Comprehensive Evaluation of Artificial Intelligence Models for Diagnosis of Multiple Sclerosis Using Information from Retinal Layers Multicenter OCT Images

Zahra Khodabandeh[a], Hossein Rabbani[a], Neda Shirani Bidabadi[b], Mehdi Bonyani[c], Rahele Kafieh[a,d,*]

[a] *School of Advanced Technologies in Medicine, Medical Image and Signal Processing Research Center, Isfahan University of Medical Sciences, Isfahan, Iran*

[b] *Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran*

[c] *Louisiana State University, USA*

[d] *Department of Engineering, Durham University, Durham, UK*

Abstract

Multiple sclerosis (MS) is a chronic inflammatory disease that affects the central nervous system. Optical coherence tomography (OCT) is a retinal imaging technology with great promise as a possible MS biomarker. Unlike other ophthalmologic diseases, the variations in shape of raw cross-sectional OCTs in MS are subtle and not differentiable from healthy controls (HCs). More detailed information like thickness of particular layers of retinal tissues or surface of individual retinal boundaries are more appropriate discriminators for this purpose. Artificial Intelligence (AI) has demonstrated a robust performance in feature extraction and classification of retinal OCTs in different ophthalmologic diseases using OCTs. We explore a comprehensive range of AI models including (1) feature extraction with autoencoder (AE) and shallow networks for classification, (2) classification with deep networks designed from scratch, and (3) fine-tuning of pretrained networks (as a generic model of the visual world) for this specific application. We also investigate different input data including thickness and surfaces of different retinal layers to find the most representative data for discrimination of MS. Moreover, channel-wise combination and mosaicing of multiple inputs are examined to find the better merging model. To address interpretability requirement of AI models in clinical applications, the visualized contribution of each input data to the classification performance is shown using occlusion sensitivity and Grad-CAM approaches. The data used in this study includes 38 HC and 78 MS eyes from two independent public and local datasets. The effectiveness and generalizability of the classification methods are demonstrated by testing the network on these independent datasets. The most discriminative topology for classification, utilizing the proposed deep network designed from scratch, is determined when the inputs consist of a channel-wise combination of the thicknesses of the three layers of the retina, namely the retinal fiber layer (RNFL), ganglion cell and inner plexiform layer (GCIP), and inner nuclear layer (INL). This structure resulted in balanced-accuracy of 97.3, specificity of 97.3, recall 97.4%, and g-mean of 97.3% in discrimination of MS and HC OCTs.

Keywords: multiple sclerosis, optical coherence tomography, artificial intelligence, autoencoder, feature extraction, convolutional neural network

I. Introduction

Multiple sclerosis (MS) is a chronic inflammatory and neurodegenerative disease characterized by demyelination and neuro-axonal degeneration in the central nervous system (CNS) resulting in progressive tissue loss and worsening neurological disability over time[1]. Symptoms like fatigue, weakness, spasticity, vision changes, dizziness, mental changes, emotional changes, and depression are among the most prevalent in MS patients[2–4]. Various parts of the nervous system are affected in MS, including myelin sheaths, nerve fibers, and even cells that make myelin[5]. A few weeks are enough to recover if the injuries are not too severe, but severe injuries can permanently change the spinal cord[6]. Sclerosis refers to these permanent changes, which sometimes occur in multiple different parts of the body[7], thus giving rise to the name "multiple sclerosis". The cause of MS is unknown[8,9]. Scientists believe there is a combination of environmental and genetic factors involved in MS[10,11]. MS may be influenced by factors such as geography, vitamin D deficiency, obesity, and smoking[12,13]. The diagnosis of MS is established by applying the McDonald criteria[14], which primarily rely on the amalgamation of clinical, imaging, and laboratory evidence. However, it should be noted that the McDonald criteria are most applicable in cases where the disease has progressed to a relatively advanced stage. While magnetic resonance imaging (MRI) is the most established method to detect inflammations and lesions in MS[15], it is not sensitive or specific enough to reveal the degree of axonal damage[16].

MS causes extensive changes in the retina and optic nerve, which can be detected by a faster alternative method. Optical coherence tomography (OCT) is an imaging technology, which enables us to measure biomarkers such as thinning of individual layers in retina. [17,18]. The OCT technique uses in-vivo non-invasive imaging principles and is fast, with a temporal resolution of a few seconds. It allows for the precise delineation and measurement of various neuro-retinal layer thicknesses.

In particular, OCT has shown that the ganglion cell layer (GCL) has a gradual thinning from the early stages of MS, but the retinal nerve fiber layer (RNFL) serves as an indicator of axonal integrity[19–21]. As a potential biomarker for neurodegeneration, thinning of the peripapillary retinal nerve fiber layer (pRNFL), GCL, and inner plexiform layer thickness (IPL) are direct indicators of axonal damage[22,23]. Activation of inflammatory disease also alters the thickness of the inner nuclear layer (INL) [24]. These findings highlight the potential usefulness of measurements obtained from OCT as biomarkers for diagnostic applications in MS[25].

Artificial Intelligence (AI) has been widely used for automatic analysis of ophthalmic imaging modalities including OCT[26]. Instead of utilizing conventional statistical analysis, AI approaches can be used to predict the progression of MS-associated disability. Traditional methods for analyzing retinal layer thicknesses obtained through OCT entail extracting relevant features from the images and subsequently classifying them using specific techniques such as support vector machines (SVM)[27,28], linear discriminant function (LDF)[29], artificial neural network (ANN) [30–32], decision tree [32], logistic regression (LR) and logistic regression regularized with the elastic net penalty (LR-EN) [33], multiple linear regression (MLR), k-nearest neighbors

(k-NN), Naïve Bayes (NB), ensemble classifier (EC), long short-term memory (LSTM) [34], and recurrent neural network [35].

Recent scientific investigations have explored the utilization of machine learning (ML) methodologies specifically focused on analyzing data pertaining to the thickness of the RNFL in each of the four quadrants delineating the peripapillary region (pRNFL), as well as GCL++ (between inner limiting membrane to INL), and the macular ganglion cell-inner plexiform layer (GCIPL). Furthermore, additional factors including age, sex, best-corrected visual acuity, and other relevant data were duly considered in the analysis[27–34]. In our prior research[36], we devised a machine learning approach to overcome limitations present in earlier automated methods for distinguishing between individuals with MS and HCs. To adhere to the principles of trustworthy AI, we improved interpretability by employing the occlusion sensitivity approach, allowing for the visualization of regional layer contributions to classification performance. The algorithm's robustness was demonstrated through testing on a new independent dataset, confirming its effectiveness. Utilizing a dimension reduction method, we identified the most discriminative features from diverse topologies of multilayer segmented OCTs. Classification was carried out using SVM, Random Forest (RF), and ANN. The performance of the algorithm was evaluated through patient-wise cross-validation, ensuring that training and test folds encompassed records from distinct subjects[36].

Deep learning (DL) is a class of state-of-the-art ML techniques that has attracted a lot of global attention in recent years[37]. This simulates the operation of the human brain using numerous layers of artificial neural networks that can produce automatic predictions from input data. Unlike ML, the structure of DL uses more hidden layers to decode raw images without using manually creating specific features or feature selection algorithms. One of the notable attributes of deep learning networks is their capacity to deduce and extract latent and intrinsic features from data[26]. As a result, DL does not require any manual intervention at the stage of feature extraction, allowing the feature extraction and classification steps to be fully integrated into a Computer-Aided Diagnosis System (CADS)[38–40].

Convolutional neural networks (CNNs) are widely used in deep learning applications. They are inspired by the human visual system's complex structure[41]. CNNs process images directly through convolutional layers, extracting features to create a predictive model. This eliminates the need for a separate detection model and provides an end-to-end solution that takes input OCT images and produces diagnostic decisions.

CNN systems have demonstrated strong diagnostic performance in applying ocular imaging, mostly OCT and fundus photographs[42]. Major ocular diseases which CNN methods have been applied for include diabetic retinopathy(DR) [43–45], glaucoma[46], age-related macular degeneration (AMD)[47,48], retinopathy of prematurity (ROP)[49], diabetic macular edema (DME), choroidal neovascularization (CNV), drusen, and others. Some of the classification methods based on CNN have been used to classify the OCT images in different ocular diseases[50–60]. Regarding MS, two studies have employed CNNs for classification. The first one that has utilized a CNN architecture for discriminating between individuals with MS and HC is [61], wherein CNNs feed with images of retinal layer thicknesses obtained with swept source OCT (SS-OCT) are used to diagnose early MS. To enhance the training dataset for CNNs,

synthetic images representing retinal thickness are generated through the utilization of generative adversarial networks (GANs). This approach allows for the augmentation of the training data by producing synthetic images that capture variations in retinal thickness. The discriminative capability of the images is assessed through the utilization of effect sizes, specifically Cohen's distance. Cohen's parameter value is computed for each pixel in layers of retina. Pixels surpassing the predetermined threshold retain their thickness values, while those falling below the threshold are assigned a value of 0. Consequently, the input images for the CNN solely contain pixel information that is pertinent to the diagnosis of MS. As a result, the images fed into CNN only contain data on pixels that are pertinent to the diagnosis of MS.

In a recent investigation by Garcia Martin et al.[62], the Posterior Pole Protocol Spectralis SD-OCT, which incorporates an anatomic positioning system, was utilized for quantifying thickness measurements of the GCL, IPL, and RNFL. These measurements were subsequently employed as input data for a CNN designed for classification purposes. By using the mean thickness of the GCL (AUROC = 0.82) and the inter-eye difference in the IPL (AUROC = 0.71) as input features for a two-layer CNN, the automated diagnostic system achieved an overall accuracy of 0.87, a sensitivity of 0.82, and a specificity of 0.92.

In this study, we investigate an extensive array of AI models including (1) feature extraction with autoencoder (AE) and shallow networks for classification, (2) classification with deep networks designed from scratch, and (3) fine-tuning of pretrained network like ResNet152V2 for this specific application. In addition, we conduct an in-depth exploration of various input data, encompassing the thickness and surface measurements of different retinal layers, with the objective of identifying the most informative and representative data for discriminating MS. Moreover, we analyze the channel-wise combination and mosaicing techniques of multiple inputs, aiming to identify an optimal merging model that achieves improved performance.

This research explores the combination of different retinal layers, emphasizing the most effective composition of layers, specifically the mRNFL, GCIP, and INL. The results of this study are outlined in this manuscript. In order to meet the interpretability demands of AI models in clinical applications, we present the visualization of the contribution made by each input data to the classification performance. This is achieved through the utilization of occlusion sensitivity and Grad-CAM approaches. The data used in this study obtained from two independent public and local datasets. The efficacy and generalizability of the classification methods are substantiated through testing of the network on independent datasets. This evaluation showcases the ability of the methods to perform effectively beyond the initial training dataset. Our goal is to develop AI-based classification systems that can aid neurologists in the early detection of MS by analyzing alterations in sub-retinal layers of OCT.

II. Materials and Methods

A. Databases:

The data used in this study includes 38 HC and 78 MS eyes that were achieved from two independent (but similar) datasets. The first dataset (Public dataset) has 14 HC and 18 MS and the second one (Local dataset) contains 24 HC and 60 MS. Demographic information of both

datasets are included in Table 1. First, the models are trained and tested on the whole dataset and then, two datasets are used as separate training and testing datasets in measuring the robustness of the best-performing set of the classification algorithm.

Public dataset: The first dataset consists of 32 human retina scans taken with Spectralis SD-OCT equipment (Heidelberg Engineering, Heidelberg, Germany), 14 HC, and 18 MS OCTs which were manually segmented to eight retinal layers using internally developed software and reviewed and revised independently. At least, 12 images at the same location have been averaged for each Bscans, and the signal-to-noise ratio of the final averaged scans is at least 20 dB[63].

Local dataset: The second OCT dataset was collected between April 2017 and March 2019 at the Kashani Comprehensive MS Center in Isfahan, Iran[64]. The images were acquired using Spectralis SD-OCT and Heidelberg HEYEX version 5.1 by one trained technician with automatic real-time (ART) of 9 frames function for image averaging. The dataset consists of 24 HC and 60 MS eyes. All scans were checked for sufficient quality using OSCAR-IB criteria[65]. OSCAR-IB criterion is a standard for evaluating the quality of OCT images. Several indicators are considered as a 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 [65]. This criterion has been validated for MS [66].

Table 1: Demographic characteristics of participants of the Local and Public datasets

Dataset	Characteristics	HC	MS
Public	Current age, Y, mean ±SD	35.77±13.03	41.97±8.77
	Sex, F, n(%)	12(86%)	14(78%)
Local	Current age, Y, mean ±SD	26.3±3.06	34.5±8.03
	Sex, F, n(%)	12(66%)	30(85%)

y: year, SD: standard deviation, n: number

B. Data Preparation:

Cross-sectional OCT scans have been effectively utilized in AI applications for the detection of various ophthalmologic diseases. However, the changes in sub-retinal thicknesses associated with MS are notably subtle, making it challenging to diagnose using unprocessed cross-sectional OCT scans alone. Hence, multilayer segmented OCTs are used rather than raw B-scans.
Segmentation of the retinal layers is the key first step for AI-based analysis of MS by OCT images. Achieving precise and automated segmentation of retinal boundaries is crucial in this context, despite the notable challenges frequently encountered in images obtained from individuals with neurodegenerative diseases (NDD). These challenges include the presence of atrophic layers and ambiguous boundaries, media opacity, and artifacts resulting from patient movement. Overcoming these hurdles is essential for accurate analysis and diagnosis in such cases. The NDD-SEG presents a viable solution to this challenge. It is an optimal algorithm would possess the capability to effectively process various OCT devices and imaging protocols characterized by a wide range of resolution, size, artifacts, and noise[67]. So, the NDD-SEG approach is utilized for the segmentation of retinal layers in this study. By applying the NDD-SEG model, 9 boundaries and 8 layers of retina are extracted. The distances between pairs of

retinal layers generates retinal thickness maps. After plotting the thickness maps, it becomes evident that certain maps exhibit destroyed regions. This destruction can be attributed to low quality, missed B-scan or the presence of noise in the B-scans used for generating the thickness maps. To address this issue and enhance the quality of the maps while simultaneously reducing noise, an inpainting algorithm is employed.

The data preparation steps are depicted in Figure 2 and are divided into the following sequential steps.

1) Segmentation of retinal layers

The fundamental design of NDD-SEG[67] consists of a U-net structure that lacks pooling and upsampling layers. An initial network, referred to as the Retinal Tissue Segmentation Network (RTSN), proficiently identifies the presence of the retina within the image, even in the presence of a noisy background. Following this, the Retinal Layer Segmentation Network (RLSN) focuses on precisely segmenting the intra-retinal layers by utilizing the probability map derived from RTSN along with the original input image. In the decoder stages of RLSN, the probability map and texture features are integrated. Rather than a straightforward concatenation, the Self-Attention Transformer Block captures more advanced representations of the combined data. To maintain accurate layer boundaries, a novel Boundary Preservation Loss (BPL) function is introduced, allowing the incorporation of high-precision layer edges into a classification loss function. The integration of texture awareness and the focused preservation of edges imparts robustness to the network against noise. This approach achieves size-independence and resilient segmentation performance, even in the presence of image artifacts and pathologies. The Figure 1 presents a comparison between NDD-SEG and a conventional UNet approach for retinal layer segmentation in the presence of image artifacts and pathologies[67].

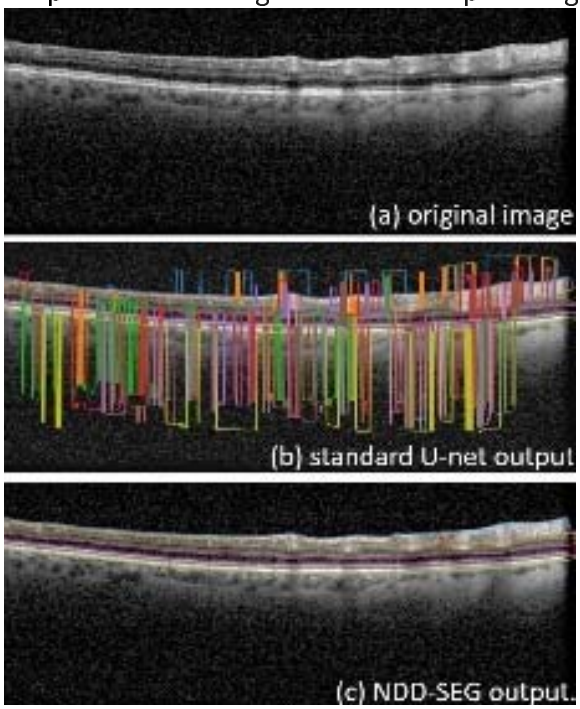


Figure 1: A comparison between NDD-SEG and a standard UNet approach for retinal layer segmentation [67]

2) Extraction of thickness and surface of retinal layers

A noticeable change in the thickness of inner retinal layers, specifically the pRNFL and the combined macular ganglion cell and inner plexiform layers (mGCIPL), is evident when comparing individuals with MS to HCs[68]. Consequently, the assessment of inner retinal layer thicknesses serves as a diagnostic metric for distinguishing individuals with MS from HCs.

While changes in thickness can distinguish areas showing signs of retinal disease from normal regions, the use of texture descriptors can provide additional insights into disease progression. To identify neurodegenerative alterations using OCT image segmentation, the investigation focuses on variations in texture descriptors and optical characteristics of retinal tissue layers in patients with MS.[69]. Research indicates the potency of texture features in diagnosing MS cases[70]. Changes in retinal texture can result in modifications to the boundaries of the retina. As a result, measuring surface across different retinal layers offers an assessment method for differentiating individuals with MS from those unaffected.

3) Quality Enhancement of the images

To enhance the quality of the images while simultaneously reducing noise, an inpainting algorithm is employed. Inpainting approach is used for quality enhancement and denoising in the data preparation step. In the inpainting algorithm, a highpass filter is initially applied to the input image, followed by the application of thresholding to generate a mask. Subsequently, Exemplar-based inpainting is employed to fill in the missing or corrupted regions of the image based on the mask.

The exemplar-based image inpainting approach utilizes a block matching algorithm to identify the most appropriate patch within the image that matches the target region. Once the optimal matching patch is identified, the pixel values from the source region are copied and pasted into the target region, effectively filling in the missing or omitted areas. The quality of the inpainted result is significantly influenced by the search for a matching patch in this technique. By selecting a patch that closely mirrors the content and structure of the missing region, the exemplar-based method aims to seamlessly integrate the inpainted region with the rest of the image[71].

4) Feed images to network

In this study, we examine channel-wise combination and mosaicing of multiple inputs to find the better merging model. These combinations are shown in the final section of Figure 2.

a) Channel-wise combination

To feed data into classifier models, we employ three thickness maps representing different retinal layers or three boundaries of retinal layers across all B-scans as individual channels in 3D color images. By exploring combinations of these retinal layers, our objective is to pinpoint the most effective discriminator for distinguishing between individuals with MS and HCs.

b) Mosaicing of multiple inputs

Another approach to identify the optimal merging model entails constructing a mosaic by combining three thickness maps or boundaries of retinal layers along all B-scans. Subsequently, a grayscale image is produced with the same width and three times the length, allowing additional analysis and comparison to determine the superior merging model.

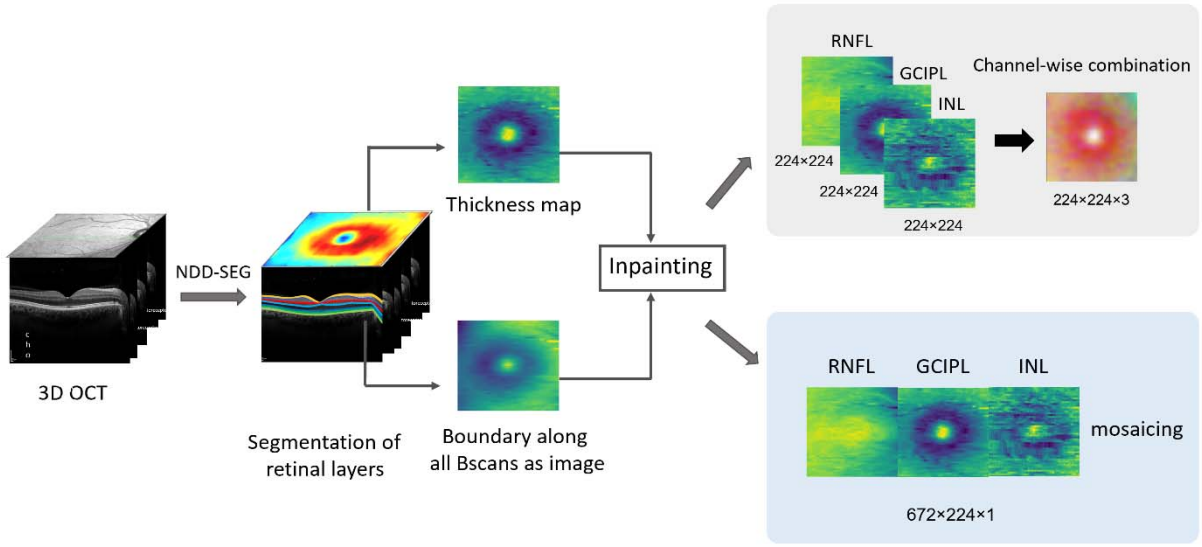


Figure 2: Input Data preparation steps

C. Classification methodology

In this research, we investigate a comprehensive range of AI models including (1) feature extraction using AE and shallow network for classification, (2) classification employing deep networks designed from scratch, and (3) fine-tuning of pretrained networks for classification.

1) feature extraction with autoencoder (AE) and shallow network for classification

The architecture of the AE is designed to identify a low-dimensional representation for input images. An AE integrates an encoder function, responsible for altering the representation of input data, with a decoder function, which transforms the new representation back to the original format. The encoder compresses high-dimensional images into lower-dimensional displays, commonly referred to as the latent space or bottleneck representation. The decoder returns the data to its original dimensions. The latent space contains sufficient information to represent the data correctly. AEs are trained to preserve as much information as possible when input is sent through the encoder and subsequently the decoder[72]. Following the training of the AE and the reduction of features to the latent space, a shallow neural network is employed to classify the low-dimensional latent space.

When developing an AE, there are multiple possibilities to explore. The depth of an AE, often termed as the number of hidden layers in both the encoder and decoder networks, plays a

crucial role. A deeper AE has the potential to learn more intricate representations, but this also increases the model's capacity and computational requirements. The optimal depth is determined by the specific task, the complexity of the data, or a combination of both factors. To achieve the right balance between model complexity and performance, various depths are examined. The compressed representation of the input data is expressed through the latent layer of an AE, and the dimensions of this representation are contingent on the number of nodes in the latent layer. To effectively capture the essential information in the data without losing crucial details, the latent layer must have the appropriate number of nodes, a decision often influenced by the data's complexity and the desired level of compression. While having too many nodes can result in overfitting or ineffective representation, too few nodes can result in information loss. In our classification framework, exploring the number of nodes in the shallow network is an additional option to be considered.

When investigating these options, hyperparameters are tuned and the model is choseb using the grid search method[73]. This method enables the exploration of various options to identify the optimal configuration.

2) classification with deep network designed from scratch

In this study, a CNN model is utilized to achieve effective performance in the classification of OCT images. The network incorporates convolutional layers with Rectified Linear Unit (ReLU) activation functions, pooling, batch normalization, dropout, and fully connected (FC) layers. The last FC layer utilizes a Softmax activation function. Filters with size of 5, 4, and 3 are applied in various layers throughout the network. The nested-cross validation method is applied to select features and tune hyperparameters in CNN through the grid search method. With this approach, the classification model is chosen based on the outer fold that gives the maximum inner-fold accuracy[74]. The Adam optimizer and binary cross-entropy loss function are employed during training. The network architecture is adjusted throughout the training process to optimize performance. Due to the imbalanced dataset where the number of MS patients is more than HCs, class weight is incorporated during the model training processes.

3) fine-tuning of pretrained Resnet152v2 network

Transfer learning (TL) in the context of CNNs is based on the premise that knowledge can be transferred at the parametric level. In this approach, pre-trained CNN models exploit the parameters of the convolutional layers to address a new task within the medical domain. Specifically, in TL with CNN for medical image classification, the task of classifying medical images can be accomplished by utilizing the generic features learned from a source task of natural image classification, where labels are available in both domains. There exist two approaches in TL for leveraging CNN models: the feature extractor approach and the fine-tuning approach. In the feature extractor approach, the convolutional layers are frozen, meaning their parameters are not updated during the model fitting process. Conversely, in the fine-tuning approach, the parameters of the convolutional layers are updated and adjusted during the model fitting process to adapt to the new task[75].

Numerous studies have consistently indicated that deeper models, such as ResNet (Residual Network) [76], exhibit superior performance compared to shallower models like VGG[75]. In the present study, we specifically investigated the performance of VGG16 and ResNet152V2 models, both pre-trained on the ImageNet dataset, through a process known as fine-tuning. Notably, ResNet152V2 demonstrated the most exceptional performance among the models evaluated, and therefore, its results are reported herein. Throughout the training process, the network architecture, including the number of nodes in the dense layers, was fine-tuned in order to attain optimal performance.

Figure 3 illustrates the AI models examined in the current study, encompassing the three approaches elucidated above.

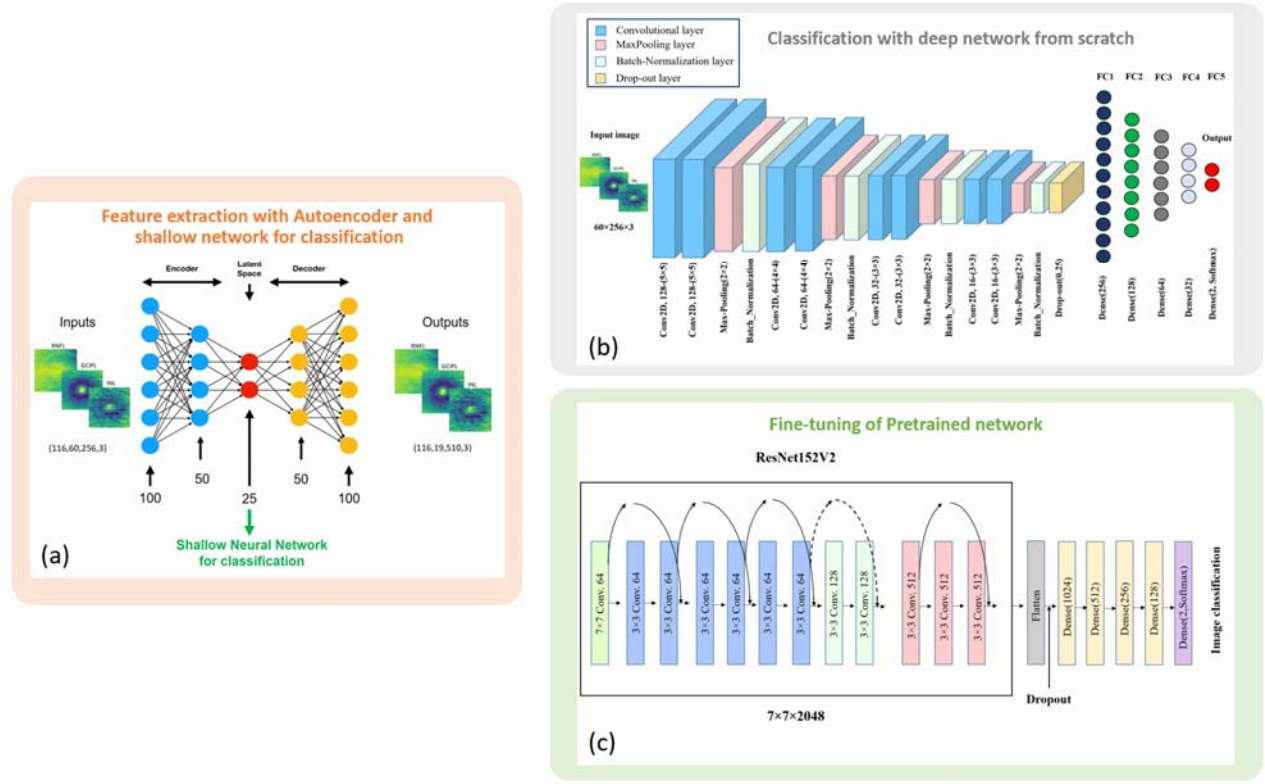


Figure 3: AI models for classification. (a) Feature extraction with autoencoder and shallow neural network for classification. (b) classification with deep network from scratch. (c) Fine-tuning of pretrained ReNet152v2 network

D. Data Augmentation:

1) Data augmentation for the first model

After training the AE and reducing image features to the latent space, data augmentation is employed as a strategy to mitigate the risk of overfitting. Two widely utilized techniques for data augmentation, extrapolation and interpolation, are employed. Interpolation is the process of creating new data points within the range of previously collected data. It is frequently employed to generate more samples by interpolating between nearby data points. In contrast, extrapolation involves expanding the dataset beyond the current set of data points. This

method is used to create samples outside the original data distribution by estimating or forecasting values based on patterns or trends that exist beyond the observable data. By combining interpolation and extrapolation techniques, data augmentation can effectively create new data instances with variations or characteristics similar to the original data.

2) Image augmentation for the second and third models

Due to the limited size of our dataset, the likelihood of overfitting is heightened. To address this concern, Geometrical transforms such as horizontal and vertical flips, and randomly rotation with 20° , width shifting and height shifting with range of 0.1, zooming with range of 0.1 are implemented. These techniques aim to augment the dataset, thereby reducing the risk of overfitting.

III. Experimental results:

A. Performance measures

Cross-validation, specifically employing a five-fold approach, serves as a means to assess results and prevent overfitting. This methodology guarantees that the ultimate outcomes remain unaffected by the initial data split. Within each fold, the training set undergoes augmentation, and the model is trained with these augmented sets, while the remaining set is reserved for testing the model without any augmentation. This process is repeated five times, altering the test set on each iteration. The final accuracy is determined by calculating the mean accuracy across these five distinct test sets.

In classification, the most widely adopted metric is accuracy, representing the rate of correct classifications. Nonetheless, there are instances, such as dealing with imbalanced datasets, where accuracy may not be the optimal choice. To address this, additional metrics such as G-means, F-measure, recall, and specificity are employed alongside accuracy to effectively handle imbalanced classifications[77].

The assessment of classification performance relies on the reporting of recall, g-mean, and specificity values, serving as quantifiable measures of the predictive capabilities of each model. The metrics are defined as follows:

Specificity: measures the ratio of negatives that are recognized correctly.

$$Specificity = \frac{TN}{TN + FP} \quad (1)$$

Recall: measures the positives ratio that are correctly recognized.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

The quantities of true positives, false positives, false negatives, and true negatives are denoted as TP, FP, FN, and TN, respectively, with TP + TN + FP + FN = n representing the overall number of observations.

G-means: following recommendations from [78], the geometric mean (G-mean) is determined as the product of prediction accuracies for both classes, namely specificity (accuracy on the negative samples), and sensitivity (accuracy on the positive samples). A low G-mean value may be attributed to inaccurate categorization of the positive samples, even if the negative samples are classified correctly[79].

$$Gmean = \sqrt{recall \times specificity} \quad (3)$$

B. Validation strategy:

The nested cross-validation (nCV) technique begins with an initial division of data into outer folds, encompassing training and test categories. Following this, each outer training category undergoes further division into inner categories dedicated to setting hyperparameters (highlighted by the red box in Figure 3). Employing data augmentation and grid search techniques, the optimal internal training model is identified and subsequently applied to the external test data. The external model parameters, demonstrating a robust correlation with the highest accuracy achieved by the internal model parameters, are then chosen as the best model parameters[74] (Figure 4).

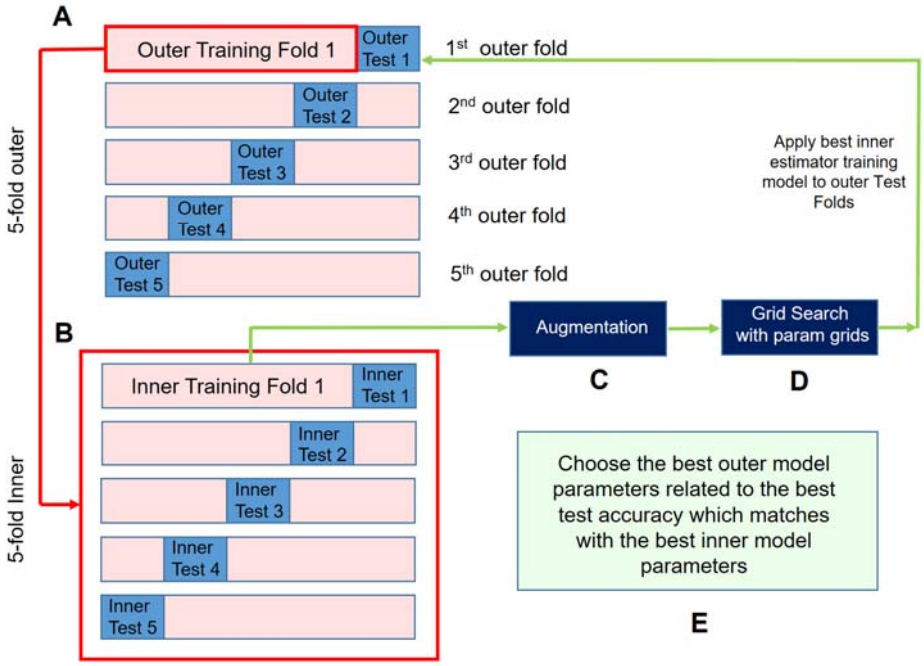


Figure 4: Nested-cross validation. (A) The data splits into outer folds containing train and test data pairs. (B) Outer training fold splits into inner folds. (C) Inner training fold augments to avoid overfitting. (D) Using grid search method for hyperparameter tuning. (E) Choose the best outer model matches with the best inner model parameters.

C. Interpretability:

Despite deep learning showcasing remarkable effectiveness across diverse tasks, particularly in image classification, interpretability has consistently presented a challenge for deep neural networks[80]. In this investigation, we enhance interpretability in deep learning algorithms through the classification process involving both a deep network built from scratch and fine-tuning of a pre-trained network. To address the inherent black box nature limitation of deep neural networks in medical applications, we employ occlusion sensitivity[81] and Gradient-weighted Class Activation Mapping (Grad-CAM)[82] methods.

Both occlusion sensitivity and Grad-CAM enhance the interpretability and clarity of deep learning networks. Occlusion sensitivity involves systematically blocking portions of the image to discern the significance of features, while Grad-CAM visualizes the areas of the image that wield the greatest influence in the model's decision-making. These methodologies offer valuable insights into comprehending the internal mechanisms of deep learning models, aiding in the identification of discriminative regions or features crucial for predictions.

1) Occlusion sensitivity

In this methodology, post-model training, a black mask measuring 40×40 pixels is generated and systematically applied to images within the test set, effectively traversing the entire image. The masked images are then fed into the model, and the resulting test accuracy is computed. A decrease in accuracy is expected if occlusion excludes regions containing crucial discriminative information. Reconstructing the occlusion with the original image size and accuracy values assigned to each pixel's location illustrates interpretability, referred to as the heatmap. An interpretability heatmap delineates the localized impact on classification, highlighting the significance of each area in relation to the respective class.

2) Grad-CAM

Grad-CAM generates a rudimentary localization map that accentuates specific areas in an image for concept prediction. This is achieved by leveraging the gradients of a chosen target concept as they flow into the concluding convolutional layer[82]. Utilizing the gradient information supplied to the final convolutional layer of the CNN, Grad-CAM discerns the significance of each neuron in relation to a specific decision[83].

D. Performance results:

The investigation involved the combination of various retinal layers through the implemented algorithms. In this section, the presented results encompass color images with resolutions of 60×256×3 and 224×224×3 pixels, where the thickness maps of the mRNFL, GCIP, and INL layers are allocated to three respective channels. Additionally, color images with resolutions of 60×256×3 and 224×224×3 pixels, incorporating the placement of the first three boundaries into their corresponding channels, along with grayscale images featuring a resolution of 60×768×1 pixels obtained by horizontally concatenating thickness maps, and grayscale images with a resolution of 60×768×1 pixels acquired through the horizontal concatenation of boundaries, are

showcased. The outcomes of all combinations are detailed in Table 3. Initially, classification models were trained and tested on the complete dataset, comprising 116 OCT images, 38 HC images, and 78 MS images, utilizing 5-fold cross-validation.

Table 3: comparison of different models and structures for classification of MS and HCs.

	Channel-wise/ Mosaic	size	Data type (Boundary/thickness)	Balanced-accuracy	Sensitivity	Specificity	g-mean
AE	Channel-wise	60×256×3	Thickness	88.2%	92.3%	84.2%	88.1%
	Channel-wise	224×224×3	Thickness	82.9%	94.8%	71.0%	82.1%
	Channel-wise	60×256×3	Boundary	78.3%	83.6%	63.1%	76.8%
	Mosaic	60×768×1	Thickness	84.9%	88.7%	73.6%	84.1%
	Mosaic	60×768×1	Boundary	79.6%	84.4%	65.7%	78.4%
From Scratch	Channel-wise	60×256×3	Thickness	**97.3%**	97.4%	**97.3%**	**97.3%**
	Channel-wise	224×224×3	Thickness	94.1%	98.7%	92.1%	94.1%
	Channel-wise	60×256×3	Boundary	88.8%	98.7%	78.9%	88.2%
	Mosaic	60×768×1	Thickness	93.4%	97.4%	89.4%	93.3%
	Mosaic	60×768×1	Boundary	94.7%	95.6%	92.1%	94.7%
Fine-tune	Channel-wise	224×224×3	Thickness	94.0%	96.1%	89.4%	93.9%
	Channel-wise	224×224×3	Boundary	91.5%	92.2%	89.4%	91.5%

To evaluate the models' generalization capabilities, the classifiers that underwent training the public dataset, consisting of 14 HC and 18 MS samples, were subjected to testing on the local dataset, which includes 24 HC and 60 MS samples. Moreover, the local dataset was utilized for training purposes, while the public dataset was employed for testing. The corresponding outcomes are detailed in Table 4.

Table 4. generalizability of the methods by training the on the public dataset and testing on the local dataset and vice versa

	Train data	Test data	Balanced-accuracy	Sensitivity	Specificity	g-mean
AE	Public	Local	84.5%	83.3%	85.7%	84.5%
	Local	Public	79.3%	94.4%	64.2%	77.9%
From scratch	Public	Local	78.9%	83.3%	85.7%	78.6%
	Local	Public	80.5%	78.1%	100%	78.1%
Fine-tune	Public	Local	66.2%	70.0%	62.5%	66.1%
	Local	Public	66.6%	83.3%	50%	64.5%

Visual interpretability

The visual interpretability is depicted through the presentation of occlusion sensitivity and Grad-CAM heatmaps, as shown in Figure 5. To calculate the overlap of the heatmap and thickness maps, the formula heatmap × α + thickness map is utilized, with α set to 0.2.

For the proposed CNN network from scratch and fine-tuning with ResNet152V2, the visual interpretability is showcased through heatmap visualization methods, such as occlusion sensitivity and Grad-CAM. This is demonstrated in both channel-wise combination and the mosaic arrangement of images as input.

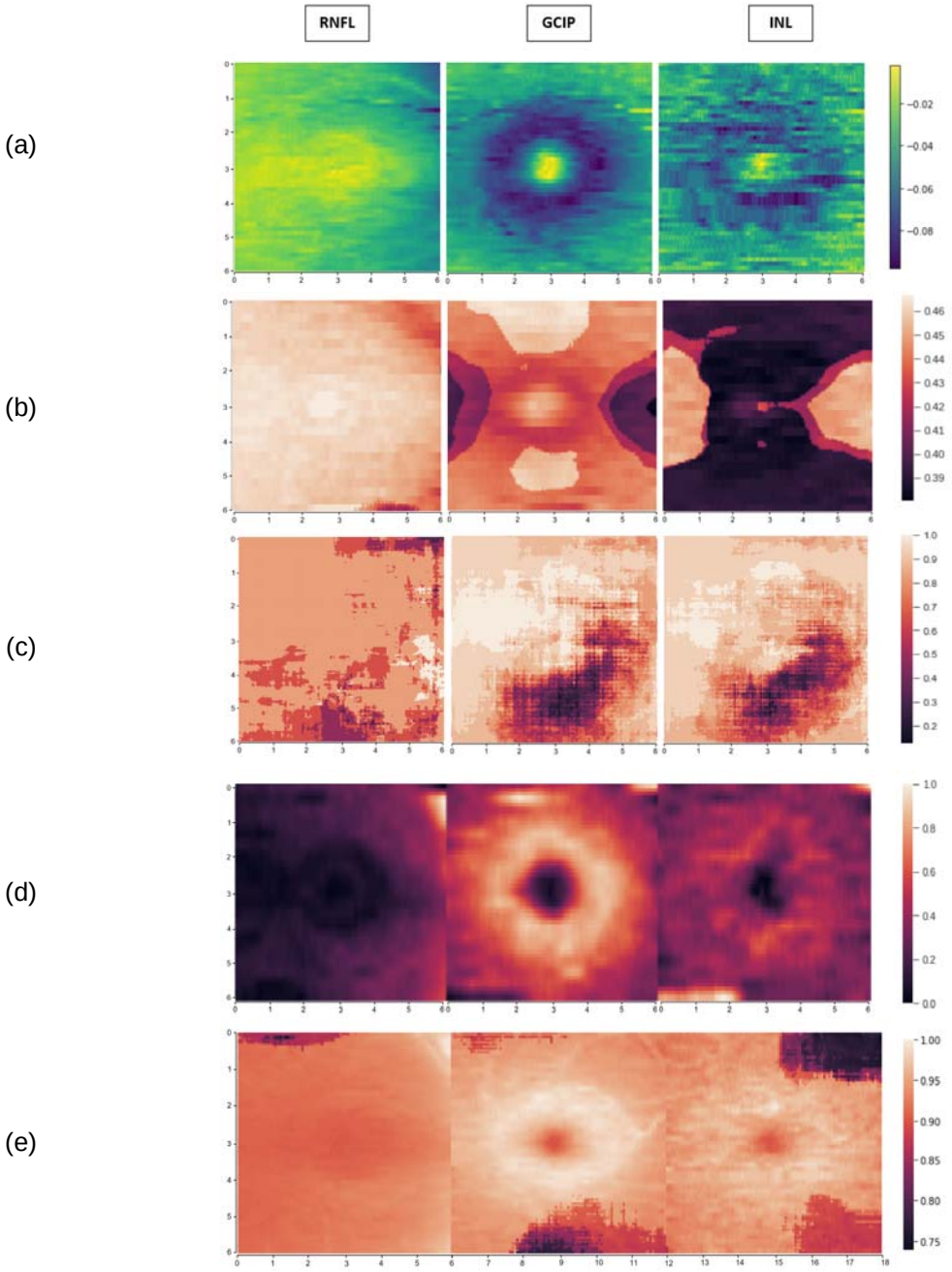


Figure 5: Visual Interpretability on Thickness Maps of mRNFL, GCIP, and INL Using Occlusion Sensitivity and Grad-CAM Methods. (a) Thickness maps in a sample from the MS dataset (x and y axes in millimeters). (b) Occlusion sensitivity heat map in MS and HC classification with the proposed CNN model trained from scratch and channel-wise combination of thickness maps. (c) Occlusion sensitivity heat map using the ResNet152V2 model and channel-wise combination of thickness maps. (d) Grad-CAM heat map

with the proposed CNN model trained from scratch and mosaic of thickness maps. (e) Occlusion sensitivity heat map with the proposed CNN model trained from scratch and mosaic of thickness maps

IV. Discussion

In this work, we have explored a diverse array of AI models, encompassing (1) utilizing AE for feature extraction and shallow network for classification, (2) developing deep network from scratch, and (3) fine-tuning of pre-trained network for the purpose of classifying retinal OCT scans for discrimination of MS.

We additionally explored various input data, encompassing the thickness and surface of diverse retinal layers, in order to identify the most representative information for distinguishing MS. Furthermore, we examined the combination of multiple inputs through channel-wise combinations and mosaicing to determine an improved merging model.

The optimal topology for classification, employing the suggested deep network developed from scratch, was identified when the inputs comprised a channel-wise combination of the thicknesses of the mRNFL, GCIP, and INL retinal layers.

The reason for this outcome can be expressed that by combining the channels of images, the interconnection between the channels is established within the network. Given that neural-cellular, the layers are not independent of each other, by merging the channels of images, we intend for the network to establish connections between the layers, resulting in a better outcome.

As regards the analysis, it was observed that the inner retinal layers (mRNFL, GCIP, and INL) in the macular region exhibited the most significant impact, thereby demonstrating the highest capability to differentiate between control subjects and patients.

Utilizing a size of 60×256×3 among the employed channel wise combination combination, the proposed deep network developed from scratch yielded a balanced accuracy of 97.3%, specificity of 97.35%, recall of 97.4%, and a g-mean of 97.3% in discriminating between MS and HC OCTs. The execution time of an experiment for this model using Google Colaboratory, was calculated as 0.008 ± 0.007 seconds. This value was obtained by running the experiment five times, calculating the mean and standard deviation.

To prevent overfitting, data augmentation was employed in all networks. Moreover, patient-wise cross-validation was implemented in the training and testing datasets to avoid the leakage of testing data information into the training dataset. The study evaluated the generalizability of the proposed CNN model by training it on a local dataset and subsequently testing it on an independent Public dataset obtained from a new device in another country, and vice versa. The interpretability of the models was demonstrated through two approaches, namely occlusion sensitivity and Grad-CAM, to illustrate the contribution of regional layers to the classification performance. The heat maps generated by GCIP and INL revealed that the temporal and central regions had a greater impact on the classification of MS and HCs, as depicted in Figure 5 that

was also achieved using machine learning algorithms in our previous study[36]. Therefore, the information from GCIP and INL layers in segregating MS and HC data, consistent with prior studies [22,24], was observed in interpretable outcomes of deep learning-based networks as well. Moreover, the thickness of the mRNFL was found to have a lesser impact on distinguishing MS disease when compared to GCIP.

Grad-CAM and occlusion sensitivity methods are common techniques for interpreting deep learning models. Despite sharing the same goal, they serve different purposes. Grad-CAM identifies crucial areas in an image by computing gradients of the target class relative to the feature maps in the network. The resulting heat map illustrates which portion of the image significantly influences the model's prediction. On the other hand, the occlusion method systematically obstructs various regions of an input image and observes the impact on the model's output. By assessing the change in the model's prediction when different parts of the image are occluded, this method offers insights into the significance of various image regions.

To address age as a potential confounding variable between the MS and control groups, stochastic age-matching was applied to the test dataset before each model run. For every MS eye, the control eye with the closest age was included in the age-matched test dataset. This age-matching process was iterated five times for each model, and the outcomes were consolidated. This approach led to a well-balanced test dataset with no noteworthy differences in age between the MS and control groups.

The findings indicate that each protocol has its own merits depending on the dataset size. Specifically, for smaller datasets, the combination of feature extraction using AE and a shallow neural network demonstrates superior performance. On the other hand, deep learning-based methods provide interpretability in the results.

Table 6 shows a summary of the previous studies on ML and DL algorithms in distinguishing MS and HCs.

Table 6. Summary of previous studies.

Previous works	Number of datasets	Input retinal layers	Performance metrics	The most discriminant retinal layer	Classification method
Garcia-Martin et al. [29] 2012	115 MS, 115 HC	Peripapillary area	Specificity=95%	pRNFL	Linear Discriminant Function (LDF)
Garcia-Martin et al. [30] 2013	106 MS, 115 HC	Peripapillary area	AUC=0.945	pRNFL	ANN
Garcia-Martin et al. [31] 2015	112 MS, 105 HC	Peripapillary Area	Recall=89.3% Specificity=87.6%	pRNFL	ANN
Palomar et al. [32] 2019	80 MS, 180 HC	Peripapillary, macular and extended (between macula and papilla) areas	Decision tree in the macular area: AUC=0.959 In the extended area: AUC=0.998	pRNFL	Decision tree, ANN, SVM
Cavaliere et al.	48 MS,	Peripapillary and	Recall=89%	GCL++	SVM

[27] 2019	48 HC	macular areas	Specificity=92%	and nasal quadrant of the outer and inner ring in pRNFL	
Garcia-Martin et al. [28] 2021	48 MS, 48 HC	Macular area	Recall=98% Specificity=98%	GCL++	SVM, ANN
Zhang et al. 2020	58 MS, 63 HC	Macular area	Recall=64% Specificity=94%	GCIPL	logistic regression (LR), logistic regression regularized with the elastic net penalty (LR-EN), SVM
Montolio et al. [34] 2021	108 MS, 104 HC	Peripapillary and macular areas	EC: Recall=87% Specificity=88.5% LSTM: Recall=81.1% Specificity=82.2%	pRNFL	MLR), SVM, decision tree, k-NN, NB, EC, LSTM recurrent neural network
López-Dorado et al. [61] 2022	48 MS, 48 HC	Macular area	Specificity = 100% Recall=100%	GCL++, whole retina GCL+ (Cohen distance)	Augmentation with GAN and classification with CNN
Khodabandeh et al. [36] 2023	Train: 106 MS,422 HC (Charite dataset) Test: 67 MS, 45 HC (Isfahan dataset)	Macular area	Precision = 89% Recall =88%	GCIPL and INL	SVM, RF, and ANN
Garcia-Martin et al. [84] 2023	100 MS, 111 HC	Posterior pole grid 8×8 and zone grouping	Recall = 90.90 % Specificity = 95%	RNFL and GCL	Gradient boosting, RF, Explainable boosting machine
Garcia-Martin et al. [85] 2023	79 MS, 69 HC	Inter-eye difference using posterior pole protocol	Recall = 86% Specificity = 90 %	RNFL, GCL, and IPL	SVM- RFE- Leave-One-Out cross validation (SVM-RFE-LOOCV)

Previous studies lacked direct utilization of images as input for models, and there was a scarcity of deep learning methodologies employed in those investigations. The results of this study cannot be directly compared to previous works due to the unavailability of codes and datasets in those studies. It is important to note that this study did not consider the presence of optic neuritis (ON), and therefore, MS patients with and without ON were combined for classification. Consequently, the lower performance compared to studies considering MS with ON is reasonable, as eyes without ON exhibit less thinning and are less distinguishable from healthy controls. Finally, some previous studies incorporated pRNFL data as input for classification, resulting in higher performance compared to the limited focus on the macular region in this study.

This study has several limitations. Firstly, it did not consider the history of ON, a prevalent clinical symptom of MS. Secondly, longitudinal follow-up of patients was not conducted, which could aid in the early diagnosis of the disease. Thirdly, the models were trained using data from a specific device, and it was not evaluated if the models are generalizable to other devices, such as TOPCON or ZEISS. To improve the model's ability to interact with new datasets, additional data from different devices should be included in the training dataset in future studies. We anticipate that addressing these limitations will lead to better results and a clearer demonstration of the generalizability of the proposed model.

References

1. Sreekantha, Vinodchandran, R. Rangaswamy, M. Amareshwara, S. S. Avinash, and Remya, "Multiple sclerosis," Biomedicine **30**(3), 400–401 (2010).

2. Y. Zhao, B. C. Healy, D. Rotstein, C. R. G. Guttmann, R. Bakshi, H. L. Weiner, C. E. Brodley, and T. Chitnis, "Exploration of machine learning techniques in predicting multiple sclerosis disease course," PLoS One **12**(4), e0174866 (2017).

3. P. A. Calabresi, "Diagnosis and management of multiple sclerosis," Am. Fam. Physician **70**(10), 1935–1944 (2004).

4. A. Eshaghi, A. L. Young, P. A. Wijeratne, F. Prados, D. L. Arnold, S. Narayanan, C. R. G. Guttmann, F. Barkhof, D. C. Alexander, and A. J. Thompson, "Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data," Nat. Commun. **12**(1), 1–12 (2021).

5. R. Dobson and G. Giovannoni, "Multiple sclerosis–a review," Eur. J. Neurol. **26**(1), 27–40 (2019).

6. A. Verma, "Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study," Yearb. Neurol. Neurosurg. **2008**(9), 114–115 (2008).

7. G. C. Ebers, "Environmental factors and multiple sclerosis," Lancet Neurol. **7**(3), 268–277 (2008).

8. D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," Med. Image Anal. **17**(1), 1–18 (2013).

9. M. Hartmann, N. Fenton, and R. Dobson, "Current review and next steps for artificial intelligence in multiple sclerosis risk research," Comput. Biol. Med. **132**, 104337 (2021).

10. J. R. Oksenberg, A. B. Begovich, H. A. Erlich, and L. Steinman, "Genetic factors in multiple sclerosis," Jama **270**(19), 2362–2369 (1993).

11. E. Canto and J. R. Oksenberg, "Multiple sclerosis genetics," Mult. Scler. **24**(1), 75–79 (2018).

12. A. Ascherio, "Environmental factors in multiple sclerosis," Expert Rev. Neurother.

13(sup2), 3–9 (2013).

13. R. Milo and E. Kahana, "Multiple sclerosis: geoepidemiology, genetics and the environment," Autoimmun. Rev. **9**(5), A387–A394 (2010).

14. A. J. Thompson, B. L. Banwell, F. Barkhof, W. M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, and M. S. Freedman, "Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria," Lancet Neurol. **17**(2), 162–173 (2018).

15. 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).

16. Y. Fu, T. M. Talavage, and J.-X. Cheng, "New imaging techniques in the diagnosis of multiple sclerosis," Expert Opin. Med. Diagn. **2**(9), 1055–1065 (2008).

17. 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).

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

19. C. A. Wicki, J. V. M. Hanson, and S. Schippling, "Optical coherence tomography as a means to characterize visual pathway involvement in multiple sclerosis," Curr. Opin. Neurol. **31**(5), 662–668 (2018).

20. A. Petzold, J. F. de Boer, S. Schippling, P. Vermersch, R. Kardon, A. Green, P. A. Calabresi, and C. Polman, "Optical coherence tomography in multiple sclerosis: a systematic review and meta-analysis," Lancet Neurol. **9**(9), 921–932 (2010).

21. J. Britze, G. Pihl-Jensen, and J. L. Frederiksen, "Retinal ganglion cell analysis in multiple sclerosis and optic neuritis: a systematic review and meta-analysis," J. Neurol. **264**(9), 1837–1853 (2017).

22. 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).

23. E. Garcia-Martin, J. R. Ara, J. Martin, C. Almarcegui, I. Dolz, E. Vilades, L. Gil-Arribas, F. J. Fernandez, V. Polo, and J. M. Larrosa, "Retinal and optic nerve degeneration in patients with multiple sclerosis followed up for 5 years," Ophthalmology **124**(5), 688–696 (2017).

24. 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).

25. R. Alonso, D. Gonzalez-Moron, and O. Garcea, "Optical coherence tomography as a biomarker of neurodegeneration in multiple sclerosis: A review," Mult. Scler. Relat. Disord. **22**, 77–82 (2018).

26. Y. Tong, W. Lu, Y. Yu, and Y. Shen, "Application of machine learning in ophthalmic imaging modalities," Eye Vis. **7**(1), 1–15 (2020).

27. 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).

28. 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).

29. E. Garcia-Martin, L. E. Pablo, R. Herrero, M. Satue, V. Polo, J. M. Larrosa, J. Martin, and J. Fernandez, "Diagnostic ability of a linear discriminant function for spectral-domain optical coherence tomography in patients with multiple sclerosis," Ophthalmology **119**(8), 1705–1711 (2012).

30. 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).

31. 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.

32. A. Perez del Palomar, J. Cegonino, A. Montolio, E. Orduna, E. Vilades, B. Sebastian, 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).

33. 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).

34. 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).

35. E. De Brouwer, T. Becker, Y. Moreau, E. K. Havrdova, M. Trojano, S. Eichau, S. Ozakbas, M. Onofrj, P. Grammond, and J. Kuhle, "Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression," Comput. Methods Programs Biomed. **208**, 106180 (2021).

36. Z. Khodabandeh, H. Rabbani, F. Ashtari, H. G. Zimmermann, S. Motamedi, A. U. Brandt, F. Paul, and R. Kafieh, "Discrimination of Multiple Sclerosis using multicenter OCT images," Mult. Scler. Relat. Disord. 104846 (2023).

37. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**(7553), 436–444 (2015).

38. F. J. Martinez-Murcia, J. M. Górriz, J. Ramírez, and A. Ortiz, "Convolutional neural networks for neuroimaging in parkinson's disease: is preprocessing needed?," Int. J. Neural Syst. **28**(10), 1850035 (2018).

39. C. Jiménez-Mesa, J. Ramírez, J. Suckling, J. Vöglein, J. Levin, J. M. Górriz, A. D. N. I. ADNI, and D. I. A. N. DIAN, "Deep Learning in current Neuroimaging: a multivariate approach with power and type I error control but arguable generalization ability," arXiv Prepr. arXiv2103.16685 (2021).

40. J. M. Górriz, J. Ramírez, A. Ortíz, F. J. Martínez-Murcia, F. Segovia, J. Suckling, M. Leming, Y.-D. Zhang, J. R. Álvarez-Sánchez, and G. Bologna, "Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications," Neurocomputing **410**, 237–270 (2020).

41. D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," J. Physiol. **160**(1), 106 (1962).

42. 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).

43. M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," Invest. Ophthalmol. Vis. Sci. **57**(13), 5200–5206 (2016).

44. D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, and S. Y. Lee, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images

from multiethnic populations with diabetes," Jama **318**(22), 2211–2223 (2017).

45. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, and J. Cuadros, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," Jama **316**(22), 2402–2410 (2016).

46. Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, "Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs," Ophthalmology **125**(8), 1199–1206 (2018).

47. F. Grassmann, J. Mengelkamp, C. Brandl, S. Harsch, M. E. Zimmermann, B. Linkohr, A. Peters, I. M. Heid, C. Palm, and B. H. F. Weber, "A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography," Ophthalmology **125**(9), 1410–1420 (2018).

48. P. M. Burlina, N. Joshi, M. Pekala, K. D. Pacheco, D. E. Freund, and N. M. Bressler, "Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks," JAMA Ophthalmol. **135**(11), 1170–1176 (2017).

49. J. M. Brown, J. P. Campbell, A. Beers, K. Chang, S. Ostmo, R. V. P. Chan, J. Dy, D. Erdogmus, S. Ioannidis, and J. Kalpathy-Cramer, "Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks," JAMA Ophthalmol. **136**(7), 803–810 (2018).

50. S. Apostolopoulos, C. Ciller, S. De Zanet, S. Wolf, and R. Sznitman, "RetiNet: Automatic AMD identification in OCT volumetric data," Invest. Ophthalmol. Vis. Sci. **58**(8), 387 (2017).

51. C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images," Ophthalmol. Retin. **1**(4), 322–327 (2017).

52. W. Lu, Y. Tong, Y. Yu, Y. Xing, C. Chen, and Y. Shen, "Deep Learning-Based Automated Classification of Multi-Categorical Abnormalities From Optical Coherence Tomography Images," Transl. Vis. Sci. Technol. **7**(6), 41 (2018).

53. O. Perdomo, S. Otalora, F. A. Gonzalez, F. Meriaudeau, and H. Muller, "OCT-NET: A convolutional network for automatic classification of normal and diabetic macular edema using sd-oct volumes," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (IEEE, 2018), pp. 1423–1426.

54. M. Awais, H. Muller, T. B. Tang, and F. Meriaudeau, "Classification of SD-OCT images using a Deep learning approach," in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)* (IEEE, 2017), pp. 489–492.

55. R. Rasti, A. Mehridehnavi, H. Rabbani, and F. Hajizadeh, "Automatic diagnosis of abnormal macula in retinal optical coherence tomography images using wavelet-based

convolutional neural network features and random forests classifier," J. Biomed. Opt. **23**(03), 1 (2018).

56. S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," Biomed. Opt. Express **8**(2), 579 (2017).

57. Q. Ji, W. He, J. Huang, and Y. Sun, "Efficient deep learning-based automated pathology identification in retinal optical coherence tomography images," Algorithms **11**(6), (2018).

58. K. Roy, S. S. Chaudhuri, P. Roy, S. Chatterjee, and S. Banerjee, "Transfer learning coupled convolution neural networks in detecting retinal diseases using OCT images," in *Intelligent Computing: Image Processing Based Applications* (Springer, 2020), pp. 153–173.

59. N. Saleh, M. Abdel Wahed, and A. M. Salaheldin, "Transfer learning-based platform for detecting multi-classification retinal disorders using optical coherence tomography images," Int. J. Imaging Syst. Technol. **32**(3), 740–752 (2022).

60. N. Y. Kang, H. Ra, K. Lee, J. H. Lee, W. K. Lee, and J. Baek, "Classification of pachychoroid on optical coherence tomography using deep learning," Graefe's Arch. Clin. Exp. Ophthalmol. **259**(7), 1803–1809 (2021).

61. A. López-Dorado, M. Ortiz, M. Satue, M. J. Rodrigo, R. Barea, E. M. Sánchez-Morla, C. Cavaliere, J. M. Rodríguez-Ascariz, L. Boquete, and E. Garcia-Martin, "Early diagnosis of multiple sclerosis using swept-source optical coherence tomography and convolutional neural networks trained with data augmentation," Sensors **22**(1), 167 (2022).

62. M. Ortiz, V. Mallen, L. Boquete, E. M. Sánchez-Morla, B. Cordón, E. Vilades, F. J. Dongil-Moreno, J. M. Miguel-Jiménez, and E. Garcia-Martin, "Diagnosis of multiple sclerosis using optical coherence tomography supported by artificial intelligence," Mult. Scler. Relat. Disord. **74**, 104725 (2023).

63. Y. He, A. Carass, S. D. Solomon, S. Saidha, P. A. Calabresi, and J. L. Prince, "Retinal layer parcellation of optical coherence tomography images: Data resource for multiple sclerosis and healthy controls," Data Br. **22**, 601–604 (2019).

64. 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).

65. 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).

66. 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).

67. R. Kafieh, S. Rakhshani, J. Hogg, R. A. Lawson, N. Pavese, W. Innes, D. H. Steel, J. Bacardit, J. Read, and A. Hurlbert, "A robust, flexible retinal segmentation algorithm designed to handle neuro-degenerative disease pathology (NDD-SEG)," Invest. Ophthalmol. Vis. Sci. **63**(7), 2080-F0069 (2022).

68. 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).

69. B. E. Varga, W. Gao, K. L. Laurik, E. Tátrai, M. Simó, G. M. Somfai, and D. Cabrera DeBuc, "Investigating tissue optical properties and texture descriptors of the retina in patients with multiple sclerosis," PLoS One **10**(11), e0143711 (2015).

70. H. D. Tazarjani, Z. Amini, R. Kafieh, F. Ashtari, and E. Sadeghi, "Retinal OCT Texture Analysis for Differentiating Healthy Controls from Multiple Sclerosis (MS) with/without Optic Neuritis," Biomed Res. Int. **2021**, (2021).

71. M. Shroff and M. S. R. Bombaywala, "A qualitative study of exemplar based image inpainting," SN Appl. Sci. **1**, 1–8 (2019).

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

73. 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).

74. S. Parvandeh, H.-W. Yeh, M. P. Paulus, and B. A. McKinney, "Consensus features nested cross-validation," Bioinformatics **36**(10), 3093–3098 (2020).

75. H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," BMC Med. Imaging **22**(1), 69 (2022).

76. K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14* (Springer, 2016), pp. 630–645.

77. V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," Inf. Sci. (Ny). **250**, 113–141 (2013).

78. M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Icml* (Nashville, USA, 1997), **97**(1), p. 179.

79. S. Hido, H. Kashima, and Y. Takahashi, "Roughly balanced bagging for imbalanced data," Stat. Anal. Data Min. ASA Data Sci. J. **2**(5-6), 412–426 (2009).

80. Q. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: a survey," Front. Inf. Technol. Electron. Eng. **19**(1), 27–39 (2018).

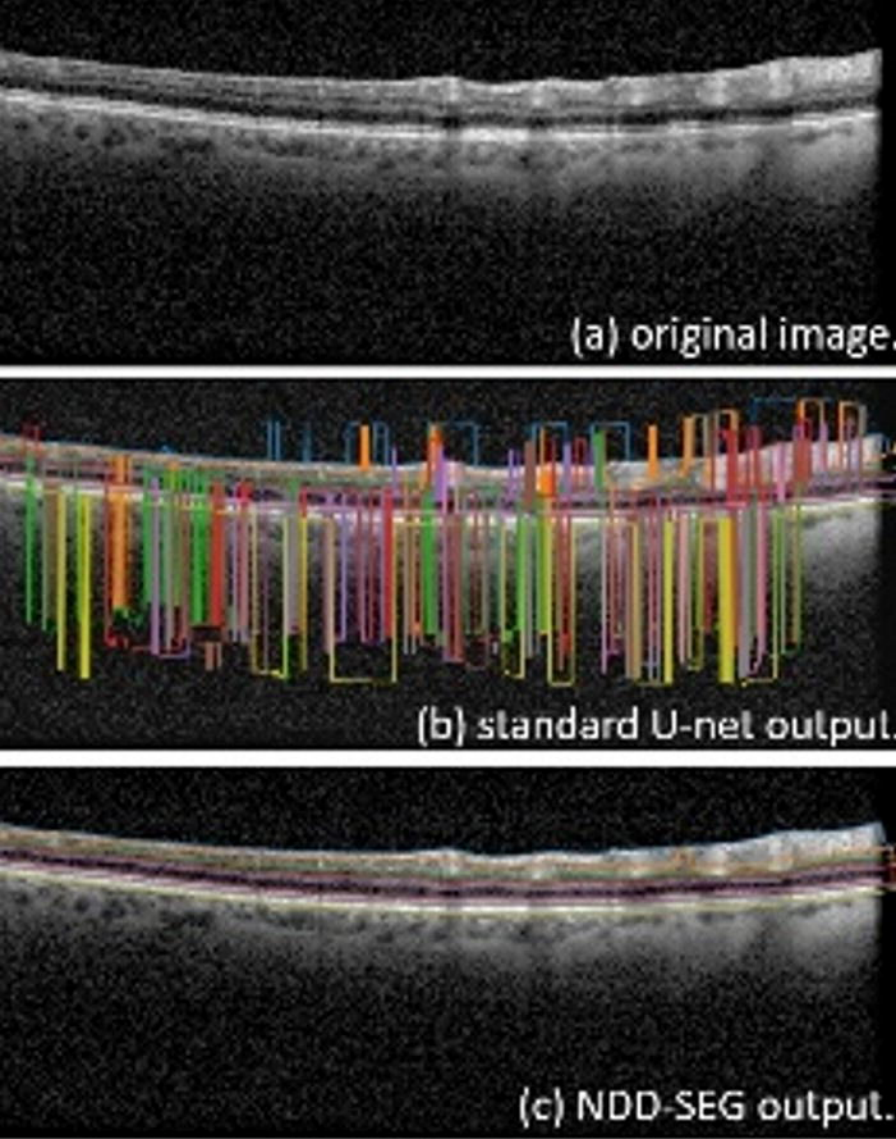
81. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Springer, 2014), **8689 LNCS**(PART 1), pp. 818–833.

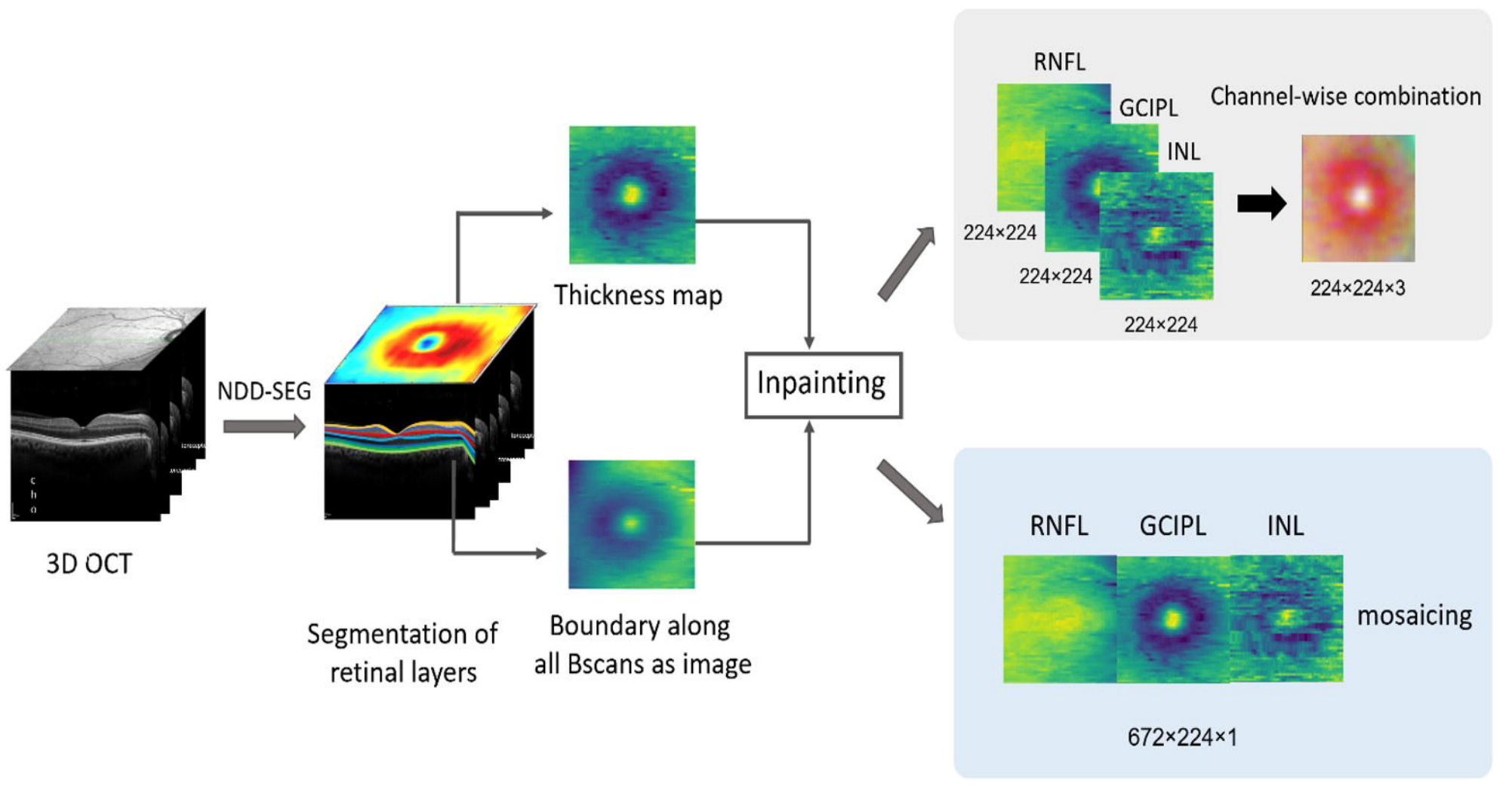
82. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 618–626.

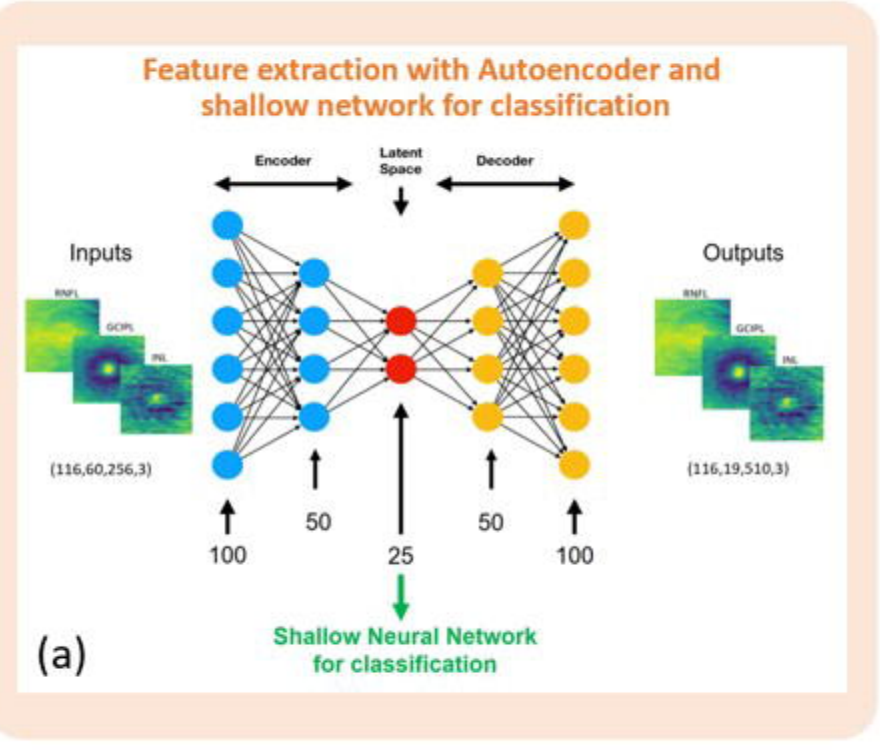
83. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, 2018), pp. 839–847.

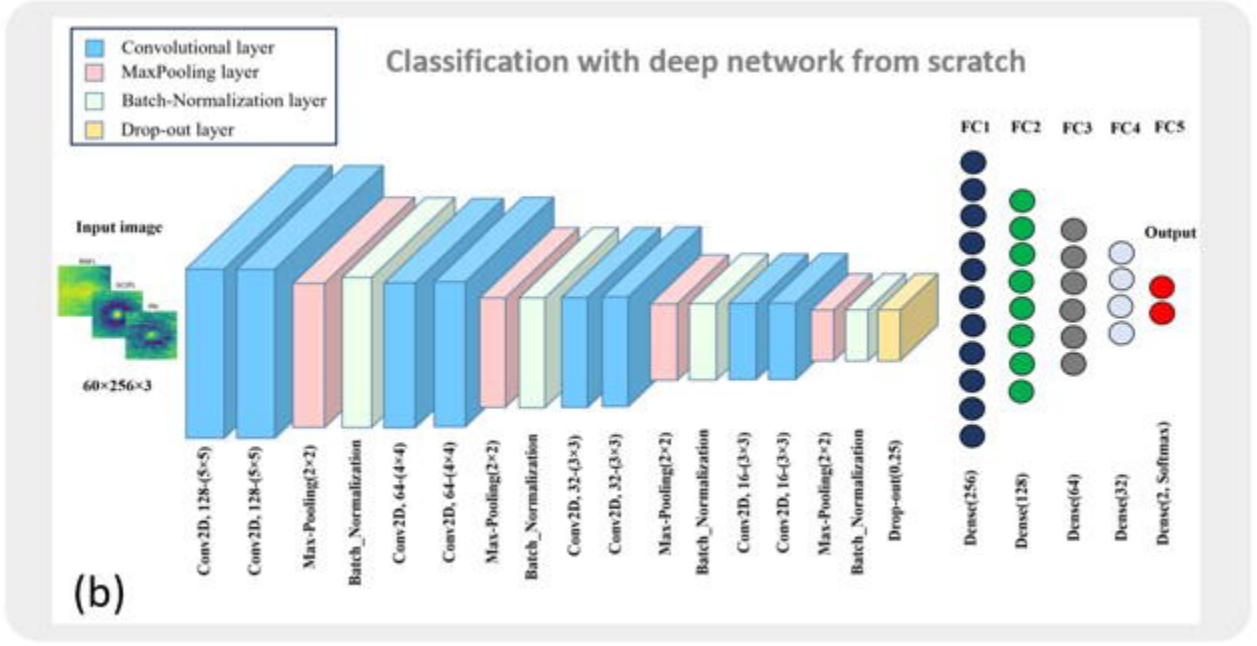
84. M. Hernandez, U. Ramon-Julvez, E. Vilades, B. Cordon, E. Mayordomo, and E. Garcia-Martin, "Explainable artificial intelligence toward usable and trustworthy computer-aided early diagnosis of multiple sclerosis from Optical Coherence Tomography," arXiv Prepr. arXiv2302.06613 (2023).

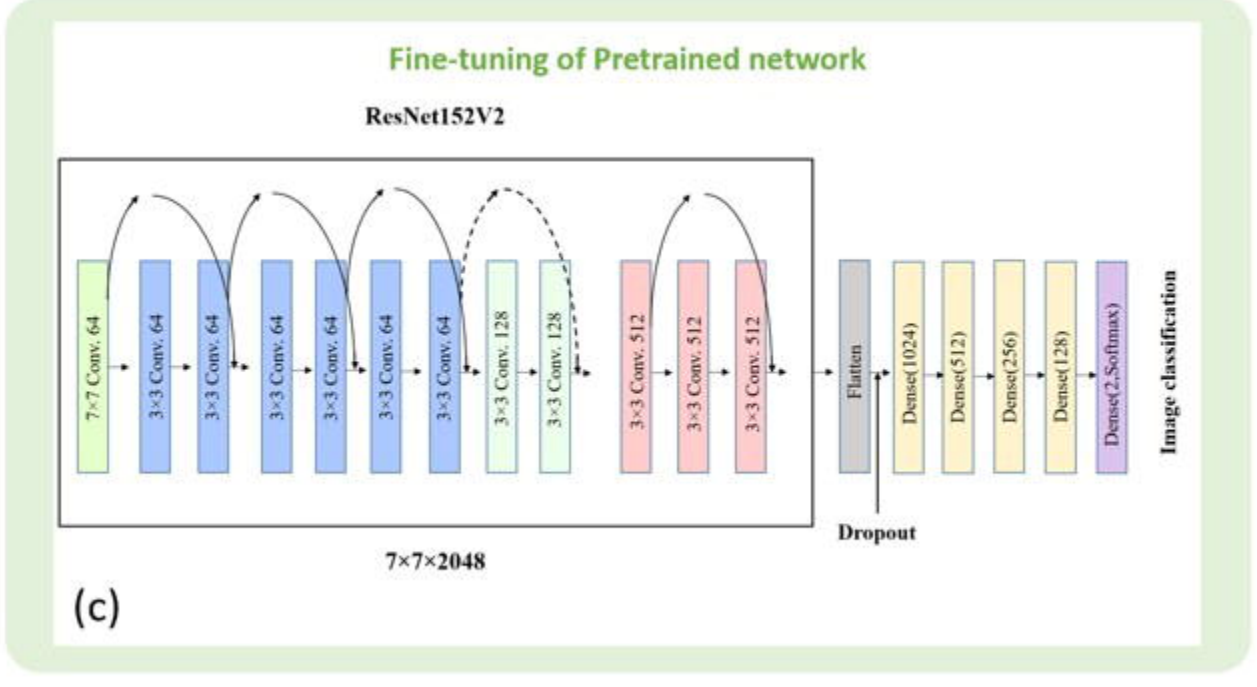
85. M. Ortiz, V. Mallen, L. Boquete, E. M. Sánchez-Morla, B. Cordón, E. Vilades, F. J. Dongil-Moreno, J. M. Miguel-Jiménez, and E. Garcia-Martin, "Diagnosis of multiple sclerosis using optical coherence tomography supported by artificial intelligence," Mult. Scler. Relat. Disord. **74**, (2023).

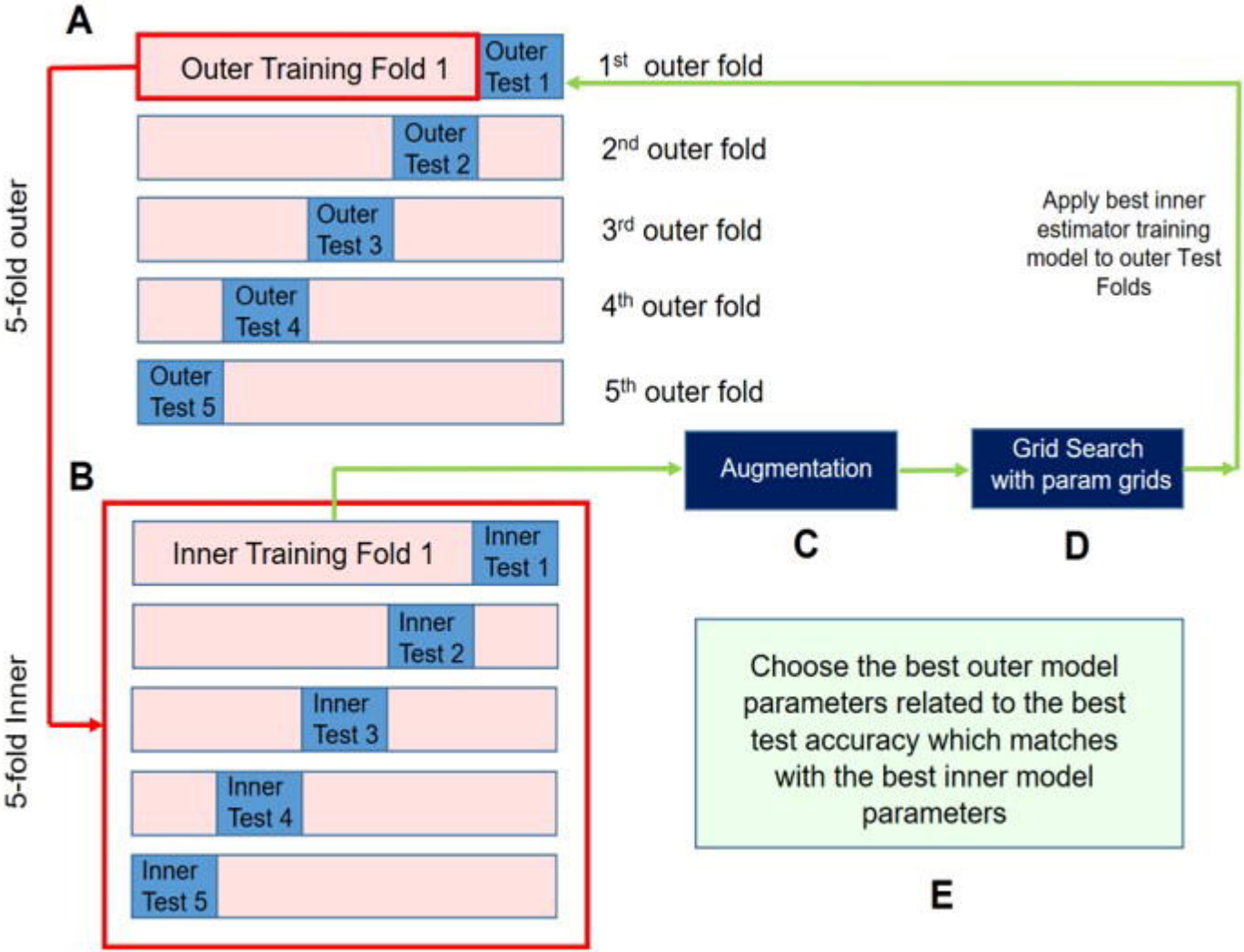

(a) original image.
(b) standard U-net output.
(c) NDD-SEG output.

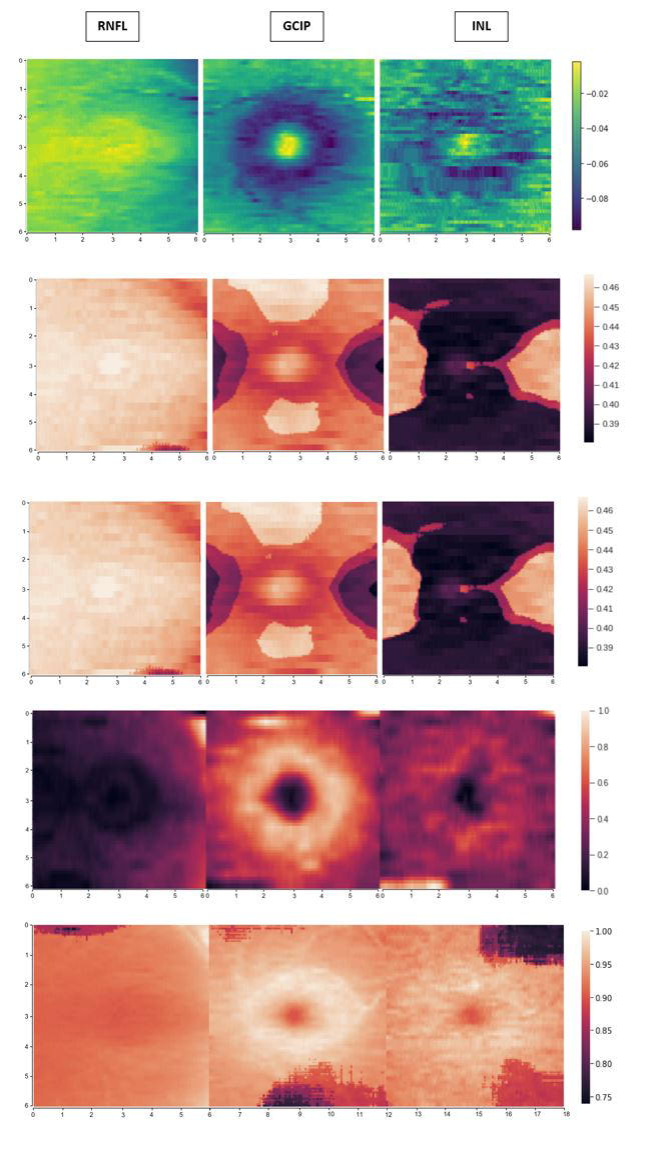

RNFL
GCIPL
INL
Channel-wise combination
224×224
224×224
224×224
224×224×3
Thickness map
NDD-SEG
Inpainting
3D OCT
Segmentation of
retinal layers
Boundary along
all Bscans as image
RNFL
GCIPL
INL
mosaicing
672×224×1


Feature extraction with Autoencoder and shallow network for classification
Encoder
Latent Space
Decoder
Inputs
Outputs
RNFL
GCIPL
INL
(116,60,256,3)
(116,19,510,3)
100
50
25
50
100
Shallow Neural Network for classification
(a)


Convolutional layer
MaxPooling layer
Batch-Normalization layer
Drop-out layer
Classification with deep network from scratch
Input image
60×256×3
Conv2D, 128-(5×5)
Conv2D, 128-(5×5)
Max-Pooling(2×2)
Batch_Normalization
Conv2D, 64-(4×4)
Conv2D, 64-(4×4)
Max-Pooling(2×2)
Batch_Normalization
Conv2D, 32-(3×3)
Conv2D, 32-(3×3)
Max-Pooling(2×2)
Batch_Normalization
Conv2D, 16-(3×3)
Conv2D, 16-(3×3)
Max-Pooling(2×2)
Batch_Normalization
Drop-out(0.25)
FC1
FC2
FC3
FC4
FC5
Output
Dense(256)
Dense(128)
Dense(64)
Dense(32)
Dense(2, Softmax)
(b)


Fine-tuning of Pretrained network
ResNet152V2
7×7 Conv. 64
3×3 Conv. 64
3×3 Conv. 128
3×3 Conv. 512
Flatten
Dropout
Dense(1024)
Dense(512)
Dense(256)
Dense(128)
Dense(2,Softmax)
Image classification
7×7×2048
(c)


A
Outer Training Fold 1
Outer Test 1
1st outer fold
Outer Test 2
2nd outer fold
Outer Test 3
3rd outer fold
Outer Test 4
4th outer fold
Outer Test 5
5th outer fold
5-fold outer
Apply best inner estimator training model to outer Test Folds
B
Augmentation
Grid Search with param grids
C
D
Inner Training Fold 1
Inner Test 1
Inner Test 2
Inner Test 3
Inner Test 4
Inner Test 5
5-fold Inner
Choose the best outer model parameters related to the best test accuracy which matches with the best inner model parameters
E


RNFL
GCIP
INL