APPENDIX

Image Masking

Image masking was attempted to remove or minimize eyelash and eyelid artifacts present in many of the ultra-widefield (UWF) images. The U-Net convolutional network for biomedical image segmentation[1] was applied to the UWF color and UWF fundus autofluorescence (FAF) images. The segmentation network uses a U-Net architecture to translate the raw UWF color and UWF FAF images into a mask map with two classes including the image for analysis and a portion discarded as noise. The U-Net model was trained with 20 UWF color images from the cognitively healthy control cohort that had undergone manual image masking to remove eyelash and eyelid artifacts. Manually masked images were binary images, with noise pixels (eyelid and eyelash artifacts) set to 0, and others pixels set to 1. The area under the curve (AUC) of the automated segmentation task was 0.970 when compared to a sample of 20 manual segmentations based on the leave-one-out cross-validation (LOOCV).[2] The results of our convolutional neural network (CNN) for prediction of a symptomatic Alzheimer's disease (AD) diagnosis were compared with and without the image masking. There was no improvement in performance in the validation or test sets when image masking was used. Thus, image masking was not used in the results reported in this study.

Image Cropping

Results of the combination model incorporating all retinal imaging modalities, OCT and OCTA quantitative data, and patient data were compared with and without cropping of UWF color images. With the entire UWF image incorporated, the model using all images and all data inputs achieved an AUC of 0.854 [95% CI: 0.717, 0.990] on the validation set and 0.826 [95% CI: 0.715, 0.937] on the test set, while use of cropped images allowed the model to achieve an AUC of 0.854 [95% CI: 0.718, 0.991] on the validation set and 0.836 [95% CI: 0.729, 0.943] on the test set. Thus, image cropping slightly improved the predictive value of the model. The model used cropped UWF color and UWF FAF images.

Image Resizing

All images were resized prior to use in the model. Images were resized to 128x128 pixels from 1800x1800 for UWF color and FAF, 409x409 – 450x450 for OCTA, and 281x281 for ganglion cell-inner plexiform layer (GC-IPL) maps. This resizing resulted in 7x7x64 convolutional features that, when flattened out, are represented as 3136-dimensional vectors. While such resizing prevents fine details from being preserved, sample size limitations did not allow us to preserve all of the fine details of each imaging modality. While model architectures with larger images were trialed, these architectures were unable to obtain significant performance improvements during the development phase, quantified on the validation set. Thus, while finer details have the potential of greatly improving model performance, we would need a significantly larger sample size to seize the benefits of such resolution gain.

Attention Maps

To interpret the classification result made by the model, attention maps were generated to visualize the discriminative image regions used by the model to distinguish symptomatic AD subjects from controls using class activation mappings (CAMs).[3] After the model was fully trained, we fed the images, including OCTA images and UWF color and UWF FAF images, into the image feature extractor and created the feature maps for each imaging modality, f_{OCTA}, $f_{\text{Color and FAF}}$, and $f_{\text{GC-IPL}}$ (Figure 1). We then projected back the weights of the corresponding fully connected layer onto the feature maps to produce the CAMs. The final attention maps were generated by scaling the CAMs to a heatmap with the same size as the input images. Figures 2 and 3 include examples of the OCTA attention maps that were generated. Supplemental Figure 2 contains attention map examples of the UWF images.

Model Structure

The image feature extractor used in this study has a similar structure to the first five layers of the ResNet18[4] neural network. The model included a total of 166,936 parameters. More specifically, the CNN-based feature extractor consists of 157,504 parameters (1 layer of 64 7x7x3 filters, 4 layers of 64 3x3x64 filters, and 5 layers of batch normalizations of 64 dimensions). The fully connected layers that output the modality-wise pre-classification probabilities consist of 9,432 parameters (3,137 weights each for UWF color and FAF, OCTA, and GC-IPL maps, and 21 weights for the quantitative features).

Two primary differences between our model and the original ResNet18[4] are: 1) Stride size of the first convolutional layer was changed from 2 to 1. As input images are compressed from large images, we found that a stride size of 2 in the first layer may lose some detailed information in the compressed images; 2) Pooling size was changed from 2x2 to 4x4 to reduce the dimension of output feature maps, which leads to fewer parameters in the following fully-connected (FC) layers and mitigates over-fitting. As our feature extractor has the same structure with ResNet18, it can still be initialized by ResNet18 pre-trained on ImageNet.[5] The detailed structure of the image feature extractor can be found in Supplemental Figure 1.

A significant concern with training a CNN with a dataset of this size is over-fitting. To mitigate over-fitting, we employed the following technologies.

1）Model Architecture

Given the limitation of our relatively smaller training dataset, we used only the first five layers of the ResNet18[4] as the feature extractor and discarded other layers. The number of layers is typically chosen based on performance of the training and validation sets after assessing different layer configurations. At the same time, we shared the feature extractor across different modalities, which increases the amount of data available to train the feature extraction component of the model (i.e. the CNN model). Provided that different imaging modalities have their own idiosyncrasies, feeding them through the same feature extractor seems counterintuitive. However, noting that modality differences are predominantly high-level image characteristics, we customized the feature extractor for each modality by appending a modality-specific FC layer at the output of the CNN-based feature extractor, as shown in Figure 1. As a result, these FC layers learned the unique characteristics of each modality. Given that the feature extractor contained only five layers, it is likely to be extracting general image features that are then adapted for each modality via FC layers. After passing through the FC layers, pre-classification signals were generated separately for each imaging modality describing how likely an eye carried an AD diagnosis. We also increased the average pooling from 2x2 to 4x4 (the “avg pool, ¼” by 4 in Supplemental Figure 1) between different convolutional layers to help reduce the dimensions of the image feature extractor’s outputs, which led to fewer parameters for the corresponding FC layers. Finally, given the dataset size, we were aware that overloading the model with excess input data from each subject could also cause over-fitting. Thus, providing the model with the entire set of volumetric OCT images would not likely improve performance. In light of this, quantitative OCT data that summarized the volumetric OCT findings with fewer data points was used in our study. In future studies, using volumetric OCT images may be further explored; however, given that our model achieved similar AUC values with inputs of all images only and inputs including quantitative OCT data, it remains unclear if volumetric data would improve performance.

2) L2 and L1 Regularization

L2 regularization is a technique where the sum of squared parameters of a model (multiplied by a coefficient) is added into the loss function as a penalty term to be minimized. It tends to cause the learning algorithm to perceive the input as having higher variance, which makes it shrink the weights on features with low covariance with the target label.[6,7] L1 regularization adds the sum of the absolute values of the individual parameters to the loss function. In comparison to L2 regularization, L1 regularization results in a solution that is sparser, which is consistent with the fact that some parameters may have an optimal value of zero. Thus, L1 regularization has also been used for feature selection.[6] Both L1 and L2 regularization make it difficult for the regularized network to learn local noise in the dataset and force the network to learn only those features which are often seen across the training set. According to the properties of L2 and L1 regularization, we added L2 regularization to the entire model, and L1 regularization on the FC layers to the final loss function to limit the capacity of the model and prevent over-fitting.[8,9] We tested different weights from 0.001 to 10 for the regularization loss and found 0.01 worked best. We also tried dropout with different dropout rates without benefit.

3) Data Augmentation

Data augmentation can be used to reduce over-fitting and increase the amount of training data. It creates new images by transforming (rotating, translating, scaling, flipping, distorting) and adding some noise (e.g. Gaussian noise) to the images in the training dataset. Both the original image and the created images are fed into the neural network. In our model, we increased the diversity of images for training models through rotating, shifting, cropping, and zooming images.

4) Transfer Learning and Fine-tuning

We chose to initialize parameters of our feature extractor with weights of the pre-trained ResNet[4] models to aid in the classification task. This model was then fine-tuned with our labeled imaging data. This method transferred the knowledge from general large-scale image datasets (ImageNet)[5] to our task. Gulshan et al.[10] also used this method to speed up the training of a deep learning model, which was designed to detect diabetic retinopathy in retinal fundus photographs, although they had a relatively larger dataset.

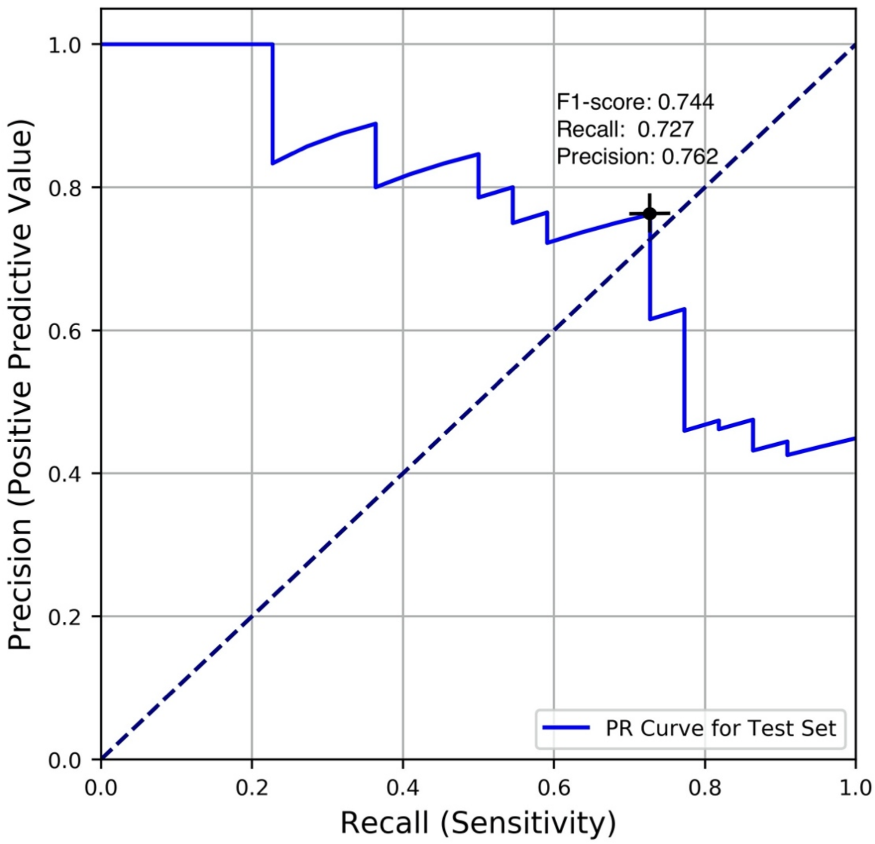
Inter-eye Correlations

We acknowledge that, when evaluating each eye separately from each patient in the test set as we have done herein, inter-eye correlation can result in inflated performance metrics. As a result, we recalculated the performance metrics by repeatedly sampling a single eye for each of the 34 subjects in the test set. We reviewed AUC summaries for 300, 500, 1000, and 10,000 samples, each of which demonstrated stable means and standard deviations. For 10,000 samples, calculated considering data from both eyes of each patient, the “patient-wise” AUC for the best performing model was 0.84 (SD 0.034). This result is very similar to the results in Table 1, which consider only a single eye for each patient and demonstrate the “eye-wise” AUC. Thus, in our case, inter-eye correlation does not seem to affect performance estimates.

Performance Reporting

We presented AUC as the main performance metric because it is widely used and accepted. However, we recognize that AUC can sometimes be misleading in situations where there is an imbalance between cases and controls. In Appendix Figure 1, we include the performance recall curve (AUPRC) for the test set for the best performing model, including GC-IPL maps, quantitative data, and patient data. Precision in classification tasks is also referred to as positive predictive value (PPV). Recall in classification tasks is also referred to as true positive rate (TPR) or sensitivity. We chose the optimal point according to F1-score, which combines precision and recall into one metric by calculating the harmonic mean between the two. The optimal point is achieved when the threshold for probability of AD, P(AD), is set to 0.224. The model achieves F1-score of 0.744, recall 0.727, and precision 0.762.

Appendix Figure 1



Age Differences in AD and Control Subjects

Supplemental Table 1 demonstrates that AD subjects in our study were significantly older (2.73 years on average) than control subjects (p=0.015). To address any concern that the model might simply be detecting the age difference between the two groups, we performed experiments using age directly to predict AD using a decision tree model (to account for nonlinear effects) using the same data partitions we used for the model (training, validation, and test sets). Results for tree models of varying depths (1-10) shown in Appendix Table 1 below indicate that the performance of age only (depth=5), 0.583 validation AUC, and 0.528 test AUC, is considerably lower than that of the image-based model.

Appendix Table 1

Max Depth	Validation Set Accuracy	Validation Set AUC	Test Set Accuracy	Test Set AUC
1	0.733	0.458	0.618	0.480
2	0.800	0.500	0.647	0.502
3	0.733	0.458	0.618	0.480
4	0.733	0.458	0.647	0.549
5	0.733	0.583	0.618	0.528
6	0.667	0.542	0.588	0.482
7	0.667	0.542	0.588	0.482
8	0.667	0.542	0.588	0.482
9	0.667	0.542	0.559	0.460
10	0.667	0.542	0.559	0.460

REFERENCES

1. Ronneberger O, Fischer P, Brox T. *U-Net: Convolutional Networks for Biomedical Image Segmentation.* Vol 9351: Springer, Cham; 2015.
2. Molinaro A, Simon R, Pfeiffer R. Prediction error estimation: a comparison of resampling methods. *Bioinformatics.* 2005;21(15):3301-3307.
3. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016:2921-2929.
4. He K, Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016:770-778.
5. Deng J, Dong W, Socher R, Li-Jia L, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition.* 2009.
6. Goodfellow I, Bengio Y, Courville A. *Deep Learning.* Cambridge, MA: The MIT Press; 2016.
7. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations (ICLR).* 2017;arXiv:1611.03530.
8. Bühlmann P, van de Geer S. *Statistics for High-Dimensional Data.* Vol 1. Heidelberg: Springer-Verlag Berlin Heidelberg; 2011.

9. Loshchilov I, Hutter F. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR).* 2019;arXiv:1711.05101.
10. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410.