An Ensemble-Based System for Microaneurysm Detection and Diabetic Retinopathy Grading

Bálint Antal*, *Student Member, IEEE*, and András Hajdu, *Member, IEEE*

***Abstract*—Reliable microaneurysm detection in digital fundus images is still an open issue in medical image processing. We propose an ensemble-based framework to improve microaneurysm detection. Unlike the well-known approach of considering the output of multiple classifiers, we propose a combination of internal components of microaneurysm detectors, namely preprocessing methods and candidate extractors. We have evaluated our approach for microaneurysm detection in an online competition, where this algorithm is currently ranked as first, and also on two other databases. Since microaneurysm detection is decisive in diabetic retinopathy (DR) grading, we also tested the proposed method for this task on the publicly available Messidor database, where a promising AUC 0.90 ± 0.01 is achieved in a "DR/non-DR"-type classification based on the presence or absence of the microaneurysms.**

***Index Terms*—Diabetic retinopathy (DR) grading, ensemble-based systems, fundus image processing, microaneurysm (MA) detection.**

I. Introduction

DIABETIC retinopathy (DR) is a serious eye disease that originates from diabetes mellitus and is the most common cause of blindness in the developed countries. Early treatment can prevent patients to become affected from this condition or at least the progression of DR can be slowed down. Thus, mass screening of patients suffering from diabetes is highly desired, but manual grading is slow and resource demanding. Therefore, much effort has been made to establish reliable computer-aided screening systems based on color fundus images [1]. The promising results reported by Fleming *et al.* [2] and Jelinek *et al.* [3] indicate that automatic DR screening systems are getting closer to be used in clinical settings.

A key feature to recognize DR is to detect microaneurysms (MAs) in the fundus of the eye. The importance of handling MAs are twofold. First, they are normally the earliest signs of DR, hence their timely and precise detection is essential. On the other hand, the grading performance of computer-aided DR screening systems highly depends on MA detection [3], [4]. In this paper, we propose a MA detector that provides remarkable results from both aspects.

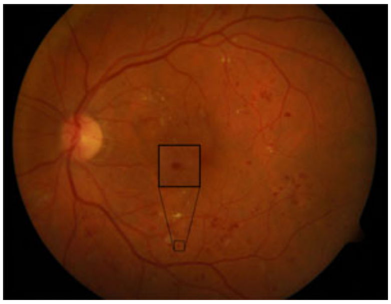

Fig. 1. Sample digital fundus image with a MA.

One way to ensure high reliability and raise accuracy in a detector is to consider ensemble-based systems, which have been proven to be efficient in several fields. However, the usual ensemble techniques aim to combine class labels or real values that cannot be adopted in our case. In MA detection, detectors provide spatial coordinates as centers of potential MA candidates. The use of well-known ensemble techniques would require a classification of each pixel, which can be misleading in our context, since different algorithms extract MAs with different approaches and the MA centers may not coincide exactly. To overcome this difficulty, we gather close MA candidates of the individual detectors and apply a voting scheme on them.

In [5], Niemeijer *et al.* showed that the fusion of the results of the several MA detectors leads to an increased average sensitivity measured at seven predefined false positive rates. In this paper, we propose a framework to build MA detector ensembles based on the combination of the internal components of the detectors not only on their output as in [5]. Some of our earlier research on combining MA detectors did not provide reassuring results [6]. To increase the accuracy of such ensembles, we must identify the weak points of MA detection. The first difficulty originates from the shape characteristics of MAs. They appear as small circular dark spots on the surface of the retina (see Fig. 1), which can be hard to distinguish from fragments of the vascular system or from certain eye features. Most MA detectors tackle this problem in the following way: first, the green channel of the fundus image is extracted and preprocessed to enhance MA like characteristics. Then, in a coarse level step (which will be referred as candidate extraction in the rest of this paper), all MA-like objects are detected in the image. Finally, a fine level algorithm (usually a supervised classifier) removes the potentially false detections based on some assumptions about

Manuscript received August 1, 2011; revised November 17, 2011 and January 11, 2012; accepted February 27, 2012. Date of publication April 3, 2012; date of current version May 18, 2012. This work was supported in part by the János Bolyai under Grant of the Hungarian Academy of Sciences, and by the TECH08-2 project DRSCREEN—Developing a computer-based image processing system for diabetic retinopathy screening of the National Office for Research and Technology of Hungary under Contract OM-00194/2008, Contract OM-00195/2008, Contract OM-00196/2008. *Asterisk indicates corresponding author.*

*B. Antal is with the Faculty of Informatics, University of Debrecen, Debrecen H-4032, Hungary (e-mail: antal.balint@inf.unideb.hu).

A. Hajdu is with the Faculty of Informatics University of Debrecen, Debrecen H-4032, Hungary (e-mail: hajdu.andras@inf.unideb.hu).

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

Digital Object Identifier 10.1109/TBME.2012.2193126

TABLE I
SUMMARY OF THE KEY DIFFERENCES OF THE PREPROCESSING METHODS

Algorithm	Aim	Method
Walter-Klein	contrast enhancement	gray level transformation
CLAHE	salient object enhancement	local histogram equalization
Vessel Removal	MA enhancement near vessels	vessel removal and inpainting
Illumination eq.	MA enhancement at the border of the ROI	vignette correction

MAs. Our former investigations showed that the low sensitivity of MA detectors originates from the candidate extractor part [7]. However, we could increase the sensitivity by applying proper preprocessing methods before candidate extraction. This technique causes a slight increment in the number of false positives, but it can be decreased by classification or voting.

In this paper, we propose an effective MA detector based on the combination of preprocessing methods and candidate extractors. We provide an ensemble creation framework to select the best combination. An exhaustive quantitative analysis is also given to prove the superiority of our approach over individual algorithms. We also investigate the grading performance of our method, which is proven to be competitive with other screening systems.

The rest of the paper is organized as follows: the selected preprocessing methods and candidate extractors are presented in Sections II and III, respectively. The details of the proposed ensemble creation framework is discussed in Section IV. We present our evaluation methodology in Section V. In Section VI, we summarize our experimental results. A detailed discussion is given in Section VII to address several issues. Finally, we draw conclusions in Section VIII.

II. PREPROCESSING METHODS

In this section, we present the selected preprocessing methods, which we consider to be applied before executing MA candidate extraction. The selection of the preprocessing method and candidate extractor components for this framework is a challenging task. Comparison of preprocessing methods dedicated to MA detection has not been published yet. Since preprocessing methods need to be highly interchangeable, we must select algorithms that can be used before any candidate extractor and do not change the characteristics of the original images (unlike, e.g., shade correction [8]). We also found some techniques to generate too noisy images for MA detection (histogram equalization [8], adaptive histogram equalization [8] or color normalization [8]). Thus, we have selected methods which are well-known in medical image processing and preserve image characteristics. Naturally, the proposed system can be improved in the future with adding new methods. A summary on the key differences of the algorithms is given in Table I.

A. *Walter–Klein Contrast Enhancement [9]*

This preprocessing method aims to enhance the contrast of fundus images by applying a gray level transformation using the following operator:

$$f' = \begin{cases} \dfrac{\frac{1}{2}\left(f'_{\max} - f'_{\min}\right)}{\left(\mu - f_{\min}\right)^r} \cdot \left(f - f_{\min}\right)^r + f'_{\min}, & f \leq \mu \\ \dfrac{-\frac{1}{2}\left(f'_{\max} - f'_{\min}\right)}{\left(\mu - f_{\max}\right)^r} \cdot \left(f - f_{\max}\right)^r + f'_{\max}, & f \geq \mu \end{cases}$$

where $\{f_{\min}, \ldots, f_{\max}\}$, $\{f'_{\min}, \ldots, f'_{\max}\}$ are the intensity levels of the original and the enhanced image, respectively, μ is the mean value of the original grayscale image and $r \in \mathbb{R}$ is a transition parameter.

B. *Contrast Limited Adaptive Histogram Equalization [10]*

Contrast limited adaptive histogram equalization (CLAHE) is a popular technique in biomedical image processing, since it is very effective in making the usually interesting salient parts more visible. The image is split into disjoint regions, and in each region a local histogram equalization is applied. Then, the boundaries between the regions are eliminated with a bilinear interpolation.

C. *Vessel Removal and Extrapolation [11]*

We investigate the effect of processing images with the complete vessel system being removed based on the idea proposed in [11]. We extrapolate the missing parts to fill in the holes caused by the removal using the inpainting algorithm presented in [12]. MAs appearing near vessels become more easily detectable in this way.

D. *Illumination Equalization [8]*

This preprocessing method aims to reduce the vignetting effect caused by uneven illumination of retinal images. Each pixel intensity is set according to the following formula:

$$f' = f + \mu_d - \mu_l$$

where f, f' are the original and the new pixel intensity values, respectively, μ_d is the desired average intensity, and μ_l is the local average intensity. MAs appearing on the border of the retina are enhanced by this step.

E. *No Preprocessing*

We also consider the results of the candidate extractors obtained for the original images without any preprocessing. That is, we formally consider a "no preprocessing" operation, as well.

III. MA CANDIDATE EXTRACTORS

Candidate extraction is a process that aims to spot any objects in the image showing MA-like characteristics. Individual MA detectors consider different principles to extract MA candidates. In this section, we provide a brief overview of the candidate extractors involved in our analysis. Again, just as for preprocessing methods, adding new MA candidate extractors may lead to further improvement in the future. A summary on the key differences of the candidate extractor algorithms and their

TABLE II
SUMMARY OF THE KEY DIFFERENCES OF THE CANDIDATE EXTRACTORS. THE SENSITIVITY AND AVERAGE NUMBER OF FALSE POSITIVES PER IMAGE (FP/I) IS MEASURED ON THE ROC TRAINING DATABASE WITH DEFAULT PARAMETER SETTINGS

Algorithm	Method	Sensitivity	FP / I
Walter	diameter closing	36%	154.42
Spencer	top-hat transformation	12%	20.3
Hough	circular Hough-transformation	28%	505.85
Zhang	matching multiple Gaussian masks	33%	328.3
Lazar	cross-section profile analysis	48%	73.94

performance measured in the Retinopathy online challenge (ROC) training dataset [13] are shown in Table II.

A. Walter et al. [14]

Candidate extraction is accomplished by grayscale diameter closing. That is, this method aims to find all sufficiently small dark patterns on the green channel. Finally, a double threshold is applied.

B. Spencer et al. [15]

From the input fundus image, the vascular map is extracted by applying 12 morphological top-hat transformations with 12 rotated linear structuring elements (with a radial resolution 15 °). Then, the vascular map is subtracted from the input image, which is followed by the application of a Gaussian matched filter. The resulting image is then binarized with a fixed threshold. Since the extracted candidates are not precise representations of the actual lesions, a region growing step is also applied to them. While the original paper [15] is written to detect MAs on fluorescein angiographic images, our implementation is based on the modified version published by Fleming *et al.* [16].

C. Circular Hough-Transformation [17]

Following the idea presented in [17], we established an approach based on the detection of small circular spots in the image. Candidates are obtained by detecting circles on the images using circular Hough transformation. With this technique, a set of circular objects can be extracted from the image.

D. Zhang et al. [18]

In order to extract candidates, this method constructs a maximal correlation response image for the input retinal image. This is accomplished by considering the maximal correlation coefficient with five Gaussian masks with different standard deviations for each pixel. The maximal correlation response image is thresholded with a fixed threshold value to obtain the candidates. Vessel detection and region growing is applied to reduce the number of candidates, and to determine their precise size, respectively.

E. Lazar et al. [19]

Pixel-wise cross-sectional profiles with multiple orientations are used to construct a multidirectional height map. This map assigns a set of height values that describe the distinction of the pixel from its surroundings in a particular direction. In a modified multilevel attribute opening step, a score map is constructed from which the MAs are extracted by thresholding.

IV. ENSEMBLE CREATION

In this section, we describe our ensemble creation approach. In our framework, an ensemble E is a set of $\langle$preprocessing method, candidate extractor$\rangle$ or shortly $\langle \text{PP}, \text{CE} \rangle$ pairs. The meaning of a $\langle$preprocessing method, candidate extractor$\rangle$ pair is that first we apply the preprocessing method to the input image and then we apply the candidate extractor to this result. That is, such a pair will extract a set of candidates H_E from the original image. If an ensemble E contains more $\langle$preprocessing method, candidate extractor$\rangle$ pairs, their outputs are fused in the following way: for each candidate c, all such candidates of the other participants are collected, whose euclidean distance d is smaller than a predefined constant $r \in \mathbb{R}$ from c. Let I_c denote that the set of these points collected for a candidate c. Then, the centroid calculated from I_c is put into H_E.

Ensemble creation is a process where all ensembles E from an ensemble pool $\mathcal{E}$ is evaluated and the best performing one $E_{\text{best}} \in \mathcal{E}$ regarding an evaluation function on a training set is selected. To evaluate an ensemble E, its output candidate set H_E must be compared to the ground truth in the following way: if for a $c \in H_E$ there exists a point in the ground truth, whose euclidean distance d from c is smaller than a predefined constant $r \in \mathbb{R}$, then c is considered as a true positive. Otherwise, c is false positive, while each ground truth point is a false negative that does not have a close candidate from H_E.

The selection of the optimal ensemble E_{best} would require each possible $\langle$preprocessing method, candidate extractor$\rangle$ ensembles to be evaluated. However, currently we consider $M = N = 5$ preprocessing methods and candidate extractors in our experiments. That is, we have 25 $\langle$preprocessing method, candidate extractor$\rangle$ pairs with 2^{25} number of possible combinations to form the ensemble. It would be very resource-demanding to evaluate such a large number of combinations, so we used simulated annealing [20] as a search algorithm to find the final ensemble, which is proven to be effective in such large search spaces. However, we describe the selection procedure as an exhaustive search in the latter parts, since it is better to evaluate all configurations if sufficient resource is available. Moreover, several other choices of search algorithms are possible.

As an energy function, we used the competition performance metric CPM [13], which is defined as the average sensitivity level at seven predefined false positive per image rate $(1/8, 1/4, 1/2, 1, 2, 4, 8)$ [13]. The process of ensemble creation is also shown in Fig. 2.

The ensemble creation part results in a set of $\langle$preprocessing method, candidate extractor$\rangle$ pairs. This ensemble E_{best} then can be used to detect MAs on unknown images. The final ensemble is applied in real detection in the same way as in the training phase. Namely, the final MAs are detected by the fusion of the MA candidates of the individual pairs building up the ensemble E_{best}. Similarly, for every detected MA, we will

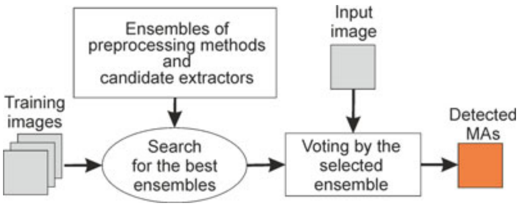
Ensembles of preprocessing methods and candidate extractors
Input image
Training images
Search for the best ensembles
Voting by the selected ensemble
Detected MAs

Fig. 2. Flow chart of the ensemble-based framework.

have a confidence value as described earlier. Thus, for the final decision on the presence of MAs, the output MA set needs to be thresholded according to the assigned confidence values. The choice of the threshold value is discussed in Section VII in detail.

The proposed ensemble creation method can be summarized through the following steps:

Algorithm 1: Selection of the optimal combination of preprocessing methods and candidate extractors.

```
$\mathcal{E} \leftarrow P\left(PP_i \times CE_j\right),\ i = 1, \ldots, M,\ j = 1, \ldots, N$
$CPM_{best} \leftarrow 0$
$E_{best} \leftarrow NULL$
for all $E \in \mathcal{E}$ do
    $H_E \leftarrow \emptyset$
    for all $p \in E$ do
        for all MA candidate $c$ detected by $p$ do
            $I_c \leftarrow \{c' | c'$ is a MA candidate found by a $p' \in E$, with $p \neq p'$ and $d\left(c, c'\right) < r\} \cup \{c\}$
            $confidence\left(c\right) = \frac{|I_c|}{|E|}$,
            $H_E \leftarrow H_E \cup centroid\left(I_c\right)$
        end for
    end for
    if $CPM\left(H_E\right) > CPM_{best}$ then
        $CPM_{best} \leftarrow CPM\left(H_E\right)$
        $E_{best} \leftarrow E$
    end if
end for
return $E_{best}$
```

V. Methodology

We have evaluated the proposed approach for both MA detection and DR grading. In this section, we present the evaluation methodology we used in each case.

A. MA Detection

We have evaluated the MA detection capabilities of the proposed method in the ROC competition for MA detectors [13], as well as on a publicly available [21] and a private database. In this section, we provide a brief overview on these databases and on the methodology we used for the evaluation of MA detection performance of the proposed approach.

1) Retinopathy Online Challenge [13]: ROC is a worldwide competition dedicated to measure the accuracy of MA detectors. The ROC database consists of 50 training and 50 test images with different resolutions (768×576, 1058×1061 and 1389×1383), 45° FOV and JPEG compression. The average number of MAs for the training and test sets are 6.72 and 6.86, respectively. There are 13 and 10 images of the training and test sets, where no MAs are marked by the experts.

2) DiaretDB1 2.1 Database [21]: The DiaretDB1 2.1 database contains 28 losslessly compressed training and 61 test images with a 1500×1152 resolution and 50° FOV. The average number of MAs for the training and test sets are 4.34 and 3.91, respectively. There are 15 and 39 images of the training and test sets, where no MAs are marked by the experts.

3) Private Database Provided by Moorfields Eye Hospital, U.K.: This database consists of 60 losslessly compressed images with a resolution 3072×2048 and 45° FOV. The average number of MAs for the training and test sets are 8.67 and 8.87, respectively. There are 10 and 8 images of the training and test sets, respectively, where no MAs are marked by the experts.

4) Testing: For each database, we provide the free-response receiver operating characteristic (FROC) curves [22], which plots the sensitivity against the average number of false positives per image. To measure the sensitivity at different average false positive per image levels, we thresholded the output set of the MA detector based on the confidence values assigned to each candidate. For the ROC dataset, we also provide the current ranking of the competition along with the CPM values (see Section VIII for details) that serves as the basis for the ranking. In addition, we also calculated a partial AUC of the algorithms in the same range (between $1/8$ and 8) by normalizing the average false positive per image figure by dividing with the maximum (8) and applying trapezoidal integration. The empirical AUC calculated this way is likely to underestimate the true AUC. However, the uncertainty for the partial AUCs may be quite high due to the low number of images.

B. DR Grading

We have also evaluated our ensemble-based approach to see its grading performance to recognize DR. For this aim, we determined the image-level classification rate of the ensemble on the Messidor[1] dataset containing 1200 images. That is, the presence of any MA means that the image contains signs of DR, while the absence of MAs indicates a healthy case. In other words, a pure yes/no decision of the system has been tested.

1) Ensemble Creation: As there is no training set provided for the Messidor database, we used an independent dataset (the ROC dataset) to train our algorithm. Note that, this is quite a strong handicap in comparison with the usual approach to train on a part of the same database. However, we feel that in this way we can get much closer to measure up the true performance of our system under real circumstances.

2) Testing: We used the publicly available Messidor database for testing. This database consists of 1200 losslessly

[1] Kindly provided by the Messidor program partners (see http://messidor.crihan.fr).

TABLE III
⟨PREPROCESSING METHOD, CANDIDATE EXTRACTOR⟩ PAIRS SELECTED AS MEMBERS OF THE ENSEMBLE FOR THE THREE DATASET. R, D, M DENOTE WHETHER THE PAIR IS SELECTED FOR THE ROC, DIARET2.1, OR THE MOORFIELDS DATASET, RESPECTIVELY

	Walter	Spencer	Hough	Lazar	Zhang
Walter-Klein		M			R
CLAHE	R, D	M		R	D
Vessel Removal	D			R, D, M	R, D
Illumination eq.				R, M	
No preprocessing	R		M	R, D	R

TABLE IV
QUANTITATIVE RESULTS OF THE ROC COMPETITION. FOR EACH PARTICIPATING TEAM, THE COMPETITION PERFORMANCE METRIC AND THE PARTIAL AUC ARE PRESENTED

Team	CPM	AUC
DRSCREEN	**0.434**	**0.551**
Niemeijer et al.	0.395	0.469
LaTIM	0.381	0.489
ISMV	0.375	0.435
OKmedical II	0.369	0.465
OKmedical	0.357	0.430
Lazar et al.	0.355	0.449
GIB	0.322	0.399
Fujita	0.310	0.378
IRIA	0.264	0.368
Waikato	0.206	0.273

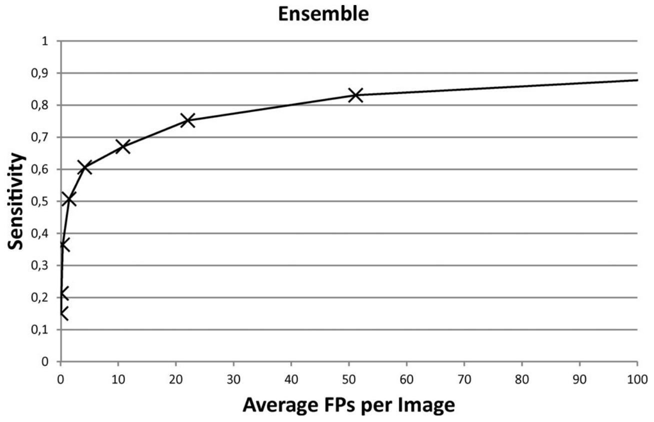

Fig. 3. FROC curve of the ensemble on the ROC dataset.

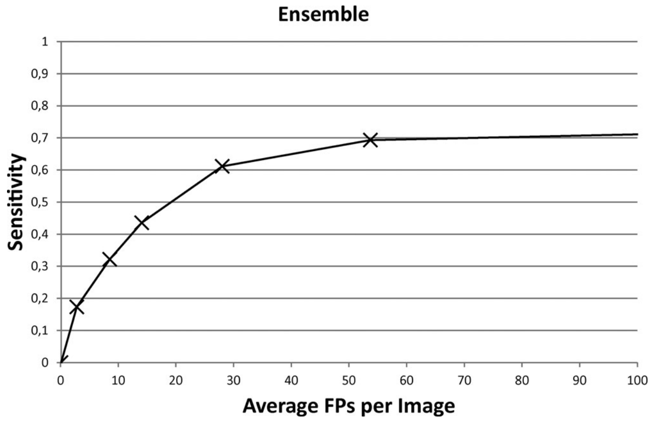

Fig. 4. FROC curve of the ensemble on the DiaretDB2.1 dataset.

compressed images with 45° FOV and different resolutions ($440 \times 960, 2240 \times 1488$, and 2304×1536). For each image, a grading score ranging from R0 to R3 is provided. These grades correspond to the following clinical conditions: a patient with an R0 grade has no DR. R1 and R2 are mild and severe cases of nonproliferative retinopathy, respectively. Finally, R3 is the most serious condition (proliferative retinopathy). The grading is based on the appearance of MAs, haemorrhages and neovascularization. The proportion of the images in the Messidor dataset: 540 R0 (46%), 153 R1 (12.75%), 247 R2 (20.58%), and 260 R3 (21.67%).

In our evaluation, we classified the retinal images whether they contain signs of DR (R1, R2, R3) or not (R0). The MA detector classifies an image as diseased if at least one MA was detected, and healthy otherwise. We measured the sensitivity, specificity, and accuracy of the detector at different levels by thresholding the confidence values assigned to the MA candidates as described in Section IV using the following formulas:

$$\text{sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

$$\text{specificity} = \frac{\text{tn}}{\text{tn} + \text{fp}}$$

and

$$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fn} + \text{tn} + \text{fp}}.$$

We also measured that the percentage of correctly recognized cases for each grade. We provided a fitted receiver operating characteristic (ROC) curve along with the empirical and fitted AUC for the proposed method on the Messidor database. For curve fitting, we used JROCFIT [23].

VI. RESULTS

In this section, we present our experimental results for both MA detection and DR grading.

A. MA Detection

In Table III, we exhibit the ⟨preprocessing method, candidate extractor⟩ pairs included in the selected ensembles for the three datasets, respectively. The rows of the table show the preprocessing methods from Section II, while the columns label the candidate extractor algorithms listed in Section III.

Table IV contains the ranked quantitative results of the participants at the ROC competition, with the proposed ensemble (DRSCREEN) highlighted as the current leader. The performance of the ensemble is also shown in Fig. 3 in terms of a FROC curve. As we can see from Table IV, the proposed ensemble earned both a higher CPM score and a higher partial AUC than the individual algorithms.

The FROC curves of the ensemble for the DiaretDB1 v2.1 and for the Moorfields database is shown in Figs. 4 and 5, respectively. To the best of our knowledge, no corresponding quantitative results have been published for these databases yet. Thus, we disclose the results of the ensemble-based method only.

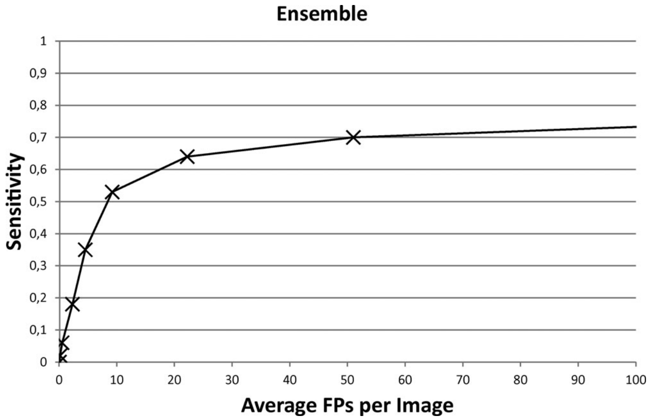

Fig. 5. FROC curve of the ensemble on the Moorfields dataset.

TABLE V
RESULTS ON THE MESSIDOR DATASET. FOR EACH THRESHOLD, SENSITIVITY, SPECIFICITY, ACCURACY AND THE PERCENTAGE OF CORRECTLY RECOGNIZED CASES FOR EACH GRADE ARE PRESENTED

Threshold	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Sensitivity	1	1	1	0.99	0.96	0.76	0.31
Specificity	0	0.01	0.03	0.14	0.51	0.88	0.98
Accuracy	0.53	0.54	0.55	0.59	0.75	0.82	0.62
R0	0.00	0.01	0.03	0.14	0.51	0.88	0.98
R1	1.00	1.00	1.00	0.97	0.92	0.60	0.18
R2	1.00	1.00	1.00	1.00	0.96	0.72	0.29
R3	1.00	1.00	1.00	1.00	0.98	0.92	0.42

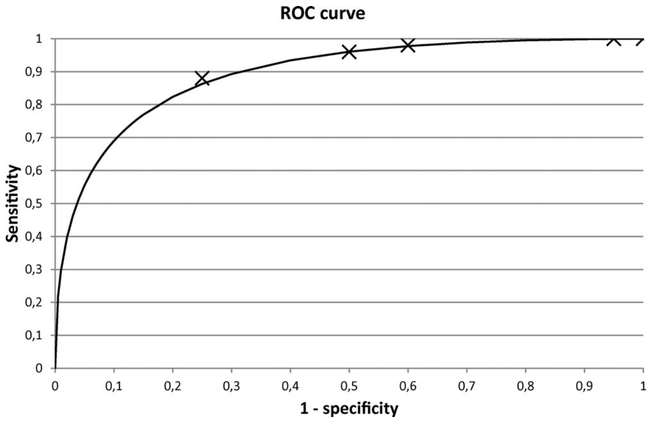

Fig. 6. ROC curve of the ensemble on the Messidor dataset.

B. DR Grading

In Table V, we provide the sensitivity, specificity and accuracy measures of our detector corresponding to different threshold values, respectively. The fitted ROC curve of the detector can be seen in Fig. 6. The empirical area under curve (AUC) is 0.875, while the AUC for the fitted curve is 0.90 ± 0.01. Table V also contains the percentage of the correctly recognized cases for each class.

VII. DISCUSSION

A strong point of the proposed method is that it performs well under difficult circumstances. Fig. 7 shows an example image where the application of CLAHE made it easier to distinguish the MAs from their background. However, the use of the vessel removal and inpainting preprocessing method caused the missing of a true MA, while the detection of the remaining MA is easier in the absence of thin retinal vessels. Thus, using different preprocessing methods with candidate extractors creates diversity among the members of the ensemble, which is desired for systems using multiple estimators [24]. This diversity ensures the suppression of false detections, since diverse detectors tend to make different mistakes. Thus, the false detections are likely to receive lower confidence values in the voting procedure.

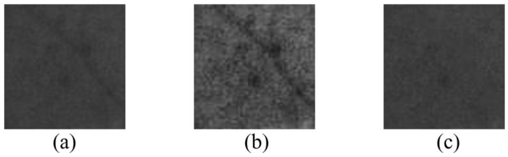

Fig. 7. Effect of different preprocessing methods where MAs are hard to detect.

Our experimental results show that the proposed ensemble-based MA detector outperforms the current individual approaches in MA detection. It has been also proven that the framework has high flexibility for different datasets. As can be seen in Table III, the ensemble members may vary, which suggests relatively high variance among databases in this field. Despite this variability, the performance of the ensemble still remained stable. In [13], the authors measured a human expert average false positive rate at the ROC dataset against the consensus of three human experts. This level is approximately 1 FP per image [13] for the ROC database, on which level our ensemble achieved the best score in the competition. Thus, we can recommend to use this level for thresholding at the ensemble creation phase and use it for detecting MAs on unknown images.

As for DR grading, our ensemble also performed well. It is also important to see how the different classes (R0, R1, R2, R3) are recognized at different levels. As can be desired, the severity of DR affects the performance of our detector. At each threshold level, where the sensitivity is less than 1.0, the more severe case recognized with higher probability.

The selection of the appropriate threshold is also an important issue for our detector to provide sufficient sensitivity and specificity rate. In [4], the authors suggest that sensitivity is more important for a screening system than specificity. In opposition, the British Diabetic Association recommends 80% sensitivity and 95% specificity for DR screening [25]. In Table V, we can see that the most accurate result is achieved with the threshold value 0.9. By applying the first idea, we might consider the results corresponding to the threshold value 0.8 as the best in our experiment, where 96% sensitivity and 51% specificity are achieved. That is, we recognized almost all of the cases where DR is present, and half of the healthy ones. The closest to the second recommendation is the performance achieved at the 0.9 level: 76% sensitivity and 88% specificity.

It is difficult to compare our method to other screening systems. First of all, to the best of our knowledge, no other results reported for the complete Messidor database. Other screening systems are tested on private images. Unfortunately, the proportion of non-DR/DR cases are varying in these experiments. Abramoff *et al*. [4] reported 0.86 AUC on a population where 4.96% of the cases had at least minimum signs of DR. The databases on which Agurto *et al*. [26] tested, 74.43% and 76.26% cases contained signs of DR and they achieved 0.81 and 0.89 AUCs, respectively. The closest to match the requirements of BDA is the system of Jelinek *et al*. [3] with a 85% sensitivity and 90% specificity, where approximately 30% of patients had DR. Similar proportion (35.88%) of patients having DR are reported by Fleming *et al*. [2] in their automatic screening system.

Despite the promising results, our system still misclassifies some stage, where serious case of DR is present. To improve grading performance, we must take into account the presence or absence of more DR-specific lesions (e.g., exudates), image quality, the recognition of anatomical parts which are essential in a clinical setting. However, our MA detector can serve as a main component of such a system.

VIII. Conclusion

In this paper, we have proposed an ensemble-based MA detector that has proved its high efficiency in an open online challenge with its first position. Our novel framework relies on a set of $\langle$preprocessing method, candidate extractor$\rangle$ pairs, from which a search algorithm selects an optimal combination. Since our approach is modular, we can expect further improvements by adding more preprocessing methods and candidate extractors. We have also evaluated the grading performance of this detector in the 1200 images of the Messidor database. We have achieved a 0.90 ± 0.01 AUC value, which is competitive with the previously reported results on other databases. The grading results presented in this paper are already promising. However, a proper screening system should contain other components, which is expected to increase the performance of this approach, as well.

References

[1] M. Abramoff, M. Niemeijer, M. Suttorp-Schulten, M. A. Viergever, S. R. Russel, and B. van Ginneken, “Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes,” *Diabetes Care*, vol. 31, pp. 193–198, 2008.

[2] A. D. Fleming, K. A. Goatman, S. Philip, G. J. Prescott, P. F. Sharp, and J. A. Olson, “Automated grading for diabetic retinopathy: A large-scale audit using arbitration by clinical experts,” *Br. J. Ophthalmol.*, vol. 94, no. 12, pp. 1606–1610, 2010.

[3] H. J. Jelinek, M. J. Cree, D. Worsley, A. Luckie, and P. Nixon, “An automated microaneurysm detector as a tool for identification of diabetic retinopathy in rural optometric practice,” *Clin. Exp. Optom.*, vol. 89, no. 5, pp. 299–305, 2006.

[4] M. Abramoff, J. Reinhardt, S. Russell, J. Folk, V. Mahajan, M. Niemeijer, and G. Quellec, “Automated early detection of diabetic retinopathy,” *Ophthalmology*, vol. 117, no. 6, pp. 1147–1154, 2010.

[5] M. Niemeijer, M. Loog, M. D. Abramoff, M. A. Viergever, M. Prokop, and B. van Ginneken, “On combining computer-aided detection systems,” *IEEE Trans. Med. Imag.*, vol. 30, no. 2, pp. 215–223, Feb. 2011.

[6] B. Antal, I. Lazar, A. Hajdu, Z. Torok, A. Csutak, and T. Peto, “A multilevel ensemble-based system for detecting microaneurysms in fundus images,” in *Proc. 4th IEEE Int. Workshop Soft Comput. Appl.*, 2010, pp. 137–142.

[7] B. Antal and A. Hajdu, “Improving microaneurysm detection using an optimally selected subset of candidate extractors and preprocessing methods,” *Pattern Recog.*, vol. 45, no. 1, pp. 264–270, 2012.

[8] A. A. A. Youssif, A. Z. Ghalwash, and A. S. Ghoneim, “Comparative study of contrast enhancement and illumination equalization methods for retinal vasculature segmentation,” in *Proc. Cairo Int. Biomed. Eng. Conf.*, 2006, pp. 21–24.

[9] T. Walter and J. Klein, “Automatic detection of microaneurysm in color fundus images of the human retina by means of the bounding box closing,” *Lecture Notes in Computer Science*, vol. 2526. Berlin, Germany: Springer-Verlag, 2002, pp. 210–220.

[10] K. Zuiderveld, “Contrast limited adaptive histogram equalization,” *Graphics Gems*, vol. 4, pp. 474–485, 1994.

[11] S. Ravishankar, A. Jain, and A. Mittal, “Automated feature extraction for early detection of diabetic retinopathy in fundus images,” in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2009, pp. 210–217.

[12] A. Criminisi, P. Perez, and K. Toyama, “Object removal by exemplar-based inpainting,” in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, vol. 2, 2003, pp. II-721–II-728.

[13] M. Niemeijer, B. van Ginneken, M. Cree, A. Mizutani, G. Quellec, C. Sanchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu, X. Wu, G. Cazuguel, J. You, A. Mayo, Q. Li, Y. Hatanaka, B. Cochener, C. Roux, F. Karray, M. Garcia, H. Fujita, and M. Abramoff, “Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs,” *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 185–195, Jan. 2010.

[14] T. Walter, P. Massin, A. Arginay, R. Ordonez, C. Jeulin, and J. C. Klein, “Automatic detection of microaneurysms in color fundus images,” *Med. Image Anal.*, vol. 11, pp. 555–566, 2007.

[15] T. Spencer, J. A. Olson, K. C. McHardy, P. F. Sharp, and J. V. Forrester, “An image-processing strategy for the segmentation and quantification of microaneurysms in fluorescein angiograms of the ocular fundus,” *Comput. Biomed. Res.*, vol. 29, pp. 284–302, 1996.

[16] A. D. Fleming, S. Philip, and K. A. Goatman, “Automated microaneurysm detection using local contrast normalization and local vessel detection,” *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1223–1232, Sep. 2006.

[17] S. Abdelazeem, “Microaneurysm detection using vessels removal and circular hough transform,” in *Proc. 19th National Radio Sci. Conf.*, pp. 421–426, 2002.

[18] B. Zhang, X. Wu, J. You, Q. Li, and F. Karray, “Detection of microaneurysms using multi-scale correlation coefficients,” *Pattern Recogn.*, vol. 43, no. 6, pp. 2237–2248, 2010.

[19] I. Lazar and A. Hajdu, “Microaneurysm detection in retinal images using a rotating cross-section based model,” in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2011, pp. 1405–1409.

[20] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *Science*, vol. 220, pp. 671–680, 1983.

[21] T. Kauppi, V. Kalesnykiene, J.-K. Kämäräinen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä, “Diaretdb1 diabetic retinopathy database and evaluation protocol,” in *Proc. 11th Conf. Med. Image Understanding Anal.*, 2007, pp. 61–65.

[22] D. Chakraborty, “Clinical relevance of the ROC and free-response paradigms for comparing imaging system efficacies,” *Radiation Protection Dosimetry*, vol. 139, no. 1–3, pp. 37–41, 2010.

[23] J. Eng. Roc analysis: Web-based calculator for roc curves. Johns Hopkins University, Baltimore. (2006). [Online]. Available: http://www.jrocfit.org

[24] L. I. Kuncheva, *Combining Pattern Classifiers. Methods and Algorithms*. Hoboken, NJ: Wiley, 2004.

[25] British Diabetic Association, “Retinal Photography Screening for Diabetic Eye Disease,” London, U.K.: British Diabetic Association, 1997, pp. 1–19.

[26] C. Agurto, E. S. Barriga, V. Murray, S. Nemeth, R. Crammer, W. Bauman, G. Zamora, M. S. Pattichis, and P. Soliz, “Automatic detection of diabetic retinopathy and age-related macular degeneration in digital fundus images,” *Invest. Ophthalmol. Vis. Sci.*, vol. 52, no. 8, pp. 5862–5871, 2011.

Authors’ photographs and biographies not available at the time of publication.