This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2795545, IEEE Journal of Biomedical and Health Informatics

Surrogate-assisted Retinal OCT Image Classification Based on Convolutional Neural Networks

Yibiao Rong, Dehui Xiang, Weifang Zhu, Kai Yu, Fei Shi, Zhun Fan, *Senior Member, IEEE* and Xinjian Chen, *Senior Member, IEEE*

***Abstract*—Optical Coherence Tomography (OCT) is becoming one of the most important modalities for the noninvasive assessment of retinal eye diseases. As the number of acquired OCT volumes increases, automating the OCT image analysis is becoming increasingly relevant. In this paper, we propose a surrogate-assisted classification method to classify retinal OCT images automatically based on convolutional neural networks (CNNs). Image denoising is first performed to reduce the noise. Thresholding and morphological dilation are applied to extract the masks. The denoised images and the masks are then employed to generate a lot of surrogate images, which are used to train the CNN model. Finally, the prediction for a test image is determined by the average of the outputs from the trained CNN model on the surrogate images. The proposed method has been evaluated on different databases. The results (AUC of 0.9783 in the local database and AUC of 0.9856 in the Duke database) show that the proposed method is a very promising tool for classifying the retinal OCT images automatically.**

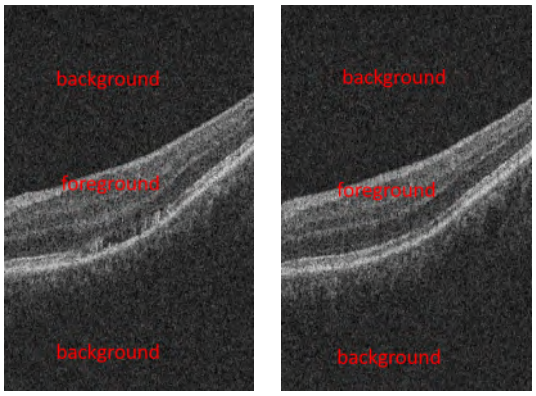
***Index Terms*—Retinal OCT image classification, surrogate-assisted, ensemble, convolutional neural networks.**

I. Introduction

As a noninvasive imaging modality, optical coherence tomography (OCT) has become an important modality for assisting the diagnosis and management of eye diseases, such as assisting the diagnosis of age-related macular degeneration (AMD) and diabetic macular edema (DME) [1]. However, many factors, including population growth, rapid aging in many countries, lead to an increase in the number of acquired OCT volumes, which causes the burden of ophthalmologists increased significantly. As a result, a computer aided system which can automate the process of eye disease analysis is desired since it can alleviate the burden on the clinicians [2].

Many technologies have been developed for the automatic analysis of eye diseases, including segmentation [3–10] and classification [11–19], in which [7–10] and [15–19] are deep learning based methods for segmentation and classification developed in recent respectively. The purpose of classification is to classify the images into several categories so that each category has the same characteristics. For instance, Alsaih et al. [11] classified the OCT images as either DME or normal with multi pyramids, LBP and HOG descriptors. Farsiu et al. [12] classified the OCT images as either AMD or normal by using 4 disease indicators, which are total retina volumes, pigment epithelium drusen complex volumes, abnormal pigment epithelium drusen complex volumes and thinning volumes. Srinivasan et al. [13] employed multi-scale histograms of oriented gradient descriptors as feature vectors of a support vector machine to classify the OCT images into three categories: AMD, DME or normal. Liu et al. [14] employed a machine learning approach to detect macular pathology automatically in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding. To learn more about the methods used to facilitate the process of automatic analysis of eye diseases, we refer readers to [1] and [20] for a comprehensive reading.

The classification methods introduced above [11–14] employed the hand-crafted features for the classifiers to identify the patterns in the images. Although promising results can be obtained using the hand-crafted features, the disadvantages are obvious, including requiring abundant expert knowledge, being time-consuming and difficult to be generalized to other domains. One of the popular methods to address these disadvantages is the convolutional neural networks (CNNs) [21], which can learn representations from raw data automatically.



(a) (b)

Fig. 1. An example of the retinal OCT image. (a) A B-scan image of an eye affected by AMD. (b) A B-scan image of a normal eye.

This study was supported in part by the National Basic Research Program of China (973 Program) under Grant 2014CB748600, in part by the National Nature Science Foundation of China for Excellent Young Scholars under Grant 61622114, in part by the National Nature Science Foundation of China under Grants 81371629, 81401472, 61401294, 61401293 and 61601317, and in part by the International Cooperation Project of Ministry of Science and Technology (2016YFE010770).

Yibiao Rong, Dehui Xiang, Weifang Zhu, Kai Yu, Fei Shi and Xinjian Chen are with School of Electrical and Information Engineering, Soochow University. 215006, Suzhou, China (e-mail: ybrong@stu.suda.edu.cn; xiangdehui@suda.edu.cn; wfzhu@suda.edu.cn; 578069383@qq.com; shifei@suda.edu.cn; xjchen@suda.edu.cn). (Corresponding authors: Xinjian Chen, Dehui Xiang.)

Zhun Fan is with key Laboratory of Digital Signal and Image Processing of Guangdong Provincial , College of Engineering, Shantou University. 515063, Shantou, China (e-mail: zfan@stu.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2795545, IEEE Journal of Biomedical and Health Informatics

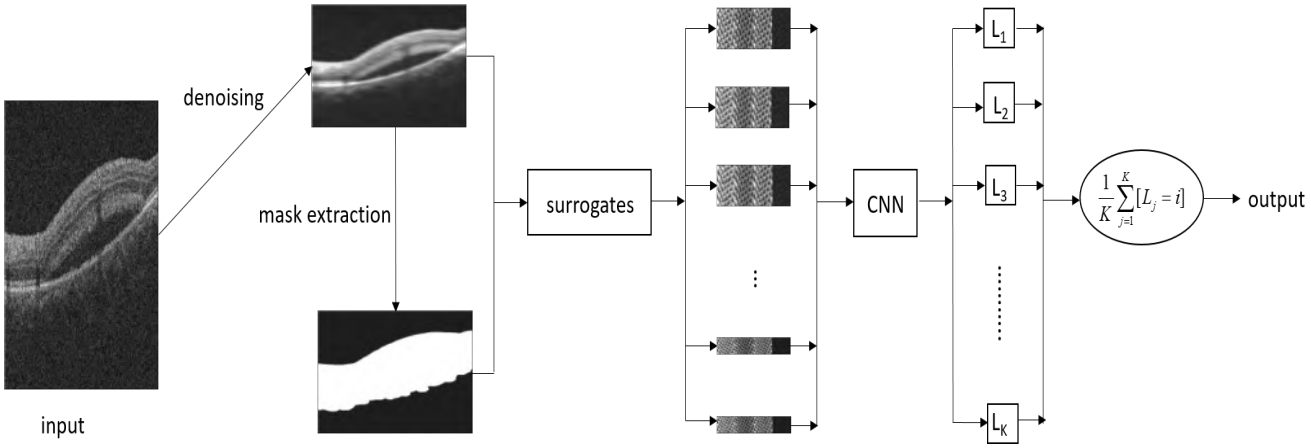


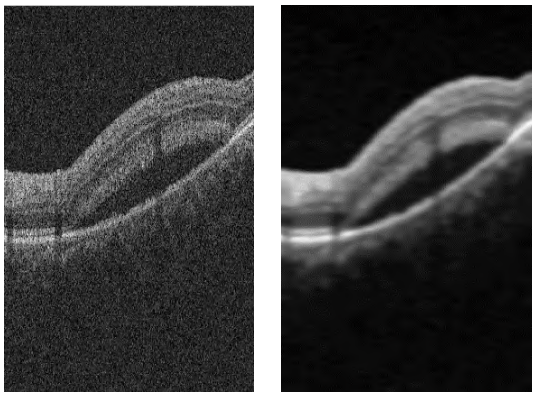
Fig. 2. The flowchart of the proposed method. In $\frac{1}{K}\sum_{j=1}^{K}[L_j = i]$, $i = 1, 2, .., n$ where n is the number of categories. K is the number of surrogate images of the input image. $[\cdot]$ is an indicator, which means that if $L_j = i$ is true, $[L_j = i]$ is equal to 1, otherwise 0.

The effectiveness of CNNs has been proved by many successful applications, such as image recognition [22], object detection [23] and semantic segmentation [24].

Recently, some researchers have employed deep learning based methods for medical image classification. Such as, in [15], Gualshan et al. developed a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Similar with [15], a deep learning based method for detecting diabetic retinopathy in retinal fundus photographs was also reported in [19]. Lee et al. [16] employed the convolutional neural networks to distinguish AMD from normal OCT images. In [18], Karri et al. fine-tuned a pre-trained CNN to identify retinal pathologies given retinal OCT images. In [17], Burlina et al. explored the appropriateness of the transfer of image features computed from pre-trained deep neural networks to the problem in AMD detection.

In these deep learning based methods [15–19], some of them used the raw images to train the CNN model from scratch, e.g. [15], [16]. Some of them used the raw images to fine tune a pre-trained CNN model, e.g. [18]. Different from these methods which used the raw images as input for the CNNs, in this paper, we propose a method using the trained CNN model to classify the surrogates of the original OCT images to achieve the original OCT images classification. The advantages of the proposed surrogate-assisted classification method are concluded as follows.

- We can generate many different surrogates of the original images. As a result, the challenge in medical imaging domain that lacking of large scale annotated images to train a CNN [25] can be addressed.
- Generally, the resolution of an original OCT image is high and the background is a large part of an OCT image, as shown in Fig. 1. Thus, it is memory-consuming using the original OCT images to train and test a CNN model. The proposed surrogate-assisted classification method can deal with this problem since the size of the surrogate image is much smaller than the original image.
- Due to the surrogates of an original OCT image are different, inspired by the ensemble methods [26] in machine learning, we take the average of the outputs from the trained CNN model on the surrogate images as the final prediction for the original image, which is proved to be an effective way to improve the performance of the proposed method by the experimental results.



(a) (b)

Fig. 3. A denoised image example. (a) Original image. (b) Denoised image.

The rest of this paper is organized as follows: in section II, the main steps of the proposed algorithm are described and explained. Section III presents the empirical study of parameter setting and experimental results obtained using different databases. We have discussion and conclusion in Section IV.

II. Method

The flowchart of the proposed method is shown in Fig. 2. Image denoising is first performed to reduce the noise. Threshhodling and morphological dilation are applied to extract the mask. The denoised image and the mask are then employed to generate the surrogate images. The final prediction for the input image is determined by the average of the outputs from the trained CNN model on the surrogate images.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2795545, IEEE Journal of Biomedical and Health Informatics

(a) (b) (c)

Fig. 4. An example to demonstrate the process of mask extraction. (a) BW. (b) $BW1$. (c) Mask.

A. Denoising

Likewise all coherent imaging systems, OCT imaging suffers from speckle noise [27]. Thus, denoising would be a useful step for the subsequent processing. Many methods have been introduced for OCT image denoising, including anisotropic diffusion filter [28], [29], optimized non-orthogonal wavelet filter [30] and bilateral filter [31]. In this work, we employ the toolbox [1] of sparse representation based method for OCT image denoising developed by Fang et al. [32] to reduce the noise in the OCT image.

Fig. 3 gives an example where Fig. 3(a) is the original image and Fig. 3(b) is the denoised image. To learn more about sparse representation based method for OCT image denoising, we refer readers to [32] for a comprehensive reading.

B. Mask Extraction

Since the background of the OCT image does not contain useful information, we extract a mask of the image to limit the processing scope of the subsequent operations. The procedures of mask extraction are described as follows.

Thresholding is first performed on the denoised image to generate a binary image. The process of the thresholding can be formulated as:

$$BW(x,y) = \begin{cases} 1 & \text{if } I(x,y) > T \\ 0 & \text{others} \end{cases} \quad (1)$$

where I is the denoised image and BW is a binary image. In this work, we employ the OSTU method [33] to compute the threshold value T. The binary image BW is shown in Fig. 4(a). All the connected components in the BW that have fewer than P ($P = 2000$ in this work) pixels are then removed, producing another binary image $BW1$, as shown in Fig. 4(b). Finally, morphological dilation [34] with a structural element of 15×15 disk is applied to fill the holes in the $BW1$, producing the final mask, as shown in Fig. 4(c).

C. Surrogate Image Generation

The purposes of generating surrogate images are to augment the data and reduce the complexity of the original image. The process of generating a surrogate image can be demonstrated by Fig. 5. N small regions $\{r_1, r_2, ..., r_N\}$ are cropped from the image I. M features $[f_{i1}, f_{i2}, ..., f_{iM}]$ are then extracted for each small region r_i to construct a matrix $\mathbf{F}$:

$$\mathbf{F} = \begin{bmatrix} f_{11}, f_{12}, ..., f_{1M} \\ f_{21}, f_{22}, ..., f_{2M} \\ f_{31}, f_{32}, ..., f_{3M} \\ f_{41}, f_{42}, ..., f_{4M} \\ f_{51}, f_{52}, ..., f_{5M} \\ ... \quad ... \quad ... \quad ... \\ f_{N1}, f_{N2}, ..., f_{NM} \end{bmatrix}$$

The matrix $\mathbf{F}$ is finally reshaped as a $\sqrt{N \times M} \times \sqrt{N \times M}$ matrix $\mathbf{S}$, which is termed surrogate image.

$$\mathbf{S} = \begin{bmatrix} f_{11}, f_{12}, f_{13}, f_{14}..., f_{1\sqrt{N\times M}} \\ f_{21}, f_{22}, f_{23}, f_{24}..., f_{2\sqrt{N\times M}} \\ ... \quad ... \quad ... \quad ... \\ f_{\sqrt{N\times M}1}, ..., f_{\sqrt{N\times M}\times\sqrt{N\times M}} \end{bmatrix}$$

We follow two rules when select the N small regions: 1) The size of each small region can be different so that we can generate many different surrogate images. 2) The area composed by the N small regions should cover the area of mask as many as possible. Fig. 6 is the schematic diagram used to demonstrate how to crop the small regions. Assume that the size of the original image is $S_x \times S_y$ and the number of pixels on the mask is M. Let $x_s = \sqrt{M/N}$ where N is the number of small regions. C columns are then selected with the interval of x_s in the range of $[x_s, S_x - x_s]$. At each column, we randomly select N/C points, which are the centers of the small regions. The width of a small region is $2x_s$. The height of a small region is 2 times the distance between the center of the small region and its adjacent center in the same column. Taking the c_1 in Fig. 6 as example, if the distance between c_1 and c_2 is d_1, then the dimension of the small region centered on c_1 is $2x_s \times 2d_1$. K different surrogates can be obtained by repeating the process of generating a surrogate image K times due to the centers are selected randomly.

In this work, we employ four properties of a small region to design the surrogates. They are maximal and minimal pixel intensity, variance and average of pixel intensity in a small

[1] http://people.duke.edu/ sf59/Fang_TMI_2013.htm

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2795545, IEEE Journal of Biomedical and Health Informatics

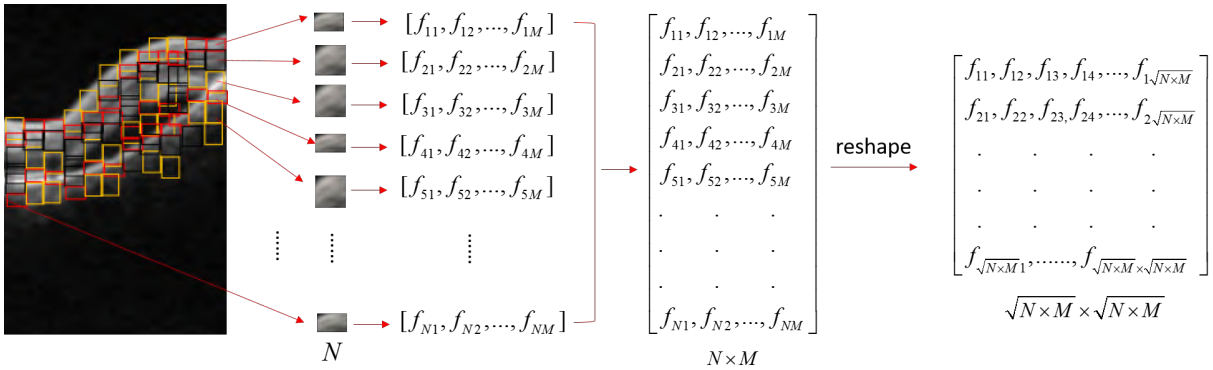


Fig. 5. The schematic diagram of generating surrogate image.

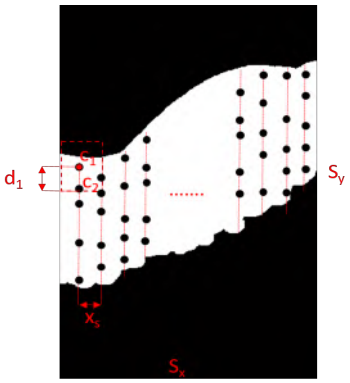
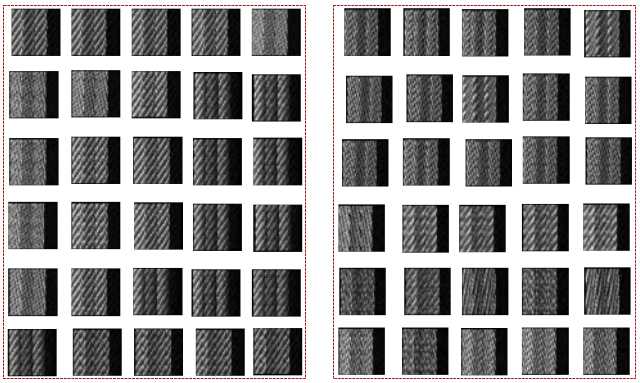


Fig. 6. The schematic diagram of cropping small regions.



(a) (b)

Fig. 7. (a) The examples of surrogates of an OCT image of a normal eye. (b) The examples of surrogates of an OCT image of an eye affected by AMD.

region. The reason for we use these four statistical parameters is based on the observation that the pixel intensity in the lesions is different from the pixel intensity in the normal tissues. Thus, when we use these four statistical parameters to design surrogates, the surrogates of abnormal images will be different from the surrogates of normal images, which is favorable for the proposed method to distinguish abnormal OCT images from normal OCT images. Another reason is that these four properties are easy to compute. We believe that other methods which can make the surrogates of abnormal images different from the surrogates of normal images are also feasible.

It is noteworthy to point out that the category of the surrogate images is same as the original image. Fig. 7 shows some examples where Fig. 7(a) are some surrogate images of an OCT image of a normal eye and Fig. 7(b) are some surrogate images of an OCT image of an eye affected by AMD.

D. Convolutional Neural Networks

Convolutional neural networks (CNNs) [35] allow computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. The architecture of a typical CNN consists of several convolutional layers and pooling layers optionally followed by at least one fully connected layer. A convolutional layer has k filters (or kernels) with trainable weights w. k feature maps are produced by convolving the image with each filter and adding bias b optionally in convolutional layer. Pooling layer is a form of non-linear down-sampling to preserve task-related information while removing irrelevant details [36]. Fully connected layer is used to map the excitations into output neurons. Each output neuron corresponds to one decision class. Fig. 8 shows the architecture of the CNN model applied in this work.

Gradient descent [21] is the common method used to learn the parameters $\theta = [w, b]$ in a CNN model. Besides, there are many tricks for improving the performance of a CNN, such as, data augmentation [22], contrast normalization and whitening [37], dropout [38], batch normalization [39], ensemble [40] and so on. To learn more about the tricks used to improve the performance of a CNN, we refer readers to [41] for a comprehensive reading.

In this work, contrast normalization and whitening [37] are performed on the surrogate images before training and testing the CNN model. In addition, inspired by the ensemble methods [26], given a test image, we take the average of the outputs from the trained model on the surrogate images of each category as the final output. It is different from the presented ensemble methods (e.g [40]) which train multiple learners and then combine them. In this work, only one CNN model is trained. The ensemble is achieved by averaging the outputs from the trained model on surrogate images of each category,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2795545, IEEE Journal of Biomedical and Health Informatics

which can be formulated as:

$$p(i) = \frac{1}{K}\sum_{j=1}^{K}[L_j = i], i = 1, 2, .., n. \quad (2)$$

where n means the number of categories. K is the number of surrogate images of the test image. L_j is the prediction label of the CNN model for the j^{th} surrogate image. $[\cdot]$ is an indicator, which means that if $L_j = i$ is true, $[L_j = i]$ is equal to 1, otherwise 0. $p(i)$ is the probability of the original image classified as the i^{th} category.

III. Results

A. Data Collection and Analysis

Two databases, a local database and a public database Duke [13], are employed to evaluate the proposed method. For the local database, they are collected from seven patients affected by choroidal neovascularization (CNV). The data of each patient consists of 12 (or 13 or 14) volumes. The 3D images with $512 \times 1024 \times 128$ voxels ($11.72\mu m \times 5.86\mu m \times 15.6\mu m$), covering the volume of $6mm \times 6mm \times 2mm$, are obtained by ZEISS scanner. It is noted that the algorithm is performed on B-scan level. Each B-scan in a volume is annotated as either normal or abnormal by a doctor.

The Duke database consists of volumetric scans acquired from 45 subjects: 15 normal subjects, 15 subjects with AMD, and 15 subjects with DME, in which 20 volumes consists of 97 B-scans, 4 volumes consists of 73 B-scans, 8 volumes consists of 61 B-scans, 3 volumes consists of 49 B-scans, 9 volumes consists of 37 B-scans and 1 volume consists of 31 B-scans.

The performance of the method has been evaluated according to different metrics, which are receiver operating characteristic (ROC) curve, the area under curve (AUC) of the ROC [43], sensitivity (Sen), specificity (Spe) and accuracy (Acc). The sensitivity, specificity and accuracy are defined as:

$$Sen = \frac{TP}{TP + FN} \quad (3)$$

$$Spe = \frac{TN}{TN + FP} \quad (4)$$

$$Acc = \frac{TN + TP}{TN + FN + TP + FP} \quad (5)$$

where TP is short for true positive, TN true negative, FP false positive and FN false negative.

The experiments are performed on a PC equipped with an Intel (R) Core (TM) i-7 4790 M CPU at 3.60 GHZ and 8 GB of RAM capacity using MATLAB. Once we have obtained the trained CNN model (It takes about 20 minutes to train the CNN model when the number of the surrogate images is about 6.8×10^4), the average computational time obtained for a B-scan image classification is 6.1426s with a standard deviation of 1.6173s.

B. Empirical Study of Parameter Setting

The parameters may affect the performance of the proposed method are the number of surrogate images K and the number of small regions N. We employ the local database to study the influence of the parameters on the performance of the proposed method. We set a default value for each parameter. Then we allow one change when another is equal to the default value. As a result, we can explore how varying the parameter may affect the performance of the proposed method. The default values of N and K are 1024 and 21 respectively.

The dataset is divided into three parts before training and testing the CNN model: test set, training set and validation set [44]. We select randomly one volume of each subject to construct the test set. The remaining is then divided into validation set and training set randomly, in which validation set accounted for 20%.

1) The influence of K on the performance: Table I summarises the average value of each metric when K is different. Fig. 9(a) are the trend curves drawn according to the results in Table I. It is observed that the performance of the proposed method improves with the increases of the K until it reaches 10. When K is larger than 10, the ACC and AUC tend to be stagnant. Spe improves, but Sen deteriorates, with the increases of K in the range of 10 to 15. After that, the performance of the algorithm becomes stable.

TABLE I
THE PERFORMANCE OF THE PROPOSED METHOD WHEN K IS DIFFERENT.

K	AUC	Acc	Sen	Spe
6	0.9591	0.9085	0.9322	0.8821
10	0.9733	0.9275	0.9428	0.9104
15	0.9768	0.9252	0.9153	0.9363
21	0.9783	0.9286	0.9237	0.9340

2) The influence of N on the performance: Table II summarises the average value of each metric when N is different. Fig. 9(b) are the trend curves drawn according to the results in Table II. It is observed that the performance of the algorithm improves with the increase of the N until it reaches 1024. After that, the performance of the algorithm tends to be stagnant.

TABLE II
THE PERFORMANCE OF THE PROPOSED METHOD WHEN N IS DIFFERENT.

N	AUC	Acc	Sen	Spe
64	0.9639	0.8996	0.9025	0.8962
256	0.9643	0.9007	0.9131	0.8869
1024	0.9783	0.9286	0.9237	0.9340
4096	0.9781	0.9284	0.9230	0.9349

It is noted that different parameters may lead to different metrics improved, e.g. when $K = 10$ ($N = 1024$), the value of Sen is better than other cases, when $K = 15$ ($N = 1024$), the value of Spe is better than other cases. According to the above analysis, we set $K = 21$ and $N = 1024$ for the following experiments, which can be regarded as a compromise among these metrics. Fig. 10 shows some detected examples. In the title of each sub figure, $gt = 1$ or $gt = 2$ means the images annotated as abnormal or normal. $scroe$ is a value between 0 and 1 determined by the proposed algorithm. The higher score, the higher probability of the image classified as abnormal. It is noted that the proposed method can classify the images properly for most of the given cases. However, the proposed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2795545, IEEE Journal of Biomedical and Health Informatics

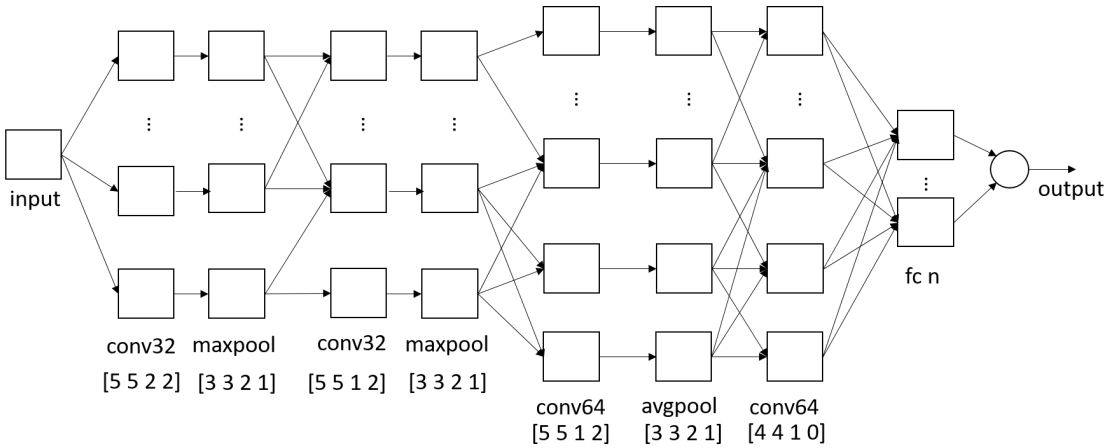
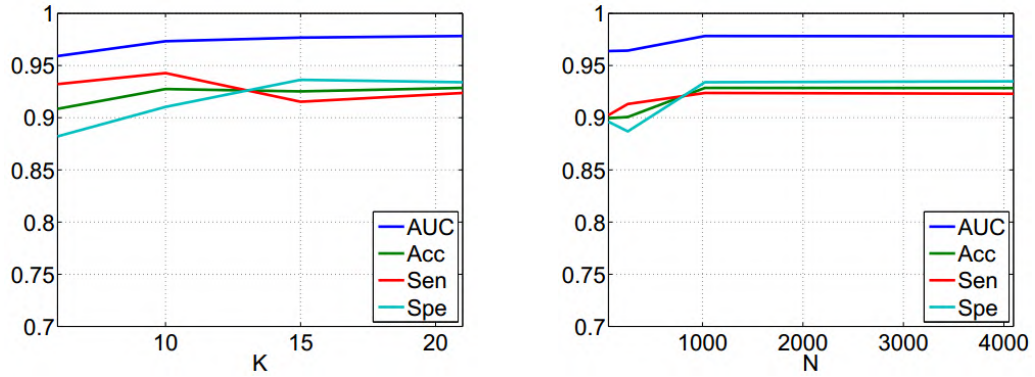


Fig. 8. The architecture of the CNN model applied in this work. The numerical values behind the layer names specify the number of feature maps. The numerical values in square brackets specify receptive field size, stride and padding. It is noted that 'conv' is short for convolutional layer, 'maxpool' and 'avgpool' are short for max pooling layer and average pooling layer respectively. 'fc' is short for fully connected layer. Each convolutional layer is followed by a ReLU layer [42] which is not displayed in the figure. In the fully connected layer, n=2 for binary classification, and n=3 for multiclass classification in this work.



(a) (b)

Fig. 9. Empirical study of parameter setting. (a) The trend curve of each metric when K is different. (b) The trend curve of each metric when N is different.

TABLE III
THE EXPERIMENTAL RESULTS OF THE MODEL FOR BINARY CLASSIFICATION

		NO.B	NO.D	*Acc*			NO.B	NO.D	*Acc*
test1	Sub1(AB)	37	35	0.9459	test6	Sub1(AB)	37	37	1
	Sub2(AB)	61	61	1		Sub2(AB)	61	61	1
	Sub3(NOR)	97	87	0.8969		Sub3(NOR)	49	43	0.8776
test2	Sub1(AB)	37	36	0.9730	test7	Sub1(AB)	37	37	1
	Sub2(AB)	61	61	1		Sub2(AB)	61	61	1
	Sub3(NOR)	97	93	0.9588		Sub3(NOR)	97	96	0.9897
test3	Sub1(AB)	49	49	1	test8	Sub1(AB)	37	36	0.9730
	Sub2(AB)	61	61	1		Sub2(AB)	61	61	1
	Sub3(NOR)	97	95	0.9794		Sub3(NOR)	97	86	0.8866
test4	Sub1(AB)	73	73	1	test9	Sub1(AB)	73	73	1
	Sub2(AB)	97	72	0.7423		Sub2(AB)	31	31	1
	Sub3(NOR)	97	93	0.9588		Sub3(NOR)	97	91	0.9381
test5	Sub1(AB)	49	40	0.8163	test10	Sub1(AB)	37	37	1
	Sub2(AB)	61	61	1		Sub2(AB)	31	31	1
	Sub3(NOR)	97	93	0.9588		Sub3(NOR)	97	86	0.8866

NO.B – the number of B-Scans in an OCT volume. NO.D – the number of B-Scans detected correctly by the proposed algorithm. Sub1(AB) (Sub2(AB)) – the B-scans are annotated as abnormal. Sub3(NOR) – the B-scans are annotated as normal. Acc – The ratio between NO.D and NO.B.

method may fail in some cases, including Fig. 10(b), Fig. 10(g) and Fig. 10(j).

C. Performance on the Public Database

The images in the public database Duke are categorized as AMD, DME or normal. In this section, we train two

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2795545, IEEE Journal of Biomedical and Health Informatics

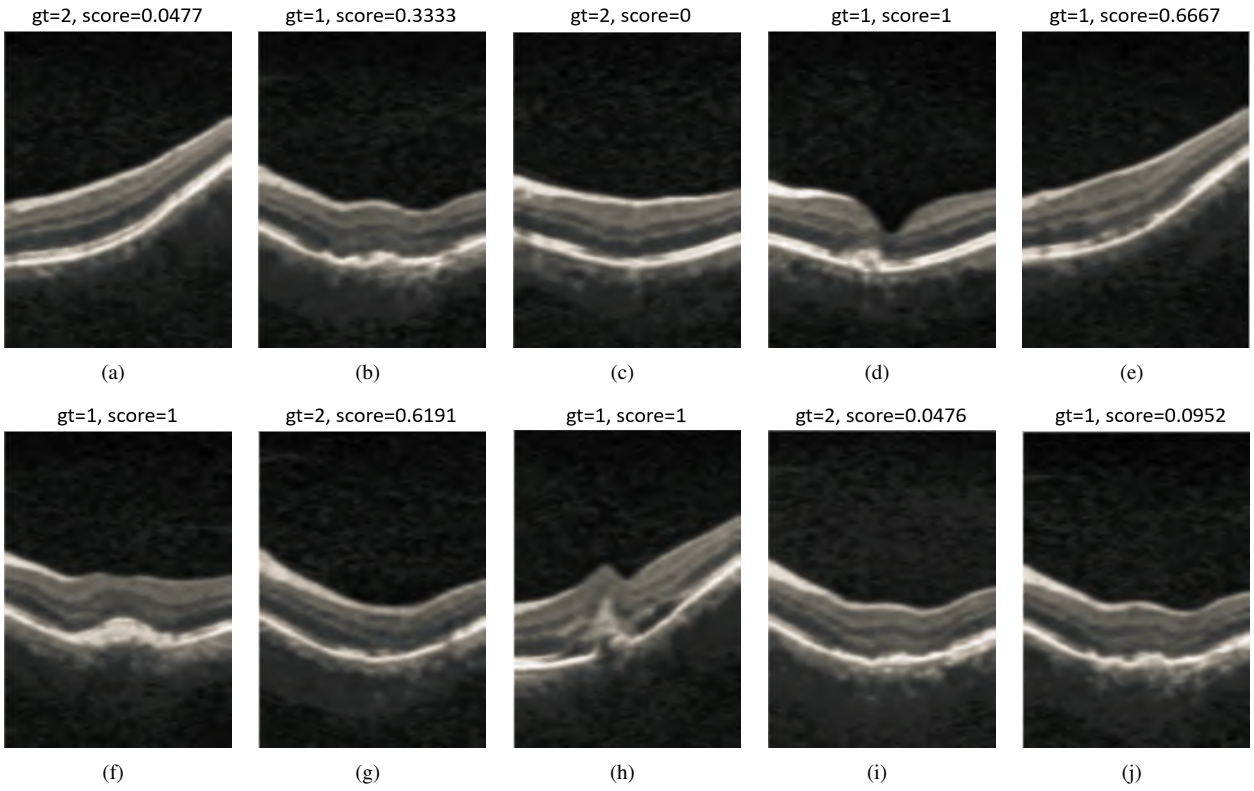


Fig. 10. Some detected examples for binary classification. In the title of each sub figure, $gt = 1$ or $gt = 2$ means the images annotated as abnormal or normal. $scroe$ is a value between 0 and 1 determined by the proposed algorithm. The higher score, the higher probability of the image classified as abnormal.

different types of models to evaluate the performance of the proposed method on this database. One model is used for binary classification, namely, the B-scans categorized as AMD or DME are classified as abnormal, others, normal. The other model is used for multiclass classification, namely the B-scans are classified as AMD, DME or normal. For each type of model, 10 experiments are carried out to evaluate the performance of the models. In each experiment, one volume of each category is selected randomly to construct the test set. The remaining is then divided into validation set and training set randomly, in which validation set accounted for 20%.

1) Model for Binary Classification: Table III summarizes the results obtained by the CNN model for binary classification. In Table III, NO.B is short for the number of B-scans in an OCT volume and NO.D means the number of the B-scans detected correctly by the proposed method. Sub1(AB) or Sub2(AB) means the B-scans are annotated as abnormal. Sub3(NOR) means the B-scans are annotated as normal. Taking the Sub1(AB) in test1 as example, there are 37 abnormal B-scans in the volume, in which 35 B-scans are classified correctly by the proposed method. The average accuracy obtained by model for binary classification is 0.9509 in Duke database.

2) Model for Multiclass Classification: The model for multiclass classification is used to classify the B-scans as AMD, DME or normal. Fig. 11 gives some examples obtained by the CNN model for multiclass classification. In the title of each sub figure, gt=1, gt=2 or gt=3 means that the images are annotated as AMD, DME or normal. 'amd', 'dme' and 'nor' are the values between 0 and 1 obtained by the proposed method. The category of an image classified by the algorithm is determined by the largest value among them. For example, in Fig. 11(e), amd=0.2381, dme=0.6190, nor=0.1429, which means that the image is classified as DME.

Table IV lists the results obtained by the model for multiclass classification. In Table IV, Sub1(AMD), Sub2(DME) or Sub3(NOR) means the B-scans are annotated as AMD, DME or normal. NO.B means the number of B-scans in a volume. AMD.D, DME.D or NOR.D means the number of B-scans are classified as AMD, DME or normal by the proposed method. Acc is the ratio between the number of the B-scans detected correctly by the proposed method and NO.B. Taking Sub1(AMD) in test1 as example, there are 37 B-scans are annotated as AMD in the volume, in which 30 B-scans are classified correctly by the proposed method, 1 B-scan is misclassified as DME and 6 B-scans are misclassified as normal. The average accuracy obtained by the proposed method for multiclass classification is 0.8845.

IV. Discussion & Conclusion

In this paper, we propose a surrogate-assisted retinal OCT image classification method based on CNNs. Two databases, a local database and a public database Duke, have been employed to evaluate the performance of the proposed algorithm in B-scan level. The results (AUC of 0.9783 in the local database and AUC of 0.9856 in the public database) show that the proposed method is a very promising tool for classifying OCT images automatically.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2795545, IEEE Journal of Biomedical and Health Informatics

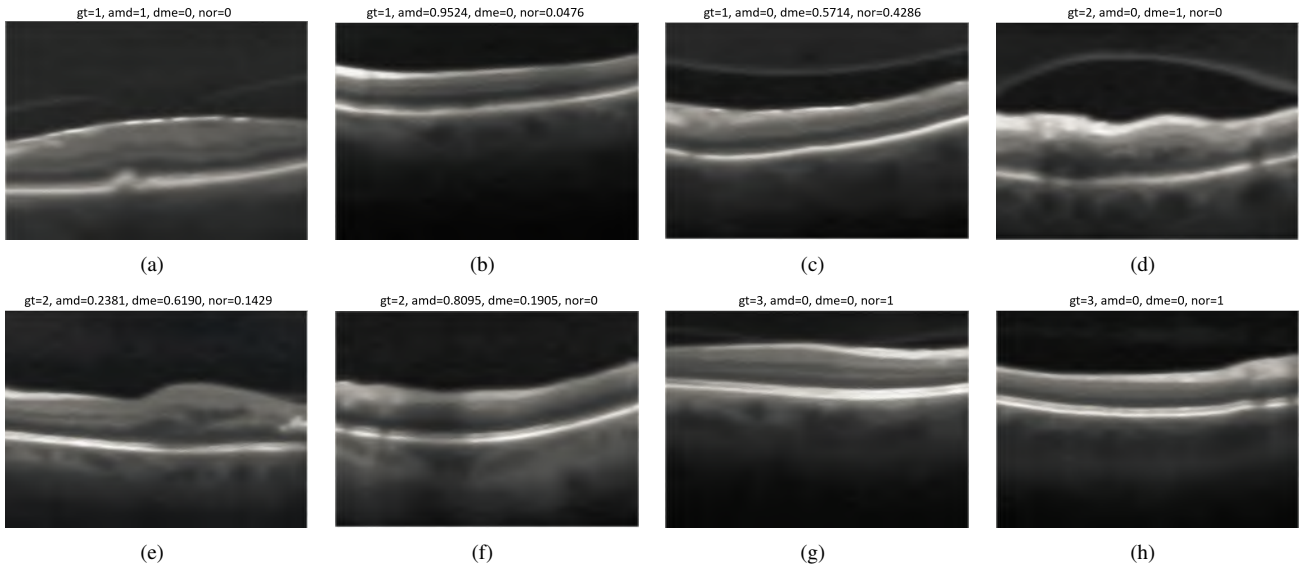


Fig. 11. Some detected examples for multiclass classification. In the title of each sub figure, gt=1, gt=2 or gt=3 means that the images are annotated as AMD, DME or normal. 'amd', 'dme' and 'nor' are the values between 0 and 1 obtained by the proposed method. The category of an image classified by the algorithm is determined by the largest value among 'amd', 'dme' and 'nor'.

TABLE IV
THE EXPERIMENTAL RESULTS OF THE MODEL FOR MULTICLASS CLASSIFICATION

		NO.B	AMD.D	DME.D	NOR.D	*Acc*			NO.B	AMD.D	DME.D	NOR.D	*Acc*
	Sub1 (AMD)	37	30	1	6	0.8108		Sub1 (AMD)	37	37	0	0	1
test1	Sub2 (DME)	61	3	57	1	0.9344	test6	Sub2 (DME)	61	7	54	0	0.8852
	Sub3 (NOR)	97	2	4	91	0.9381		Sub3 (NOR)	49	4	2	43	0.8776
	Sub1 (AMD)	37	17	16	4	0.4595		Sub1 (AMD)	37	31	6	0	0.8378
test2	Sub2 (DME)	61	0	60	1	0.9836	test7	Sub2 (DME)	61	12	49	0	0.8038
	Sub3 (NOR)	97	3	4	90	0.9278		Sub3 (NOR)	49	0	0	49	1
	Sub1 (AMD)	49	30	9	10	0.6122		Sub1 (AMD)	37	29	4	4	0.7838
test3	Sub2 (DME)	61	12	49	0	0.8033	test8	Sub2 (DME)	61	1	60	0	0.9836
	Sub3 (NOR)	97	2	2	93	0.9588		Sub3 (NOR)	97	1	5	91	0.9381
	Sub1 (AMD)	37	26	11	0	0.7027		Sub1 (AMD)	73	71	2	0	0.9726
test4	Sub2 (DME)	61	0	61	0	1	test9	Sub2 (DME)	31	1	30	0	0.9677
	Sub3 (NOR)	97	1	1	95	0.9794		Sub3 (NOR)	97	2	0	95	0.9794
	Sub1 (AMD)	49	28	13	8	0.5714		Sub1 (AMD)	37	34	2	1	0.9189
test5	Sub2 (DME)	61	0	61	0	1	test10	Sub2 (DME)	61	1	59	1	0.9672
	Sub3 (NOR)	97	6	0	91	0.9381		Sub3 (NOR)	97	0	0	97	1

NO.B – the number of B-scans in an OCT volume. AMD.D – the number of B-scans classified as AMD by the proposed algorithm. DME.D – the number of B-scans classified as DME by the proposed algorithm. NOR.D – the number of B-scans classified as normal by the proposed algorithm. Sub1 (AMD) (Sub2 (DME)) – the B-scans are annotated as AMD (DME). Sub3 (NOR) – the B-scans are annotated as normal. Acc – The ratio between the number of B-scans detected correctly by the algorithm and NO.B.

A. *Effectiveness of denoising*

In the proposed method, denoising is first employed to reduce the noise. To verify the effectiveness of denoising, we compare with the results of using the original images to generate the surrogates. The results are summarized in Table V. It is observed that the results of using denoised OCT images to generate the surrogates are better than the results of using original OCT images to generate the surrogates. The reason is that maximal intensity and minimal intensity of the pixels in a small region are employed as features for generating the surrogates. Imagine that if there is speckle noise in the images, the maximal and minimal intensity of pixels in the small regions from normal images and abnormal images may be no significant difference, which may lead to no significant difference between the surrogates of normal images and the surrogates of abnormal images, which is detrimental to the performance of the CNN model. Hence, denoising is a helpful step to improve the performance of the proposed method.

TABLE V
RESULTS OF USING THE DENOISED IMAGES AND ORIGINAL IMAGES TO GENERATE THE SURROGATES RESPECTIVELY.

	Local Database		Duke Database	
	Denoised	Original	Denoised	Original
AUC	0.9783	0.9687	0.9856	0.9707
Acc	0.9286	0.9196	0.9509	0.9264
Sen	0.9237	0.9237	0.9639	0.9021
Spe	0.9340	0.9151	0.9360	0.9588

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2795545, IEEE Journal of Biomedical and Health Informatics

B. Effectiveness of the ensemble approach

The final prediction for a test image is determined by the average of outputs from the trained CNN model on the surrogate images. To verify the effectiveness of this ensemble approach, we draw the ROC curves of the ensemble approach and the ROC curves of using single surrogate for the CNN to make the final prediction in Fig. 12. The dotted lines in Fig. 12 are the ROC curves of using single surrogate, and the red solid line are the ROC curves of the ensemble approach. It is observed that the AUC of the ensemble approach is larger than the AUC of using any single surrogate, which means that the proposed ensemble approach is effective.

C. Effectiveness of surrogate-assisted classification

To verify the effectiveness of the proposed surrogate-assisted classification method, we compare with the results of using raw images to train and test the CNN. Table VI summarizes the results. It is observed that when the number of images is large (there are 11904 B-scans in the local database), the results of using the surrogates are very competitive to the results of using the raw images. When the number of images is small (there are 3231 B-scans in the Duke database), the results can be improved by the proposed surrogate-assisted classification method compared with using the raw images to train the CNN model.

TABLE VI
RESULTS OF USING THE SURROGATE IMAGES AND RAW IMAGES TO TRAIN AND TEST THE CNN MODEL RESPECTIVELY.

	Local Database		Duke Database	
	surrogates	raw images	surrogates	raw images
AUC	0.9783	0.9800	0.9856	0.9491
Acc	0.9286	0.9297	0.9509	0.9205
Sen	0.9237	0.9576	0.9639	0.9059
Spe	0.9340	0.8986	0.9360	0.9371

D. Performance in volume level

It is noteworthy to point out that the analysis of a single B-scan is often not sufficient for the diagnosis of retinal diseases. Srinivasan et al. [13] advocated utilizing several B-scans for the detection of retinal diseases. According to this advocating, in this work, the category of a volume is determined by the maximum value among the number of B-scans of different categories detected by the proposed method. For example, for the Sub1(AMD) in test1 in Table IV, the AMD.D is 30, DME.D is 1 and NOR.D is 6, thus the subject is classified as affected by AMD. Based on this protocol and the results shown in Table IV, it is observed that the accuracy obtained by the proposed method is 100% in the volume level.

Table VII summarises the results obtained by the proposed method and the method proposed by Srinivasan et.al. [13] in the Duke database. Although promising results can be obtained by the proposed method in volume level, it is time-consuming. For a volume consisting of 97 B-scans, it will take about 10min to make a diagnosis. Future work will contain make the aided diagnosis faster.

TABLE VII
COMPARISONS WITH THE METHOD PROPOSED BY SRINIVASAN ET.AL. [13] IN VOLUME LEVEL IN DUKE DATASET.

	AMD	DME	Normal	average
proposed	100%	100%	100%	100%
Srinivasan et.al. [13]	100%	100%	86.67%	95.56%

In summary, in this paper, we propose a surrogate-assisted classification method for the automatic classifying the retinal OCT images based on CNNs. The results show that the proposed method is reliable since it works properly on different databases.

ACKNOWLEDGMENT

We would like to thank the three anonymous reviewers whose remarks and suggestions have helped us greatly in improving the quality of the paper.

REFERENCES

[1] M. D. Abramoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Reviews in Biomedical Engineering*, vol. 3, no. 1, pp. 169–208, December 2010.

[2] Z. Zhang, R. Srivastava, H. Liu, X. Chen, L. Duan, D. W. K. Wong, C. K. Kwoh, T. Y. Wong, and J. Liu, "A survey on computer aided diagnosis for ocular diseases," *BMC Medical Informatics & Decision Making*, vol. 14, no. 1, pp. 169–176, 2014.

[3] G. Quellec, K. Lee, M. Dolejsi, M. K. Garvin, M. D. Abramoff, and M. Sonka, "Three-dimensional analysis of retinal layer texture: identification of fluid-filled regions in sd-oct of the macula," *IEEE transactions on medical imaging*, vol. 29, no. 6, pp. 1321–1330, June 2010.

[4] X. Xu, K. Lee, L. Zhang, M. Sonka, and M. D. Abràmoff, "Stratified sampling voxel classification for segmentation of intraretinal and subretinal fluid in longitudinal clinical oct data," *IEEE transactions on medical imaging*, vol. 34, no. 7, pp. 1616–1623, July 2015.

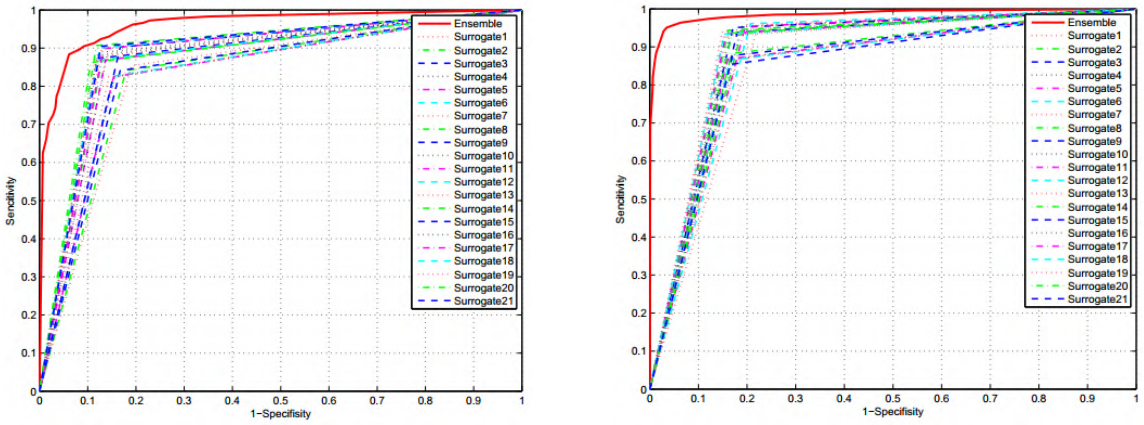
[5] G. R. Wilkins, O. M. Houghton, and A. L. Oldenburg, "Automated segmentation of intraretinal cystoid fluid in optical coherence tomography," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 1109–1114, April 2012.

[6] X. Chen, M. Niemeijer, L. Zhang, and K. Lee, "Three-dimensional segmentation of fluid-associated abnormalities in retinal oct: Probability constrained graph-search-graph-cut," *IEEE transactions on medical imaging*, vol. 31, no. 8, pp. 1521–1531, August 2012.

[7] A. G. Roy, S. Conjeti, S. 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." *Biomedical Optics Express*, vol. 8, no. 8, pp. 3627–3642, 2017.

[8] B. Liefers, F. G. Venhuizen, V. Schreur, B. van Ginneken, C. Hoyng, S. Fauser, T. Theelen, and C. I. Sánchez, "Automatic detection of the foveal center in optical coherence

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2795545, IEEE Journal of Biomedical and Health Informatics



(a) Local Database

(b) Duke Database

Fig. 12. The ROC curves.

tomography," *Biomedical Optics Express*, vol. 8, no. 11, pp. 5160–5178, 2017.

[9] 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." *Biomedical Optics Express*, vol. 8, no. 5, pp. 2732–2744, 2017.

[10] X. Sui, Y. Zheng, B. Wei, H. Bi, J. Wu, X. Pan, Y. Yin, and S. Zhang, "Choroid segmentation from opticalcoherence tomography with graph-edge weights learned from deep convolutional neural networks," *Neurocomputing*, vol. 237, no. C, pp. 332–341, 2017.

[11] K. Alsaih, G. Lemaître, J. M. Vall, M. Rastgoo, D. Sidibé, T. Y. Wong, E. Lamoureux, D. Milea, C. Y. Cheung, and F. Mériaudeau, "Classification of sd-oct volumes with multi pyramids, lbp and hog descriptors: application to dme detections," in *2016 IEEE 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 1344–1347.

[12] S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, E. Yuan, J. A. Izatt, C. A. Toth, A.-R. E. D. S. . A. S. D. O. C. T. S. Group *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.

[13] P. P. Srinivasan, L. A. Kim, P. S. Mettu, S. W. Cousins, G. M. Comer, J. A. Izatt, and S. Farsiu, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," *Biomedical optics express*, vol. 5, no. 10, pp. 3568–3577, 2014.

[14] 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," *Medical Image Analysis*, vol. 15, no. 5, pp. 748–759, 2011.

[15] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, and J. Cuadros, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[16] C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration optical coherence tomography images," *Ophthalmology Retina*, vol. 124, no. 8, pp. 1090–1095, 2017.

[17] P. Burlina, D. E. Freund, N. Joshi, Y. Wolfson, and N. M. Bressler, "Detection of age-related macular degeneration via deep learning," in *IEEE International Symposium on Biomedical Imaging*, 2016, pp. 184–188.

[18] S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," *Biomedical Optics Express*, vol. 8, no. 2, pp. 579–592, 2017.

[19] M. D. Abramoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning." *Investigative Ophthalmology & Visual Science*, vol. 57, no. 13, pp. 5200–5206, 2016.

[20] N. Patton, T. M. Aslam, T. Macgillivray, I. J. Deary, B. Dhillon, R. H. Eikelboom, K. Yogesan, and I. J. Constable, "Retinal image analysis: concepts, applications and potential." *Progress in Retinal & Eye Research*, vol. 25, no. 1, pp. 99–127, 2006.

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

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on*

Pattern Analysis & Machine Intelligence, vol. 38, no. 1, pp. 142–158, January 2016.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[25] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[26] Z. H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Taylor & Francis, 2012, pp. 1–20.

[27] A. Pizurica, L. Jovanov, B. Huysmans, V. Zlokolica, P. De Keyser, F. Dhaenens, and W. Philips, "Multiresolution denoising for optical coherence tomography: A review and evaluation," *Current Medical Imaging Reviews*, vol. 4, no. 4, pp. 270–284, 2008.

[28] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 12, no. 7, pp. 629–639, July 1990.

[29] D. C. Fernndez, "Delineating fluid-filled region boundaries in optical coherence tomography images of the retina," *IEEE Transactions on Medical Imaging*, vol. 24, no. 8, pp. 929–945, August 2005.

[30] M. Gargesha, M. W. Jenkins, A. M. Rollins, and D. L. Wilson, "Denoising and 4d visualization of oct images," *Optics Express*, vol. 16, no. 16, pp. 12 313–12 333, 2008.

[31] F. Shi, X. Chen, H. Zhao, W. Zhu, D. Xiang, E. Gao, M. Sonka, and H. Chen, "Automated 3-d retinal layer segmentation of macular optical coherence tomography images with serous pigment epithelial detachments." *IEEE Transactions on Medical Imaging*, vol. 34, no. 2, pp. 441–452, February 2014.

[32] L. Fang, S. Li, R. P. McNabb, Q. Nie, A. N. Kuo, C. A. Toth, J. A. Izatt, and S. Farsiu, "Fast acquisition and reconstruction of optical coherence tomography images via sparse representation," *IEEE transactions on medical imaging*, vol. 32, no. 11, pp. 2034–2049, November 2013.

[33] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[34] R. C. Gonzalez and P. Wintz, *Digital image processing*. Addison-Wesley, 1977, pp. 519–560.

[35] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2014.

[36] Y. L. Boureau, J. Ponce, and Y. Lecun, "A theoretical analysis of feature pooling in visual recognition," in *International Conference on Machine Learning*, 2010, pp. 111–118.

[37] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," *Journal of Machine Learning Research*, vol. 15, pp. 215–223, 1991.

[38] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[40] X.-S. Wei, B.-B. Gao, and J. Wu, "Deep spatial pyramid ensemble for cultural event recognition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 38–44.

[41] G. Montavon, G. Orr, and K. R. Mller, *Neural Networks: Tricks of the trade*. Springer Berlin Heidelberg, 2012, pp. 7–735.

[42] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[43] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[44] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, pp. 224–249.