Macular OCT Classification Using a Multi-Scale Convolutional Neural Network Ensemble

Reza Rasti, *Student Member, IEEE*, Hossein Rabbani, *Senior Member, IEEE*, Alireza Mehridehnavi, *Member, IEEE*, and Fedra Hajizadeh

***Abstract*— Computer-aided diagnosis (CAD) of retinal pathologies is a current active area in medical image analysis. Due to the increasing use of retinal optical coherence tomography (OCT) imaging technique, a CAD system in retinal OCT is essential to assist ophthalmologist in the early detection of ocular diseases and treatment monitoring. This paper presents a novel CAD system based on a multi-scale convolutional mixture of expert (MCME) ensemble model to identify normal retina, and two common types of macular pathologies, namely, dry age-related macular degeneration, and diabetic macular edema. The proposed MCME modular model is a data-driven neural structure, which employs a new cost function for discriminative and fast learning of image features by applying convolutional neural networks on multiple-scale sub-images. MCME maximizes the likelihood function of the training data set and ground truth by considering a mixture model, which tries also to model the joint interaction between individual experts by using a correlated multivariate component for each expert module instead of only modeling the marginal distributions by independent Gaussian components. Two different macular OCT data sets from Heidelberg devices were considered for the evaluation of the method, i.e., a local data set of OCT images of 148 subjects and a public data set of 45 OCT acquisitions. For comparison purpose, we performed a wide range of classification measures to compare the results with the best configurations of the MCME method. With the MCME model of four scale-dependent experts, the precision rate of 98.86%, and the area under the receiver operating characteristic curve (AUC) of 0.9985 were obtained on average.**

***Index Terms*— CAD system, classification, macular pathology, Multi-scale Convolutional Mixture of Experts (MCME), Optical Coherence Tomography (OCT).**

I. Introduction

THE retina in human eyes receives the focused light by the lens, and converts it into neural signals. The main sensory region for this purpose is the macula which is located in the central part of the retina. The macula processes light through special layers of photoreceptor nerve cells which are responsible for detecting light-intensity, color, and fine visual details. The retina processes the information gathered by the macula, and sends them to the brain via the optic nerve for visual recognition. In fact, necessary features of visual perception are traced to the retinal processing of encoding light into neural signals within the macula region.

The macula healthiness can be affected by a number of pathologies, including age-related macular degeneration (AMD), and diabetic macular edema (DME). AMD is a retinal disease resulting in blurred, blind spots or even no vision in the center of the visual field and it was the fourth most common cause of blindness in 2013 [1]. According to [2], about 0.4% of people between 50-60 and approximately 15% of people over 60 years old are suffering from AMD pathology. Moreover, diabetic retinopathy accounts for working-age blindness. In the United States, 12% of all new cases of blindness (people aged between 20-64 years) in each year is due to diabetic retinopathy [3]. Generally, DME is the most common diabetic cause of vision loss in different societies [4]. In the early stages of retinopathy, central vision may be affected by DME. In diabetic patients, particularly in type II ones, DME is the most frequent sight-threatening complication [5].

It has been shown that the blindness rate would be reduced by comprehensive screening programs, and early treatment of the eyes with effective diagnostic tools [1]. In ophthalmology, one of the most commonly used imaging technique is optical coherence tomography (OCT) with more than 5 million acquisitions in the US in 2014 [6]. OCT is a non-invasive imaging technique, which captures cross-sectional images at microscopic resolution of biological tissues [7]. This developing and high-speed diagnostic imaging technology has a major contribution to the early identification and treatment of retinal pathologies today.

Since 3-D OCT image interpretation is a time-consuming and tedious process for ophthalmologists, different computer-aided diagnosis (CAD) systems for semi/fully automatic analysis of OCT data have been developed during current years [8]–[27]. To this end, various groups have developed computerized algorithms for pre-processing, including denoising and curvature correction [8]–[11], intra-retinal and pathological area segmentation [12]–[18], and 2-D or 3-D OCT

Manuscript received October 3, 2017; revised November 27, 2017; accepted December 1, 2017. Date of publication December 6, 2017; date of current version April 2, 2018. This work was supported by the Department of Biomedical Engineering, Isfahan University of Medical Sciences, under Grant 395645. *(Corresponding author: Hossein Rabbani.)*

R. Rasti, H. Rabbani, and A. Mehridehnavi are with the Medical Image and Signal Processing Research Center, Department of Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan 8174673461, Iran (e-mail: mr.r.rasti@ieee.org; rabbani.h@ieee.org; mehri@med.mui.ac.ir).

F. Hajizadeh is with the Noor Ophthalmology Research Center, Noor Eye Hospital, Tehran 1968653111, Iran (e-mail: fedra_hajizadeh@yahoo.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2017.2780115

TABLE I
RECENT WORKS ON CAD SYSTEMS IN RETINAL OCT IMAGING

Author	Database	Method	Result	Notes
Liu et, al. [19]	Cirrus HD-OCT: 326 cases (4 classes: AMD, MH, ME, and Normal)	Retinal layers alignment + Multi-scale LBP feature extraction + RBF-SVM classification	AUC=0.93 based on the 10-fold CV	The analysis is limited to a single foveal slice which is manually selected by expert ophthalmologists.
Albarrak et, al. [20]	3-D OCT: 140 cases (2 classes: AMD and Normal)	B-scan denoising + LBP and HOG feature extraction + PCA + Bayesian classification	AUC=0.944 based on 10-fold CV	The presented method relies on a denoising step.
Farsiu et, al. [21]	Bioptigen SD-OCT: 384 cases (2 classes: AMD and Control)	A semi-automatic segmentation of BM, ILM, and RPE layers + Manually feature extraction + Linear regression	AUC=0.99 based on leave-one-out CV	The algorithm relies on a precise segmentation of retinal layers and requires manual corrections to avoid misleading outcomes.
Srinivasan et, al. [22]	Heidelberg SD-OCT: 45 cases (3 classes: AMD, DME, and Normal)	B-scan denoising + Retinal layers alignment + Multi-scale HOG feature extraction + Linear SVM classification	CR of 86.67%, 100%, and 100% for Normal, AMD, and DME based on leave-three-out CV	The method relies on a denoising step. Also, the case results depend on a threshold of 33% of the B-scans in the volumes as abnormal ones.
Venhuizen et, al. [23]	Bioptigen SD-OCT: 384 cases (2 classes: AMD and Control)	An unsupervised feature learning based on Bag-of-Words (BoWs) method + Random forests (RF) classification	AUC=0.984	The method does not need accurate layer segmentation.
Lematre et, al. [24]	Cirrus,SD-OCT: 32 cases (2 classes: DME and Normal)	NLM filtering + Retinal layers flattening + LBP-TOP feature extraction + Local mapping + BoWs features + RBF-SVM classification	Sensitivity=81.2% and Specificity=93.7% based on leave-one-out CV	The method relies on a denoising step. Indeed, the algorithm results did not report for any AMD data (the most relevant diagnostic problem in retinal OCT).
Apostolopoulos et, al. [25]	Bioptigen SD-OCT: 384 cases (2 classes: AMD and Control)	B-scan mosaicking of 3D OCT data + OCT classification by a 2-D deep CNN model (RetiNet C)	AUC=0.997 based on five-fold CV	The model is limited to 3D OCT data with a same number of B-scans in different volumes.
Venhuizen et, al. [26]	Heidelberg Spectralis HRA-OCT: 3265 eyes (5 classes: No AMD, Early AMD, Intermediate AMD, Advanced AMD GA, Advanced AMD CNV)	OCT volume resampling + Detection of AMD affected regions + BoWs feature learning + Multi-class RF classification	AUC=0.980 with a sensitivity of 98.2% at a specificity of 91.2%.	The method does not need accurate layer segmentation. To automatically identify AMD affected regions, a simple and coarse layer segmentation method is needed.
Sun et, al. [27]	Heidelberg SD-OCT: Set1 [22] composed of 45 OCT volumes and Set2 included of 678 retinal B-scans (3 classes: AMD, DME, and Normal)	B-scan denoising + Retinal layers alignment + Image partitioning + SIFT feature extraction + Dictionary learning and sparse coding + Multi-scale max pooling + Linear SVM classification	CR of 93.33%, 100%, 100% on set1 and CR of 100%, 99.67%, 99.67% on set2 for Normal, AMD, and DME respectively.	The method relies on a denoising step. Three 2-class SVM are used in max-out strategy and the results reported based on leave-three-out CV. For volumetric diagnosis on dataset1 the maximum vote strategy is chosen.

ME=Macular Edema, **MH**=Macular Hole, **CV**=Cross-Validation, **CR**=Classification Rate, **BM**=Bruch's Membrane, **ILM**=Inner Limiting Membrane, **RPE**=Retinal Pigmented Epithelium, **GA**=Geographic Atrophy, **CNV**=Choroidal Neovascularization.

classification [19]–[27]. Table I summarizes the recent works on CAD systems in retinal OCT. The table highlights different aspects or strengths of the reviewed works.

In the present study, we propose a novel CAD system for automatic macular OCT classification. The proposed system consists of two main steps. First, in the preprocessing step, a graph-based curvature correction algorithm is used to remove the retinal distortions, and to yield a set of standard region/volume of interests (ROIs/VOIs). Second, in the classification step, a data-driven representation solution with a full-training approach is introduced and used. In this step, the system includes a new deep ensemble method based on convolutional neural networks (CNNs) [28]–[31]. The presented method uses the idea of a mixture of multi-scale CNN experts as a robust combining method. Generally, the goal of combining methods is to improve the performance in prediction and classification tasks particularly for complicated problems involving limited number of patterns, high dimensional feature sets, and highly overlapping classes [32]–[34].

The proposed ensemble model of multi-scale convolutional mixture of expert (MCME) is a fast and scale-dependent CNN-based classifier which employs a prior decomposition, and a new cost function for discriminative and fast learning of image representative features.

Compared to the reviewed works in Table. I, the main contributions of the present research are: (i) to analyze the proposed MCME model to address a fully automatic classification approach with minimum pre-processing requirements, (ii) the capability of multi-slice analysis of volumetric OCTs, including different slicing and acquisition protocols, (iii) promising sensitivity, and (iv) robustness in the retinal OCT diagnostic problem.

Basically, the proposed CAD system is designed to analyze the macular OCT volumes. For this purpose, the MCME model in the classification stage performs a slice-based analysis by assessing the retinal B-scans in volumes and making a final diagnosis decision on subjects (cases). There are two main reasons for choosing this strategy: (i) slice-based analysis of retinal OCT volumes is the clinical routine in ophthalmology, and (ii) different imaging protocols in retinal OCT (as we see in the datasets) don't yield consistent slicing and unique volume sizes to design a full 3-D diagnostic system.

The outline of this paper is organized as follows: Section II describes the database and proposed MCME classification method used in this study. In Section III the ability of proposed method is investigated and results are presented. Section IV presents a comprehensive discussion of the method and experimental studies, and finally, this paper is concluded in Section V.

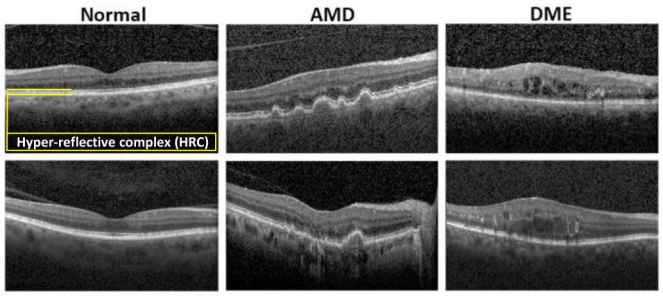

Fig. 1. Example B-scans from Normal, AMD, and DME subjects in dataset 1 (Top row) and dataset 2 (Bottom row).

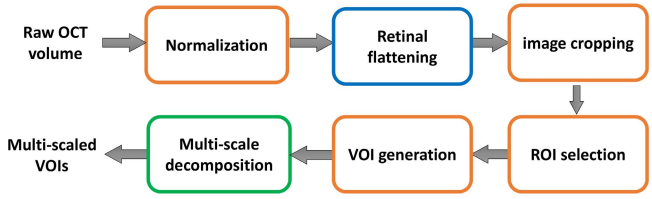

Fig. 2. General pipeline of the preprocessing algorithm.

II. Material and Methods

A. Database

For this research study, the proposed algorithm was designed and evaluated on two different datasets acquired by Heidelberg SD-OCT imaging systems. The first one is acquired at Noor Eye Hospital in Tehran consisting of 50 normal, 48 dry AMD, and 50 DME OCTs. For this dataset, the axial resolution is $3.5\mu m$ with the scan-dimension of $8.9\times7.4\ mm^2$, but the lateral and azimuthal resolutions are not consistent for all patients. So, the number of A-scans varies among 512 or 768 scans where 19, 25, 31, and 61 B-scans per volume are acquired from different patients. The dataset is available at http://www.biosigdata.com.

The second one is a publicly available dataset containing 45 OCT acquisitions obtained at Duke University, Harvard University, and the University of Michigan [22]. This dataset consists of volumetric scans with non-unique protocols of normal, dry AMD, and DME classes which includes 15 subjects for each class. Fig. 1 displays example B-scans from different SD-OCT volumes of each class.

In addition to labeling at the patient level, all the 4142 B-scans in dataset 1, and 3247 B-scans in dataset 2 were annotated by an ophthalmologist experienced in OCT imaging to train the proposed 2-D ensemble models. In total, for dataset 1, the labeled B-scans consist of 862 DME and 969 AMD B-scans. These labeled samples are 856 DME and 711 AMD B-scans for dataset 2, where the other B-scans are considered as normal ones.

B. Data Preprocessing and VOI Extraction

A general overview of the preprocessing algorithm is shown in Fig. 2. The algorithm consists the following steps.

1) Normalization: To obtain a unique field of view for OCT images, all the B-scans in different volumes are resized to 496 × 512 pixels. In addition, in order to handle the intensity variations in OCT images from different patients, a normalization step is done to remove the intensity mean value of each B-scan (zero-mean), and to scale it to have a standard deviation of one.

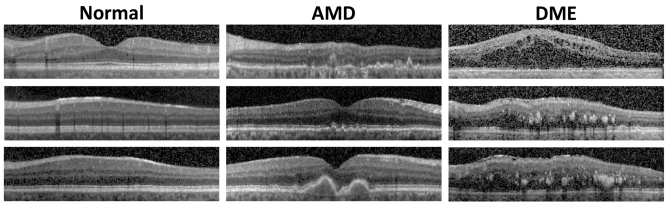

Fig. 3. Output samples of the image cropping block. The flattened and cropped B-scans belong to the dataset 1. The image size of the flattened outputs is 128 × 512 pixels.

2) Retinal Flattening Algorithm and Image Cropping: In OCT images, due to the anatomical structures and acquisition distortions, the retinal layers in B-scans may be shifted or oriented randomly. Consequently, it causes high variability in locations in B-scans. In order to deal with this issue, a graph based geometry curvature correction algorithm [11] is used for the retinal flattening based on the detection of the hyper-reflective complex (HRC) band in the retina without any denoising consideration (see Fig. 1). The main idea behind this algorithm is the construction of graph nodes from each pixel of the OCT image. Each node is defined in three-dimensional space determined by normalized intensity, horizontal value and vertical value of each pixel. The graph edges are then created according to measures of brightness similarity and closeness, after using a pipeline of morphological operations and candidate region sorting. These processes are designed to reduce the graph size to restrict the graph point selection to an area around the HRC. After generating a matrix of connectivity, a random walk method is then employed to jump from one node to another. Finally, based on fitting a second-order polynomial on the detected HRC, retinal boundary points are shifted vertically so that HRC points can lie on a horizontal band.

Using this algorithm, all the B-scans are flattened to conquer the misalignments of retinal layers. To this aim, in each flattened image, the estimated HRC is warped to a vertical line that is located at 70% of the image height. Also, in order to focus on the region of the retina that contains main morphological structures and to reduce the dimension of the image, at first, each B-scan is cropped vertically with considering 200 pixels above and 35 pixels below the estimated HRC. These values were selected for cropping step via a visual inspection on both datasets to preserve all the retinal B-scan information. Finally, each cropped SD-OCT image is resized to 128 × 512 pixels for further processes. Fig. 3 demonstrates examples of the retinal flattening and cropping steps.

3) ROI Selection, VOI Generation, and Augmentation Strategy: In this step, the ROI is selected by cropping a centered 128 × 470 pixels bounding box from each B-scan in a given case. In next step, all the extracted ROIs in all B-scans are resized to 128 × 256 pixels and concatenated to generate the

case VOI. In learning phase and for training VOIs, in order to have an efficient training process, the centered bounding box is horizontally flipped and/or translated by ±20 pixels to generate an augmented training collection of ROIs. This strategy increases the number of samples with a factor of six in our training dataset, reduces the chance of over-fitting, and also degrades inconsistency in the data due to a different number of right and left eyes [25].

4) Multi-Scale Spatial Decomposition: According to the following motivations, the multi-scale spatial pyramid (MSSP) decomposition [35] is applied to macular OCT B-scans before feeding them to convolutional ensemble models: (i) some retinal pathologies such as DME exhibit key characteristics at different scales; therefore, it is expected that multi-scale views of the retina should be presented to the model [19], and (ii) although CNN-based algorithms benefit from spatial pooling for providing some built-in invariance to distorted, scaled, and translated inputs [36], but in the proposed ensemble model, the prior MSSP decomposition can be used to reduce the time complexity and effective parameters of the overall model to reduce the chance of over-fitting and to obtain a promising performance in practice. Therefore, to assess the main hypothesis of the study, four levels of the multi-scale version of Gaussian low-pass image pyramids (i.e., $l_0 \sim l_3$) are considered to simulate the multi-view perception of the proposed model. To do this, the image pyramids are calculated for each slice within the VOIs using a symmetric pyramidal decomposition method.

Suppose that the B-scan I is represented by a 2-D array with C columns and R rows of pixels. This image is the zero level (l_0) of the Gaussian pyramid. Pyramid level l contains image I_l, which is a reduced or low-pass filtered version of I_{l-1}. Each value within level I_l is computed as a weighted average of values in level I_{l-1} within a 5×5 window [35]. Therefore, the different level pyramids are calculated by:

$$\mathbf{I}_l(i,j) = \sum_{m=-2}^{2} \sum_{n=-2}^{2} \mathbf{W}(m,n) \cdot \mathbf{I}_{l-1}(2i+m, 2j+n), \quad (1)$$

Here l refers to the level of pyramid, and positive integers of i and j are respectively less than C_l and R_l, where C_l and R_l are the dimensions of the l^{th} scale. In this study, the separable kernel $W(m,n) = w(m).w(n)$ with $w = [(1/4) - (a/2), 1/4, a, 1/4, (1/4) - (a/2)]$ and $a = 0.375$ are used for simulation.

C. Classification Methodology

In the present work, a new modular ensemble method based on a mixture of CNN experts is used as a robust image-based classifier. This combined model is originated from the concept of divide-and-conquer approach in machine learning literature. It benefits from multi-view decomposition of input patterns to fuse key information and to solve the classification problem in a sparse and efficient manner. In the following sub-sections, the MCME model is presented in detail.

1) Mixture of Experts (ME) Background: The traditional ME structure was introduced by Jordan and Jacobs in 1991 [37], and extended by several research groups later for combining the outputs of multi-layered perceptrons (MLPs) [38]–[40].

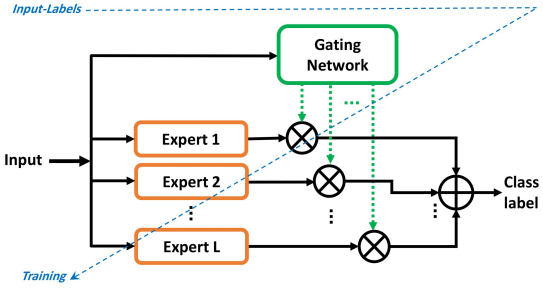

Fig. 4. The conventional mixture of L experts (classifiers) structure: a common signal supplies the input of all modules i.e., the experts and gating network.

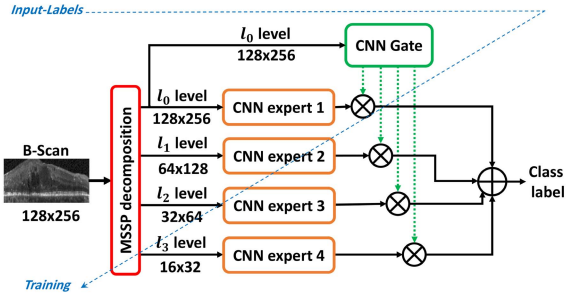

Fig. 5. The MCME structure: the CNN experts and gating network are fed by specific scales of the input pattern.

As illustrated in Fig. 4, this model combines the outputs of several expert (classifier) networks by training a gating network. Given an input pattern, each expert network estimates the conditional posterior distribution on the partitioned feature space which is separated by a gating network. In fact, the gating network plays a weighting role and causes the overall model to execute a competitive learning process for experts' modules [37]. To do that, this module tries to maximize the likelihood function of the training dataset and ground truth by considering a Gaussian mixture model (GMM), in which each Gaussian component in the mixture model corresponds to one expert module [41], [42].

2) The MCME Model: In the proposed MCME method for retinal classification, the experts and gating network are constructed by CNNs. The scale-dependent modules used in MCME enable the model to have a multi-scale perception of input patterns inspired from visual attention systems. So, with a symmetric Gaussian decomposition of the input ROIs, specific pyramidal scales (views) are fed to regular CNNs' experts and convolutional gating network (CGN). In contrast to the traditional ME, suggesting a prior decomposition of the inputs can be useful for dividing the task among simpler experts in MCME, where the complexity of the overall model would be reduced. This is possible by selective information fusion of different scales with the gating network. The proposed MCME model is depicted in Fig. 5. It is noted that all modules in this structure are learned simultaneously rather than independently or sequentially based on an end-to-end optimization procedure. This strategy provides an ability for the modules to interact with each other and to specialize in different parts of automatically extracted feature space.

a) MCME signal forward propagation: The MCME model learns a set of scale-dependent convolutional expert networks f_i with scale $l_{i-1} : i = 1, \ldots, L$ (in which each expert has its own specific input scale) along with a CGN g with original input scale (l_0). Each f_i maps the scaled input x_i to C (one output for each class: AMD, DME, and Normal), while $g_i(x_0)$ is a conditional posterior distribution over CNN experts. The CGN has a Softmax output layer. It assigns a probabilistic weight g_i to each expert's output f_i. The g_i fusion weights are optimized in the learning process where the CGN is simultaneously and interactively trained with the other individual experts to adaptively predict the contribution of each scale-dependent CNN experts by encoding the l_0-scale input into a vector of g_i values. The final output of the entire MCME structure is then given by Eq. 2 for k^{th} input image:

$$\begin{aligned}\mathbf{F}_{MCME}(x^k) &= \sum_{i=1}^{L} g_i(x_0^k) \cdot f_i(x_i^k), \\ &= \sum_{i=1}^{L} P(e_i|x_0^k) \cdot P(c|e_i, x_i^k), \\ &= P(c|x^k). \end{aligned} \tag{2}$$

where L is the number of CNN experts (different scales) in the model, e_i indicates the i^{th} expert module, and c is the output class label.

b) MCME error back propagation: The traditional mixture of experts method can train de-correlated individual experts modules in a suitable fashion, and benefits from it [43], but the ME error back propagation procedure does not include any control parameter to keep the experts uncorrelated. Inspired from [44], to control and monitor this ability of the ME method in the training algorithm, we added a cross-correlation penalty term to MCME error cost function. Therefore, total error cost function of MCME for k^{th} input pattern is defined as:

$$E_{MCME}(x^k) = -\ln\left(\sum_{i=1}^{L} g_i(x_0^k) \cdot e^{\left(-\frac{1}{2}\|\mathbf{d}^k - \mathbf{f}_i(x_i^k)\|^2 + \lambda \cdot \rho_i^k\right)}\right), \tag{3}$$

where $\mathbf{f}_i$, $\mathbf{d}$, and ρ_i are output vectors of CNN expert i, the desire output of the input sample x^k, and the cross-correlation penalty term, respectively. Here, ρ_i is defined as follows:

$$\rho_i^k = \frac{1}{L-1} \sum_{j=1; j\neq i}^{L} \left(\mathbf{f}_i(x_i^k) - \mathbf{F}_{MCME}(x^k)\right) \cdot \left(\mathbf{f}_j(x_j^k) - \mathbf{F}_{MCME}(x^k)\right)^T, \tag{4}$$

The strength of this penalty is adjusted explicitly with the control parameter $0 \leq \lambda \leq 1$. It is obvious that for $\lambda = 0$ there is no cross-correlation penalty term in the MCME cost functions.

From the technical point of view, the purpose behind adding ρ penalty term is to negatively correlate each CNN expert's error with the other CNN experts' errors in the MCME ensemble. As explained before, ME model tries to maximize the likelihood function of the training dataset and ground truth by considering GMM, in which each expert module is modeled by an independent multivariate Gaussian. However, in MCME, correlated multivariate components are employed which also model the joint interaction between individual experts instead of only modeling the marginal distributions.

By calculating the error gradient for all free parameters of the experts and gating networks, different optimization methods can be used for learning the MCME structure. In this study, the mini-batch root mean square propagation (RMSprop) procedure [45] is used for parameters updating. In the present research in order to classify the input VOIs as AMD, DME or Normal cases, we hypothesize that an adequate number of scale-dependent CNN experts and a CGN (with l_0 input scale) have the potential to execute a competitive feature representation process and to yield an efficient combination of key intensity, shape, texture, and context features with reserving speed considerations. So, with extending scale fusion, if the MCME modules split the learned feature space properly, the overall classification rate will be increased.

III. Experimental Study and Results

A. Performance Measures

Classification performance in this problem was computed based on the following evaluation measures. According to the 3-classes confusion matrix and receiver operating characteristic (ROC) analyses, the values of precision, recall, F1-score, and average area under the ROC curves (AUC) are used for performance evaluation of all implemented structures at the patient level.

For the evaluated ensemble models, in order to explore the ability of the experts to partition the feature space of the problem, average correlation coefficient, Cohen's kappa (κ), and distance-based disagreement (DbD) factors are considered. Direction and strength of a linear relationship between CNN experts in the model can be indicated by the correlation coefficient. Indeed, κ is a statistical measure of classifiers agreement that expresses the level of agreement ($-1 \leq \kappa \leq 1$) between each pair of different estimators on a classification problem [46]. The value of 1 means a complete agreement (the lower the κ, the less the agreement). Moreover, the DbD factor represents experts' disagreements in which the confusion matrices are used to compute distances for each individual experts [47]. For the MCME model with experts $1, 2, \ldots, L$, a distance measure D^l between expert l and all other experts is calculated by Eq. 5 in which $cm_{i,j}^l$ are the confusion matrix elements for expert l:

$$D^l = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} |cm_{i,j}^l - \sum_{k=1; k\neq l}^{L} cm_{i,j}^k|, \quad l = 1, 2, \ldots, L. \tag{5}$$

In the above formula, all confusion matrices are considered with N classes. Therefore, for the MCME model with L experts, the DbD measure ($DbD \geq 0$) can be expressed by the following definition:

$$DbD = \frac{1}{L} \sum_{l=1}^{L} D^l, \tag{6}$$

TABLE II
DETAILS OF SINGLE-SCALE CONVOLUTIONAL EXPERT NETWORKS STRUCTURES

Module	Input scale	Input size	Number of layers	First convolutional mask size	Other convolutional mask size	Max-pooling size	Number of FMs in C and P layers	Number of FC1 neurons	Number of FC2(Output) neurons	Number of parameters
CNN1	$l0$	128×256	19	5×5	3×3	2×2	3	15	3	2993
CNN2	$l1$	64×128	16	5×5	3×3	2×2	3	15	3	1901
CNN3	$l2$	32×64	13	5×5	3×3	2×2	3	15	3	1381
CNN4	$l3$	16×32	10	5×5	3×3	2×2	3	15	3	997

Note: FM is the number of feature maps. C, P and FC indicate the convolutional, pooling, and fully-connected layers respectively.

In the DbD formula, confusion matrix information of individual experts is used to express average experts' disagreement. A higher DbD value indicates more disagreement in class prediction for experts.

B. Validation and Diagnostic Strategy

In this study, to evaluate and generalize the performance of the model to an independent dataset, the unbiased 5-fold cross-validation method is considered and applied at the patient level (OCTs). For this purpose, SD-OCT volumes of each dataset in the study are first stratified and partitioned into 5 equally sized folds to ensure that each fold is a good representative of the whole. Subsequently, 5 iterations of training and validation are performed such that within each iteration a different fold of the data is held out for validation while the remaining 4 folds are used for the learning process.

Furthermore, to perform a balanced learning process, an approximately comparable number of normal B-scans with AMD or DME sets are considered randomly for selected volumes in each dataset. Also, any dataset-dependent bias is removed by preserving the random seed across all iterations.

As mentioned before, the learning process is performed based on the selected cases in training folds with considering the augmentation method for their B-scans. Finally, the diagnosis decision by a trained model for test VOIs is obtained by the following role: in a given 3-D volume if more than 15 percent of B-scans ($\tau = 15\%$) are predicted as abnormal ones, the maximum probability of B-scans' votes (according to AMD or DME likelihood scores) determines the type of patient retinal disease. This threshold value will be further evaluated in Sub-section III-E.

C. Study Design

1) Baseline Studies: In this study, to illustrate the proficiency of the proposed strategy as well as to get a criterion for the comparison of MCME's performance and complexity in retinal OCT classification, the following baselines are evaluated:

- Feature-based HOG-LSVM classification method [22].
- Convolutional non-ensemble methods:
  - Off-the-shelf competitive CNNs: The VGG19 [48], ResNet50 [49], and InceptionV3 [50] models are evaluated for the comparison purpose. To do this, and for each considered deep network, after modification of the l_0 scale according to the input receptive field of the model, the CNN codes are extracted as the visual features. The extracted features are then classified by a Softmax classifier including three output neurons and an optimized dropout factor of 30% for its visible layer. Here, the input modification includes: (i) l_0 resizing according to the network input size (e.g., 224×224 pixels for the VGG19 model), and (ii) duplication of the gray channel for 3 times to construct the RGB input (e.g., $224 \times 224 \times 3$ input dimensions for the VGG19 model).
  - Single-scale CNNs: 4 different structures are considered according to Table II to see single-scale CNN classification performance. All of the CNNs are constructed based on sequences of CONV-BN[1]-POOL combination in hidden layers and ended up with two stacks of fully-connected (FC)-BN layers, including 15 and 3 output neurons, respectively. In order to reduce over-fitting probability during the learning process, an optimized dropout factor of 70% is considered for all FC1 layers too.
- Convolutional ensemble methods:
  - Ave-ensemble model: A combination method commonly used in the convolutional ensemble literature is the averaging of different CNN subnetworks in the output layers. Generally, this method trains several CNNs from the available ground truth labels. For a given input image, the real outputs of the trained CNNs are averaged to generate the final output of the ensemble [52], [53]. To consider this method, we used 4 different scales (i.e., l_0 to l_3) and the CNNs in Table II. This ensemble method is called as the Ave-ensemble model in the rest of this paper.
  - Full-rank convolutional mixture of experts (FCME) model: To get better insight into the proposed prior decomposition in convolutional ME model [54], and also the proposed cost function, FCME model is considered and analyzed (see Fig. 4). Following this purpose, the FCME structure is investigated using the full-rank combination of 2, 3, and 4 similar experts with l_0 input scale (i.e., CNN1 in Table II).

2) Characterization of the MCME Model: This experiment is designed to assess the potential of proposed MCME model with considering different scales. To this end, the number of experts (scales) influencing the performance of

[1]Batch normalization layer [51].

TABLE III

DETAILS AND THE AVERAGE PERFORMANCE OF THE BASELINES AND THE PROPOSED STRUCTURES ON DATASET 1 ACCORDING TO THE 5-FOLD CROSS-VALIDATION, THE THRESHOLD OF 15% FOR DECISION MAKING, AND OPTIMUM λ VALUES FOR ME MODELS

Method	Configuration	Best λ	Performance								
			Precision(%)	Recall(%)	F1(%)	AUC	MSE	Correlation	DbD	Kappa	Tr.time(s/ROI)
Feature-based	HOG+LSVM [22]	–	85.35±9.51	82.56±11.2	82.09±11.1	0.903	0.311	–	–	–	–
Off-the-shelf CNN models	VGG19 [48]	–	92.65±4.00	91.39±5.69	91.27±5.65	0.935	0.224	–	–	–	–
	ResNet50 [49]	–	95.31±3.40	94.67±4.00	94.62±3.99	0.960	0.115	–	–	–	–
	InceptionV3 [50]	–	93.32±2.97	92.06±3.97	91.80±4.24	0.941	0.263	–	–	–	–
Single-scale CNNs	CNN1 (l_0)	–	97.50±1.31	97.33±1.33	97.28±1.38	0.991	0.027	–	–	–	0.086
	CNN2 (l_1)	–	96.95±2.11	96.63±2.11	96.55±2.35	0.995	0.034	–	–	–	0.041
	CNN3 (l_2)	–	95.58±6.21	93.32±10.1	92.06±12.7	0.986	0.068	–	–	–	0.026
	CNN4 (l_3)	–	78.80±17.3	77.17±7.42	73.55±12.4	0.963	0.351	–	–	–	0.015
Ave-ensemble model [52], [53]	$l_3-l_2-l_1-l_0$	–	96.45±0.98	95.96±1.33	95.95±1.33	0.994	0.040	0.31	0.67	0.77	0.169
FCME-model [54]	l_0-l_0	0.1,0.2	98.24±1.48	97.97±1.63	98.01±1.64	0.994	0.020	−0.03	0.34	0.57	0.150
	$l_0-l_0-l_0$	0.1	98.83±1.48	98.64±1.63	98.67±1.64	0.992	0.013	0.11	0.58	−0.02	0.193
	$l_0-l_0-l_0-l_0$	0.3	97.05±2.76	96.62±2.98	96.57±3.13	0.993	0.034	0.02	0.77	0.07	0.265
MCME-model	l_1-l_0	0.4	98.83±1.48	98.66±1.63	98.68±1.64	0.995	0.013	0.04	0.51	0.41	0.120
	l_2-l_0	0.2,0.5,0.7	98.24±1.48	97.97±1.63	98.01±1.64	0.994	0.020	−0.05	0.70	0.26	0.109
	l_2-l_1	0.4	97.93±3.01	97.29±3.89	97.39±3.81	0.994	0.027	0.03	0.52	0.39	0.088
	l_3-l_0	0.3	98.34±2.25	97.99±2.67	98.01±2.69	0.992	0.020	0.07	0.56	0.34	0.115
	l_3-l_1	0.3	98.21±1.48	97.97±1.63	97.99±1.64	0.995	0.040	−0.01	0.62	0.29	0.107
	l_3-l_2	0.2	98.34±2.25	97.96±2.67	97.99±2.69	0.994	0.020	−0.04	0.98	0.14	0.085
	$l_2-l_1-l_0$	0.1	98.83±1.48	98.64±1.63	98.67±1.64	0.997	0.013	0.03	0.53	0.13	0.143
	$l_3-l_1-l_0$	0.1	98.83±1.48	98.64±1.63	98.67±1.64	0.995	0.013	−0.01	0.47	0.30	0.139
	$l_3-l_2-l_0$	0.4	98.24±1.48	97.97±1.63	98.01±1.64	0.992	0.020	0.07	0.51	0.09	0.127
	$l_3-l_2-l_1$	0.6	97.76±2.11	97.30±2.49	97.38±2.45	0.993	0.027	−0.06	0.51	0.30	0.101
	$l_3-l_2-l_1-l_0$	0.2	99.39±1.21	99.36±1.33	99.34±1.34	0.998	0.006	−0.04	0.79	0.03	0.170

Note: l_i indicates the MSSP decomposition level of the input ROI for scale-dependent CNNs in models.

convolutional mixture structure is investigated versus simple experts. Following this purpose, to get better insight into the MCME performance in the classification of retinal pathologies in the datasets, a low to high-resolution strategy is performed. Therefore, the single-scale CNNs are considered for expert modules (according to Table II). Moreover, the CGN module is designed to have a similar topology such as CNN1 but its output layer contains 2, 3 or 4 Softmax neurons according to 2-, 3-, or 4-scale MCMEs, respectively.

Consequently, the MCME structure is investigated by testing any combination of 2, 3, or 4 experts (scales). According to a grid search on a nested 5-fold cross-validation within the folds training set, parameter optimization is carried out considering the precision metric. The considered structures are:

- Two-scale MCME: 6 different structures of two-scale MCME are considered, i.e., $l_3-l_2, l_3-l_1, l_3-l_0, l_2-l_1, l_2-l_0$, and l_1-l_0.
- Three-scale MCME: 4 different structures of three-scale MCME are explored, i.e., $l_3-l_2-l_1, l_3-l_2-l_0, l_3-l_1-l_0$ and $l_2-l_1-l_0$.
- Four-scale MCME: the complete combination of 4 experts ($l_3-l_2-l_1-l_0$) is evaluated for the comparison purpose.

For the convolutional structures, the training process is executed based on the mini-batch RMSprop algorithm [45] (training parameters: $lr = 0.001$, $rho = 0.9$, $batch\_size = 32$, $Max\_epoch = 50$, and $decay = 10^{-5}$). In off-the-shelf CNN baseline study, the "cross-entropy" cost function is considered for the training of the Softmax classifiers. Furthermore, the proposed cost function is optimized by the control parameter λ varying between 0 and 1 with a step of 0.1 for the mixture ensemble models. In all experts, the "Sigmoid" activation function is considered for output layers (i.e., FC2 layers), while the corresponding function is the "Softmax" for the CGN. Moreover, the "ReLU" function is selected for all hidden layers, and the modules are initialized by the Glorot-Uniform method [55].

D. Performance Results

1) Dataset 1: Table III shows the details and average results of the evaluated structures according to the 5-fold cross validation method on dataset 1 and the diagnostic threshold (τ) of 15%. Following the reported performances, the MCME-model with the fusion of $l_3-l_2-l_1-l_0$ scales, and at $\lambda = 0.2$, outperforms the other methods and configurations. It presents an AUC of 0.998 with a precision of 99.39% at a recall of 99.36% on this dataset.

The precision results of the best MCME models and the FCME baselines are compared based on λ factor in Fig. 6.

2) Dataset 2: In this experiment, the best multi-scale structures on dataset 1, i.e., the l_1-l_0 MCME, $l_2-l_1-l_0$ MCME, and $l_3-l_2-l_1-l_0$ MCME models are evaluated by exploring optimal λ parameter at $\tau = 15\%$ on dataset 2. To this end, the precision measure is calculated to be 95.67%, 98.33%, and 96.67%, respectively for the models. So, the $l_2-l_1-l_0$ MCME model at $\lambda = 0.7$ outperforms the other models with AUC of 0.999, recall of 97.78%, and F1-score of 97.71% where the MSE, mean correlation coefficient, DbD, and κ values of this model are 0.02, −0.05, 0.51, and 0.16, respectively on this dataset.

E. Diagnostic Sensitivity

In this experiment, different values of the diagnostic threshold value τ are explored to investigate the effects of this

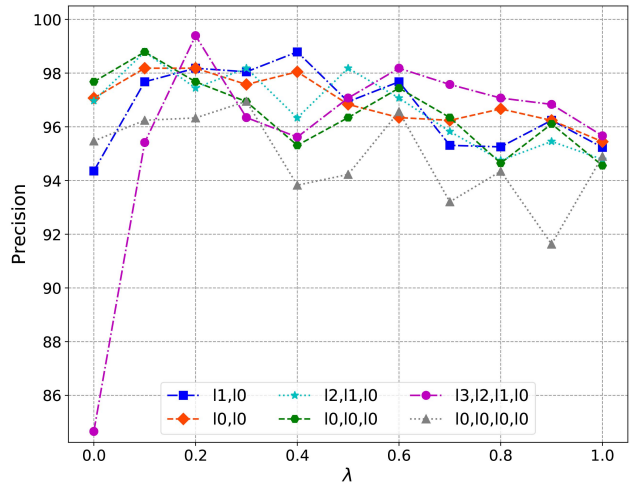

Fig. 6. Comparison of the precision rate of the best ME structures with different numbers of experts/scales on dataset 1 at $\tau = 15\%$.

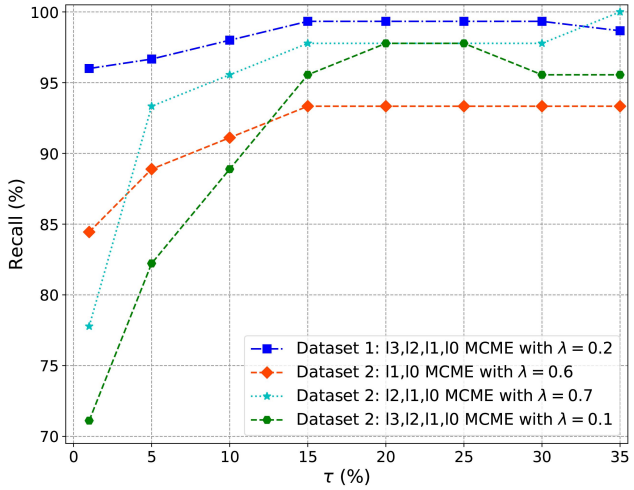

Fig. 7. Recall vs. τ curves of the selected MCME models on the datasets evaluated by the 5-fold cross-validation method.

parameter on the performance of the selected MCME models at the patient level. Fig. 7 demonstrates the CAD system *Recall* changes versus τ on *Dataset*1 and *Dataset*2.

The evaluated models were coded in Python 2.7 (based on Theano v0.8 [56] and Keras v1.2 [57] Toolkits) and trained using the NVIDIA Pascal GTX 1080 GPU card with NVIDIA Cuda v8.0 and cuDNN v5.1 accelerated library. It should be noted that when using the testing code, the classification average time was about 8 $msec/ROI$ on average for the $l_3 - l_2 - l_1 - l_0$ MCME model. This time is approximately 1.4 times faster than the best FCME model with three l_0-dependent CNN experts.

IV. Discussion

In the present study, we proposed a novel and automatic CAD system in retinal OCT images using a new deep ensemble model (i.e., MCME) in classification stage. To assess the CAD system, the use of different scales (experts) in the MCME model, and the effects on diagnostic performance and time complexity were explored based on two different datasets. The first one was a local dataset of 148 subjects. This dataset was used to investigate the best configurations of the MCME method. The second dataset was utilized to evaluate the proficiency of the selected MCME configuration. This set is a publicly available dataset of 45 volumetric SD-OCT acquisitions [22]. To this end, the evaluation of the proposed ensemble method was done by the ROC, and confusion matrix analyses.

A. MCME Analysis Based on Dataset 1

For dataset 1, by considering the 5-fold cross validation and the diagnostic threshold $\tau = 15\%$, an exhaustive experimental study was carried out for all the possible scale-fusion configurations for the proposed MCME model. The results are summarized in Table III. As the best results, an overall precision of 99.39%, recall of 99.36%, F1-score of 99.34%, and AUC of 0.998 was achieved for the classification of the patients with an MCME model including 4 experts with $l_3 - l_2 - l_1 - l_0$ input scales and also a gating network with l_0 input. Regarding the evaluation of the feature space partitioning hypothesis by MCME components; mean correlation coefficient, DbD, and Cohen's kappa (κ) factors were assessed which were obtained to be -0.04, 0.79, and 0.03 respectively for built-in CNN experts in the model.

For the benchmark study, to get better insight into the performance of the MCME model, five different techniques were considered as the baselines: (1) HOG-LSVM, (2) Off-the-shelf CNNs, (3) Single-scale CNNs, (4) Convolutional Ave-ensemble model, and (5) Full-rank convolutional mixture of experts model.

For this purpose, in the first step the HOG-LSVM method [22] was tried as a classic feature-based technique with a precision of 85.35% and AUC of 0.903.

Moreover, the recent competitive deep networks VGG19, ResNet50, and InceptionV3 were evaluated separately in this problem as the off-the-shelf visual feature extractors. So, the extracted features were classified by the Softmax classifier. As reported in Table III, the best results belonged to the ResNet50 model with a precision of 95.31%. Indeed, the model's execution time in the feature extraction step was about 33.2 $sec/image$ on average. It should be noted that, as pointed out in [58], applications of the off-the-shelf deep CNNs to computer aided diagnostic or detection systems can be improved by either exploring the complementary role of hand-crafted features or by training CNNs from scratch on the target dataset. The first solution is not the main focus in this paper. For the latter, since the training of such the models needs very large databases along with special hardware requirements; the data-driven representation approach via full-training of an ensemble of CNNs was chosen as the alternative strategy for this classification problem.

In the next baseline study, the performance of single-scale CNNs was explored for $l0$, $l1$, $l2$ and $l3$ input scales separately. To this purpose, four different CNNs were designed and optimized with considering a minimum number of free parameters along with acceptable performance as described

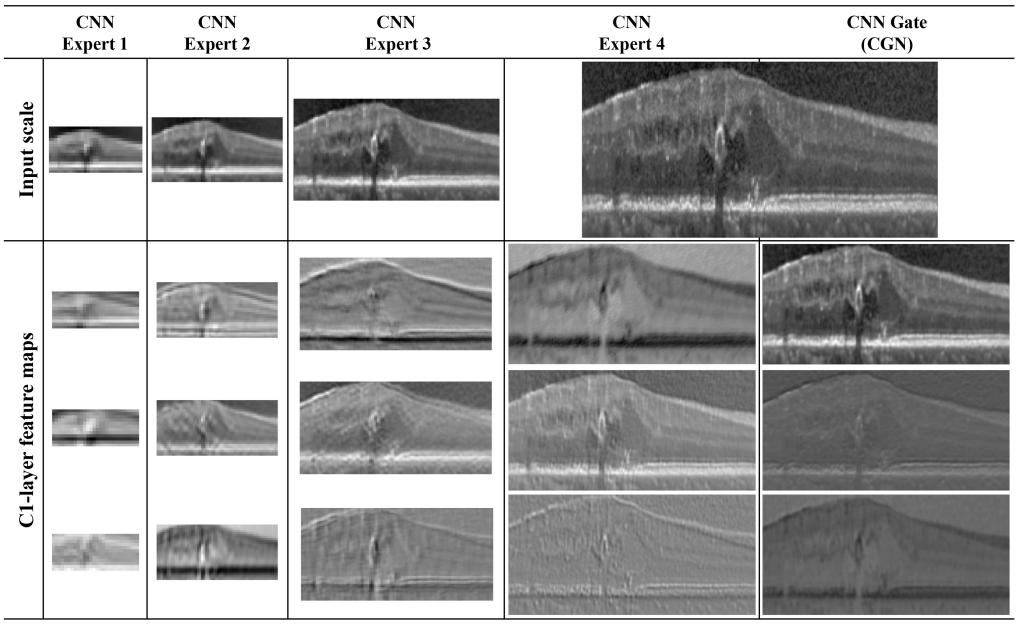

Fig. 8. Extracted feature maps of C1 layers in the trained $l_3 - l_2 - l_1 - l_0$ MCME model for an input ROI in DME class.

in subsection III-C.1. As expected, the CNN1 with 19-active layers and l_0-scale receptive field had better performance than the other evaluated single-scale CNNs. For this model, the overall precision and the training time were 97.50% and 0.086 sec/ROI, respectively. Compared to the evaluated MCME models, CNN1's time complexity is comparable with the $l_3 - l_2$ MCME while the CNN1 performance is less than this model. This outcome proves the efficiency of the selected ensemble approach versus the comparable vanilla CNNs.

For the comparison purpose, the common averaging ensemble method of single-CNNs (the Ave-ensemble) was also evaluated on dataset 1. This model, with averaging of the CNN1, CNN2, CNN3, and CNN4 output-maps yielded an AUC of 0.994 and an overall precision of 96.45%. This model couldn't attain to a comparable performance with the $l_3 - l_2 - l_1 - l_0$ MCME model, although the training time of the two models was roughly the same (169 vs. 170 ms/ROI). In fact, in the Ave-ensemble, all the independent CNNs have the same combination weights, and are treated equally. However, not all the scale-dependent CNNs are equally important. In contrast with the Ave-ensemble, different CNNs in the MCME can be adaptively specialized to different parts of the feature space by means of the CGN weighting role in the learning phase. So, making the final decision of the ensemble with these weighted outputs can effectively enhance the performance. On the other hand, as discussed in [44] and [59], the combining results in neural ensemble models are weakened specifically if the errors of individual nets are positively correlated. This phenomenon occurred here too because for the subnetworks in the Ave-ensemble model the average correlation coefficient and κ factors were 0.31 and 0.77 respectively, while these values were -0.04 and 0.03 for the selected MCME model on dataset 1.

Furthermore, the FCME structure was assessed by fusion of 2, 3, and 4 experts with l_0 input scale. So, the best results were obtained using 3-experts with an overall precision of 98.83% at $\lambda = 0.1$, where the training time for this model was about 0.193 sec/ROI. In comparison to the mixture ensemble of full-ranked CNNs, using the new penalty term could effectively enhance the performance of the FCME model. Confirming this claim, the 3-experts FCME at $\lambda = 0$ (i.e., without the penalty term ρ) had a precision of 97.75% while the average correlation coefficient, DbD, and κ factors for the CNN experts were 0.07, 0.45, and 0.43, respectively. Based on the results in Table III, since the MCME model with the fusion of $l_1 - l_0$ scales had a performance same as the $l_0 - l_0 - l_0$ FCME but in a training time of 0.120 sec/ROI, proposing the prior MSSP decomposition in the MCME model (see Section II-C) is an impressive strategy for a fast and discriminative information fusion. Based on the preliminary analyses on this dataset, the MCME results and complexity were improved in comparison to the best FCME model on the retinal OCT database. So, the proposed MCME has a greater potential to analyze the retinal VOIs in comparison to the best FCME method. For this MCME model with fewer parameters, less training time-complexity (0.170 vs. 0.193 *sec/ROI*) and fewer memory requirements, the performance of the MCME was higher than the best FCME model. These outcomes are mainly valuable in case of encountering a medical diagnosis problem, including a limited number of samples and with an in-depth full-training procedure. Fig. 8 shows 2-D extracted feature maps of C1 layers in the trained MCME model for an input ROI. This figure shows that each trained module in the MCME model generates different feature maps based on its specific input scale.

As demonstrated in Fig. 7, regarding the recall measure, $\tau = 15\%$ is the best value for volumetric diagnosis at the patient level based on predicted abnormal B-scans by the $l_3 - l_2 - l_1 - l_0$ MCME on dataset 1. It should be noted that, in some volumetric OCT acquisitions (those with 19 B-scans), the threshold of 15% includes only 3 abnormal

B-scans for decision making. Furthermore, the model obtained a recall of 95.97% at $\tau = 1\%$ on the dataset.

B. MCME Analysis Based on Dataset 2

For dataset 2, the $l_2 - l_1 - l_0$ MCME (with 9268 free parameters) outperformed the $l_3 - l_2 - l_1 - l_0$ MCME at $\tau = 15\%$ where the precision and AUC values were 98.33% and 0.999, respectively for this model. Although the $l_3 - l_2 - l_1 - l_0$ MCME at $\lambda = 0.7$ and $\tau = 20\%$ results in a precision of 98.33%, the ratio of its free parameters (i.e., 10285) to the number of augmented training samples (10940 ROIs on average in the CV5 method) makes it more prone to be over-fitted, therefore it presents a lower diagnostic sensitivity than the $l_2 - l_1 - l_0$ MCME model at $\tau = 15\%$. It should be noted that both the latter MCME models have a lower false positive diagnostic rate rather than the HOG-based classification method by Srinivasan *et al.* [22] on this database because they just misclassified a normal case as DME one in the CV5 process. Compared to the sparse coding approach presented by Sun *et al.* [27], although the classification performance of the methods is similar on this dataset, the MCME approach does not rely on any denoising step in the preprocessing algorithm. This is where the performance of [27] algorithm is evaluated by the leave-three-out method.

Generally, based on Table III and Fig. 6, we observe that the adjusting parameter (λ) in the proposed cost function can effectively impress the MCME performance. As pointed out in [44], this phenomenon occurs due to the control ability of the bias-variance-covariance trade-off in the learning process by λ. It seems that an optimum increasing diversity among the scale-dependent experts by the λ leads to promising improvements in the performance of the MCME model with different scales fusion.

Moreover, the performance of the proposed MCME models is impressively influenced by an intrinsic competition of information fusion by scale-dependent CNN experts for input pattern classification, while the CGN rewards the winner of each competition with stronger and specific error feedback signals [54]. Thus, the CGN can partition the feature space according to the experts' performance, decrease the experts' correlations and κ, and increase the overall diagnosis performance consequently.

Besides, by considering the time complexity in Table III, the MCME also addresses an efficient full-training strategy and a fast and powerful representation model in pathologic OCT diagnosis.

Among several promising directions, one could extend the MCME approach in macular OCT image classification without engaging the retinal alignment processing, by using more complex CNN modules and large OCT databases.

Finally, the experimental studies on two distinct datasets showed that with an optimum number of scale-dependent convolutional experts, and a near-optimum balance in the bias-variance-covariance trade-off provided by λ, the MCME structure with in-depth and full-training approach, would result in an efficient, fast and reliable diagnostic classifier for macular OCT screening.

V. Conclusion

The present study introduced and evaluated a novel CAD system for retinal pathology diagnosis in macular OCT, which does not rely on image denoising, full retinal layers segmentation, or lesions detection processes. The main contribution of this study was to introduce and analyze the MCME model in retinal OCT classification problem. Therefore, the mathematical model of the novel classifier was introduced which is coupled with a new cost function based on the addition of a cross-correlation penalty term. Data-driven features and the representative ability of the proposed model benefit to reduce the complexity and diagnosis error and to obtain an overall average precision rate of 98.86% on two datasets of 148 and 45 retinal OCT volumes including dry AMD, DME, and normal subjects. As the future work, it is expected that with a larger database including more retinal pathologies with a larger amount of different cases, and extended convolutional modules, the performance of the proposed MCME model should be significantly improved.

References

[1] T. Vos *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: A systematic analysis for the global burden of disease study 2013," *Lancet*, vol. 386, no. 9995, pp. 743–800, 2015.

[2] S. Mehta, "Age-related macular degeneration," *Primary Care, Clin. Office Pract.*, vol. 42, no. 3, pp. 377–391, 2015.

[3] M. M. Engelgau *et al.*, "The evolving diabetes burden in the United States," *Ann. Internal Med.*, vol. 140, no. 11, pp. 945–950, 2004.

[4] S. Pershing, E. A. Enns, B. Matesic, D. K. Owens, and J. D. Goldhaber-Fiebert, "Cost-effectiveness of treatment of diabetic macular edema," *Ann. Internal Med.*, vol. 160, no. 1, pp. 18–29, 2014.

[5] R. Bernardes and J. Cunha-Vaz, *Optical Coherence Tomography: A Clinical and Technical Update*. Springer, 2012.

[6] (2016). [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.htm

[7] M. E. van Velthoven, D. J. Faber, F. D. Verbraak, T. G. van Leeuwen, and M. D. de Smet, "Recent developments in optical coherence tomography for imaging the retina," *Prog. Retinal Eye Res.*, vol. 26, no. 1, pp. 57–77, 2007.

[8] H. Rabbani, M. Sonka, and M. D. Abramoff, "Optical coherence tomography noise reduction using anisotropic local bivariate Gaussian mixture prior in 3D complex wavelet domain," *Int. J. Biomed. Imag.*, vol. 2013, Jun. 2013, Art. no. 417491.

[9] R. Kafieh, H. Rabbani, and I. Selesnick, "Three dimensional data-driven multi scale atomic representation of optical coherence tomography," *IEEE Trans. Med. Imag.*, vol. 34, no. 5, pp. 1042–1062, May 2015.

[10] Z. Amini and H. Rabbani, "Statistical modeling of retinal optical coherence tomography," *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1544–1554, Jun. 2016.

[11] R. Kafieh, H. Rabbani, M. D. Abramoff, and M. Sonka, "Curvature correction of retinal OCTs using graph-based geometry detection," *Phys. Med. Biol.*, vol. 58, no. 9, pp. 2925–2938, 2013.

[12] M. D. Abràmoff *et al.*, "Automated segmentation of the cup and rim from spectral domain OCT of the optic nerve head," *Inv. Ophthalmol. Vis. Sci.*, vol. 50, no. 12, pp. 5778–5784, 2009.

[13] Q. Yang *et al.*, "Automated layer segmentation of macular OCT images using dual-scale gradient information," *Opt. Exp.*, vol. 18, no. 20, pp. 21293–21307, 2010.

[14] G. Quellec, K. Lee, M. Dolejsi, M. K. Garvin, M. Abràmoff, and M. Sonka, "Three-dimensional analysis of retinal layer texture: Identification of fluid-filled regions in SD-OCT of the macula," *IEEE Trans. Med. Imag.*, vol. 29, no. 6, pp. 1321–1330, Jun. 2010.

[15] 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, 2013.

[16] P. A. Dufour *et al.*, "Graph-based multi-surface segmentation of OCT data using trained hard and soft constraints," *IEEE Trans. Med. Imag.*, vol. 32, no. 3, pp. 531–543, Mar. 2013.

[17] Y. Sun, T. Zhang, Y. Zhao, and Y. He, "3D automatic segmentation method for retinal optical coherence tomography volume data using boundary surface enhancement," *J. Innov. Opt. Health Sci.*, vol. 9, no. 2, p. 1650008, 2016.

[18] M. Esmaeili, A. M. Dehnavi, and H. Rabbani, "3D curvelet-based segmentation and quantification of drusen in optical coherence tomography images," *J. Elect. Comput. Eng.*, vol. 2017, Jan. 2017, Art. no. 4362603.

[19] Y.-Y. Liu, M. Chen, H. Ishikawa, G. Wollstein, J. S. Schuman, and J. M. Rehg, "Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding," *Med. Image Anal.*, vol. 15, no. 5, pp. 748–759, 2011.

[20] A. Albarrak, F. Coenen, and Y. Zheng, "Age-related macular degeneration identification in volumetric optical coherence tomography using decomposition and local feature extraction," in *Proc. Int. Conf. Med. Image, Understand. Anal.*, 2013, pp. 59–64.

[21] S. Farsiu *et al.*, "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography," *Ophthalmology*, vol. 121, no. 1, pp. 162–172, 2014.

[22] P. P. Srinivasan *et al.*, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," *Biomed. Opt. Exp.*, vol. 5, no. 10, pp. 3568–3577, 2014.

[23] F. G. Venhuizen *et al.*, "Automated age-related macular degeneration classification in OCT using unsupervised feature learning," *Med. Imag.*, vol. 9414, p. 94141I, Mar. 2015.

[24] G. Lemaître *et al.*, "Classification of SD–OCT volumes using local binary patterns: Experimental validation for DME detection," *J. Ophthalmol.*, vol. 2016, May 2016, Art. no. 3298606.

[25] S. Apostolopoulos, C. Ciller, S. I. De Zanet, S. Wolf, and R. Sznitman. (Oct. 2016). "RetiNet: Automatic AMD identification in OCT volumetric data." [Online]. Available: https://arxiv.org/abs/1610.03628

[26] F. G. Venhuizen *et al.*, "Automated staging of age-related macular degeneration using optical coherence tomography," *Invest. Ophthalmol. Vis. Sci.*, vol. 58, no. 4, pp. 2318–2328, 2017.

[27] Y. Sun, S. Li, and Z. Sun, "Fully automated macular pathology detection in retina optical coherence tomography images using sparse coding and dictionary learning," *J. Biomed. Opt.*, vol. 22, no. 1, p. 16012, 2017.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[29] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May/Jun. 2010, pp. 253–256.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[31] R. Rasti, M. Teshnehlab, and R. Jafari, "A CAD system for identification and classification of breast cancer tumors in DCE-MR images based on hierarchical convolutional neural networks," *Comput. Intell. Elect. Eng.*, vol. 6, no. 1, pp. 1–14, 2015.

[32] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[33] S. Kotsiantis, "Combining bagging, boosting, rotation forest and random subspace methods," *Artif. Intell. Rev.*, vol. 35, no. 3, pp. 223–240, 2011.

[34] S. B. Kotsiantis, "An incremental ensemble of classifiers," *Artif. Intell. Rev.*, vol. 36, no. 4, pp. 249–266, 2011.

[35] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. COM-31, no. 4, pp. 532–540, Apr. 1983.

[36] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[37] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.

[38] M. H. Nguyen, H. A. Abbass, and R. I. Mckay, "A novel mixture of experts model based on cooperative coevolution," *Neurocomputing*, vol. 70, nos. 1–3, pp. 155–163, 2006.

[39] R. Ebrahimpour, E. Kabir, H. Esteky, and M. R. Yousefi, "View-independent face recognition with mixture of experts," *Neurocomputing*, vol. 71, nos. 4–6, pp. 1103–1107, 2008.

[40] M. Javadi, S. A. A. A. Arani, A. Sajedin, and R. Ebrahimpour, "Classification of ECG arrhythmia by a modular neural network based on mixture of experts and negatively correlated learning," *Biomed. Signal Process. Control*, vol. 8, no. 3, pp. 289–296, 2013.

[41] M. N. Dailey and G. W. Cottrell, "Organization of face and object recognition in modular neural network models," *Neural Netw.*, vol. 12, nos. 7–8, pp. 1053–1074, 1999.

[42] S. Masoudnia and R. Ebrahimpour, "Mixture of experts: A literature survey," *Artif. Intell. Rev.*, vol. 42, no. 2, pp. 275–293, 2014.

[43] R. A. Jacobs, "Bias/variance analyses of mixtures-of-experts architectures," *Neural Comput.*, vol. 9, no. 2, pp. 369–383, 1997.

[44] Y. Liu and X. Yao, "Simultaneous training of negatively correlated neural networks in an ensemble," *IEEE Trans. Syst., Man, B, Cybern.*, vol. 29, no. 6, pp. 716–725, Dec. 1999.

[45] G. Hinton, N. Srivastava, and K. Swersky, "Lecture 6a overview of mini–batch gradient descent," *Coursera Lect. Slides*, 2012. [Online]. Available: https://class.coursera.org/neuralnets-2012-001/lecture

[46] J. Cohen, "A coefficient of agreement for nominal scales," *Edu. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[47] C. O. A. Freitas, J. M. de Carvalho, J. Oliveira, Jr., S. B. K. Aires, and R. Sabourin, "Confusion matrix disagreement for multiple classifiers," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2007, pp. 387–396.

[48] K. Simonyan and A. Zisserman. (Apr. 2015). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[51] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[52] P. Buyssens, A. Elmoataz, and O. Lézoray, "Multiscale convolutional neural networks for vision–based classification of cells," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 342–352.

[53] H. Chen, Q. Dou, X. Wang, J. Qin, and P. A. Heng, "Mitosis detection in breast cancer histology images via deep cascaded networks," in *Proc. Assoc. Adv. Artif. Intell.*, 2016, pp. 1160–1166.

[54] R. Rasti, M. Teshnehlab, and S. L. Phung, "Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks," *Pattern Recognit.*, vol. 72, pp. 381–390, Dec. 2017.

[55] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, vol. 9. 2010, pp. 249–256.

[56] F. Bastien *et al.*, "Theano: New features and speed improvements," in *Proc. Deep Learn. Unsupervised Feature Learn. NIPS Workshop*, 2012.

[57] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[58] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[59] M. P. Perrone and L. N. Cooper, "When networks disagree: Ensemble methods for hybrid neural networks," Inst. Brain Neural Syst., Brown Univ., Providence, RI, USA, Tech. Rep., 1992.