IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 6, 2021, accepted April 21, 2021, date of publication April 29, 2021, date of current version May 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3076427

Ensemble Learning Approach to Retinal Thickness Assessment in Optical Coherence Tomography

ALEX CAZAÑAS-GORDÓN, ESTHER PARRA-MORA, AND LUÍS A. DA SILVA CRUZ, (Senior Member, IEEE)
Department of Electrical and Computer Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
Instituto de Telecomunicações, University of Coimbra, 3030-290 Coimbra, Portugal

Corresponding authors: Alex Cazañas-Gordón (acazanas@deec.uc.pt) and Luís A. Da Silva Cruz (lcruz@deec.uc.pt)

This work was supported in part by the Secretariat of Higher Education, Science, Technology and Innovation of the Republic of Ecuador, and in part by the Portuguese research funding agency Fundação para a Ciência e a Tecnologia (FCT) under Project UIDB/EEA/50008/2020.

ABSTRACT Manual assessment of the retinal thickness in optical coherence tomography images is a time-consuming task, prone to error and inter-observer variability. The wide variability of the retinal appearance makes the automation of retinal image processing a challenging problem to tackle. The difficulty is even more accentuated in practice when the retinal tissue exhibits large structural changes due to disruptive pathology. In this work, we propose an ensemble-learning-based method for the automated segmentation of retinal boundaries in optical coherence tomography images that is robust to retinal abnormalities. The segmentation accuracy of the proposed algorithm was evaluated on two publicly available datasets that included cases of severe retinal edema. Moreover, the performance of the proposed method was compared to two existing methods, widely referenced in the relevant literature. The proposed algorithm outperformed reference methods at segmenting the retinal boundaries in both normal and pathological images. Furthermore, a thorough reliability analysis showed a strong agreement between the retinal thickness measurements derived from the segmentation obtained with the proposed method and corresponding manual measurements computed with the manual annotations.

INDEX TERMS Deep learning, ensemble learning, semantic segmentation, image processing, retinal thickness, optical coherence tomography.

I. INTRODUCTION

Optical coherence tomography (OCT) is a non-invasive imaging technology widely used in clinical practice to diagnose retinal pathology [1], [2]. The technology allows visualizing the internal structure of the retina by acquiring high-resolution cross-sectional images of the back of the eye. Retinal OCT scans are extensively used in the monitoring of sight-threatening diseases such as age-related macular degeneration (AMD), retinal vein occlusion (RVO), diabetic macular edema (DME), and glaucoma [3], [4]. Measurements derived from the analysis of OCT images are pivotal for the evaluation of disease progression and treatment effectiveness [5]. Retinal thickness and central macular thickness (CMT) are two of such measurements that are highly regarded as markers in the progression of various ocular diseases [6].

The quantitative analysis of the retinal thickness involves segmenting the extent of the retina from other anatomical structures in the OCT scans. Commonly, OCT scanners include image processing tools that provide reasonably accurate segmentation of the retinal boundaries on healthy and minimal distorted retinas [7]. However, recent studies found that these tools perform poorly in the presence of degenerative diseases such as AMD [3], [8]. Similarly, other studies reported that segmentation errors frequently occur in the presence of disruptive pathologies like macular edema, and retinal detachment [9], [10].

The reliability of the retinal-thickness assessment depends largely on the accurate segmentation of the inner and outer retinal boundaries. Segmentation errors render the retinal-boundary delineation and derived measurements unreliable. Depending on the extent of the segmentation error, the course of action might involve manual correction which besides being impractical in clinical practice is labor-intensive and prone to inter-observer variability [11]. Therefore, there is an unmet demand for fully automated retinal segmentation

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

methods that are robust to the presence of severe pathology which has been reported as the cause of segmentation error in commercially available OCT scans.

A. RELATED WORK

Much effort has been dedicated to developing techniques to segment retinal structures. In particular, work focused on OCT retinal images produced a wide range of methods in previous years [12].

Earlier methods were based on classical image processing techniques, including sparse higher-order potentials [13], diffusion maps [14], variational methods [15], kernel regression classification [16], and graph-theory based methods [17]–[19]. A common characteristic of earlier methods is their reliance on predefined rules, which helped algorithms to fit curves directly to the data. Although these rules proved to be effective in enhancing the segmentation performance, they also limited the flexibility of the algorithms. As a result, these methods performed well in normal and mildly damaged retinas, but poorly in the presence of severe pathology [20].

Later work addressed this limitation by building upon machine learning methods. In contrast to earlier methods, machine learning approaches segmentation as a classification problem. Machine learning methods in conjunction with graph-search algorithms have been demonstrated to improve the segmentation of retinal layers in the presence of retinal abnormalities [21], [22]. Similarly, machine-learning classifiers such as support vector machine [23], [24], random forest classifiers [25] or neural networks [26], [27] have shown good performance in normal retinas.

More recently, the outstanding performance of deep learning in natural image classification has motivated the application of deep neural networks to retinal image analysis. Deep learning algorithms, particularly convolutional neural networks (CNN), have been increasingly applied to retinal segmentation in OCT images [28]–[31]. Methods that combine CNN with graph-search algorithms are a common approach to improve the segmentation performance [32]–[34]. Other deep learning approaches to retinal segmentation include recurrent neural networks [35], and fully convolutional networks [36], [37].

B. ENSEMBLE LEARNING

Deep neural networks (DNN) are at the core of deep learning. These mathematical models ordinarily have a large number of parameters, which are tuned to optimize an objective function — minimize the prediction error. The sheer number of adjustable parameters in DNN architectures makes these algorithms highly effective at learning non-linear, complex relationships in the data, but at the same time renders them prone to overfitting [38].

Overfitting causes distinct models of the same DNN to perform inconsistently on the test data. This is more formally referred to as high variance. An alternative to reduce variance is to combine the prediction of multiple models. This approach is known as Ensemble learning and relies on the observation that is unlikely that different models make the same mistake on the test set. As such, the combined prediction of a set of models that are good in different ways is usually more accurate than any prediction of any single model [39]–[41].

Ensemble learning encompasses a wide range of methods to combine the predictions of multiple models. The key to effective ensembles is gathering a set of models that disagree [42]. Common approaches to enforce differences between models are resampling-based methods such as Bootstrap aggregating (Bagging) and k-fold cross validation [43], [44]. These methods produce base models by training the same learning algorithm with different partitions of the training set. Training subsets are obtained by sampling with replacement, like in Bagging; or without replacement, like in k-fold cross-validation. Another approach to ensemble learning is Boosting. In this approach, ensemble members are generated sequentially to correct the errors of earlier models. Tracking of prediction errors determines the training set of subsequent models such that incorrectly predicted inputs are emphasized in later training iterations. Algorithms in this category are AdaBoost [45], Gradient Boosting [46], and XGBoost [47].

Model predictions are aggregated using several methods. A common approach in segmentation tasks is averaging the predictions of the ensemble members [48]–[50]. The advantage of this method is its simplicity, but it gives the same weight to all models regardless of how good their predictions are. An alternative to averaging is weighting the predictions of the ensemble members based on their performance on a holdout set. In this way, the predictions of high-performing models are privileged over the predictions of inferior models. Another approach to combining the ensemble predictions is using a machine learning algorithm. This approach is known as Stacking and the learning algorithm is termed meta-classifier [51]. Common choices for the meta-classifier are fully connected networks which allow complex, non-linear combinations of the ensemble predictions [52].

C. CONTRIBUTION

The wide variability of the appearance of retinal structures is challenging to capture in a single model. This variability is accentuated in the presence of severe pathology, where large abnormalities, like macular edema, disrupt the normal alignment of retinal structures. This work addresses the problem of segmenting the retina in OCT images exhibiting disruptive retinal pathology. We propose a deep learning approach that uses an ensemble of convolutional neural networks to delineate the retinal boundary. In contrast to single-classifier methods, our approach does not optimize a single classifier but instead leverages the predictions of multiple classifiers to improve the segmentation performance.

The main contribution of this work is the development of a fully automated method for accurate segmentation of the retina that is robust to the presence of severe retinal pathology.

We also introduce a framework that builds upon ensemble learning to train classifiers that generalize better than standalone models but require less annotated data. In addition, we investigated the impact of various data augmentation techniques to capture the variability of the retinal tissue appearance in OCT images.

Experiments in two independent datasets demonstrated that our method outperforms reference algorithms at segmenting the retinal extent in OCT scans. Furthermore, the retinal thickness measurements derived from the segmentation of the proposed algorithm showed a strong agreement with the corresponding measurements obtained with manual annotations.

II. MATERIALS AND METHODS

A. OCT RETINAL IMAGE DATASET

To develop and evaluate the proposed algorithm we used OCT retinal images from three independent datasets. All datasets are publicly available and contain healthy and pathological cases (Fig. 1).

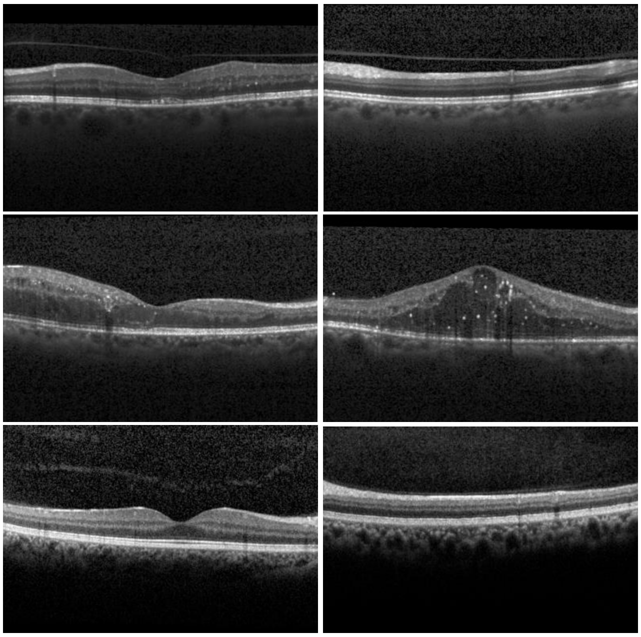

FIGURE 1. **OCT B-scans examples of the data used to train and evaluate the proposed method. Top row: Examples of the training set. Middle row: Examples of the test dataset containing DME cases. Bottom row: Examples of the test dataset containing controls.**

1) TRAINING DATA

The data used to train the proposed algorithm were sourced from the RETOUCH dataset [53]. This dataset contains 112 macula-centered OCT volumes of 112 patients. Half of the volumes are from patients with macular edema secondary to AMD, and the other half from patients with macular edema secondary to RVO. A third of the OCT volumes were acquired with a spectral-domain SD-OCT Spectralis device (Heidelberg Engineering, Heidelberg, Germany). Each volume in this set has 49 B-scan with 512 × 496 pixels with axial resolution 3.9 µm and covers a macular area of 6×6 mm^2. From this dataset, we randomly selected 110 B-scans and split them into two parts: 100 B-scans for training the algorithm and 10 B-scans for monitoring the learning progress.

2) TESTING DATA

To evaluate the algorithm's performance we used two datasets. The first is a set of 10 OCT volumes from 10 patients acquired with a Spectralis HRA+OCT device (Heidelberg Engineering, Heidelberg, Germany) [16]. The volumes were obtained from patients with DME and include B-scans showing severe macular edema. Each volume comprises 61 B-scans of 768×496 pixels, axial resolution 3.87 µm. The second is a set of 10 OCT volumes of 10 healthy patients acquired with an SD-OCT Spectralis device (Heidelberg Engineering, Heidelberg, Germany) [17]. The volumes in this set contain 61 B-scans, 496 pixels in height, axial resolution 3.87 µm; and variable-width ranging from 543 pixels to 644 pixels, lateral resolution 10-12 µm.

For details of the acquisition protocol and the exclusion criteria, the interested reader is referred to the references describing the corresponding datasets. Table 1 presents a summary of the data used in this study.

B. DATA PRE-PROCESSING

1) DENOISING AND CONTRAST ENHANCEMENT

All B-scans in the training and testing data were preprocessed to reduce speckle noise and to enhance the contrast. To reduce speckle noise, we applied a median filter with kernel size 3 × 3 pixels followed by a mean filter with kernel size 7×3 pixels. Then, to increase contrast, a power-law transformation [54] was applied to the normalized pixel intensity values. The dimensions of the kernels and the exponent of the power-law transformation were estimated empirically to preserve the continuity of the retinal boundaries in the horizontal direction. Edge detection and morphological operations were employed to fill blanks spaces at the top and bottom of the B-Scans.

2) OBSERVER MANUAL ANNOTATIONS

The retinal boundaries and the retinal thickness have multiple definitions in clinical practice and consequently vary among OCT manufacturers [55]. In this study, we defined the retina as the region between the inner limiting membrane (ILM) and the retinal pigment epithelium (RPE). Based on this definition, an experienced grader segmented the retina in each B-scan of the training and validation sets. To speed up the annotation process, a preliminary segmentation was conducted with a publicly available algorithm [56]. Upon careful examination of the preliminary boundaries, the grader adjusted the initial delineation to produce the final segmentation.

TABLE 1. **Summary of the training, validation, and testing sets.**

Dataset	Volumes	B-scans	Annotated B-scans	Dimensions (pixels)
Training	10	100	100	512 x 496
Validation	1	10	10	512 x 496
Testing (DME)	10	610	110	768 x 496
Testing (controls)	10	100	100	543 - 644 width, 496 height

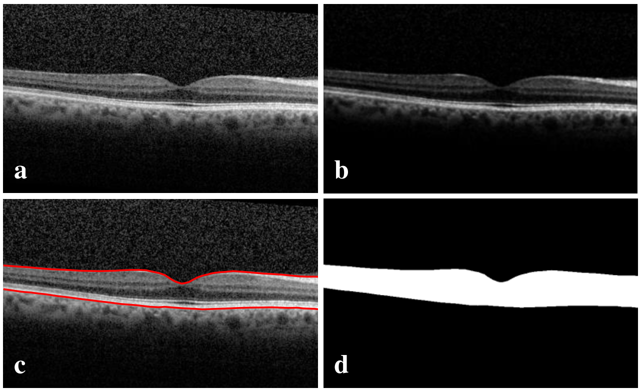

FIGURE 2. **Input image pre-processing. (a) Input, (b) Input after median filtering and contrast enhancement, (c) Boundary annotations, (d) Binary ground-truth mask. Retinal boundaries are shown as red lines in (c), top line ILM boundary, bottom line RPE boundary.**

As per the testing data, both datasets provide manual annotations of retinal layers. The dataset containing DME cases has 110 annotated B-scans, whereas the dataset containing controls includes annotations for 100 B-scans. Further information on the annotation protocol of the test data can be found in the papers describing the corresponding datasets [16], [17].

3) SEGMENTATION GROUND TRUTH

With the manual annotations, we produced ground-truth segmentation masks to train and evaluate the proposed algorithm. Pixels centered within the retina were assigned to the positive class, whereas pixels beyond the ILM or the RPE boundaries were assigned to the negative class. Labeling schemes in similar classification tasks typically use more classes [28], [35]. In this study, we opted for a two-label approach to maximize class membership while keeping the training set small.

Fig. 2 shows an example OCT and results of the pre-processing.

C. ENSEMBLE-LEARNING-BASED SEGMENTATION

The proposed method uses ensemble learning to conduct a semantic segmentation of the input OCT images. The segmentation is approached as a binary classification problem where every pixel $\mathbf{x}$ in a given input image is map to a class label $\hat{y} \in \{0, 1\}$. The ensemble-learning algorithm is implemented as a two-tier classifier in which the predictions of the first-tier classifiers $M(\mathbf{x}) = p$ are combined in the second tier through an aggregation rule to obtain the ensemble prediction $\hat{y}$. The output of the ensemble-learning algorithm is a binary segmentation map of the input image built upon the class label predictions of the ensemble. Fig. 3 shows the major components of the proposed algorithm.

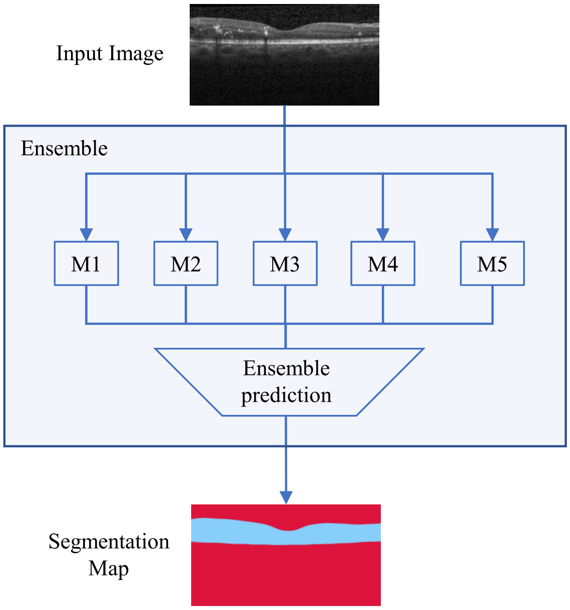

FIGURE 3. **Overview of the proposed method. The input OCT image is fed to the first-tier classifiers, which independently predict class labels for every pixel of the input. The first-tier predictions are then combined in the ensemble-prediction block to construct a segmentation map of the input OCT image. M_j: First-tier classifiers, $j \in$ [1-5].**

1) ENSEMBLE PREDICTION

To decide the ensemble prediction we used three schemes to combine the first-tier predictions: majority voting, weighted averaging, and stacking. In the majority-voting scheme the ensemble prediction is decided by averaging out the predicted probabilities of the first-tier classifiers. More formally, the predicted class label is defined as follows:

$$\hat{y} = \arg\max_{i} \frac{1}{N} \sum_{j=1}^{N} p_{ij} \tag{1}$$

where, p_{ij} is the predicted class probability of the jth first-tier classifier for class label i, and N is the number of first-tier classifiers.

In the weighting-averaging scheme, the predicted class probabilities are weighted according to the classification performance of the first-tier classifiers in a holdout set. In this scheme, the predicted class label of the ensemble is given by:

$$\hat{y} = \arg\max_{i} \sum_{j=1}^{N} w_j p_{ij} \tag{2}$$

where, w_j is the weight of the jth first-tier classifier.

In the stacking scheme, the ensemble prediction is determined by feeding the predictions of the first-tier classifiers to another machine learning algorithm termed meta-classifier. This learning algorithm is trained to optimize a non-linear combination of the first-tier predictions.

D. RETINAL THICKNESS ASSESSMENT

The retinal thickness is determined by computing the distance between the ILM and RPE boundaries at every column in the segmentation map. To compute the distance we first identify the coordinates of the pixels in the retinal boundaries. Let be $\mathbf{S_{mxn}}$ the matrix that represents the segmentation map, s_{ij} an element of $\mathbf{S}$, and $c_j = [s_{1j}, s_{2j}, \ldots, s_{mj}]^\top$ a column of $\mathbf{S}$. Then, each boundary $B^{(k)} = [r_1, r_2, \ldots, r_n]$ is a row vector of length n, $k \in \{ILM, RPE\}$. Coordinates in the ILM boundary correspond to the topmost pixels of label $\hat{y} = 1$ in the segmentation map $\mathbf{S}$. Thus, the coordinate of the ILM boundary in the column c_j is given by:

$$r_j = \arg\min_{i} ic_j \tag{3}$$

For the RPE boundary, the coordinates r_j belong to the lowest pixels of the class label $\hat{y} = 1$ in $\mathbf{S}$ and are given by:

$$r_j = \arg\max_{i} ic_j \tag{4}$$

With the coordinates of the ILM and RPE boundaries we determined the total retinal thickness T as follows:

$$T = |B^{(RPE)} - B^{(ILM)}| \tag{5}$$

Like the retinal boundaries, the total retinal thickness is a row vector of length n that contains thickness measurements at every column of a given segmentation map. In clinical practice, these measurements are aggregated to obtain thickness profiles of different retinal regions such as the central macular thickness (CMT), which is the region within 1 mm around the center of the macula.

E. ENSEMBLE ARCHITECTURE

The proposed ensemble-learning algorithm comprises five first-tier classifiers. All first-tier classifiers are convolutional neural networks and have the same base architecture. The number of first-tier classifiers is a hyperparameter of the ensemble-learning algorithm and it was empirically determined. The first-tier-CNN architecture was also empirically determined by observing the performance of several network candidates of varying depth and complexity.

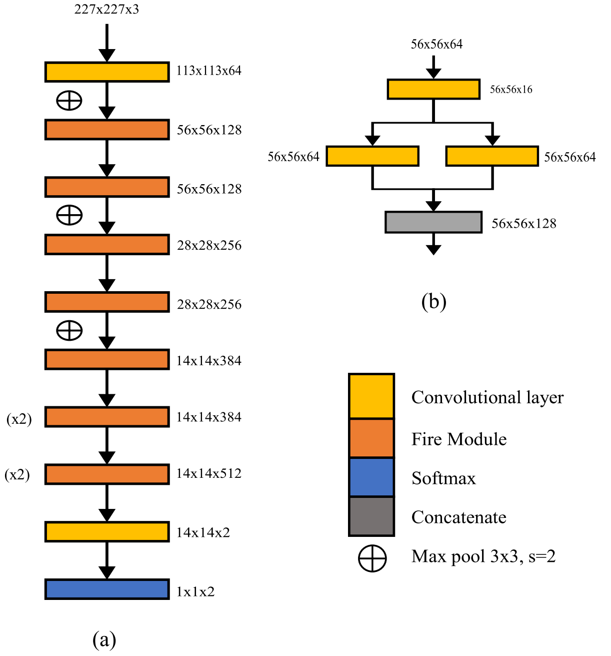

FIGURE 4. Simplified diagram of the base convolutional neural network used to train the first-tier classifiers. (a) Network architecture, (b) Detail of the topmost Fire module. Activation size is indicated next to the corresponding layer.

1) FIRST-TIER-CLASSIFIER BASE ARCHITECTURE

The base architecture of the first-tier classifiers is the SqueezeNet version 1.1 network [57]. The architecture is 20 layers deep and consists of a convolutional layer that takes the input image, followed by 9 fire modules, a convolutional layer, and the output layer. The fire modules comprise convolutional layers that perform squeeze and expand operations over their corresponding inputs. The squeeze convolutional layers have only 1 × 1 filters, whereas the expansion convolutional layers may have 1 × 1 or 3 × 3 filters. The output layer applies a softmax function to the output of the last convolutional layer to produce the probability scores for each class.

2) META-CLASSIFIER

In the stacking variant of the ensemble, the meta-classifier is a deep neural network that receives the stacked probability scores of the five first-tier CNN. The meta-classifier architecture comprises an input layer followed by a fully-connected layer with 512 nodes and ReLU activation, a fully connected layer with 256 nodes and ReLU activation, a softmax layer, and a classification layer.

F. TRAINING

1) TRANSFER LEARNING

Compared to stand-alone classifiers, ensemble learning has a drawback: the number of computations grows linearly with the number of first-tier classifiers. To reduce the computational complexity we used transfer learning to train the first-tier CNNs. To train the first CNN we fine-tuned all layers of a model pre-trained in the ImageNet dataset. Then, to train the other four networks, we initialize each network with the first CNN, froze all convolutional layers, and fine-tuned the remaining layers.

2) DATA AUGMENTATION

Central to the success of ensemble learning is generating classifiers that disagree. In our approach, we enforce diversity by training the ensemble CNNs with augmented versions of the training set. This allowed us to use the whole training set to generate all first-tier classifiers, as opposed to the resampling approach where each ensemble member is trained with a fraction of the training data. To augment the data, we used geometric and pixel-intensity value transformations. Geometric transformations included random rotation and random vertical reflection. For the pixel-intensity we used power-law transformations [54].

3) TRAINING SET

To produce training samples, we extracted small overlapping patches from the OCT images and labeled them according to the class of the central pixel. In this type of approach, patch size has been demonstrated to influence the classification accuracy [58]. In related work, patch sizes ranged from 33 × 33 [28] pixels to whole B-scans [37]. In this study, we determined empirically the best patch size to be 65 × 65 pixels. Experiments with other patch sizes showed that large patches increase the classification performance but at cost of increasing the computational complexity.

To reduce the computational workload, we restricted the patch extraction to the regions where pixels were not trivially identified by standard image processing methods, e.g. by thresholding. We defined this region of interest (ROI) as two strips surrounding the ILM and RPE boundaries (see Fig. 5(a)). The height of the strips was determined empirically to be the height of one patch. Furthermore, to reduce the occurrence of very similar patches, we limited the creation of patches to evenly spaced columns in each B-scan.

From the OCT B-scans in the training and validation sets, we created patches for the positive and negative classes. To balance the number of samples per class, we randomly selected 100 patches per class in each B-scan (200 patches per image). The resulting training set consisted of 20,000 patches, whereas the validation set comprised 2,000 patches. This last set was used to enforce regularization by early stopping.

III. EXPERIMENTS

A. ENSEMBLE SETUP

1) SELECTION OF THE FIRST-TIER-CNN ARCHITECTURE

To select the architecture of the first-tier classifiers we fine-tuned ten networks pre-trained in the ImageNet dataset and evaluate their performance on the validation set. To adapt the pre-trained models to our classification task, in each model we removed the last layer and replaced it for a blank fully connected layer with an output size equal to two. All pre-trained models were sourced from [59]. The models were trained in a supervised fashion to perform binary classification, being the classification task to label pixels as belonging to the retina or otherwise. The network performances were evaluated with 5-fold cross-validation. We also observed the training time for all folds and architectures

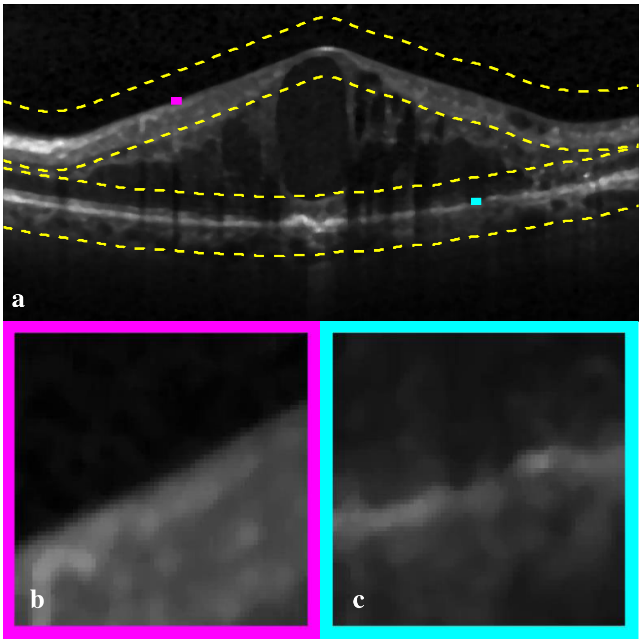

FIGURE 5. Patch extraction. (a) Input, (b) and (c) patch examples extracted from the region of interest (ROI), edge color correspond to pixel color of the point of extraction. Dashed lines delineate the ROI for patch extraction. Train patches were randomly extracted from this region only.

TABLE 2. Hyperparameters of the first-tier CNN training.

Hyperparameter	Value
Optimizer	Adam
Loss function	Cross entropy
Max. Epochs	100
Minibatch size	[8, 32, 128]
Learning rate	$10^{-6} - 10^{-1}$
Regularization	Early stop

All CNNs were trained with stochastic gradient descent for a maximum of 100 epochs. The learning rate and the batch size were determined empirically through a grid search, as recommended in [60]–[62]. The network parameters were updated with the Adam gradient-based optimization algorithm [63], to minimize the cross-entropy loss function [64]. The Adam optimizer has been demonstrated to perform well, and to converge faster than other stochastic-gradient-descent methods [65]. The training set was shuffled every epoch to prevent overfitting by data-order memorization. The validation loss was checked twice per epoch, and the training was stopped if the validation loss failed to improve for five consecutive checkpoints. Table 2 summarize the hyperparameters of the first-tier-classifier training.

2) DETERMINATION OF THE NUMBER OF FIRST-TIER CLASSIFIERS

After defining the CNN base architecture, we investigated the influence of the number of first-classifier on the ensemble performance. To that end, we trained four more first-tier classifiers with the best combination of hyperparameters found in the grid search. To enforce predictive diversity between classifiers, we trained every network with different augmented versions of the training set.

To observe the influence of the number of models we evaluated the performance of every combination of 2, 3, 4, and 5 different models. The first-tier-classifier predictions were aggregated with three schemes: majority voting, weighted average, and stacking. Furthermore, to compare the ensemble performance to single-model performance, we evaluated every first-tier classifier individually.

3) META-CLASSIFIER

The meta-classifier of the stacking variant of the ensemble was trained for a maximum of 100 epochs with a batch size of 32. The optimization algorithm was Adam and the loss function cross-entropy. Considering the adaptive mechanism of the parameter updates in the Adam algorithm, we selected an initial learning rate of 0.1 and left the remaining parameters at the default values.

B. COMPARISON OF METHODS

To further evaluate the performance of the proposed method, we compared our segmentation results with corresponding results from three other methods. The first was a fully-convolutional network (FCN) with a U-net architecture [66] modified as proposed in [37]. Different from our approach, this method conducted semantic segmentation on whole OCT inputs. The FCN in this method included four downsampling units, each comprising two 3 × 3 convolutions, one ReLU, and a max-pooling layer with filter size 2 × 2. The network was trained for 100 epochs with minibatch size 8, learning rate starting at 5e-4 up to 1.8e-4 as specified in [37]. The loss function was the cross-entropy and the optimization algorithm Adam [63]. We trained the network with the same dataset and the same development environment that we used to train our models. The other two algorithms in the comparison were two existing graph-based segmentation methods widely referenced in research: the OCT Explorer tool version 3.8, part of the Iowa Reference Algorithms (Retinal Image Analysis Lab, Iowa Institute for Biomedical Imaging, Iowa City, IA) [18], [67], [68]; and the JHU OCT Segmentation Version 2.11, part of the AURA tools [25].

C. PERFORMANCE METRICS

To evaluate the classification and segmentation performance, we used several metrics of performance. The formal definition of the performance metrics is presented below.

1) CLASSIFICATION ACCURACY

To evaluate the classification performance we computed the accuracy of the class-label prediction. This metric is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

where TP is the number true positives, TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives.

2) SEGMENTATION

To quantify the segmentation performance we used the Sørensen–Dice similarity coefficient between the segmentation maps obtained with the proposed algorithm and the segmentation ground truths. The score es defines as:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{7}$$

To measure the error of the segmentation we computed the mean absolute error of the location of the ILM and the RPE boundaries. This error was computed as follows:

$$MAE = \frac{1}{NM} \sum_{j=1}^{N} \sum_{i=1}^{M} |B_{ij}^{(k)} - \hat{B}_{ij}^{(k)}| \tag{8}$$

where, B_{ij} represents the ground-truth location of the boundary k in the A-scan i of the B-scan j, $\hat{B}_{ij}$ is the predicted location of the boundary k in the A-scan i of the B-scan j, N is the number of B-scans, M is the number of A-scans, and $k \in \{ILM, RPE\}$.

3) RELIABILITY OF THE RETINAL THICKNESS MEASUREMENTS

To evaluate the reliability of the retinal thickness measurements, we computed the mean absolute difference, and Bland-Altman statistics of the CMT measurements.

The mean absolute difference was computed in a similar way to the boundary location error, by comparing the CMT obtained with the manual annotations to corresponding values computed with the predicted boundary locations, as follows:

$$MAE = \frac{1}{N} \sum_{j=1}^{N} |C_j - \hat{C}_j| \tag{9}$$

where, C_j represents the ground-truth CMT of B-scan j, $\hat{C}_j$ is the estimated CMT of B-scan j, and N is the number of B-scans.

The Bland-Altman statistics, also called limits of agreement, measures the magnitude of the agreement between two sets of measurements. This statistics comprises the mean of the difference between pairs of corresponding measurements, or bias, and the limits of the 95% confidence interval of the bias. Let be d_j be the difference between the estimated

TABLE 3. Mean (std) accuracy and similarity score per CNN architecture. Metrics were measured by k-cross-validation, with k=5 on the holdout set. Training time is the mean fold-training duration in minutes. All models were trained for 100 epochs.

Network	Accuracy	Dice	Training time (min)
VGG16	**97.32 (04.29)**	**0.92 (0.06)**	54.03
Squeezenet	96.10 (07.65)	0.91 (0.08)	**25.87**
Alexnet	93.45 (08.00)	0.88 (0.08)	30.37
Googlenet	91.48 (12.48)	0.88 (0.10)	34.01
Inceptionv3	91.05 (11.20)	0.85 (0.09)	94.44

CMT $\hat{C}_j$ and the corresponding ground truth CMT C_j, then the bias is given by:

$$b = \mu(d_j) \tag{10}$$

where, $\mu(.)$ is the mean function, and N is the number of CMT measurement pairs. The limits of agreement are defined as:

$$[u, l] = [b + \sigma(d_j), b - \sigma(d_j)] \tag{11}$$

where, u is the upper limit, l is the lower limit, and $\sigma(.)$ is the standard deviation.

D. HARDWARE AND SOFTWARE TOOLS

The development and testing environment was MATLAB® release 2019b and CUDA® library version 9.0. All models were trained on a desktop computer with Windows 10 operating system, processor Intel i7 8700K CPU @ 3.7 GHz - 6 cores and 32 GB RAM, using a GPU NVIDIA® GeForce GTX® 1080 Ti with 11 GB RAM.

IV. RESULTS

A. ENSEMBLE SETUP

1) FIRST-TIER CLASSIFIER ARCHITECTURE AND HYPERPARAMETERS

To find the first-tier CNN architecture, we evaluated ten convolutional neural networks pre-trained in the ImageNet dataset. The networks we fine-tuned by grid search with different combinations of the minibatch size and the learning rate hyperparameters. The prediction accuracy and the Dice coefficient were evaluated on the validation set with k-fold cross-validation (k=5). We also observed the training time for all folds and architectures. Table 3 shows the top-five networks for the best combination of hyperparameters, minibatch size = 32 and learning rate = 10^{-2}.

2) NUMBER OF FIRST-TIER CLASSIFIERS

Upon the definition of the base-CNN architecture, we trained four more classifiers to join the ensemble. To obtain the ensemble prediction we evaluated three schemes to combine the predictions of the first-tier classifiers: majority vote, weighted average, and stacking. The classification accuracy and the similarity Dice score were evaluated on the validation set. Both metrics were evaluated for every possible combination of M models (M ∈ [2,5]) and every aggregation method.

TABLE 4. Mean (std) classification accuracy per combination of M first-tier classifiers, from M=2 to M=5. Best results in each combination of M classifiers are listed. The classification performance was evaluated on the test set.

Number of models	Majority vote	Weighted average	Stacking
5	99.48 (0.16)	99.47 (0.16)	99.51 (0.13)
4	99.42 (0.19)	99.41 (0.19)	99.46 (0.16)
3	99.46 (0.21)	99.45 (0.20)	99.49 (0.14)
2	99.07 (0.46)	99.07 (0.46)	99.33 (0.33)

TABLE 5. Mean and standard deviation of the classification accuracy and the Dice similarity score of individual first-tier classifiers. The performance metrics were computed on the test set.

Model	Accuracy	Dice
1	94.39 (02.98)	0.836 (0.07)
2	98.11 (01.66)	0.938 (0.05)
3	88.52 (07.39)	0.742 (0.12)
4	82.22 (13.21)	0.674 (0.18)
5	98.33 (01.46)	0.941 (0.06)

Table 4 shows the highest scores in every combination of M models. The performances of individual models are shown in Table 5 for comparison. As can be seen, the classification accuracy of the ensemble is higher than any of the individual classifiers. Also, gains in performances are observed from the combination of only 2 models.

B. COMPARISON OF METHODS

1) TOTAL RETINA SEGMENTATION

The performances of the proposed algorithm and reference methods were evaluated on two independent datasets containing normal and DME cases. Part of the test data, particularly the DME set proved to be challenging for the reference algorithms. OCT-Explorer produced no segmentation for seven whole DME volumes and left out large areas of the retina in 26% of the remaining B-scans. These results were excluded from the evaluation to not distort the comparison between algorithms. Fig. 6 shows an example of the DME group showing a severely damaged retina. The segmentation performance in the control group was generally higher and the failure rate of reference algorithms were considerably lower. Fig. 7 shows examples of the segmentation in the control group.

The Sørensen–Dice similarity coefficient between the algorithm-segmentation maps and the ground-truth-segmentation masks in the whole test data was 0.991 ± 0.002 for the proposed algorithm, whereas reference methods Aura tools, OCT-Explorer and U-net attained 0.960 ± 0.051, 0.974 ± 0.018, 0.029 ± 0.004, and 0.906 ± 0.027 respectively. Boxplots in Fig. 8(a) show the distribution of the similarity scores in the control set, whereas boxplots in Fig. 8(b) show the score distribution in the DME set.

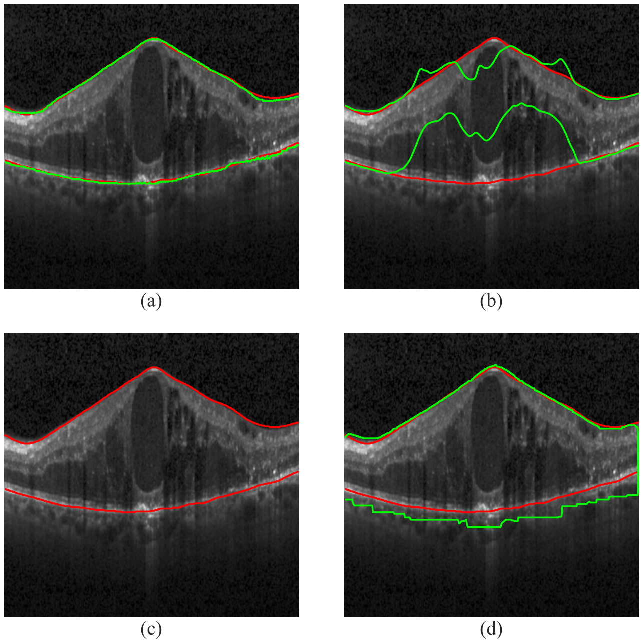

FIGURE 6. Example of the segmentation obtained by evaluated algorithms in the DME set. (a) the proposed method, (b) Aura tools, (c) OCT Explorer, and (d) the U-net-based algorithm. The algorithm boundaries are shown in green and the manual annotations in red.

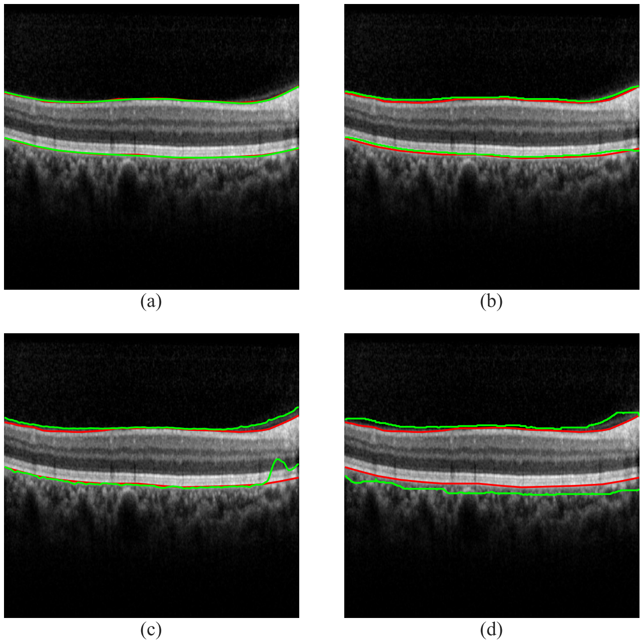

FIGURE 7. Example of the segmentation obtained by evaluated algorithms in the normal set. (a) the proposed method, (b) Aura tools, (c) OCT Explorer, and (d) the U-net-based algorithm. The algorithm boundaries are shown in green and the manual annotations in red.

As for the error of the delineation of the retinal boundaries. The overall mean absolute error of the proposed method was 0.9 ± 0.4 pixels for the IML, and 1.0 ± 0.5 pixels for the RPE. Corresponding MAE values for Aura tools were 2.4 ± 1.9 pixels for the ILM, and 4.8±9.3 pixels for the RPE. The MAE of the OCT-Explorer algorithm for the ILM and RPE were 3.1 ± 4.7 pixels, and 2.6 ± 4.8 pixels respectively; whereas the U-net network obtained MAE values of 4.9 ± 2.6 pixels for the ILM, and 11.9 ± 4.1 pixels for the RPE. The distribution of the boundary error grouped by dataset is shown in Fig. 9.

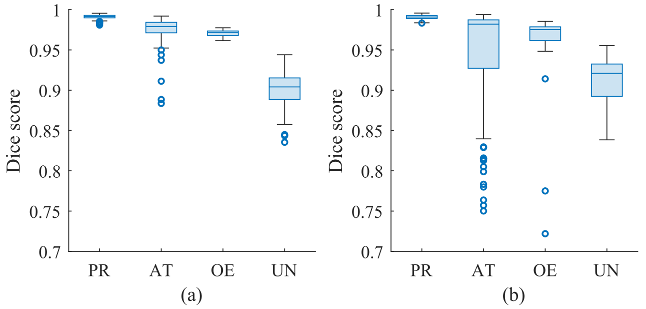

FIGURE 8. Boxplots of the distribution of the Dice coefficient of the similarity between the algorithms' segmentation and the ground truths, as measured in (a) the normal set, (b) the DME set. PR: the proposed method, AT: Aura tools, OE: OCT Explorer, UN: U-net based algorithm.

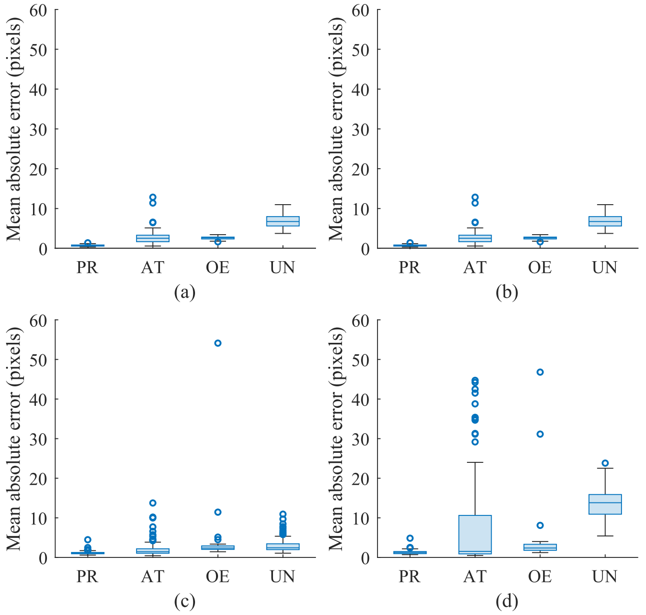

FIGURE 9. Boxplots of the distribution of the ILM and RPE location error. The error was measured in the normal and DME sets by comparing the algorithm predictions with the manual annotations. (a) Mean absolute ILM error in the normal set, (b) Mean absolute RPE error in the normal set, (c) Mean absolute ILM error in the DME set, and (d) Mean absolute RPE error in the DME set. PR: the proposed method, AT: Aura tools, OE: OCT Explorer, UN: U-net based algorithm.

2) RETINAL THICKNESS ASSESSMENT

The mean CMT computed with the ground truth in the normal set was 69.8 ± 8.2 pixels, whereas the mean estimate of the proposed algorithm was 70.4 ± 8.0 pixels. Reference algorithms Aura tools, OCT-Explorer, and the U-net network obtained mean CMT values of 73.1 ± 7.7 pixels, 71.2 ± 7.4 pixels, 86.6 ± 5.5 pixels respectively. In the DME

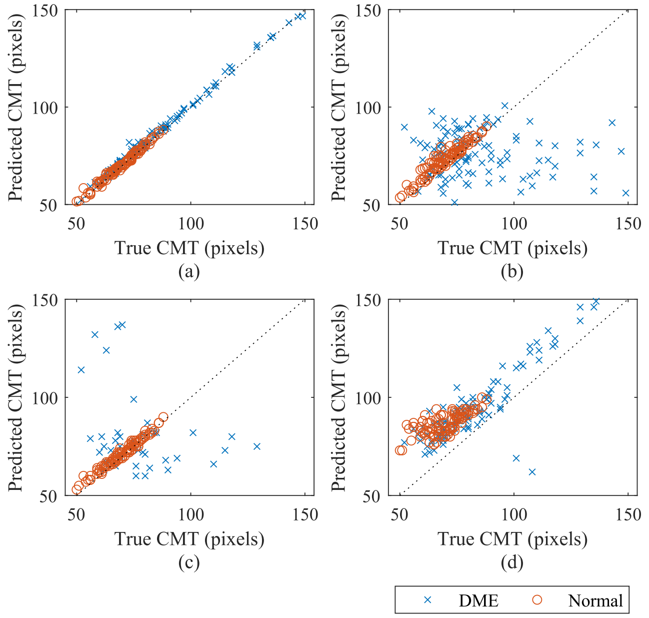

FIGURE 10. Agreement between the true central macular thickness (CMT) and corresponding measurements derived from (a) the proposed method, (b) Aura tools, (c) OCT Explorer, and (d) the U-net-based algorithm. DME and normal subgroups are identified by marker color and marker symbol. The dotted line represents a perfect agreement of the two measurements.

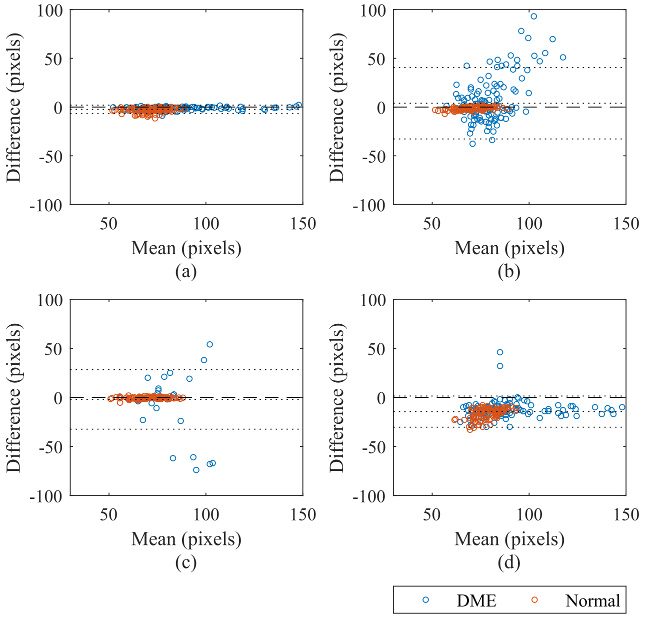

FIGURE 11. Bland-Altman plots of the systematic difference between the estimated central macular thickness (CMT) and the ground truth for (a) the proposed method, (b) Aura tools, (c) OCT Explorer, and (d) the U-net network. The 95% limits of agreement are shown in dotted lines. The zero bias line is indicated by the dashed line. DME and normal subgroups are identified by color.

group, the true mean CMT was 84.8±21.0 pixels, whereas the algorithm estimates were 86.3 ± 20.6 pixels for the proposed method, 75.9±11.5 pixels for Aura tools, 86.17±23.9 pixels for the OCT Explorer, and 97.2 ± 20.5 pixels for the U-net network.

The overall mean absolute error of the CMT was 1.3 ± 1.2 pixels for the proposed algorithm, 11.5 ± 15.5 pixels for AURA tools, 6.3 ± 14.4 pixels for OCT-Explorer, and 15.2 ± 6.6 pixels for the U-net-based algorithm. Fig. 10 shows scatter plots of the algorithm estimates of the CMT versus corresponding true values. The CMT estimates of the proposed algorithm and corresponding true values showed a strong correlation with a Pearson correlation coefficient of 0.997. By contrast, the correlation coefficients for reference algorithms were 0.154 for Aura tools, 0.246 for OCT Explorer, and 0.891 for the U-net network.

3) RELIABILITY OF THE RETINAL THICKNESS MEASUREMENTS

Bland-Altman plot analysis of CMT measurements showed a strong agreement between the CMT obtained with the proposed algorithm and those of the human graders (Fig. 11(a)). The proposed algorithm's mean difference to the true CMT was -1.09 pixels with 95% limits of agreement between 1.78 and −3.95 pixels. Whereas, the mean differences, and corresponding 95% limits of agreement of reference algorithms were 3.03, [40.34 −34.28] pixels for AURA tools; −2.78, [27.47, −33.04] pixels for OCT-Explorer; and −14.49, [1.39, −30.38] pixels for the U-net-based method. Corresponding bias values in metrics units were −4.12 µm for the proposed method, 11.53 µm for Aura tools, −10.57 µm for OCT Explorer, and −55.08 µm for the U-net based method. The difference between the predicted CMT of reference methods and corresponding truths increased with the macular thickness. This confirms that segmenting severe DME cases is challenging to said algorithms. Furthermore, the 95% interval of agreement of the proposed method was significantly narrower than the intervals of reference algorithms and there were fewer data points outside of the limits of agreement.

The intra-class correlation coefficient (ICC) between the manual CMT measurements and corresponding measurements derived from the proposed method was 0.961 with 95% CI [0.827 0.984] in the normal set and 0.994 with 95% CI [0.953 0.998] in the DME set. By contrast, the ICC for reference algorithms in the DME set was 0.1 for Aura tools, 0.03 for OCT Explorer, and 0.77 for the U-net network. In the normal set, the corresponding ICC values were 0.87, 0.98, and 0.16 respectively.

Table 6 summarizes the performance of the different algorithms in the normal and DME groups, along with corresponding values of performance for the entire test data.

V. DISCUSSION

The CNN architecture and hyperparameters of the first-tier classifiers were empirically selected by grid search. For the same learning rate, we observed the impact of the minibatch size on the prediction accuracy and the training duration. Large minibatch sizes resulted in shorter training but higher accuracy, whereas short minibatch sizes resulted in

TABLE 6. **Mean absolute error and similarity score of the retina segmentation obtained by evaluated algorithms. Metrics were computed in the entire test data, as well as in the DME and the normal test sets.**

	Proposed	Aura Tools	OCT Explorer[a]	U-net
Overall				
Mean absolute CMT error	1.3 ± 1.2 pixels	11.5 ± 15.5 pixels	6.3 ± 14.4 pixels	15.2 ± 6.6 pixels
Mean absolute ILM error	0.9 ± 0.4 pixels	2.40 ± 1.90 pixels	3.1 ± 4.70 pixels	4.9 ± 2.60 pixels
Mean absolute RPE error	1.0 ± 0.5 pixels	4.80 ± 9.30 pixels	2.6 ± 4.80 pixels	11.9 ± 4.1 pixels
Dice score	0.991 ± 0.003	0.960 ± 0.051	0.967 ± 0.029	0.906 ± 0.027
DME				
Mean absolute CMT error	0.9 ± 0.7 pixels	3.30 ± 2.40 pixels	1.7 ± 1.4 pixels	16.8 ± 6.0 pixels
Mean absolute ILM error	0.7 ± 0.2 pixels	2.70 ± 1.80 pixels	2.6 ± 0.3 pixels	6.9 ± 1.60 pixels
Mean absolute RPE error	0.8 ± 0.3 pixels	1.30 ± 1.10 pixels	1.9 ± 0.4 pixels	9.9 ± 3.20 pixels
Dice score	0.991 ± 0.003	0.974 ± 0.018	0.971 ± 0.004	0.900 ± 0.022
Normal				
Mean absolute CMT error	1.7 ± 1.4 pixels	18.9 ± 18.4 pixels	25.3 ± 25.0 pixels	13.8 ± 6.9 pixels
Mean absolute ILM error	1.1 ± 0.4 pixels	2.10 ± 2.00 pixels	5.0 ± 10.70 pixels	3.1 ± 1.80 pixels
Mean absolute RPE error	1.3 ± 0.5 pixels	8.00 ± 12.0 pixels	5.6 ± 10.60 pixels	13.6 ± 4.1 pixels
Dice score	0.990 ± 0.003	0.948 ± 0.066	0.952 ± 0.065	0.911 ± 0.029

[a] OCT Explorer failed to produce any segmentation for 86 DME B-scan. These results were excluded from the statistics to not distort the comparison

the opposite. This observation is consistent with prior research regarding this hyper-parameter optimization [60], [69]. To inform the selection of the base architecture, we observe the performance and training time (see Table 3). For the best set of hyperparameters, we found that the VGG16 architecture attained the highest accuracy, but it was the second slowest of all. On the other hand, the SqueezeNet architecture ranked second in classification accuracy and was the fastest. With this, we chose SqueezeNet as the architecture for first-tier classifiers.

Regarding the influence of the number of first-tier classifiers on the ensemble performance, we observed the performance grows with the number of models (Table 4). The trend is consistent across all three aggregation schemes, with the stacking scheme showing the biggest classification performances in every combination of models. However, we also noted that although the classification performance grows with the number of models, the gains in performance tend to stabilize past 3 models. Similarly, increasing the complexity of the aggregation scheme shows no clear incentive in that gains in performance of the most complex scheme (stacking) are not significantly higher than those of the simplest scheme (majority vote). These observations evidence a trade-off between computational complexity and classification performance.

The proposed method used supervised learning to train the first-tier classifiers of the deep ensemble learning. Ordinarily, supervised learning is conducted with large sets of annotated data to achieve high classification performance [70], [71]. Recent works on retinal segmentation through supervised learning reported dataset sizes in the range of tens of thousands of annotated B-scans [28], [37], [58]. By contrast, our algorithm was trained with a dataset much smaller, yet it accomplished high classification accuracy and outperformed reference methods. Furthermore, we compared the classification performance of our transfer-learning-based approach against the performance of a U-net model trained from scratch. Although both models were trained with the same dataset, the performance of our model was significantly superior in both test sets (see Table 6).

A. COMPARISON OF METHODS

The segmentation performance of our method was evaluated with two test sets and contrasted against corresponding performances of three algorithms, one based on a deep learning network and two other non-deep-learning methods which are highly referenced in related research. These methods, implement graph-cut algorithms where the retinal layers are individually segmented by fitting segmenting surfaces to estimates of the layer boundaries. As a result, these algorithms are highly sensitive to large disruptions such as those of retinas with severe DME. The segmentation error is typically large, and occasionally the algorithms failed to segment the B-scan completely. Fig. 6 shows an example of severe DME, where large intraretinal fluid pockets broke the boundaries of the retinal layers. Graph-cut-based algorithms left parts of the retina out (see Fig. 6(b)), or produced no segmentation at all (see Fig. 6(c)). By contrast, the proposed method properly segmented the retina, as shown in Fig. 6(a).

Examining the algorithm performance in the control group, we observed an overall higher performance. Particularly, the segmentation error and the failure rate of reference algorithms were considerably lower than corresponding values in the DME set. However, the performance of the proposed method was significantly higher across all metrics, as shown in Table 6. Despite the performance improvement, noticeable errors occurred in the segmentation obtained with reference methods, especially in those from the OCT Explorer algorithm. Such errors appeared more frequently towards

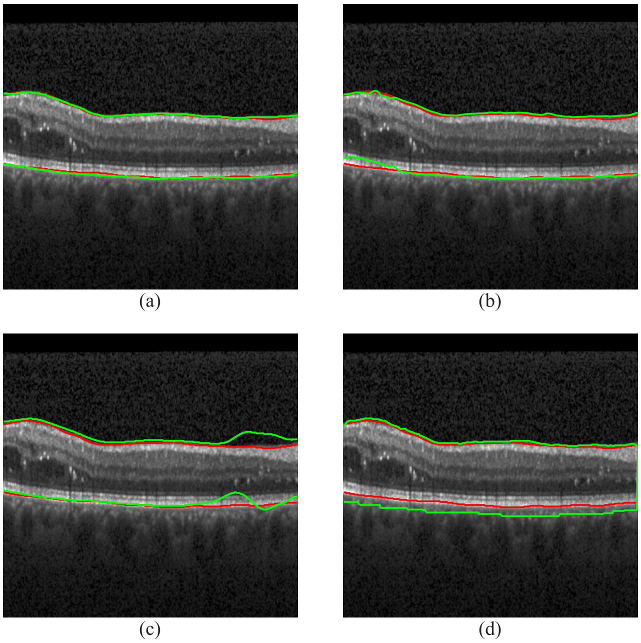

FIGURE 12. Example of the segmentation obtained by evaluated algorithms in the DME set. (a) the proposed method, (b) Aura tools, (c) OCT Explorer, and (d) the U-net-based algorithm. The algorithm boundaries are shown in green and the manual annotations in red.

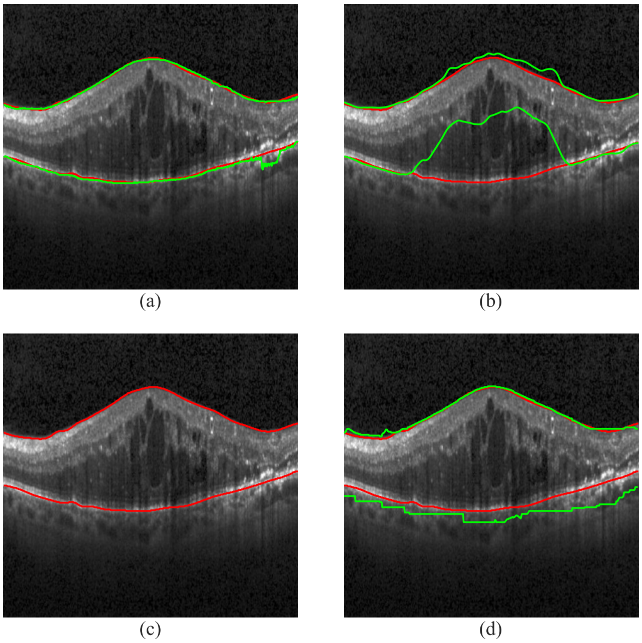

FIGURE 13. Example of segmentation errors produced by the proposed method (a) in the DME set. Corresponding segmentation from Aura tools (b), OCT Explorer (c), and the U-net-based algorithm (d) are also presented. The algorithm boundaries are shown in green and the manual annotations in red.

the nasal quadrant, where the boundary delineations of the ILM and RPE layers diverged from the expected location (see Fig. 7(c) and Fig. 12(c)).

Careful examination of the OCT volumes that end up badly segmented, led us to note that the cause of the errors is related to the presence of abrupt transitions between adjacent B-scans. Reference algorithms leverage 3D features of the OCT volumes to estimate initial boundaries upon which they determine the final segmentation. Whereas this approach might work well on dense volumes where neighboring B-scans exhibit a strong correspondence, it results in poor boundary estimates in volumes with a low number of B-scans like the ones used in this study. This evidences a limitation of reference algorithms considering that the amount of B-scans per volume is typically an uncontrollable variable.

The proposed method performed consistently high in both normal and DME sets, however, a few B-scans in the latter group presented minor segmentation errors. These errors occurred around hyperreflective foci (HF) clusters outside the retina. HF is associated with intermediate stages of DME and appears in the inner and outer layers as highly reflective spots, brighter than the RPE [72]. In the presence of HF clusters, the proposed algorithm classifies pixels in the groupings as belonging to the retina. As a consequence, the resulting segmentation oversteps the target boundary (see Fig. 13(a)). The cause of the segmentation errors is the under-representation of these abnormal formations in the train data, and it should be addressed by adding more examples of mild DME to the training set.

Looking at the reliability of the reference algorithms we observed a high incidence of large deviations in the DME group (see Figs. 10(b) and 10(c). Similarly, corresponding Bland-Altman plots show that the reference algorithm's CMT measurements deviate in greater magnitude with thicker retinas (Fig. 11). This evidences that segmentation errors occurred frequently in the presence of macular edema. OCT Explorer failed to produce any segmentation for 86 DME B-scans, whereas Aura tools left out parts of the retina in 29 B-scans of the DME group. The U-net-based algorithm performed slightly better in the set of DME cases but the segmentation error was 10%. By contrast, our algorithm successfully segmented all DME test data with a mean segmentation error of 1%.

VI. CONCLUSION

A fully automated method based on patch classification was developed to segment the region in between the outermost layers of the retina in OCT B-Scans. The proposed algorithm outperformed reference methods in the presence of macular fluid, a retinal pathology in which existing algorithms have been observed to have a low-performance [3].

Reference algorithms, particularly those based on graph theory, rely on expected features of the retina layers to fit segmentation curves or surfaces to the OCT scans or volumes. Assumptions regarding the smoothness of the retinal contour and the retinal thickness allow these algorithms to limit the amount of deformation of the retina. Whereas this constraint-driven approach has proved to be effective in segmenting normal retinas, we observed that it fails in retinas exhibiting macular edema, particularly in cases with large intraretinal cysts.

Contrary to reference algorithms, our method implements a patch-based semantic segmentation of the ROI that classifies pixels according to their local context rather than predetermined constraints. This allows our algorithm to capture pathological changes in the retinal boundaries, which are typically variable in shape, orientation, and location. As a result, the proposed method produces accurate segmentation boundaries regardless of the presence of macular edema.

In conclusion, we introduced a robust method to automatically segment the neurosensory retina in OCT images. Our algorithm accurately delineates the boundaries of the outer retinal layers, even in the presence of severe pathology. Moreover, the proposed method can produce reliable clinical measurements derived from the segmentation such as the central macular thickness. Our algorithm was evaluated on two independent OCT datasets and outperformed reference methods at segmenting normal and edematous retinas. This suggests that our method can be reliably used in clinical practice as an alternative to labor-intensive manual processing of retinal OCT images.

ACKNOWLEDGMENT

The authors would like to thank the Institute for Telecommunications for providing the computational and physical resources needed to develop this work.

REFERENCES

[1] J. F. De Boer, R. Leitgeb, and M. Wojtkowski, "Twenty-five years of optical coherence tomography: The paradigm shift in sensitivity and speed provided by Fourier domain OCT," *Biomed. Opt. Exp.*, vol. 8, no. 7, pp. 3248–3280, 2017.

[2] M. Adhi and J. S. Duker, "Optical coherence tomography–current and future applications," *Current Opinion Ophthalmol.*, vol. 24, no. 3, pp. 213–221, 2013.

[3] S. M. Waldstein, B. S. Gerendas, A. Montuoro, C. Simader, and U. Schmidt-Erfurth, "Quantitative comparison of macular segmentation performance using identical retinal regions across multiple spectral-domain optical coherence tomography instruments," *Brit. J. Ophthalmol.*, vol. 99, no. 6, pp. 794–800, Jun. 2015.

[4] W. Geitzenauer, C. K. Hitzenberger, and U. M. Schmidt-Erfurth, "Retinal optical coherence tomography: Past, present and future perspectives," *Brit. J. Ophthalmol.*, vol. 95, no. 2, pp. 171–177, Feb. 2011.

[5] S. M. Waldstein, J. Wright, J. Warburton, P. Margaron, C. Simader, and U. Schmidt-Erfurth, "Predictive value of retinal morphology for visual acuity outcomes of different ranibizumab treatment regimens for neovascular AMD," *Ophthalmology*, vol. 123, no. 1, pp. 60–69, Jan. 2016.

[6] A. Wood, A. Binns, T. Margrain, W. Drexler, B. Považay, M. Esmaeelpour, and N. Sheen, "Retinal and choroidal thickness in early age-related macular degeneration," *Amer. J. Ophthalmol.*, vol. 152, no. 6, pp. 1030–1038, 2011.

[7] A. Aojula, S. P. Mollan, J. Horsburgh, A. Yiangou, K. A. Markey, J. L. Mitchell, W. J. Scotton, P. A. Keane, and A. J. Sinclair, "Segmentation error in spectral domain optical coherence tomography measures of the retinal nerve fibre layer thickness in idiopathic intracranial hypertension," *BMC Ophthalmol.*, vol. 17, no. 1, Dec. 2017, pp. 1–7, doi: 10.1186/s12886-017-0652-7.

[8] R. A. Alshareef, A. Goud, M. Mikhail, H. Saheb, H. K. Peguda, S. Dumpala, S. Rapole, and J. Chhablani, "Segmentation errors in macular ganglion cell analysis as determined by optical coherence tomography in eyes with macular pathology," *Int. J. Retina Vitreous*, vol. 3, no. 1, Dec. 2017, pp. 1–8, doi: 10.1186/s40942-017-0078-7.

[9] P. J. Patel, F. K. Chen, L. da Cruz, and A. Tufail, "Segmentation error in stratus optical coherence tomography for neovascular age-related macular degeneration," *Investigative Ophthalmol. Vis. Sci.*, vol. 50, no. 1, pp. 399–404, 2009.

[10] S. Sadda, Z. Wu, A. Walsh, L. Richine, J. Dougall, R. Cortez, and L. Labree, "Errors in retinal thickness measurements obtained by optical coherence tomography," *Ophthalmology*, vol. 113, no. 2, pp. 285–293, Feb. 2006.

[11] D. Hanumunthadu, J. P. Wang, W. Chen, E. N. Wong, Y. Chen, W. H. Morgan, P. J. Patel, and F. K. Chen, "Impact of retinal pigment epithelium pathology on spectral-domain optical coherence tomography-derived macular thickness and volume metrics and their intersession repeatability," *Clin. Exp. Ophthalmol.*, vol. 45, no. 3, pp. 270–279, Apr. 2017.

[12] D. C. DeBuc, "A review of algorithms for segmentation of retinal image data using optical coherence tomography," in *Image Segmentation*, vol. 1. Rijeka, Croatia: IntechOpen, 2011, pp. 15–54.

[13] J. Oliveira, S. Pereira, L. Gonçalves, M. Ferreira, and C. A. Silva, "Multi-surface segmentation of OCT images with AMD using sparse high order potentials," *Biomed. Opt. Exp.*, vol. 8, no. 1, pp. 281–297, 2017.

[14] R. Kafieh, H. Rabbani, M. D. Abramoff, and M. Sonka, "Intra-retinal layer segmentation of 3D optical coherence tomography using coarse grained diffusion map," *Med. Image Anal.*, vol. 17, no. 8, pp. 907–928, Dec. 2013.

[15] F. Rathke, S. Schmidt, and C. Schnörr, "Probabilistic intra-retinal layer segmentation in 3-D OCT images using global shape regularization," *Med. Image Anal.*, vol. 18, no. 5, pp. 781–794, Jul. 2014.

[16] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomed. Opt. Exp.*, vol. 6, no. 4, pp. 1172–1194, 2015, doi: 10.1364/BOE.6.001172.

[17] J. Tian, B. Varga, G. M. Somfai, W.-H. Lee, W. E. Smiddy, and D. C. DeBuc, "Real-time automatic segmentation of optical coherence tomography volume data of the macular region," *PLoS ONE*, vol. 10, no. 8, Aug. 2015, Art. no. e0133908, doi: 10.1371/journal.pone.0133908.

[18] K. Li, X. Wu, D. Z. Chen, and M. Sonka, "Optimal surface segmentation in volumetric images—A graph-theoretic approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 119–134, Jan. 2006.

[19] H. Xue, L. Srinivasan, S. Jiang, M. Rutherford, A. D. Edwards, D. Rueckert, and J. V. Hajnal, "Automatic segmentation and reconstruction of the cortex from neonatal MRI," *NeuroImage*, vol. 38, no. 3, pp. 461–477, Nov. 2007.

[20] J. Oliveira, S. Pereira, L. Gonçalves, M. Ferreira, and C. A. Silva, "Sparse high order potentials for extending multi-surface segmentation of OCT images with drusen," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 2952–2955.

[21] S. P. K. Karri, D. Chakraborthi, and J. Chatterjee, "Learning layer-specific edges for segmenting retinal layers with large deformations," *Biomed. Opt. Exp.*, vol. 7, no. 7, pp. 2888–2901, 2016.

[22] A. Montuoro, S. M. Waldstein, B. S. Gerendas, U. Schmidt-Erfurth, and H. Bogunović, "Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context," *Biomed. Opt. Exp.*, vol. 8, no. 3, pp. 1874–1888, 2017.

[23] K. A. Vermeer, J. van der Schoot, H. G. Lemij, and J. F. de Boer, "Automated segmentation by pixel classification of retinal layers in ophthalmic OCT images," *Biomed. Opt. Exp.*, vol. 2, no. 6, pp. 1743–1756, 2011.

[24] P. P. Srinivasan, S. J. Heflin, J. A. Izatt, V. Y. Arshavsky, and S. Farsiu, "Automatic segmentation of up to ten layer boundaries in SD-OCT images of the mouse retina with and without missing layers due to pathology," *Biomed. Opt. Exp.*, vol. 5, no. 2, pp. 348–365, 2014.

[25] A. Lang, A. Carass, M. Hauser, E. S. Sotirchos, P. A. Calabresi, H. S. Ying, and J. L. Prince, "Retinal layer segmentation of macular OCT images using boundary classification," *Biomed. Opt. Exp.*, vol. 4, no. 7, pp. 1133–1152, 2013.

[26] K. McDonough, I. Kolmanovsky, and I. V. Glybina, "A neural network approach to retinal layer boundary identification from optical coherence tomography images," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Aug. 2015, pp. 1–8.

[27] S. Thangaraj, V. Periyasamy, and R. Balaji, "Retinal vessel segmentation using neural network," *IET Image Process.*, vol. 12, no. 5, pp. 669–678, May 2018.

[28] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search," *Biomed. Opt. Exp.*, vol. 8, no. 5, pp. 2732–2744, 2017.

[29] K. Hu, B. Shen, Y. Zhang, C. Cao, F. Xiao, and X. Gao, ''Automatic segmentation of retinal layer boundaries in OCT images using multi-scale convolutional neural network and graph search,'' *Neurocomputing*, vol. 365, pp. 302–313, Nov. 2019.

[30] M. Pekala, N. Joshi, T. Y. A. Liu, N. M. Bressler, D. C. DeBuc, and P. Burlina, ''Deep learning based retinal OCT segmentation,'' *Comput. Biol. Med.*, vol. 114, Nov. 2019, Art. no. 103445, doi: 10.1016/j.compbiomed.2019.103445.

[31] A. Shah, M. D. Abramoff, and X. Wu, ''Simultaneous multiple surface segmentation using deep learning,'' in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017, pp. 3–11.

[32] D. Alonso-Caneiro, J. Kugelman, J. Hamwood, S. A. Read, S. J. Vincent, F. K. Chen, and M. J. Collins, ''Automatic retinal and choroidal boundary segmentation in OCT images using patch-based supervised machine learning methods,'' in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 215–228.

[33] J. Kugelman, D. Alonso-Caneiro, Y. Chen, S. Arunachalam, D. Huang, N. Vallis, M. J. Collins, and F. K. Chen, ''Retinal boundary segmentation in stargardt disease optical coherence tomography images using automated deep learning,'' *Transl. Vis. Sci. Technol.*, vol. 9, no. 11, p. 12, Oct. 2020.

[34] P. Zang, J. Wang, T. T. Hormel, L. Liu, D. Huang, and Y. Jia, ''Automated segmentation of peripapillary retinal boundaries in OCT combining a convolutional neural network and a multi-weights graph search,'' *Biomed. Opt. Exp.*, vol. 10, no. 8, pp. 4340–4352, 2019.

[35] J. Kugelman, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, ''Automatic segmentation of OCT retinal boundaries using recurrent neural networks and graph search,'' *Biomed. Opt. Exp.*, vol. 9, no. 11, pp. 5759–5777, 2018.

[36] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, ''ReLayNet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks,'' *Biomed. Opt. Exp.*, vol. 8, no. 8, pp. 3627–3642, 2017.

[37] F. G. Venhuizen, B. van Ginneken, B. Liefers, M. J. van Grinsven, S. Fauser, C. Hoyng, T. Theelen, and C. I. Sánchez, ''Robust total retina thickness segmentation in optical coherence tomography images using convolutional neural networks,'' *Biomed. Opt. Exp.*, vol. 8, no. 7, pp. 3292–3316, Jul. 2017.

[38] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, ''Understanding deep learning requires rethinking generalization,'' 2016, *arXiv:1611.03530*. [Online]. Available: http://arxiv.org/abs/1611.03530

[39] A. Krogh and J. Vedelsby, ''Neural network ensembles, cross validation and active learning,'' in *Proc. 7th Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 1994, pp. 231–238.

[40] Z.-H. Zhou, J. Wu, and W. Tang, ''Ensembling neural networks: Many could be better than all,'' *Artif. Intell.*, vol. 137, nos. 1–2, pp. 239–263, 2002.

[41] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, ch. 7, pp. 256–258.

[42] Y. Ren, L. Zhang, and P. N. Suganthan, ''Ensemble classification and regression-recent developments, applications and future directions [review article],'' *IEEE Comput. Intell. Mag.*, vol. 11, no. 1, pp. 41–53, Feb. 2016, doi: 10.1109/MCI.2015.2471235.

[43] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL, USA: CRC Press, 1994, ch. 6, pp. 45–56.

[44] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, vol. 26. New York, NY, USA: Springer, 2013, pp. 192–198.

[45] Y. Cao, Q.-G. Miao, J.-C. Liu, and L. Gao, ''Advance and prospects of AdaBoost algorithm,'' *Acta Automatica Sinica*, vol. 39, no. 6, pp. 745–758, Mar. 2014.

[46] A. Natekin and A. Knoll, ''Gradient boosting machines, a tutorial,'' *Frontiers Neurorobot.*, vol. 7, p. 21, Dec. 2013, doi: 10.3389/fnbot.2013.00021.

[47] T. Chen and C. Guestrin, ''XGBoost: A scalable tree boosting system,'' in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[48] A. Lahiri, A. G. Roy, D. Sheet, and P. K. Biswas, ''Deep neural ensemble for retinal vessel segmentation in fundus images towards achieving label-free angiography,'' in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 1340–1343.

[49] Z. Ji, Q. Chen, S. Niu, T. Leng, and D. L. Rubin, ''Beyond retinal layers: A deep voting model for automated geographic atrophy segmentation in SD-OCT images,'' *Transl. Vis. Sci. Technol.*, vol. 7, no. 1, p. 1, Jan. 2018, doi: 10.1167/tvst.7.1.1.

[50] Y. Guo, Ü. Budak, and A. Şengür, ''A novel retinal vessel detection approach based on multiple deep convolution neural networks,'' *Comput. Methods Programs Biomed.*, vol. 167, pp. 43–48, Dec. 2018.

[51] D. H. Wolpert, ''Stacked generalization,'' *Neural Netw.*, vol. 5, no. 2, pp. 241–259, Jan. 1992.

[52] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[53] H. Bogunović *et al.*, ''RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge,'' *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1858–1874, Aug. 2019.

[54] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. New York, NY, USA: Pearson, 2018, pp. 125–129.

[55] NEMA. (2015). *III.6 Retinal Thickness Definition*. Accessed: Feb. 28, 2021. [Online]. Available: http://dicom.nema.org/DICOM/2013/output/chtml/part17/sect_III.6.html

[56] M. K. Garvin, M. D. Abramoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka, ''Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images,'' *IEEE Trans. Med. Imag.*, vol. 28, no. 9, pp. 1436–1447, Sep. 2009.

[57] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, ''SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size,'' 2016, *arXiv:1602.07360*. [Online]. Available: http://arxiv.org/abs/1602.07360

[58] J. Hamwood, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, ''Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers,'' *Biomed. Opt. Exp.*, vol. 9, no. 7, pp. 3049–3066, 2018.

[59] MATLAB. *Pretrained Deep Neural Networks*. Accessed: Feb. 28, 2020. [Online]. Available: https://www.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html

[60] J. Bengio, ''Practical recommendations for gradient-based training of deep architectures,'' in *Neural Networks: Tricks of the Trade* (Lecture Notes in Computer Science), G. Montavon, G. B. Orr, and K. R. Müller, Eds. Berlin, Germany: Springer, 2012, pp. 437–478.

[61] T. M. Breuel, ''The effects of hyperparameters on SGD training of neural networks,'' 2015, *arXiv:1508.02788*. [Online]. Available: http://arxiv.org/abs/1508.02788

[62] L. N. Smith, ''Cyclical learning rates for training neural networks,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 464–472.

[63] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[64] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006, ch. 4, pp. 206–212.

[65] S. Ruder, ''An overview of gradient descent optimization algorithms,'' 2016, *arXiv:1609.04747*. [Online]. Available: http://arxiv.org/abs/1609.04747

[66] O. Ronneberger, P. Fischer, and T. Brox, ''U-Net: Convolutional networks for biomedical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[67] H. Bogunovic, M. Sonka, Y. H. Kwon, P. Kemp, M. D. Abramoff, and X. Wu, ''Multi-surface and multi-field co-segmentation of 3-D retinal optical coherence tomography,'' *IEEE Trans. Med. Imag.*, vol. 33, no. 12, pp. 2242–2253, Dec. 2014.

[68] X. Chen, M. Niemeijer, L. Zhang, K. Lee, M. D. Abràmoff, and M. Sonka, ''Three-dimensional segmentation of fluid-associated abnormalities in retinal OCT: Probability constrained graph-search-graph-cut,'' *IEEE Trans. Med. Imag.*, vol. 31, no. 8, pp. 1521–1531, Aug. 2012.

[69] A. Cazañas-Gordón, E. Parra-Mora, and L. A. da Silva Cruz, ''Evaluating transfer learning for macular fluid detection with limited data,'' in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 1348–1352, doi: 10.23919/Eusipco47968.2020.9287859.

[70] C. Shorten and T. M. Khoshgoftaar, ''A survey on image data augmentation for deep learning,'' *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[71] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, ''Predicting sample size required for classification performance,'' *BMC Med. Informat. Decis. Making*, vol. 12, no. 1, pp. 1–10, Dec. 2012, doi: 10.1186/1472-6947-12-8.

[72] V. Schreur, L. Altay, F. van Asten, J. M. M. Groenewoud, S. Fauser, B. J. Klevering, C. B. Hoyng, and E. K. de Jong, ''Hyperreflective foci on optical coherence tomography associate with treatment outcome for anti-VEGF in patients with diabetic macular edema,'' *PLoS ONE*, vol. 13, no. 10, Oct. 2018, Art. no. e0206482, doi: 10.1371/journal.pone.0206482.

ALEX CAZAÑAS-GORDON received the B.E. degree in electrical engineering from National Polytechnic School, Quito, Ecuador, in 2003, and the M.Sc. degree in information technology from The University of Queensland, Brisbane, QLD, Australia, in 2015. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Coimbra, Coimbra, Portugal.

Since 2018, he has been a Researcher with the Multimedia Signal Processing Laboratory, Department of Electrical and Computer Engineering, University of Coimbra. His research interests include signal processing, deep learning, optical coherence tomography, scanning laser ophthalmoscopy, and fundus photography.

ESTHER PARRA-MORA received the bachelor's degree in electronics and information networks from National Polytechnic School, Quito, Ecuador, in 2007, and the master's degree in computer science from The University of Queensland, Brisbane, QLD, Australia, in 2015. She is currently pursuing the Ph.D. degree with the University of Coimbra, Coimbra, Portugal.

Since October 2017, she has been a Researcher with the Department of Electrical and Computer Engineering, University of Coimbra. Her research interests include automatic diagnosis of retinal diseases using deep learning techniques and different modalities of retinal images.

LUÍS A. DA SILVA CRUZ (Senior Member, IEEE) received the Licenciado and M.Sc. degrees in electrical engineering from the University of Coimbra, Portugal, in 1989 and 1993, respectively, and the M.Sc. degree in mathematics and the Ph.D. degree in electrical computer and systems engineering from Rensselaer Polytechnic Institute (RPI), Troy, NY, USA, in 1997 and 2000, respectively. Since 1990, he has been a Teaching Assistant with the Department of Electrical and Computer Engineering, University of Coimbra, and as an Assistant Professor, since 2000. He is currently a Researcher with the Institute for Telecommunications of Coimbra, where he works on image and video processing and coding and medical image processing. He is also a member of the EURASIP, SPIE, and IEEE technical societies.