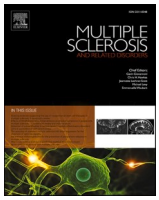
Multiple Sclerosis and Related Disorders 77 (2023) 104846

Contents lists available at ScienceDirect

Multiple Sclerosis and Related Disorders

journal homepage: www.elsevier.com/locate/msard



Discrimination of multiple sclerosis using OCT images from two different centers

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

ARTICLE INFO

Keywords:
Multiple sclerosis
Optical coherence tomography
Interpretable artificial intelligence
Generalizable
Patient-wise cross-validation

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 thicknesses of various retinal layers 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 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.

* Corresponding author at: Department of Engineering, Durham University, Durham, UK.
E-mail address: Raheleh.kafieh@durham.ac.uk (R. Kafieh).

https://doi.org/10.1016/j.msard.2023.104846
Received 22 May 2023; Accepted 19 June 2023
Available online 24 June 2023

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 (Reich et al., 2018). While the most established method to monitor the degree of CNS damage in MS is magnetic resonance imaging (MRI) (Filippi et al., 2019), 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 (Graves et al., 2022, Costello and Burton, 2018). 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 (Paul et al., 2021, Oertel et al., 2019). Inflammatory disease activity may also lead to changes in inner nuclear layer thickness (INL) (Oertel et al., 2019). Layer thinning can be measured by aligning and subtracting retinal layer thicknesses from a normal healthy population (Hu et al., 2019, Shi et al., 2019).

Artificial intelligence (AI) is a promising area of health innovation (Pesapane et al., 2018, Oren et al., 2020). Its application in ophthalmology is also evident in analysis of different ocular images (Li et al., 2021, Ting et al., 2019), with purpose of segmenting the retinal boundaries (Maloca et al., 2021), discriminating different diseases (Yoon et al., 2020, De Fauw et al., 2018) or interpretation of neurological diseases using quality control (QC) criteria (Petzold et al., 2021). Cross-sectional OCTs are successfully employed in AI for detection of ophthalmologic diseases. However, the alteration in thicknesses of various retinal layers 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 (Saeb et al., 2017).

Here we propose an AI method that aims to capture ultra-fine changes in thicknesses of various retinal 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 retinal layers. This study aims to obtain a classification algorithm using AI method based on changes in different 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 (Tewarie et al., 2012, Schippling et al., 2015). 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 (Motamedi et al., 2019). 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 (Ashtari et al., 2021). 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 (Tewarie et al., 2012) The dataset consists of 45 HC and 45 MS eyes. The automated segmentation was carried out using a graph-based method (Kafieh et al., 2013, Kafieh et al., 2015). All segmentation results were quality controlled and manually corrected in case of errors by an experienced grader using custom-developed software (Ashtari et al., 2021, Montazerin et al., 2021). 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 (Tewarie et al., 2012). This criteria has been validated for MS (Schippling et al., 2015).

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

2.2. Preprocessing and feature extraction

Intra-retinal thickness changes in MS are often noticeably subtle compared to primary eye disorders (Petzold et al., 2017). Multilayer segmented OCTs are therefore used and the distances between pairs of retinal layers, called retinal thickness maps, are calculated (Fig. 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*) (Fig. 2(a)); and the relative direction of each B-scan to a horizontal line through the disc center (*slope_Bscan*) (Fig. 2(b)) are calculated. The left eyes are also flipped. The value of correcting rotation (*rotation angle* – Fig. 2(c)) is calculated by:

$$\begin{aligned} rotation\ angle &= angle\_Fovea\_ONH\_SLine\ if\ slope\_Bscans = 0 \\ rotation\ angle &= abs(-90\ -\ angle\_Fovea\_ONH\_SLine)\ if\ 88 < slope\_Bscan <= 90 \\ rotation\ angle &= \left(angle\_Fovea\_ONH\_SLine\_\mathrm{deg}\right) + (slope\_Bscan)\ otherwise \end{aligned} \tag{1}$$

The rotated thickness maps are cropped to a unique size of 450 × 450 pixels (Fig. 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.

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) (Ashtari et al., 2021). 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 Fig. 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 Fig. 3.

2.3. Dimension reduction

To decrease the model complexity and avoid overfitting, we used principal component analysis (PCA) (Shlens, 2014) that deduces information from the feature set to make a new feature subspace. Recursive feature elimination (RFE) (Chen, 2003) 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 (G et al., 2016). 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

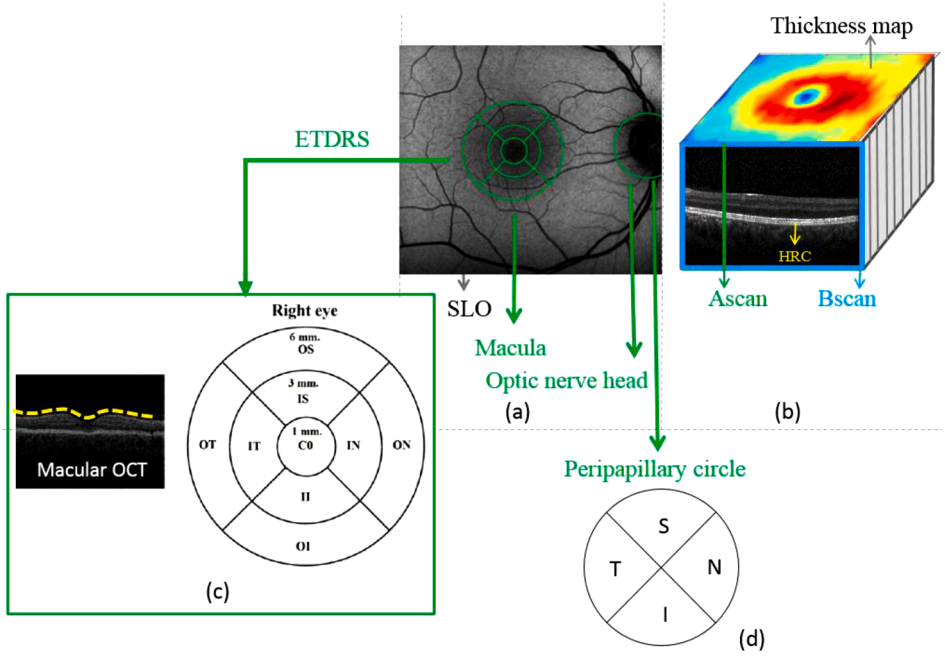


Fig. 1. Retinal parameters acquired by OCT. (a) location of sectors and ring scan on SLO image. (b) A-scan, B-scan, and thickness map of OCT data. (c) quadrants in ETDRS: central fovea (CF), inner superior (IS), inner nasal (IN), inner inferior (II), inner temporal (IT), outer superior (OS), outer superior (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) (Ashtari et al., 2021).

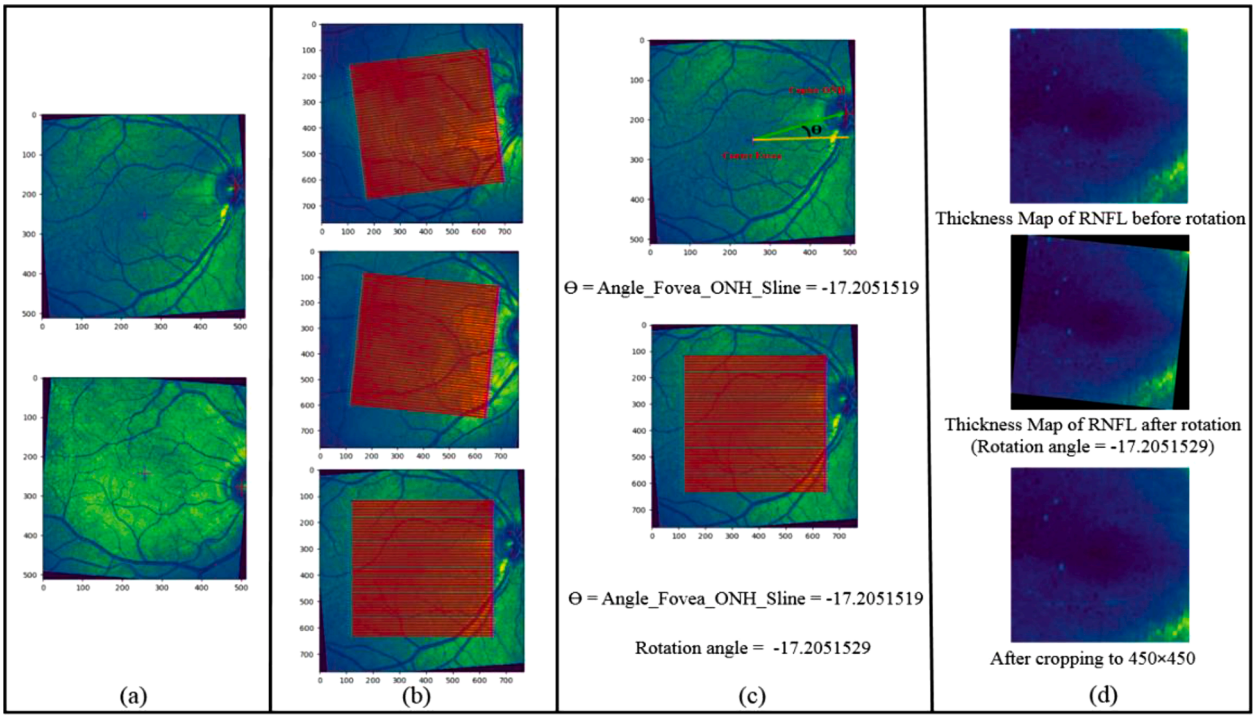


Fig. 2. (a) SLO image in clockwise and counterclockwise rotations. (b) B-scans in different directions. (c) Example of finding the rotation angle), (d) Process of rotating a thickness map.

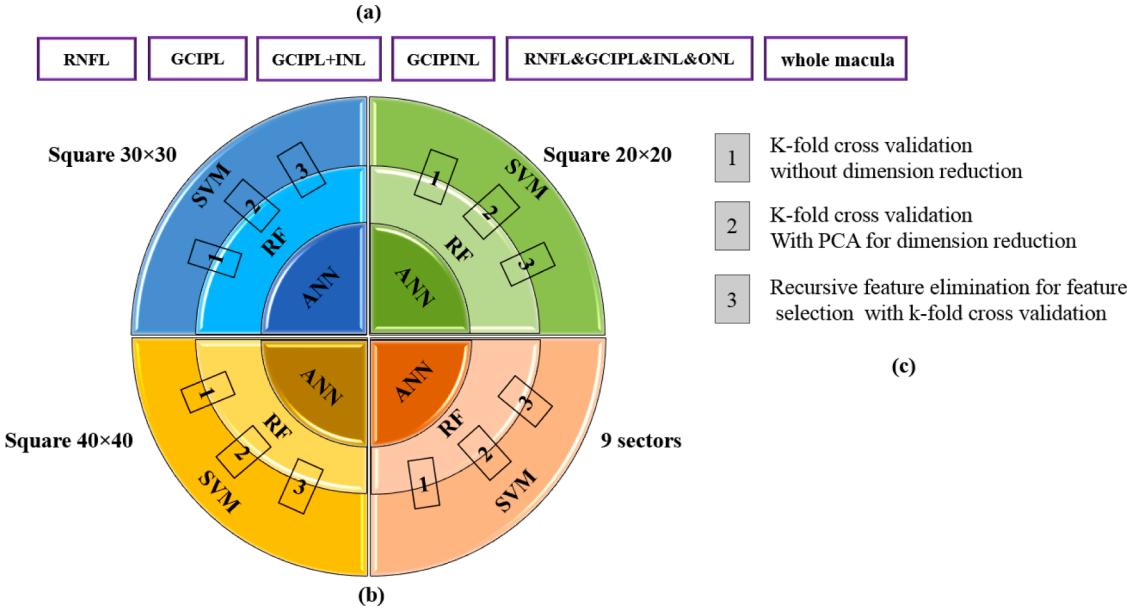


Fig. 3. (a) Retinal layers investigated by the proposed method. (b) Classification model: SVM, RF, and ANN on four groups of extracted features: square 20 × 20, square 30 × 30, square 40 × 40, and 9 ETDRS sectors. (c) Different evaluation and dimension reduction methods in the study.

be used to transfer the feature space to a higher dimension space with good separability (Cavaliere et al., 2019). 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 (Mao and Wang, 2012, Zheng et al., 2017). We used the grid search method (Syarif et al., 2016) 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 (Asadollahfardi, 2015). 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 unit (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 (Saeb et al., 2017) 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 the black box nature that 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 (Shukla et al., 2020).

In this study, we propose a novel approach to add interpretability to current machine learning approaches. We used occlusion sensitivity (Zeiler and Fergus, 2014) 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 (Fig. 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.

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 the 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,

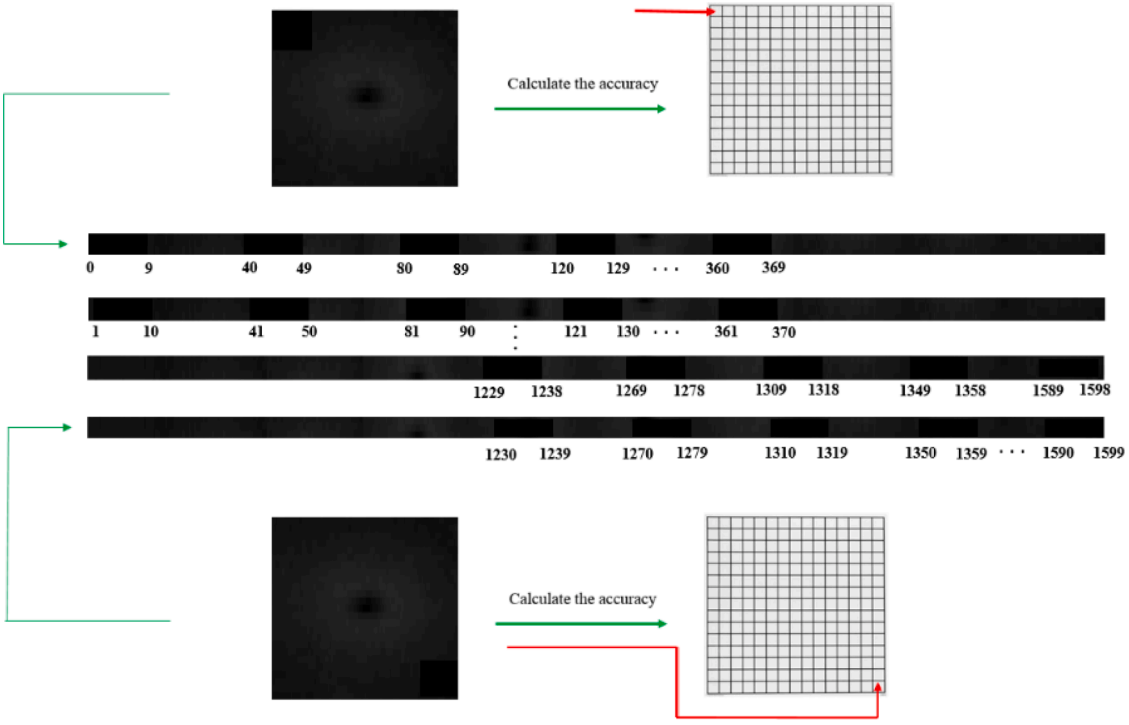


Fig. 4. The proposed process for creating a black mask, moving it to the test set, and transferring the locations to vector-shaped positions

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%

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).

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 Fig. 5.

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 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%

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 Fig. 6.

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 thicknesses of various retinal layers 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.

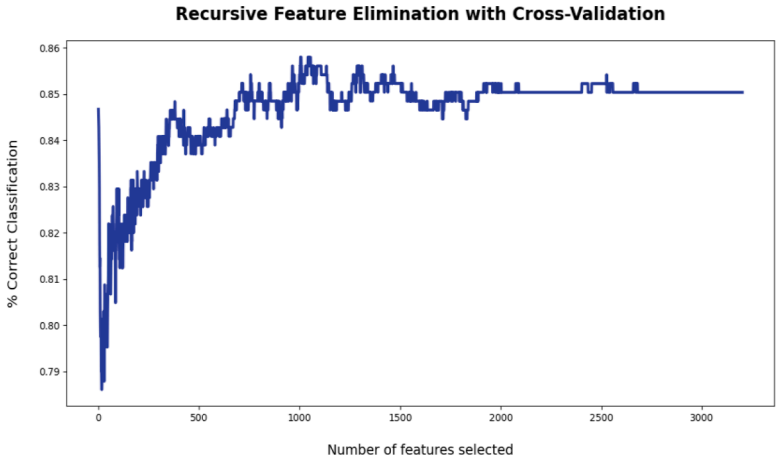


Fig. 5. RFE with cross-validation diagram that shows accuracy with the number of features.

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%

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 (Petzold et al., 2017, Oertel et al., 2019). 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 Fig. 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 (Bock et al., 2010).

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 (Hsu et al., 2003). 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

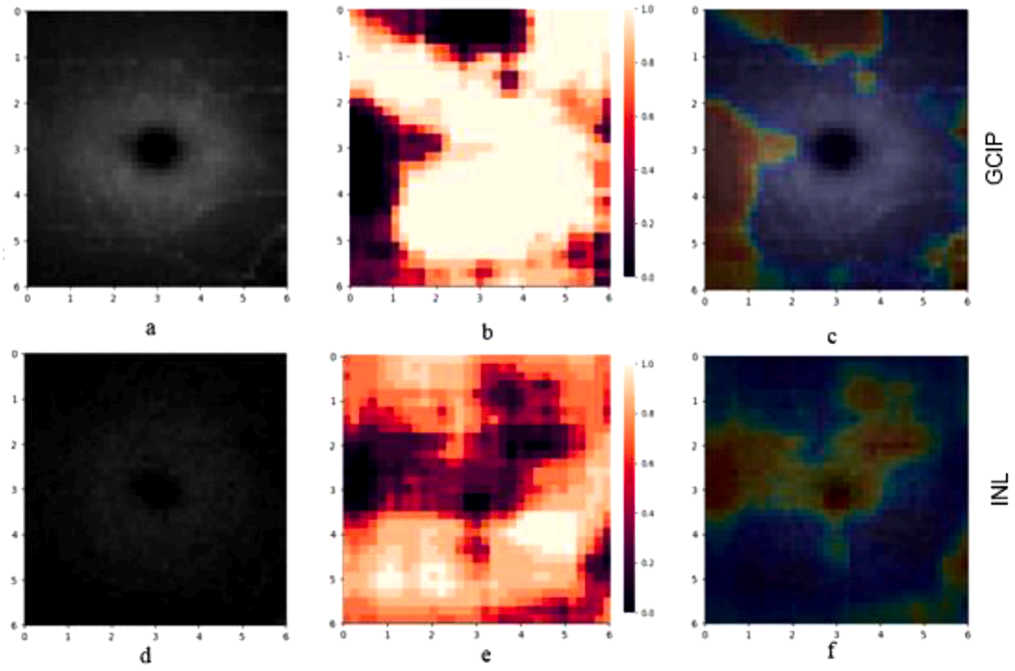


Fig. 6. Visual interpretability on thickness maps of GCIP and INL. (a) Thickness map of GCIP in one sample from MS dataset (x and y axis in mm), (b) heatmap of occlusion sensitivity in the classification of MS and HC, (c) overlap of the heatmap and the GCIP layer, (d) Thickness map of INL in one sample 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

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. (Garcia-Martin et al., 2013) 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. (Garcia-Martin et al., 2015) 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. (del Palomar et al., 2019) 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% AUC=0.998	pRNFL	Decision tree, ANN, SVM
Cavaliere et al. (Cavaliere et al., 2019) 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. (Garcia-Martin et al., 2021) 2020	48 MS, 48 HC	Macular area	Without ON	instance-wise	Recall=98% Specificity=98% AUC=0.83	GCL++	SVM, ANN
Zhang et al. (Zhang et al., 2020) 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. (Montolío et al., 2021) 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

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 (Oertel et al., 2019, Petzold et al., 2017, Aly et al., 2022). 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.

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 (Petzold et al., 2022, Denis et al., 2022). 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 thicknesses maps of various retinal layers 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.

Supplementary materials

The codes are deposited in: https://zenodo.org/record/8092074.

Institutional review board statement

Not applicable.

Informed consent statement

Not applicable.

Data availability statement

Not applicable.

CRediT authorship contribution statement

Zahra Khodabandeh: Conceptualization, Methodology, Validation, Supervision, Formal analysis, Software, Visualization, Writing – original draft, Writing – review & editing. **Hossein Rabbani:** Conceptualization, Methodology, Validation, Supervision, Writing – review & editing. **Fereshteh Ashtari:** Writing – review & editing, Resources. **Hanna G. Zimmermann:** Validation, Data curation, Resources, Writing – review & editing. **Seyedamirhosein Motamedi:** Validation, Data curation, Resources, Writing – review & editing. **Alexander U. Brandt:** Conceptualization, Methodology, Validation, Supervision, Writing – review & editing, Funding acquisition. **Friedemann Paul:** Conceptualization, Methodology, Validation, Supervision, Writing – review & editing, Resources, Funding acquisition. **Rahele Kafieh:** Conceptualization, Methodology, Validation, Supervision, Writing – original draft, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

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

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.

Acknowledgments

The manuscript is presented as a preprint in https://assets.researchsquare.com/files/rs-1547669/v1_covered.pdf?c=1650558772 (Khodabandeh et al., 2022).

References

Aly, L., Noll, C., Wicklein, R., Wolf, E., Romahn, E.F., Wauschkuhn, J., Hosari, S., Mardin, C., Berthele, A., Hemmer, B., 2022. Dynamics of retinal vessel loss after acute optic neuritis in patients with relapsing multiple sclerosis. Neurol. Neuroinflam. 9 (3).

Asadollahfardi, G., 2015. Artificial Neural Network. Interdisciplinary Computing in Java Programming. Springer, pp. 77–91.

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

Bock, M., Brandt, A.U., Dörr, J., Kraft, H., Weinges-Evers, N., Gaede, G., Pfueller, C.F., Herges, K., Radbruch, H., Ohlraun, S., 2010. 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.

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

Chen, X.W., 2003. Gene selection for cancer classification using bootstrapped genetic algorithms and support vector machines. In: Proc. 2003 IEEE Bioinforma. Conf. CSB 2003, 46, pp. 504–505.

Costello, F., Burton, J.M., 2018. Retinal imaging with optical coherence tomography: a biomarker in multiple sclerosis? Eye Brain 10, 47–63.

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

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

Denis, M., Woillez, J.-P., Smirnov, V.M., Drumez, E., Lannoy, J., Boucher, J., Zedet, M., Pruvo, J.-P., Labreuche, J., Zephir, H., 2022. Optic nerve lesion length at the acute phase of optic neuritis is predictive of retinal neuronal loss. Neurol. Neuroinflam. 9 (2).

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

G, I., B, Y., Courville, A., 2016. Deep Learning. MIT press Cambridge, p. 29.

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

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

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

Graves, J.S., Oertel, F.C., Van der Walt, A., Collorone, S., Sotirchos, E.S., Pihl-Jensen, G., Albrecht, P., Yeh, E.A., Saidha, S., Frederiksen, J., Newsome, S.D., Paul, F., 2022. Leveraging visual outcome measures to advance therapy development in neuroimmunologic disorders. Neurol. Neuroimmunol. neuroinflam. 9 (2).

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

Hu, H., Jiang, H., Gameiro, G.R., Hernandez, J., Delgado, S., Wang, J., 2019. 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.

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

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

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).

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

Maloca, P.M., Müller, P.L., Lee, A.Y., Tufail, A., Balaskas, K., Niklaus, S., Kaiser, P., Suter, S., Zarranz-Ventura, J., Egan, C., Scholl, H.P.N., Schnitzer, T.K., Singer, T., Hasler, P.W., Denk, N., 2021. Unraveling the deep learning gearbox in optical

coherence tomography image segmentation towards explainable artificial intelligence. Commun. Biol. 4 (1), 1–12.
Mao, W., Wang, F.-Y., 2012. Cultural modeling for behavior analysis and prediction. Adv. Intell. Secur. Informatics 91–102.
Montazerin, M., Sajjadifar, Z., Khalili Pour, E., Riazi-Esfahani, H., Mahmoudi, T., Rabbani, H., Movahedian, H., Dehghani, A., Akhlaghi, M., Kafieh, R., 2021. 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.
Montolío, A., Martín-Gallego, A., Cegoñino, J., Orduna, E., Vilades, E., Garcia-Martin, E., Del Palomar, A.P., 2021. Machine learning in diagnosis and disability prediction of multiple sclerosis using optical coherence tomography. Comput. Biol. Med. 133, 104416.
Motamedi, S., Gawlik, K., Ayadi, N., Zimmermann, H.G., Asseyer, S., Bereuter, C., Mikolajczak, J., Paul, F., Kadas, E.M., Brandt, A.U., 2019. Normative data and minimally detectable change for inner retinal layer thicknesses using a semi-automated OCT image segmentation pipeline. Front. Neurol. 10, 1117.
Oertel, F.C., Zimmermann, H.G., Brandt, A.U., Paul, F., 2019. Novel uses of retinal imaging with optical coherence tomography in multiple sclerosis. Expert Rev. Neurother. 19 (1), 31–43.
Oren, O., Gersh, B.J., Bhatt, D.L., 2020. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. Lancet Digit. Heal. 2 (9), e486–e488.
Paul, F., Calabresi, P.A., Barkhof, F., Green, A.J., Kardon, R., Sastre-Garriga, J., Schippling, S., Vermersch, P., Saidha, S., Gerendas, B.S., 2021. Optical coherence tomography in multiple sclerosis: a 3-year prospective multicenter study. Ann. Clin. Transl. Neurol. 8 (12), 2235–2251.
Pesapane, F., Codari, M., Sardanelli, F., 2018. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur. Radiol. Exp. 2 (1), 1–10.
Petzold, A., Albrecht, P., Balcer, L., Bekkers, E., Brandt, A.U., Calabresi, P.A., Deborah, O. G., Graves, J.S., Green, A., Keane, P.A., 2021. Artificial intelligence extension of the OSCAR-IB criteria. Ann. Clin. Transl. Neurol. 8 (7), 1528–1542.
Petzold, A., Balcer, L., Calabresi, P.A., Costello, F., Frohman, T., Frohman, E., Martinez-Lapiscina, E.H., Green, A., Kardon, R., Outteryck, O., Paul, F., Schippling, S., Vermersch, P., Villoslada, P., Balk, L., Aktas, O., Albrecht, P., Ashworth, J., Asgari, N., Black, G., Boehringer, D., Behbehani, R., Benson, L., Bermel, R., Bernard, J., Brandt, A., Burton, J., Calabresi, P., Calkwood, J., Cordano, C., Courtney, A., Cruz-Herranz, A., Diem, R., Daly, A., Dollfus, H., Fasser, C., Finke, C., Frederiksen, J., Garcia-Martin, E., Suárez, I.G., Pihl-Jensen, G., Graves, J., Havla, J., Hemmer, B., Huang, S.C., Imitola, J., Jiang, H., Keegan, D., Kildebeck, E., Klistorner, A., Knier, B., Kolbe, S., Korn, T., LeRoy, B., Leocani, L., Leroux, D., Levin, N., Liskova, P., Lorenz, B., Preiningerova, J.L., Martínez-Lapiscina, E.H., Mikolajczak, J., Montalban, X., Morrow, M., Nolan, R., Oberwahrenbrock, T., Oertel, F.C., Oreja-Guevara, C., Osborne, B., Papadopoulou, A., Ringelstein, M., Saidha, S., Sanchez-Dalmau, B., Sastre-Garriga, J., Shin, R., Shuey, N., Soelberg, K., Toosy, A., Torres, R., Vidal-Jordana, A., Waldman, A., White, O., Yeh, A., Wong, S., Zimmermann, H., 2017. Retinal layer segmentation in multiple sclerosis: a systematic review and meta-analysis. Lancet Neurol. 16 (10), 797–812.
Petzold, A., Fraser, C.L., Abegg, M., Alroughani, R., Alshowaeir, D., Alvarenga, R., Andris, C., Asgari, N., Barnett, Y., Battistella, R., 2022. Diagnosis and classification of optic neuritis. Lancet Neurol. 21 (12), 1120–1134.
Reich, D.S., Lucchinetti, C.F., Calabresi, P.A., 2018. Multiple Sclerosis. N. Engl. J. Med. 378 (2), 169–180.
Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C., Kording, K.P., 2017. The need to approximate the use-case in clinical machine learning. Gigascience 6 (5), 1–9.
Schippling, S., Balk, L.J., Costello, F., Albrecht, P., Balcer, L., Calabresi, P.A., Frederiksen, J.L., Frohman, E., Green, A.J., Klistorner, A., 2015. Quality control for retinal OCT in multiple sclerosis: validation of the OSCAR-IB criteria. Mult. Scler. J. 21 (2), 163–170.
Shi, C., Jiang, H., Gameiro, G.R., Hu, H., Hernandez, J., Delgado, S., Wang, J., 2019. 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.
J. Shlens, "A tutorial on principal component analysis," arXiv Prepr. arXiv1404.1100 (2014).
Shukla, P., Verma, A., Verma, S., Kumar, M., 2020. Interpreting SVM for medical images using Quadtree. Multimed. Tools Appl. 79 (39), 29353–29373.
Syarif, I., Prugel-Bennett, A., Wills, G., 2016. SVM parameter optimization using grid search and genetic algorithm to improve classification performance. TELKOMNIKA (Telecommun. Comput. Electron. Control. 14 (4), 1502.
Tewarie, P., Balk, L., Costello, F., Green, A., Martin, R., Schippling, S., Petzold, A., 2012. The OSCAR-IB consensus criteria for retinal OCT quality assessment. PLoS One 7 (4), e34823.
Ting, D.S.W., Pasquale, L.R., Peng, L., Campbell, J.P., Lee, A.Y., Raman, R., Tan, G.S.W., Schmetterer, L., Keane, P.A., Wong, T.Y., 2019. Artificial intelligence and deep learning in ophthalmology. Br. J. Ophthalmol. 103 (2), 167–175.
Yoon, J., Han, J.M.J., Park, J.I., Hwang, J.S., Han, J.M.J., Sohn, J., Park, K.H., Hwang, D. D.-J.J., 2020. Optical coherence tomography-based deep-learning model for detecting central serous chorioretinopathy. Sci. Rep. 10 (1), 1–9.
Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8689 LNCS. Springer, pp. 818–833.
Zhang, J., Wang, J., Wang, C., Delgado, S., Hernandez, J., Hu, H., Cai, X., Jiang, H., 2020. Wavelet features of the thickness map of retinal ganglion cell-inner plexiform layer best discriminate prior optic neuritis in patients with multiple sclerosis. IEEE Access 8, 221590–221598.
Zheng, G., Li, S., Székely, G., 2017. Statistical Shape and Deformation Analysis: Methods, Implementation and Applications. Academic Press.