Discrimination of Multiple Sclerosis using multicenter OCT images

Zahra Khodabandeh , Hossein Rabbani , Fereshteh Ashtari , Hanna G. Zimmermann , Seyedamirhosein Motamedi , Alexander U. Brandt , Friedemann Paul , Rahele Kafieh

PII: S2211-0348(23)00348-6
DOI: https://doi.org/10.1016/j.msard.2023.104846
Reference: MSARD 104846

To appear in: *Multiple Sclerosis and Related Disorders*

Received date: 22 May 2023
Accepted date: 19 June 2023

Please cite this article as: Zahra Khodabandeh , Hossein Rabbani , Fereshteh Ashtari , Hanna G. Zimmermann , Seyedamirhosein Motamedi , Alexander U. Brandt , Friedemann Paul , Rahele Kafieh , Discrimination of Multiple Sclerosis using multicenter OCT images, *Multiple Sclerosis and Related Disorders* (2023), doi: https://doi.org/10.1016/j.msard.2023.104846

Discrimination of Multiple Sclerosis using multicenter OCT images

Zahra Khodabandeh[a] , Hossein Rabbani[a] , Fereshteh Ashtari[b] , Hanna G.Zimmermann[c,d] , Seyedamirhosein Motamedi[c,d] , Alexander U.Brandt[c,d,e] , Friedemann Paul[c,d,f] , Rahele Kafieh[a,d,g,*]

a School of Advanced Technologies in Medicine, Medical Image and Signal Processing Research Center, Isfahan University of Medical Sciences, Isfahan, Iran
b Isfahan Neurosciences Research Center, Isfahan University of Medical Sciences, Isfahan, Iran
c Experimental and Clinical Research Center, Max Delbrück Center for Molecular Medicine and Charité- Universitätsmedizin Berlin, Berlin, Germany
d NeuroCure Clinical Research Center- Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany
e Department of Neurology, University of California, Irvine, CA, USA
f Department of Neurology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany
g Department of Engineering, Durham University, Durham, UK

* Corresponding Author: Department of Engineering | Christopherson Building | Durham University | UK
*Email address:*Raheleh.kafieh@durham.ac.uk (R.Kafieh)

ABSTRACT

Background: Multiple sclerosis (MS) is one of the most prevalent chronic inflammatory diseases caused by demyelination and axonal damage in the central nervous system. Structural retinal imaging via optical coherence tomography (OCT) shows promise as a noninvasive biomarker for monitoring of MS. There are successful reports regarding the application of Artificial Intelligence (AI) in the analysis of cross-sectional OCTs in ophthalmologic diseases. However, the alteration of sub-retinal thicknesses in MS is noticeably subtle compared to other ophthalmologic diseases. Therefore, raw cross-sectional OCTs are replaced with multilayer segmented OCTs for discrimination of MS and healthy controls (HCs).

Methods: To conform to the principles of trustworthy AI, interpretability is provided by visualizing the regional layer contribution to classification performance with the proposed occlusion sensitivity approach. The robustness of the classification is also guaranteed by showing the effectiveness of the algorithm while being tested on the new independent dataset. The most discriminative features from different topologies of the multilayer segmented OCTs are selected by the dimension reduction method. Support vector machine (SVM), random forest (RF), and artificial neural network (ANN) are used for classification. Patient-wise cross-validation (CV) is utilized to evaluate the performance of the algorithm, where the training and test folds contain records from different subjects.

Results: The most discriminative topology is determined to square with a size of 40 pixels and the most influential sub-retinal layers are the ganglion cell and inner plexiform layer (GCIPL) and inner nuclear layer (INL). Linear SVM resulted in 88% Accuracy (with standard deviation (std) = 0.49 in 10 times of execution to indicate the repeatability), 78% precision (std=1.48), and 63% recall (std=1.35) in the discrimination of MS and HCs using macular multilayer segmented OCTs.

Conclusion: The proposed classification algorithm is expected to help neurologists in the early diagnosis of MS. This paper distinguishes itself from other studies by employing two distinct datasets, which enhances the robustness of its findings in comparison with previous studies with

lack of external validation. This study aims to circumvent the utilization of deep learning methods due to the limited quantity of the available data and convincingly demonstrates that favorable outcomes can be achieved without relying on deep learning techniques.

Keywords: multiple sclerosis, optical coherence tomography, interpretable Artificial Intelligence, generalizable, patient-wise cross-validation

1. Introduction

Multiple sclerosis (MS) is a chronic inflammatory and neurodegenerative disease of the central nervous system (CNS) that causes progressive neurological disability over time. MS is determined by demyelination and neuro-axonal damage that results in tissue loss and progressive neurologic deficits [1]. While the most established method to monitor the degree of CNS damage in MS is magnetic resonance imaging (MRI) [2], MS leads to widespread changes in the retina and optic nerve, which may be assessed with optical coherence tomography (OCT) to obtain useful disease biomarkers [3,4]. OCT-derived imaging markers like peripapillary retinal nerve fiber layer thickness (pRNFL) and composite thickness of macular ganglion cell layer (GCL) and Inner plexiform layer (IPL) (named GCIPL) have been proposed as promising biomarkers for neurodegeneration [5,6]. Inflammatory disease activity may also lead to changes in inner nuclear layer thickness (INL) [6]. Layer thinning can be measured by aligning and subtracting retinal layer thicknesses from a normal healthy population [7,8].

Artificial intelligence (AI) is a promising area of health innovation[9,10]. Its application in ophthalmology is also evident in analysis of different ocular images[11,12], with purpose of segmenting the retinal boundaries[13], discriminating different diseases [14,15] or interpretation of neurological diseases using quality control (QC) criteria[16]. Cross-sectional OCTs are successfully employed in AI for detection of ophthalmologic diseases. However, the alteration of sub-retinal thicknesses in MS are noticeably subtle to be diagnosed with raw cross-sectional OCTs. The other limitation of AI in medical applications is its black box nature which contradicts with interpretability in trustworthy AI. Furthermore, limiting the training and testing datasets to single clinical centers leads to less generalizable algorithms. Finally, cross-validation (CV) in most of medical AI works is performed instance-wise, which overestimates algorithm prediction accuracy[17].

Here we propose an AI method that aims to capture ultra-fine changes in retinal sub-layers by using multilayer segmented OCT. The method is interpretable using a novel proposed approach, which means regional layer contribution to classification performance is visualized using the proposed occlusion sensitivity approach. We test the trained model on an independent second dataset to show robustness. The patient-wise CV is used where the training and test folds contain eyes from different subjects; therefore, in testing stage, the performance is measured on a new subject whose data from the fellow eye has not been used for training.

By considering the mentioned concepts, feature selection from different topologies of multilayer segmented OCTs is done. We compare the performances of support vector machine (SVM), random forest (RF), and artificial neural network (ANN), and identify the most discriminative topology and the most influential sub-retinal layers. This study aims to obtain a classification algorithm using AI method based on changes in sub-retinal layers of OCT in the neurodegeneration process to help neurologists in the early diagnosis of MS disease.

2. Materials and Methods

2.1. Structure of the datasets

Generalizable algorithms are of interest in medical AI, but when both training and testing datasets come from single clinical centers, attaining this goal cannot be evaluated. We therefore concentrate on two independent datasets with different devices in different countries to be used as separate training and testing datasets in measuring the robustness of the algorithm.

2.1.1. Charité dataset

The first OCT dataset is from the NeuroCure Clinical Research Center (NCRC) at Charité – Universitätsmedizin Berlin, Berlin, Germany. It consists of 422 HC and 106 MS OCTs from two multimodal register studies to evaluate quantitative measurements of neuro-axonal damage in MS. The OCT data in this dataset includes 40 to 51 B-scans with a size of 496 × (479 to 555) pixels for each B-scan. All OCT measurements were carried out with an Spectralis SD-OCT and Heidelberg Eye Explorer (HEYEX) version 5.7.5.0 by eight individual operators and an automatic real-time function for image averaging and an activated eye tracker in a dimly lit room. All scans were quality controlled according to the OSCAR-IB criteria [18,19]. Retrospective inclusion criteria for the study were participants in a healthy condition, aged between 18 and 70 years, Caucasian ethnicity, and high-quality macular OCT scans. Collecting this dataset was approved by the ethics committee of Charité - Universitätsmedizin Berlin and was conducted according to the Declaration of Helsinki in the applicable version. The macular OCT scans were produced from the device and stored in HEYEX vol file format and then a segmentation approach was carried out using a segmentation pipeline. All segmentation results were quality controlled and manually corrected [20]. Demographic features of the subjects in this dataset are summarized in Table 1.

2.1.2. Isfahan dataset

The second OCT dataset is from the Kashani Comprehensive MS center in Isfahan, Iran, between April 2017 and March 2019 [21]. The images were obtained using Spectralis SD-OCT and Heidelberg HEYEX version 5.1 by one trained technician with an automatic real-time (ART) of 9 frames function for image averaging. All scans were checked for sufficient quality using OSCAR-IB criteria[18]. The dataset consists of 45 HC and 45 MS eyes. The automated segmentation was carried out using a graph-based method [22,23]. All segmentation results were quality controlled and manually corrected in case of errors by an experienced grader using custom-developed software [21,24]. However, because of using high-quality OCT images, the segmentation errors are not significant and in average, don't have significant effect on classification results. Demographic and clinical features of the subjects in this dataset are summarized in Table 2.

2.1.3. Standardized quality control criteria

OSCAR-IB criteria is a standard for quality assessment of OCT images based on manual evaluation by expert grader. Several indicators are considered as quality indicator, forming the abbreviation OSCAR-IB: (O) obvious problems, (S) poor signal strength, (C) centration of scan, (A) algorithm failure, (R) unrelated retinal pathology, (I) illumination and (B) beam placement [18]. This criteria has been validated for MS[19].

Table 1. Demographic and clinical characteristics in participants of Charité dataset

	HC	MS
Current age, y, mean ±SD	36.5±12.3	41.42±10.11
Sex, F, n (%)	280 (66%)	70(66%)

Table 2. Demographic and clinical characteristics in participants of Iran dataset

	HC	MS
Current age, y,mean ±SD	26.3±3.06	34.5±8.03
Sex, F, n (%)	12 (66%)	30 (85%)
Disease duration, y, mean ±SD	NA	7.67±1.37

AQP4-Ab: aquaporin 4 antibody, y: year, SD: standard deviation

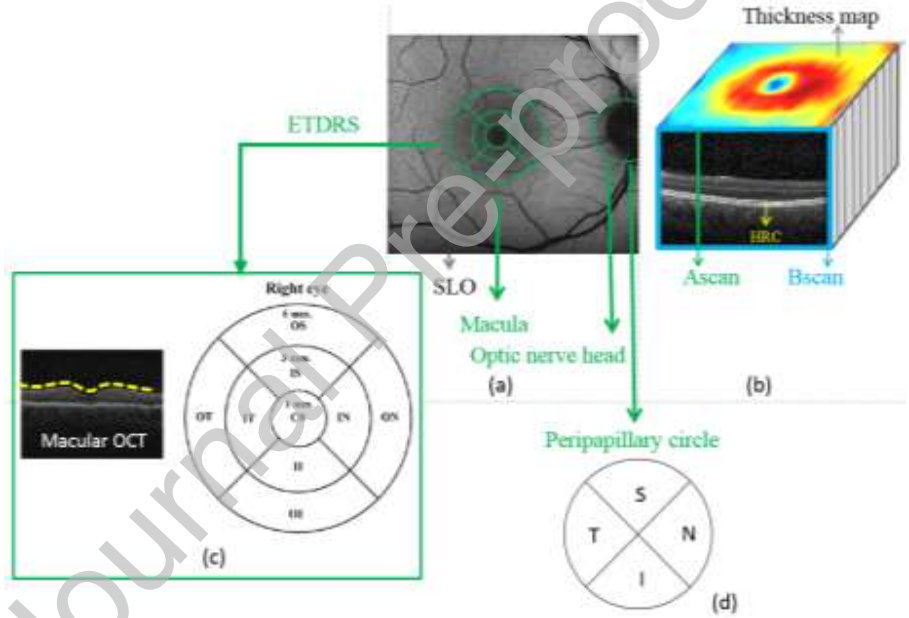

Figure 1. Retinal parameters acquired by OCT. (a) location of sectors and ring scan on SLO image. (b) A-scan, B-scan, and thickness map of OCT data. (c) quadrants in ETDRS: central fovea (CF), inner superior (IS), inner nasal (IN), inner inferior (II), inner temporal (IT), outer superior (OS), outer superior (OS), outer nasal (ON), outer inferior (OI) and outer temporal (OT). (d) quadrants in the peripapillary circle: superior (S), inferior (I), temporal (T), and nasal(N) [21].

2.2. Preprocessing and feature extraction

Intra-retinal thickness changes in MS are often noticeably subtle compared to primary eye disorders [25]. Multilayer segmented OCTs are therefore used and the distances between pairs of retinal layers, called retinal thickness maps, are calculated (Figure 1). The area covered by B-scans around the macula may be oriented. As one possible hypothesis, the effect of compensating the orientation angle is studied in this work. For this purpose, the thickness maps are rotated to have a unique format as input to the next processing steps. The angle between a horizontal line

through the disc center and the disc–foveal line ($angle\_Fovea\_ONH\_SLine$) (Figure 2(a)); and the relative direction of each B-scan to a horizontal line through the disc center ($slope\_Bscan$) (Figure 2(b)) are calculated. The left eyes are also flipped. The value of correcting rotation ($rotation\ angle -$ Figure 2(c)) is calculated by:

$$\begin{aligned} rotation\ angle &= angle\_Fovea\_ONH\_SLine && if\ \ slope\_Bscans\ \ = 0 \\ rotation\ angle &= abs(-90\ -\ angle\_Fovea\_ONH\_SLine) && if\ \ 88 < slope\_Bscan <= 90 \\ rotation\ angle &= (angle\_Fovea\_ONH\_SLine\_deg) + (slope\_Bscan) && otherwise \end{aligned} \quad (1)$$

The rotated thickness maps are cropped to a unique size of 450×450 pixels (Figure 2(d)). The thickness maps (with/without rotation) from different retinal layers including mRNFL, GCIP, sum of GCIP and INL layers (GCIP+INL), parallel use of GCIP and INL (GCIP/INL), parallel use of mRNFL, GCIP, INL, ONL, and the total macular thickness are considered as input to the classification stage.

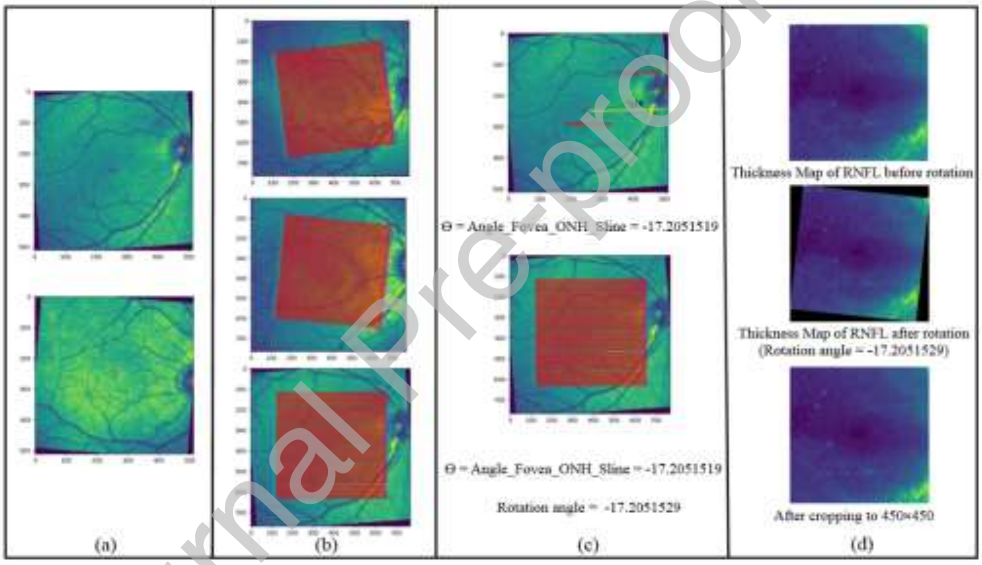

Figure 2. (a) SLO image in clockwise and counterclockwise rotations. (b) B-scans in different directions. (c) Example of finding the rotation angle), (d) Process of rotating a thickness map.

To extract different topological information from each thickness map, the regions of interest typically follow those defined by the Early Treatment Diabetic Retinopathy Study (ETDRS) [21]. ETDRS concentric circles are calculated with diameters of 1 mm, 3mm, and 6 mm around the fovea, divided into quadrants and forming nine macular areas demonstrated in Figure 1. As alternative topologies, we also used different resolutions of the thickness maps in squares ranging between 20×20, 30×30, and 40×40 pixels. A combination of retinal layers, classifiers, evaluation, and dimension reduction approaches are used and summarized in Figure 3.

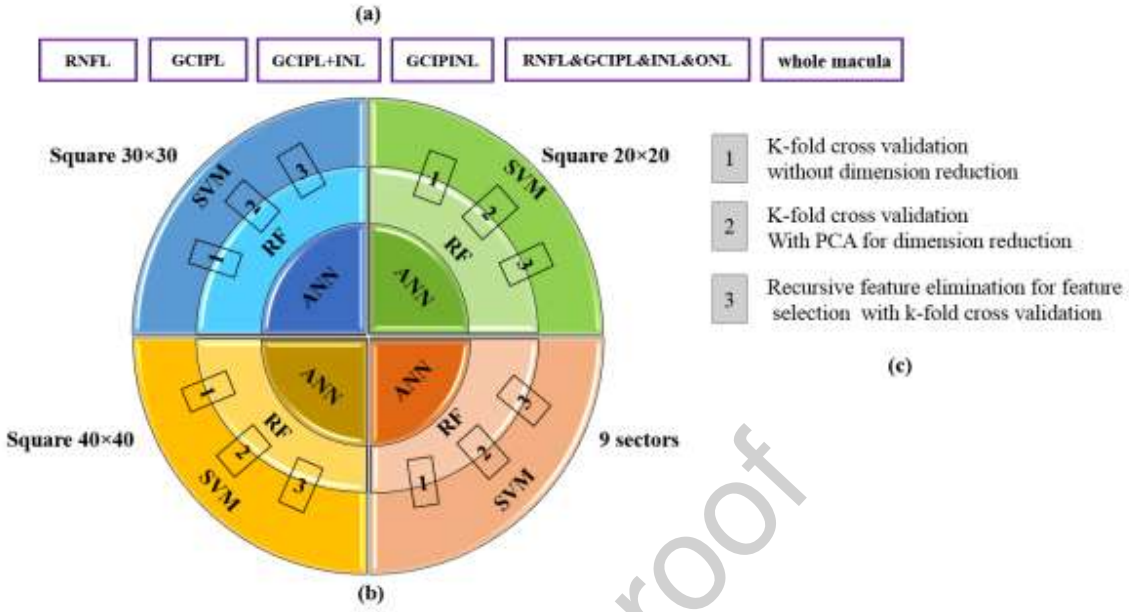

Figure 3. (a) Retinal layers investigated by the proposed method. (b) Classification model: SVM, RF, and ANN on four groups of extracted features: square 20×20, square 30×30, square 40×40, and 9 ETDRS sectors. (c) Different evaluation and dimension reduction methods in the study.

2.3. Dimension reduction

To decrease the model complexity and avoid overfitting, we used principal component analysis (PCA) [26] that deduces information from the feature set to make a new feature subspace. Recursive feature elimination (RFE) [27] is also used to select subsets of the main features.

2.4. Machine learning algorithms and evaluation method

Machine learning algorithms are used to explain the patterns in the data and to extract information from it. The algorithms in this study are SVM, RF, and ANN.

2.4.1. *Support Vector Machine(SVM)*

Support Vector Machine is driven by a linear function $w^T x + b$ that predicts the classes according to the sign of this function [28]. In two-class problems, SVM looks for a hyper-plane to divide two different classes with a maximum margin. When the original data is not separable linearly, a nonlinear transformation with a kernel function can be used to transfer the feature space to a higher dimension space with good separability [29]. Kernel functions used in this study are linear, polynomial, radial basis, and sigmoid.

2.4.2. *Random Forest (RF)*

RF includes many decision trees, and each decision tree prepares a classification for input data. RF gathers the trees and chooses the most voted prediction as the result. The input of each tree is the sampled data from the whole dataset. Moreover, a subset of features is randomly chosen from the optimal features to grow the tree at each node[30,31]. We used the grid search method [32] to optimize parameters of a random forest like the number of trees, criterion (the function to measure the quality of a split) including Gini and entropy, and maximum features

(the number of features to be considered when looking for the best split) such as sqrt, log2, and auto modes.

2.4.3. Artificial Neural Network (ANN)

An artificial neural network (ANN) includes an input layer of neurons, one or two hidden layers, and an output layer that is the universal function approximator of the interconnection of human neurons [33]. To avoid overfitting due to complex networks or getting low accuracy due to simple networks with few layers, we used the grid search method to find the model with the best performance. In this study, we found a good performance with a sequential model with four dense layers. The neurons in each layer are 100, 80, 20, and 1, respectively, by grid search method. Rectified linear activation (ReLu) is used as the activation function in the first three layers, and the last layer uses the sigmoid activation function.

2.5. Evaluation methods

Ten-fold patient-wise CV is used with no combination of subjects' eyes in the training and test folds. This approach reduces the overestimation of prediction accuracy [17] in instance-wise CV with leakage of information between training and testing phases. Classification performance is evaluated according to the confusion matrix and the values of accuracy, precision, recall, and f1-score are reported. The reproducibility of the results is checked by removing the constant random state in the k-fold CV and executing the model ten times and calculating the standard deviation of the results.

2.6. Interpretability

One of the main limitations of AI in medical applications is black box nature contradicts with interpretability of trustworthy AI. Conventional machine learning methods are mostly designed to work with vectors as input. Therefore, the images are changed into vectors, and the original image structure is ignored. On the other hand, recent methods like Convolutional Neural Networks are introduced as powerful competitors, preserving the image structure and providing image-based interpretability, expected to be humanly interpretable [34].

In this study, we propose a novel approach to add interpretability to current machine learning approaches. We used occlusion sensitivity [35] and modified it to fit the vector-like inputs. After training the model, we created a black mask with the size of 10×10 pixels and moved it to the test set with a single step to sweep the whole image. The locations of the pixels covered by the mask are transferred to vector-shaped positions (Figure 4). The masked vector is sent as input to the model and the accuracy is calculated. It is expected that the occlusion of regions with important discriminative information leads to lower accuracy. The interpretability is shown by regenerating the occlusion with the original image size, with the value of accuracy in the location of each pixel (called the heat map). An interpretability heatmap indicates how important each location is concerning the class and visualizes the regional contribution to classification.

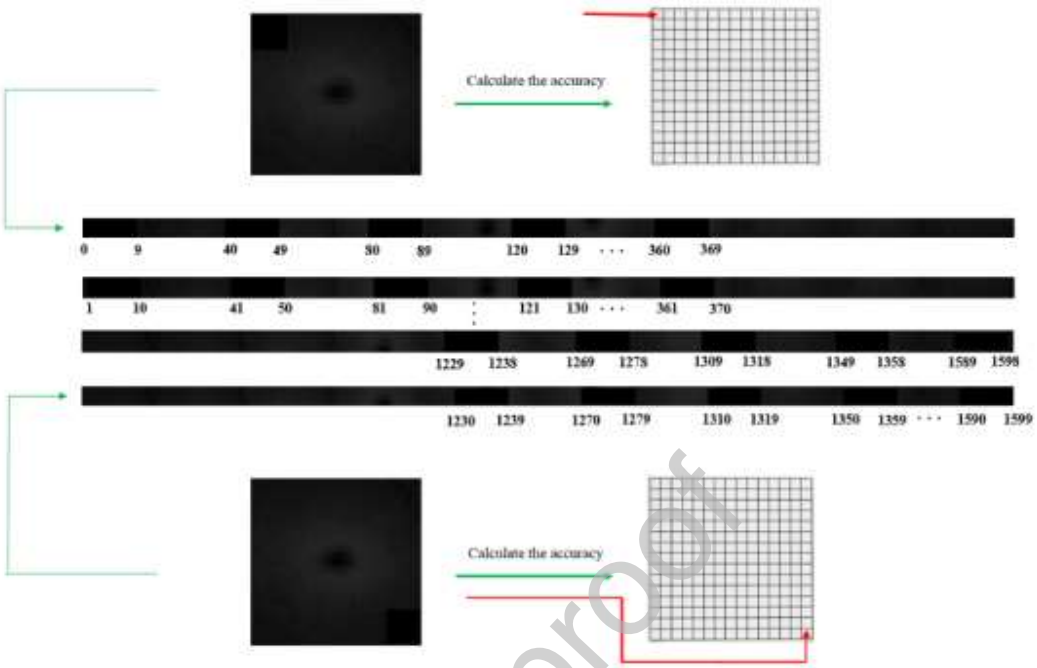

Figure 4. The proposed process for creating a black mask, moving it to the test set, and transferring the locations to vector-shaped positions

3. Results

For classification purposes, different topologies of the thickness maps around the macula in squares with resolutions of 20×20, 30×30, 40×40 pixels, and mean thicknesses in 9 sectors of ETDRS are considered. The effect of compensating rotation on thickness maps is examined. The classification models are first trained and tested on Charité (first) dataset. To show the generalizability of the method, the trained classifier with the best performance (on the first dataset) is tested on the Isfahan (second) dataset. The proposed occlusion sensitivity is also shown for interpretability.

Different combinations of features, two different dimension-reduction methods and different machine learning methods are used with 10-fold patient-wise CV on Charité (first) dataset. The comparison of metrics on each parameter is presented by keeping the other parameters fixed on the best-performing set. Table 3 compares the effectiveness of different retinal layers and the effect of rotation in the correct classification. In this comparison, other parameters are fixed on the best performing set including square size of 40×40 pixels, linear SVM as classifier, 10-fold cross-validation, and PCA for dimension reduction. The results are compared in both situations (with and without rotation in the preprocessing step). As can be seen, GCIP&INL (GCIP/INL) without rotation is the most informative combination of the retinal layers. The selection of the best topology is performed based on Table 4. When the best set of parameters are fixed (GCIP/INL without rotation as input feature, Linear SVM as a classifier with 10-fold cross-validation, and PCA for dimension reduction), the best topology is related to the square size of 40×40 pixels. Accuracy of different classification methods with different topologies is presented in Table 5. As can be seen, linear SVM has better results than RF and ANN in terms of accuracy. For SVM method in classification, Table 6 compares the performance of the different kernels. Moreover, the dimension reduction methods are compared in Table 7. Ten-fold cross validation

with PCA for dimension reduction has the best results when other parameters are fixed on the best-performing set of information (GCIP/INL with square size of 40×40 without rotation as input feature, Linear SVM as a classifier).

Table 3. Comparison of input features in the classification of MS and HC. The other parameters are fixed on the best-performing set of information (square size of 40×40, Linear SVM as classifier with 10-fold cross-validation, PCA for dimension reduction). The effect of rotation is shown in the upper and lower part of the table, respectively.

Square 40×40 – (10-fold with PCA) - without rotation				
	Accuracy	**Precision**	**Recall**	**F1-score**
mRNFL	79%	47%	41%	43%
GCIP	87%	72%	60%	64%
GCIP&INL(GCIP/INL)	**88%**	**78%**	**63%**	**68%**
GCIP+INL	82%	56%	51%	52%
mRNFL&GCIP&INL&ONL	84%	64%	58%	59%
Whole macular volume	80%	52%	45%	47%
GCIP & whole macular volume	80%	51%	49%	49%
GCIP & INL & macular volume	81%	54%	52%	52%
Square 40×40 - (10-fold with PCA) - with rotation				
	Accuracy	**Precision**	**Recall**	**F1-score**
mRNFL	74%	33%	33%	33%
GCIP	82%	56%	56%	55%
GCIP&INL(GCIP/INL)	82%	56%	51%	52%
GCIP+INL	83%	63%	50%	54%
mRNFL&GCIP&INL&ONL	82%	58%	47%	50%
Whole macular volume	80%	53%	44%	45%
GCIP & whole macular volume	80%	51%	47%	48%
GCIP & INL & macular volume	79%	47%	51%	48%

Table 4. Comparison of square size in the classification of MS and HC. The other parameters are fixed on the best-performing set of information (GCIP/INL without rotation as input feature, Linear SVM as a classifier with 10-fold cross-validation, and PCA for dimension reduction)

	Accuracy (SVM-linear)	**Precision (SVM-linear)**	**Recall (SVM-linear)**	**F1-score (SVM-linear)**
Square 20×20	84%	64%	57%	57%
Square 30×30	86%	74%	57%	61%
Square 40×40	**88%**	**78%**	**63%**	**68%**
9 sectors	84%	75%	33%	44%

Table 5. Comparison of machine learning methods in classification of MS and HC. The other parameters are fixed on the best-performing set of information (GCIP/INL for a square size of 40×40 without rotation as an input feature, and PCA for dimension reduction)

	Accuracy (SVM-liner)	**Accuracy (RF)**	**Accuracy (ANN)**
Square 20×20	**84%**	84%	82%
Square 30×30	**86%**	85%	84%
Square 40×40	**88%**	85%	85%
9 sectors	84%	85%	82%

Table 6. Comparison of kernels for SVM method in the classification of MS and HC. The other parameters are fixed on the best-performing set of information (GCIP/INL for a square size of 40×40 without rotation as an input feature, SVM as a classifier with 10-fold cross-validation, and PCA for dimension reduction

	Accuracy	Precision	Recall	F1-score
Linear	**88%**	78%	**63%**	**68%**
Polynomial	85%	87%	28%	41%
Radial basis (RBF)	86%	91%	33%	47%
Sigmoid	83%	67%	38%	47%

Table 7. Comparison of different dimension reduction methods in the classification of MS and HC. The other parameters are fixed on the best-performing set of information (GCIP/INL for a square size of 40×40 without rotation as input feature, Linear SVM as a classifier with 10-fold cross-validation)

	Accuracy	Precision	Recall	F1-score
10-fold cross validation without dimension reduction	88%	79%	59%	65%
10-fold cross validation with PCA for dimension reduction	**88%**	78%	**63%**	**68%**
10-fold cross validation with RFE for dimension reduction	86%	76%	50%	57%

To explore the application of RFE in cross-validation, the importance of each feature is obtained through a coefficient attribute and features with a correlation coefficient above a threshold of 0.8 are removed. The diagram of accuracy against the number of features is shown in Figure 5.

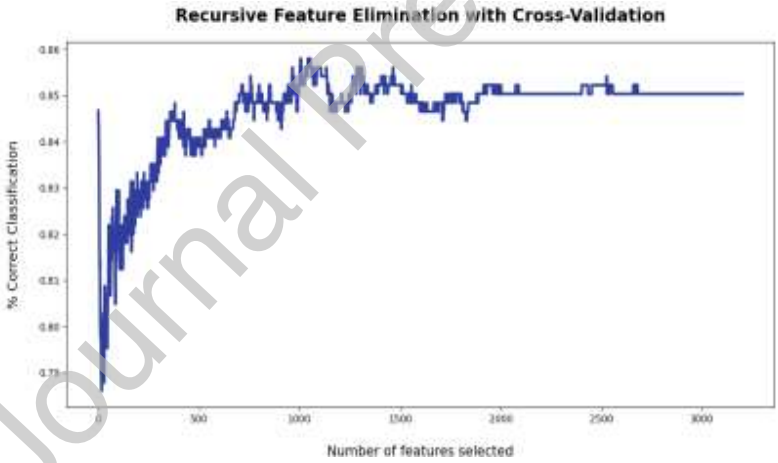

Figure 5. RFE with cross-validation diagram that shows accuracy with the number of features.

To show the generalizability of the method, the trained classifier with the best performance on the Charité dataset (GCIP/INL with square size of 40×40 without rotation as input feature, Linear SVM as a classifier, and PCA for dimension reduction) is tested on Isfahan dataset and the performance is shown in Table 8.

Table 8. Classification of MS and HC on Isfahan (second) dataset using a best-performing classifier trained on Charité (first) dataset.

	Accuracy	Precision	Recall	F1-score
SVM(linear) with PCA for dimension reduction and	88%	89%	88%	84%

3.1. Visual Interpretability

The proposed method for visual interpretability is demonstrated by plotting the heatmap of the occlusion sensitivity. The results in the previous section showed that GCIP/INL (parallel use of GCIP and INL) are the most effective layers in distinguishing MS patients from HCs; therefore, these two layers of the best-performing set of hyperparameters are used for analyzing the interpretability in Figure 6.

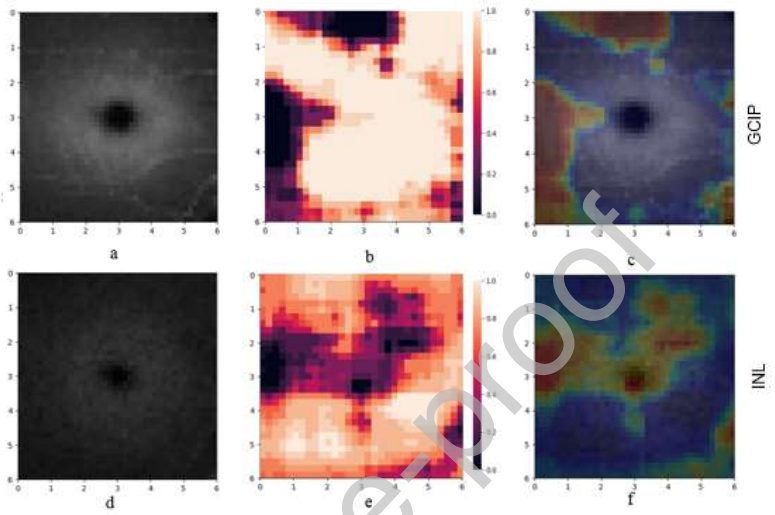

Figure 6. Visual interpretability on thickness maps of GCIP and INL. (a) Thickness map of GCIP in one sample from MS dataset (x and y axis in mm), (b) heatmap of occlusion sensitivity in the classification of MS and HC, (c) overlap of the heatmap and the GCIP layer, (d) Thickness map of INL in one sample from MS dataset, (e) heatmap of occlusion sensitivity in the classification of MS and HC, (f) overlap of the heatmap and the INL layer. As can be seen, the temporal region in the thickness map of GCIP have more important information in classification of MS disease

4. Discussion

The model with the highest accuracy based on our optimization approach is able to discriminate MS and HCs with an accuracy of 88% and F1-score of 68% with standard deviation of 0.48 and 0.94 in 10 times of execution to indicate repeatability, using GCIPL and INL information. Indistinct changes in retinal sublayers are captured with multilayer segmented OCT. An interpretable result is acquired to indicate the regional layer contribution to classification performance using occlusion sensitivity. The generalizability is evaluated by training on a first dataset and then testing on a second independent dataset with a new device from another country. The performance is similar (accuracy of 88% and F1-score of 84%) when testing on data, which proves the generalization ability of the proposed method (more detail is presented in Table 8). To avoid overestimation, patient-wise CV is used to a separate set of patients in the training and test datasets. Different combinations of the retinal layers as input features, two different dimension reduction methods and different machine learning methods are compared.

Simultaneous data from GCIP and INL (GCIP/INL) were found to be the most informative combination of the retinal layers (Table 3). This finding is in accordance with clinical studies [25] [6]. The rotation of the thickness maps did not improve the performance. One possible

reason for this finding is using the traditional machine learning methods which change the image format to vectorized data. This vectorization process may be responsible for reducing the effect of the rotation.

The best topology is a square size of 40×40 (Table 4). It seems that this resolution is relevant to the number of B-scans in each OCT data (40 to 51 B-scans). Namely, 40×40 square extracts the most possible information without suppressing the data between the B-scans.

The interpretability heatmap of classification with this novel proposed algorithm is a new strategy in conventional machine learning methods and makes them comparable to their main competitors like CNN. As demonstrated in Figure 6, the temporal region in the thickness map of GCIP is found to have more effect on the classification of MS disease. It is related to occurring the most degree of loss in the temporal preponderance of RNFL in MS eyes[36].

Among the machine learning methods, SVM achieved the best results (Table 5) with linear kernel (Table 6). This finding seems reasonable since linear kernels are proven to be more effective when the number of features is large in comparison to the training samples [37]. Dimension reduction improved the results and PCA method was found more appropriate (Table 6). The selected model with the highest accuracy based on our optimization approach discriminates MS and HCs with an accuracy of 88% and F1-score of 68%, using GCIPL and INL information. Table 9 shows a summary of previous similar methods in comparison with the proposed algorithm. Direct comparison of the results with these works is not possible since the codes and datasets are not released in any of those works. Furthermore, none of the previous works considered the patient-wise CV and accordingly higher performance is reported with leakage of information between train and test data in instance-wise approaches. It should also be noted that in this work, the state of being affected by optic neuritis (ON) was not considered and accordingly, MS patients with/ without ON are combined for classification. Therefore, compared to work considering MS with ON, a lower performance is convincing since the eyes without ON show less thinning and are less discriminable from the HCs [6,25,38]. Finally, some previous works include the pRNFL data as the input of the classification and a correspondingly higher performance is achieved compared to the limited focus of macular region.

Table 9. Summary of previous similar methods

Previous works	Number of datasets	Input retinal layers	Being affected by ON	patient-wise/ instance-wise cross validation	Performance metrics	The most discriminant retinal layer	Classification method
Garcia-Martin et al. [39] 2013	106 MS, 115 HC	Peripapillary area	29% (31) with ON, 71% (75) without ON	instance-wise	AUC=0.945	pRNFL	ANN
Garcia-Martin et al. [40] 2015	112 MS, 105 HC	Peripapillary area	36.6%(41) with ON, 63.4% (71) without ON	instance-wise	Recall=89.3% Specificity=87.6% Precision=88.5%	pRNFL	ANN
Palomar et al. [41] 2019	80 MS, 180 HC	Peripapillary, macular and extended (between macula and papilla) areas	with ON	instance-wise	Decision tree in macular area: Accuracy=97.24% AUC=0.959 In extended area: Accuracy=95.3%	pRNFL	Decision tree, ANN, SVM

					AUC=0.998		
Cavaliere et al. [29] 2019	48 MS, 48 HC	Peripapillary and macular areas	Without ON	instance-wise	Accuracy=91% Recall=89% Specificity=92% AUC=0.97	GCL++ (between inner limiting membrane to INL) and nasal quadrant of outer and inner ring in pRNFL	SVM
Garcia-Martin et al. [42] 2020	48 MS, 48 HC	Macular area	Without ON	instance-wise	Recall=98% Specificity=98% AUC=0.83	GCL++	SVM, ANN
Zhang et al. [43] 2020	58 MS, 63 HC	Macular area	33 with ON, 25 without ON	instance-wise	Recall=64% Specificity=94%	GCIPL	LR, LR-EN, SVM
Montolio et al. [44] 2021	108 MS, 104 HC	Peripapillary and macular areas	34 with ON, 74 without ON	instance-wise	EC: Accuracy=87.7% Recall=87% Specificity=88.5% Precision=88.7% AUC=0.8775 K-NN: Accuracy=85.4% SVM: Accuracy=84.4% LSTM: Accuracy=81.7% Recall=81.1% Specificity=82.2% Precision=78.9% AUC=0.8165	pRNFL	MLR, SVM, decision tree, k-NN, NB, EC, LSTM recurrent neural network
Proposed algorithm With training and testing on first dataset	106 MS, 422 HC	Macular area	With and without ON	Patient-wise	Accuracy = 88% Precision = 78% Recall =63% F1-score = 68%	GCIPL and INL	Elaborated in the text
Proposed algorithm With training on first dataset and testing on second dataset	Train: 106 MS, 422 HC Test: 67 MS 45 HC	Macular area	With and without ON	Patient-wise	Accuracy = 88% Precision = 89% Recall =88% F1-score = 84%	GCIPL and INL	Elaborated in the text

LR: logistic regression, LR-EN: logistic regression regularized with the elastic net penalty, MLR: multiple linear regression, k-NN: k-nearest neighbors, NB: Naïve Bayes, EC: ensemble classifier, LSTM: long short-term memory

This article differentiates itself from prior investigations by utilizing two separate datasets, thereby augmenting the reliability and validity of its results. Conversely, a noteworthy constraint of previous studies pertains to their absence of external validation. The objective of this study is to overcome the use of deep learning methods, given the scarcity of accessible data. The research convincingly exhibits that desirable outcomes can be attained without dependence on deep learning techniques.

There are several limitations to the present study. First, the state of having a history of ON – a frequent clinical feature in MS - has not been considered[45,46]. Second, a longitudinal follow-up data from patients were not taken into consideration. Third, as we didn't have access to other devices, two devices that were used are Heidelberg with different HEYEX versions (5.7.5 and

5.1). If we had access to other devices like TOPCON or ZEISS, we had to test the trained model first. It is possible that the model is not generalizable in this stage. In the next step, it should add a limited number of the new data in train set to help the model get involved with this new dataset. We expect that would yield to better results that can somehow indicate the generalizability. In conclusion, this machine learning approach is designed to fill the gap in previous automatic methods for discrimination of MS and HCs. The relatively big sample size with manually corrected multilayer segmented OCT is used. Various topologies from sub-retinal thicknesses maps are individually analyzed to find the best combination. Interpretability and generalizability are guaranteed with the proposed approaches and the overestimated results are avoided with patient-wise techniques. Future work should be done on more comprehensive datasets to prove the effectiveness of such methods in clinical applications.

Declarations of interest: none..

Supplementary Materials:

The codes are deposited in:

https://github.com/shimakhs/Interpretable-classification-OCT-in-patients-with-Multiple-Sclerosis-and-healthy-controls

Author Contributions: RK led the study and supervised the work. ZK, AUB, FP, HR, and RK contributed to the conception, and design of the study. ZK performed the software, formal analysis, and visualization. ZK and RK wrote the first draft of the manuscript, HZ and SM contributed to data analysis and evaluation. All authors contributed to manuscript revision and read and approved the submitted version. Collection of the German data set was supported by German Research Foundation DFG Excellence Cluster NeuroCure to FP and collection of Iranian data was done by FA.

Funding: The study was funded in part by a grant from Einstein Foundation (Einstein Junior Scholar (EJS-2018-446)) to RK. This work was supported in part by the Vice-Chancellery for Research and Technology, Isfahan University of Medical Sciences, under Grant 398995.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
Data Availability Statement: Not applicable.

Acknowledgments: The manuscript is presented as a preprint in https://assets.researchsquare.com/files/rs-1547669/v1_covered.pdf?c=1650558772 [47].

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this article.

References

1. D. S. Reich, C. F. Lucchinetti, and P. A. Calabresi, "Multiple Sclerosis," N. Engl. J. Med. **378**(2), 169–180 (2018).

2. M. Filippi, P. Preziosa, B. L. Banwell, F. Barkhof, O. Ciccarelli, N. De Stefano, J. J. G. Geurts, F. Paul, D. S. Reich, A. T. Toosy, A. Traboulsee, M. P. Wattjes, T. A. Yousry, A. Gass, C. Lubetzki, B. G. Weinshenker, and M. A. Rocca, "Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines," Brain **142**(7), 1858–1875 (2019).

3. J. S. Graves, F. C. Oertel, A. Van der Walt, S. Collorone, E. S. Sotirchos, G. Pihl-Jensen, P. Albrecht, E. A. Yeh, S. Saidha, J. Frederiksen, S. D. Newsome, and F. Paul, "Leveraging Visual Outcome Measures to Advance Therapy Development in Neuroimmunologic Disorders," Neurol. Neuroimmunol. neuroinflammation **9**(2), (2022).

4. F. Costello and J. M. Burton, "Retinal imaging with optical coherence tomography: A biomarker in multiple sclerosis?," Eye Brain **10**, 47–63 (2018).

5. F. Paul, P. A. Calabresi, F. Barkhof, A. J. Green, R. Kardon, J. Sastre-Garriga, S. Schippling, P. Vermersch, S. Saidha, and B. S. Gerendas, "Optical coherence tomography in multiple sclerosis: A 3-year prospective multicenter study," Ann. Clin. Transl. Neurol. **8**(12), 2235–2251 (2021).

6. F. C. Oertel, H. G. Zimmermann, A. U. Brandt, and F. Paul, "Novel uses of retinal imaging with optical coherence tomography in multiple sclerosis," Expert Rev. Neurother. **19**(1), 31–43 (2019).

7. H. Hu, H. Jiang, G. R. Gameiro, J. Hernandez, S. Delgado, and J. Wang, "Focal thickness reduction of the ganglion cell-inner plexiform layer best discriminates prior optic neuritis in patients with multiple sclerosis," Investig. Ophthalmol. Vis. Sci. **60**(13), 4257–4269 (2019).

8. C. Shi, H. Jiang, G. R. Gameiro, H. Hu, J. Hernandez, S. Delgado, and J. Wang, "Visual function and disability are associated with focal thickness reduction of the ganglion cell-inner plexiform layer in patients with multiple sclerosis," Invest. Ophthalmol. Vis. Sci. **60**(4), 1213–1223 (2019).

9. F. Pesapane, M. Codari, and F. Sardanelli, "Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine," Eur. Radiol. Exp. **2**(1), 1–10 (2018).

10. O. Oren, B. J. Gersh, and D. L. Bhatt, "Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints," Lancet Digit. Heal. **2**(9), e486–e488 (2020).

11. J. P. O. Li, H. Liu, D. S. J. Ting, S. Jeon, R. V. P. Chan, J. E. Kim, D. A. Sim, P. B. M. Thomas, H. Lin, Y. Chen, T. Sakomoto, A. Loewenstein, D. S. C. Lam, L. R. Pasquale, T. Y. Wong, L. A. Lam, and D. S. W. Ting, "Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective," Prog. Retin. Eye Res. **82**, 100900 (2021).

12. D. S. W. Ting, L. R. Pasquale, L. Peng, J. P. Campbell, A. Y. Lee, R. Raman, G. S. W. Tan, L. Schmetterer, P. A. Keane, and T. Y. Wong, "Artificial intelligence and deep learning in ophthalmology," Br. J. Ophthalmol. **103**(2), 167–175 (2019).

13. P. M. Maloca, P. L. Müller, A. Y. Lee, A. Tufail, K. Balaskas, S. Niklaus, P. Kaiser, S. Suter, J. Zarranz-Ventura, C. Egan, H. P. N. Scholl, T. K. Schnitzer, T. Singer, P. W. Hasler, and N. Denk, "Unraveling the deep learning gearbox in optical coherence tomography image segmentation towards explainable artificial intelligence," Commun. Biol. **4**(1), 1–12 (2021).

14. J. Yoon, J. M. J. Han, J. I. Park, J. S. Hwang, J. M. J. Han, J. Sohn, K. H. Park, and D. D.-J. J. Hwang, "Optical coherence tomography-based deep-learning model for detecting central serous chorioretinopathy," Sci. Rep. **10**(1), 1–9 (2020).

15. J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger, "Clinically applicable deep learning for diagnosis and referral in retinal disease," Nat. Med. **24**(9), 1342–1350 (2018).

16. A. Petzold, P. Albrecht, L. Balcer, E. Bekkers, A. U. Brandt, P. A. Calabresi, O. G. Deborah, J. S. Graves, A. Green, and P. A. Keane, "Artificial intelligence extension of the OSCAR-IB criteria," Ann. Clin. Transl. Neurol. **8**(7), 1528–1542 (2021).

17. S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," Gigascience **6**(5), 1–9 (2017).

18. P. Tewarie, L. Balk, F. Costello, A. Green, R. Martin, S. Schippling, and A. Petzold, "The OSCAR-IB consensus criteria for retinal OCT quality assessment," PLoS One **7**(4), e34823 (2012).

19. S. Schippling, L. J. Balk, F. Costello, P. Albrecht, L. Balcer, P. A. Calabresi, J. L. Frederiksen, E. Frohman, A. J. Green, and A. Klistorner, "Quality control for retinal OCT in multiple sclerosis: validation of the OSCAR-IB criteria," Mult. Scler. J. **21**(2), 163–170 (2015).

20. S. Motamedi, K. Gawlik, N. Ayadi, H. G. Zimmermann, S. Asseyer, C. Bereuter, J. Mikolajczak, F. Paul, E. M. Kadas, and A. U. Brandt, "Normative Data and Minimally Detectable Change for Inner Retinal Layer Thicknesses Using a Semi-automated OCT Image Segmentation Pipeline," Front. Neurol. **10**, 1117 (2019).

21. F. Ashtari, A. Ataei, R. Kafieh, Z. Khodabandeh, M. Barzegar, M. Raei, A. Dehghani, and M. Mansurian, "Optical Coherence Tomography in Neuromyelitis Optica spectrum disorder and Multiple Sclerosis: A population-based study," Mult. Scler. Relat. Disord. **47**, 102625 (2021).

22. R. Kafieh, H. Rabbani, M. D. Abramoff, and M. Sonka, "Intra-retinal layer segmentation of 3D optical coherence tomography using coarse grained diffusion map," Med. Image Anal. **17**(8), 907–928 (2013).

23. R. Kafieh, H. Rabbani, F. Hajizadeh, M. D. Abramoff, and M. Sonka, "Thickness mapping of eleven retinal layers segmented using the diffusion maps method in normal eyes," J. Ophthalmol. **2015**, (2015).

24. M. Montazerin, Z. Sajjadifar, E. Khalili Pour, H. Riazi-Esfahani, T. Mahmoudi, H. Rabbani, H. Movahedian, A. Dehghani, M. Akhlaghi, and R. Kafieh, "Livelayer: a semi-automatic software program for segmentation of layers and diabetic macular edema in optical coherence tomography images," Sci. Rep. **11**(1), 1–13 (2021).

25. A. Petzold, L. Balcer, P. A. Calabresi, F. Costello, T. Frohman, E. Frohman, E. H. Martinez-Lapiscina, A. Green, R. Kardon, O. Outteryck, F. Paul, S. Schippling, P. Vermersch, P. Villoslada, L. Balk, O. Aktas, P. Albrecht, J. Ashworth, N. Asgari, G. Black, D. Boehringer, R. Behbehani, L. Benson, R. Bermel, J. Bernard, A. Brandt, J. Burton, P. Calabresi, J. Calkwood, C. Cordano, A. Courtney, A. Cruz-Herranz, R. Diem, A. Daly, H. Dollfus, C. Fasser, C. Finke, J. Frederiksen, E. Garcia-Martin, I. G. Suárez, G. Pihl-Jensen, J. Graves, J. Havla, B. Hemmer, S. C. Huang, J. Imitola, H. Jiang, D. Keegan, E. Kildebeck, A. Klistorner, B. Knier, S. Kolbe, T. Korn, B. LeRoy, L. Leocani, D. Leroux, N. Levin, P. Liskova, B. Lorenz, J. L. Preiningerova, E. H. Martínez-Lapiscina, J. Mikolajczak, X. Montalban, M. Morrow, R. Nolan, T. Oberwahrenbrock, F. C. Oertel, C. Oreja-Guevara, B. Osborne, A. Papadopoulou, M. Ringelstein, S. Saidha, B. Sanchez-Dalmau, J. Sastre-Garriga, R. Shin, N. Shuey, K. Soelberg, A. Toosy, R. Torres, A. Vidal-Jordana, A. Waldman, O. White, A. Yeh, S. Wong, and H. Zimmermann, "Retinal layer segmentation in multiple sclerosis: a systematic review and meta-analysis," Lancet Neurol. **16**(10), 797–812 (2017).

26. J. Shlens, "A Tutorial on Principal Component Analysis," arXiv Prepr. arXiv1404.1100 (2014).

27. X. W. Chen, "Gene selection for cancer classification using bootstrapped genetic algorithms and support vector machines," Proc. 2003 IEEE Bioinforma. Conf. CSB 2003 **46**(1–3), 504–505 (2003).

28. I. G. and Y. B. and A. Courville, *Deep Learning* (MIT press Cambridge, 2016), **29**(7553).

29. C. Cavaliere, E. Vilades, M. A. C. Alonso-Rodríguez, M. J. Rodrigo, L. E. Pablo, J. M. Miguel, E. López-Guillén, E. M. A. Sánchez Morla, L. Boquete, and E. Garcia-Martin, "Computer-aided diagnosis of multiple sclerosis using a support vector machine and optical coherence tomography features," Sensors (Switzerland) **19**(23), 5323 (2019).

30. W. Mao and F.-Y. Wang, "Cultural Modeling for Behavior Analysis and Prediction," Adv. Intell. Secur. Informatics 91–102 (2012).

31. G. Zheng, S. Li, and G. Székely, *Statistical Shape and Deformation Analysis: Methods, Implementation and Applications* (Academic Press, 2017).

32. I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance," TELKOMNIKA (Telecommunication Comput. Electron. Control. **14**(4), 1502 (2016).

33. G. Asadollahfardi, "Artificial Neural Network," in *Interdisciplinary Computing in Java Programming* (Springer, 2015), pp. 77–91.

34. P. Shukla, A. Verma, S. Verma, and M. Kumar, "Interpreting SVM for medical images using Quadtree," Multimed. Tools Appl. **79**(39), 29353–29373 (2020).

35. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Springer, 2014), **8689 LNCS**(PART 1), pp. 818–833.

36. M. Bock, A. U. Brandt, J. Dörr, H. Kraft, N. Weinges-Evers, G. Gaede, C. F. Pfueller, K. Herges, H. Radbruch, and S. Ohlraun, "Patterns of retinal nerve fiber layer loss in multiple sclerosis patients with or without optic neuritis and glaucoma patients," Clin. Neurol. Neurosurg. **112**(8), 647–652 (2010).

37. C. Hsu, C. Chang, and C. Lin, "A practical guide to support vector machines," (2003).

38. L. Aly, C. Noll, R. Wicklein, E. Wolf, E. F. Romahn, J. Wauschkuhn, S. Hosari, C. Mardin, A. Berthele, and B. Hemmer, "Dynamics of Retinal Vessel Loss After Acute Optic Neuritis in Patients With Relapsing Multiple Sclerosis," Neurol. Neuroinflammation **9**(3), (2022).

39. E. Garcia-Martin, L. E. Pablo, R. Herrero, J. R. Ara, J. Martin, J. M. Larrosa, V. Polo, J. Garcia-Feijoo, and J. Fernandez, "Neural networks to identify multiple sclerosis with optical coherence tomography," Acta Ophthalmol. **91**(8), e628–e634 (2013).

40. E. Garcia-Martin, R. Herrero, M. P. Bambo, J. R. Ara, J. Martin, V. Polo, J. M. Larrosa, J. Garcia-Feijoo, and L. E. Pablo, "Artificial neural network techniques to improve the ability of optical coherence tomography to detect optic neuritis," in *Seminars in Ophthalmology* (Taylor & Francis, 2015), **30**(1), pp. 11–19.

41. A. P. del Palomar, J. Cegoñino, A. Montolío, E. Orduna, E. Vilades, B. Sebastián, L. E. Pablo, and E. Garcia-Martin, "Swept source optical coherence tomography to early detect multiple sclerosis disease. The use of machine learning techniques," PLoS One **14**(5), e0216410 (2019).

42. E. Garcia-Martin, M. Ortiz, L. Boquete, E. M. Sánchez-Morla, R. Barea, C. Cavaliere, E. Vilades, E. Orduna, and M. J. Rodrigo, "Early diagnosis of multiple sclerosis by OCT analysis using Cohen's d method and a neural network as classifier," Comput. Biol. Med. **129**, 104165 (2021).

43. J. Zhang, J. Wang, C. Wang, S. Delgado, J. Hernandez, H. Hu, X. Cai, and H. Jiang, "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 (2020).

44. A. Montolío, A. Martín-Gallego, J. Cegoñino, E. Orduna, E. Vilades, E. Garcia-Martin, and A. P. Del Palomar, "Machine learning in diagnosis and disability prediction of multiple sclerosis using optical coherence tomography," Comput. Biol. Med. **133**, 104416 (2021).

45. A. Petzold, C. L. Fraser, M. Abegg, R. Alroughani, D. Alshowaeir, R. Alvarenga, C. Andris, N. Asgari, Y. Barnett, and R. Battistella, "Diagnosis and classification of optic neuritis," Lancet Neurol. (2022).

46. M. Denis, J.-P. Woillez, V. M. Smirnov, E. Drumez, J. Lannoy, J. Boucher, M. Zedet, J.-P. Pruvo, J. Labreuche, and H. Zephir, "Optic Nerve Lesion Length at the Acute Phase of Optic Neuritis Is Predictive of Retinal Neuronal Loss," Neurol. Neuroinflammation **9**(2), (2022).

47. Z. Khodabandeh, H. Rabbani, F. Ashtari, H. G. Zimmermann, S. Motamedi, A. U. Brandt, F. Paul, and R. Kafieh, "Interpretable classi cation using occlusion sensitivity on multilayer segmented OCT from patients

with Multiple Sclerosis and healthy controls," (2022).

'Declarations of interest: none'.