AMERICAN ACADEMY OF OPHTHALMOLOGY®

Improving Interpretability in Machine Diagnosis

Detection of Geographic Atrophy in OCT Scans

Xiaoshuang Shi, PhD,[1,] Tiarnan D.L. Keenan, MD,[2,*] Qingyu Chen, PhD,[1] Tharindu De Silva, PhD,[2] Alisa T. Thavikulwat, MD,[2] Geoffrey Broadhead, MD,[2] Sanjeeb Bhandari, MD,[2] Catherine Cukras, MD,[2] Emily Y. Chew, MD,[2] Zhiyong Lu, PhD[1]*

Purpose: Manually identifying geographic atrophy (GA) presence and location on OCT volume scans can be challenging and time consuming. This study developed a deep learning model simultaneously (1) to perform automated detection of GA presence or absence from OCT volume scans and (2) to provide interpretability by demonstrating which regions of which B-scans show GA.

Design: Med-XAI-Net, an interpretable deep learning model was developed to detect GA presence or absence from OCT volume scans using only volume scan labels, as well as to interpret the most relevant B-scans and B-scan regions.

Participants: One thousand two hundred eighty-four OCT volume scans (each containing 100 B-scans) from 311 participants, including 321 volumes with GA and 963 volumes without GA.

Methods: Med-XAI-Net simulates the human diagnostic process by using a region-attention module to locate the most relevant region in each B-scan, followed by an image-attention module to select the most relevant B-scans for classifying GA presence or absence in each OCT volume scan. Med-XAI-Net was trained and tested (80% and 20% participants, respectively) using gold standard volume scan labels from human expert graders.

Main Outcome Measures: Accuracy, area under the receiver operating characteristic (ROC) curve, F_1 score, sensitivity, and specificity.

Results: In the detection of GA presence or absence, Med-XAI-Net obtained superior performance (91.5%, 93.5%, 82.3%, 82.8%, and 94.6% on accuracy, area under the ROC curve, F_1 score, sensitivity, and specificity, respectively) to that of 2 other state-of-the-art deep learning methods. The performance of ophthalmologists grading only the 5 B-scans selected by Med-XAI-Net as most relevant (95.7%, 95.4%, 91.2%, and 100%, respectively) was almost identical to that of ophthalmologists grading all volume scans (96.0%, 95.7%, 91.8%, and 100%, respectively). Even grading only 1 region in 1 B-scan, the ophthalmologists demonstrated moderately high performance (89.0%, 87.4%, 77.6%, and 100%, respectively).

Conclusions: Despite using ground truth labels during training at the volume scan level only, Med-XAI-Net was effective in locating GA in B-scans and selecting relevant B-scans within each volume scan for GA diagnosis. These results illustrate the strengths of Med-XAI-Net in interpreting which regions and B-scans contribute to GA detection in the volume scan. *Ophthalmology Science 2021;1:100038 Published by Elsevier on behalf of the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Age-related macular degeneration (AMD) is a leading cause of vision loss in industrialized countries.[1] Late AMD has 2 forms, atrophic and neovascular; geographic atrophy (GA) is the defining lesion of atrophic disease.[2] Geographic atrophy is thought to affect more than 5 million people worldwide.[3] In GA, confluent atrophy of the retinal pigment epithelium (RPE) typically is accompanied by atrophy of adjacent photoreceptors and choriocapillaris and is associated with dense scotomata.[4,5] For this reason, central GA usually is accompanied by very poor visual acuity.[6] Geographic atrophy represents an important research priority because no treatments are available routinely in clinical practice to prevent its occurrence or to restore lost vision, although recent trials of local complement inhibition have shown some success in slowing down its enlargement rate.[7,8] In this context, rapid and accurate identification of eyes with GA could lead to improved clinical diagnosis and decision making, enhanced recruitment of eligible patients for future clinical trials, and eventually to early identification of appropriate patients for proven treatments.

https://doi.org/10.1016/j.xops.2021.100038
ISSN 2666-9145/21

A traditional clinical definition for GA (based on clinical examination or color fundus photography) has been a sharply demarcated, usually circular zone of partial or complete depigmentation of the RPE, typically with exposure of underlying large choroidal blood vessels, in the absence of neovascular changes in the same eye.[9] However, recent years have seen the advent of spectral-domain (SD) OCT[10] as an essential imaging method in ophthalmology. Spectral-domain OCT volume scans consist of a large number of 2-dimensional images (i.e., B-scans) captured in a raster pattern to form a cube. As a 3-dimensional imaging method, SD OCT has advantages over 2-dimensional methods such as color fundus photography that include detailed characterization of the multiple layers of the inner and outer retina at high resolution.[11,12] This allows atrophy to be studied in 3 dimensions and the involvement of specific retinal layers to be assessed quantitatively. Indeed, an international group of retinal experts has proposed SD OCT as the reference standard to diagnose GA.[13] The SD OCT term for GA is *complete RPE and outer retinal atrophy* (cRORA).

However, diagnosing cRORA on OCT scans sometimes can be challenging for ophthalmologists, particularly in cases of early disease, because each of 4 anatomic criteria (i.e., a region of hypertransmission of at least 250 μm in diameter in any lateral dimension, a zone of attenuation or disruption of the RPE of at least 250 μm in diameter, evidence of overlying photoreceptor degeneration, and absence of scrolled RPE or other signs of an RPE tear) must be met.[13] In addition, because each SD OCT volume scan consists of a large number of B-scans (often 100 or more), the diagnostic process can be time-consuming for ophthalmologists. In this context, automated deep learning approaches to the detection of AMD and GA could be highly useful.[14–18] Several previous approaches to semiautomated or automated detection of GA have been proposed.[13] Specifically, several segmentation algorithms[19–21] have used a partial summed voxel projection of the choroid, relying on the increased OCT signal intensity observed beneath Bruch's membrane in GA. With the development of deep neural networks, which can extract powerful feature representations automatically, various deep models—including convolutional neural networks (CNN),[22] sparse autoencoders,[23,24] 3-dimensional CNNs,[25] and adversarial networks[26]—have been applied to segment and classify GA. However, most of these models have required ground truth bounding boxes and typically used small datasets (e.g., fewer than 100 scans) for training and testing. Indeed, manually annotating bounding boxes is impractical for large datasets (e.g., more than 1000 volume scans, with each containing 100 or more B-scans). Moreover, none of those studies focused on model interpretability, a major barrier to the widespread adoption of automated machine learning systems in medical diagnosis and health care. However, other approaches to annotation have been used in some studies, with larger datasets in some cases. This includes GA annotation at the B-scan level,[27] which allows for GA segmentation, and annotation at the pixel level for features including those constituting cRORA.[28] To reduce annotation costs, deep multiple instance learning frameworks[29,30] first extract features from B-scans using volume-level labels for OCT classification and then use Gradient-weighted Class Activation Mapping (Grad-CAM)[31] or Class Activation Mapping (CAM)[32] to generate class activation maps for identifying significant regions in the B-scan. This process involves 2 separate stages, and it is different from the human diagnostic process that locates the most relevant regions in each B-scan, followed by selecting the most relevant B-scans for classifying GA presence or absence in each OCT volume scan. Recently, loss-based attention mechanisms[33,34] have been proposed to select significant regions and to classify images simultaneously and have demonstrated better model interpretability than methods with 2 separate stages, but they originally were designed for 2-dimensional natural images and cannot be applied directly to OCT scans.

To detect GA directly on OCT scans and also to provide interpretability for the decision making without any ground truth bounding boxes, we aimed to develop a novel deep learning algorithm to simulate the human diagnostic process and to perform several tasks simultaneously on an SD OCT volume scan: the localization of GA within each B-scan (if GA was present in the volume scan), the selection of the most representative B-scans in each volume scan, and the diagnosis of GA (i.e., presence or absence in the volume scan). Specifically, we aimed to train the deep learning algorithm to do this using only the semantic labels: the presence or absence of GA at the level of the entire OCT volume scan. As such, we proposed a novel CNN, namely Med-XAI-Net (explainable artificial intelligence for medical images analysis) (Fig 1). Importantly, Med-XAI-Net was designed to include 2 different loss-based attention modules, called image attention and region attention. The purposes of these 2 attention modules, respectively, were (1) to interpret the contribution of each B-scan in determining GA presence or absence in a volume scan and (2) to interpret the contribution of each region for locating GA in a B-scan. The specific aims of this study were (1) to assess the performance of the proposed framework on GA detection at the level of the entire OCT volume scan and (2) to validate the interpretation capability of the proposed framework by comparing its performance with that of ophthalmologists.

Methods

Image Datasets for Training and Testing

The dataset used for this study was from the Age-Related Eye Disease Study 2 (AREDS2) Ancillary SD OCT Study. The AREDS2 Ancillary SD OCT Study participants were a subset of the AREDS2 participants. The study designs and protocols for both studies were described previously.[35,36] The AREDS2 was a multicenter, phase 3, randomized controlled clinical trial designed to assess the effects of nutritional supplements on the course of AMD in people at moderate to high risk of progression to late AMD. It enrolled participants between the ages of 50 and 85 years with bilateral large drusen or large drusen in 1 eye and advanced AMD in the fellow eye. At baseline and annual study visits, comprehensive eye examinations were performed by certified study personnel using standardized protocols.

The AREDS2 Ancillary SD OCT Study enrolled AREDS2 participants from 4 study sites (Devers Eye Institute, Duke Eye Center, Emory Eye Center, and the National Eye Institute). The

study was approved by the institutional review boards of the 4 study sites and was registered at ClinicalTrails.gov (identifier, NCT00734487). It adhered to the tenets of the Declaration of Helsinki, and written informed consent was obtained from all participants. The participants underwent imaging using the Bioptigen Tabletop SD OCT system (Research Triangle Park, NC) at each annual study visit, as described previously.[35] For each eye, 6.7 × 6.7-mm SD OCT volume scans were captured (with 1000 A-scans per B-scan and 67-μm spacing between each B-scan). The ground truth grading of the SD OCT scans for the presence or absence of cRORA or GA was described previously.[13,37] In brief, the OCT scans were displayed with the Duke OCT Retinal Analysis Program and were graded independently by 2 human experts, with any disagreement adjudicated by another human expert. These grades (for the presence or absence of cRORA or GA at the level of each volume scan) provided the ground truth labels used for training and testing purposes in this study.

The SD OCT dataset consisted of 1284 volume scans from 311 participants (because participants contributed volume scans from multiple study visits over consecutive years, with a median number of 4 volume scans per participant). This comprised 321 volume scans with GA and 963 without GA. The dataset was split randomly into 2 independent subsets, the training subset (80%) and the testing subset (20%); this split was made at the participant level, so that all volume scans of each participant were in the same subset. Then, we randomly selected 10% participants of the training subset as a validation set. This process was repeated 10 times.

Composition of Med-XAI-Net

Med-XAI-Net (Fig 1) was designed as a deep learning model with 3 major parts: (1) a backbone network consisting of multiple convolutional layers to extract powerful feature representations from each B-scan; (2) a region-attention layer, which is an attention module using the parameters of the fully connected layer to select the region with GA in each B-scan; and (3) an image-attention layer, which is an attention module with the fully connected layer to select the B-scans that contributed most to the GA classification for the volume scan. In addition, we proposed a novel loss function to guarantee consistency between region or B-scan selection and volume scan classification during model training. The details of the backbone network, region-attention module, image-attention module, and loss function are shown in Appendix 1 (Supplemental information). Note that region-attention and image-attention use the same parameters to calculate the weights of regions and B-scans as used for classifying volume scans (i.e., loss-based attention).

We adopted a deep model, ResNet26 (which is derived from the state-of-the-art CNN ResNet50[38] and shown in Supplemental Table 1), as the backbone network. ResNet26 comprises 25 convolutional layers and 1 fully connected layer, comprising a total of more than 11 million parameters.

Before training, we scaled all images to a resolution of 224 × 224 pixels and augmented each image with random translations ($\{\Delta x, \Delta y\} \sim [-16, 16]$). We trained our model using the PyTorch platform. During the training stage, we updated the model parameters using the Adam optimizer for every minibatch of 1 volume scan (i.e., 100 B-scans) because labels were available only at the level of each volume scan. In total, we trained the model for 100 epochs using a learning rate of 10^{-4} for the first 50 epochs and 10^{-7} for the second 50 epochs. During the first 50 epochs, we used only the region-attention module for model training, while assigning each B-scan the same weight; during the second 50 epochs, we used both the region-attention and image-attention modules to update the model parameters. All experiments were conducted on a server with 48 Intel Xeon CPUs, using an NVIDIA GeForce GTX 1080 Ti 32Gb GPU for training and testing, with 754 Gb available in RAM memory.

Performance Evaluation and Comparison

To evaluate the performance of Med-XAI-Net on classifying GA presence or absence on SD OCT volume scans, we compared its performance with that of 2 popular methods: (1) Baseline, which is a baseline method that uses ResNet26 directly for volume scan classification and assigns equal weight to the B-scans in each volume scan and to the regions in each B-scan; (2) Inflated 3D Convnet (I3D),[39] which is a 3-dimensional CNN based on Inception[40]; and (3) AttentionNet, which adopts a state-of-the art attention mechanism,[41] namely, gated attention. Specifically, 2 gated-attention modules are embedded into ResNet26 for region and image selection, respectively (i.e., using a similar structure to that of Med-XAI-Net). Gated-attention uses different parameters to calculate the weights of regions and B-scans from those used for classification of volume scans. In this way, gated attention differs from the loss-based attention described above for Med-XAI-Net. For fairness, both Baseline and AttentionNet adopt the same experimental setting as Med-XAI-Net. Overall accuracy, area under receiver operating characteristic (ROC) curve, F_1 score, sensitivity, specificity, and precision-recall (PR) curves were used to evaluate the performance of Med-XAI-Net and the 2 comparative methods against the ground truth of human exert grading. All trained models used the softmax function to generate binary predictions (where the class with the highest output was selected).

To assess the performance of Med-XAI-Net in providing interpretability, that is, in selecting the relevant B-scans from a volume scan and localizing GA within a B-scan, we first randomly selected 100 different volume scans from the testing sets comprising 50 negative and 50 positive cases (i.e., we randomly selected 5 negative and 5 positive volume scans from each testing set, ensuring that no duplicates were present). For each trained model applied to each volume scan, the weights of each B-scan in a volume scan and of each region in a B-scan were recorded. Next, we selected the 5 B-scans with the highest weights in each volume scan and the region (of size 64 × 64 pixels) with the highest weight in each B-scan. Three ophthalmologists independently evaluated these selected B-scans and regions without access to any additional clinical information. Specifically, (1) to evaluate the performance of Med-XAI-Net in B-scan selection, the ophthalmologists recorded GA presence or absence in each volume scan based on the 5 B-scans selected, that is, recorded as GA present in the volume scan if at least 1 B-scan showed GA; and (2) to evaluate the performance of Med-XAI-Net in GA localization, the ophthalmologists recorded GA presence or absence in each B-scan based on the selected region, that is, more than half of the GA or cRORA should be in the selected region. If at least 2 of the ophthalmologists recorded GA as present in the volume scan, GA was recorded as present; if not, GA was recorded as absent.

Results

Performance of Med-XAI-Net in Identifying the Presence or Absence of Geographic Atrophy in Spectral-Domain OCT Volume Scans

The mean performance metrics of the 4 models in correctly classifying GA presence or absence from SD OCT volume scans are shown in Table 1. Med-XAI-Net achieved the highest overall accuracy (0.915), area under the ROC curve (0.935), F_1 score (0.823), and specificity (0.946) among the

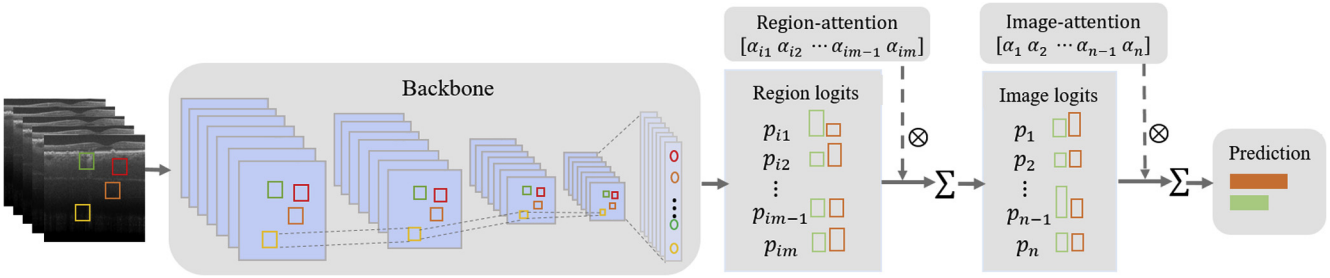

Figure 1. Overview of the proposed framework, Med-XAI-Net, for detecting geographic atrophy in a SD OCT cube scan by mining the relevant B-scans and regions. $[\alpha_{i1}\ \alpha_{i2}\ \ldots\ \alpha_{im-1}\ \alpha_{im}]$ denotes the weights of region logits $[p_{i1}\ p_{i2}\ \ldots\ p_{im-1}\ p_{im}]$ in the ith image, and $[\alpha_1\ \alpha_2\ \ldots\ \alpha_{n-1}\ \alpha_n]$ represents the weights of B-scan logits $[p_1\ p_2\ \ldots\ p_{n-1}\ p_n]$. m and n are the number of regions in each B-scan and the number of B-scans in each cube, respectively.

4 models. Its sensitivity (0.828) was intermediate between that of AttentionNet (0.796) and Baseline (0.853). Baseline obtained the highest sensitivity (0.853), but the lowest specificity (0.729). Additionally, I3D achieved a much better performance than Baseline, probably because 3-dimensional CNNs can capture more useful structural information than 2-dimensional CNNs. Figures 2 and 3 show the ROC and PR curves of the 4 models on GA classification at the level of volume scans, respectively.

Interpretability: the Performance of Med-XAI-Net in Identifying the Representative B-Scans and Localizing of Geographic Atrophy

Table 2 shows the performance metrics of the ophthalmologists in correctly classifying GA presence or absence at the level of the volume scan, based on either all B-scans, only the 5 B-scans selected by Med-XAI-Net, or only 1 region of 1 B-scan (both selected by Med-XAI-Net) using the testing set of 50 positive and 50 negative volume scans. Alongside these are shown the performance metrics of Med-XAI-Net on the same testing set based on all B-scans.

As shown in Table 2, the performance of the ophthalmologists was essentially identical when using either all B-scans or only the 5 B-scans selected by Med-XAI-Net as most representative. In the former case, the performance metrics were 96.0%, 95.7%, 91.8%, and 100.0% for accuracy, F_1 score, sensitivity, and specificity, respectively. In the latter case, the metrics were 95.7%, 95.4%, 91.2%, and 100.0%, respectively. The performance of the ophthalmologists was lower (driven by lower sensitivity) when using only the 1 region of 1 B-scan selected by Med-XAI-Net compared with using all B-scans or the 5 B-scans. In this case, the metrics were 89.0%, 87.4%, 77.6%, and 100.0%, respectively. The performance metrics of Med-XAI-Net on its own were 91.0%, 90.5%, 86.0%, and 96.0%, respectively. Figure 4 presents 3 representative examples of volume scans with GA present; in each case, the 5 B-scans selected by Med-XAI-Net from the volume scan are shown, as well as the region with GA selected by Med-XAI-Net from the B-scan.

Which Attention Module Is More Important?

Table 3 shows the performance of Med-XAI-Net in correctly classifying GA presence or absence according to 3 different versions of Med-XAI-Net: (1) region attention, that is, using only the region-attention layer; (2) image attention, that is, using only the image-attention layer; and (3) dual attention, that is, using both layers. Med-XAI-Net with region attention generally obtained almost the same performance metrics as Med-XAI-Net using dual attention. For all performance metrics, the absolute difference was less than 1%. Given the size of the test set and the overlapping 95% confidence intervals, these very small numerical differences do not seem meaningful. This strongly suggests that the region-attention module is a key contributor for correctly classifying the volume scans for GA presence or absence. However, because performance was not meaningfully lower with dual attention than with region attention, there seems to be no substantial tradeoff between performance and interpretability. Hence, with dual attention, interpretability is gained without meaningful loss of performance.

Table 1. Performance of Med-XAI-Net and 3 Comparative Methods on the Full Testing Sets of Spectral Domain OCT Scans

Method	Accuracy (95% Confidence Interval)	Area under the Receiver Operating Characteristic Curve (95% Confidence Interval)	F_1 Score (95% Confidence Interval)	Sensitivity (95% Confidence Interval)	Specificity (95% Confidence Interval)
Baseline	0.764 (0.727–0.802)	0.770 (0.732–0.808)	0.705 (0.673–0.737)	0.853 (0.827–0.879)	0.729 (0.684–0.774)
I3D	0.895 (0.875–0.919)	0.932 (0.915–0.949)	0.797 (0.759–0.835)	0.855 (0.815–0.897)	0.912 (0.885–0.937)
AttentionNet	0.858 (0.831–0.885)	0.876 (0.848–0.934)	0.752 (0.726–0.778)	0.796 (0.771–0.813)	0.880 (0.849–0.911)
Med-XAI-Net	0.915 (0.905–0.928)	0.935 (0.917–0.953)	0.823 (0.799–0.846)	0.828 (0.784–0.872)	0.946 (0.933–0.959)

I3D = Inflated 3D Convnet.

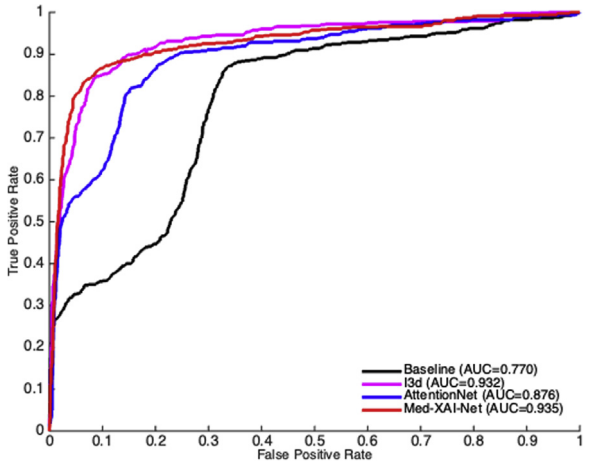

Figure 2. Receiver operator characteristic curves of 4 deep models on the full testing sets of spectral-domain OCT volume scans. AUC = area under the receiver operating characteristic curve; I3D = Inflated 3D Convnet.

Discussion

Table 1 demonstrates that the proposed framework, Med-XAI-Net, achieved superior performance to that of ResNet26 and AttentionNet. This is because Med-XAI-Net effectively can select relevant B-scans in each volume scan and relevant regions in each B-scan (as shown in Table 2), but Baseline and AttentionNet easily generate classification bias, for example, Baseline often misclassifies OCT scans without GA as having GA present (i.e., high false-positive rate), whereas AttentionNet sometimes misclassifies OCT scans with GA as not having GA (i.e., false-negative results). Additionally, I3D can obtain almost the same area under the ROC curve and superior sensitivity to Med-XAI-Net, and it has slightly worse performance on accuracy, F_1 score, specificity, and area under the PR curve. Regarding the relative performance of Med-XAI-Net and I3D, although

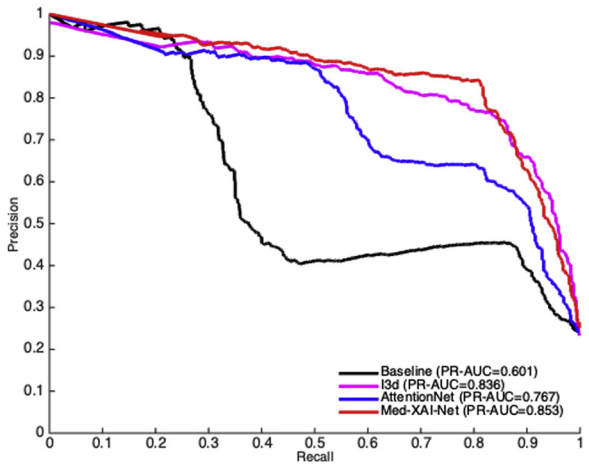

Figure 3. Precision-recall curves of 4 deep models on the full testing sets of spectral-domain OCT volume scans. I3D = Inflated 3D Convnet; PR-AUC = area under the precision-recall curve.

Table 2. Performance of Ophthalmologists on a Subset of the Full Testing Sets of Spectral Domain OCT Scans

Method	Data	Accuracy	F_1 Score	Sensitivity	Specificity
Ophthalmologists	Volume	0.960	0.957	0.918	1.00
	B-scans	0.957	0.954	0.912	1.00
	Region	0.890	0.874	0.776	1.00
Med-XAI-Net		0.910	0.905	0.860	0.960

The table shows the full volume scan (first row), 5 B-scans selected by Med-XAI-Net from the volume scan (second row), 1 region of 1 B-scan selected by Med-XAI-Net from the volume scan (third row), and performance of Med-XAI-Net on the full volume scan (fourth row). In total, 100 volume scans are from the testing sets, comprising 50 negative and 50 positive cases.

most of the performance metrics of I3D were numerically inferior, no statistically significant difference was found in the area under the ROC curve and area under the PR curve metrics of the 2 algorithms. In terms of interpretability, the proposed method is still more beneficial. Inflated 3D Convnet could be combined with Grad-CAM to provide interpretable predictions; however, in contrast to our proposed method, this would involve 2 separate stages: (1) training a model to extract features from volume scans for OCT classification and (2) using GradCAM to generate class activation maps to identify significant regions in the B-scan. Hence, it could not simulate the human diagnostic process (i.e., locating the most relevant region in each B-scan, followed by selecting the most relevant B-scans for classifying GA presence or absence in each OCT volume scan), thereby potentially decreasing model interpretability on decision making. Table 3 implies that region attention was a key contributor for Med-XAI-Net to classify OCT volume scans correctly. This means that locating GA was more relevant for GA diagnosis than B-scan selection and that image attention for B-scan selection is used primarily to reduce the workload of ophthalmologists with interpreting the significance of B-scans.

Several previous attention-based deep multiple instance learning methods,[41,42] which embed attention mechanisms using auxiliary layers into neural networks, also can be used for B-scan and region selection. However, their interpretation capability usually is inferior to that of loss-based attention mechanisms[33,34] because they use different parameters to calculate the weights of regions or B-scans from those used to classify the volume scans, potentially leading to inconsistency between B-scan or region selection and volume scan classification, that is, the selected region or B-scan does not contain GA for the volume scan with GA. Med-XAI-Net develops the loss-based attention mechanism[33,34] to maintain consistency between selection and classification. Importantly, unlike the previous loss-based attention mechanism that only contained one attention module for region selection, the proposed Med-XAI-Net used 2 different attention modules simultaneously for region and B-scan selection. Moreover, we proposed a new and different loss function (equation (3) in Appendix 1) in this work to connect the region or B-scan selection with the volume scan classification to maximize this consistency.

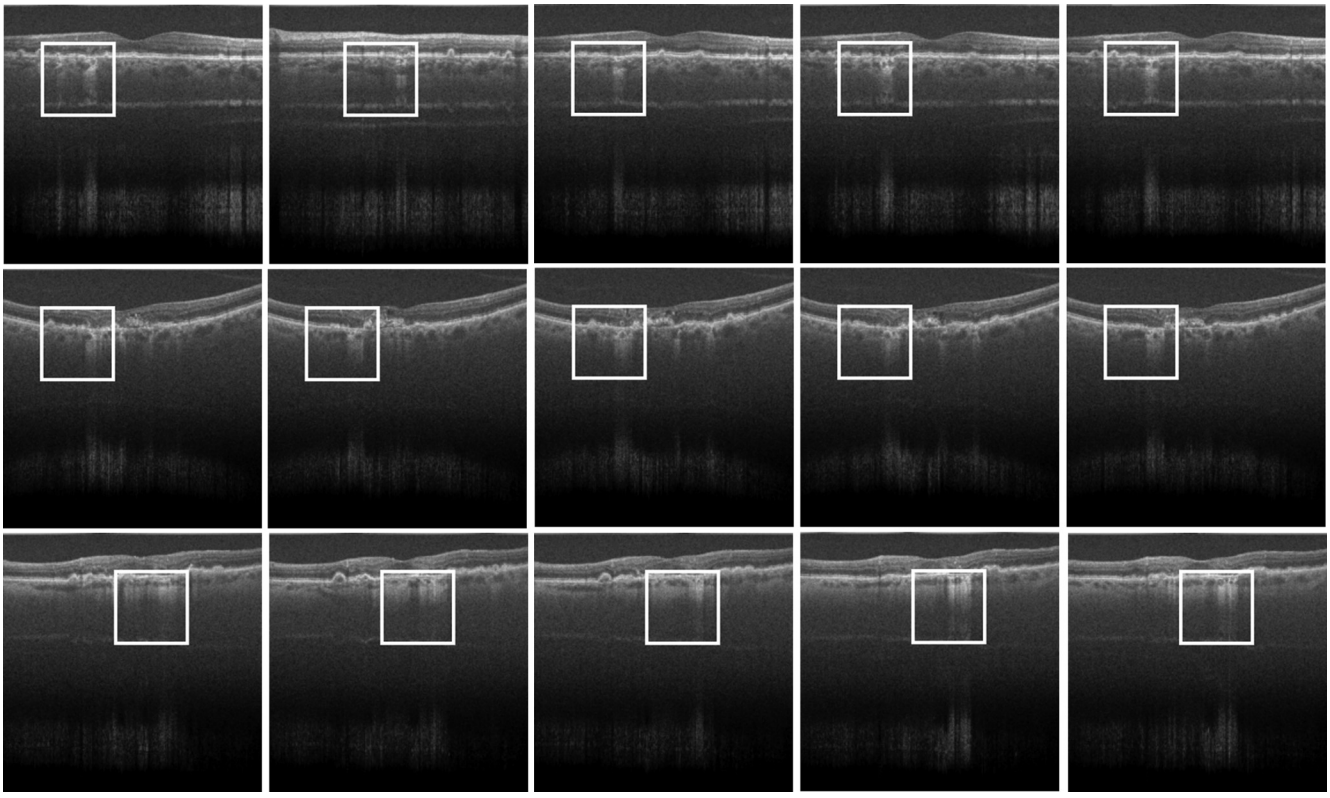

Figure 4. The selected images and located geographic atrophy (GA) by Med-XAI-Net. Each row represents 1 volume scan with 5 selected images, in which 1 box with a size of 64 × 64 is to locate GA.

The proposed method has potential clinical applicability in assisting physicians and other health care professionals with the detection of GA from OCT scans. For retinal specialists and general ophthalmologists in routine clinical practice, this may lead to increased accuracy and earlier diagnosis of GA (because detecting cRORA is sometimes challenging, particularly in early cases), as well as increased speed of workflow (because scrolling through many B-scans is time-consuming). In addition, the method's ability to localize GA to the most relevant B-scans (and affected regions) has several advantages: in addition to saving time, it assists the physician with disease characterization (e.g., assessing central involvement) and interpretability (to ensure the physician agrees with the diagnosis). For other health care professionals such as optometrists, detecting GA would prompt referral to retinal specialists; this will become increasingly important as therapies to slow GA enlargement become approved and widely available. In addition, outside routine clinical practice, the proposed method could be applied at scale to large datasets of OCT scans. This could be used, for example, to estimate the prevalence and incidence of GA in population-based or other epidemiologic studies or to identify individuals from datasets in a clinical setting who are eligible for a particular clinical trial or a licensed therapy.

This potential clinical applicability may differ partially from that of methods described in some recent studies. For example, the studies described above, based on GA annotations at the B-scan level[27] or AMD-related features at the pixel level,[28] may have overlapping but partially distinct potential clinical applicability. In particular, these methods allow for GA segmentation or quantification, which is advantageous in

Table 3. Ablation Study of Med-XAI-Net on Spectral Domain OCT Volume Scans

Method	Accuracy (95% Confidence Interval)	Area under the Receiver Operating Characteristic Curve (95% Confidence Interval)	F_1 Score (95% Confidence Interval)	Sensitivity (95% Confidence Interval)	Specificity (95% Confidence Interval)
Region attention	0.920 (0.904–0.937)	0.942 (0.925–0.958)	0.829 (0.800–0.858)	0.820 (0.783–0.857)	0.953 (0.942–0.964)
Image attention	0.732 (0.621–0.842)	0.791 (0.700–0.890)	0.584 (0.439–0.730)	0.704 (0.602–0.805)	0.734 (0.616–0.860)
Dual attention*	0.915 (0.905–0.928)	0.935 (0.917–0.953)	0.823 (0.799–0.846)	0.828 (0.784–0.872)	0.946 (0.933–0.959)

*Denotes Med-XAI-Net using both region-attention and image-attention layers.

some clinical or research scenarios. For example, they may allow estimation of GA enlargement over time (an important end point in clinical trials), and so may be particularly useful in the reading center and academic settings. Overall, we consider that these 2 sets of methods may be complementary in their applicability.

Limitations and Future Work

One potential limitation of Med-XAI-Net arises from the imbalance of the volume scans in the dataset, where volume scans without GA outnumber those with GA. This may have contributed to the relatively lower sensitivity (Table 1). This limitation might be addressed by using 2 different augmentation methods (e.g., the translation method used in this study, together with another method, such as AutoAugment[43] with rotation and shearing) to augment volume scans and increase the number of volume scans with GA or by generating soft labels for volume scans with GA so as to leverage their diagnosis information.

The second limitation of Med-XAI-Net is relatively worse performance on GA localization (demonstrated by the lower performance of ophthalmologists on classifying GA presence or absence from 1 region selected by the model). One contributing factor may be that, in these relatively large B-scan images, areas without GA are likely to outnumber substantially those with GA, which makes this task relatively difficult. Additionally, areas without GA may be learned easily and unconsciously as noisy features, which could lower the GA classification performance. One possible solution is to use preprocessing image techniques to enhance the most relevant regions for GA detection or denoising methods to alleviate the effects of irrelevant regions before feeding images to the network.

Finally, Med-XAI-Net is validated only for GA diagnosis using SD OCT volume scans in this work because of the limitation of available data sources. It would be interesting in future studies to apply Med-XAI-Net to different diseases and various imaging methods to investigate its potential generalization capability further. We also recommend that other groups evaluate Med-XAI-Net for GA diagnosis from SD OCT volume scans from multiple diverse sources.

An important aspect in the future will be optimization of the operating point, based on the ROC curve, to maximize clinical applicability for particular task and setting. Additionally, in the future, it would be promising to extend Med-XAI-Net to 3-dimensional CNNs to explore the possibility of obtaining better classification and interpretability performance. Also, we aim to improve the model performance further by incorporating multiple other techniques, such as augmentation, soft labels, and preprocessing image techniques. Moreover, we plan to apply Med-XAI-Net to AMD diagnosis based on multimodal imaging, that is, by using SD OCT scans and color fundus photographs in combination.

Footnotes and Disclosures

Originally received: March 12, 2021.
Final revision: July 2, 2021.
Accepted: July 2, 2021.
Available online: July 13, 2021. Manuscript no. D-21-00041.

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland.

[2] Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, Maryland.

*Both authors contributed equally as first authors.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s): T.D.L.K.: Patent — Coinventor on a patent application "Methods and Systems for Predicting Rates of Progression of Age-Related Macular Degeneration"

This work was supported by the Intramural Research Program of the National Library of Medicine and the National Eye Institute, National Institutes of Health, Bethesda, Maryland.

Emily Chew and Catherine Cukras, members of the editorial board of this journal, were recused from the peer-review process of this article and had no access to information regarding its peer-review.

HUMAN SUBJECTS: Human subjects were included in this study. The human ethics committees at Devers Eye Institute, Duke Eye Center, Emory Eye Center, and the National Eye Institute approved the study. All research adhered to the tenets of the Declaration of Helsinki. All participants provided informed consent.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Chew, Lu

Analysis and interpretation: Shi, Keenan, Chen, Thavikulwat, Broadhead, Bhandari, Cukras

Data collection: Shi, Keenan, De Silva, Chew

Obtained funding: N/A; Study was performed as part of regular employment duties at National Center for Biotechnology Information, National Library of Medicine, National Instistutes of Health, and Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health. No additional funding was provided.

Overall responsibility: Shi, Keenan, Lu

Abbreviations and Acronyms:

AMD = age-related macular degeneration; **AREDS2** = Age-Related Eye Disease Study 2; **AUC** = area under curve; **CAM** = class activation mapping; **CFP** = color fundus photograph; **CNN** = convolutional neural network; **cRORA** = complete retinal pigment epithelium and outer retinal atrophy; **GA** = geographic atrophy; **Grad-CAM** = gradient-weighted class activation mapping; **I3D** = Inflated 3D Convnet; **PR** = precision-recall; **PR-AUC** = area under PR curve; **ROC** = receiver operating characteristic; **RPE** = retinal pigment epithelium; **SD** = spectral-domain; **XAI** = explainable artificial intelligence.

Keywords:
Deep learning, GA detection, Interpretable, OCT.

Correspondence:
Emily Y. Chew, MD, National Eye Institute, National Institutes of Health, 9000 Rockville Pike, Bethesda, MD 20892. E-mail: echew@nei.nih.gov; and Zhiyong Lu, PhD, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894. E-mail: zhiyong.lu@nih.gov.

References

1. Steinmetz JD, Bourne RR, Briant PS, et al. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. *Lancet Glob Health*. 2021;9(2): e144–e160.
2. Blair CJ. Geographic atrophy of the retinal pigment epithelium: a manifestation of senile macular degeneration. *Arch Ophthalmol*. 1975;93:19–25.
3. Wong WL, Su X, Li X, 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*. 2014;2(2):106–116.
4. Bhutto I, Gerard L. Understanding age-related macular degeneration (AMD): relationships between the photoreceptor/retinal pigment epithelium/Bruch's membrane/choriocapillaris complex. *Mol Aspects Med*. 2012;33:295–317.
5. Guymer RH, Rosenfeld PJ, Curcio CA, et al. Incomplete retinal pigment epithelial and outer retinal atrophy in age-related macular degeneration: classification of atrophy meeting report 4. *Ophthalmology*. 2020;127:394–409.
6. Shen LL, Sun M, Ahluwalia A, et al. Relationship of topographic distribution of geographic atrophy to visual acuity in nonexudative age-related macular degeneration. *Ophthalmol Retina*. 2020 Nov 17;S2468-6530(20):30446–2. https://doi.org/10.1016/j.oret.2020.11.003. Online ahead of print.
7. Ammar MJ, Hsu J, Chiang A, et al. Age-related macular degeneration therapy: a review. *Curr Opin Ophthalmol*. 2020;31(3):215–221.
8. Shen LL, Sun M, Ahluwalia A, et al. Natural history of central sparing in geographic atrophy secondary to non-exudative age-related macular degeneration. *Br J Ophthalmol*; 2020 Dec 23; bjophthalmol-2020-317636. https://doi.org/10.1136/bjophthalmol-2020-317636. Online ahead of print.
9. Age-Related Eye Disease Study Research Group. The Age-Related Eye Disease Study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: the Age-Related Eye Disease Study report number 6. *Am J Ophthalmol*. 2001;132(5):668–681.
10. De Boer JF, Cense B, Park BH, et al. Improved signal-to-noise ratio in spectral-domain compared with time-domain optical coherence tomography. *Opt Lett*. 2003;28: 2067–2069.
11. Sadda SR, Guymer R, Holz FG, et al. Consensus definition for atrophy associated with age-related macular degeneration on OCT: classification of atrophy report 3. *Ophthalmology*. 2018;125(4):537–548.
12. Spaide RF, Jaffe GJ, Sarraf D, et al. Consensus nomenclature for reporting neovascular age-related macular degeneration data: consensus on neovascular age-related macular degeneration nomenclature study group. *Ophthalmology*. 2020;127(5): 616–636.
13. Arslan J, Samarasinghe S, Benke KK, et al. Artificial intelligence algorithms for analysis of geographic atrophy: a review and evaluation. *Transl Vis Sci Technol*. 2020;9(2):57.
14. Keenan TD, Dharssi S, Peng Y, et al. A deep learning approach for automated detection of geographic atrophy from color fundus photographs. *Ophthalmology*. 2019;126(11): 1533–1540.
15. Peng Y, Dharssi S, Chen Q, et al. DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*. 2019;126(4):565–575.
16. Chen Q, Peng Y, Keenan TD, et al. A multi-task deep learning model for the classification of age-related macular degeneration. *AMIA Jt Summits Trans Sci Proc*. 2019;(2019):505–514.
17. Chen Q, Keenan TD, Allot A, et al. Multi-modal, multi-task, multi-attention (M3) deep learning detection of reticular pseudodrusen: 1 towards automated and accessible classification of age-related macular degeneration. *J Am Med Inform Assoc*. 2021;28(6):1135–1148.
18. Keenan TD, Chen Q, Peng Y, et al. Deep learning automated detection of reticular pseudodrusen from fundus autofluorescence images or color fundus photographs in AREDS2. *Ophthalmology*. 2020;127(12):1674–1687.
19. Chen Q, de Sisternes L, Leng T, et al. Semi-automatic geographic atrophy segmentation for SD-OCT images. *Biomed Opt Exp*. 2013;4:2729–2750.
20. Hu Z, Medioni GG, Hernandez M, et al. Segmentation of the geographic atrophy in spectral-domain optical coherence tomography and fundus autofluorescence images. *Invest Ophthalmol Vis Sci*. 2013;54:8375–8383.
21. Niu S, de Sisternes L, Cheng Q, et al. Automated geographic atrophy segmentation for SD-OCT images using region-based CV model via local similarity factor. *Biomed Opt Exp*. 2016;7: 581–600.
22. Fang L, Cunefare D, Wang C, et al. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomed Opt Exp*. 2017;8(5):2732–2744.
23. Ji Z, Chen Q, Niu S, et al. Beyond retinal layers: a deep voting model for automated geographic atrophy segmentation in SD-OCT images. *Transl Vis Sci Technol*. 2018;7:1.
24. Xu R, Niu S, Chen Q, et al. Automated geographic atrophy segmentation for SD-OCT images based on two-stage learning model. *Comput Biol Med*. 2019;105:102–111.
25. Xu R, Niu S, Gao K, Chen Y. Multi-path 3D convolution neural network for automated geographic atrophy segmentation in SD-OCT images. In: International Conference on Intelligent Computing. Cham: Springer; 2018:493–503.
26. Wu M, Cai X, Chen Q, et al. Geographic atrophy segmentation in SD-OCT images using synthesized fundus autofluorescence imaging. *Comput Methods Programs Biomed*. 2019;182, 105–101.
27. Liefers B, Colijn JM, Gonzalez-Gonzalo C, et al. A deep learning model for segmentation of geographic atrophy to study its long-term natural history. *Ophthalmology*. 2020;127(8):1086–1096.
28. Liefers B, González-Gonzalo C, Klaver C, et al. Dense segmentation in selected dimensions: application to retinal optical coherence tomography. International Conference on Medical Imaging with Deep Learning. *PMLR*. 2019:337–346.
29. Das V, Prabhakararao E, Dandapat S, Bora PK. B-Scan attentive CNN for the classification of retinal optical coherence tomography volumes. *IEEE Signal Process Lett*. 2020;27: 1025–1029.
30. Wang X, Tang F, Chen H, et al. UD-MIL: uncertainty-driven deep multiple instance learning for OCT image classification. *IEEE J Biomed Health Inform*. 2020;24(12):3431–3442.

31. Selvaraju RR, Cogswell M, Das A, Vedantam R, et al. Grad-cam: visual explanations from deep networks via gradient-based localization. *IEEE International Conference on Computer Vision*. 2017:618−626.
32. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. *IEEE Conference on Computer Vision and Pattern Recognition*. 2016:2921−2929.
33. Shi X, Xing F, Xie Y, et al. Loss-based attention for deep multiple instance learning. *AAAI Conference on Artificial Intelligence*. 2020;34(04):5742−5749.
34. Shi X, Xing F, Xu K, et al. Loss-based attention for interpreting image-level prediction of convolutional neural networks. *IEEE Trans Image Process*. 2021;30:1662−1674.
35. Leuschen JN, Schuman SG, Winter KP, et al. Spectral-domain optical coherence tomography characteristics of intermediate age-related macular degeneration. *Ophthalmology*. 2013;120(3): 140−150.
36. AREDS2 Research Group, Chew EY, Clemons T, SanGiovanni JP, et al. The age-related eye disease study 2 (AREDS2): study design and baseline characteristics (AREDS2 report number 1). *Ophthalmology*. 2012;119(11):2282−2289.
37. Christenbury JG, Folgar FA, O'Connell RV, et al. Progression of intermediate age-related macular degeneration with proliferation and inner retinal migration of hyperreflective foci. Ophthalmology. 2013;120(5):1038−1045.
38. Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. *IEEE Conference on Computer Vision and Pattern Recognition*. 2017:6299−6308.
39. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning. 2015:448−456.
40. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*. 2016:770−778.
41. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. *International Conference on Machine Learning*. 2018:2127−2136.
42. Pappas N, Andrei P. Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research*. 2017;58:591−626.
43. Cubuk ED, Zoph B, Mane D, et al. Autoaugment: learning augmentation strategies from data. *IEEE Conference on Computer Vision and Pattern Recognition*. 2019:113−123.