eyes of 140 individuals were included in this analysis (1,910 volumes were collected, but 26 volumes were excluded from the study due to the development of neovascular AMD in the eye). In this study, the development of nGA was assessed by manual grading of all 49 B-scans of each OCT volume scans, and nGA was defined by the subsidence of the outer plexiform layer and inner nuclear layer, and/or the presence of a hyporeflective wedge-shaped band within Henle's fiber layer, as per the original definition [1]. All OCT volume scans were initially assessed by a senior grader, and all visits of any eye deemed to have questionable or definite nGA were then reviewed by two further experienced graders [2].

Overall, nGA was graded as being absent and present in 1,766 and 118 OCT volume scans respectively. In the context of this study, note that nGA also includes lesions that could also meet the criteria for having complete retinal pigment epithelium and outer retinal atrophy (cRORA), if the lesion also had choroidal signal hypertransmission and retinal pigment epithelium (RPE) attenuation or disruption of $\geq$250 μm [15]. Graders also concurrently graded the location of nGA lesions by identifying the B-scans with nGA lesions and by drawing bounding boxes on the B-scans horizontally covering the subsidence, vertically from the inner limiting membrane (ILM) to Bruch's membrane. For the weakly supervised model, the bounding boxes were used only in evaluating the weakly supervised lesion localization, not in model training. The bounding boxes were then used to train a fully supervised object detector to compare the results of the weakly supervised and fully supervised methods.

2.2 Deep Learning Architecture

A late-fusion model with a 2D ResNet backbone was developed to classify 3D OCT volumes, considering their anisotropic nature. As shown in Fig. 1a, B-scans from a 3D OCT volume were fed into a B-scan detector, and the outputs, which are vectors of classification logits for each B-scan, were averaged to generate prediction scores for the volume. Thinking of the B-scans as instances and the OCT volumes as bags, this framework can be categorized as simplified multi-instance learning [16] in which the network was trained on weakly labeled data, using labels on bags (OCT volumes) only. During the training process, given an OCT volume annotated as nGA, the network was forced to identify as many B-scans with nGA lesion to improve the final prediction of nGA, thus the trained model allows prediction of nGA labels on OCT volumes as well as on individual B-scans.

The details of the B-scan classifier are shown in Fig. 1b. An individual B-scan of size 1024 $\times$ 496 from the volume is downsampled to 512 $\times$ 496 and passed through the ResNet-18 backbone, which outputs activation maps of 512 $\times$ 16 $\times$ 16. A max-pooling layer and an average pooling layer are concatenated to generate a feature vector of 1024. Then a fully connected layer was applied to generate the classification logit for the B-scan.

The classification model was evaluated on its own both in terms of volume-wise and slice-wise performance in classifying nGA. After it was confirmed that the classification model worked well, the ability of the model to localize the lesions within individual OCT slices was evaluated.

Given an OCT volume and a trained model, saliency maps were generated with the GradCAM technique [8] to visualize regions making larger contributions to the final classification. In Fig. 2a, GradCAM output was overlaid as the yellow channel on the input images for easy visualization of the saliency as well as the original grayscale image. The saliency map from a legitimate model should highlight nGA lesions, thus GradCAM output can help localize nGA lesions. The objective was to localize nGA lesions in the 3D OCT volume, i.e. to identify which B-scans have nGA lesions and to generate a bounding box surrounding the lesions in those B-scans with confidence scores.

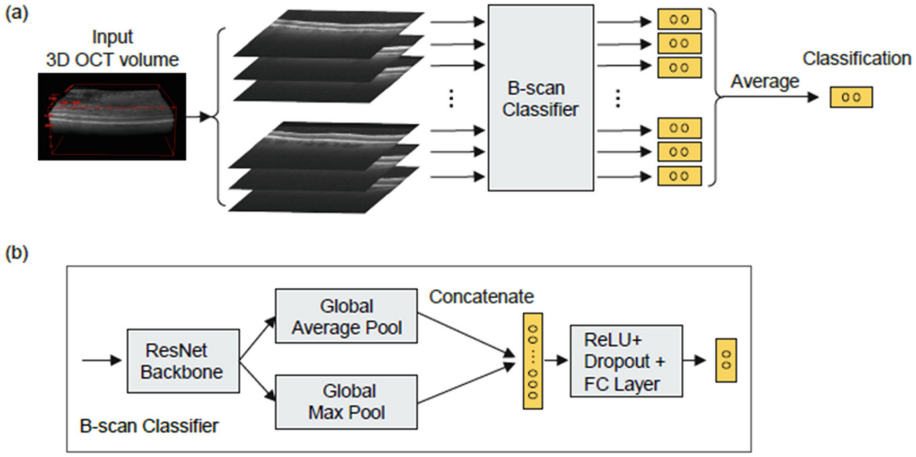

Fig. 1. Deep learning architecture of the nGA classification model. (a) Model network for 3D OCT volume. (b) The B-scan classifier. ReLU = rectified linear unit. FC = fully connected.

As illustrated in Fig. 2, the automated image processing pipeline was built upon adaptive thresholding and connected component analysis [17]. For each B-scan with positive logit, one or multiple bounding boxes covering potential lesions were detected. The confidence score for each bounding box was estimated from the individual classification logit of the B-scan classifier as Eq. (1).

$$S\left(\frac{l}{n}\frac{h}{\Sigma h}\right) \tag{1}$$

where S is the sigmoid function, l is the individual B-scan classification logit, and n is the number of B-scans in a volume, h is the mean saliency in the detected region and Σh is the total mean saliency of all detected regions within the B-scan. A higher confidence score implies a higher possibility that the detected region covers nGA lesions. Since only class labels of 3D OCT volume are required for training, the proposed lesion localization algorithm was weakly supervised.

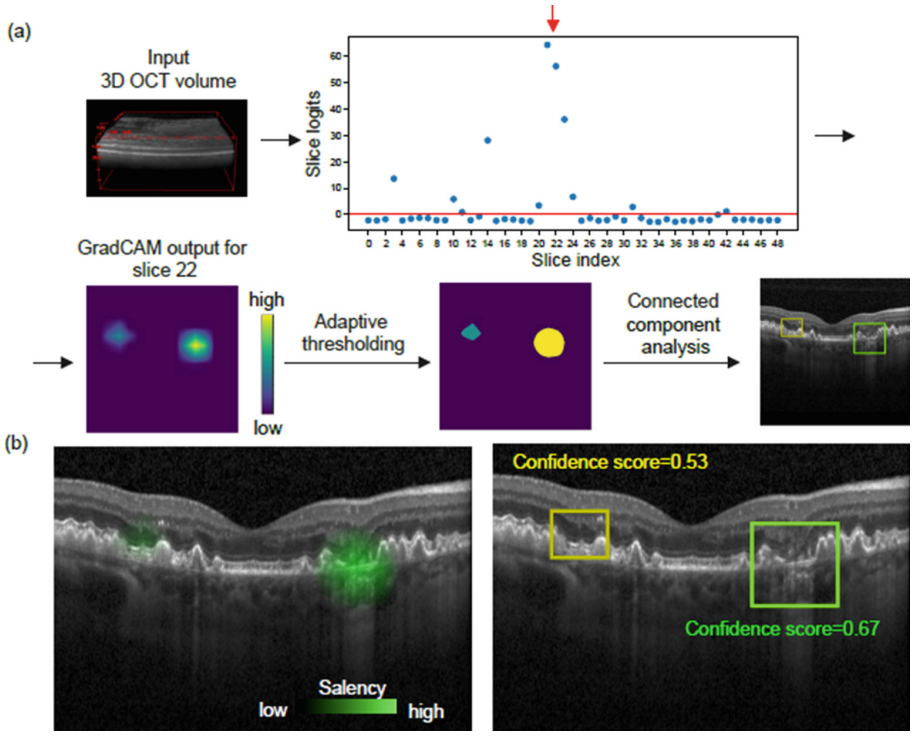

Fig. 2. (a) Illustration of locating nGA lesions with B-scan logit and GradCAM. B-scans with positive logits were selected, their GradCAM outputs (in the viridis colormap) were thresholded adaptively, then a bounding box was generated by the connected component analysis. A confidence score of the bounding box was estimated based on its average saliency and the corresponding B-scan logit. (b) Examples of a B-scan overlaid with GradCAM output (left) and bounding boxes with confidence scores (right), the yellow bounding box with confidence score below threshold was removed in following processing. (Color figure online)

2.3 Model Training, Tuning, and Validation Test

Considering the relatively small number of participants in the dataset, a five-fold cross-validation was applied to evaluate the proposed method's performance. We followed the nomenclature for data splitting as recommended previously [18]. For each fold, the validation test set of OCT volumes were obtained from roughly 20% of the participants stratified on whether the individual developed nGA. The OCT volumes from the remaining 80% individuals were further split into training (64%) and tuning sets (16%), with volumes from one individual only existing in one of the sets. With the proposed data split strategy, the corresponding validation test set was not used in the training or hyperparameter tuning process.

Pre-processing was performed on B-scans for standardization. The B-scans were first resized to 512 $\times$ 496, followed by rescaling the intensity range to [0, 1]. Data augmentation, including rotation of small angles, horizontal flips, add on of Gaussian

noises, and Gaussian blur were randomly applied to improve the model's invariance to those transformations.

A Resnet-18 backbone pre-trained on the ImageNet dataset was used. During the model training, the Adam optimizer was used to minimize focal loss. The L2 weight decay regularization was used to improve the model's generalization.

As a benchmark for the weakly supervised lesion localization, a fully supervised YOLOv3 object detector [19] with a Resnet-18 backbone was trained using the bounding box information for each B-scan.

A successful lesion localization was recorded only if the bounding box output overlapped with the bounding boxes annotated by clinicians with an intersection over union (IoU) value of at least 0.05. The area under the Precision-Recall curve (AUPRC) was calculated to evaluate the model performance. In patient screening, a high recall is preferred over precision. Considering the difference of the two methods, different strategies were used to determine the confidence threshold in calculating the precision and recall values in the validation test dataset. For the weakly supervised method, the threshold for confidence score that would achieve a recall value of 0.98 for nGA volume classification in the training and tuning sets is used. For the supervised method, the confidence threshold which would achieve a recall value of 0.9 for bounding box detection in the turning set is used.

3 Results

3.1 Performance and Saliency Map Analysis of the nGA Classification Model on OCT Volumes

The deep learning based nGA classification model achieved an AUPRC of 0.83(±0.09) in classifying 3D OCT volumes. Based on the trained 3D OCT volume classification model and input OCT volumes, we generated the corresponding saliency map using GradCAM technique. Examples of GradCAM output are shown in Fig. 2. Values in the GradCAM output indicate the importance of the corresponding pixel in the input B-scan to the model's prediction. A higher (brighter) value means the corresponding pixel contributes more to the model's prediction that the input OCT volume is positive. A thresholding was applied to the pixel values to determine the region where the bounding box delineating nGA should be drawn.

3.2 Performance of the Weakly Supervised Localization of nGA Lesions

The weakly supervised algorithm achieved a similar level of performance for localizing nGA when compared to the YOLOv3 based fully supervised method, without utilizing bounding box annotations and these findings are illustrated in Fig. 3. The YOLOv3 based method achieved an AUPRC of 0.77(±0.07) compared to 0.72(±0.08) for the weakly supervised model; no statistically significant difference was observed between the two methods (Wilcoxon signed-rank test for AUPRC, $p = 1.0$).

In the patient screening setting described previously, the YOLOv3 based method achieved a precision and recall of 0.53(±0.16) and 0.88(±0.06), compared to 0.39(±0.13) and 0.88(±0.02) for the weakly supervised method.

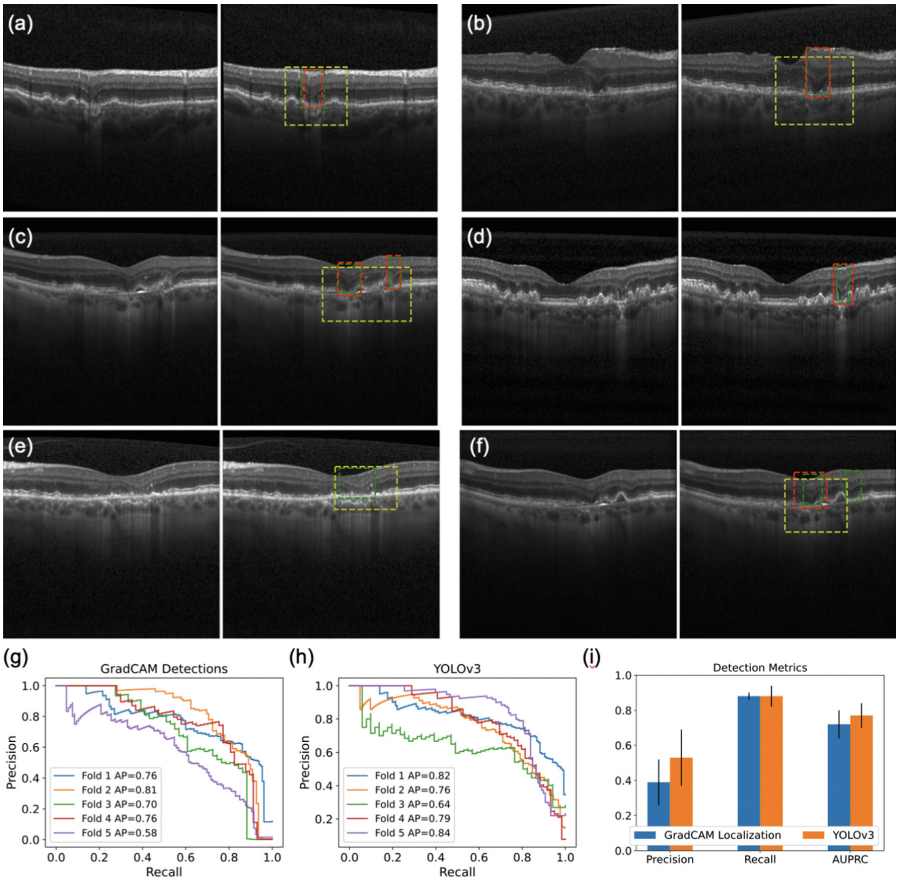

Fig. 3. (a)–(f) Example results from the weakly and fully supervised detection methods alongside the ground truth. The ground truth is in green, the YOLOv3 detector output in red, and the GradCAM based detector output is in yellow. (a–c) A true positive case for both methods. (d) A true positive case for the fully supervised method, but false negative for the weakly supervised method. (e) A true positive case for the weakly supervised method, but false negative for the fully supervised method. (f) A case where there are two ground truth lesions, but both methods detected only the left one. (g) The per fold precision-recall curve for the GradCAM based detector along with the AUPRC values. (h) The per fold precision-recall curve for YOLOv3 detector along with the AUPRC values. (i) A comparison between the GradCAM based detector and YOLOv3 on the metrics of AUPRC, as well as recall and precision in the patient screening setting described in Methods. (Color figure online)

4 Conclusion and Discussion

This study demonstrates that the performance for localizing nGA lesions by only using OCT volume-wise classification labels with the GradCAM technique was on par with a fully supervised approach using B-scan level annotations with the YOLOv3 detector. These findings therefore underscore the potential of a weakly supervised approach for

enabling the development of a robust model for lesion localization without the need for laborious, lesion-level annotations on OCT B-scans.

One limitation of the GradCAM-based lesion localization is its relatively large bounding box size, often exceeding the annotated region. This is expected, considering the low spatial resolution of GradCAM saliency map, but also potentially because this approach identified contextual features that are distinguishing of nGA lesions that were not annotated by the graders. In addition, the weakly supervised model uses adaptive threshold of the saliency to determine the bounding box size, which was not optimized to match the ground truth grading. This limitation with the larger bounding box size could impact the quantification of the number of nGA lesions present, but it would unlikely have a substantial impact on the task of identifying a subset of OCT B-scans requiring manual review in an AI-assisted evaluation.

In conclusion, this study demonstrates that a weakly supervised method, requiring only volume-wise tags, can achieve a similar level of performance for localizing lesions compared to a fully supervised method using slice-wise bounding box labels. A weakly supervised approach could thus minimize the labeling burden when seeking to develop a lesion localization model, and could even leverage existing volume-wise labels for its development.

Acknowledgements. We would like to thank all of the study participants and their families, and all of the site investigators, study coordinators, and staff. We also appreciate the analysis support from biostatistician Ling Ma.

Funding. Supported by the National Health and Medical Research Council of Australia (project grant no.: APP1027624 [R.H.G.], and fellowship grant nos.: GNT1103013 [R.H.G.], #2008382 [Z.W.]; the BUPA Health Foundation (Australia) (R.H.G.) and the Macular Disease Foundation Australia (Z.W. and R.H.G.). The Centre for Eye Research Australia receives operational infrastructure support from the Victorian Government. Ellex R&D Pty Ltd (Adelaide, Australia) provided partial funding of the central coordinating center and the in-kind provision the Macular Integrity Assessment microperimeters for the duration of the LEAD study. The web-based Research Electronic Data Capture application and open- source platform OpenClinica allowed secure electronic data capture. The LEAD study was sponsored by the Centre for Eye Research Australia, East Melbourne, Australia, an independent medical research institute and a not-for-profit company. This sub-study was supported by Genentech, Inc.

Heming Yao, Adam Pely, Simon S. Gao, Hao Chen, Mohsen Hejrati, and Miao Zhang, are employees of Genentech, Inc. and shareholders in F. Hoffmann La Roche, Ltd.

References

1. Wu, Z., et al.: Optical coherence tomography-defined changes preceding the development of drusen-associated atrophy in age-related macular degeneration. Ophthalmology **121**(12), 2415–2422 (2014)
2. Wu, Z., et al.: Prospective longitudinal evaluation of nascent geographic atrophy in age-related macular degeneration. Ophthalmol. Retina **4**(6), 568–575 (2020)
3. Wu, Z., Guymer, R. H.: Can the onset of atrophic age-related macular degeneration be an acceptable endpoint for preventative trials?. ophthalmologica. J. Int. d'ophtalmologie. Int. J. Ophthalmol. Zeitschrift fur Augenheilkunde **243**(6), 399–403 (2020)

4. Derradji, Y., Mosinska, A., Apostolopoulos, S., Ciller, C., De Zanet, S., Mantel, I.: Fully-automated atrophy segmentation in dry age-related macular degeneration in optical coherence tomography. Sci. Rep. **11**(1), 21893 (2021)
5. Corradetti, G., et al.: Automated identification of incomplete and complete retinal epithelial pigment and outer retinal atrophy using machine learning. Investig. Ophthalmol. Vis. Sci. **63**(7), 3860 (2022)
6. Chiang, J.N., et al.: Automated identification of incomplete and complete retinal epithelial pigment and outer retinal atrophy using machine learning. Ophthalmol. Retina **7**(2), 118–126 (2023)
7. Yang, H.L., et al.: Weakly supervised lesion localization for age-related macular degeneration detection using optical coherence tomography images. PLoS ONE **14**(4), e0215076 (2019)
8. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
9. Shi, X., et al.: Improving interpretability in machine diagnosis: detection of geographic atrophy in OCT scans. Ophthalmol. Sci. **1**(3), 100038 (2021)
10. Yoon, J., et al.: Optical coherence tomography-based deep-learning model for detecting central serous chorioretinopathy. Sci. Rep. **10**(1), 18852 (2020)
11. Wang, Y., Lucas, M., Furst, J., Fawzi, A.A., Raicu, D.: Explainable deep learning for biomarker classification of OCT images. In: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), Cincinnati, OH, pp. 204–210 (2020)
12. Li, Y., et al.: Development and validation of a deep learning system to screen vision-threatening conditions in high myopia using optical coherence tomography images. Br. J. Ophthalmol. **106**(5), 633–639 (2022)
13. Guymer, R.H., et al.: Subthreshold nanosecond laser intervention in age-related macular degeneration: the lead randomized controlled clinical trial. Ophthalmology **126**(6), 829–838 (2019)
14. Wu, Z., Bogunović, H., Asgari, R., Schmidt-Erfurth, U., Guymer, R.H.: Predicting progression of age-related macular degeneration using OCT and fundus photography. Ophthalmol. Retina **5**(2), 118–125 (2021)
15. Guymer, R.H., et al.: Incomplete retinal pigment epithelial and outer retinal atrophy in age-related macular degeneration: classification of atrophy meeting report 4. Ophthalmology **127**(3), 394–409 (2020)
16. Carbonneau, M.-A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: a survey of problem characteristics and applications. Pattern Recog. **77**, 329–353 (2018)
17. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**, 62–66 (1979)
18. Liu, X., et al.: A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit. Health **1**(6), e271–e297 (2019)
19. Redmon, J., Farhadi, F.: YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767 (2018)