Supplemental Material: Annu. Rev. Vis. Sci. 2026. 12: https://doi.org/10.1146/annurev-vision-110423-025837

Supplemental Text

1. Optical Coherence Tomography Technology

The advent of Optical Coherence Tomography (OCT) ophthalmic imaging provided a non-contact method for acquisition of optical cross-sections of the retina and optic nerve using scanning lasers of safe energies (Huang et al., 1991). Interferometry of short coherence length light scanned across the retina in a sweeping pattern enables the creation of two-dimensional images of retinal tissue sections by computation from the returning reflectance signals. These cross-sectional images are referred to as B scans. The years since the release of the first time-domain ophthalmic imaging systems for clinical use have seen extraordinary progress in the technology and resulting imaging resolution. Current commercially available ophthalmic OCT systems typically achieve 5um axial imaging resolution, allowing analysis of retinal structure in exquisite detail. Alignment of many adjacent two-dimensional sections permits the construction of three-dimensional representations of imaged tissue and volumetric analysis (Aumann S et al., 2019; Huang et al., 1991). Further subtraction analysis of flow voids left by reflections from blood constituents moving through blood vessels additionally permits the construction of three-dimensional models of the vascular networks in living retinal tissue, a technique known as OCT Angiography (OCT-A) (Aumann S et al., 2019).

2. Machine Learning, Deep Learning, and Convolutional Neural Networks

Since its inception in 1956, artificial intelligence (AI) has evolved as a promising group of technologies capable of complex information-processing tasks previously believed to be the exclusive cognitive domain of humans (Kaplan and Haenlein, 2019). Machine learning (ML) is a crucial subset of AI and refers to the ability of computer systems to learn from data, identify patterns, and make decisions on their own without being explicitly programmed. Various algorithms have been proposed for ML, including support vector machine (SVM), decision trees, and neural networks. Inspired by the human brain, neural networks involve several inter-connected processing units, named neurons, organized in consecutive layers, including the input layer, hidden layer(s), and the final output layer. Each artificial neuron processes input signals through mathematical operations and sends the output to the following layer. In simple terms, deep neural networks are essentially neural networks with three or more hidden layers, giving rise to the term deep learning (DL) which is now widely utilized in various applications, including natural language processing, speech recognition, and computer vision. Convolutional neural networks (CNNs) are DL-based models that have gained much attention in recent years for their astonishing performance in computer vision tasks. The design of CNNs mimics the human visual cortex, with low-level neurons detecting simple features like corners and edges, and high-level neurons capturing complex patterns such as shape and texture. CNNs typically consist of three major parts, namely, convolutional layers, pooling layers, and fully-connected layers. An easy-to-understand review of the fundamentals of ML, DL, and CNNs is provided in the work by Alzubaidi et al. (2021) (Alzubaidi et al., 2021).

3. Brief Description of the Machine Learning Studies Aimed at Classifying Multiple Sclerosis Using Retinal Images

OCT studies

Spectral-Domain OCT Studies

In SD-OCT studies, the parameters employed for MS classification were total macular volume (Khodabandeh et al., 2023) and thickness of the total retina (Arian et al., 2024; Ciftci Kavaklioglu et al., 2022), RNFL (Ciftci Kavaklioglu et al., 2022; Garcia-Martin et al., 2015, 2013; Hernandez et al., 2023; Khodabandeh et al., 2023; Montolío et al., 2022, 2021; Ortiz et al., 2023), GCIPL (Ciftci Kavaklioglu et al., 2022; Khodabandeh et al., 2023; Montolío et al., 2024; Zhang et al., 2020), GCL (Hernandez et al., 2023; Ortiz et al., 2023), and other outer retinal layers (Khodabandeh et al., 2023; Ortiz et al., 2023), alone or in combination, obtained from peripapillary, macular, and posterior pole protocols. Some studies also considered participants' clinical and/or demographic data (Kenney et al., 2022; Montolío et al., 2024, 2022, 2021).

Garcia Martin et al. (2013) developed a simple neural network for classifying 106 MS patients and 115 HCs using 768 pRNFL thickness points distributed in 24 sectors, each of which had an equal size of 15 degrees of angle (360° / 24 = 15°). Their model showed an area under the receiver operating characteristics (AUC) of 0.945 (Garcia-Martin et al., 2013). In a similar study, using 16 uniform-sized groups (each consisting of 48 pRNFL thickness points), the same authors achieved an accuracy of 88.5% (sensitivity = 89.3%, specificity = 87.6%) (Garcia-Martin et al., 2015). Garcia Martin et al. (2015)'s second model (Garcia-Martin et al., 2015) was also capable of

detecting a history of ON among the MS population, but expectedly, with a higher misclassification rate (sensitivity = 85%, specificity=83%).

Using a custom ultra-high-resolution OCT device, Zhang et al. (2020) applied a discrete wavelet transform algorithm to extract 22 texture features from mGCIPL thickness maps. They precisely defined three binary classification problems, i.e., ON-positive MS vs. HC eyes (AUC = 0.950), ON-negative MS vs. HC eyes (AUC = 0.849), and ON-positive vs. ON-negative MS eyes (AUC = 0.632) (Zhang et al., 2020).

In addition to RNFL and GCIPL, total macular volume and thickness measurements of outer retinal layers have also been employed for MS detection (Khodabandeh et al., 2023; Ortiz et al., 2023). Khodabandeh et al. (2023) observed that the best classification results are achieved when the most important features extracted from GCIPL/inner nuclear layer (INL) thickness maps in a 40 × 40-pixel area centered around the fovea are classified using support vector machine (SVM). They reported an accuracy of 88%, even when they used independent datasets from two countries for train and test phases (Khodabandeh et al., 2023).

Likewise, Kenney et al. (2022) performed another multi-center study including 1568 MS patients and 552 HC individuals from the USA, Canada, Europe, and the Middle East. They used different demographic, visual acuity, and OCT parameters, including pRNFL and mGCIPL thickness measurements of each eye and the average values of both eyes, for distinguishing between 1) MS patients and HCs, 2) MS patients with and without antecedent unilateral ON, and 3) MS eyes with and without antecedent ON. This work was also notable for using inter-eye thickness difference as an independent discriminating feature given to ML models, as studies have shown

that inter-eye difference cut-offs of 5 µm for pRNFL and 4 µm of GCIPL thicknesses are able to reliably detect unilateral ON (Nolan et al., 2018; Nolan-Kenney et al., 2019). The authors identified that three features, namely, mGCIPL thickness of both eyes on average, inter-eye mGCIPL thickness difference, and binocular 2.5% low-contrast letter acuity were the most important variables for MS classification, resulting in an AUC of 89% and an accuracy of 81% when logistic regression was applied to the whole dataset. SVM classifier was only used for 23% of study participants whose data were not missing (482/2,120), resulting in an accuracy of 88%. Both models also performed well in diagnosing ON by patient or by eye (Kenney et al., 2022).

Furthermore, MS could be discriminated from other demyelinating diseases of the CNS (DDs) using OCT measurements. Kavaklioglu et al. (2022) obtained 24 thickness values from different sectors of pRNFL, mGCIPL, and the whole retina in the macular region, and employed nine conventional ML algorithms to classify MS from HC (accuracy = 80%, sensitivity = 80%, specificity = 80%) and DDs (accuracy = 68%, sensitivity = 69%, specificity = 67%), including NMOSD, myelin oligodendrocyte glycoprotein antibody-associated disease, and monophasic acquired demyelinating syndromes in a pediatric population (Ciftci Kavaklioglu et al., 2022). Importantly, RNFL thickness in temporal and superior regions had the greatest capacity to differentiate between MS and other DDs. Also, the authors had excluded MS patients with a prior history of ON, but only considered the most recent three-month period; therefore, the results of this study may be interpreted with more caution, since there is a possibility that the retina may still be under affectation of the ON episode (s) having occurred before the specified threshold (Ciftci Kavaklioglu et al., 2022).

Another SD-OCT scanning protocol recently incorporated into the Spectralis devices, i.e., the posterior pole protocol (Asrani, 2011), was employed in studies by Hernandez et al. (2023) (Hernandez et al., 2023) and Ortiz et al. (2023) (Ortiz et al., 2023). The posterior pole protocol offers a unique property named anatomical positioning system in which a line is drawn between two fixed anatomical landmarks, i.e., the center of the fovea and the center of the Bruch's membrane opening. An 8 × 8 grid holding 64 average thickness measurements of the segmented retinal layers and the total retina is then superimposed on a 25° × 30° area around this line (Asrani, 2011; Hernandez et al., 2023; Ortiz et al., 2023).

Hernandez et al. (2023) utilized the 64 thickness values both individually and organized in six groups (Zones 1 to 6) as the input features to their decision tree-based classifiers (Lou et al., 2013). A major contribution of their work was the detailed explanation of the model predictions, providing an in-depth analysis of the features crucial for the decision-making process (Scott M. Lundberg and Lee, 2017). Also, thickness data from participants' right eyes, left eyes, randomly selected eyes whether right or left, and the average of both eyes were independently analyzed. Over ten folds of cross-validation, a best accuracy of 93.5% (fold 5) and a best mean accuracy of 81.56 ± 7.94 were achieved using an extreme gradient boosting (XGB) algorithm applied to the dataset of randomly selected eyes (Hernandez et al., 2023).

Similar to Kenny et al. (2022)'s study (Kenney et al., 2022), the discriminant capacity of inter-eye thickness difference was investigated by Ortiz et al. (2023) (Ortiz et al., 2023), but this time, using the posterior pole protocol. The authors first analyzed the AUC of the average thickness of both eyes and the inter-eye thickness differences for each of the nine segmented retinal layers and identified the

most important features accordingly; GCL average thickness and IPL inter-eye thickness difference were finally selected to be used for training a CNN from scratch; an accuracy of 87%, a sensitivity of 82%, and a specificity of 92% were finally achieved (Ortiz et al., 2023).

In addition to diagnosis, the role of ML in predicting MS course was also evaluated in three 10-year longitudinal studies by Montolio et al (Montolío et al., 2024, 2022, 2021). Inputs to their models were patients' demographic and clinical data, as well as OCT measurements, consisting of pRNFL,mRNFL, and mGCIPL thickness values. Initially, the authors developed models to differentiate between MS and HC individuals based on data from the visit at year 0, and were able to achieve accuracies of 87.7% (Montolío et al., 2021), 95.8% (Montolío et al., 2022), and 90.3% (Montolío et al., 2024). In their second experiment, Montolío et al. (2021, 2022, 2024) trained ML models with data from the visits at year 0, 1, and/or 2 to distinguish between MS patients with and without disease progression at year 10 of the follow-up, yielding promising results (81.7% (Montolío et al., 2021), 91.3% (Montolío et al., 2022), 81.9% (Montolío et al., 2024)). Disease progression was defined as an increase of at least one point on the Expanded Disability Status Scale between the year 2 visit and the final visit at year 10 (there were slight variations in the criteria for progression across the studies). The authors demonstrated that adding input data from the visit at year 2 will enhance model performance for predicting the disease course, achieving accuracies of 91.3 % (data from years 0, 1, and 2) and 86.3% (data from years 0 and 1) (Montolío et al., 2022). Of note, results from the third study (Montolío et al., 2024) that utilized mGCIPL data were inferior to the previous two studies that employed RNFL thickness measurements, suggesting that RNFL may provide more useful features for both diagnosing MS and predicting its disease course compared to GCIPL.

Swept-Source OCT Studies

SS-OCT studies have been generally successful in MS detection, with best accuracies ranging from 91% to even 100% (Cavaliere et al., 2019; Garcia-Martin et al., 2021; López-Dorado et al., 2021; Pérez del Palomar et al., 2019). All studies used the Topcon DRI Triton device (Topcon, Tokyo, Japan), capable of generating grids of 200 × 200 μm cubes of thickness measurements. In the first SS-OCT study, Del Palomar et al. (2019) applied decision trees, SVM, and neural networks on patients' age and gender, as well as mean RNFL and GCL+ (equivalent to GCIPL in SD-OCT) thicknesses in each grid box captured via the macular, peripapillary, and wide scanning protocols. When selected features from RNFL thickness values were given to a random forest classifier, accuracies of 97.2 % and 95.7% were achieved using the macular and wide protocols, respectively (Pérez del Palomar et al., 2019). Using a wide scanning protocol, Cavaliere et al. (2019) found that the thickness of 1) peripapillary GCL++ (equivalent to RNFL plus GCIPL), as well as that of the total retina in the 2) inner and 3) outer nasal sectors (according to the Early Treatment Diabetic Retinopathy Study) had highest AUCs for classifying MS. Applying SVM to the three features, they reached an accuracy of 91%, a sensitivity of 89%, and a specificity of 92% (Cavaliere et al., 2019). Two more recent studies (Garcia-Martin et al., 2021; López-Dorado et al., 2021) used the same dataset as Cavaliere et al. (2019)'s (Cavaliere et al., 2019), i.e., 48 HC and 48 early-diagnosed MS patients. In both of these works, Cohen's d technique (Cohen, 1962) was used to determine the most discriminant points across 45 × 60 thickness maps. Briefly, with respect to the mean and standard deviation of each thickness point, d values were calculated, and subsequently, were compared to a threshold of 1.02. Points with d values greater than the threshold (689 points in total) were directly fed to the models in Garcia Martin et al. (2021)'s study

(Garcia-Martin et al., 2021). López-Dorado et al. (2021) followed a slightly different approach, i.e., they applied Cohen's d method to the above thickness maps and generated new images that consisted of pixels with 1) d values greater than 1.02 with their intensities (thickness values) being unchanged and 2) lesser-than-threshold d values which were set to zero (López-Dorado et al., 2021). Both research teams (Garcia-Martin et al., 2021; López-Dorado et al., 2021) showed that GCL++, total retina, and GCL+ thickness maps contained more than 94% of the points with the most diagnostic capacity. In contrast to Del Palomar et al. (2019)'s study, RNFL thickness data was of the lowest importance; however, it should be noted that RNFL may still have made a valuable contribution to the final prediction since the highly differentiating GCL++ also includes RNFL. Garcia Martin et al. (2021) were finally able to successfully classify MS patients with an accuracy of 98% using a simple neural network (Garcia-Martin et al., 2021). More promising results were achieved by López-Dorado et al. (2021), i.e., sensitivity and specificity of 100%. Input images to their CNN were the newly generated thickness map images for each of the GCL++, total retina, and GCL+ layers, accounting for a final image size of 45×60×3. Notably, they used a generative adversarial neural network (GAN) for generating synthetic images in the training phase to avoid overfitting (López-Dorado et al., 2021). The model architecture was similar to Ortiz et al. (2023)'s study (Ortiz et al., 2023), consisting of two successive blocks, each consisting of a convolutional layer, followed by a ReLU activation function and a max pooling layer; the extracted features were then given to the fully-connected part. The higher performance of the CNN model in this study compared to Ortiz et al. (2023)'s (Ortiz et al., 2023) could be attributed to many reasons, e.g., the intrinsic difference between the two datasets particularly when considering that both had low sample sizes, more accurate OCT acquisition using an SS device, the superiority of Cohen's d technique over AUC analysis for feature selection, or the selected hyperparameters for each CNN model. Overall, the results

of such ML models certainly need to be validated using larger and more heterogeneous datasets, although SS-OCT devices are less accessible worldwide. Furthermore, the ability of SS-OCT data to detect ON in MS patients should also be assessed, since none of the prior studies included patients with a history of ON.

IR-SLO studies

Recently, two studies have utilized infrared reflectance scanning laser ophthalmoscopy (IR-SLO) images for classifying MS (Aghababaei et al., 2024; Arian et al., 2024). IR-SLO technology uses near-infrared laser beams to illuminate the retina in a raster pattern and then captures the backscattered light to create monochromatic, en face images of the retina, similar in appearance to fundus camera photographs. IR-SLO images are often captured alongside OCT, enabling the operator to fix the B-scan position, which appears as a line superimposed on the IR-SLO image. This approach helps eliminate eye movement artifacts, as the system uses an algorithm to monitor the IR-SLO images and reposition the B-scans in real time. Additionally, the combination of IR-SLO imaging with OCT enables consistent evaluation of disease progression by allowing the same B-scan to be examined at subsequent follow-up visits (Aumann S et al., 2019).Aghababaei et al. (2024) (Aghababaei et al., 2024) utilized IR-SLO images from two independent centers (Iran and the U.S. centers [total number of images = 164 MS and 150 HC images]) to train CNN models based on state-of-the-art architectures (VGG-16, VGG-19, ResNet-50, InceptionV3) and a custom CNN with five convolutional layers. Additionally, they created an autoencoder model for extracting the relevant features from the images, and the resulting features were then given to conventional ML classifiers, namely, k-nearest neighborhood, support vector machine, random forest, and neural network. The top-performing model was a combination of

autoencoder for feature extraction plus a neural network for final classification (accuracy = 88%, sensitivity = 86%, specificity = 91%) for classifying MS.

Multi-Modal Studies

Building on the work of Aghababaei et al. (2024) (Aghababaei et al., 2024), Arian et al. (2024) combined IR-SLO images with OCT thickness maps of the total retina to determine if this integration could further enhance model performance (Arian et al., 2024). They used datasets similar to those in Aghababaei et al. (2024) (Aghababaei et al., 2024)'s study, but to assess the generalization ability of their model, the authors assigned data from the Iranian center for training and internal validation, while data from the U.S. center was designated for external testing. Additionally, the authors used a bootstrapping strategy during the test phase. This approach involved repeatedly drawing subsamples from the test set with replacement, assessing the model performance on each sample, and ultimately presenting a distribution of the results. Arian et al. (2024) (Arian et al., 2024) developed CNN models based on VGG-16, VGG-19, ResNet-50, ResNet-101, and a custom architecture, and were able to achieve an accuracy of 92.40% ± 4.1% and 85.43% ± 0.08% for classifying MS using both modalities (IR-SLO and OCT) on the internal and external test datasets, respectively. Although identifying MS using IR-SLO images were associated with appealing results (82.57% ± 0.42% in the internal test dataset) similar to Aghababaei et al. (2024)' s work (Aghababaei et al., 2024), OCT data alone led to superior performance (94.32% ± 1.12% in the internal test dataset). The best-performing model, however, employed both modalities, suggesting that each may detect unique retinal changes indicative of

MS aiding in better classification outcomes; this noteworthy result calls for more focus in future studies, especially since IR-SLO and OCT are typically acquired together, and combining their data requires minimal additional effort.

Finally, two issues in MS studies worth noting:

First, although retinal thinning occurs in both eyes of MS patients, the damage is more pronounced in the eyes with a prior history of ON; this distinction is critical when determining how to train ML models. One strategy is to analyze MS eyes by creating two sub-datasets: those eyes with a history of ON and those without (Garcia-Martin et al., 2015; Kenney et al., 2022; Zhang et al., 2020). Alternatively, most other researchers opt to exclude ON-positive eyes entirely (Andorra et al., 2024; Cavaliere et al., 2019; Ciftci Kavaklioglu et al., 2022; Garcia-Martin et al., 2021; Hernandez et al., 2023; López-Dorado et al., 2021; Ortiz et al., 2023; Pérez del Palomar et al., 2019). In none of the approaches, however, the presence of antecedent subclinical ON in the ON-negative eyes can be ruled out. Other studies, ignored the ON status of MS eyes (Aghababaei et al., 2024; Arian et al., 2024; Garcia-Martin et al., 2013; Khodabandeh et al., 2023; Montolío et al., 2024, 2022, 2021), which may have led to inaccurate results.

Additionally, most studies included data from the same participant in both the train and test datasets. This is poor practice since the resulting ‘data leakage’ can lead one to overestimate the model’s performance on real patients who won’t be included in the training set. A smaller number of the studies has addressed this issue (data leakage between train and test datasets) in several ways (Aghababaei et al., 2024; Arian et al., 2024; Hernandez et al., 2023; Kenney et al., 2022; Khodabandeh et al., 2023; Ortiz et al., 2023): 1) Some studies (Aghababaei et al., 2024; Arian et al., 2024; Khodabandeh et al., 2023) adopted a “subject-wise” data-splitting strategy. In this method,

if data from one eye is included in the training set, data from the other eye of the same participant is prevented to appear in the test set. Arian et al. demonstrated that this subject-wise approach yields lower but more realistic performance metrics compared to the conventional data-splitting method (named as "record-wise" in Arian et al.'s work) that ignores possible data leakage between train and test datasets (Arian et al., 2024). 2) Some other studies (Kenney et al., 2022; Ortiz et al., 2023) utilized features like inter-eye thickness difference and/or the average thickness of both eyes as the input data to their models; obviously, both of these parameters (average thickness and inter-eye thickness difference) are unique for each participant and therefore, cannot be selected twice during train and test data-splitting. 3) In their study, Hernandez et al. (Hernandez et al., 2023) prevented data leakage by including only one randomly selected eye per subject, thereby ensuring that no data from the same individual is utilized in both the training and test datasets.

4. Brief Description of the Machine Learning Studies Aimed at Classifying Alzheimer's Disease Using Retinal Images

Fundus Camera Studies

The largest multi-center study on AD patients using retinal biomarkers was performed by Cheung et al. (2022) (Cheung et al., 2022). The participants consisted of 648 AD patients and 3240 HCs from 11 previous studies performed in 8 different countries. The authors developed a four-phase methodology; in the first phase, they used macula- and optic disc-centered fundus camera images from both eyes of each subject as the inputs to a bilateral model based on EfficientNet-B2 CNN (Tan and Le, 2020). A unilateral model was also developed using the fundus images from a single eye, since in a primary care setting for the elderly population, data on both eyes may

not be available, due to issues such as severe cataracts. In the third phase, a hybrid model was also trained, integrating extracted features from the images with demographic and clinical data considered as AD risk factors. Finally, the diagnostic capability of AD risk factors alone was tested in the fourth phase. By testing the models on five independent test datasets, the bilateral model generally performed better, with accuracies ranging from 79.6% to 92.1%. The promising performance of Cheung et al. (2022)'s models is of high importance in AD screening at a community-care level, because they were trained on a relatively large multi-country dataset and a more readily accessible imaging technology, compared to OCT, was used. The model was also capable of effectively differentiating between AD patients with positive and negative amyloid-beta in PET imaging (accuracy = 85.4-90.8%). As aggregation of amyloid-beta within the CNS may occur years to decades before a clinical diagnosis of AD (Dubois et al., 2016), identifying individuals positive for amyloid-beta using retinal imaging, which is accessible, non-invasive, and less costly compared to PET or CSF investigations, shows great potential for screening AD at a population level, enabling timely initiating therapeutic interventions.

In two other studies, fundus camera images were utilized to create retinal vessel maps and classify AD/MCI accordingly (Tian et al., 2021; Zhang et al., 2021). Both works followed a two-stage procedure, i.e., first, appropriate features were selected (Tian et al., 2021; Zhang et al., 2021) from the generated vessel maps, and subsequently, conventional ML classifiers were trained with the resulting features. Note that this assumes AD manifests only in the pattern of the blood vessels, and ignores any other information potentially available in the fundus image. Tian et al. (2020) were able to detect AD with 82.4% accuracy using the UK Biobank (UKBB) dataset which is a large-scale biomedical database containing extensive genetic and health data on over half a million individuals aged between 40-69 years from the UK (Sudlow et al., 2015). First, the authors took advantage of a strict image quality control process; tackling the

problem as a binary classification task (sufficient vs low quality), five CNN models with the same architecture and hyper-parameters except for weight initialization were trained with 150 fundus images labeled as “sufficient quality” and another 150 images labeled as “low quality”. Sources of low quality of fundus camera images may be related to both the scanning process such as over- and under-exposure of light and inadequate field of view, or the comorbidities in the examined patients, e.g., the cataract leads to an opaque lens worsening the image quality. When testing the resulting classifier on the whole dataset, an image was considered to have sufficient quality only if every one of the five models validated it as such.. This process led to the exclusion of over 70% of the total images, and the final dataset consisted of 122 fundus images from 87 AD patients and 122 images of HCs. Subsequently, after applying U-Net (Ronneberger et al., 2015) for vessel segmentation, the authors employed a t-test for feature selection, and finally, the SVM algorithm was used for the classification. In Zhang et al. (2021)’s study (Zhang et al., 2021), original fundus camera images were shown to result in superior performance compared to their corresponding vessel maps, suggesting that some discriminative features for MCI/dementia may specifically be found in the original images. The authors first created vessel maps and then employed histogram of oriented gradients (HOG) as a popular feature extraction technique in image processing (Dalal and Triggs, 2005) on both the original fundus camera and vessel map images. SVM and a neural network were applied to the HOG-derived features, finally achieving an AUC of 0.85 for HC (n = 38), 0.87 for MCI (n = 26), and 0.86 for dementia (n = 22) individuals. HOG works by first calculating the gradients (edges) within each cell (small boxes) in an image and then determining their orientation in degrees; finally, a histogram with its X axis consisting of bins of orientations and its Y axis consisting of the number of gradients with orientations corresponding to specific bins in the X axis is calculated for each cell and these histograms are finally integrated for the whole image. Therefore, as the features obtained by applying

HOG are related with shape and the edges, the models in Zhang et al. (2021)' study probably have focused on information from the retinal vessels. The fact that models based on original fundus images showed superior performance compared to vessel maps may be caused by the low efficacy of the vessel segmentation algorithm employed by the authors, missing vessel information when generating the vessel maps. Additionally, the selection of MCI/dementia patients was not limited to those caused by AD; for example, patients with a history of stroke and thus probable vascular dementia were also included. Therefore, the results of this study cannot be reliably compared with most other studies (Kim et al., 2022; Lemmens et al., 2020a; Nunes et al., 2019; Tian et al., 2021; Wang et al., 2022; Wisely et al., 2022, 2023) in this respect that only included patients based on the National Institute on aging-Alzheimer's Association diagnostic criteria for MCI (Albert et al., 2011) and dementia (McKhann et al., 2011) due to AD. Applying an strict exclusion criteria for quality control in Tian et al. (2021)'s study (Tian et al., 2021) allows for the ML classifier to be trained with data not confounded with irrelevant changes caused by scanning errors or comorbid pathologies other than AD; however, this process concurrently make the results of this study far from real-world scenarios where many of the problems with scanning devices and the process cannot be resolved especially in less developed regions where there is lack of high-technology devices and expert technicians. Also, conditions such as cataract that lowers image quality have a high prevalence among the elderly population who are also the biggest sufferers from AD; excluding the eyes with such conditions will mean that the models cannot be applied to many of cases we want to assess whether having AD or not. Therefore, like Cheung et al. (2022) (Cheung et al., 2022) and Shi et al. (2024)'s (Shi et al., 2024) works, future studies should be focus on training ML models with images of different qualities, allowing for their applicability in routine clinical practice.

Another potential drawback of Tian et al. (2020)'s study is the extensive computational requirements. This is because five CNN models were used for quality control and also, a complex model like U-Net was adopted for segmentation, both with a large number of parameters needed to be trained. To overcome such limitations, Kim et al. (2022) designed a simpler model with MobileNetV3-Large (Howard et al., 2019) as the backbone and modified it with a U-Net-like architecture, finally reaching a promising AUC of 0.93 (Kim et al., 2022).

OCT Studies

In addition to DL approaches, traditional ML models have also been utilized for classifying AD using OCT, in both mouse and human studies. OCT data given to the classifiers were in the form of B-scans (Gao et al., 2023; Shi et al., 2024), fundus-like images generated by averaging the A-scan values between different segmented layers (mean value fundus [MVF] images) (Ferreira et al., 2022; Trindade, 2021), extracted features from such MVF images by applying texture analysis methods (Nunes et al., 2019), and thickness measurements of different retinal layers (Lemmens et al., 2020a; Wang et al., 2022), also inputted in the image format in Wisely et al.'s studies (Wisely et al., 2023, 2022).

Wang et al. (2022) applied six different traditional ML models on OCT data obtained from a cohort of 159 AD and 299 HCs (Wang et al., 2022). The most significant decrease in OCT measurements among AD patients was observed in three pRNFL and five macular parameters. Training the ML models with these eight highly discriminant variables, the authors achieved a 0.74 accuracy and 0.75 AUC, with XGBoost as the best model. MVF images have also been a target for the automatic detection of AD and PD. Nunes et al. (2019)

generated MVF images for each of the six segmented retinal layers in the macular region, extracted 86 features per image using texture analysis techniques, and selected the best discriminating features. Ultimately, three binary classification problems were defined, i.e., AD vs. PD, AD vs. HC, and PD vs. HC., and 18 SVM models were trained in total (6 en face images per eye × 3 classification problems) with the texture features; using a majority-vote algorithm, assigning the final label to each eye in the test dataset was made based on the class most frequently predicted by the models. Although RNFL is considered a key layer affected by neurodegenerative diseases, the authors found that excluding RNFL from the analysis yielded more favorable results on the validation dataset, i.e., a median accuracy of 82.2% (across 100 consecutive runs). In particular, AD was discriminated from HC and PD with a sensitivity of 79.5% and a specificity of 92.5%. Furthermore, in cases where both eyes of each patient received the same label, the model predictions were correct more than 94% of the time (Nunes et al., 2019).

Hyperspectral Imaging Studies

Hyperspectral imaging data have also been considered as inputs to ML models for differentiating between individuals positive and negative for amyloid-beta (Hadoux et al., 2019; Sharafi et al., 2019), like Cheung et al. (Cheung et al., 2022) 's study. Unlike conventional fundus photography which provides limited spectral details from just three distinct color ranges (red, green, and blue), hyperspectral imaging captures reflectance data across tens to hundreds of continuous wavelength bands. This technique, therefore, combines both spectral and spatial information into a single 3-dimensional image, allowing for detecting potential reflectance changes in retinal tissue in the presence of molecules like Aβ in vivo (Lemmens et al., 2020b). Sharafi et al. (2019) applied SVM to the texture

features derived from hyperspectral images in addition to vessel tortuosity and diameter measurements of 46 individuals (16 positive and 30 negative for Aβ based on PET), finally reaching an accuracy of 85% (sensitivity = 82%, specificity = 86%) (Sharafi et al., 2019). Similar results (AUC = 0.87) were reported by Hadoux et al. (2019) who trained a linear ML algorithm (Dimension Reduction by Orthogonal Projection for Discrimination (Hadoux et al., 2015)) with mean values of spectral data from separate anatomical regions surrounding the fovea and the optic nerve (Hadoux et al., 2019).

Multi-Modality Studies

Five studies used a combination of retinal imaging modalities for automated diagnosis of AD/MCI. Lemmens et al. (2020) showed that a linear discriminant analysis classifier can achieve an AUC of 0.74 when pRNFL thickness measurements are added to hyperspectral imaging data (Lemmens et al., 2020a). Two other multi-modal studies were undertaken by Wisely et al. (2022, 2023) to discriminate MCI (Wisely et al., 2023) and AD (Wisely et al., 2022) patients from HCs. The authors incorporated mGCIPL thickness maps (in image format) and superior capillary plexus OCT-A en face images with (Wisely et al., 2022) or without (Wisely et al., 2023) ultra-widefield (UWF) color SLO and fundus autofluorescence images. The images were fed into a CNN model as the feature extractor with a ResNet-18 (He et al., 2015)-like architecture, sharing the same weight matrix for all input images. The resulting features from the images were then merged with numerical parameters obtained from OCT (e.g., mean GCIPL thickness) and OCT-A (e.g., fovea avascular zone area) as well as patients' general data (demographic and clinical), and finally, were given to the fully-connected part. Wisely et al. (2022, 2023) (Wisely et al., 2023, 2022) combined input data (image data, OCT and OCT-A numerical parameters and patients' general data)

in different ways and reported classification performance for each of these combinations. Input data leading to the highest accuracies were finally reported to be mGCIPL thickness map images and OCT/OCT-A numerical data, combined with patients' data in the first (AUC = 0.84) or OCT-A en face images in the second study (AUC = 0.81). Interestingly, the statistical analysis revealed no significant difference in almost all OCT and OCT-A measurements between MCI and HC individuals; however, CNNs were able to effectively differentiate between the two groups using the same images, highlighting the potential superiority of ML in detecting retinal features indicative of MCI compared to the parameters obtained from conventional image processing algorithms inherent to the commercial software in OCT/OCT-A devices.

Importantly, model generalization was more robustly validated in the second study (Wisely et al., 2023), since a post-hoc age-matching for the test dataset was performed, thereby minimizing the effects of age as a major confounding factor.

Gao et al. (2023) (Gao et al., 2023) utilized fundus camera images and OCT B scans of MCI (58 eyes), AD (76 eyes) and HC (100 eyes) individuals to train a custom CNN-based model incorporating attention mechanisms to focus on the most important features within each modality (self-attention) and to capture meaningful correlations between modalities (cross-attention), thereby minimizing irrelevant information for classifying AD and MCI. Additionally, the researchers applied contrast enhancement techniques like contrast-limited adaptive histogram equalization (CLAHE) to fundus camera images and employed noise reduction methods for refining OCT B scans. They were finally able to achieve a best accuracy of 92% (AUC = 0.97) for classifying AD, MCI, and HC subjects. However, using either fundus camera images or OCT B scans alone resulted in less satisfactory outcomes (accuracy = 78% for OCT and = 79% for

fundus camera images) compared to the bi-modal method. They further fine-tuned their pretrained model on the initial dataset to classify MCI in an external test dataset, achieving favorable results with 83% accuracy and an AUC of 0.84. They also showed that their model had superior performance over state-of-the-art attention-based models such as the Vision Transformer (Dosovitskiy et al., 2021). Notably, one of the key achievements of Gao et al. (2023) is their innovative use of OCT B-scans to identify cognitive impairment. The importance of using B-scans is that in neurodegenerative diseases, B-scans are not known for specific changes that can be visible to a human physician, unlike to retinal thickness measurements. To our knowledge, this is the first study that took advantage of ML to identify probable neurodegenerative processes by analyzing B-scans.

A large-scale study, similar to Cheung et al. (2022) (Cheung et al., 2022)'s, was recently performed. Shi et al. (2024) (Shi et al., 2024) trained CNN models with fundus camera images (optic disc-centered and macula-centered) (n = 8248) and OCT B scans (n = 4124) from 2062 Chinese individuals, consisting of 1807 HC subjects and 255 individuals with cognitive dysfunction (defined as Mini-Mental State Examination Score below 24). The models were based on VGG-19, ResNet-50, InceptionV3, and DensNet-121 CNNs; to enhance performance and reduce the risk of overfitting, the authors employed an ensemble approach, combining the predictions of individual models to generate the final outcome. To assess the model generalization ability, datasets from two independent centers were also considered for the test phase. Shi et al. (2024) were able to achieve an AUC of 0.82, 0.79, and 0.78 using the internal validation, external test 1, and external test 2 datasets, respectively. Like Cheung et al. (2022) (Cheung et al., 2022), the authors did not employ a strict exclusion criterion, i.e., individuals with concomitant conditions that may also affect the retina were also included. To achieve a more robust model applicable to routine clinical practice, it is important that the model be able to detect cognition impairment even in the

presence of diseases like diabetic retinopathy, age-related macular degeneration, and glaucoma, which are very common among the elderly population. Although the study did not determine the exact reasons for lower cognitive test scores, it is well-established that AD is the most prevalent cause of dementia. This implies that the retinal changes identified by the models might reflect neurodegenerative processes impacting both the retina and the brain. Using Grad-CAM, the authors also highlighted the potential of the choroid as a distinguishing feature between HCs and individuals with cognitive impairment, in line with previous studies indicating a reduction in the choroid thickness among AD and MCI patients (Alber et al., 2020; Gupta et al., 2021). Like Gao et al. (2023) (Gao et al., 2023)' study, OCT B scans were utilized for identifying cognitive impairment, further suggesting the probable retinal changes that can be detected by ML models but not a human. The authors also demonstrated that the model could reliably distinguish cognitive dysfunction from the HC state, regardless of age or gender, suggesting that that retinal changes associated with cognitive dysfunction may be consistent between males and females, as well as among individuals of varying ages. Additionally, no significant change was observed in model performance when fundus camera images and OCT scans were utilized independently as the input. This indicates that fundus camera imaging, which is more affordable and accessible compared to OCT, may be sufficient to detect retinal changes linked to cognitive impairment. However, the best-performing model required both modalities, suggesting that each one may capture unique features, and combining these features may improve the detection of cognitive dysfunction. A similar observation has been reported in Arian et al. (2024)'s study, in which features from both OCT thickness map and an en face retinal image (IR-SLO image) were useful for classifying MS (Arian et al., 2024).

Supplemental Material: Annu. Rev. Vis. Sci. 2026. 12: https://doi.org/10.1146/annurev-vision-110423-025837

Animal Studies

Furthermore, OCT-based models have also been used for the classification of wild-type and triple-transgenic mouse models of AD (3xTg-AD) using MVF images. The latter are genetically engineered mouse models of tauopathies, with mutations in the genes that are linked with familial AD, leading to progressive aggregation of amyloid-beta and changes in p-tau deposits (Yokoyama et al., 2022).

Ferreira et al. (2022) (Ferreira et al., 2022) generated MVF images of five layers per eye in an OCT dataset from 57 WT and 57 3xTg-AD mice, and designed a network with inceptionV3 (Szegedy et al., 2015) architecture as the backbone. First, they utilized images from mice of similar age (three-, four-, and eight-month-old) for both training/validation and the test phases, achieving a best accuracy of 91.3% (sensitivity = 100%, specificity = 82%, F1-score = 92%) when MVF images of the IPL were the model inputs. In their second experiment, the authors used images of three-, four-, and eight-month-old mice as their training data set and those of one-, two-, and twelve-month-old mice as their test set. This approach allowed them to determine whether their model could identify retinal changes caused by AD regardless of the mice's age. As expected, the results were less promising compared to when the training, validation, and test datasets were from mice of similar age; still, an accuracy as high as 88.5% was yielded when the input images were of RNFL-GCL. Model misclassification was more frequent among one-month-old mice, which was partly attributed to the resemblance of retinal features between healthy and diseased mice at a young age, as their CNS has not fully been developed yet. Also, model inaccuracy was shown to consistently decrease when moving from the outermost (ONL) to innermost (RNFL-GCL) retinal layers, which is in line with

clinical evidence suggesting that RNFL and GCL, the layers that consist of ganglion cells and their axons, become most affected in neurodegenerative diseases including AD (Alber et al., 2020; Gupta et al., 2021; Kashani et al., 2021).

Trindade (2021)'s study was very similar to Ferreira et al. (2022)'s (Ferreira et al., 2022), i.e., they used the same model architecture, and applied similar methods for contrast and intensity normalizations in the pre-processing stage. Six independent CNN models were developed for the classification of 783 MVF images produced for each layer; again, RNFL-GCL had the most discriminant characteristics (accuracy = 89.2%). To combine the information from all layers, the six scores predicted by CNN models for each eye were given to a simple neural network with a single hidden layer, resulting in an accuracy of 85.4%. The number was not different from the average accuracy of all CNNs when considered independently. The most important layers for AD classification were INL and OPL. Also, the model performed inadequately when right eyes were used exclusively for training and left eyes for testing; the authors attributed this to the intrinsic differences between the opposite eyes of each mouse; however, no robust clinical evidence has been found that confirms this hypothesis.

5. Brief Description of the Machine Learning Studies Aimed at Classifying Parkinson's Disease Using Retinal Images

OCT Studies

In Nunes et al. (2019)'s study, PD was differentiated from AD and HC state with a median sensitivity of 77.8% and a specificity of 97.8%, using texture parameters retrieved from MVF images of GCL, IPL, INL, OPL, ONL, and OPL; similar results were also achieved for AD detection as mentioned in the Section 4 (sensitivity = 79.5, specificity = 92.5) (Nunes et al., 2019).

Fundus Camera Studies

More recent studies published in February 2023 have adopted fundus images for classifying PD patients from HCs (Tran et al., 2023) or individuals with non-PD atypical motor problems (Ahn et al., 2023). Tran et al. (2023) manually selected 246 (123 PD and 123 HC) high-quality fundus images from the UKBB dataset and used them to train CNN and traditional ML models, including logistic regression and SVM. Well-known CNN architectures, such as AlexNet (Krizhevsky et al., 2012), VGG-16 (Simonyan and Zisserman, 2015), GoogleNet (InceptionV1) (Szegedy et al., 2014), InceptionV3 (Szegedy et al., 2015), and ResNet-50 (He et al., 2015), were employed, with AlexNet reaching the best performance (AUC = 0.78, accuracy = 71%). Also, VGG-16-based models were capable of distinguishing incident and prevalent PD from HCs with an AUC of 0.80 and 0.79, respectively. Patients with incident PD were defined as those participants who were not diagnosed prior to the baseline assessments in the UKBB study; the diagnosis of incident PD was essentially in line with prodromal and/or pre-symptomatic PD. Achieving diagnostic accuracy for pre-symptomatic PD similar to that for established cases is critically important. This finding implies that retinal changes associated with PD may occur regardless of clinical symptoms. Future research should prioritize this observation, and if confirmed in studies involving larger patient cohorts, it could pave the way for a new approach to early PD diagnosis. Additionally, relatively simple algorithms were also able to yield satisfactory results

comparable to DL models. For instance, SVM resulted in an AUC of 0.79 and an accuracy of 74% in detecting prevalent PD. The authors also used guided backpropagation (Fong and Vedaldi, 2019) for interpreting the model predictions and showed that the most robust explanation can be reached when AlexNet is utilized (Tran et al., 2023).

In a prospective hospital-based study by Ahn et al. (2023), fundus images of 266 PD patients and 349 non-PD individuals with atypical motor abnormalities were utilized to train five classic CNN architectures, including VGG-19 (Simonyan and Zisserman, 2015), ResNet-18 and -101 (He et al., 2015), and EfficientNet-B0 and -B7 (Tan and Le, 2020). The authors observed that ResNet-18 was the most appealing classifier (sensitivity = 77.1%, specificity = 76.5%, AUC = 0.76, accuracy = 76.6%). Furthermore, participants were divided into three categories, i.e., control, early-stage PD, and moderate/advanced-stage PD, according to their neurologic disability scores obtained from two well-known rating scales. As a multi-modal approach, the ResNet-18 (He et al., 2015) model was trained with fundus images and clinical/demographic data, reaching an accuracy of 70.5% for classifying between the three groups (Ahn et al., 2023).

Multi-Modal Studies

A recent study by Richardson et al. (2024) (Richardson et al., 2024) is the first to use multiple imaging modalities for classifying PD. Their approach closely resembles the work of Wisely et al. (2022, 2023), who used similar methods to differentiate MCI (Wisely et al., 2023) and AD (Wisely et al., 2022) from HC. In all the three studies, a similar combination of imaging modalities was utilized, including OCT GCIPL thickness maps, OCT-A en face 6 × 6 images of the SCP, with (Richardson et al., 2024; Wisely et al., 2022) and without (Wisely et al., 2023)UWF color and fundus autofluorescence SLO images. These images were input into a CNN model as a feature

extractor; the resulting features were then passed through fully-connected layers for the final classification. Richardson et al. (2024) adopted a subject-wise data splitting strategy, similar to approaches used in previous studies on MS classification (Aghababaei et al., 2024; Arian et al., 2024; Khodabandeh et al., 2023). This method ensures that data from the same subject's eyes is not duplicated across both the training and test sets, eliminating the risk of data leakage and allowing for a more accurate and fair assessment of the model performance. Another strategy that was employed for addressing the issue of overfitting was bootstrapping, similar to Arian et al. (2024)'s work (Arian et al., 2024). Interestingly, unlike Wisely et al. (2022)'s study (Wisely et al., 2022), UWF color SLO images led to the highest performance when used as the single input data, compared to when other modalities were utilized independently. This indicates that retinal changes unique to PD, not MCI or AD, can be most effectively identified with light reflectance/absorbance patterns recorded by color SLO imaging. SLO images may also contain retinal biomarkers that are beneficial for classifying MS, based on the favorable results achieved by Aghababaei et al. (2024) (Aghababaei et al., 2024) and Arian et al. (2024) (Arian et al., 2024) that utilized IR-SLO. Indeed, future research should pay more attention to the use of UWF SLO images for distinguishing between PD and HC or other neurodegenerative diseases.

6. Brief Description of the Studies That Utilized CNNs for Classifying Neurodegenerative Diseases

To the best of our knowledge, there have been fifteen studies that utilized CNNs to classify AD/MCI (Cheung et al., 2022; Ferreira et al., 2022; Gao et al., 2023; Kim et al., 2022; Shi et al., 2024; Trindade, 2021; Wisely et al., 2023, 2022), MS (Aghababaei et al., 2024;

Arian et al., 2024; López-Dorado et al., 2021; Ortiz et al., 2023), and PD (Ahn et al., 2023; Richardson et al., 2024; Tran et al., 2023) patients/diseased mice with retinal images as their inputs (Supplemental Table 3). Except for López-Dorado et al. (2021)'s (López-Dorado et al., 2021),Ortiz et al. (2023)'s (Ortiz et al., 2023) and Gao et al. (2023)'s (Gao et al., 2023), all studies took advantage of transfer learning using architecture and pre-trained weights of AlexNet (Tran et al., 2023), VGG-16 (Aghababaei et al., 2024; Arian et al., 2024; Tran et al., 2023), VGG-19 (Aghababaei et al., 2024; Ahn et al., 2023; Arian et al., 2024; Richardson et al., 2024; Shi et al., 2024), ResNet-18 (Ahn et al., 2023; Wisely et al., 2023, 2022), ResNet-50 (Aghababaei et al., 2024; Arian et al., 2024; Shi et al., 2024; Tran et al., 2023), ResNet-101 (Ahn et al., 2023; Arian et al., 2024), GoogleNet (InceptionV1) (Tran et al., 2023), InceptionV3 (Aghababaei et al., 2024; Ferreira et al., 2022; Shi et al., 2024; Tran et al., 2023; Trindade, 2021), MobileNet (Kim et al., 2022), DenseNet-121 (Shi et al., 2024), and EfficientNet-B0 (Ahn et al., 2023), -B2 (Cheung et al., 2022), and B-7 (Ahn et al., 2023) networks.

7. Approaches to Increase Interpretability of the Classification Models for Neurodegenerative Disorders

Both local and global interpretability techniques have been employed in the studies that classified neurogenerative diseases based on retinal images using conventional ML (Hernandez et al., 2023; Khodabandeh et al., 2023; Tian et al., 2021) and DL (Aghababaei et al., 2024; Ahn et al., 2023; Arian et al., 2024; Cheung et al., 2022; Gao et al., 2023; Shi et al., 2024; Tran et al., 2023; Trindade, 2021; Wisely et al., 2022) models.

Before diving into the studies in detail, it is essential to highlight that perturbation- and propagation-based methods are frequently used for local interpretability. Perturbation-driven approaches involve iteratively modifying (perturbing) the input data in a controlled way to observe the resulting changes in the output, while propagation-based algorithms use a single forward and/or backward propagation (Fong and Vedaldi, 2019).

Khodabandeh et al. (2023) (Khodabandeh et al., 2023) used a known perturbation-based technique, i.e., the occlusion method, and modified it to fit vector-shape inputs to make their SVM model interpretable. This approach involves a sliding window that moves across the input image, modifying the pixel intensities within each region to a predetermined value, such as 0 (representing a black patch). This iterative process continues until the entire image has been covered. By comparing the model output before and after masking specific regions of the input, regions that significantly influence the model predictions can be identified. Regions of greater importance to the model will exhibit more pronounced differences, while less influential regions will show smaller disparities (Zeiler and Fergus, 2013). Khodabandeh et al. (2023) (Khodabandeh et al., 2023) observed that the macular area (its temporal region) within GCIPL and INL thickness maps has the greatest impact on classifying an eye as having MS. Although the findings of Khodabandeh et al. (2023) is along the evidence that GCIPL thinning may be a more sensitive indicator of neurodegeneration in MS, compared to pRNFL (Saidha et al., 2015) and INL thickness is correlated with MRI lesion volumes in the brain (Saidha et al., 2013), we found no robust evidence that could explain the tendency of the higher involvement of the temporal areas around the macula.

The occlusion method was also used to explain another SVM model developed by Tian et al. (2021) (Tian et al., 2021). They tried black patches with different sizes of 1×1, 2×2, 4×4, and 8×8 pixels, sliding each of which over the entire vessel map to calculate the saliency scores for each image region. Visualizing the resulting heatmaps, the authors discovered that small retinal vessels had a more significant impact on the final classification compared to their larger counterparts; consistently with prior studies, this observation suggests that cognitive decline has a stronger association with changes in retinal "micro"-vascular system (Heringa et al., 2013; Santos et al., 2018; Tian et al., 2021).

As Fong & Vedaldi (2019) suggested (Fong and Vedaldi, 2019), the propagation-based strategy for CNN interpretation is composed of three main sub-categories, namely, gradient-based, activation-based, and a combination of gradient- and activation-based methods. Gradient-based methods, such as sensitivity analysis (also known as saliency maps) (Simonyan et al., 2014), rely on calculating the gradient of the output class scores with respect to the input features, with greater gradient values representing a more significant contribution to the model prediction. For instance, in the case of Alzheimer's disease (AD) fundus image classification, a sensitivity analysis algorithm computing the attribution scores from AD fundus images answers the question "What regions make this image more or less belong to an AD patient". However, one major pitfall in saliency map interpretation is the presence of visual noise, prompting extensive research efforts to remove noise and enhance the visualization quality (Montavon et al., 2018). Other gradient-based methods to tackle such a problem include DeConvNet (Zeiler and Fergus, 2013) and Guided Backprop (Springenberg et al., 2015).

Tran et al. (2023) (Tran et al., 2023) employed a guided backpropagation algorithm (Springenberg et al., 2015) to explain their CNN predictions. Guided backpropagation takes advantage of both sensitivity analysis and DeConvNet, in which the gradient signals are set to zero at each ReLU function, if the gradient itself is negative (DeConvNet (Zeiler and Fergus, 2013)) or the input to the same ReLU function was negative during the forward pass (sensitivity analysis (Simonyan et al., 2014)). Therefore, this approach helps in removing unimportant features, reducing noise, and improving the visualization of attribution maps. Interestingly, Tran et al. (2023) also computed scores for infidelity and sensitivity of their interpretation technique to objectively determine its reliability (Yeh et al., 2019). In their work, Yeh et al. (2019) (Yeh et al., 2019) defined explanation infidelity as the difference between 1) the dot-product of the input perturbation to the attribution scores and 2) the output perturbation, i.e., changes in the output values when "significant" perturbations are introduced to the input. These perturbations may encompass both random and non-random variations aimed at driving the input toward single or multiple reference values. Yeh et al. (2019) also introduced another novel metric, named sensitivity, that quantifies the extent to which "insignificant" perturbations in the test point affect the explanation. The ideal explanation should possess low sensitivity, which is crucial for generating attributions that remain consistent, even when confronted with minor perturbations applied to the input. Also importantly, through a combination of theoretical reasoning and empirical investigation, the authors (Yeh et al., 2019) effectively demonstrated that by incorporating an appropriate level of smoothing into the explanation method, it becomes feasible to attain an explanation that simultaneously showcases both low infidelity and low sensitivity (Yeh et al., 2019). Finally, Tran et al. (2023) showed that the CNN models primarily focused on retinal vessels for

distinguishing between both pre-clinical/prodromal and prevalent PD and HC individuals (Tran et al., 2023), in line with previous studies indicating retinal vascular changes in PD (Kwapong et al., 2018; Robbins et al., 2021; Shi et al., 2020).

Three other studies used interpretation algorithms like class activation mapping (CAM) (Zhou et al., 2015) and gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2020). CAM does not use backpropagation to create saliency heatmaps; instead, it combines the weights of the fully connected layer pertaining to a certain target class with the activation (feature) maps of the penultimate convolutional layer, using a dot-product operation. The obtained activation heatmaps are then summed and up-sampled to match the original input image size, revealing the regions that make the most significant contribution to predicting the target class (Zhou et al., 2015). However, an important consideration in CAM is that it could only be used for CNNs with an average pooling layer. Grad-CAM (Selvaraju et al., 2020) is a known example of the third group classified by Fong & Veldaldi (2019) (i.e., a combination of activation and gradient) (Fong and Vedaldi, 2019), and is a generalization of CAM that can be applied to any CNN architecture without modification. Grad-CAM works by calculating the gradients of the output score of the desired class with respect to the activations of the last convolutional layer. The gradients corresponding to each feature map are then averaged and used to compute a weighted sum of all maps of the last convolutional layer, ultimately resized to the original input image size. Essentially, the gradients here play a similar role to the weights of the fully connected layer in CAM. Grad-CAM provides a more accurate and fine-grained representation of the salient features for the target predicted class, compared to CAM. Additionally, Grad-CAM can generate similar saliency heatmaps for the intermediate convolutional layers, allowing for a more complete analysis of the model decision-making process (Selvaraju et al., 2020).

Applying Grad-CAM to IR-SLO images, Aghababaei et al. (2024) (Aghababaei et al., 2024) and Arian et al. (2024) (Arian et al., 2024) showed that the peripapillary area is the central region where MS-related changes occur, enabling the ML model to distinguish between MS and HC individuals. This is in line with statistical studies indicating that pRNFL become thinned in MS eyes with and even without history of ON (Petzold et al., 2017). Notably, MS is not typically associated with retinal changes within en face images that can be discerned by a human physician, at least in its early stages. RNFL damage cannot be detected during an ophthalmoscopic examination (providing en face view of the retina similar to IR-SLO or fundus camera images) until at least half of the layer thickness has been decreased (Quigley et al., 1982). However, Aghababaei et al. (2024) and Arian et al. (2024) highlighted the promising potential of ML models to uncover such pathologies.

Regarding AD, Wisely et al. (2022) showed that regions in en face OCT-A images of superior capillary plexus that had the greatest impact on AD classification were situated within the fovea avascular zone and in areas exhibiting reduced vessel density using CAM (Wisely et al., 2022). The importance of retinal vessel changes, including but not limited to vascular bifurcation abnormalities, in classifying AD and MCI has also been shown in Gao et al. (2024)'s study (Gao et al., 2023) that applied Grad-CAM to fundus camera images. Although the specific method of generating saliency heatmaps were not discussed in Shi et al. (2024)'s study (Shi et al., 2024), they demonstrated that the peripapillary area in fundus images and the thickness of the retina and choroid around the macula in OCT B scans were the regions identified by the model for classifying AD. All these findings are in line with prior evidence obtained from statistical studies, suggesting the structural (e.g., reduced thickness of pRNFL and GCIPL) and vascular changes (e.g., reduced vessel density, increased fovea avascular zone) in AD (Alber et al., 2020). In contrast to prior studies, Gao et al. observed

that the peripapillary area in fundus camera images did not play a role in identifying AD/MCI individuals; they attributed this, at least partially, to the absence of retinal capillary changes due to the amyloid-beta accumulation-induced inflammation at early stages of AD (Gao et al., 2023).

Regarding PD, utilizing CAM also revealed that staging the severity of PD is largely dependent on the features extracted from the macular area (Ahn et al., 2023), in line with previous studies reporting a negative correlation between mRNFL thickness and volume and PD severity (Unlu et al., 2018) .

Compared to the above "local" techniques which offer explanations specific for each input example, "global" interpretation methods provide a holistic perspective of the model predictions. There exists a variety of global interpretation strategies, including SHapley Additive exPlanations (SHAP) (Scott M Lundberg and Lee, 2017) and explainable boosting machine (EBM) (Lou et al., 2013) that were recently employed in Hernandez et al. (2023)'s study (Hernandez et al., 2023). SHAP is a popular model-agnostic interpretability method that originated from the cooperative game theory. In this method, the marginal contribution of a feature is calculated for all possible combinations of the features with that specific feature being included; the average of the resulting scores corresponds to the Shapley value of that feature (Scott M Lundberg and Lee, 2017). EBM is an inherently explainable ML model like linear and logistic regression algorithms, providing very accurate results while still maintaining a high level of interpretability. EBM is essentially a variant of generalized additive models that also calculates the interactions between the features in an efficient way. The training process is similar to that of the gradient boosting algorithm, where a shallow decision tree specific for each feature

is trained in an iterative “round-robin” manner. Similar to SHAP, EBM determines the pure effect of each feature on the final prediction by comparing the prediction scores with and without that feature being added (Lou et al., 2013). It is of note to mention that both methods are also capable of providing local interpretations. Applying SHAP on eXtreme Gradient Boosting (XGB) and random forest classifiers, Hernandez et al. (2023) (Hernandez et al., 2023) showed that thickness measurements of GCL in papillomacular bundle (Zone 1) and macular area (Zone 2) had the most contribution to model decisions, i.e., a thinner GCL in the above regions predicted a higher probability of MS. Among the 64 thickness values of the posterior pole grid, the most important features were points 5.4 and 5.3 which are again located in Zones 1 and 2, respectively. Interestingly, a boundary feature in Zone 5, i.e., point 3.1., also possessed a significant discriminating capability; this observation could be explained by the presence of medium to large vessels in this region that have probably been damaged during the disease course, ultimately leading to neuronal atrophy and GCL thinning. Results of EBM interpretation were also consistent with SHAP; however, interaction analysis revealed that EBM may be more susceptible to classification errors since thickness measurements from the peripheral areas were also shown to be discriminating, an observation that is inconsistent with clinical knowledge (Hernandez et al., 2023).

8. Supplemental Tables

Supplemental Table 1. Summary of machine learning studies for classification of multiple sclerosis using retinal images as their inputs

Authors (Year of publication)	Participants	Eyes with a positive history of ON (%)	Country	Device	OCT scanning protocol	Diagnosis criteria	Data given to feature selection/extraction algorithms	Major preprocessing tasks	Final inputs given to the classifier with best results	Classifier (s) [a]	Best performance
SD-OCT studies											
(Garcia-Martin et al., 2013)	106 MS 115 HC	31/106 (29%)	Spain	Spectralis SD-OCT	peripapillary	2001 McDonald Criteria	pRNFL thickness in 24 separate sectors of the same size	none	pRNFL thickness in 24 separate sectors defined by the authors	NN	AUC = 0.945
(Garcia-Martin et al., 2015)	112 MS 105 HC	41/112 (36.6%) [b]	Spain	Spectralis SD-OCT	peripapillary	ND	pRNFL thickness in 16 separate groups of the same size	performing experiments to get the optimal size of the groups containing pRNFL thickness points	pRNFL thickness in 16 separate sectors defined by the authors	NN	**MS vs HC:** SEN = 89% SPE = 88% **ON MS vs non-ON MS:** SEN = 85% SPE=83%
(Zhang et al., 2020)	29 MS 63 HC	33/58 (56.9)	USA	UHR-OCT (custom device)	macular	2010 McDonald Criteria	mGCIPL thickness maps	feature extraction using DWT	22 wavelet features from the mGCIPL thickness maps	LR LR-EN* SVM	**ON MS vs HC:** AUC=0.950 SEN = 88% SPE = 94% **non-ON MS vs ON MS:** AUC= 0.632 SEN = 91% SPE = 28% [c]
(Montolío et al., 2021)	108 MS 104 HC	34/108 (31.5)	Spain	Cirrus HD-OCT 4000	macular peripapillary	2001 McDonald Criteria	pRNFL thickness in superior, nasal, inferior, and temporal regions, and in average, mRNFL thickness in central fovea, age, sex, BCVA	none	pRNFL thickness in superior, nasal, inferior, temporal regions and in average, mRNFL thickness in central fovea, age, sex, BCVA	linear regression SVM DT K-NN NB LogitBoost*	ACC = 88% AUC = 0.878 SEN = 87% SPE = 89%
(Montolío et al., 2022)	72 MS 30 HC	15/72 (20.8%)	Spain	Spectralis SD-OCT	macular peripapillary	2010 McDonald criteria	RNFL thickness in macular (ETDRS sectors [d]) and peripapillary (six sectors [e] and 768 points) regions,	oversampling of the minority class using SMOTE	mRNFL thickness in superior outer, inferior outer, inferior inner, and	linear regression SVM K-NN* DT NB	ACC = 96% AUC = 0.958 SEN = 94% SPE = 97%

							papillomacular bundle, nasal/temporal ratio	feature selection using LASSO	central fovea regions, and in average	LogitBoost	
(Kenney et al., 2022)	1568 MS 552 HC	620/1568 (39.5%) [f]	11 centers from the USA, Canada, Europe, and the Middle East	Spectralis SD-OCT Cirrus HD-OCT	peripapillary macular	2010 McDonald Criteria	age, sex, ethnicity, disease duration, OCT [g], and visual acuity parameters	feature selection through CART and LR-based SFS	mGCIPL IED, average mGCIPL thickness of both eyes, binocular 2.5% LCLA	SVM [h] LR*	**MS vs HC:** AUC = 0.89 SEN = 81% SPE = 80% **ON-MS vs non-ON MS [i]:** AUC = 0.73, SEN = 19, SPE = 46
Kavaklioglu et al. (2022)	187 DDs (57 MS) 69 HC	none [j]	Canada	Cirrus HD-OCT 4000 Cirrus HD-OCT 5000	peripapillary macular	2017 McDonald Criteria	pRNFL thickness in four quadrants [e] and in average, mGCIPL thickness [k], macular retinal thickness in average and in EDTRS sectors [d], total macular volume	under-sampling of the dominant class feature selection using RFE	mGCIPL thickness in superior temporal, superior nasal, and inferior nasal sectors and in average, macular retinal thickness in superior inner and nasal outer sectors pRNFL in temporal quadrant [l]	SVM DT K-NN SGD RF * XGB AdaBoost NB LR	**MS vs HC:** ACC= 80% SEN = 80% SPE = 80% AUC = 0.85 **MS VS DDs:** ACC = 75% SEN = 65% SPE = 85% AUC = 0.84
(Khodabandeh et al., 2023)	**Charité dataset:** 106 MS 422 HC **Isfahan dataset:** 45 MS 45 HC	ND	Germany Iran	Spectralis SD-OCT	macular	**Charité dataset:** ND **Isfahan dataset:** 2017 McDonald criteria	total macular volume thickness maps of retinal layers [m] generated from 20×20, 30×30, and 40×40-pixel squares around the fovea, and from mean of ETDRS sectors [d]	selecting independent train and test datasets rotation and cropping of thickness maps feature extraction using PCA feature selection using RFE	extracted/selected features from thickness maps of mGCIPL and mINL	SVM* RF NN	ACC = 88% SEN = 88% PER = 89% F1 = 84%
(Ortiz et al., 2023)	79 MS 69 HC	none	Spain	Spectralis SD-OCT with posterior pole protocol	posterior pole	2017 McDonald Criteria	median thickness of both eyes and inter-eye thickness difference for each retinal layer [n]	feature selection using AUC analysis	GCL median thickness of both eyes, IPL inter-eye thickness difference	CNN	ACC= 87% SEN = 82% SPE = 92%
(Hernandez et al., 2023)	59 MS 111 HC	none	Spain	Spectralis SD-OCT with posterior pole protocol	posterior pole	2010 McDonald Criteria	RNFL and GCL thicknesses of each of the 64 cells inside the 8 × 8 grid around the macula, both	oversampling of the minority class with SMOTE generating four independent thickness	RNFL and GCL thickness measurements of each of the 64 cells inside the 8 × 8 grid around the macula,	XGB * RF EBM	ACC= 81.56 ± 7.94 SEN= 75.22 ± 13.14 SPE= 84.87 ± 9.59

							individually and grouped in six zones	datasets of the right eye, left eye, randomly selected eye, and the average of both eyes	both individually and grouped in six zones		F1= 69.36 ± 13.26 AUC = 0.80 ± 0.08 [o]
(Montolío et al., 2024)	72 MS 30 HC	22/72 (30.6%)	Spain	Spectralis SD-OCT	macular peripapillary	2010 McDonald criteria	GCL thickness in macular regions (ETDRS sectors [d] plus center thickness, center thickness minimum and maximum) total macular volume	oversampling of the minority class with SMOTE feature selection using LASSO	GCL thickness in macular regions (ETDRS sectors [d] plus center thickness, center thickness minimum and maximum) total macular volume	NN	ACC = 90.3% SEN = 88.9% SPE = 91.7% AUC = 0.90
SS-OCT studies											
(Pérez del Palomar et al., 2019)	80 MS 180 HC	none	Spain	Topcon DRI Triton SS-OCT	peripapillary macular wide	2017 McDonald Criteria	RNFL and GCL+ [q] thicknesses in macular and peripapillary regions using the three scanning protocols	feature selection using Attribute Classifier algorithms in WEKA software [r]	mRNFL thickness	SVM NN DT*	ACC = 97% AUC = 0.995 SEN = 96% SPE = 98%
(Cavaliere et al., 2019) [p]	48 MS 48 HC	none	Spain	Topcon DRI Triton SS-OCT	peripapillary macular wide	2010 McDonald Criteria	thickness of total retina and choroid in ETDRS sectors [d] thickness of total retina, RNFL, GCL+, GCL++ [s], and choroid in peripapillary quadrants and sectors [e]	feature selection using AUC analysis	thickness of GCL ++ and total retina in nasal inner and nasal outer ETDRS sectors	SVM	ACC = 0.91 AUC = 0.97 SEN =0.89 SPE = 0.92
(Garcia-Martin et al., 2021) [p]	48 MS 48 HC	none	Spain	Topcon DRI Triton SS-OCT	wide	2010 McDonald Criteria	thickness of total retina, RNFL, GCL+, GCL++, and choroid	feature selection using Cohen's d method	689 points selected from Cohen's d values of GCL++, GCL+, RNFL, and total retina	SVM NN*	ACC= 98% SEN = 0.98 SPE = 0.98
(López-Dorado et al., 2021) [p]	48 MS 48 HC	none	Spain	Topcon DRI Triton SS-OCT	wide	2010 McDonald criteria	thickness of total retina, RNFL, GCL+, GCL++, and choroid	feature selection using Cohen's d method generation of thickness map images according the Cohen's d values image augmentation using GAN	thickness maps generated from applying Cohen's d method on thickness values of GCIPL+, GCIPL++, and total retina	CNN	ACC = 100% SEN = 100% SPE = 100%

IR-SLO studies											
(Aghababaei et al., 2024)	**Isfahan dataset:** 35 MS 71 HC **Johns' Hopkins dataset:** 18 MS 14 HC	ND	Iran United states	Spectralis SD-OCT	macular	**Isfahan dataset:** 2017 McDonald Criteria **Johns' Hopkins dataset:** ND	IR-SLO images	image augmentation feature extraction using CAE	extracted features from IR-SLO images using CAE	NN	ACC = 88 % AUC = 0.94 SEN = 86 % SPE = 91%
Multi-Modal studies											
Arian et al. (2024)	**Isfahan dataset:** 32 MS 70 HC **Johns' Hopkins dataset:** 18 MS 14 HC	ND	Iran United states	Spectralis SD-OCT	macular	**Isfahan dataset:** 2017 McDonald Criteria **Johns' Hopkins dataset:** **ND**	OCT thickness maps of the total retina IR-SLO images	image augmentation	combination of OCT thickness maps of the total retina and IR-SLO images	CNN	ACC^{t} = 92.40% ± 4.1% % AUC = 96.99% ± 2.99% SEN = 95.43% ± 5.75% SPE = 92.82% ± 3.72%

a. the best classifier is noted with an asterisk sign.
b. patients with a previous history of ON in the recent six months were excluded.
c. model performance was also desirable for classifying non-ON MS vs HC (AUC = 0.849, SEN = 64%, SPE = 94%).
d. ETDRS sectors consist of central fovea, inner superior, inner nasal, inner inferior, inner temporal, outer superior, outer superior, outer nasal, outer inferior, and outer temporal.
e. peripapillary quadrants are made up of superior, inferior, temporal, and nasal quadrants, constituting six sectors, namely, superior nasal, superior temporal, temporal, nasal, inferior nasal, inferior temporal.
f. the reported ratio corresponds to the number of "patients" with a history of unilateral/bilateral ON, not the eyes. Also, patients with a history of acute ON (up to three months before data gathering) were excluded.
g. inter-eye pRNFL thickness difference, inter-eye GCIPL thickness difference, average pRNFL thickness of both eyes, average GCIPL thickness of both eyes, pRNFL thickness for each eye, GCIPL thickness for each eye
h. only participants with non-missing data were considered for SVM classification (n = 482). The results were as follows: ACC= 84%, AUC = 0.95, SEN = 86%, SPE = 79%, which was similar to logistic regression performance: ACC = 88%, AUC = 0.96, SEN = 86%, SPE = 89%.
i. the best inputs to the logistic regression classifier for obtaining such performance metrics (according to AUCs) were mGCIPL and pRNFL thickness separately. When using mGCIPL thickness, the metrics were as follows: AUC = 0.73, SEN = 19, SPE = 46. When using pRNFL thickness, the metrics were as follows: AUC = 0.73, SEN = 27, SPE = 36.
j. patients with a previous history of ON in the recent three months were excluded.

k. average and minimum mGCIPL thickness was measured, in addition to its thickness in superior, superior temporal, inferior, inferior nasal, and superior nasal sectors.
l. the features reported here were selected for MS-vs-HC (not MS-vs-DDs, for example) classification problem.
m. the input data were 1) RNFL thickness maps, 2) GCIPL, 3) sum of GCIPL and INL thickness maps, 4) parallel use of GCIPL and INL thickness maps, 5) parallel use of RNFL, GCIPL, INL, and ONL thickness maps, 6) total macular volume, 7) parallel use of GCIPL thickness maps and total macular volume, and 8) parallel use of GCIPL and INL thickness maps plus total macular volume.
n. thickness measurements of the following layers were utilized before feature selection: RNFL, GCL, IPL, INL, outer plexiform layer (OPL), outer nuclear layer (ONL), retinal pigment epithelium (RPE) layer, mean retinal thickness between the internal and external limiting membranes, and mean thickness of retinal layers between the external limiting and Bruch membranes.
o. the performance metrics reported here are mean accuracies across ten-fold cross validation using the dataset of randomly selected eyes (whether right or left).
p. the same dataset was used in these three studies.
q. GCL+ is equivalent to GCIPL in SD-OCT.
r. no specific feature selection method was clearly mentioned.
s. GCL++ is equivalent to GCIPL plus RNFL in SD-OCT.
t. Results reported here are model performance metrics with 95% confidence intervals, obtained after applying bootstrapping on the external test data set (the Johns Hopkins dataset).

Supplemental Table 2. Summary of machine learning studies for classification of Alzheimer's disease/mild cognitive impairment using retinal images as their inputs

Modality	Authors (Year of publication)	Participants	Dataset source (Country)	Device	Major preprocessing tasks	Final inputs given to the classifier with best results	Classifier (s) [a]	Best performance
Fundus camera image	Human studies							
	(Tian et al., 2021)	87 AD 87 HC	UKBB (England)	Topcon 3D OCT1000 Mark II	selection of five separate HC datasets with their age, gender, and eye side (right/left) matched with the AD dataset image quality control via a CNN model vessel segmentation using U-Net feature selection using t-test	t-test-selected features from the U-Net-generated vessel maps	SVM	ACC = 82.4% SEN = 79.2% SPE = 84.8%
	(Zhang et al., 2021)	22 dementia 26 MCI 38 HC	Sun Yatsen University (China)	RetiCam 3100	vessel segmentation using AHE, binarization, and median filtering feature extraction using HOG	HOG-extracted features from the original (not segmented) fundus camera images	SVM * ELM	AUC = 0.85 (HC) AUC = 0.87 (MCI) AUC = 0.86 (dementia)
	(Cheung et al., 2022)	648 AD 3240 HC	11 previous studies from 8 countries	Canon CR-1 40D Canon CR-DGi Canon CR-DGi10D Topcon TRC 50DX TRC-NW400 (Topcon) Topcon Maestro2	selecting independent train and test datasets	fundus images	CNN	ACC = 79.6 to 92.1% AUC = 0.73 to 0.91 SEN = 72 to 100% SPE = 90.9 to 100%
	(Kim et al., 2022)	111 AD 111 HC	UKBB (England)	Topcon 3D OCT1000 Mark II	random erasing data augmentation with random values	fundus images	CNN	AUC = 0.929 SEN = 84% SPE = 75%
OCT	(Wang et al., 2022)	159 AD 299 HC	Xiangya hospital (China)	Cirrus HD-OCT 4000	feature selection using covariance analysis	total macular volume thickness of macular RNFL, GCL, IPL, and total retina mean pRNFL, superior pRNFL, and inferior pRNFL	XGBoost *, Light GBM, KNN, RF, Gradient Boost, AdaBoost	ACC = 74% AUC = 0.75
	(Nunes et al., 2019)	20 AD 28 PD 27 HC	University of Coimbra (Portugal)	Cirrus SD-OCT 5000	generation of MVF image feature extraction using GLCM and DTCWT feature selection using SFS	selected features from GLCM- and DTCWT-extracted features of RNFL MVF images	SVM	SEN = 80% SPE = 93% ACC = 82%

Hyperspectral Imaging	(Sharafi et al., 2019)	16 Aβ positive 30 Aβ negative	Polytechnique Montreal, McGill University, etc. (Canada)	Metabolic Hyperspectral Retinal Camera (MHRC) by Optina Diagnostics	vessel segmentation measurement of vessel tortuosity and diameter texture feature extraction sequential feature selection feature extraction using PCA	PCA-extracted features from the eight selected variables, including texture (energy, contrast, correlation, and homogeneity) features plus vessel tortuosity and diameter	SVM	SEN = 82% SPE = 86% ACC = 85%
	(Hadoux et al., 2019)	15 Aβ positive 20 Aβ negative	Royal Melbourne Hospital and the AIBL study (Australia)	Metabolic Hyperspectral Retinal Camera (MHRC) by Optina Diagnostics	defining 6 ROIs segmentation and removal of blood vessels averaging the log-transformed spectral values spectral smoothing	smoothed form of the average spectral values from each ROI after log transformation	Dimension Reduction by Orthogonal Projection for Discrimination	AUC = 82%, 87% [b]
OCT + Fundus camera images	(Gao et al., 2023)	**Daping hospital dataset:** 29 MCI 38 AD 50 HC **WMU eye hospital:** 140 MCI 133 HC	Daping Hospital WMU eye hospital (China)	RetiCam 3100 VG 200 SS-OCT by Topcon	contrast enhancement of fundus camera images using CLAHE noise reduction for OCT B scans using bilateral and medial filtering creating masks for OCT B scans using Otsu's thresholding	fundus camera images OCT B scans	self- and cross-attention-enhanced CNN	ACC = 92% AUC = 0.97 ACC = 83% AUC = 0.84[c]
	(Shi et al., 2024)	**The Beijing Eye Study:** 301 MCI 2055 HC **Beijing Tongren Eye Center:** 35 MCI 260 HC **Beijing Tongren Hospital Physical Examination Center:** 30 MCI	The Beijing Eye Study Beijing Tongren Eye Center Beijing Tongren Hospital Physical Examination Center	Type CR6-45NM by Canon (fundus camera) Spectralis SD-OCT	image augmentation	fundus camera images OCT B scans	CNN	AUC = 0.84, 0.80, 0.75[d]

		265 HC	(China)					
OCT + hyperspectral	(Lemmens et al., 2020a)	17 AD 22 HC	University Hospitals UZ Leuven (Belgium)	XIMEA SNm4x4 VIS hyperspectral snapshot camera RT-vue XR Avanti	calculation of relative spectral values removal of blood vessels defining 4 ROIs, averaging the related values and a final standardization	standardized mean relative reflectance spectral values of I2 ROIs (14 features) pRNFL thickness in superior, inferior, temporal, and nasal areas and in average (5 features)	LDA	AUC = 0.74
OCT + OCT-A + UWF SLO and FAF	(Wisely et al., 2022)	36 AD 123 HC	Duke University (USA)	Cirrus HD-OCT 5000 with AngioPlex OCT Angiography Optos California	cropping UWF color and FAF images image resizing	GCIPL thickness map, OCT/OCTA quantitative data, patient data	CNN	AUC = 0.841 SEN = 100% SPE = 75%
OCT + OCT-A	(Wisely et al., 2023)	129 HC 80 MCI	Duke University (USA)	Cirrus HD-OCT 5000 with AngioPlex OCT Angiography	age-matching for the test dataset images	OCT/OCTA quantitative data [c]	CNN	AUC = 0.960
						GCIPL thickness map, OCTA SCP en face images, OCT/OCTA quantitative data		ACC = 81.3%, SEN = 79%, SPE = 83%, AUC = 0.809
Animal studies								
OCT (MVF images)	(Trindade, 2021)	57 WT mice 57 3xTg-AD mice (1, 2, 3, and 4-months old)	University of Coimbra (Portugal)	Micron IV OCT System	generation of MVF image from the RNFL-GCL, IPL, INL, OPL, ONL, and total retina contrast enhancement using CLAHE normalization using a Gaussian filter	MVF images of RNFL-GCL	CNN	ACC = 89% SEN = 98% SPE = 80% F1 = 90% [f]
	(Ferreira et al., 2022)	57 WT mice 57 3xTg-AD mice (1, 2, 3, 4, 8, and 12-months old)	University of Coimbra (Portugal)	Micron IV OCT System	generation of MVF image from the RNFL-GCL, IPL, INL, OPL, ONL, and total retina contrast enhancement using CLAHE normalization using a Gaussian filter	MVF images of RNFL	CNN	ACC = 88.1% SEN = 87.7% SPE = 88.5% F1 = 88.5% [g]

a. the best classifier is noted using an asterisk sign.

b. AUC was reported to be 0.82 and 0.87 when the training and validation data (internal validation and an independent dataset from Canada) were utilized, respectively.
c. the results reported here are obtained when the model performance was evaluated using the internal test (accuracy = 92% , AUC = 0.97) and external test datasets, i.e., data from the WMU eye hospital (accuracy = 83% , AUC = 0.84).
d. the results reported here, i.e., AUCs of 0.84, 0.80, and 0.75, are obtained when the model performance was evaluated using the internal test, external test 1 (Beijing Tongren Eye Center), and external test 2 (Beijing Tongren Hospital Physical Examination Center) datasets, respectively.
e. while the best-performing model that received images as input data (GCIPL thickness map and OCTA SCP en face images, plus OCT/OCTA quantitative data) reached an AUC of 0.809 [95% CI: 0.681, 0.937], the sole use of OCT and OCT-A quantitative data led to even a superior performance, i.e., AUC = 0.960 [95% CI: 0.902, 1.00].
f. six CNN models were trained with MVF images of each retinal layer/layer aggregate (i.e., RNFL-GCL, IPL, INL, OPL, ONL, and the total retina). As mentioned in the table, best results were achieved for the CNN trained with MVF images of the RNFL-GCL. Similarly, an accuracy of 89% was achieved when the predictions made by the all six CNN models were combined using a majority-vote algorithm, i.e., each mouse was labeled as diseased or healthy based on the class predicted by the majority of CNNs.
g. these results were obtained when train and test MVF images came from different age groups, i.e., images from 3-, 4-, and 8-months old mice were used in the training phase and those from 1- and 12-months old mice were used in the test phase. When both train and test image datasets were from mice with 3, 4, and 8 months of age, the best results were achieved for the CNN trained with IPL MVF images (ACC = 91%, SEN = 100%, SPE = 82%, F1 = 92%).

Supplemental Table 3. Characteristics of convolutional neural network-based classifiers used for automatic diagnosis of neurodegenerative disorders with retinal images as their input

Image type (image size)	Number of images	Architecture	Batch size	Optimizer (Learning rate)	Number of epochs	Specific techniques used to avoid over-fitting	Techniques used for increasing interpretability	Image augmentation methods	Best performance	Reference
AD/MCI										
Fundus images	5,598 AD 7,351 HC	EfficientNet-B2	40, 60 [a]	RAdam	160	transfer learning domain-specific batch normalization domain adaptation independent train and test datasets	Grad-CAM	affine horizontal/vertical flip color jitter	ACC= 80 to 92% SEN = 72 to 100% SPE = 90 to 100% AUC = 0.73 to 0.91	(Cheung et al., 2022)
SCP 3×3 mm en face OCTA (128 × 128) UWF color/FAF SLO (128 × 128) GCIPL thickness maps (128 × 128)	62 AD [b] 222 HC	ResNet-18	ND	Adam (0.01 for fully-connected layers and 0.0001 for others)	100	transfer learning with model complexity reduction [c] increasing training data through a shared feature extractor L1 and L2 regularizations image augmentation batch normalization	CAM	rotate shift crop zoom	AUC = 0.841 SEN = 100% SPE = 75%	(Wisely et al., 2022)
SCP 6×6 mm en face OCTA GCIPL thickness maps	154 MCI[b] 236 HC	ResNet-18	ND	Adam	100	transfer learning with model complexity reduction [c] increasing training data through a shared feature extractor batch normalization dropout	None	none	ACC = 81% SEN = 79% SPE = 83% AUC = 0.81	(Wisely et al., 2023)
fundus images	219 AD		16	AdaBelief	ND	transfer learning image augmentation	None	rotation horizontal/vertical flip	AUC = 0.929	

(830 × 830, 1500 × 1500)	226 HC	U-Net with MobileNetV3 as the backbone		(6e-4)		batch normalization		random size masking crop	SEN = 84% SPE = 89%	(Kim et al., 2022)
MVF images (512 × 512)	411 AD 372 HC	InceptionV3	[d]	RMSProp	ND	transfer learning batch normalization early stopping image augmentation	Grad-CAM	rotation horizontal/vertical flip	ACC = 89% [e] SEN = 98% SPE = 80% F1 = 90%	(Trindade, 2021)
MVF images	581 AD 563 HC	InceptionV3	ND	ND	ND	transfer learning heterogenous train and test datasets from different age groups image augmentation	none	rotation vertical flip	ACC = 88% [f] SEN = 87% SPE = 89% F1 = 89% [g]	(Ferreira et al., 2022)
Fundus camera images OCT B scans	76 AD [b] 58 MCI 100 HC	Custom CNN enhanced with self- and cross-attention modules	8	SGD (0.003 for the initial learning rate)	400	early stopping batch normalization independent train and test datasets using attention-based mechanisms to focus on the most relevant features within each image modality (self-attention) and combining them to reveal inter-modality correlations (cross-attention) and make a holistic analysis transfer learning utilized for the test dataset	Grad-CAM	Random crop Random flip	ACC [h] = 92% AUC = 0.97 ACC = 83% AUC = 0.84	(Gao et al., 2023)
Fundus camera images (224 × 224) OCT B scans (224 × 224)	510 MCI [b] 3614 HC	VGG-19 ResNet-50 InceptionV3 DenseNet-121	64	Adam (0.001 for the initial learning rate)	50	independent train and test datasets	ND	Horizontal flip Random crop	AUC [i] = 0.84, 0.80, 0.75	(Shi et al., 2024)
MS										
thickness maps after applying Cohen's d method (45 × 60 × 3)	100 MS 100 HC [j]	(convolutional layer + ReLU + pooling layer) × 2	ND	Adam	maximum number = 1000	feature selection using Cohen's d technique image augmentation through GAN batch normalization	none	none	ACC = 100% SEN = 100% SPE = 100%	(López-Dorado et al., 2021)
GCL thickness map calculated by taking the median of the thickness values from both eyes	79 MS 69 HC	(convolutional layer + non-linear activation function + batch	ND	Adam (1e-5)	200	feature selection using AUC analysis	none	none	ACC= 87% SEN = 82% SPE = 92%	(Ortiz et al., 2023)

IPL inter-eye difference thickness map (8×8×2)		normalization) × 2				batch normalization				
IR-SLO images (128 × 128 × 1)	164 MS 150 HC	VGG-16 VGG-19 ResNet-50 InceptionV3 Custom CNN CAE plus conventional classifier	16	Adam (1e-5)	ND	feature extraction using CAE transfer learning batch normalization dropout image augmentation independent train and test datasets	Grad-CAM	Rotation Zoom Vertical flip Width shift Height shift Brightness change	ACC = 88 % AUC = 0.94 SEN = 86 % SPE = 91%	(Aghababaei et al., 2024)
IR-SLO images (128 × 128 × 1) OCT thickness maps of the total retina (60 × 256 × 1)	MS: 133 IR-SLO 60 OCT map HC: 132 IR-SLO 124 OCT map	VGG-16 VGG-19 ResNet-50 ResNet-101 Custom CNN	16	Adam (2.75e-4)	ND	transfer learning batch normalization dropout image augmentation independent train and test datasets overcoming class imbalance using augmentation	Grad-CAM	rotation zoom vertical flip width shift height shift brightness change	ACC = 92% AUC = 97% SEN = 95% SPE = 93%	(Arian et al., 2024)
PD										
fundus images (256 × 256 × 3)	123 PD 123 HC	AlexNet GoogleNet ResNet-50 InceptionV3 VGG-16	64	Adam (1e-4)	100	transfer learning early stopping image augmentation	guided backpropagation	rotation horizontal flip vertical flip	AUC = 0.78 ACC = 71% SEN = 82% SPE = 60% F1 = 0.73 [k]	(Tran et al., 2023)
fundus images	539 PD 700 non-PD motor deficits	ResNet-18 ResNet-101 VGG-19 EfficientNet-B0 EfficientNet-B7	32	Adam (1e-4)	150	transfer learning image augmentation	CAM	rotation flip background color change	SEN= 77.1% SPE= 76.5% ACC= 76.8% AUC= 0.76	(Ahn et al., 2023)
SCP 3×3 mm en face OCTA (224 × 224) UWF color/FAF SLO (224 × 224) GCIPL thickness maps (224 × 224)	75 PD[b] 371 HC	VGG1-9	ND	Adam	50	Batch normalization Dropout Overcoming class imbalance using weighted random sampler and modified cross entropy loss	none	Rotation Width shift zoom	SEN = 100% SPE = 86% ACC = 91% AUC = 0.92	(Richardson et al., 2024)

a. the bilateral, unilateral, and hybrid models were trained with a batch size of 40, 60, and 40, respectively.
b. here, the number of eyes pertaining to each class is reported.
c. only the first five layers of ResNet-18 were used as the backbone model; model parameter reduction was further achieved by changing the average pooling layers from 2×2 to 4×4 size.
d. a grid search was performed for hyperparameter tuning for each of six CNNs trained with MVF images specific to each retinal layer. The grid values included 0.01, 0.001, and 0.0001 for learning rate, 8, 16, 32, and 64 for batch size, and 20, 30, 40, and 50 for number of epochs. The final selected values were specific for each CNN; for example, best values for the model trained with RNFL images included a learning rate of 0.0001, a number of epochs of 30, and a batch size of 8.
e. best performance was achieved for the CNN trained with MVF images of the RNFL-GCL. Also, using a majority-vote algorithm, the predictions made by the six CNNs trained with all MVF images of each mouse's retina were combined, resulting in a similar accuracy of 89%.
f. best performance was achieved for the CNN trained with MVF images of the RNFL-GCL.
g. these results were achieved when train and test MVF images came from different age groups, i.e., images from 3-, 4-, and 8-months old mice were used for training and those from 1- and 12-months old mice were used for the test datasets. When both train and test image datasets were from mice with 3, 4, and 8 months of age, best results were achieved for the CNN trained with IPL MVF images (ACC = 91%, SEN = 100%, SPE = 82%, F1 = 92%).
h. the results reported here are obtained when the model performance was evaluated using the internal test (accuracy = 92%, AUC = 0.97) and external test datasets, i.e., data from the WMU eye hospital (accuracy = 83% , AUC = 0.84).
i. the results reported here, i.e., AUCs of 0.84, 0.80, and 0.75, are obtained when the model performance was evaluated using the internal test, external test 1 (Beijing Tongren Eye Center), and external test 2 (Beijing Tongren Hospital Physical Examination Center) datasets, respectively.
j. the original participants included 48 MS and 48 HC eyes, which were fed into a generative adversarial neural network to generate a new dataset consisting of 100 MS and 100 HC eyes.
k. these results were achieved for discriminating overall PD and HC individuals; prevalent PD-vs-HC classification results were as follows: AUC = 0.79, ACC = 71%, SEN = 73% , SPE = 69%, and F1 = 71%; incident PD-vs-HC classification results were as follows: AUC =0.80, ACC = 71% , SEN = 78%, SPE = 64%, and F1 = 72%.

9. Literature Cited

Aghababaei A, Arian R, Soltanipour A, Ashtari F, Rabbani H, Kafieh R. 2024. Discrimination of multiple sclerosis using scanning laser ophthalmoscopy images with autoencoder-based feature extraction. *Mult Scler Relat Disord*. 88: 105743. https://doi.org/10.1016/j.msard.2024.105743

Ahn S, Shin J, Song SJ, Yoon WT, Sagong M, et al. 2023. Neurologic Dysfunction Assessment in Parkinson Disease Based on Fundus Photographs Using Deep Learning. *JAMA Ophthalmol*. 141(3): 234. https://doi.org/10.1001/jamaophthalmol.2022.5928

Alber J, Goldfarb D, Thompson LI, Arthur E, Hernandez K, et al. 2020. Developing retinal biomarkers for the earliest stages of Alzheimer's disease: What we know, what we don't, and how to move forward. *Alzheimer's & Dementia*. 16(1): 229–243. https://doi.org/10.1002/alz.12006

Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, et al. 2011. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*. 7(3): 270–279. https://doi.org/10.1016/j.jalz.2011.03.008

Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, et al. 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data*. 8(1): 53. https://doi.org/10.1186/s40537-021-00444-8

Andorra M, Freire A, Zubizarreta I, De Rosbo NK, Bos SD, et al. 2024. Predicting disease severity in multiple sclerosis using multimodal data and machine learning. *J Neurol*. 271(3): 1133–1149. https://doi.org/10.1007/s00415-023-12132-z

Arian R, Aghababaei A, Soltanipour A, Khodabandeh Z, Rakhshani S, et al. 2024. SLO-Net: Enhancing Multiple Sclerosis Diagnosis Beyond Optical Coherence Tomography Using Infrared Reflectance Scanning Laser Ophthalmoscopy Images. *Translational Vision Science & Technology*. 13(7): 13. https://doi.org/10.1167/tvst.13.7.13

Asrani S. 2011. Novel Software Strategy for Glaucoma Diagnosis: Asymmetry Analysis of Retinal Thickness. *Arch Ophthalmol*. 129(9): 1205. https://doi.org/10.1001/archophthalmol.2011.242

Aumann S, Donner S, Fischer J, Müller F. 2019. Optical Coherence Tomography (OCT): Principle and Technical Realization, In: *High Resolution Imaging in Microscopy and Ophthalmology*. pp. 59–85. Springer Cham.

Cavaliere C, Vilades E, Alonso-Rodríguez M, Rodrigo M, Pablo L, et al. 2019. Computer-Aided Diagnosis of Multiple Sclerosis Using a Support Vector Machine and Optical Coherence Tomography Features. *Sensors*. 19(23): 5323. https://doi.org/10.3390/s19235323

Cheung CY, Ran AR, Wang S, Chan VTT, Sham K, et al. 2022. A deep learning model for detection of Alzheimer's disease based on retinal photographs: a retrospective, multicentre case-control study. *Lancet Digit Health*. 4(11): e806–e815. https://doi.org/10.1016/S2589-7500(22)00169-8

Ciftci Kavaklioglu B, Erdman L, Goldenberg A, Kavaklioglu C, Alexander C, et al. 2022. Machine learning classification of multiple sclerosis in children using optical coherence tomography. *Mult Scler*. 28(14): 2253–2262. https://doi.org/10.1177/13524585221112605

Cohen J. 1962. The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*. 65(3): 145–153. https://doi.org/10.1037/h0045186

Dalal N, Triggs B. 2005. Histograms of oriented gradients for human detection, In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. . pp. 886–893 vol. 1. . https://doi.org/10.1109/CVPR.2005.177

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929

Dubois B, Hampel H, Feldman HH, Scheltens P, Aisen P, et al. 2016. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimers Dement*. 12(3): 292–323. https://doi.org/10.1016/j.jalz.2016.02.002

Ferreira H, Serranho P, Guimarães P, Trindade R, Martins J, et al. 2022. Stage-independent biomarkers for Alzheimer's disease from the living retina: an animal study. *Sci Rep*. 12(1): 13667. https://doi.org/10.1038/s41598-022-18113-y

Fong R, Vedaldi A. 2019. Explanations for Attributing Deep Neural Network Predictions, In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. eds. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-Rpp. 149–167. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_8

Gao H, Zhao S, Zheng G, Wang X, Zhao R, et al. 2023. Using a dual-stream attention neural network to characterize mild cognitive impairment based on retinal images. *Computers in Biology and Medicine*. 166: 107411. https://doi.org/10.1016/j.compbiomed.2023.107411

Garcia-Martin E, Herrero R, Bambo MP, Ara JR, Martin J, et al. 2015. Artificial Neural Network Techniques to Improve the Ability of Optical Coherence Tomography to Detect Optic Neuritis. *Seminars in Ophthalmology*. 30(1): 11–19. https://doi.org/10.3109/08820538.2013.810277

Garcia-Martin E, Ortiz M, Boquete L, Sánchez-Morla EM, Barea R, et al. 2021. Early diagnosis of multiple sclerosis by OCT analysis using Cohen's d method and a neural network as classifier. *Computers in Biology and Medicine*. 129: 104165. https://doi.org/10.1016/j.compbiomed.2020.104165

Garcia-Martin E, Pablo LE, Herrero R, Ara JR, Martin J, et al. 2013. Neural networks to identify multiple sclerosis with optical coherence tomography. *Acta Ophthalmologica*. 91(8): e628–e634. https://doi.org/10.1111/aos.12156

Gupta VB, Chitranshi N, den Haan J, Mirzaei M, You Y, et al. 2021. Retinal changes in Alzheimer's disease— integrated prospects of imaging, functional and molecular advances. *Progress in Retinal and Eye Research*. 82: 100899. https://doi.org/10.1016/j.preteyeres.2020.100899

Hadoux X, Hui F, Lim JKH, Masters CL, Pébay A, et al. 2019. Non-invasive in vivo hyperspectral imaging of the retina for potential biomarker use in Alzheimer's disease. *Nat Commun*. 10(1): 4227. https://doi.org/10.1038/s41467-019-12242-1

Hadoux X, Rutledge D, Rabatel G, Roger J-M. 2015. DROP-D: dimension reduction by orthogonal projection for discrimination. *Chemometr Intell Lab Syst*. 146: 221–231. https://doi.org/10.1016/j.chemolab.2015.05.021

He K, Zhang X, Ren S, Sun J. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/arXiv.1512.03385

Heringa SM, Bouvy WH, van den Berg E, Moll AC, Kappelle LJ, Biessels GJ. 2013. Associations Between Retinal Microvascular Changes and Dementia, Cognitive Functioning, and Brain Imaging Abnormalities: A Systematic Review. *J Cereb Blood Flow Metab*. 33(7): 983–995. https://doi.org/10.1038/jcbfm.2013.58

Hernandez M, Ramon-Julvez U, Vilades E, Cordon B, Mayordomo E, Garcia-Martin E. 2023. Explainable artificial intelligence toward usable and trustworthy computer-aided diagnosis of multiple sclerosis from Optical Coherence Tomography. *PLoS ONE*. 18(8): e0289495. https://doi.org/10.1371/journal.pone.0289495

Howard A, Sandler M, Chu G, Chen L-C, Chen B, et al. 2019. Searching for MobileNetV3. https://doi.org/10.48550/arXiv.1905.02244

Huang D, Swanson EA, Lin CP, Schuman JS, Stinson WG, et al. 1991. Optical coherence tomography. *Science*. 254(5035): 1178–1181. https://doi.org/10.1126/science.1957169

Kaplan A, Haenlein M. 2019. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*. 62(1): 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kashani AH, Asanad S, Chan JW, Singer MB, Zhang J, et al. 2021. Past, present and future role of retinal imaging in neurodegenerative disease. *Progress in Retinal and Eye Research*. 83: 100938. https://doi.org/10.1016/j.preteyeres.2020.100938

Kenney RC, Liu M, Hasanaj L, Joseph B, Abu Al-Hassan A, et al. 2022. The Role of Optical Coherence Tomography Criteria and Machine Learning in Multiple Sclerosis and Optic Neuritis Diagnosis. *Neurology*. 99(11): e1100–e1112. https://doi.org/10.1212/WNL.0000000000200883

Khodabandeh Z, Rabbani H, Ashtari F, Zimmermann HG, Motamedi S, et al. 2023. Discrimination of Multiple Sclerosis using multicenter OCT images. *Multiple Sclerosis and Related Disorders*. 0(0). https://doi.org/10.1016/j.msard.2023.104846

Kim DY, Lim YJ, Park JH, Sunwoo MH. 2022. Efficient Deep Learning Algorithm for Alzheimer's Disease Diagnosis using Retinal Images, In: *2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. . pp. 254–257. Incheon, Korea, Republic of: IEEE. https://doi.org/10.1109/AICAS54282.2022.9869953

Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. *Commun. ACM*. 60(6): 84–90. https://doi.org/10.1145/3065386

Kwapong WR, Ye H, Peng C, Zhuang X, Wang J, et al. 2018. Retinal Microvascular Impairment in the Early Stages of Parkinson's Disease. *Invest. Ophthalmol. Vis. Sci.* 59(10): 4115. https://doi.org/10.1167/iovs.17-23230

Lemmens S, Van Craenendonck T, Van Eijgen J, De Groef L, Bruffaerts R, et al. 2020a. Combination of snapshot hyperspectral retinal imaging and optical coherence tomography to identify Alzheimer's disease patients. *Alzheimers Res Ther*. 12: 144. https://doi.org/10.1186/s13195-020-00715-1

Lemmens S, Van Eijgen J, Van Keer K, Jacob J, Moylett S, et al. 2020b. Hyperspectral Imaging and the Retina: Worth the Wave? *Trans. Vis. Sci. Tech.* 9(9): 9. https://doi.org/10.1167/tvst.9.9.9

López-Dorado A, Ortiz M, Satue M, Rodrigo MJ, Barea R, et al. 2021. Early Diagnosis of Multiple Sclerosis Using Swept-Source Optical Coherence Tomography and Convolutional Neural Networks Trained with Data Augmentation. *Sensors*. 22(1): 167. https://doi.org/10.3390/s22010167

Lou Y, Caruana R, Gehrke J, Hooker G. 2013. Accurate intelligible models with pairwise interactions, In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. . pp. 623–631. Chicago Illinois USA: ACM. https://doi.org/10.1145/2487575.2487579

Lundberg Scott M., Lee S-I. 2017. A unified approach to interpreting model predictions, In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. . pp. 4768–4777. Red Hook, NY, USA: Curran Associates Inc.

Lundberg Scott M, Lee S-I. 2017. A Unified Approach to Interpreting Model Predictions, In: *A Unified Approach to Interpreting Model Predictions*. . p. Long Beach, CA, USA.

McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, et al. 2011. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*. 7(3): 263–269. https://doi.org/10.1016/j.jalz.2011.03.005

Montavon G, Samek W, Müller K-R. 2018. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing*. 73: 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Montolío A, Cegoñino J, Garcia-Martin E, Pérez del Palomar A. 2022. Comparison of Machine Learning Methods Using Spectralis OCT for Diagnosis and Disability Progression Prognosis in Multiple Sclerosis. *Ann Biomed Eng*. 50(5): 507–528. https://doi.org/10.1007/s10439-022-02930-3

Montolío A, Cegoñino J, Garcia-Martin E, Pérez del Palomar A. 2024. The macular retinal ganglion cell layer as a biomarker for diagnosis and prognosis in multiple sclerosis: A deep learning approach. *Acta Ophthalmologica*. 102(3). https://doi.org/10.1111/aos.15722

Montolío A, Martín-Gallego A, Cegoñino J, Orduna E, Vilades E, et al. 2021. Machine learning in diagnosis and disability prediction of multiple sclerosis using optical coherence tomography. *Computers in Biology and Medicine*. 133: 104416. https://doi.org/10.1016/j.compbiomed.2021.104416

Nolan RC, Galetta SL, Frohman TC, Frohman EM, Calabresi PA, et al. 2018. Optimal Intereye Difference Thresholds in Retinal Nerve Fiber Layer Thickness for Predicting a Unilateral Optic Nerve Lesion in Multiple Sclerosis. *Journal of Neuro-Ophthalmology*. 38(4): 451–458. https://doi.org/10.1097/WNO.0000000000000629

Nolan-Kenney RC, Liu M, Akhand O, Calabresi PA, Paul F, et al. 2019. Optimal intereye difference thresholds by optical coherence tomography in multiple sclerosis: An international study. *Ann Neurol*. 85(5): 618–629. https://doi.org/10.1002/ana.25462

Nunes A, Silva G, Duque C, Januário C, Santana I, et al. 2019. Retinal texture biomarkers may help to discriminate between Alzheimer's, Parkinson's, and healthy controls. *PLoS One*. 14(6): e0218826. https://doi.org/10.1371/journal.pone.0218826

Ortiz M, Mallen V, Boquete L, Sánchez-Morla EM, Cordón B, et al. 2023. Diagnosis of multiple sclerosis using optical coherence tomography supported by artificial intelligence. *Multiple Sclerosis and Related Disorders*. 74: 104725. https://doi.org/10.1016/j.msard.2023.104725

Pérez del Palomar A, Cegoñino J, Montolío A, Orduna E, Vilades E, et al. 2019. Swept source optical coherence tomography to early detect multiple sclerosis disease. The use of machine learning techniques. *PLoS ONE*. 14(5): e0216410. https://doi.org/10.1371/journal.pone.0216410

Petzold A, Balcer LJ, Calabresi PA, Costello F, Frohman TC, et al. 2017. Retinal layer segmentation in multiple sclerosis: a systematic review and meta-analysis. *Lancet Neurol*. 16(10): 797–812. https://doi.org/10.1016/S1474-4422(17)30278-8

Quigley HA, Addicks EM, Green WR. 1982. Optic nerve damage in human glaucoma. III. Quantitative correlation of nerve fiber loss and visual field defect in glaucoma, ischemic neuropathy, papilledema, and toxic neuropathy. *Arch Ophthalmol*. 100(1): 135–146. https://doi.org/10.1001/archopht.1982.01030030137016

Richardson A, Kundu A, Henao R, Lee T, Scott BL, et al. 2024. Multimodal Retinal Imaging Classification for Parkinson's Disease Using a Convolutional Neural Network. *Trans. Vis. Sci. Tech.* 13(8): 23. https://doi.org/10.1167/tvst.13.8.23

Robbins CB, Thompson AC, Bhullar PK, Koo HY, Agrawal R, et al. 2021. Characterization of Retinal Microvascular and Choroidal Structural Changes in Parkinson Disease. *JAMA Ophthalmology*. 139(2): 182–188. https://doi.org/10.1001/jamaophthalmol.2020.5730

Ronneberger O, Fischer P, Brox T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation.

Saidha S, Al-Louzi O, Ratchford JN, Bhargava P, Oh J, et al. 2015. Optical coherence tomography reflects brain atrophy in multiple sclerosis: A four-year study: Retinal Atrophy Reflects Brain Atrophy in MS. *Ann Neurol.* 78(5): 801–813. https://doi.org/10.1002/ana.24487

Saidha S, Sotirchos ES, Oh J, Syc SB, Seigo MA, et al. 2013. Relationships between retinal axonal and neuronal measures and global central nervous system pathology in multiple sclerosis. *JAMA Neurol*. 70(1): 34–43. https://doi.org/10.1001/jamaneurol.2013.573

Santos CY, Johnson LN, Sinoff SE, Festa EK, Heindel WC, Snyder PJ. 2018. Change in retinal structural anatomy during the preclinical stage of Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*. 10(1): 196–209. https://doi.org/10.1016/j.dadm.2018.01.003

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int J Comput Vis*. 128(2): 336–359. https://doi.org/10.1007/s11263-019-01228-7

Sharafi SM, Sylvestre J, Chevrefils C, Soucy J, Beaulieu S, et al. 2019. Vascular retinal biomarkers improves the detection of the likely cerebral amyloid status from hyperspectral retinal images. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*. 5(1): 610–617. https://doi.org/10.1016/j.trci.2019.09.006

Shi C, Chen Y, Kwapong WR, Tong Q, Wu S, et al. 2020. CHARACTERIZATION BY FRACTAL DIMENSION ANALYSIS OF THE RETINAL CAPILLARY NETWORK IN PARKINSON DISEASE. *Retina*. 40(8): 1483–1491. https://doi.org/10.1097/IAE.0000000000002641

Shi XH, Ju L, Dong L, Zhang RH, Shao L, et al. 2024. Deep Learning Models for the Screening of Cognitive Impairment Using Multimodal Fundus Images. *Ophthalmology Retina*. 8(7): 666–677. https://doi.org/10.1016/j.oret.2024.01.019

Simonyan K, Vedaldi A, Zisserman A. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. https://doi.org/10.48550/arXiv.1312.6034

Simonyan K, Zisserman A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556

Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. 2015. Striving for Simplicity: The All Convolutional Net.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, et al. 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med*. 12(3): e1001779. https://doi.org/10.1371/journal.pmed.1001779

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, et al. 2014. Going Deeper with Convolutions. https://doi.org/10.48550/arXiv.1409.4842

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. 2015. Rethinking the Inception Architecture for Computer Vision. https://doi.org/10.48550/arXiv.1512.00567

Tan M, Le QV. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. https://doi.org/10.48550/arXiv.1905.11946

Tian J, Smith G, Guo H, Liu B, Pan Z, et al. 2021. Modular machine learning for Alzheimer's disease classification from retinal vasculature. *Sci Rep*. 11: 238. https://doi.org/10.1038/s41598-020-80312-2

Tran C, Shen K, Liu K, Fang R. 2023. Deep Learning Predicts Prevalent and Incident Parkinson's Disease From UK Biobank Fundus Imaging. https://doi.org/10.48550/arXiv.2302.06727

Trindade RC. 2021. Identification of Changes in Retinal Images of Animal Models of Alzheimer's Disease Using a Deep Learning Approach. Master's Thesis. Coimbra, Portugal: University of Coimbra.

Unlu M, Gulmez Sevim D, Gultekin M, Karaca C. 2018. Correlations among multifocal electroretinography and optical coherence tomography findings in patients with Parkinson's disease. *Neurol Sci*. 39(3): 533–541. https://doi.org/10.1007/s10072-018-3244-2

Wang X, Jiao B, Liu H, Wang Y, Hao X, et al. 2022. Machine learning based on Optical Coherence Tomography images as a diagnostic tool for Alzheimer's disease. *CNS Neurosci Ther*. 28(12): 2206–2217. https://doi.org/10.1111/cns.13963

Wisely CE, Richardson A, Henao R, Robbins CB, Ma JP, et al. 2023. A convolutional neural network using multimodal retinal imaging for differentiation of mild cognitive impairment from normal cognition. *Ophthalmology Science*. 100355. https://doi.org/10.1016/j.xops.2023.100355

Wisely CE, Wang D, Henao R, Grewal DS, Thompson AC, et al. 2022. Convolutional neural network to identify symptomatic Alzheimer's disease using multimodal retinal imaging. *Br J Ophthalmol*. 106(3): 388–395. https://doi.org/10.1136/bjophthalmol-2020-317659

Yeh C-K, Hsieh C-Y, Suggala AS, Inouye DI, Ravikumar P. 2019. On the (In)fidelity and Sensitivity for Explanations.

Yokoyama M, Kobayashi H, Tatsumi L, Tomita T. 2022. Mouse Models of Alzheimer's Disease. *Front. Mol. Neurosci.* 15: 912995. https://doi.org/10.3389/fnmol.2022.912995

Zeiler MD, Fergus R. 2013. Visualizing and Understanding Convolutional Networks.

Zhang J, Wang J, Wang C, Delgado S, Hernandez J, et al. 2020. Wavelet Features of the Thickness Map of Retinal Ganglion Cell-Inner Plexiform Layer Best Discriminate Prior Optic Neuritis in Patients With Multiple Sclerosis. *IEEE Access*. 8: 221590–221598. https://doi.org/10.1109/ACCESS.2020.3041291

Zhang Q, Li J, Bian M, He Q, Shen Y, et al. 2021. Retinal Imaging Techniques Based on Machine Learning Models in Recognition and Prediction of Mild Cognitive Impairment. *NDT*. Volume 17: 3267–3281. https://doi.org/10.2147/NDT.S333833

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. 2015. Learning Deep Features for Discriminative Localization. https://doi.org/10.48550/arXiv.1512.04150