

1 Computerized Macular Pathology Diagnosis in Spectral Domain Optical Coherence Tomography  
2 Scans Based on Multi-Scale Texture and Shape Features

3

4 Yu-Ying Liu,<sup>1</sup> Hiroshi Ishikawa,<sup>2,3</sup> Mei Chen,<sup>4</sup> Gadi Wollstein,<sup>2</sup> Jay S. Duker,<sup>5</sup> James G. Fujimoto,<sup>6</sup>  
5 Joel S. Schuman,<sup>2,3</sup> James M. Rehg<sup>1</sup>

6

7 <sup>1</sup> School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA;

8 <sup>2</sup>UPMC Eye Center, Eye and Ear Institute, Ophthalmology and Visual Science Research Center,  
9 Department of Ophthalmology, University of Pittsburgh School of Medicine, Pittsburgh, PA;

10 <sup>3</sup>Department of Bioengineering, Swanson School of Engineering, University of Pittsburgh;

11 <sup>4</sup>Intel Science & Technology Center, Carnegie Mellon University, Pittsburgh, PA;

12 <sup>5</sup>New England Eye Center, Tufts Medical Center, Tufts University School of Medicine, Boston, MA;

13 <sup>6</sup>Department of Electrical Engineering and Computer Science and Research Laboratory of  
14 Electronics, Massachusetts Institute of Technology, Cambridge, MA.

15

16 Early results presented in part at the Annual Meeting of the Association for Research in Vision and

17 Ophthalmology, Fort Lauderdale, Florida, May 2010, and the International Conference on Medical

18 Image Computing and Computer Assisted Intervention, Beijing, China, September 2010.

19

20 Supported in part by National Institutes of Health contracts NIH R01-EY013178, R01-EY011289 and

21 P30-EY008098 (Bethesda, MD), The Eye and Ear Foundation (Pittsburgh, PA), Research to Prevent

22 Blindness, Inc. (New York, NY), and Intel Labs, Intel Corporation.

23

1 Financial Interest Disclosure: Drs. Ishikawa, Schuman and Wollstein received royalties for  
2 intellectual property licensed by University of Pittsburgh to Bioptigen, Inc. Dr. Wollstein received  
3 research funding from Carl Zeiss Meditec and Optovue. Dr. Duker received research funding from  
4 Topcon Medical Systems, and is a consultant to Alcon, Genetech and Ophthotech and is a member of  
5 the advisory board of Paloma Pharmaceuticals. Dr. Fujimoto is a scientific advisor and has stock  
6 options in Optovue. Drs. Fujimoto and Schuman receive royalties for intellectual property licensed by  
7 Massachusetts Institute of Technology to Carl Zeiss Meditec.

8

9 Corresponding author: Hiroshi Ishikawa, UPMC Eye Center, Eye and Ear Institute, Ophthalmology  
10 and Visual Science Research Center, Department of Ophthalmology, University of Pittsburgh School  
11 of Medicine, Pittsburgh, PA, [ishikawah@upmc.edu](mailto:ishikawah@upmc.edu).

12

13 Word Count: 3942

14 Includes 7 Table and 7 Color Figures

15

1 **Abstract**

2 **Purpose:** To develop an automated method to identify the normal macula and three macular  
3 pathologies (macular hole (MH), macular edema (ME), and age-related macular degeneration  
4 (AMD)) from the fovea-centered cross sections in three-dimensional (3D) spectral domain optical  
5 coherence tomography (SD-OCT) images.

6 **Methods:** A sample of SD-OCT macular scans (Macular Cube 200x200 scan protocol; Cirrus HD-  
7 OCT; Carl Zeiss Meditec, Inc., Dublin, CA) were obtained from healthy subjects and subjects with  
8 MH and/or ME and/or AMD (dataset for development (326 scans from 136 subjects (193 eyes)), and  
9 dataset for testing (131 scans from 37 subjects (58 eyes))). Fovea-centered cross-sectional slice for  
10 each of SD-OCT images was encoded using spatially-distributed multi-scale texture and shape  
11 features. Three ophthalmologists labeled each fovea-centered slice- independently and the majority  
12 opinion for each pathology was used as the ground truth. Machine learning algorithms were used to  
13 identify the discriminative features automatically. Two-class Support Vector Machine classifiers were  
14 trained to identify the presence of normal macula and each of the three pathologies separately. The  
15 area under the receiver operating characteristic curve (AUC) was calculated to assess the  
16 performance.

17 **Results:** The cross-validation AUC result on the development dataset was 0.976, 0.931, 0.939, and  
18 0.938, and the AUC result on the hold-out testing set was 0.978, 0.969, 0.941, and 0.975, for  
19 identifying normal macula, MH, ME, and AMD, respectively.

20 **Conclusion:** The proposed automated data-driven method successfully identified various macular  
21 pathologies (all AUC > 0.94). This method may effectively identify the discriminative features  
22 without relying on a potentially error-prone segmentation module.

23

## 1 INTRODUCTION

2

3 Spectral-domain optical coherence tomography (SD-OCT) is a non-contact, non-invasive  
4 three-dimensional (3D) imaging technique, which performs optical sectioning at micron resolution. It  
5 is widely used in ophthalmology for identifying the presence of disease and its progression.<sup>1</sup> This  
6 technology measures the optical back scattering of the tissues, making it possible to visualize  
7 intraocular structures and diagnose ocular diseases, such as glaucoma and macular hole, objectively  
8 and quantitatively.

9 Although OCT imaging technology continues to evolve, technology for automated OCT  
10 image analysis and interpretation has not kept pace. With OCT data being generated at increasingly  
11 larger amount and higher sampling rates, there is a strong need for automated analysis to support  
12 disease diagnosis and tracking. This need is further amplified by the fact that an ophthalmologist  
13 making a diagnosis under standard clinical conditions does not have assistance of a specialist in  
14 interpreting OCT data beforehand. A software system which is capable of automated interpretation  
15 can potentially assist clinicians in making clinical decisions efficiently in busy daily routines.

16 To our knowledge, *there has been no prior work on automated macular pathology*  
17 *identification in OCT images, with the goal to directly predict the presence probability for each*  
18 *macular pathology in a given cross-sectional frame*; this automated method can be helpful to support  
19 disease diagnosis especially in situations where qualified readers are not easily accessible.

20 Automated pathology identification in ocular OCT images is complicated by three factors.  
21 First, the co-existence of pathologies accompanying with other pathological changes (epiretinal  
22 membrane, vitreous hemorrhage, etc.) can complicate the overall appearance, making it challenging  
23 to model each pathology individually. Second, there is high appearance variability within each

1 pathology, e.g., in macular hole cases, the holes can have different widths, depths, and shapes, and  
2 some can be covered by incompletely detached tissues, making explicit pathology modeling difficult.  
3 Third, the measurement of reflectivity of the tissue is affected by the optical properties of the  
4 overlying tissues, e.g., opaque media in the vitreous area or blood vessels around retinal surfaces will  
5 block or absorb much of the transmitted light respectively, and thus produce shadowing effects. As a  
6 result of these factors, attempts to hand-craft a set of features or rules to identify each pathology are  
7 unlikely to generalize well. Instead, direct encoding of the statistical distribution of low-level image  
8 features and training discriminative classifiers based on a large expert labeled dataset might achieve  
9 more robust performance.

10 In this study, a machine learning based method for automatically identifying the presence of  
11 pathologies from a fovea-centered cross section in a macular SD-OCT scan was developed.  
12 Specifically, the presence of the normal macula (NM) and each of the following three macular  
13 pathologies: macular hole (MH), macular edema (ME), and age-related degeneration (AMD) were  
14 identified separately in the cross section through the foveal center. This single-frame based method  
15 can serve as a basic component for examining the complete set of frames from the volume.

16 In this work, the automated software that makes diagnostic suggestions is solely based on the  
17 interpretation of image appearances, so as to serve as a stand-alone component for OCT image  
18 interpretation. Note that for a true clinical diagnosis, all of the available information (e.g. the results  
19 of OCT image analysis in conjunction with other ancillary tests) would be considered together to  
20 make the final diagnostic decision.

21 A preliminary version of this work was presented in our prior paper.<sup>2</sup> This paper significantly  
22 extends our previous publication in several areas: improved automated method, detailed labeling  
23 agreement analysis among three ophthalmologists, new ground truth based on majority opinion and

1 complete consensus, evaluation of our original and new method, and several additional experiments,  
2 such as effect of training set size, effect of data with inconsistent labeling, and performance on a  
3 separate testing dataset collected after the method development stage, which is representative of  
4 future unseen data.

5

## 6 **METHODS**

7

### 8 **Subjects and Image Acquisition**

9       The study subjects were enrolled at the University of Pittsburgh Medical Center Eye Center or  
10 at the New England Eye Center. All subjects had comprehensive ophthalmic examination followed  
11 by SD-OCT macular cube scan (Cirrus HD-OCT; Carl Zeiss Meditec, Dublin, CA). The training  
12 dataset (named Dataset A), consisting of 326 macular SD-OCT scans from 136 subjects (193 eyes),  
13 was used for deriving the best algorithmic and parameter settings by cross-validation. The testing  
14 dataset (named Dataset B), containing another 131 macular SD-OCT scans from 37 subjects (58 eyes)  
15 collected after the method development stage, was used for testing the performance on novel images.

16       Since the OCT manufacturer's recommended signal strength (SS) is 8 or above in 1-10 scale,  
17 all of our enrolled images were qualified  $SS \geq 8$  criteria. The original scan density was either  
18  $200 \times 200 \times 1024$  or  $512 \times 128 \times 1024$  samplings in  $6 \times 6 \times 2$  mm volumes. All horizontal cross-section  
19 images were rescaled to  $200 \times 200$  for computational efficiency. For each of the scans, the horizontal  
20 cross section through the foveal center was then manually selected by one expert ophthalmologist,  
21 and this image served as the basis for analysis in this study.

22       The study was approved by the Institutional Review Board committees of the University of  
23 Pittsburgh, Pittsburgh, PA and Tufts Medical Center, Boston, MA, and adhered to the Declaration of

1 Helsinki and Health Insurance Portability and Accountability Act regulations, with informed consent  
2 obtained from all subjects.

3

#### 4 **Subjective Classification of Images**

5 A group of OCT experts masked to any clinical information independently identified the  
6 presence or absence of normal macula and each of MH, ME and AMD in the fovea-centered frame.  
7 Note that a combination of pathologies can coexist in one cross section. For MH category, both  
8 macular hole and macular pseudohole were included in order to simplify the discrimination of all  
9 "hole-like" structures from the other cases. Dedicated labeling software was developed where only  
10 the pre-selected fovea-centered frame was presented in a randomized order.

11 For dataset A, three OCT experts gave the pathology labels for each scan, and the majority  
12 opinion of the three experts was identified for each pathology and used as the "ground truth" in our  
13 method development stage. For dataset B, two of the three experts provided the labels for each scan.  
14 For each pathology, the scans with consistent labels were selected for performance evaluation while  
15 the scans with different labels were excluded.

16

#### 17 **Automated Classification Method**

18 Our automated method encodes the appearance properties of the retinal images directly, by  
19 constructing a global image descriptor based on spatially-distributed multi-scale texture and shape  
20 features, combined with machine learning techniques to automatically learn the classifiers for  
21 identifying each pathology from a large expert labeled training set. This method does not rely on a  
22 retinal segmentation module to extract the features, and thus avoids a major source of analysis failure.

1           A previous formulation of our automated method, which was previously published,<sup>2</sup> uses only  
2 texture features. Here, we extend and enhance the approach by incorporating the shape property of  
3 the retinal images in addition to the texture. In brief, our method consists of three main steps as  
4 illustrated in Fig. 1. First, image alignment is performed to remove the curvature and center the image  
5 in order to reduce the appearance variation across scans. Second, a global image descriptor is  
6 constructed from the aligned image and its derived Canny edge image.<sup>3</sup> Multi-scale spatial pyramid  
7 (MSSP)<sup>4</sup> is used as the global representation for capturing the spatial organization of the retina in  
8 multiple scales and spatial granularities. To encode each local block, the dimension-reduced local  
9 binary pattern (LBP) histogram<sup>5</sup> based on principle component analysis (PCA) is used as the local  
10 block descriptor. The local features derived from each spatial block in the multiple rescaled images  
11 and their edge images are concatenated in a fixed order to form the overall global descriptor. These  
12 histogram-based local features are utilized to encode the *texture and shape characteristics* of the  
13 retinal image, respectively. Finally, for each pathology, a two-class non-linear support vector  
14 machine (SVM)<sup>6</sup> classifier with radial basis function (RBF) kernel and estimated probability values is  
15 trained using the image descriptors and their labels from the training set.

16           For detailed information in validating each component (MSSP, LBP) of our approach, please  
17 refer to our prior study.<sup>2</sup>

18

## 19 **Experimental Settings**

20           In developing the method on dataset A, ten-fold cross validation was used at the subject level  
21 where all images from the same subject were put together in either the training or testing set in each  
22 fold. Note that by cross validation, each testing fold is novel with respect to its corresponding training  
23 fold. In our training phase, both the original image and its horizontal flip are used as the training

1 instances for enriching our training set. After running all ten-fold training and testing, the ten-fold  
2 testing results are aggregated and the area under the receiver operator characteristic curve (AUC) is  
3 computed. To get a more reliable assessment of the performance, 6 different random ten-fold data  
4 splitting were generated, and the above procedure is run for each of the six splits. The mean and  
5 standard deviation of the 6 AUCs were reported as the performance metric on the developing set.

6 To test the statistical significance of performance difference between any two algorithmic  
7 settings, DeLong<sup>7</sup> test was adopted to compare the two receiver operating characteristics (ROC)  
8 curves. If under DeLong test, one setting is better than the other ( $p \leq 0.05$ ) for all of the 6 different  
9 data splits, then the performance of the former setting is claimed better; otherwise, the performance of  
10 the two settings is declared no significant difference.

11 After performing detailed analysis on dataset A, the best algorithmic settings and parameters  
12 determined for identifying each pathology are then applied to the test dataset B\*. The performance on  
13 dataset B is thus representative for the generalization ability of the proposed approach. (\*: for MH  
14 category, unfortunately, dataset B didn't contain macular hole cases, which coincides with real  
15 clinical situations, since macular hole has low occurrence (approximately 3.3 cases in 1000 in those  
16 persons older than 55 years)<sup>8</sup>. To deal with this situation, for MH performance testing only, the  
17 training and testing dataset was reorganized such that 80% of MH cases originated from dataset A  
18 were randomly sampled and included in the training set and the rest were included in the testing set).

19

## 20 **RESULTS**

21

### 22 **Inter-Expert Labeling Agreement and the Ground truth on Dataset A**

1           The labeling agreement among the three experts on dataset A was illustrated in Fig. 2 using  
2 the Venn diagram. The complete agreement among the experts for NM, MH, ME, and AMD was  
3 96.9%, 91.1%, 80.1%, 87.1%, respectively. The Kappa statistics ( $k$ ) was calculated to assess the pair-  
4 wise experts' labeling agreement, as listed in Table 1. All kappa values for identification of normal  
5 macula were high (all  $k > 0.93$ ) and for MH, the kappa value from one expert-pair (expert 1 and 2)  
6 was high (0.92). However, all kappa values for ME and AMD were within 0.61-0.80 range, which  
7 represented substantial but imperfect agreement.

8           The majority opinion of the image labeling was used as the ground truth so that the standard is  
9 not biased towards any specific expert. The number of images for each macular category as defined  
10 by the ground truth was listed in Table 2.

11           In order to further assess how many positive and negative cases in our ground truth result  
12 from inconsistent labeling, in Fig. 3, the statistics and several representative examples were shown for  
13 each pathology where all three experts, only two experts, or just one expert gave the “positive” label.  
14 Note that the images labeled as positive by only one expert were treated as “negative” cases in our  
15 ground truth. It was found that the quantity of images having only one positive vote is considerable  
16 for ME and AMD (31 and 18 cases, respectively), revealing larger ambiguity in their identification.  
17

## 18 **Performance of Automated Classification Method on Dataset A**

19           Different feature settings: texture (T) alone, shape (S) alone, and in combination (TS) were  
20 tested on dataset A, so that the discriminative power of each feature type for each pathology can be  
21 evaluated. For shape features, the edge detection threshold, denoted as  $t$ , was tested at various values  
22 so that different quantities of edges were obtained and encoded (Fig. 4). The AUCs for the different  
23 feature settings were reported in Table 3. The best AUCs for NM, MH, ME, and AMD, were 0.976,

1 0.931, 0.939, and 0.938, derived from the setting: TS( $t=0.4$ ), S( $t=0.4$ ), TS( $t=0.4$ ), and TS( $t=0.2$ ),  
2 respectively; their ROC curves generated from one of six random data splits were shown in Fig. 5.

3       Regarding the edge detection thresholds  $t$  for shape features, it was discovered that for NM,  
4 ME, and AMD, the AUC results under different  $t$  settings were all within 1% in AUC; but for MH,  
5 the performance is much more sensitive to the choice of  $t$  (AUC was 0.888, 0.901, 0.931, and 0.911  
6 when  $t$  varied from 0.2 to 0.5) with the best performance at  $t=0.4$ ; this suggests that for MH,  
7 encoding the stronger edges is more helpful in identifying the hole structures; the weaker edges  
8 ( $t=0.2$ ) might add noises instead and distract the classifiers.

9       The statistical significance of the performance difference under different feature settings was  
10 also evaluated. It was found that for NM, TS outperformed T though the absolute gain is small (0.7%  
11 in AUC); thus, including shape features can provide additional useful information. For MH, S is  
12 significantly better than using T and TS, with a large AUC difference (8.5%) between S and T; this  
13 reveals that using shape feature alone is sufficient to capture the distinct contours of MH. For ME, T  
14 and TS was significantly better than using S (1.6% AUC difference), but TS and T has similar  
15 performance; this suggests that encoding the intensity patterns (textures) is more informative than just  
16 describing the shapes. For AMD, all three feature settings (T, S, TS) had no significant difference, but  
17 using combined features (TS) achieved the best AUC performance, suggesting that both feature types  
18 are useful.

19       In implementation, for NM, ME, and AMD, the feature vectors are computed from the aligned  
20 retinal image directly, which is 200 pixels in width; for MH, the features are extracted from the  
21 further down-sized image (rescaled to 100 pixels in width). This rescaling for MH improves the  
22 performance by 3% consistently under different feature type settings. This suggests that removing the  
23 details or noises residing in the original resolution can help identification of the hole structures.

1

**2 Performance Comparison between Experts and Automated Method on Dataset A**

3 To compare the labeling performance of the automated method to that of each expert against  
4 the majority-based ground truth, the balanced accuracy (average of sensitivity and specificity) of the  
5 automated method and each expert was computed. For the automated method, the best balanced  
6 accuracy was derived from the ROC curve. The results were detailed in Table 4. Overall, the  
7 automated analysis method achieved good balanced accuracy for NM (95.5%), but relatively lower  
8 performance for MH, ME, and AMD (89.7%, 87.3%, and 89.3%). The automated software was  
9 inferior to the experts in most cases, but when compared to expert 3 the performance differences were  
10 all within 5% for all categories (the difference is -3.9%, +3.2%, -4.4%, -2.7% for NM, MH, ME, and  
11 AMD, respectively).

12

**13 Performance Using Varied Training Set Size on Dataset A**

14 The AUC performances of the automated method with respect to varied training size on  
15 dataset A were also studied. The ten-fold cross-validation setting was still used, but now for each  
16 training fold, k% of positive and negative subjects was sampled and utilized for training, while the  
17 testing fold remained the same. The results with settings of k=10, 20, ..., 100 were plotted in Fig. 6.  
18 The AUC results of 10%, 50%, and 100% training set were 0.906, 0.970, 0.976 for NM, 0.745, 0.904,  
19 0.931 for MH, 0.868, 0.924, 0.939 for ME, and 0.746, 0.920, 0.938 for AMD. These results showed  
20 that using more training data can improve the performance in all categories. For MH, a larger gain  
21 (2.7%) and clearer increasing trend from 50% to 100% is observed, suggesting that adding more  
22 training instances for MH can improve the performance the most.

1 From the theoretical viewpoint, utilizing more training data is always desirable for learning-  
2 based approaches since this can help discover the true discriminative information from more  
3 representative images, mitigate the over-fitting problems, and thus achieve better generalization  
4 performance.

5

### 6 **Performance Using Only Images with Complete Consensus on Dataset A**

7 In order to understand the influence of cases where there is inconsistent labeling, an  
8 experiment using only images with complete labeling agreement for each pathology separately was  
9 also conducted. In this setting, 316 (96.9%), 297 (91.1%), 261 (80.1%), 284 (87.1%) images from  
10 original 326 images were selected for NM, MH, ME, and AMD identification, respectively (as  
11 illustrated in the Venn diagram in Fig. 3). The AUC results were listed in Table 5.

12 It was found that when using only images with complete consensus, the performance for NM  
13 and MH is slightly enhanced ( $\sim 1\%$ ) but it is much better for ME (from 0.939 to 0.985) and AMD  
14 (from 0.938 to 0.968). This suggests that the larger ambiguity in ME and AMD identification, as  
15 noted in their lower Kappa values, is indeed a major factor in influencing the performance of the  
16 automated method.

17

### 18 **Performance on the Separate Testing Dataset B**

19 To test the performance on the hold-out dataset B, the pathology classifiers were trained using  
20 the images from dataset A, with the best algorithmic settings determined in analyzing dataset A  
21 ( $TS(t=0.4)$ ,  $S(t=0.4)$ ,  $TS(t=0.4)$ , and  $TS(t=0.2)$  for NM, MH, ME, and AMD, respectively). For this  
22 experiment, the ground truth was defined by the consensus between the same two experts for both  
23 datasets. The consensus includes 96.9%, 95.4%, 88.0%, and 90.5% of 326 scans from dataset A for

1 training, and 94.7%, 100%, 90.0%, and 84.7% of 131 scans from dataset B for testing, for NM, MH,  
2 ME, and AMD, respectively. The pathology distribution for both datasets is detailed in Table 6. The  
3 AUC result and the ROC curve were shown in Table 7 and Fig. 7, respectively.

4 The AUC is 0.978, 0.969, 0.941, and 0.975, and the best balanced accuracy is 95.5%, 97.3%,  
5 90.5%, and 95.2% for NM, MH, ME, and AMD, respectively. The AUC performance on all  
6 pathologies are good ( $AUC > 0.94$ ) and comparable to the cross-validation AUC results on the training  
7 dataset A ( $AUC > 0.93$ ). Our results suggest that the proposed method is effective in identifying  
8 pathologies for future unseen images.

9

## 10 **DISCUSSION**

11

12 In this study, a machine-learning based approach was proposed to identify the presence of  
13 normal macula and several macular pathologies, MH, ME, and AMD, from a fovea-centered cross  
14 section in a macular SD-OCT scan. To our knowledge, this study is the first to automatically classify  
15 OCT images for various macular pathologies. A large dataset (dataset A) containing 326 scans from  
16 136 subjects with healthy macula or assorted macular pathologies was used for developing the  
17 methodology, and a separate dataset (dataset B), with 131 scans from 37 subjects, was used as a  
18 hold-out testing set.

19 On the developing dataset (dataset A), the performance of our automated analysis achieved  
20  $> 0.93$  AUC results for all macular pathologies using 10-fold cross-validation, with particularly good  
21 performance on identifying the normal macula ( $AUC = 0.976$ ). This can be attributed to the reduced  
22 variation in normal appearance across scans. For pathology identification, the performance reduced  
23 somewhat, likely due to the greater within-category appearance variations, lack of sufficient training

1 data especially for MH and AMD, and the ambiguity existing in the majority-opinion based ground  
2 truth as shown in the kappa agreement analysis between the experts.

3 By analyzing the performance on dataset A, the discriminative power of using texture or  
4 shape features alone as well as their combination was studied. It is discovered that under DeLong test,  
5 for MH the use of shape features is more effective than texture features, while for ME texture features  
6 outperforms shapes. This makes sense since macular holes are marked by the distinct contours of  
7 holes while detection of edema requires intensity comparison information, e.g., dark cystic areas  
8 embedded in the lighter retinal layers. For NM and AMD, the combined features achieved the highest  
9 AUC results, but this setting didn't significantly outperform using either feature alone. However, it is  
10 possible that when a larger training set is available, utilizing all complimentary features can result in  
11 superior performance, since then the over-fitting phenomenon in the high dimensional feature space  
12 can be mitigated and the true discriminative information can be more effectively represented.

13 The AUC results with respect to varied training set size (10%, 20%, ..., 100%) on dataset A  
14 were also presented. It was discovered that exploiting more training data can consistently enhance the  
15 performance in all categories, especially for MH. Training on additional MH cases can boost the  
16 performance the most.

17 To understand the influence of inconsistent labeling in our majority-opinion based ground  
18 truth from dataset A, the AUC results from using only images with complete consensus for each  
19 pathology were also presented. The much higher AUC results for ME and AMD (0.985 and 0.968,  
20 respectively) suggest that our current method is more effective when the two classes (presence and  
21 absence) can be well separated. However, in reality, there are always subtle cases residing in the gray  
22 area in between, causing ambiguity in dichotomy labeling. One possible future direction is to use  
23 refined labeling, e.g., by "pathological degree" ("absent", "early" or "advanced"), and to explore

1 whether this setting can result in improved labeling consistency and superior performance in  
2 automated software. This new methodology might demand a larger amount of training data in  
3 discriminating different pathological stages.

4 Our method achieved good AUC results ( $> 0.94$  for all pathology categories) on the hold-out  
5 testing set (dataset B), when using images from the developing dataset (dataset A) for classifier  
6 training. This performance is promising in classifying future unseen images.

7 The proposed method has several advantages. First, our histogram-based image features  
8 directly capture the statistical distribution of appearance characteristics, resulting in objective  
9 measurements and straightforward implementation. Second, our method is not limited to any one  
10 pathology and can be applied to identify additional pathologies. Third, the same approach can be  
11 utilized to examine other cross sections besides the foveal slice, as long as the labeled cross-sections  
12 from the desired anatomical location are also collected for training.

13 The limitation of the current study is that only the fovea-centered frame for each 3D SD-OCT  
14 scan was analyzed and that frame was manually selected. In practice every slice in the 3D scan data  
15 should be examined so that any abnormality can be identified, even when no pathology is observed at  
16 the fovea-centered frame(an unlikely event).This study is designed as a foundation for extending to  
17 analyzing each slice in the volume.

18 In our future work, the present slice diagnosis method will be extended to analyze each slice  
19 in the entire cube, once the pathology labeling for each cross section can be gathered. The most  
20 straightforward way is to train a set of y-location indexed pathology classifiers using the labeled slice  
21 set from the same quantized y location relative to the fovea. By using location-specific classifiers, the  
22 normal and abnormal anatomical structures around similar y locations can be modeled more  
23 accurately and the entire volume can be examined. Once the eye motion artifacts in the macular scans

1 can be reliably corrected, the efficacy of volumetric features will be investigated for pathology  
2 identification. An automated method for fovea localization is also desirable so that the entire process  
3 is fully-automated.

4 In conclusion, an effective approach was proposed to computerize diagnosis of multiple  
5 macular pathologies in retinal OCT images. Our results ( $AUC > 0.94$ ) demonstrate that the proposed  
6 spatially-distributed multi-scale texture and shape descriptors combined with a data-driven  
7 framework can effectively identify the discriminative features without relying on a potentially error-  
8 prone segmentation module. Our method may provide clinically useful tools to support disease  
9 identification, improving the efficiency of OCT based examination.

10

11

1 **References**

2

3 1. Schuman JS. Spectral domain optical coherence tomography for glaucoma. *Trans Am*  
4 *Ophthalmol Soc* 2008;106:426-458.

5 2. Liu Y-Y, Chen M, Ishikawa H, Wollstein G, Schuman J, Rehg JM. Automated Macular  
6 Pathology Diagnosis in Retinal OCT Images Using Multi-Scale Spatial Pyramid with Local Binary  
7 Patterns. *Intl Conf on Medical Image Computing and Computer Assisted Intervention* 2010;6361:1-9.

8 3. Canny J. A Computational Approach To Edge Detection. *IEEE Trans Pattern Analysis and*  
9 *Machine Intelligence* 1986;8:679-698.

10 4. Wu J, Rehg JM. Where am I: Place instance and category recognition using spatial {PACT}.  
11 *IEEE Computer Vision and Pattern Recognition* 2008;1-8.

12 5. Ojala T, Pietikainen M, Maenpaa T. Multiresolution Gray-Scale and Rotation Invariant  
13 Texture Classification with Local Binary Patterns. *IEEE Trans on Pattern Analysis and Machine*  
14 *Intelligence* 2002;24:971.

15 6. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. 2001.

16 7. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more  
17 correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*  
18 1988;44:837-845.

19 8. Luckie A, Heriot W. Macular holes. Pathogenesis, natural history, and surgical outcomes.  
20 *Aust N Z J Ophthalmol* 1995;23:93-100.

21

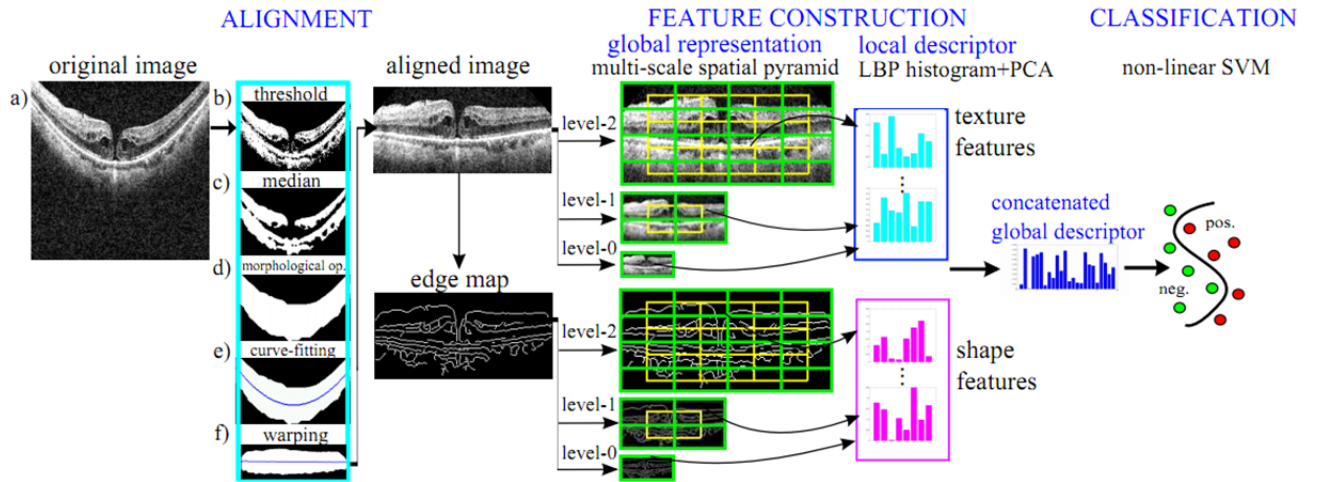
22

1 **Figures**

2 Figure 1.

3 Stages of our approach (morphological op.: morphological operations; LBP: local binary patterns;

4 PCA: principle component analysis; SVM: support vector machine).



5

6

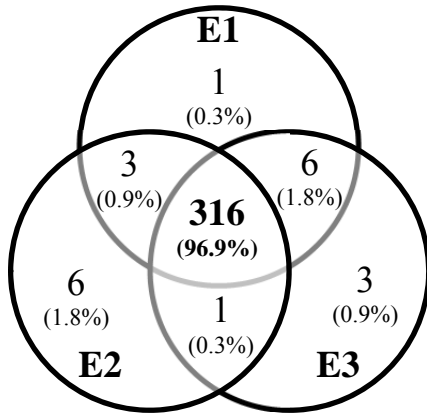
7 Figure 2.

8 On Dataset A: Venn diagram of labeling agreement among the three experts on all macular

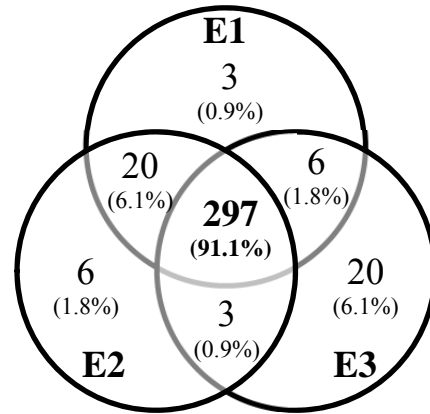
9 categories. The actual scan numbers and the percentage were both shown. E1-E3 represents expert 1

10 to expert 3.

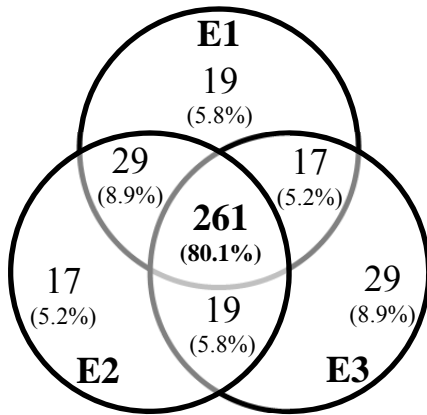
11



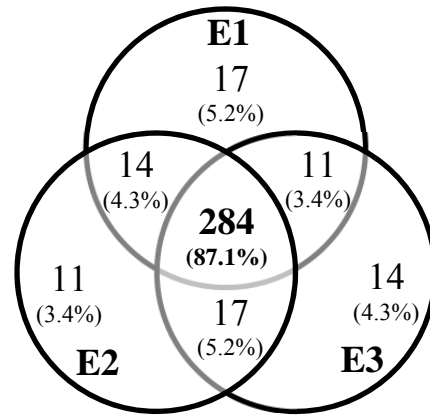
(a) Normal macula (NM)



(b) Macular hole (MH)



(c) Macular edema (ME)



(d) Age-related macular degeneration (AMD)

1

2 Figure 3.

3 On Dataset A: The number of cases and representative examples where all 3 experts, 2 experts, or

4 only one expert gave “positive” labels for the presence of normal macula and each pathology,

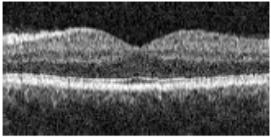
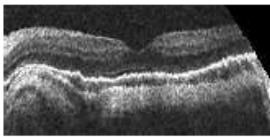
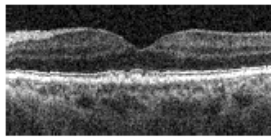
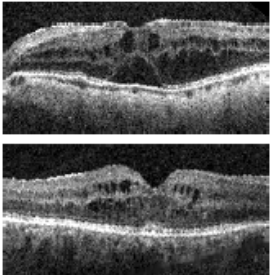
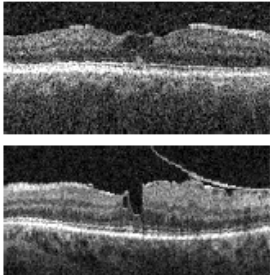
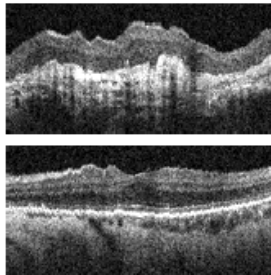
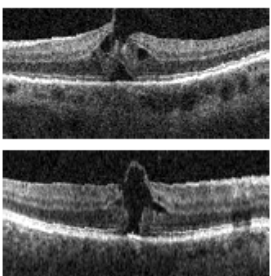
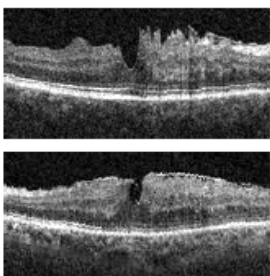
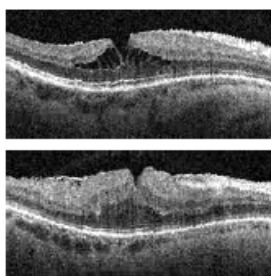
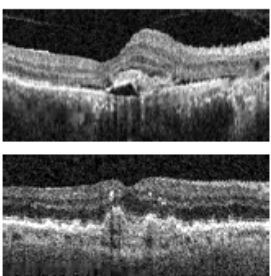
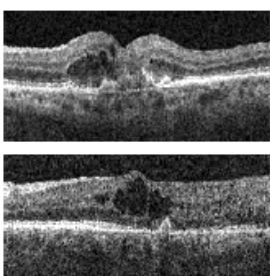
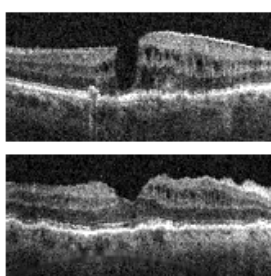
5 respectively. Note that images in the first two columns were defined as “positive” while the ones in

6 the last column were regarded as “negative” in our majority-opinion based ground truth. The images

7 without total agreement usually contain early pathologies which were subtle and occupied small

1 areas. (NM: normal macula, MH: macular hole, ME: macular edema, AMD: age-related macular  
 2 degeneration.)

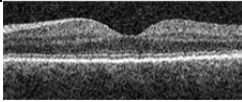


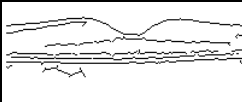

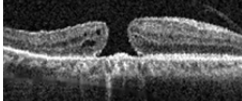



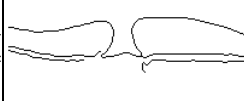
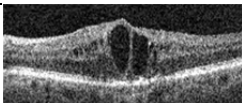



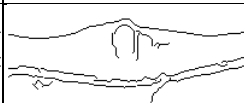
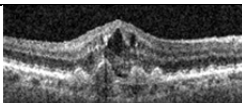



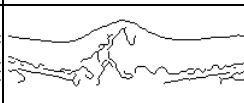
3

Cases	“Positive”		“Negative”
	3 pos. votes	2 pos. votes	1 pos. vote
NM	80 cases 	1 case 	9 cases 
ME	169 cases 	34 cases 	31 cases 
MH	53 cases 	21 cases 	8 cases 
AMD	50 cases 	24 cases 	18 cases 

4

5 Figure 4.

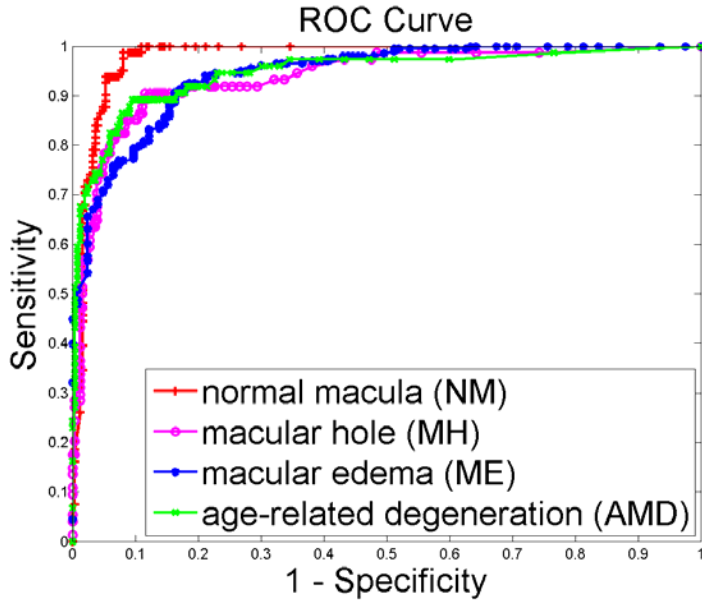
1 Examples of the aligned retinal images and their Canny edge maps derived under different edge  
 2 detection thresholds  $t$  for each macular category. The smaller value of  $t$ , the more edges are retained.  
 3 (NM: normal macula, MH: macular hole, ME: macular edema, AMD: age-related macular  
 4 degeneration.)

	Aligned image	Edge ( $t = 0.2$ )	Edge ( $t = 0.3$ )	Edge ( $t = 0.4$ )	Edge ( $t = 0.5$ )
NM					
MH					
ME					
AMD					

5

6 Figure 5.

7 On Dataset A: ROC curve of one run of 10-fold cross validation on all images in dataset A. The best  
 8 feature setting for each macular pathology was used. (Feature setting: TS ( $t = 0.4$ ), S ( $t=0.4$ ), TS, and  
 9 TS ( $t = 0.2$ ) for normal macula (NM), macular hole (MH), macular edema (ME), and age-related  
 10 macular degeneration (AMD), respectively.)

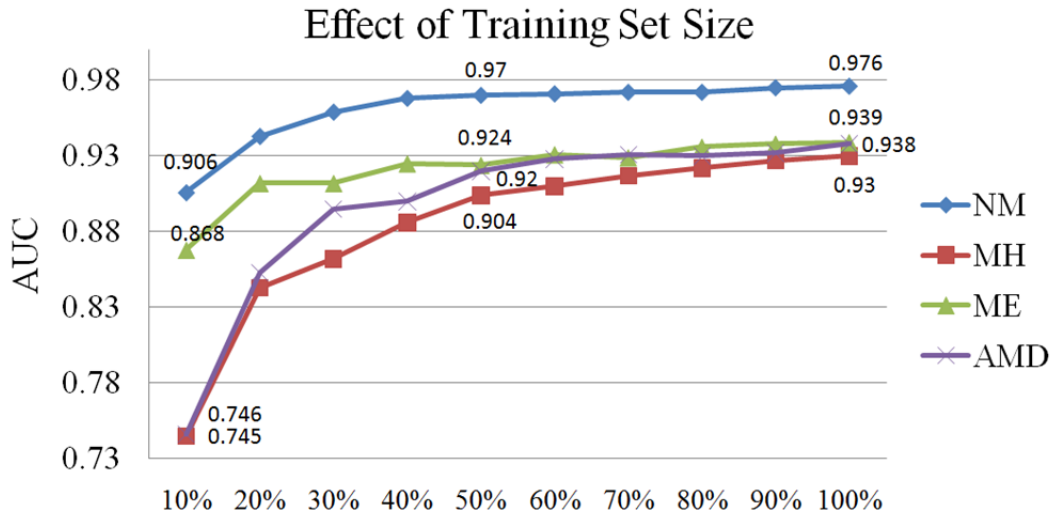


1

2 Figure 6.

3 On Dataset A: AUC results with respect to varied training set size from dataset A. For each training  
 4 fold, 10%, 20%, ..., 100% of the positive and negative subjects were sampled and used for training  
 5 while the testing fold was unchanged. (Feature setting: TS ( $t = 0.4$ ), S ( $t=0.4$ ), TS, and TS ( $t = 0.2$ ) for  
 6 normal macula (NM), macular hole (MH), macular edema (ME), and age-related macular  
 7 degeneration (AMD), respectively.)

8

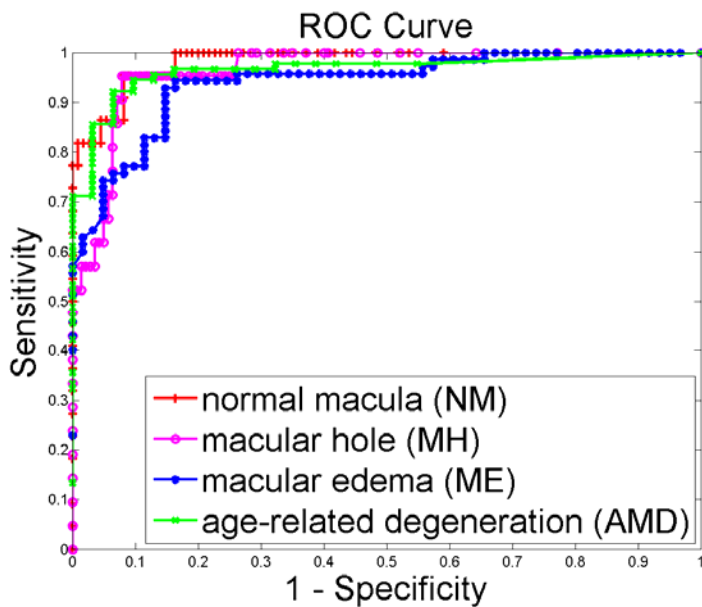


9

1

2 Figure 7.

3 On Dataset B: ROC curve of testing on dataset B, based on the pathology classifiers trained using  
 4 images from dataset A. The ground truth for this experiment was defined by the consensus of the two  
 5 experts (expert 1 and 2) on both datasets. The statistics of pathology distribution was listed in Table  
 6 6. The feature and parameter setting for each pathology was determined using dataset A only.  
 7 (Feature setting: TS ( $t = 0.4$ ), S ( $t=0.4$ ), TS, and TS ( $t = 0.2$ ) for normal macula (NM), macular hole  
 8 (MH), macular edema (ME), and age-related macular degeneration (AMD), respectively.)



9

1 **Tables**

2

3 Table 1.

4 On Dataset A: Kappa values ( $k$ ) of pair-wise expert labeling agreement for normal macula (NM),  
 5 macular hole (MH), macular edema (ME), and age-related macular degeneration (AMD) on dataset  
 6 A. E1-E3 represents expert 1-3. The values within 0.60-0.80 represent substantial but imperfect  
 7 agreements.

Kappa ( $k$ )	NM	MH	ME	AMD
E1, E2	0.94	0.92	0.76	0.76
E1, E3	0.97	0.78	0.69	0.73
E2, E3	0.93	0.76	0.71	0.77

8

9 Table 2.

10 On Dataset A: Number of positive scans, eyes and subjects from dataset A (a total of 326 scans from  
 11 136 subjects(193 eyes)) as defined by the majority opinion of the three experts for normal macula  
 12 (NM), macular hole (MH), macular edema (ME), and age-related macular degeneration (AMD). Note  
 13 that each scan was labeled with coinciding macular findings.

<b>Statistics</b>	NM	MH	ME	AMD
Scan	81	74	203	74
Eye	66	36	116	37
Subject	65	33	90	26

14

15 Table 3.

1 On Dataset A: AUC results (with standard deviation) of texture features (T), shape features (S), and  
 2 their combinations (TS) under the best edge detection threshold  $t$ . The best results for each macular  
 3 category were shown in bold face. The last column shows the results of statistical significant test  
 4 using DeLong test at  $p = 0.05$  significant level, where “<” represents that the right setting performs  
 5 better than the left, and “ $\approx$ ” represents no significant differences. (NM: normal macula, MH: macular  
 6 hole, ME: macular edema, AMD: age-related macular degeneration.)

AUC	Texture (T)	Shape (S)	Texture + Shape (TS)	Significant Test
NM	0.969±0.002	0.971±0.002 (t=0.4)	<b>0.976±0.002</b> (t=0.4)	T $\approx$ S, <b>T &lt; TS</b> , S $\approx$ TS
MH	0.846±0.011	<b>0.931±0.005</b> (t=0.4)	0.919±0.005 (t=0.4)	<b>T &lt; S</b> , <b>T &lt; TS</b> , S $\approx$ TS
ME	<b>0.939±0.004</b>	0.923±0.005 (t=0.3)	<b>0.939±0.004</b> (t=0.4)	<b>S &lt; T</b> , T $\approx$ TS, <b>S &lt; TS</b>
AMD	0.925±0.008	0.931±0.005 (t=0.2)	<b>0.938±0.006</b> (t=0.2)	T $\approx$ S $\approx$ TS
Average	0.908±0.008	0.932±0.005	<b>0.936±0.006</b>	

7

8 Table 4.

9 On Dataset A: Balanced accuracy of each of the three experts and the automated method against the  
 10 majority-opinion based *ground truth*. In each cell, the *balanced accuracy* was listed along with the  
 11 corresponding *sensitivity* and *specificity* values in the parenthesis. For the automated method, the best  
 12 feature setting for each pathology was adopted (TS(t=0.4), S(t=0.4), TS, TS(t=0.2) for normal macula  
 13 (NM), macular hole (MH), macular edema (ME), age-related macular degeneration (AMD),  
 14 respectively); the best balanced accuracy was derived from the mean of the outputs of the 6 runs.

<b>B. Accuracy</b>	Expert 1	Expert 2	Expert 3	Auto. method
NM	<b>99.8</b> (100, 99.6)	98.4 (98.8, 98.0)	99.4 (100, 98.8)	95.5 (99.4, 91.5)
MH	<b>99.4</b> (100, 98.8)	98.3 (98.6, 98.0)	86.5 (73.0, 100)	89.7 (89.1, 90.3)

ME	92.4 (99.5, 85.4)	<b>94.9</b> (94.6, 95.1)	91.7 (89.2, 94.3)	87.3 (87.5, 87.0)
AMD	<b>94.2</b> (93.2, 95.2)	94.0 (89.2, 98.8)	92.0 (85.1, 98.8)	89.3 (89.7, 88.8)
Average	<b>96.5</b> (98.2, 94.8)	96.4 (95.3, 97.5)	92.4 (86.8, 98.0)	90.5 (91.4, 89.4)

1

2 Table 5.

3 On Dataset A: AUC results of using the entire dataset (326 images) in comparison with that of using  
4 only the images of complete consensus from the 3 experts for each pathology separately. (316  
5 (96.9%), 297 (91.1%), 261 (80.1%), and 284 (87.1%) images were included for normal macula (NM),  
6 macular hole (MH), macular edema (ME), and age-related macular degeneration (AMD),  
7 respectively).

<b>AUC on Dataset A</b>	NM	MH	ME	AMD
All images (326 scans)	0.976	0.931	0.939	0.938
Images of Complete Consensus from 3 Experts (316, 297, 261, 284 scans)	0.984	0.932	0.985	0.968

8

9 Table 6.

10 On dataset A (for training) and Dataset B (for testing): number of positive scans, eyes and subjects  
11 versus the total cases as defined by the consensus of the two experts (expert 1 and 2) for normal  
12 macula (NM), macular hole (MH), macular edema (ME), and age-related macular degeneration  
13 (AMD). The consensus includes 96.9%, 95.4%, 88.0%, and 90.5% of 326 scans from dataset A, and  
14 94.7%, 100%, 90.0%, and 84.7% of 131 scans from dataset B, for NM, MH, ME, and AMD,  
15 respectively. In each cell, the number of positive cases versus that of the total cases was listed.  
16 (\*: for MH category, unfortunately, dataset B didn't contain macular hole cases, which coincides with

1 real clinical situations, since macular hole has very low prevalence rate. Therefore, for MH  
 2 diagnosing performance testing only, the training and testing dataset was organized in a way that 80%  
 3 of MH cases originated from dataset A are randomly sampled and included in the training set and the  
 4 rest (6 subjects) are included in the testing set.)

Training Statistics	NM	MH*	ME	AMD
Scan	80 / 316	49 / 287	190 / 287	59 / 295
Eye	66 / 187	27 / 176	109 / 180	27 / 178
Subject	65 / 133	26 / 128	84 / 130	21 / 133

5

Testing Statistics	NM	MH*	ME	AMD
Scan	22 / 124	21 / 153	59 / 118	81 / 111
Eye	13 / 54	8 / 66	29 / 54	31 / 50
Subject	10 / 36	6 / 43	23 / 34	20 / 33

6

7 Table 7.

8 Testing performance (AUC and best balanced accuracy) on dataset B, based on the pathology  
 9 classifiers trained using dataset A. The ground truth for this experiment was defined by the consensus  
 10 from the two experts (expert 1 and 2) for both datasets. The scans with different labels between the  
 11 two experts were excluded for each pathology separately.

Performance on Dataset B	NM	MH	ME	AMD
AUC	0.978	0.969	0.941	0.975
Best B. Accuracy	95.5%	97.3%	90.5%	95.2%

12