

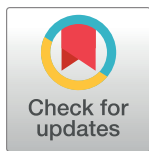
RESEARCH ARTICLE

Explainable artificial intelligence toward usable and trustworthy computer-aided diagnosis of multiple sclerosis from Optical Coherence Tomography

Monica Hernandez ^{1,2*}, Ubaldo Ramon-Julvez ^{1,2}, Elisa Vilades³, Beatriz Cordon ³, Elvira Mayordomo^{1,2}, Elena Garcia-Martin³

1 Computer Science Department, University of Zaragoza, Zaragoza, Spain, **2** Aragon Institute on Engineering Research, Zaragoza, Spain, **3** Ophthalmology Department, Miguel Servet Hospital, Zaragoza, Spain

* mhg@unizar.es



Abstract

OPEN ACCESS

Citation: 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>

Editor: Karim Adly Raafat, Cairo University Kasr Alainy Faculty of Medicine, EGYPT

Received: February 21, 2023

Accepted: July 19, 2023

Published: August 7, 2023

Copyright: © 2023 Hernandez et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data cannot be shared publicly because of the data sharing agreement conducted among the clinical and the technical authors of the study. Data comes from clinical practice in Miguel Servet Hospital, Zaragoza, Spain. Data are available from the Miguel Servet Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data after signing a collaboration agreement with the authors via ceica@aragon.es.

Background

Several studies indicate that the anterior visual pathway provides information about the dynamics of axonal degeneration in Multiple Sclerosis (MS). Current research in the field is focused on the quest for the most discriminative features among patients and controls and the development of machine learning models that yield computer-aided solutions widely usable in clinical practice. However, most studies are conducted with small samples and the models are used as black boxes. Clinicians should not trust machine learning decisions unless they come with comprehensive and easily understandable explanations.

Materials and methods

A total of 216 eyes from 111 healthy controls and 100 eyes from 59 patients with relapsing-remitting MS were enrolled. The feature set was obtained from the thickness of the ganglion cell layer (GCL) and the retinal nerve fiber layer (RNFL). Measurements were acquired by the novel Posterior Pole protocol from Spectralis Optical Coherence Tomography (OCT) device. We compared two black-box methods (gradient boosting and random forests) with a glass-box method (explainable boosting machine). Explainability was studied using SHAP for the black-box methods and the scores of the glass-box method.

Results

The best-performing models were obtained for the GCL layer. Explainability pointed out to the temporal location of the GCL layer that is usually broken or thinning in MS and the relationship between low thickness values and high probability of MS, which is coherent with clinical knowledge.

Funding: This work was partially supported by the national research grants PID2019-104358RB-I00 (DL-Ageing project), PID2022-138703OB-I00 (Trust-B-EyE project), Government of Aragon Group Reference T64_20R (COS2MOS research group), Carlos III Health Institute grants PI17/01726 and PI20/00437, and by the Inflammatory Disease Network (RICORS) (RD21/0002/0050) (Carlos III Health Institute). Ubaldo Ramon-Julvez work is granted by Government of Aragon. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Conclusions

The insights on how to use explainability shown in this work represent a first important step toward a trustworthy computer-aided solution for the diagnosis of MS with OCT.

1 Introduction

1.1 Multiple sclerosis

Multiple sclerosis (MS) is a neurodegenerative disease affecting the central nervous system (CNS). In MS the immune system attacks nerve fibers and myelin sheathing in the brain and spinal cord. The consequences are inflammation, demyelination, and axonal degeneration in the whole CNS, destroying nerve cell processes and altering the electrical messages in the brain. A confirmed diagnosis of MS is difficult, especially in the early stages of the disease when symptoms could be minor, sporadic, or even resemble other disease conditions. The diagnosis is based on the McDonald criteria, consisting of clinical, radiographic, and laboratory parameters extracted from a neurological examination and the history of neurological symptoms [1]. The initial version of the McDonald criteria was proposed in 2001 and it has been revised multiple times. Most recent criteria date from 2017.

To fulfill a diagnosis of MS based on the 2017 McDonald criteria, an individual must show evidences of CNS damage due to inflammation that is disseminating in space and in time. Dissemination in space occurs when the neurological damage is found in multiple parts of the CNS or multiple regions of the nervous system. Specifically, McDonald 2017 criteria establish that lesions should appear in at least two of the following four regions in the nervous system: periventricular, juxtacortical or cortical, and infratentorial areas of the brain, and the spinal cord. Dissemination in time occurs when the neurological damage is happening at more than one temporal point of the history of the patient. Damage can be evidenced by a second disease exacerbation, the appearance of new lesions, or evidence that damage in the same regions happened at different times (e.g. new inflammatory lesions surrounding older lesions that are no longer actively inflamed). Magnetic Resonance Imaging (MRI) of the brain and the spinal cord is used to detect the plaques or scars typical of MS damage. MRI with gadolinium can be used to distinguish between active and inactive lesions, since gadolinium enhancing lesions are areas of active inflammation.

In addition, the presence of oligoclonal bands (OB) in the cerebrospinal fluid (CSF) is indicative of CNS inflammation. Individuals with clinically isolated syndrome (CIS) have experienced a single episode of MS symptoms, therefore, dissemination in time criteria are not fulfilled. For these individuals, OB has been established as an independent predictor of relapse. Therefore, McDonald 2017 criteria establish testing positive for OB as a sufficient criterion that replaces dissemination in time criteria even in patients that show evident damage in just one temporal point of their history.

Unfortunately, MRI and CSF evaluation are time-consuming, costly, and invasive. For example, magnetic resonance devices are expensive, and image acquisition may range from 10 to 30 minutes. Gadolinium injections have side effects such as injection site pain, nausea, itching, dizziness, and headaches. Cerebrospinal fluid samples are acquired through a lumbar puncture that takes about half an hour under local anaesthetic. It can be described as unpleasant and painful and side effects may include infection of the puncture area and headaches. For these reasons it is worth exploring complementary or alternative criteria to be included in the McDonald criteria once properly validated.

1.2 MS and ophthalmic region deterioration

In the last decades, several studies indicate that the anterior visual pathway provides information about the dynamics of axonal degeneration in MS [2–11]. The hypothesis is that MS may be early diagnosed and then followed up from the quantification of the axonal damage in the neuroretina [3, 12]. The optic nerve is being extensively investigated as a potential location in future updates to the McDonald criteria [13].

Optical coherence tomography (OCT) is a fast (order of seconds) and non-invasive imaging technique that allows the segmentation and quantification of the thickness of the different neuroretinal layers with less than 2 microns of precision. The discomfort for the patient during the acquisition is minimum, since he/she only has to stare at a fixed point for a few seconds while the images are acquired. In particular, OCT has revealed that the retinal nerve fiber layer (RNFL) reflects the axonal integrity, while the ganglion cell layer (GCL) shows a progressive thinning since the early stages of MS [2, 14, 15]. These findings point out OCT-derived measurements as biomarkers potentially useful in the diagnosis of the disease.

Commercial OCT devices are distributed with accurate retinal layer segmentation algorithms. They also provide quantification of the layer thicknesses in different ways and locations. One-dimensional measurements of the layer thicknesses around the optic nerve head are taken using circular sections such as the peripapillary circle scan. Using this protocol, the thickness of the peripapillary RNFL (pRNFL) in the temporal sector has been associated with a level of cognitive and physical disability in MS [2, 16]. In its wide scan protocol, Triton OCT device (Topcon, Tokyo, Japan) provides a two dimensional grid of measurements of the thickness in macular and peripapillary areas (45×60 points). Posterior Pole (PPole) is a new two-dimensional quantification protocol for Spectralis OCT devices (Heidelberg Engineering, Germany). Similarly to Triton super grid, PPole protocol quantifies the different layer thicknesses in an 8×8 grid. PPole has been satisfactorily used in early glaucoma assessment [17, 18] and recent studies are conducted to establish its potential to detect changes in RNFL and GCL due to MS [19, 20]. Current research in the field is focused on the quest for the most discriminative measurements among MS and controls.

In particular, the PPole protocol of Spectralis OCT constitutes a significant advance in measurement acquisition, because it is based on the high reliability provided by applying the Anatomical Positioning System (APS) to OCT imaging. The APS works by locating points on the eye using two fixed structural landmarks: the centre of the fovea and the centre of the Bruch membrane opening. This allows much more precise monitoring, detecting even the slightest alteration or inter-eye difference because using the APS, in conjunction with the True Track eye-tracking system (TruTrack, Heidelberg Engineering), ensures accurate identification of macula position in each eye based on head tilt and eye cyclotorsion. With other protocols, part of the differences observed in the measurements could be due to differing alignment between the right and left eyes, to minor anatomical or refractive differences between them, or even to the patient's head being positioned on the OCT device's chin rest at a slight angle imperceptible to the clinician.

1.3 Machine learning in the diagnosis of MS from OCT. Usable solutions

The different ways of quantifying the layer thicknesses have been used as feature sets for building machine learning systems toward the computer-aided diagnosis of MS. The comprehensive review recently provided by Aslam et al. in [21] describes the advances in the performance of these methods and discusses some limitations.

Garcia-Martin et al. [22] pioneered the studies of the discrimination capability of multiple sclerosis vs healthy control (MS vs HC) subjects using ML systems. In this work, artificial

neural networks (ANN) were used with measurements from the circumpapillary RNFL thickness from Spectralis SD-OCT. The study involved 106 MS patients and 115 age-matched HC. The best-performing model obtained an AUC of 0.94.

Palomar et al. [23] studied the diagnostic capacity of traditional machine learning methods in different locations of the neuroretina, centered on the macula (macular protocol) or on the optic nerve (peripapillary protocol). The authors used a two-dimensional grid of measurements from Triton SS-OCT device. The thickness was computed for the RNFL and GCL+ (GCL + IPL layers, IPL standing for the inner plexiform layer). The study involved 80 MS patients and 180 age-matched HC. RNFL showed as the best feature set for the prediction of the disease with a 97.24% of accuracy with decision trees (DT).

Cavaliere et al. [24] combined support vector machines (SVM) with the mean values over right and left eyes of the macular thickness and the peripapillary area layer thickness from Triton SS-OCT. Feature selection was performed and the models were built over a feature set of dimension three. The most discriminant values were the total GCL++ thickness (between the inner limiting membrane to the inner nuclear layer boundaries) and the macular retina thickness in the nasal quadrant of the outer and inner rings. The study involved 48 MS and 48 HC subjects. The best-performing model obtained a 91% of accuracy.

Montolio et al. [25] performed an exhaustive study involving different ML methods in the diagnosis of MS from the RNFL thickness averaged over the peripapillary, superior, nasal, inferior, temporal, and foveal areas. The images were acquired with a Cirrus high definition OCT device. The study involved 108 MS and 104 HC subjects. The study also included a temporal analysis of the evolution of the measurements with a follow-up of 10 years. The best-performing model was obtained from a model ensemble of the best-performing models generated in the cross-validation. The accuracy of the ensemble was 87.7% while kNN reached an accuracy of 85.4% and SVM of 84.4%.

Garcia-Martin et al. [26], proposed a method for the early diagnosis of MS. The authors used a two dimensional grid of measurements from Triton OCT device acquired using wide protocol (12 × 9 mm). The thickness was computed for the RNFL, GCL+, and GCL++ layers and features were extracted using a thresholding over the Cohen's d metric. The study involved 48 MS patients and 48 healthy controls. GCL++ layer showed to be the one with the highest discriminant capacity, with a 98.00% of accuracy in this case with a shallow neural network (NN). It should be noticed the contradiction with Palomar et al. findings [23].

Montolio et al. [27] reproduced the study in [25] with the measurements of the peripapillary circle scan in the RNFL layer and a different population of subjects. The study involved 72 MS patients and 30 HC. The authors reported kNN as the best-performing method with a 95% of accuracy on a validation set balanced with SMOTE oversampling method.

Lopez Dorado et al. [28] implemented a system based on a convolutional neural network (CNN) using the two-dimensional grid of measurements from Triton OCT device acquired using wide protocol. The study involved 48 MS patients and 48 healthy controls. The dataset was augmented using a generative adversarial network (GAN) and Cohen's d metric was used to select the retinal structures with the greatest discriminant capacity (resulting in GCL++, complete retina, and GCL+). The final model obtained a 100% of accuracy in a leave-one-out evaluation.

The major criticism from Aslam et al. in [21] is that these studies have been conducted with a very small number of samples. Consequently, the models may not be robust presenting a high variance. In addition, the models may be biased. We agree with these appreciations. Indeed, the population sample and utilized features are usually quite different among the studies. This makes it unfeasible to establish which could be the best combination of feature set and learning system for the task. Some of these works show comparative tables of the

performance of the different methods when they cannot be objectively compared. These problems are not exclusive to this area of research and they are found more frequently than is desirable in different applications of ML in clinical contexts [29, 30].

The valuable information from these works relies on which features can be potential biomarkers for which stages of the disease, but the proposed systems are still far from computer-aided solutions usable for the general clinical practice. To ensure the reliability of the proposed systems, the models should be built on a large dataset, demonstrate a high generalization capability, and make decisions coherent with clinical knowledge. However, this is not a simple task in the great majority of clinical applications and this one is no exception.

The only study to date involving a large dataset has been recently conducted by Kenney et al. [31]. The authors used Cirrus high-definition OCT and converted Spectralis HP-OCT measurements of the pRNFL and GCL+ layers. The absolute difference between eyes and the average between eyes were compared as features. The study involved 1568 MS patients and 552 HC. The authors reported the best results for the GCL+ layer with a logistic regression, obtaining an AUC of 0.77. The AUC obtained for the pRNFL measurements was 0.65. These results were improved up to an AUC of 0.89 with a composite score combining OCT measurements with the score of an ophthalmologic test (a not anatomical feature). These results are far from the impressive performances over the 90% provided in the small-dataset studies, but it is not immediate to assess on the causes of the divergence of these results.

1.4 Explainable AI. Trustworthy solutions

Medical experts usually do not trust ML decisions unless they come with comprehensive and easily understandable explanations [32]. It is well-known that the most accurate ML methods are black boxes. This means that it is not comprehensive or easily understandable why or how the system reached a specific decision or whether the system is making decisions with arguments that are coherent with medical knowledge. In contrast to black-box models, glass-box models offer full transparency since all parameters are known and it is understandable how the system comes to its conclusion. However, glass-boxes tend to be much less accurate than black boxes.

Attending the taxonomy of explainable methods, the trade-off between performance and explainability is known. There is an inverse relationship between the accuracy of a machine-learning algorithm and our ability to understand and interpret its output. Therefore, one may engage the clinical experts with glass-box methods at the expense of decreasing the accuracy.

In the last years, the novel discipline of explainable AI (xAI) has been developed with interesting methods that try to open the black box of the most complex methods and provide explanations of their decisions. Among them, SHAP [33] (<https://shap.readthedocs.io>) is an xAI method based on game theory that provides information among the relationship between the feature values and the predictions of an ML system. SHAP was proposed with many different ways to graphically represent the SHAP values. Some of them allow the global identification of the most important features involved in the predictions and assess whether high or low feature values contribute to high or low probability predictions for a given class. For particular cases, the SHAP values allow knowing which features contributed to their correct or wrong classification.

Another approach in xAI consists of the proposal of inherently explainable methods that are performance competitive with traditional black boxes, breaking the performance-explainability trade-off. Explainable Boosting Machine (EBM) (<https://interpret.ml>) has been recently proposed as a glass-box method intended to achieve an accuracy similar to competing black-

box methods. Explanations from EBM can be graphically represented in a similar way to SHAP.

The information provided by the different xAI approaches can be analyzed with clinical experts to find out whether the ML models are making decisions coherent with clinical knowledge, detect biases, or even point out flawed experimental designs. Thanks to SHAP and EBM, xAI is nowadays in a mature state to bridge the gap between models generated in an academic research environment and their effective utilization in clinical practice as usable and trustworthy solutions.

1.5 Our contribution to the state of the art

The objective of our work is to study the potential of the PPole protocol from Spectralis SD-OCT combined with ML methods for the diagnosis of MS with a novel approach based on xAI. As feature sets, we consider two different ways of representing the measurements of the GCL and RNFL thickness. The ML methods have been selected among those that have obtained the best performances in other medical applications [29, 30]. Explainability has been provided using SHAP and EBM. The models have been built using a local database with the problems usually found in clinical practice: small size, unbalancing, heterogeneity, etc. For a better insight on the capabilities of our systems, we are showing the average and variance of the performance measurements for the different ML methods, the different combinations of measurement representation, and sample selection. We are describing how to provide detailed global and local explanations for the ML decisions and how can they be interpreted in an understandable way. We analyze whether the obtained explanations are coherent with clinical knowledge. Finally, the local explanations of the failure cases provide interesting insights into the decision-making process of the different methods. We believe that the explainability study shown in this work represent a first important step toward a trustworthy computer-aided solution for the diagnosis of MS that may help neurologist to get an earlier definitive diagnosis in patients with clinical suspect of MS or even to get an estimation of the risk of presenting the disease in subjects with unspecific neurological symptoms or with familiar antecedents of MS.

1.6 Manuscript organization

In the following, Section 2 describes the methods used in our study, including the data (sec 2.1), the ML and xAI methods used in our work (secs 2.2 and 2.3) for the discrimination of MS and HC subjects, and the experimental design of our study (sec 2.4). Section 3 evaluates the models obtained with our data and presents and discusses the explainability results. Finally, Section 4 gathers the most remarkable conclusions of our work.

2 Methods

2.1 Dataset description

The clinical procedures in this study adhered to the tenets of the Declaration of Helsinki. The experimental protocol was approved by the Ethics Committee of the Miguel Servet Hospital (CEICA), Zaragoza, Spain, and all participants provided written informed consent to participate in the study. We included patients with definite relapsing-remitting MS (RR MS). Diagnosis was based on the 2010 revision of the McDonald Criteria and confirmed by a specialized neurologist.

Our dataset comprises a total of 316 eyes from 111 healthy controls and 59 patients of relapsing-relapsing MS (RR MS). The dataset used in [20] is contained into ours. [Table 1](#)

Table 1. Demographics of the data used in our work. In addition, we show whether statistical significance can be obtained for the zone-based Posterior Pole (PPole) measurements of ganglion cell layer (GCL) and retinal nerve fiber layer (RNFL) between our populations of healthy controls (HC) and multiple sclerosis (MS) subjects. Bold numbers indicate p values <0.05.

	Healthy Controls	Multiple Sclerosis	p-value
Subjects	111	59	n.a.
Left eyes	109	50	n.a.
Right eyes	107	50	n.a.
Age (years)	44.55 ± 11.14	41.47 ± 14.06	0.15
Male/Female	25/86	10/49	n.a.
Z1 GCL thickness	41.78 ± 2.89	36.70 ± 6.62	< 0.001
Z2 GCL thickness	49.22 ± 3.68	43.50 ± 7.71	< 0.001
Z3 GCL thickness	28.52 ± 2.36	27.56 ± 3.40	0.015
Z4 GCL thickness	27.23 ± 2.46	25.42 ± 2.55	< 0.001
Z5 GCL thickness	28.46 ± 2.35	26.38 ± 2.87	< 0.001
Z6 GCL thickness	28.39 ± 2.62	27.13 ± 3.43	0.02
Z1 RNFL thickness	31.87 ± 4.26	30.12 ± 5.69	0.031
Z2 RNFL thickness	28.32 ± 2.99	26.66 ± 4.17	0.004
Z3 RNFL thickness	63.94 ± 10.25	60.58 ± 12.09	0.070
Z4 RNFL thickness	80.37 ± 12.93	72.40 ± 13.40	< 0.001
Z5 RNFL thickness	30.46 ± 4.38	28.63 ± 4.88	0.018
Z6 RNFL thickness	22.39 ± 2.73	22.91 ± 5.66	0.431

<https://doi.org/10.1371/journal.pone.0289495.t001>

shows the demographics of our data. Mean disease duration in patients was 7.73±3.98 years and median EDSS score was 1.35 (range: 0–5). Treatment distribution in MS group was 3 patients with Avonex (intramuscular interferon beta-1a), 4 patients with Betaseron (interferon beta-1b), 2 patients with Rebif (subcutaneous interferon beta-1a), 3 patients with Copaxone (glatiramer acetate), 44 patients received oral last-generation treatments: Tecfidera (dimethyl fumarate) for 11 patients, Gilenya (fingolimod) for 27 patients, Mayzent (siponimod) for 2 patients and Aubagio (teriflunomida) for 4 patients and 3 patients did not receive treatment. The patients did not have a flare-up of optic neuritis. They were not in an active outbreak and had not had an active outbreak for at least 6 months.

Eyes with previous episodes of optic neuritis were excluded from the study in order to avoid bias when assessing how neurodegeneration is appreciable in the retinal layers. In addition, the exclusion criteria included: best-corrected visual acuity lower than 0.5 according to Snellen charts, eyes longer than 25.2 mm, refractive errors higher than 5 diopters of spherical equivalent refraction or 3 diopters of astigmatism, intraocular pressure higher than 20 mmHg, media opacifications (nuclear color/opalescence, cortical or posterior subcapsular lens opacity lower than 2 according to the Lens Opacities Classification System III) (Chylack, 1993), concomitant ocular disease (including glaucoma or retinal pathology), and other systemic conditions potentially affecting the visual system.

Structural measurements of the retina were obtained using the Spectralis OCT device (Heidelberg Engineering, Germany). The Posterior Pole protocol (PPole), recently included in the device software, was used for all subjects (see Fig 1). This protocol incorporates the Anatomic Positioning System (APS), which describes a horizontal line between the fovea and the entrance of Bruch's Membrane opening. Based on this reference line 61 parallel explorations are performed inside an area of 25° × 30°. APS combined with the True Track system for eye tracking ensures an accurate position of the macula for each individual based on their head tilt and eye cyclotorsion.

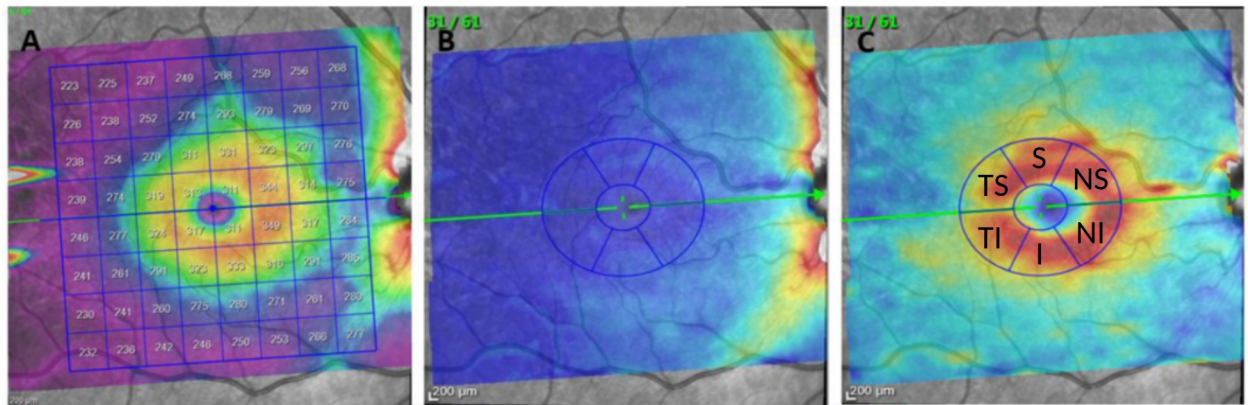


Fig 1. Illustration of the Posterior Pole (PPole) protocol. Left, standard positioning of the 8×8 grid centered over the macula. Center, color thickness map of the retinal nerve fiber layer (RNFL). Right, color thickness map of the ganglion cell layer (GCL). The green line indicates the direction of Spectralis Anatomical Positioning System (APS). The blue sectors indicate the distribution of superior (S), nasal superior (NS), nasal inferior (NI), inferior (I), and temporal inferior (TI), and temporal superior (TS) areas.

<https://doi.org/10.1371/journal.pone.0289495.g001>

The OCT images were obtained by a single operator blinded to group classification. Images have an axial (in tissue) resolution of $3.9 \mu\text{m}$. Low-quality images (quality score higher than 25/40) or images with movement artifacts were excluded from the analysis. The OCT device performed a detailed segmentation of the retinal layers that was used for obtaining the PPole measurements as follows. The $25^\circ \times 30^\circ$ area was divided into an 8×8 grid of 0.86×0.86 mm cells, and the average thickness of each segmented layer was recorded in the 64 cells. In addition, the 64 cells were grouped into 6 zones previously proposed in [20] for a better clinical understanding of the information. The criteria for zone grouping clusters the upper (Zones 4 and 5), lower (Zones 3 and 6), and papillomacular bundle (Zone 1) arch patterns that follows the RNFL, the concentric pattern of the GCL dividing quadrants (Zones 3, 4, 5, and 6), and the macular area, with the highest density of ganglion cells (Zone 2). Both the 64-dimensional and the 6-dimensional measurements will be considered feature sets in our work. Fig 2 illustrates the division of the PPole grid into zones. In this work, we will focus on the measurements of the ganglion cell layer (GCL) and the retinal nerve fiber layer (RNFL). Table 1 shows whether statistical significance can be obtained for the zone-based measurements, following the study in [19, 20]. Fig 3 shows typical examples of the PPole data for multiple sclerosis (MS) and healthy control (HC) records in our database.

Visual inspection of the whole database allowed us to obtain initial insights into the group differences that may be interesting for the diagnosis:

- There are visual differences between left and right eyes (e.g. subject 5 from MS group and subject 8 from HC group).
- The GCL layer shows a torus-like shape that is frequently broken in the temporal-inferior region for the MS group (e.g. subject 1).
- We can find this GCL break in the control group although to a lesser extent (e.g. subject 8).
- The GCL torus thickness is usually greater in the HC group, although there are subjects with a thin thickness (e.g. subject 9).
- The RNFL layer shows a wider dark blue extension for the MS subjects, which means that the thickness in the MS group tends to be smaller than in the HC group (e.g. visual comparison between subject 3 and subject 7).

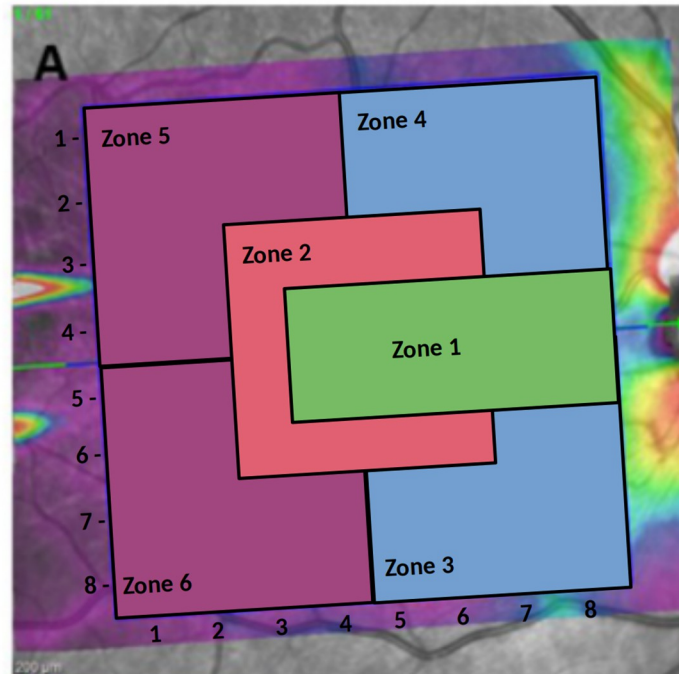


Fig 2. Coordinate system and zone grouping of the 8×8 Posterior Pole (PPole) grid. In our work we use a coordinate system with origin located at point (1,1) of the PPole grid. This way, the coordinate naming convention follows the one used by Spectralis optical coherence tomography device.

<https://doi.org/10.1371/journal.pone.0289495.g002>

2.2 Machine learning methods considered in this work

Our problem can be classified as a very small dataset problem. The Zones feature set can be represented as tabular data. Although the 8×8 PPole grid feature set could be represented as imaging data, the low resolution of the images made us lean towards the tabular representation. We included in our study two black-box methods, gradient boosting and random forest, ranking as the best-performing methods in different applications and challenges [29]. In addition, we included one good-performing glass-box method, explainable boosting machine, in order to assess the accuracy-explainability trade-off for our application. We discarded the use of neural networks architectures from our study since the superiority of traditional methods in datasets with similar characteristics to ours has been extensively demonstrated [34].

2.2.1 Gradient Boosting (XGB). Gradient Boosting is an ensemble learning technique that combines several weak learners (with poor accuracy) into a strong learner (with high accuracy). In particular, gradient boosting is a model ensemble of individual decision trees that are trained iteratively such that each new tree improves the residual errors of the previous tree ensemble. XGBoost (XGB) is an optimized distributed gradient boosting library available in the popular *scikit-learn* library. Its high efficiency is achieved through a parallel tree boosting algorithm that is known to accurately solve data science problems involving billions of examples. In the last years, XGB is ranked among the best methods for ML classification and regression on tabular data [34]. To the date, neural networks are not able to outperform XGB in many different tasks. For example, the winner of the Alzheimer's Disease diagnosis forecast in the TADPOLE Challenge was Frog, a system built on a gradient boosting machine with XGBoost [29].

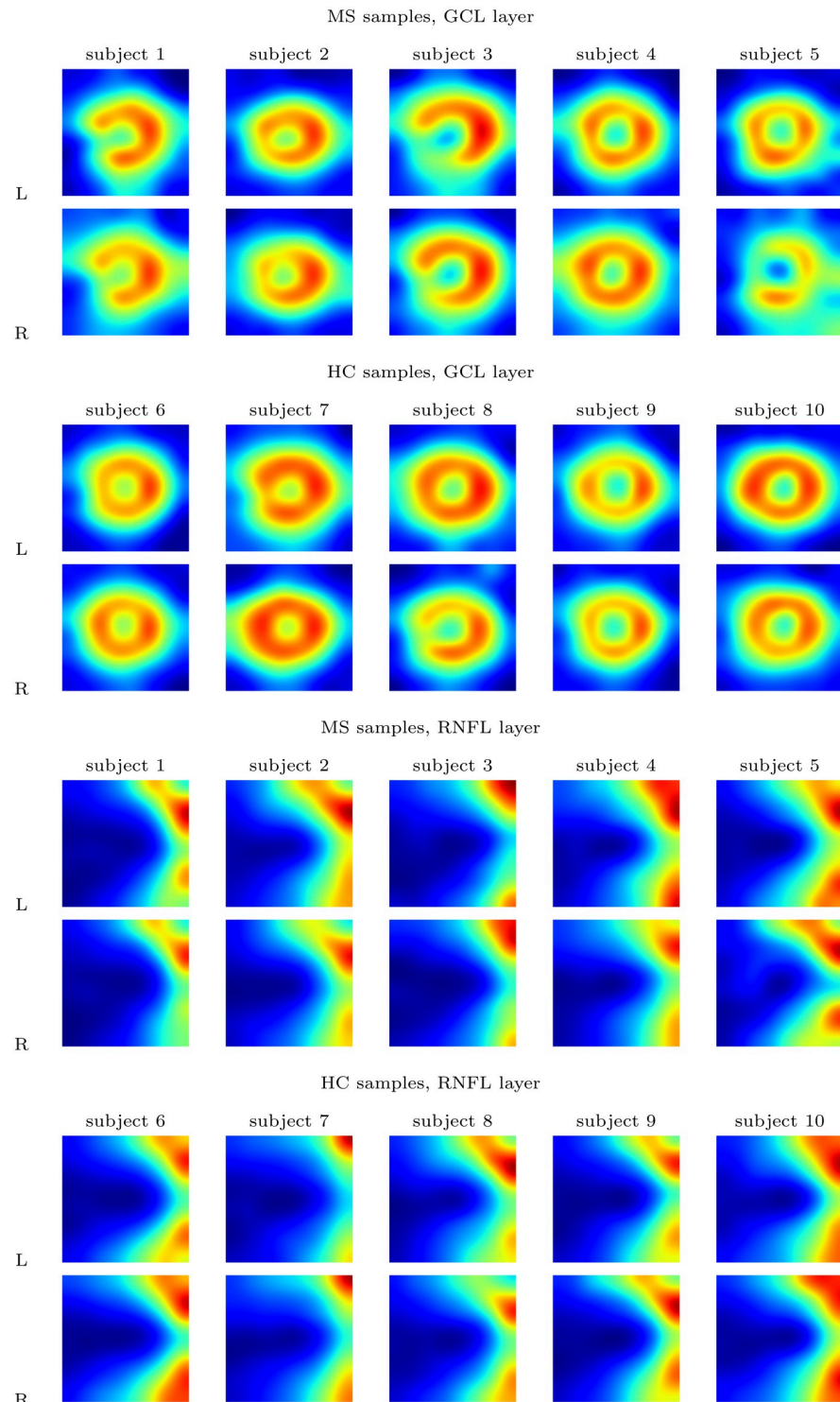


Fig 3. Selection of illustrative samples of the Posterior pole (PPole) grid values found in our database. The first row in each group shows the data acquired from the left eye while the second row shows the data from the right eye. Bicubic interpolation is used for enhancing visualization.

<https://doi.org/10.1371/journal.pone.0289495.g003>

The most relevant hyperparameters of gradient boosting are the learning rate (learning rate), maximum tree depth (max depth), number of trees to fit (n estimators), and L2 regularization weight (reg lambda). We performed hyperparameter selection using a Bayesian search with *optuna* package (<https://optuna.org/>). We found that the hyperparameter selection greatly improved the validation accuracy but the default hyperparameters provided a system whose test accuracy is close to that of the best-performing models in validation. For the sake of reproducibility, XGBoost has been executed in our study with the default hyperparameter values: learning rate = 0.3, max depth = 6, n estimators = 100, and reg lambda = 1.

2.2.2 Random Forests (RF). Random Forests (RF) is a model ensemble that consists of a large number of decision trees. Each individual tree in the forest performs one class prediction, and the class with the majority of the votes becomes the model prediction. The trees in a forest are uncorrelated models that together produce ensemble predictions that are more accurate than any of the individual predictions. Uncorrelation is the key to a successful ensemble. It is obtained from random selection of different sets of data and different features for each tree. RF has been proposed in different applications for building computer-aided diagnosis systems [32, 35]. The runner-up of the diagnosis forecast in the TADPOLE Challenge was ThreeDays, a random forest (RF) machine.

The most relevant hyperparameters of RF are the method for sampling data points (bootstrap), maximum number of levels in the tree (max depth), number of features to be considered at every split (max features), minimum number of samples required at each leaf (min samples leaf) and to split a node (min samples split), and number of trees in the forest (n estimators). As with XGB, we performed hyperparameter selection using *optuna*. We also found that the default hyperparameters provided the same accuracy as those of the best-performing model. Therefore, RF has been executed in our study with the default hyperparameter values: bootstrap = True, max depth = None, max features = auto, min samples leaf = 1, min samples split = 2, n estimators = 100.

2.2.3 Explainable Boosting Machine (EBM). Explainable Boosting Machine (EBM) is a glassbox model. It was designed with the idea of obtaining accuracies comparable to state-of-the-art ML methods like boosted trees and random forests, while being highly intelligible and explainable. EBM fits under the family of Generalized Additive Models (GAM)

$$g(E[y]) = \beta_0 + \sum f_j(x_j), \quad (1)$$

where the feature functions f_j are estimated using small trees in a way that it resembles gradient boosting. The boosting procedure is restricted to train one feature at a time in all vs all fashion using a very low learning rate so that feature order does not matter. Thus, the feature functions are composed of sums of small trees. Therefore, many complicated functions can be modeled very accurately due to the versatility of trees. In addition, this way of training mitigates the effects of co-linearity. The best feature functions f_j provide measurements of how each feature contributes to the model's prediction of the problem yielding the explainability of the model. Indeed, EBM can automatically detect and include pairwise interaction terms by using models of the form

$$g(E[y]) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j) \quad (2)$$

which further increases the accuracy while maintaining explainability in an analogous way. In this work, we will refer with EBM to the model given by Eq 1 and with EBM + i to the model with interactions given by Eq 2.

In general, the default hyperparameters for EBM should perform reasonably well on most problems. The authors recommend in <https://interpret.ml/docs/faq.html> to train a model with

the defaults and use explainability to catch features with abnormal behavior. Using the defaults, we obtained a performance comparable with XGB and RF and we did not detect any problem requiring hyperparameter tuning.

2.3 Explainable machine learning

Explainable AI (xAI) is a recent subfield of AI that aims to provide explanations of general machine learning algorithms. xAI methods are intended to be combined with the most complex families of methods that have been traditionally considered as black boxes [36, 37]. The objective of xAI is to overcome the trade-off traditionally existing between accuracy and explainability and provide both powerful and explainable systems with arguments for increasing the confidence in the output of the algorithms.

Explainability is an important first step toward fully usable and trustworthy AI systems. The most important current xAI techniques facilitate the establishment of the importance given to the different features by the system, assessing how a particular feature affects model predictions, or which feature values favor or hinder correct or incorrect predictions. All this information can help ensure whether only significant features coherent with current knowledge are used to infer the output for the application of interest, or, on the other hand, to assess whether the system is biased or the experimental design is flawed.

SHapley Additive exPlanations (SHAP) has attracted increasing attention in the field of xAI (<https://shap.readthedocs.io>). The core idea of SHAP is to transfer ideas from cooperative game theory to the attribute feature importance of a model output given an input [38]. SHAP values represent the distribution of the contributions toward game success or failure amongst all the players working in cooperation.

In the context of explainability of machine learning methods, SHAP values represent the change in each feature in the expected model prediction under conditioning on that feature. The explainability from SHAP is achieved through different graphical representations derived from the SHAP values. In this work, we will represent the mean of the SHAP values as box plots, and represent the impact of feature values on the probability of the MS class using violin plots and partial dependence plots. In addition, SHAP allows providing local explanations on the decisions taken by the models. In this work, we will use waterfall plots. In <https://www.kaggle.com/code/meliao/shap-on-titanic-why-is-rose-alive-but-jack-not> we can find an easy to follow example on the use of SHAP and the graphical representation of the SHAP values for understanding the survival chances of Rose and Jack in Titanic's tragedy.

In EBM, the contribution of each feature to a final prediction can be visualized and understood by plotting f_j or $f_{i,j}$. Since EBM is an additive modeling, each feature contributes to predictions in a modular way that makes it easy to reason about the contribution of each feature to the prediction (similarly to SHAP). For example, term contributions can be sorted and visualized to show which features have the most impact on any individual prediction.

2.4 Experimental design

Ideally, we would like to work with a train/validation/test split. However, our data set is small. Therefore, it is more convenient to evaluate the performance of our methods and study the explainability of our models using ten-fold cross-validation splits. Thus, the original set was divided into ten different train-test splits of subjects with 80% – 20% proportionality. This means that if a subject falls into the train/test set of a given fold, the eyes for this subject belong to the train/test set. Since we were not separating a hold-out test from the very beginning of the process, we were extremely careful with our experimental setup in order to avoid any kind of data leakage.

Since each subject can contribute to the dataset with one or two eyes, we generated four different datasets and assessed the performance of the corresponding models: left eye (L), right eye (R), randomly selected eye (rand), and both eyes (LR) (independent sample assumption). For the randomly selected eye models, random selection was only performed for subjects with both eyes. We carried out a comparison of the performance metrics and selected the most appropriate models for our interpretability study, after making sure that there was not statistical significance with the other candidates. We ended up with ten different models with different accuracies for assessing explainability. We opted for showing the explainability results for the best-performing fold. The [S1 File](#) shows the explainability results for the worst-performing fold.

Although the PPole grid features have a two-dimensional structure, we decided to represent them as tabular data, given the low resolution and the small number of samples. Future work will address the problem for 2D.

Both XGB, RF, and EBM require little data preparation in contrast to other methods like neural networks (e.g. no normalization). In addition, the default parameters should perform reasonably well on most problems. As we mentioned in Section 2, preliminary exploration with Bayesian search did not provide a better hyperparameter setting than default values. We found similar results in a previous study on the Alzheimer's Disease Neuroimaging Initiative (ADNI) [30]. Since further dividing our train set into train and validation sets would aggravate the small sample problem, and there would be a combinatorial explosion in the number of experiments, we decided to use the default hyperparameters.

Finally, we considered both undersampling and oversampling for dealing with the problem of data unbalancing. Preliminary exploration indicated that oversampling using SMOTE on the minority group (MS) was preferable. It should be noticed that the test set should be unbalanced, preserving the actual distribution of the overall population of our dataset. Otherwise, we have the risk of providing an overestimation of the true capacities of the models in the evaluation.

Our methods were implemented in Python 3.6 using the libraries in *scikit-learn* (<https://scikit-learn.org>). For XGB and RF, SHAP values were obtained from tree explainers, a fast implementation of SHAP for tree-based methods. Our experiments were run on a machine with an Intel Core i7 working at 3.70 GHz and 32 GBS of DDR3 RAM.

3 Results

3.1 Performance study

The performance in this study has been measured from the metrics typically used in ML studies that are computed from the number of true positives, true negatives, false positives, and false negatives in the test set. These metrics are the accuracy, sensitivity, specificity, F1-score, and area under the curve (AUC). [Fig 4](#) shows the distribution of the obtained results for the populations with left eyes (L), right eyes (R), randomly selected eyes (rand), and both eyes (LR). The [S1 File](#) gathers the results obtained through the ten-fold cross-validation experiments. From the analysis of the plots, we identify the most remarkable results:

- The models built with the left eyes show in general a performance lower than the models built with the right eyes.
- The models built with randomly selected eyes show a performance similar to the models built with both eyes.
- The variance in accuracy is close to or greater than 5% in the majority of models.

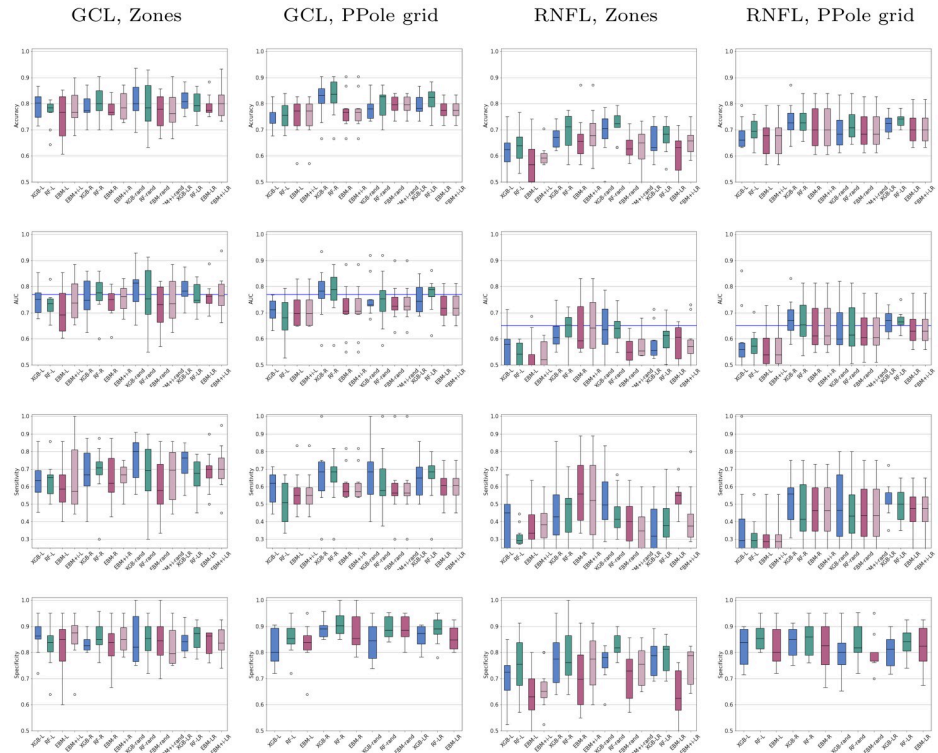


Fig 4. Distribution of the performance metrics in the test sets of the ten-fold cross-validation experiments. For each retinal layer, the left column shows the results obtained with the Zones feature set. The right column shows the results obtained with the Posterior Pole (PPole) grid feature set. L indicates the model built with the population of left eyes. R indicates the model built with right eyes. Rand indicates the model built with randomly selected eyes. LR indicates the model with both eyes. Blue boxes show the results obtained with gradient boosting (XGB), green with random forests (RF), and magenta with explainable boosting machine (EBM). The horizontal line in the area under the curve (AUC) plot indicates the test set baseline provided in [31].

<https://doi.org/10.1371/journal.pone.0289495.g004>

- The use of the GCL layer measurements as a feature set greatly outperforms the use of the RNFL layer.
- For the GCL layer,
 - The Zones feature set provides a slightly improved performance with respect to the PPole grid feature set.
 - The sensitivity indicates that some models obtain a low accuracy in the diagnosis of MS. This could be due to the small size of the dataset and the unbalanced nature of the sample.
 - The specificity indicates that the models are in general good in the identification of HC.
 - No method consistently stands out in performance above the rest.

Welch’s t-test on the GCL measurements and the Zones feature set did not find any statistical significance between accuracies. The same test on the PPole feature set showed statistical significance between L and R eyes for XGB ($p = 0.02$) and RF ($p = 0.02$), L and rand eyes for RF ($p = 0.07$), EBM ($p = 0.09$), and EBM + i ($p = 0.07$), and L and LR eyes for XGB ($p = 0.03$) and RF ($p = 0.01$). Therefore, no statistical significance was found between rand and LR eyes.

Regarding the rule of thumb of having a number of samples per class of at least ten times the number of features, we would like to remark that the Zones feature set meets this

requirement. Although the PPole grid feature set does not meet the requirement, it seems that the performance is not greatly affected by the ML methods used in this study. We observed similar behavior in our previous work for XGB and RF with the Alzheimer's Disease Neuroimaging Initiative (ADNI) tabular data [30].

Comparing our AUCs with the baseline provided in the test set in Kenney et al. [31] for the average of GCL+ thickness from both eyes, we can see that the rand models from the Zones feature set achieve a close mean. It is over the baseline for XGB. LR models from the Zones feature set achieve a comparable mean in all cases. The LR models from the PPole grid feature set slightly perform under with the exception of RF for the LR models. For the RNFL thickness, our rand models for the Zones feature set show a comparable mean for XGB and RF. The LR models from the Zones feature set underperforms the baseline. All the LR models from the PPole grid feature set show a comparable mean. It should be noticed that Kenney et al. used random sampling and the pRNFL thickness instead of RNFL.

Table 2 shows the results of the best-performing random and LR models for each feature set configuration. For the GCL layer, the models built with the random dataset greatly outperformed the models built with the LR dataset for the Zones feature set in XGB and RF. The remaining models performed similarly regardless of the use of random or LR datasets. Therefore, in the following, we proceed with the interpretability study for the models built with the LR dataset since the performance is similar and it may give us more opportunities to analyze interesting cases. Furthermore, the fold identifier (id) of the best-performing model is the same for all three methods, and therefore the results can be compared fairly. With the exception of Subsection 3.2.6, our interpretability will focus on assessing the measurements obtained on the GCL layer.

3.2 Explainability study

3.2.1 Global SHAP-based explainability. Fig 5 shows different representations of the average SHAP values for the best-performing XGB and RF models with the Zones feature set. These representations include horizontal bar plots of the SHAP values ranked by order of importance, violin plots with the distribution of the SHAP values colored by the magnitude of the feature values, and partial dependence plots for assessing the marginal effect of the two most important features on the predicted outcome of the ML models.

For both XGB and RF, zones Z1 and Z2 are the most relevant for the discrimination of MS from HC. The order of importance is similar. The violin plots show that medium and low values of the average thicknesses contribute to a higher probability for MS. The only exception is for Z6 in the case of XGB, where high values of the average thickness contribute to a higher probability for MS. Our clinical experts indicated that this zone corresponds with the most inferior and temporal paramacular area. This is a zone with medium and large size blood vessels (see Fig 2) and, therefore, GCL measurements may be less accurate. The partial dependence plots show that the slope of the distribution of the SHAP values for Z1 is negative. This means that high values of the average thickness contribute to a lower probability for MS, corroborating the violin plot for Z1. The same happens with Z2.

Analogously for PPole grid features, Fig 6 shows different representations of the average of the SHAP values for the best-performing XGB and RF models. In addition to the bar, violin, and dependence plots, we show the SHAP values on the 8×8 grid for a finer location of the anatomical part identified with our models as mostly affected by the disease.

For both XGB and RF, features 5.4 and 5.3 are the most relevant for the discrimination of MS from HC. These features are located at Z1 and Z2 zones, which were marked by SHAP as the most important ones in the corresponding models (Fig 5). The violin plots show that low

Table 2. Results of the best-performing random and LR (both eyes) models for each feature set configuration considered in this work. We indicate the id of the training and test set used to generate the model and obtain the metrics.

Random								
Method	Layer	Feature set	fold id	Acc (%)	Sens (%)	Spec (%)	F1-score (%)	AUC
XGB	GCL	Zones	5	93.54	90.90	95.00	90.90	0.92
XGB	GCL	PPole grid	5	87.09	81.81	90.00	81.81	0.85
RF	GCL	Zones	1	92.85	87.50	95.00	87.50	0.91
RF	GCL	PPole grid	5	87.09	81.81	90.00	81.81	0.85
EBM	GCL	Zones	1	85.71	60.00	100.00	66.66	0.75
EBM	GCL	PPole grid	4	83.33	100.00	80.00	66.66	0.90
EBM + i	GCL	Zones	5	90.32	81.81	95.00	85.71	0.88
EBM + i	GCL	PPole grid	5	83.87	81.81	85.00	78.26	0.83
XGB	RNFL	Zones	1	78.57	75.00	80.00	66.66	0.77
XGB	RNFL	PPole grid	5	77.41	72.72	80.00	69.56	0.76
RF	RNFL	Zones	0	79.31	66.66	82.60	57.14	0.74
RF	RNFL	PPole grid	5	83.87	72.72	90.00	76.19	0.81
EBM	RNFL	Zones	0	72.41	50.00	78.26	42.85	0.64
EBM	RNFL	PPole grid	0	82.75	66.66	86.95	61.53	0.76
EBM + i	RNFL	Zones	4	73.33	60.00	76.00	42.85	0.68
EBM + i	RNFL	PPole grid	4	76.66	80.00	76.00	53.33	0.78
LR								
Method	Layer	Feature set	fold id	Acc (%)	Sens (%)	Spec (%)	F1-score (%)	AUC
XGB	GCL	Zones	6	85.00	78.57	86.95	70.96	0.82
XGB	GCL	PPole grid	5	86.66	80.00	90.00	80.00	0.85
RF	GCL	Zones	5	86.66	75.00	92.50	78.94	0.83
RF	GCL	PPole grid	5	88.33	80.00	92.50	82.05	0.86
EBM	GCL	Zones	5	88.33	90.00	87.50	83.72	0.88
EBM	GCL	PPole grid	5	83.33	75.00	87.50	75.00	0.81
EBM + i	GCL	Zones	5	93.33	95.00	92.50	90.47	0.93
EBM + i	GCL	PPole grid	5	86.66	70.00	95.00	77.77	0.82
XGB	RNFL	Zones	5	75.00	28.57	89.13	34.78	0.58
XGB	RNFL	PPole grid	5	78.33	55.00	90.00	62.85	0.72
RF	RNFL	Zones	0	75.00	35.71	86.95	40.00	0.61
RF	RNFL	PPole grid	5	78.33	65.00	85.00	66.66	0.75
EBM	RNFL	Zones	0	71.66	57.14	76.08	48.48	0.66
EBM	RNFL	PPole grid	5	81.66	65.00	90.00	70.27	0.77
EBM + i	RNFL	Zones	4	68.33	80.00	66.00	45.71	0.73
EBM + i	RNFL	PPole grid	1	76.66	60.00	85.00	63.15	0.72

<https://doi.org/10.1371/journal.pone.0289495.t002>

average thicknesses contribute to a higher probability for MS. Some exceptions to the rule can be identified for features located at the boundary points (e.g. feature 8.4). The partial dependence plots show a negative slope corroborating the relationship between high thickness values and low probability values for MS given by the models. The grids show that RF feature importance is distributed among features 5.4 and 5.3 and some features located over the superior and nasal superior locations. Therefore, RF seems to realize the 2D geometry of the data. It is worth noticing that feature 3.1 is considered as important by both methods. Our clinical experts found a plausible interesting interpretation of this finding that is discussed in Section 4. However, we do not rule out either a bias of our models due to the characteristics of our dataset.

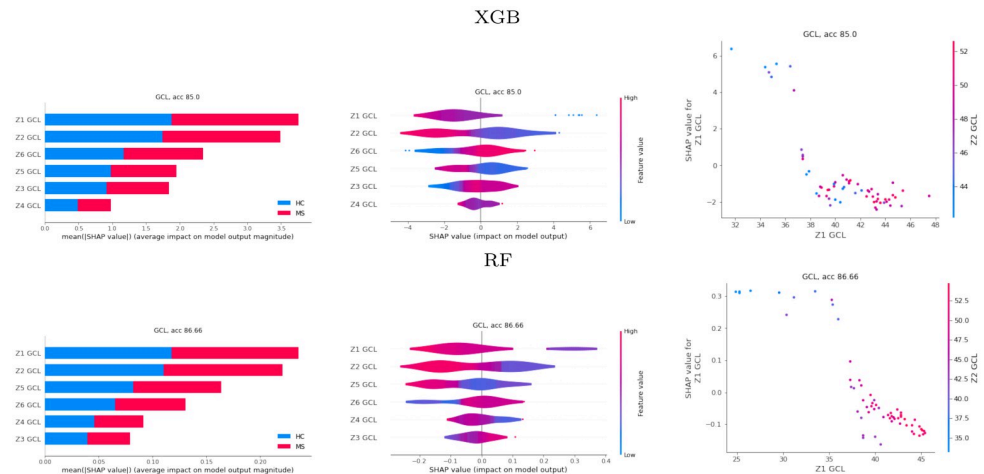


Fig 5. Global SHAP results. SHAP results for the gradient boosting (XGB) and random forests (RF) best models obtained with the Zones feature set. Left, horizontal bar plots of the SHAP values. Center, violin plots. Right, partial dependence plot for the two most important features.

<https://doi.org/10.1371/journal.pone.0289495.g005>

3.2.2 Local SHAP-based explainability. Figs 7 and 8 show the local explainability associated with the failed test set subjects for the best-performing XGB and RF models and PPole grid feature set. The S1 File shows the local explainability for the worst-performing XGB and RF models. For each sample we show the PPole grid data, the grid of the corresponding SHAP values, and the waterfall plot, which indicates the contribution of each feature toward (red) or against (blue) MS. The corresponding confusion matrices can be found in Fig 9.

Analyzing the MS subjects that were incorrectly classified as HC for XGB, we can see that three of the MS subjects show a shape typically found in the HC set (test set ids 5.1, 5.16, and 5.17). Surprisingly, the 5.13 subject shows the typical break observed in the MS set, but it was classified as HC. From the corresponding waterfall plot, we can see that feature 5.3 contributed to the probability of MS, but it was neutralized by feature 5.4. The same happened with the following meaningful features. In the end, less meaningful features against the correct classification contributed to the final outcome. The probability for HC is low, therefore the decision made by the model should not be trusted.

From the HC subjects incorrectly classified as MS by XGB, we can see that subjects 5.28 and 5.35 show the typical torus shape break or a thinning in feature 5.4. Subjects 5.47 and 5.52 show a similar SHAP grid, and it seems that features 5.3, 5.4 and 5.6 were mostly responsible of the erroneous classification. Although the overall shape is typically observed in the HC group, considering the thickness values individually may explain the error in classification.

Similarly analyzing the subjects that were incorrectly classified as HC for the best-performing RF model, we find again subjects 5.1, 5.16, and 5.17 that showed a shape typically found in the HC group. Subject 5.6 shows the typical torus shape break and feature 5.3. contributed to the probability of MS. However, the remaining 55 features surprisingly summed up against the correct classification. This may be an indication that the decision should not be trusted.

From the HC subjects that were incorrectly classified as MS, subjects 5.35 and 5.52 match with the XGB subjects. Subject 5.39 shows a torus thinning in feature 5.6, which was considered among the most important features. The waterfall plot shows oscillating contributions, which explains why it ended up in the incorrect outcome. Our clinical experts indicated that this torus configuration is typical from early glaucoma. The probabilities estimated by the model are low in all cases, therefore the decisions made by the model should not be trusted.

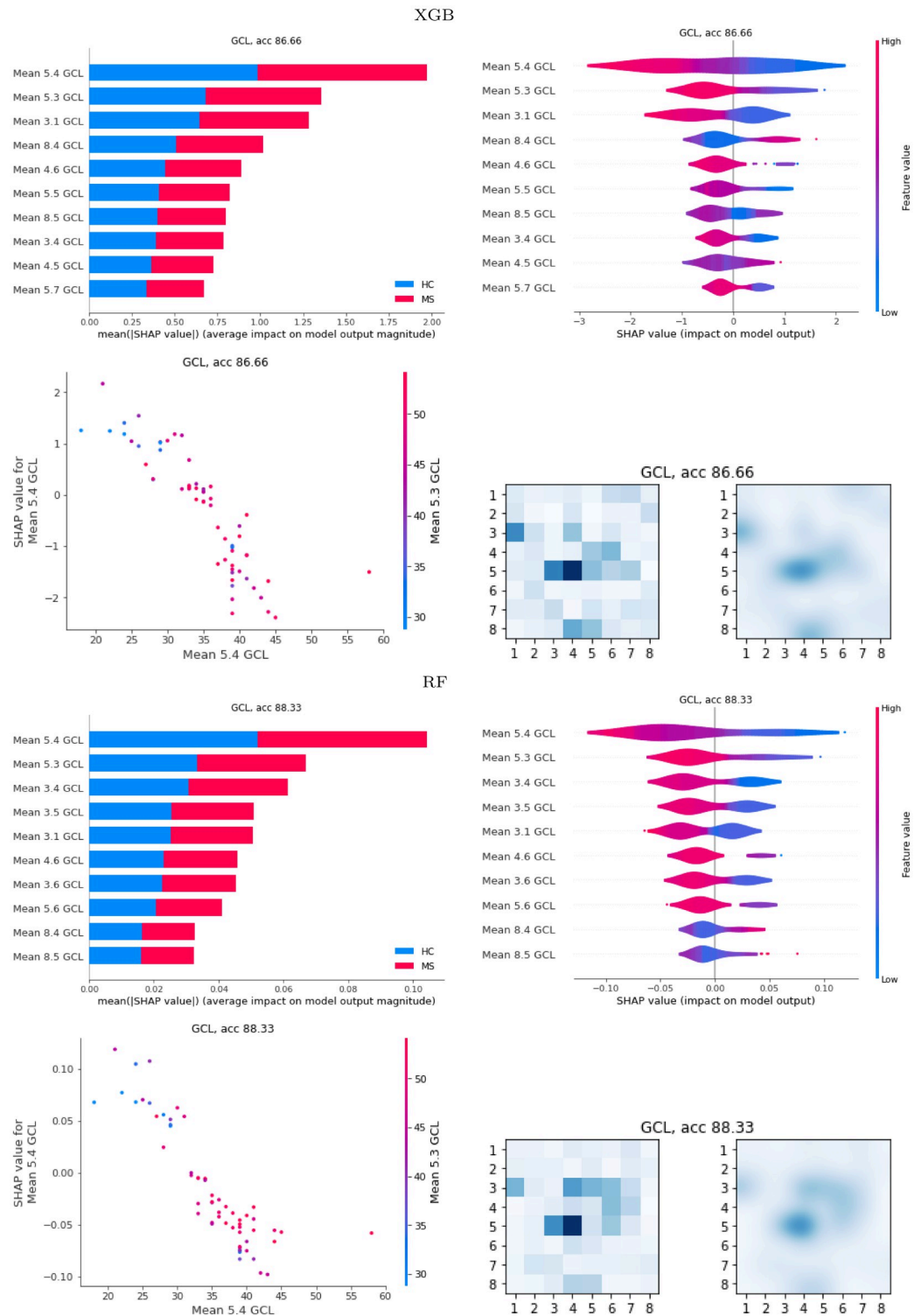


Fig 6. Global SHAP results. SHAP values for the gradient boosting (XGB) and random forests (RF) best models obtained with the posterior pole (PPole) grid feature set. The grid showing the SHAP values obtained in each spatial location is shown in its original resolution (left) and after bicubic interpolation (right).

<https://doi.org/10.1371/journal.pone.0289495.g006>

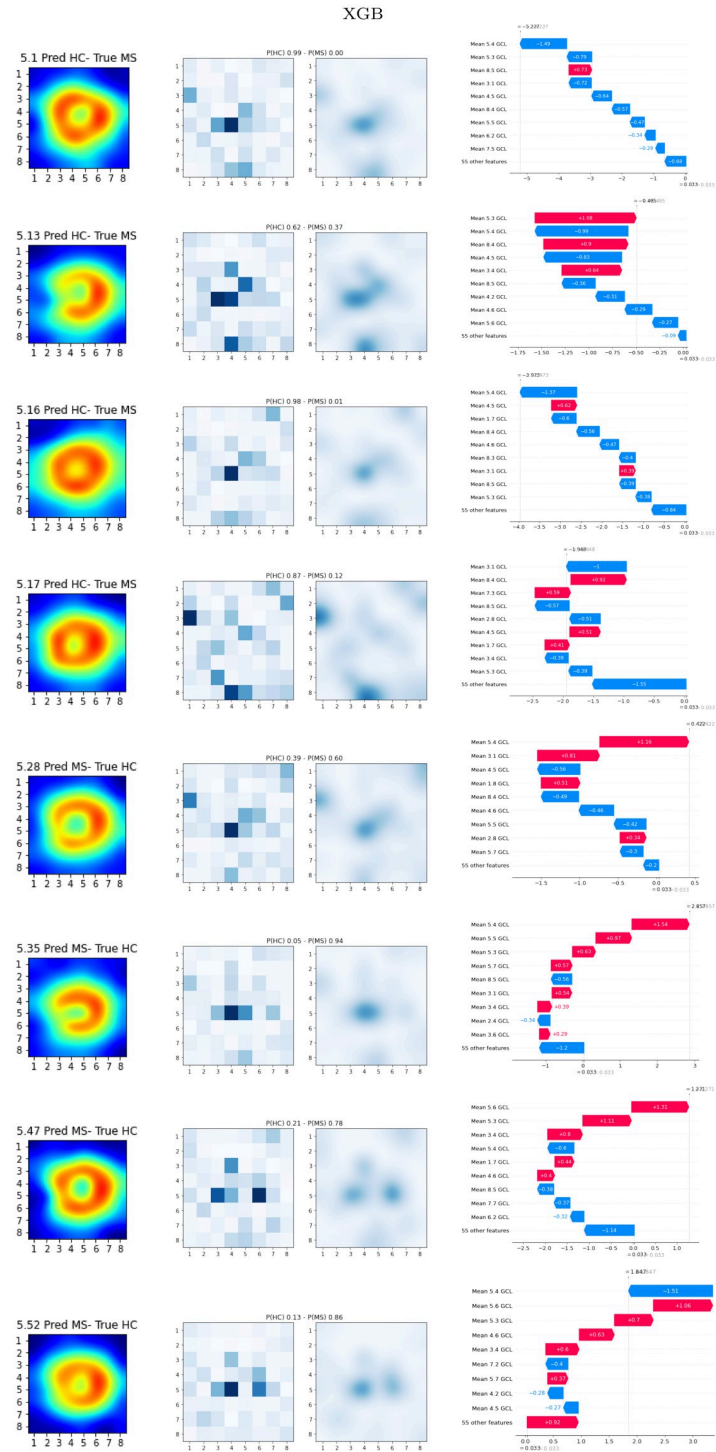


Fig 7. Local SHAP results. Gradient boosting (XGB) best-performing model. Left, Posterior Pole (PPole) grid sample. Middle, grid of the local SHAP values. Right, waterfall plot for assessing feature contribution towards or against multiple sclerosis (MS).

<https://doi.org/10.1371/journal.pone.0289495.g007>

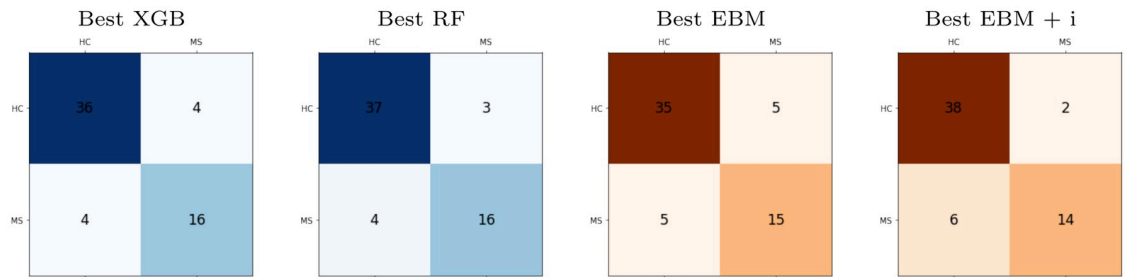


Fig 9. Confusion matrices for the best models. Gradient boosting (XGB), random forests (RF), explainable boosting machine (EBM), and EBM with interactions (EBM + i) models with the PPole grid feature set.

<https://doi.org/10.1371/journal.pone.0289495.g009>

As with XGB and RF, zones Z1 and Z2 are the most relevant for the discrimination of MS from HC. The order of importance matches the order of XGB. The curves show that high values of Z1 contribute to a lower probability of MS. The curves from Z2 show oscillating contributions from low to medium values, while the highest values contribute to a lower probability of MS. The pattern is similar for Z3 and Z4 while it is reversed for Z5 and Z6. These results are analogous to the violin plots shown in Fig 5.

Fig 11 shows the scores for the best-performing EBM model with the PPole grid feature set. Additionally to the bar plots, we show the score values on the 8 × 8 grid. As with XGB and RF, features 5.4 and 5.3 are in the top-three of relevance for the discrimination of MS from HC. It drives our attention that features in the boundary of the grid such as 8.5, 3.1, and 8.4 occupy the top positions. The grid shows that feature importance is distributed among features 5.4 and 5.3 and the torus-like shape. However, feature importance is also located in two points of the boundary. The curves for features 5.4 and 5.3 show that high values contribute to a low



Fig 10. Global explainable boosting machine (EBM) results. Scores for the best EBM model obtained with the Zones feature set. Up, horizontal bar plots of the scores. Down, curves of the score values depending on each feature value.

<https://doi.org/10.1371/journal.pone.0289495.g010>

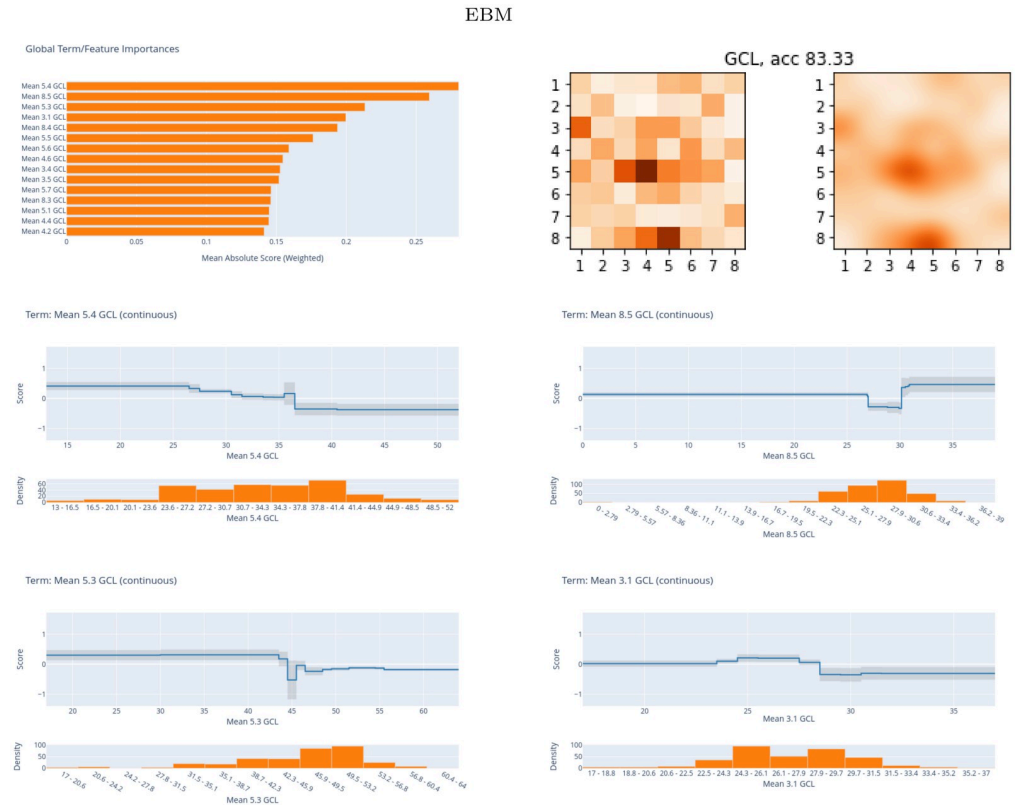


Fig 11. Global explainable boosting machine (EBM) results. Scores for the best EBM model obtained with the posterior pole (PPole) grid feature set. Up, horizontal bar plots of the scores and grids. Down, curves of the score values of the top-four feature set.

<https://doi.org/10.1371/journal.pone.0289495.g011>

probability of MS. The same happens with the curves for feature 3.1. On the contrary, the curves for feature 8.5 show that high values contribute to a high probability of MS. We find again a possible bias of the model that considers this location useful for the discrimination between groups.

3.2.4 Local EBM-based explainability. Fig 12 shows the local explainability associated with the failed test set subjects for the best-performing EBM model and PPole grid feature set. For each sample, we show the PPole grid data, the 2D grid of scores, and the local importance scores. The corresponding confusion matrix can be found in Fig 9. The S1 File shows the local explainability for the worst-performing EBM model.

Analyzing the MS subjects incorrectly classified as HC for EBM, we find subjects 5.16 and 5.17 that showed a shape typically found in the HC group. Subjects 5.11, 5.12, and 5.13 should be clearly identified as MS patients, but the importance given by the model to the boundary points 7.8, 8.3, and 8.4 contributed to an incorrect decision. For the HC subjects incorrectly classified as MS, subject 5.21 has been incorrectly classified using information from feature points 7.8, 8.5, and 8.4 and subject 5.35 from point 1.1. The remaining subjects show patterns observed in the MS population, and the model attributed the greatest importance to the locations where these patterns are found, e.g., subject 5.38 and features 6.5 and 3.5; subject 5.39 and features 6.6. and 5.6; subject 5.52 and features 5.4, 4.6, and 5.3.

3.2.5 EBM with interactions. Fig 13 shows the average of the scores for the best performing EBM + i model with the Zones and PPole grid feature sets. While for the Zones feature set

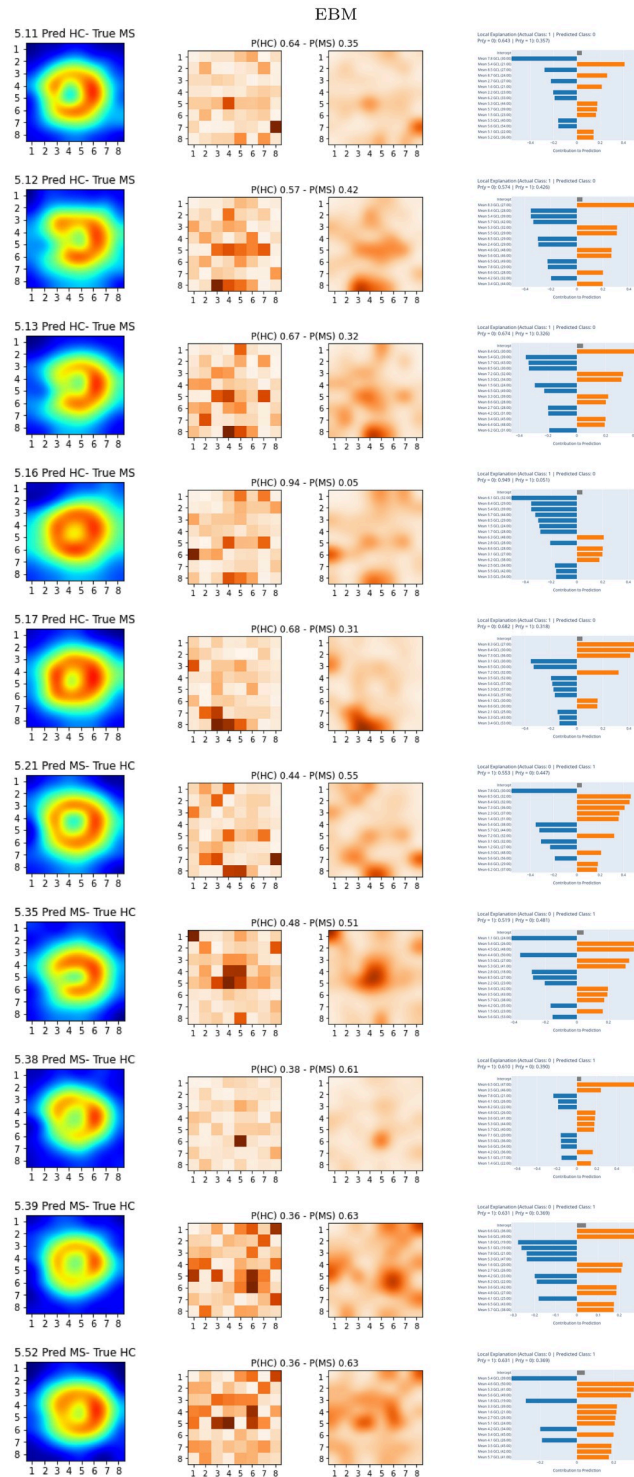


Fig 12. Local explainable boosting machine (EBM) results. EBM best-performing model. For each subject, left figure shows the posterior pole (PPole) grid sample, middle figure shows the grids, and right figure shows the feature contribution towards or against multiple sclerosis (MS).

<https://doi.org/10.1371/journal.pone.0289495.g012>

the importance rank is similar between EBM and EBM + i models, pairwise interactions occupy the top-ten positions for the EBM + i model with the PPole grid feature set. Fig 13 also shows the interaction plot for the top-four positions. The top-one interaction is between points 4.5 and 5.4. This is consistent with the explanations given by XGB and RF. The interaction plot shows that for low values of thickness in 4.5 and 5.4 the probability of MS is high, as expected. For low values in 5.4 and high values in 4.5, the probability of MS is high. This means that the model uses the torus shape break or thinning in this location for taking the decision. On the contrary, for high values in 5.4 and low values in 4.5, then the probability of MS is low. Finally, for high values in 5.4 and 4.5 the probability of MS is nearly zero. In this case, these features are not relevant for the model to distinguish between groups.

The second interaction is between the boundary points 3.1 and 8.4. As happened in the case of the top-one interaction, for low values of both features, the probability of MS is high, and for high values of both features the probability of MS is nearly zero. For low values in 8.4 and high values of 3.1 the probability of MS is low. For high values in 8.4 and low values in 3.1 the probability of MS is high. Recalling the suspected bias found in points 8.4 and 8.5 for the model without interactions, it seems that in this case the bias is shown in this interaction of highly unlikely related features. The same reasoning can be applied to the interaction between points 2.1 and 5.4 and between points 4.5 and 8.3.

Fig 14 shows the local explainability associated with the failed test set subjects for the best-performing EBM + i model and PPole grid feature set. Both EBM and EBM + i models failed

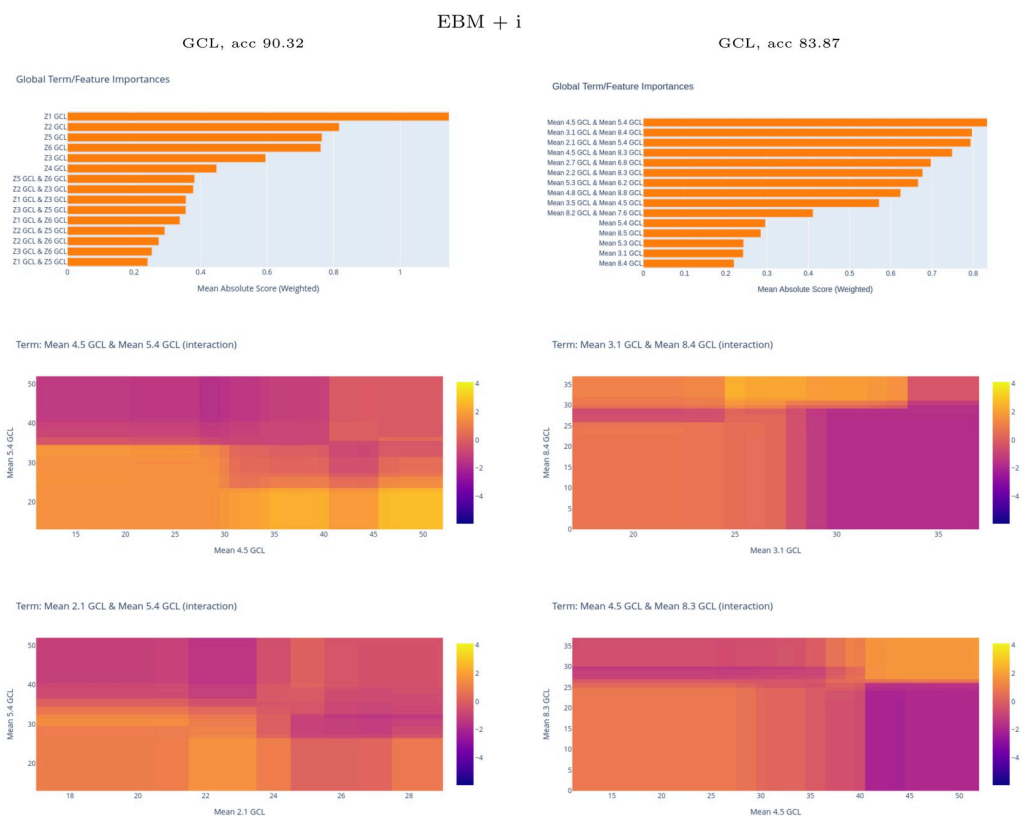


Fig 13. Global explainable boosting machine with interactions (EBM + i) results. Up, horizontal bars of the scores for the best EBM + i model obtained with Zones and posterior pole (PPole) grid feature sets. Middle and down, interaction maps of the top-four feature set for the PPole grid feature set.

<https://doi.org/10.1371/journal.pone.0289495.g013>

in subjects 5.12, 5.13, 5.16, and 5.39. It seems that the most important features involving boundary points were the responsible of the failed decisions.

3.2.6 Global explainability: Layers and models. To finish our study, we provide a comparison of the global SHAP values and EBM scores from the PPole grid feature set for the GCL and the RNFL layers. Figs 15 and 16 show the grids obtained for the models generated with the ten test sets considered in our work. For the GCL feature set, all the models give consistently importance to the location of the 5.4 feature. For XGB this seems to be the single most remarkable feature in the majority of models. RF shows importance in several locations of the torus ridge. EBM also shows importance in several locations of the torus ridge but it is also giving importance to other locations. We realize that the majority of models find feature 3.1 as important, and this becomes more apparent for XGB and EBM. EBM tends to remark as important many more points on the boundary.

For the RNFL feature set, all models consistently point out to feature 5.4. Indeed, we can appreciate two or three diagonally aligned points consistently remarked by all the models: feature 5.4, 6.5, and 7.6. These points correspond to the locations of some of the nerve bundle fibers (see Fig 1 from [39]). From the boundary points, feature 1.8 is remarked as important by all the models. It seems that the system sees some relevant information that can be used to

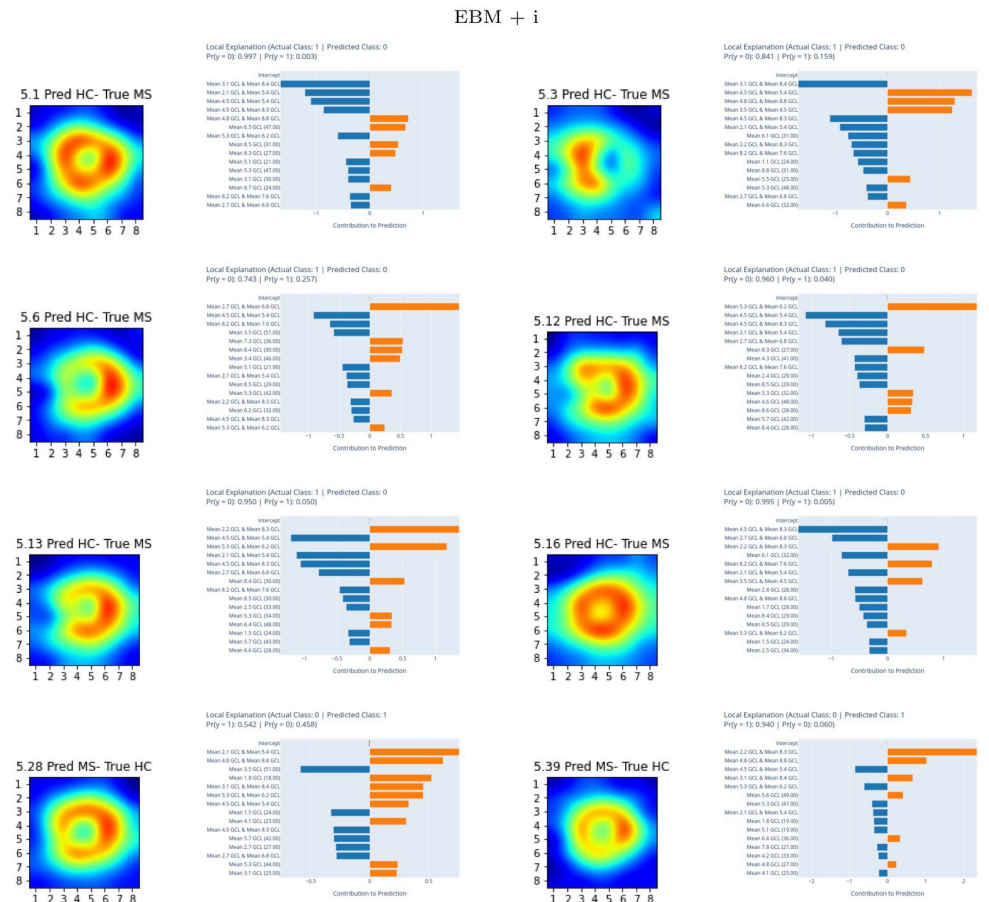


Fig 14. Local explainable boosting machine with interactions (EBM + i) results. EBM + i best-performing model. For each subject, left figure shows the posterior pole (PPole) grid sample and right figure shows the feature contribution towards or against multiple sclerosis (MS).

<https://doi.org/10.1371/journal.pone.0289495.g014>

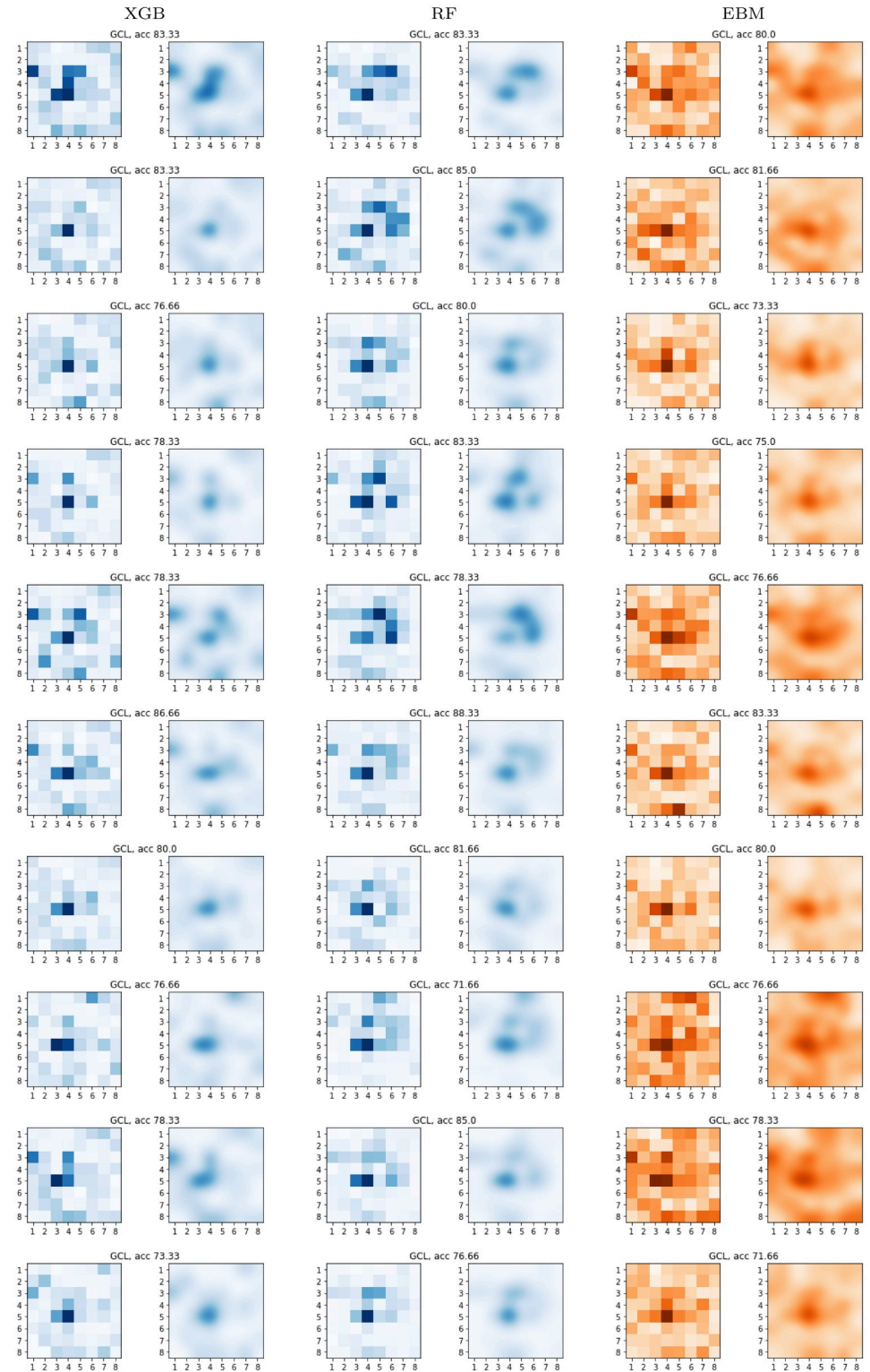


Fig 15. Global SHAP and explainable boosting machine (EBM) results. Ganglion cell layer (GCL). SHAP and EBM grids obtained with the models of the ten-fold cross-validation.

<https://doi.org/10.1371/journal.pone.0289495.g015>

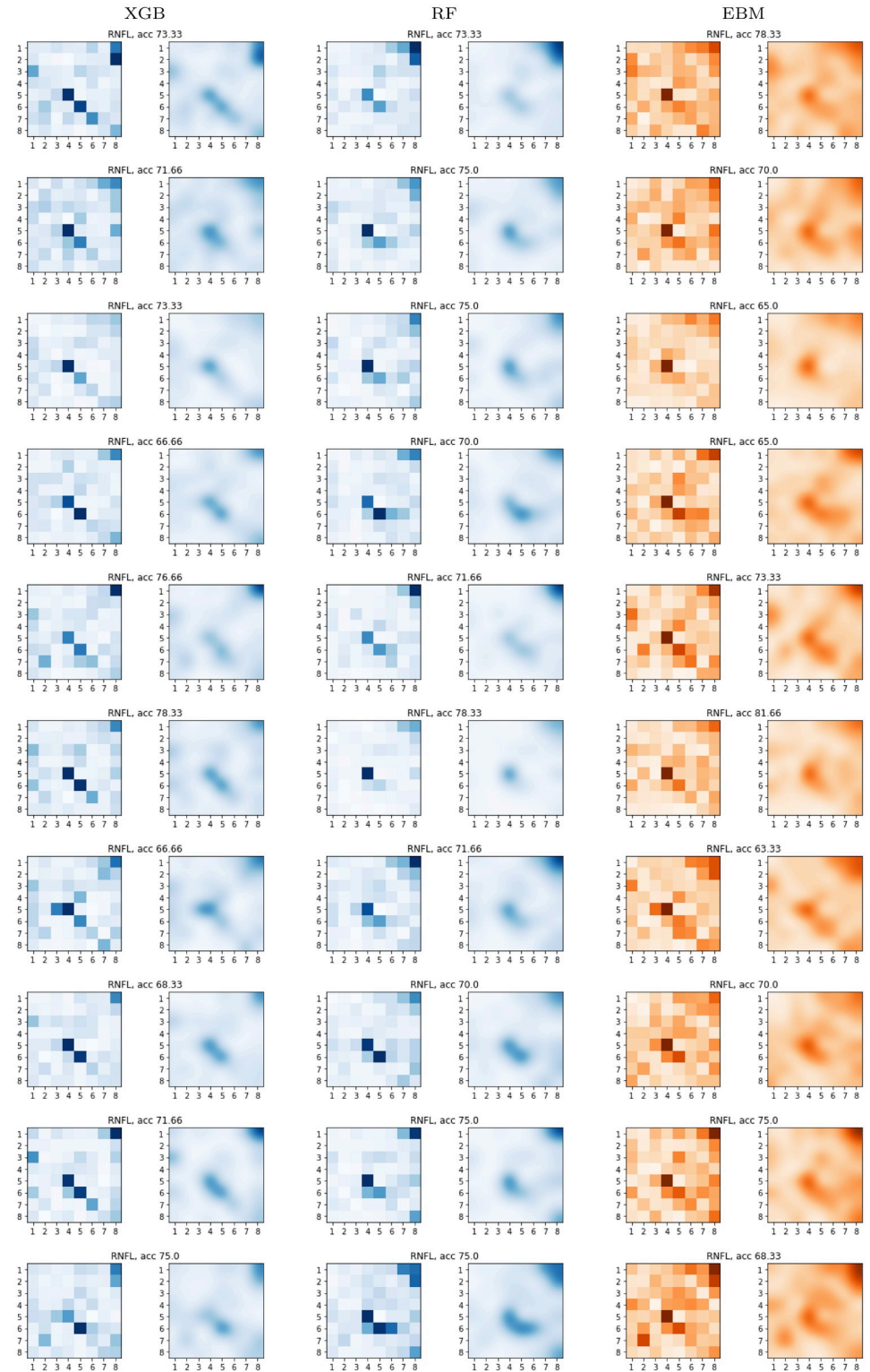


Fig 16. Global SHAP and explainable boosting machine (EBM) results. Retinal nerve fiber layer (RNFL). SHAP and EBM grids obtained with the models of the ten-fold cross-validation.

<https://doi.org/10.1371/journal.pone.0289495.g016>

discriminate between both groups. Our clinical experts indicated that feature 1.8 corresponds to the exit zone of the papillomacular bundle. This bundle is made up of ganglion cell axons which, coming from the most central area of the retina (macula), reach the temporal border of the optic nerve by means of a practically rectilinear fascicle. It is therefore highly logical that this is an important feature in determining the presence/absence of the pathology.

4 Discussion

The state of the art in ML diagnosis from OCT is made of studies that are focused exclusively on providing high values of performance while eluding a proper discussion on the usability of the models and the trust on the decisions made by the proposed solutions. The major criticism from these works is that these studies have been conducted with a very small number of samples and the results from the best models are shown without analyzing means, medians, and variances. These models have low generalization capability and may be establishing unrealistic baselines. In addition, the methods are used as black-boxes, therefore, we don't know whether the models are taking decisions in accordance to clinical knowledge that we can trust. We believe that there is still a long way to go toward solutions for MS diagnosis widely usable and trustworthy in clinical practice. To ensure the reliability of the systems, the models should be built up on a large dataset, demonstrate a high generalization capability, and yield explanations coherent with clinical knowledge.

Our study is a first step toward a trustworthy solution for computer-aided diagnosis of MS from OCT that will be usable in ophthalmological clinical practice. Our models are built using the thickness of GCL and RNFL layers, measured with the PPole protocol of Spectralis SD-OCT devices. Since every subject can contribute to the dataset with two samples, we have analyzed the performance of the models built with left, right, random, and both eyes. We did not find any significant difference in accuracy between using random or both eyes. We have compared three of the best-performing methods in different applications and challenges with datasets similar to ours: XGB, RF, and EBM. Our models confirmed measurements given in PPole protocol being as informative as other measurements in the discrimination of MS from HC subjects, with performance in the order of the state of the art baseline, with large sample sizes established in [31]. We were able to reach an accuracy greater than 0.9 in some folds for some configurations of feature set, sampling, and methods. We found that the measurements from the GCL layer were much more useful than RNFL in the discrimination of MS from HC subjects. This is in agreement with the latest advances in the state of the art [28, 31]. From these results, we believe that any of the considered configurations of feature set (Zones or PPole grid), sampling (rand or LR), and machine learning method could be used as a starting point for a usable computer-aided solution, once an appropriate large dataset sample has been gathered.

The limitation of our solution is with the applicability to individuals fulfilling the exclusion criteria. For example, eyes longer than 25.2 mm tend to show systematic layer thinning due to extreme deformations, OCT measurements are not reliable, and the images show artifacts. Eyes with acute myopia or astigmatism may show abnormal shapes of GCL and RNFL thickness. Including this kind of data in the training phase of ML models would harm the generalization capability of the systems. It is worth noticing that experienced clinicians would find difficulties in identifying damage due to MS. Eye diseases like glaucoma show a thinning effect on the neuroretinal layers, similarly to MS. In this case, the overall damage cannot be attributed to a single pathology. This is a difficult problem for ML systems.

Our study has revealed the combination of XGB or RF and SHAP as a tandem that may perfectly serve the specialist to make decisions aided by the model explanations. Global

explanations pointed out to zones Z1 and Z2 and points 5.4 and 5.3 (from 64 features), where a visual inspection of the dataset revealed a break or thinning of the torus shape of the GCL measurements. We also found that point 3.1 in the boundary of the PPole grid was mostly considered as a relevant feature. In addition, a reduction of the thickness in these locations was related with a high probability for the MS class. This is consistent with current knowledge on how MS affects neuroretina damage.

The local analysis of the SHAP values for the failed test samples revealed, for the best-performing models, subjects with features of the wrongly predicted class that would also confound an expert in most cases. We also found cases where there is no reason to give a wrong prediction. In these cases SHAP waterfall plots indicated that the most important features contributed to antagonist decisions, and the final decision ended up in less relevant feature values.

EBM without interactions obtained an accuracy similar to XGB and RF. The explanations were given in a similar way to SHAP, which proves that this glass-box method could compete or complement XGB and RF in the task. Global explanations also pointed out zones Z1 and Z2 and points 5.4 and 5.3 with the highest scores. Point 3.1 in the boundary was also considered as a relevant feature together with other boundary points. It seems that, while feature importance is located over the ridge of the torus shape for XGB and RF, it is spread out all over the grid for EBM. Adding pairwise interactions to the model ended up with pairwise explanations as the most relevant features. The interaction pair 4.5×5.4 was revealed as the most important. However, the remaining interactions involved boundary features, which may reinforce our suspicions of bias in EBM models. From these results we believe that EBM may be more sensitive to developing bias than XGB or RF for this kind of data.

Our clinical experts indicated that a plausible pathophysiological explanation for the importance of feature 3.1 could be based on the vascular theory that accompanies neurodegenerative diseases, which argues that axonal loss in these pathologies is associated with a loss of vascular supply, so that both mechanisms (neurodegeneration and vascular loss) are associated in the pathophysiology of MS. Feature 3.1 is crossed by medium and large calibre peripapillary blood vessels (see Fig 2). When the thickness of the GCL is reduced, it is due to the presence of thinner blood vessels and our machine learning methods may be able to capture this relationship.

From our results, we glimpse a computer-aided solution for the diagnosis of MS where the practitioner is informed with the PPole grid image, the probability of the ML methods, the local SHAP and EBM explanations, and some recommendations learned from the waterfall plots in negative cases that would guide the process of decision making. In the case of witnessing signs of doubt as the ones shown in this work, complementary examination beyond the neuroophthalmological eye anatomy or maintaining a wait-and-see approach should be recommended. This way, OCT can be regarded as a quick test for the assessment of MS, with no side effects, non-invasive and therefore can be repeated on a serial basis in cases of unclear diagnosis or to see the progression of axonal damage. OCT is very economical and cost-effective for healthcare systems and can be performed by non-specialised personnel, making it a tool that in many cases can be used to reduce the number of magnetic resonance imaging requested, and thus alleviate the healthcare and economic burden on healthcare systems. Despite computer-aided solutions not able to reach a 100% accuracy, they would provide a first trustable assessment in the great majority of cases that would save additional expensive, time-consuming, and invasive tests.

Possible room for improvement lies in the fact that we have worked with PPole grid features as raw data. The decision was made due to the small sample size and the low resolution of the data. Given the two-dimensional nature of PPole features, we believe that the performance and

explainability may be improved with models that take into account the spatial representation of the data. As future work, we will explore convolutional neural networks, vision transformers (ViT), or swim transformers using the PPole grid features as images.

Although our study shows the methodology to follow to develop this solution, our results need to be confirmed by a larger sample of patients. Therefore, we are currently working on recruiting more populated and balanced datasets. We expect that the improved solution will provide models with a good balance between accuracy and explainability.

Supporting information

S1 File. Contains all the supporting tables and figures.
(PDF)

Acknowledgments

The authors would like to acknowledge the anonymous reviewers for their revision of the manuscript.

Author Contributions

Conceptualization: Monica Hernandez.

Data curation: Elisa Vilades, Beatriz Cordon, Elena Garcia-Martin.

Formal analysis: Monica Hernandez.

Funding acquisition: Monica Hernandez, Elena Garcia-Martin.

Investigation: Monica Hernandez, Ubaldo Ramon-Julvez, Elisa Vilades, Beatriz Cordon, Elvira Mayordomo, Elena Garcia-Martin.

Methodology: Monica Hernandez.

Project administration: Monica Hernandez, Elena Garcia-Martin.

Software: Monica Hernandez.

Supervision: Elvira Mayordomo, Elena Garcia-Martin.

Validation: Ubaldo Ramon-Julvez, Elisa Vilades, Beatriz Cordon.

Visualization: Monica Hernandez.

Writing – original draft: Monica Hernandez.

Writing – review & editing: Ubaldo Ramon-Julvez, Elvira Mayordomo.

References

1. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*. 2018; 17(2):162–173. [https://doi.org/10.1016/S1474-4422\(17\)30470-2](https://doi.org/10.1016/S1474-4422(17)30470-2) PMID: 29275977
2. Petzold A, de Boer JF, Schippling S, Vermersch P, Kardouk R, A AG, et al. Optical coherence tomography in multiple sclerosis: a systematic review and meta-analysis. *Lancet Neurol*. 2010; 9(9):921–932. [https://doi.org/10.1016/S1474-4422\(10\)70168-X](https://doi.org/10.1016/S1474-4422(10)70168-X) PMID: 20723847
3. Garcia-Martin E, Pueyo V, Martin J, Almarcegui C, Ara JR, Dolz I, et al. Progressive changes in the retinal nerve fiber layer in patients with multiple sclerosis. *FJ Eur J Ophthalmol*. 2010; 20(1):167–173. PMID: 19882541

4. Knier B, Berthele A, B D et al. Optical coherence tomography indicates disease activity prior to clinical onset of central nervous system demyelination. *Mult Scler*. 2015; 22(7):893–900. <https://doi.org/10.1177/1352458515604496> PMID: 26362905
5. Martinez-Lapiscina EH, Arnow S, W JA et al. IMSVISUAL consortium. Retinal thickness measured with optical coherence tomography and risk of disability worsening in multiple sclerosis: a cohort study. *Lancet Neurol*. 2016; 15(6):574–584. [https://doi.org/10.1016/S1474-4422\(16\)00068-5](https://doi.org/10.1016/S1474-4422(16)00068-5) PMID: 27011339
6. Britze J, Frederiksen JL. Optical coherence tomography in multiple sclerosis. *Eye (Lond)*. 2018; 32(5):884–888. <https://doi.org/10.1038/s41433-017-0010-2> PMID: 29391574
7. Sotirchos ES, Saidha S. OCT is an alternative to MRI for monitoring MS—YES. *Mult Scler*. 2018; 24(6):701–703. <https://doi.org/10.1177/1352458517753722> PMID: 29552938
8. London F, Zephir H, Drumez E. Optical coherence tomography: a window to the optic nerve in clinically isolated syndrome. *Brain*. 2019; 142(4):903–915. <https://doi.org/10.1093/brain/awz038> PMID: 30847470
9. Manogaran P, Samardzija M, Schad AN, Wicki CA, Walker-Egger C, M MR, et al. Retinal pathology in experimental optic neuritis is characterized by retrograde degeneration and gliosis. *Acta Neuropathol Commun*. 2019; 7(1):116. <https://doi.org/10.1186/s40478-019-0768-5> PMID: 31315675
10. Aly L, Havla J, L G et al. Inner retinal layer thinning in radiologically isolated syndrome predicts conversion to multiple sclerosis. *Eur J Neurol*. 2020; 27(11):2217–2224. <https://doi.org/10.1111/ene.14416> PMID: 32589804
11. Orbus M, Niewiadomska E, D P et al. The Usefulness of Optical Coherence Tomography in Disease Progression Monitoring in Younger Patients with Relapsing-Remitting Multiple Sclerosis: A Single-Centre Study. *J Clin Med*. 2022; 12(1):93. <https://doi.org/10.3390/jcm12010093>
12. Balk LJ, Cruz-Herranz A, Albrecht P, Arnow S, Gelfand JM, Tewarie P, et al. Timing of retinal neuronal and axonal loss in MS: a longitudinal OCT study. *J Neurol*. 2016; 263(7):1323–1331. <https://doi.org/10.1007/s00415-016-8127-y> PMID: 27142714
13. Toosy AT, Eshaghi A. Machine Learning Utility for Optical Coherence Tomography in Multiple Sclerosis. Is the Future Now? *Neurology*. 2019; 99(11). <https://doi.org/10.1212/WNL.000000000000200862>
14. Britze J, Pihl-Jensen G, Frederiksen JL. Retinal ganglion cell analysis in multiple sclerosis and optic neuritis: a systematic review and meta-analysis. *J Neurol*. 2017; 264(9):1837–1853. <https://doi.org/10.1007/s00415-017-8531-y> PMID: 28567539
15. Wicki CA, Hanson JVM, Schippling S. Optical coherence tomography as a means to characterize visual pathway involvement in multiple sclerosis. *Current Opinion in Neurology*. 2018; 31(5):662–668. <https://doi.org/10.1097/WCO.0000000000000604> PMID: 30074495
16. Birkeldh U, Manouchehrinia A, Hietala MA, Hillert J, Olsson T, et al. The Temporal Retinal Nerve Fiber Layer Thickness Is the Most Important Optical Coherence Tomography Estimate in Multiple Sclerosis. *Front Neurol*. 2017; 8. <https://doi.org/10.3389/fneur.2017.00675> PMID: 29326643
17. Zhang Y, Li N, Chen J, Wei H, Jiang SM, Chen XM. A new strategy to interpret OCT posterior pole asymmetry analysis for glaucoma diagnosis. *Int J Ophthalmol*. 2017; 10(12):1857–1863. <https://doi.org/10.18240/ijo.2017.12.11> PMID: 29259904
18. Casado A, Cervero, de Eguileta AL, Fernandez R, Fonseca S, Gonzalez JC, et al. Topographic correlation and asymmetry analysis of ganglion cell layer thinning and the retinal nerve fiber layer with localized visual field defects. *PLoS One*. 2019; 14(9):e0222347. <https://doi.org/10.1371/journal.pone.0222347> PMID: 31509597
19. Martucci A, Landi D, Cesareo M, DiCarlo E, DiMauro G, et al. Complex Rearrangement of the Entire Retinal Posterior Pole in Patients with Relapsing Remitting Multiple Sclerosis. *J Clin Med*. 2021; 10(20):4693. <https://doi.org/10.3390/jcm10204693> PMID: 34682817
20. Vilades E, Cordon B, Perez-Velilla J, Orduna E, Satue M, Polo V, et al. Evaluation of multiple sclerosis severity using a new OCT tool. *PLOS ONE*. 2023; 18(7):e0288581. <https://doi.org/10.1371/journal.pone.0288581> PMID: 37440532
21. Aslam N, Khan IU, Bashamakh A, Alghool FA, Abounour M, Alsawayan NM, et al. Multiple Sclerosis Diagnosis Using Machine Learning and Deep Learning: Challenges and Opportunities. *Sensors (Basel)*. 2022; 22:7856. <https://doi.org/10.3390/s22207856> PMID: 36298206
22. Garcia-Martin E, Pablo LE, Herrero R, Ara JR, Martin J, Larrosa JM, et al. Neural networks to identify multiple sclerosis with optical coherence tomography. *Acta Ophthalmol*. 2013; 91(8):628–34. <https://doi.org/10.1111/aos.12156> PMID: 23647619
23. del Palomar AP, Cegonino J, Montolio A, Orduna E, Vilades E, Sebastian B. Swept source optical coherence tomography to early detect multiple sclerosis disease. The use of machine learning techniques. *PLoS ONE*. 2019; 14(5):e0216410. <https://doi.org/10.1371/journal.pone.0216410>

24. Cavaliere C, Vilades E, Alonso-Rodriguez MC, Rodrigo MJ, Pablo LE, Miguel JM, et al. Computer-Aided Diagnosis of Multiple Sclerosis Using a Support Vector Machine and Optical Coherence Tomography Features. *Sensors (Basel)*. 2019; 19(23):5323. <https://doi.org/10.3390/s19235323> PMID: 31816925
25. Montolio A, Martin-Gallego A, Cegonino J, Orduna E, Vilades E, Garcia-Martin E, et al. Machine learning in diagnosis and disability prediction of multiple sclerosis using optical coherence tomography. *Comput Biol Med*. 2021; 133:104416. <https://doi.org/10.1016/j.combiomed.2021.104416> PMID: 33946022
26. Garcia-Martin E, Ortiz M, Boquete L, Sanchez-Morla EM, Barea R, Cavaliere C, et al. Early diagnosis of multiple sclerosis by OCT analysis using Cohen's d method and a neural network as classifier. *Comput Biol Med*. 2021; 129:104165. <https://doi.org/10.1016/j.combiomed.2020.104165> PMID: 33302162
27. Montolio A, Cegonino J, Garcia-Martin E, Palomar APD. Comparison of Machine Learning Methods Using Spectralis OCT for Diagnosis and Disability Progression Prognosis in Multiple Sclerosis. *Ann Biomed Eng*. 2022; 50(5):507–528. <https://doi.org/10.1007/s10439-022-02930-3> PMID: 35220529
28. Lopez-Dorado A, Ortiz M, Satue M, Rodrigo MJ, Barea R, Sanchez-Morla EM, et al. Early Diagnosis of Multiple Sclerosis Using Swept-Source Optical Coherence Tomography and Convolutional Neural Networks Trained with Data Augmentation. *Sensors (Basel)*. 2022; 22:167. <https://doi.org/10.3390/s22010167>
29. Marinescu RV, Oxtoby NP, Young AL, Bron EE, AW AWT, et al. The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge: Results after 1 Year Follow-up. *ArXiv*. 2020;.
30. Hernandez M, Ramon-Julvez U, Ferraz F, ADNI. Explainable AI toward understanding the performance of the top three TADPOLE Challenge methods in the forecast of Alzheimer's disease diagnosis. *PLoS ONE*. 2022. <https://doi.org/10.1371/journal.pone.0264695> PMID: 35522653
31. Kenney RC, Liu M, H L et al. The Role of Optical Coherence Tomography Criteria and Machine Learning in Multiple Sclerosis and Optic Neuritis Diagnosis. *Neurology*. 2022; 99(11):e1100–e1112. <https://doi.org/10.1212/WNL.000000000000200883> PMID: 35764402
32. ElSappagh S, Alonso JM, Riazul SM, Sultan AM, Kwak KS. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Nature Scientific Reports*. 2021; 11:2660. <https://doi.org/10.1038/s41598-021-82098-3> PMID: 33514817
33. Merrick L, Taly A. The explanation game: explaining machine learning models using shapley values. *Machine Learning and Knowledge Extraction*. 2020; p. 17–38. https://doi.org/10.1007/978-3-030-57321-8_2
34. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? *ArXiv*. 2022;.
35. Fabrizio C, Termine A, Caltagirone C, Sancesario G. Artificial Intelligence for Alzheimer's Disease: Promise or Challenge? *Diagnostics (Basel)*. 2021; 11(8):1473. <https://doi.org/10.3390/diagnostics11081473> PMID: 34441407
36. Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*. 2017; 6:52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
37. Molnar C. A guide for making black box models explainable. <https://christophm.github.io/interpretable-ml-book>. 2021;.
38. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Proc of the 31st Conference on Neural Information Processing Systems*. 2017;.
39. Qiu K, Schiefer J, Jansonius N. Influence of the Retinal Blood Vessel Topography on the Variability of the Retinal Nerve Fiber Bundle Trajectories in the Human Retina. *Invest Ophthalmol Vis Sci*. 2015; 56(11):6320–6325. <https://doi.org/10.1167/iovs.15-17450> PMID: 26436884