Ensemble Methods in Medical Decision Making

Doktori (PhD) értekezés

Antal Bálint

Témavezető: Dr. Hajdu András

Debreceni Egyetem
Természettudományi Doktori Tanács
Informatikai Tudományok Doktori Iskola
Debrecen, 2012.

Ensemble Methods in Medical Decision Making

Doktori (PhD) értekezés

Antal Bálint

Témavezető: Dr. Hajdu András

Debreceni Egyetem
Természettudományi Doktori Tanács
Informatikai Tudományok Doktori Iskola
Debrecen, 2012.

Ezen értekezést a Debreceni Egyetem Természettudományi Doktori Tanács Informatikai Tudományok Doktori Iskola Diszkrét matematika, képfeldolgozás és komputergeometria programja keretében készítettem a Debreceni Egyetem természettudományi doktori (PhD) fokozatának elnyerése céljából.
Debrecen, 2012.

.....
a jelölt aláírása

Tanúsítom, hogy Antal Bálint doktorjelölt 2009 - 2012 között a fent megnevezett Doktori Iskola Diszkrét matematika, képfeldolgozás és komputergeometria programjának keretében irányításommal végezte munkáját. Az értekezésben foglalt eredményekhez a jelölt önálló alkotó tevékenységével meghatározóan hozzájárult. Az értekezés elfogadását javasolom.
Debrecen, 2012.

.....
a témavezető aláírása

Ensemble Methods in Medical Decision Making

Értekezés a doktori (Ph.D.) fokozat megszerzése érdekében az informatika tudományágban

Írta: Antal Bálint okleveles programtervező matematikus

Készült a Debreceni Egyetem Informatikai Tudományok doktori iskolája Diszkrét matematika, képfeldolgozás és komputergeometria programja keretében

Témavezető: Dr. Hajdu András

A doktori szigorlati bizottság:

elnök:	Dr. Kruppa András	.....
tagok:	Dr. Fazekas Gábor	.....
	Dr. Palágyi Kálmán	.....

A doktori szigorlat időpontja: 2011. június 1.

Az értekezés bírálói:

.....	.....
.....	.....

A bírálóbizottság:

elnök:	.....	.....
tagok:	.....	.....
	.....	.....
	.....	.....
	.....	.....

Az értekezés védésének időpontja:

Contents

List of Figures

List of Tables

Abstract

Computer-aided decision support in medical problems is a prominent research area nowadays. In this PhD thesis, two approaches are shown to support the medical decisions for diabetic retinopathy (DR). This disease is one of the most common causes of blindness in the developed countries. Thus, timely and precise detection is essential for a large population. Furthermore, high reliability of the diagnosis is also desired.

The first major contribution of this thesis is an approach to the early detection of DR on color fundus images. This approach aims to detect the earliest signs of DR, namely microaneurysms (MAs). Since MA detection in color fundus images is a very challenging task, we propose a novel ensemble-based framework to ensure reliable fusion of MA detection output, namely ⟨preprocessing method, candidate extractor⟩ ensembles. This approach proved its capabilities in exhaustive evaluation, including a competition dedicated to the comparison of MA detectors (Retinopathy Online Challenge), where it is currently ranked as first.

The second major contribution of this thesis is an approach to retinal image grading based on the detailed analysis of color fundus images. Both the detected anatomical parts and presence of lesions are considered as features. We use an ensemble of machine learning classifiers for grading the retinal images based on the extracted features. As the results on a publicly available database show, a highly accurate grading system is achieved in this way.

1 Introduction

Nowadays, computer-aided systems in medical decision making are of large interest in the research community [1]. Although there is no

intention to leave out human experts from the diagnostic process, the evolution of decision support systems reached the point where they actually influence physicians in their diagnoses, especially in difficult cases [2]. Thus, high reliability must be ensured for computer-aided medical decision making systems.

Regarding the nature of decision making, clinical decision support systems primarily can be categorized as knowledge-based and non-knowledge-based approaches [1]. Knowledge-based systems are direct implementations of clinical protocols, thus, their suggestions are based on a predefined set of rules. Opposed to that, non-knowledge-based approaches extract decision rules using sophisticated artificial intelligence tools like machine learning. Thus, ensuring high reliability for non-knowledge-based systems is more challenging for computer science research than to medicine.

In this thesis, two non-knowledge-based automatic decision-making approaches to diabetic retinopathy (DR) grading are presented. DR is a consequence of diabetes mellitus which manifests itself in the retina. This disease is one of the most frequent causes of visual impairment in developed countries and is the leading cause of new cases of blindness in the working age population [3]. In 2011, 366 million people are diagnosed with diabetes and a further 280 million people are having risk to develop it [4]. At any point in time, approximately 40% of diabetic persons have diabetic retinopathy, of which an estimated 5% have the sight-threatening form of this disease. Altogether, every day nearly 75 people go blind as a consequence of DR even though treatment is available [5].

The retina of the human eye is a layered tissue, where the perceived light is conversed into neural signals [6]. The neural signals are sent to the brain via the fibres of the optic nerve. Thus, retina plays an essential role in the human vision. As a natural consequence of its purpose, the retina must be optically transparent [6]. This property makes it possible to use non-invasive imaging techniques to observe the

condition of the retinal tissue [6]. See Figure 1 for a visual presentation of the path of the light through the eye.

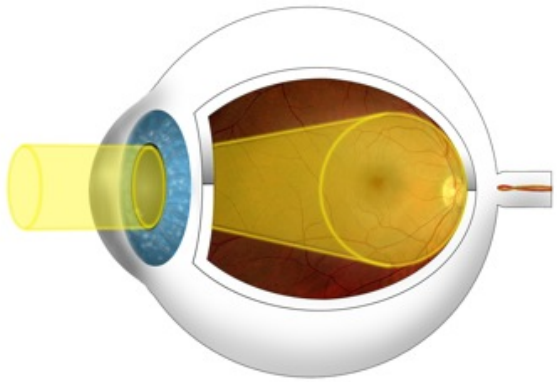


Figure 1: The path of the light through the eye. Courtesy of the UBC Department of Ophthalmology and Visual Sciences, Vancouver, Canada. Source: `http://www.ophthalmology.ubc.ca/facilities/images/section-f/img-03.jpg`.

Retinal imaging is widely used in the diagnostics and the regular controls of the consequent treatment of various eye diseases. The most widely used retinal imaging devices are fundus cameras, which are based on the invention of Gullstrand [6] [7]. Nowadays, there are techniques which possess better properties to support medical decision-making [6] (e.g. stereo fundus photography [8], optical coherence tomography [9], ultra-wide field retinal imaging [10]). However, due to its cost-effectiveness, color fundus images are widely used in medical practice. For a sample color fundus image, see Figure 2.

Timely detection, organized and practised screening programs are the mainstay of identifying patients at risk to be effected by any symptoms of DR. Several countries elaborated nation-wide or region-

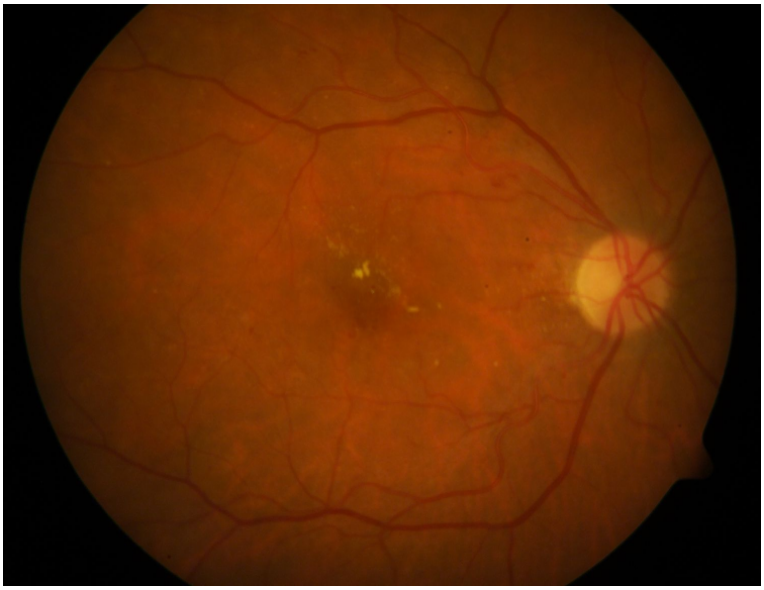


Figure 2: A sample fundus image from the DiaretDB1 database.

wide programs to fulfil this goal. In the United States [11] [12], in the United Kingdom [13] and in The Netherlands [14] there are digital photography acquisition and reader centre sites already available in daily routine. Color digital retinal images are captured at service sites even outside of health care settings and data will then be transferred to central locations, where they are double read and evaluated by specially trained graders. Further health provision of the patient depends upon the outcome of the grading process.

Automated grading of DR based on the detection of the characteristic lesions would safely reduce the burden of manual grading in screening programs. Promising results on higher sensitivity (the ratio of the correctly identified patients having DR in the investigated population) compared with manual graders have already been reported in

[15] for patients having referable diabetic retinopathy. Although the overall specificity (the ratio of correctly identified patients not showing signs of DR in the investigated population) of automated grading was lower than the manual analysis, remarkable financial savings could be achieved by reducing the grading workload. Screening programs can be organized to reduce the risk of the disease within the population. Though with the automated screening systems we have to make a compromise between sensitivity and specificity, an alternate approach with high performance is currently not available to provide mass screening. To evaluate automatic screening systems, they should be compared to human experts routinely involved. Note that, manual grading is imperfect, since graders missed more than 5% of the cases of referable DR in a study [15].

The process of analysing fundus images can be performed through a series of steps. In [16], the following steps were identified: preprocessing, localization and segmentation of the optic disc, segmentation of the retinal vasculature, localization of the macula and fovea, detection of the signs of DR (e.g. exudates, microaneurysm, haemorrhages). The anatomical parts of the retina can be seen in Figure 3, while a sample for the aforementioned lesions is shown in Figure 4. While there is a broad literature of this topic, we only deal with those aspects which are in the scope of the thesis. For the interested readers, excellent reviews are available in [16] and [17].

With an automated decision support of the grading process, wider access could be provided to the service and improvements could be realized at personal and community DR care level. Graders in DR reading centres are taught to recognize patterns which represent lesions like microaneurysm, dot and blot haemorrhages, lipid exudates and cotton wool spots. With the implementation of computer aided decision support algorithms, the detection of the above mentioned lesions would be theoretically possible. In the past, much effort has been made by different research groups to develop advanced algorithms with

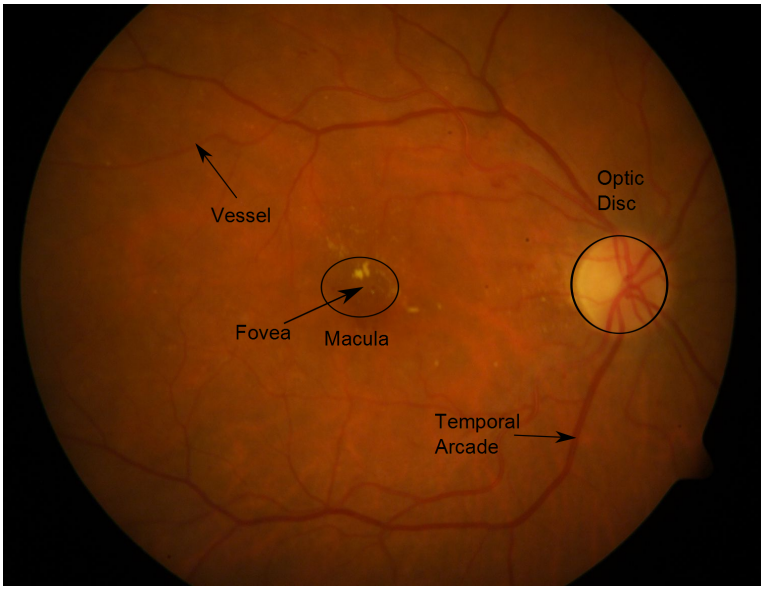


Figure 3: A sample fundus image with the main anatomical parts annotated.

sensitivity and specificity close to that of human graders [16].

A common way to improve reliability in machine learning based applications is to use ensemble-based approaches [18]. Ensemble systems combine the output of multiple learners with a specific strategy. In this thesis, two ensemble-based approaches are proposed to DR screening.

First, a method for the early detection of DR is presented. The key to early detection is the timely recognition of a lesion called microaneurysm (MA). Since MA detectors provide the spatial locations of MA candidates as output, the application of standard ensemble-based strategies do not provide sufficient solution, as we can see it later on. Thus, a spatial combination method is presented in this thesis. This approach is based on the novel concept of ⟨preprocessing method,

(a) Exudate (b) Microaneurysm (c) Haemorrhage

Figure 4: Examples of retinal lesions caused by diabetic retinopathy.

candidate extractor⟩ (⟨PP, CE⟩ for short) ensembles, which is shown to be effective in improving the sensitivity of MA candidate extraction and with the use of spatial voting, the false detections are also suppressed. For the creation of ⟨PP, CE⟩ ensembles, two approaches are introduced. First, a search-based selection approach is presented based on the performance of the ensembles on a training set. On the other hand, we investigate the effect of using context-dependent information and use the information to assign weights to the participating ⟨PP, CE⟩ pairs in the ensembles. Experimental results show superiority over individual approaches for both cases.

As the second main focus of the thesis, an ensemble-based approach to the grading of DR is also proposed. This approach is based on the output of several retinal image processing algorithms, such as lesion detection (microaneurysms, exudates), anatomical parts (macula, optic disc, vascular system), image features (diameter of the region-of-interest) and global DR descriptors (AM/FM filtering, quality assessment). From the output of the aforementioned algorithms, certain features are extracted and an ensemble of classifiers is trained. As it can be seen later, reassuring results are obtained using this technique for DR grading.

The main contributions (also including the corresponding publica-

tions) of the thesis can be summarized as follows:

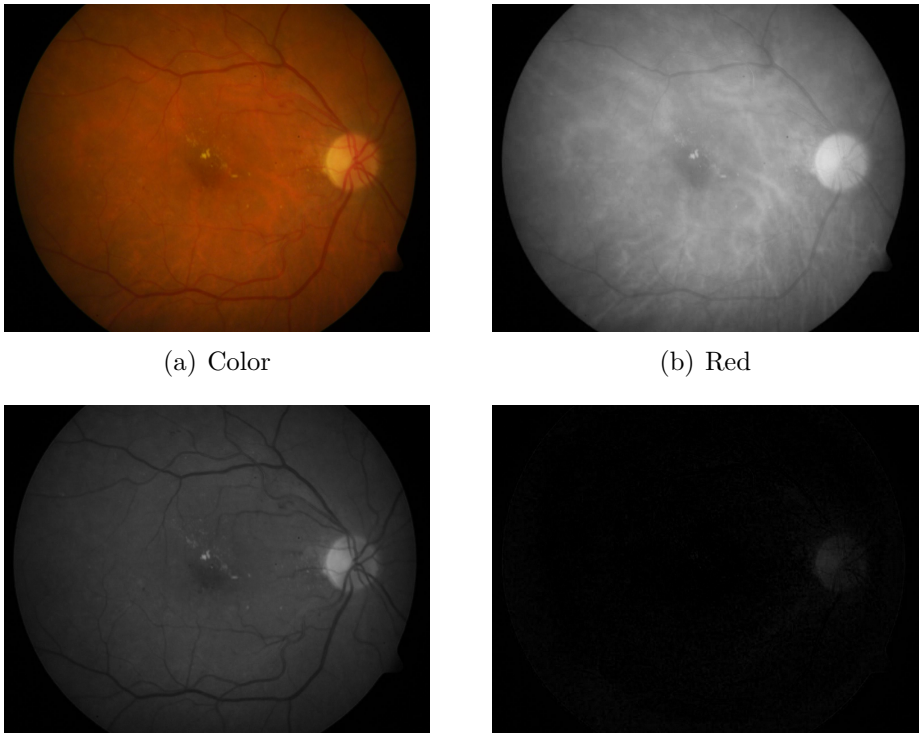
1. The introduction of ⟨PP, CE⟩ ensembles [19].
2. A ⟨PP, CE⟩ ensemble framework for microaneurysm detection [19], [20], [21], [22], [23].
3. A selection approach for ⟨PP, CE⟩ ensembles [19], [24], [25], [26], [27].
4. A context-adaptive weighting approach for ⟨PP, CE⟩ ensembles [28], [29].
5. A classifier ensemble-based approach to diabetic retinopathy grading [30].
6. A macula detection algorithm used in the diabetic retinopathy grading approach [31].
7. A method to recognize highly diseased cases used in the diabetic retinopathy grading approach [32], [33], [34].

While the thesis is written in plural form, the author has a principal contribution to the presented results.

The rest of the thesis is organized in the following way: section 2 provides a brief summary on the basic concepts and methodology used in this thesis. Sections 3 and 4 contain the main body of the thesis by describing the proposed early and final decision making approaches, respectively. Finally, conclusions are drawn in section 5.

2 Basic concepts and methodology

Most color retinal image processing algorithms rely on the processing of the green channel of the RGB image [17]. As it can be seen in Figure 5, the green channel has the highest contrast and visual information content among the three. In the rest of the thesis, only the green channel of the retinal image is used for processing.



(a) Color (b) Red

(c) Green (d) Blue

Figure 5: Channels of a sample image from the DiaretDB1 [35] database.

Due to the special properties of fundus photography, the acquired image contains several irrelevant background pixels. Thus, the region of interest (ROI) for each image is needed to be extracted, which excludes these irrelevant outer areas. The ROI of the sample image shown in Figure 2 extracted by the method presented in [36] can be seen in Figure 6.

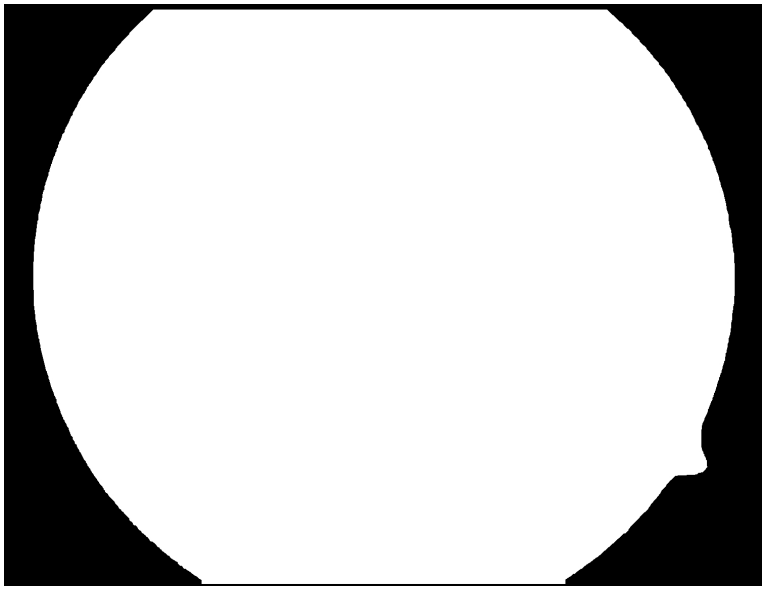


Figure 6: Region of interest (ROI, shown in white) for the sample image shown in Figure 2.

In the rest of the thesis, for all retinal image analysis methods, the pixels outside the ROI are excluded from further processing.

2.1 Color fundus image databases

In this section, we list the test databases and give the characteristics of the color fundus images used in this thesis.

2.1.1 Retinopathy Online Challenge (ROC) database

Retinopathy Online Challenge (ROC) [37] is a worldwide online competition dedicated to measure the accuracy of microaneurysm detectors. The ROC database consists of 50 training and 50 test images having different resolutions (768×576, 1058×1061 and 1389×1383 pixels), 45° field-of-view (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, respectively, where no MAs are marked by the clinical experts.

2.1.2 DiaretDB0 database

The DiaretDB0 [38] database contains 130 losslessly compressed color fundus images with a resolution of 1500×1152 pixels and 50° FOV.

2.1.3 DiaretDB1 2.1 database

The DiaretDB1 2.1 [35] database contains 28 losslessly compressed training and 61 test images, respectively with a resolution of 1500×1152 pixels 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, respectively, where no MAs are marked by the experts.

2.1.4 Database provided by Moorfields Eye Hospital, UK

This database consists of 60 losslessly compressed images with a resolution 3072×2048 pixels 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 in the training and test sets, respectively, where no MAs are marked by the experts.

2.1.5 Messidor database

This database consists of 1200 losslessly compressed images with 45° FOV and different resolutions (440×960, 2240×1488 and 2304×1536 pixels). For each image, a grading score ranging from R0 to R3 is also 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 non-proliferative retinopathy, respectively. Finally, R3 is the most serious condition (proliferative retinopathy). The grading is based on the appearance of MAs, haemorrhages and neovascularization. The corresponding proportion of the images in the Messidor dataset: 540 R0 (46%), 153 R1 (12.75%), 247 R2 (20.58%) and 260 R3 (21.67%). This database is kindly provided by the Messidor program partners (see `http://messidor.crihan.fr`).

2.1.6 DRIVE database

The DRIVE [39] database contains 40 JPEG-compressed color fundus images with 768×584 pixels resolution and 45° FOV.

3 Ensemble-based methods for the early detection of diabetic retinopathy

Microaneurysms (MAs) are normally the earliest signs of DR, so the detection of these lesions is essential in an efficient screening program [40]. MAs appear as small circular dark spots on the surface of the retina, as it can be observed in Figure 7. The most common place of occurrence of MAs is near thin vessels, though they cannot actually lie on the vessels.

MA detection is based on the detailed analysis of digital fundus images. State-of-the-art detection approaches usually start with the preprocessing of images, which is followed by candidate extraction. Finally, the extracted candidates are classified as MAs or non-MAs. The reason to separate the latter two steps is that the pixel-wise classification of the whole image would be very resource-demanding.

The vast majority of microaneurysm detectors can be organized into two categories: the ones based on mathematical morphology, and the others based on shape analysis with non-morphological tools. The largest family of morphology-based candidate extractors are originated from Lay [41] and Baudion [42]. These methods extract the vascular system by taking the maximum of multiple top-hat transformations with rotated linear structuring elements and subtract the resulting image from the original one. The candidates are then extracted by thresholding after applying a Gaussian filter. Oien et al. [43] was the first to apply similar techniques to color images. Spencer et al. [44] proposed a preceding shade correction step to this algorithm, while Frame et al. [45], Mendonca et al. [46], Hipwell et al. [47], Yang et al. [48], Cree et al. [49], Streeter et al. [50] and Fleming et al. [51] proposed modified variants. Mendonca et al. [46], Hipwell et al. [47], Yang et al. [48] and Fleming et al. [51] introduced an extension to this technique to decrease the number of false candidates. Niemeijer et al.

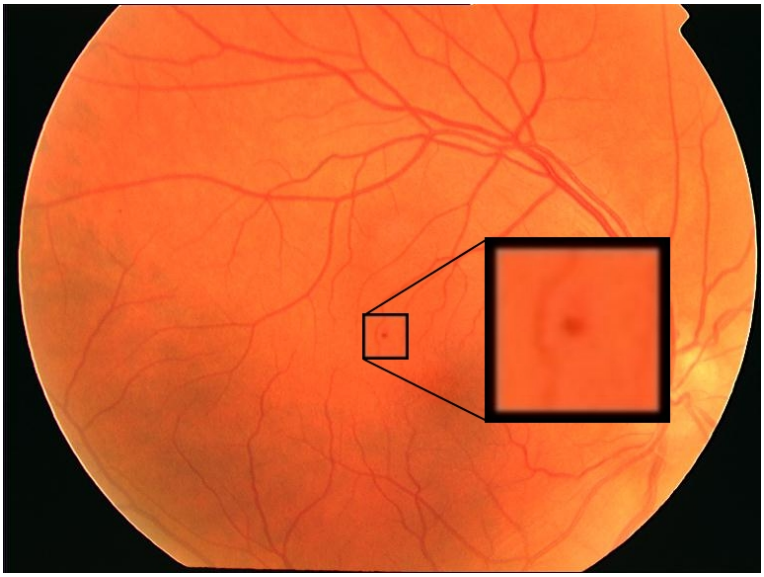


Figure 7: A sample color fundus image from the ROC database [37] with a microaneurysm highlighted.

[52] also considers a complementary machine learning-based candidate extractor and merge their output. Walter et al. [53] presented a different morphology-based approach, which relies on diameter closing, while in [54], Ravishankar applied morphological filling for establishing a candidate extraction algorithm. Other morphology-based approaches have been proposed in [55], [56] and [57].

The family of the non-morphological shape analysis-based approaches are more diverse. Marino et al. [58] and Bhalerao et al. [59] use a Gaussian mask to match the shape of MAs, while Zhang et al. [60], and Quellec et al. [61] apply multiple Gaussian masks for this purpose. Hahn et al. [62] consider the red/green ratio intensity values and select MA-candidates by applying a shape factor. Other circularity-based operators are used in [63] (double-ring filter), [64]

(circular Hough transformation) and [65] (Radon cliff operator).

There are approaches which cannot fit in any of the categories mentioned above, but also worth considering in an ensemble-based system. Grisan et al. [66] extracted candidates based on their densities after a local thresholding. Balasubramanian et al. [67] applied grey-level grouping and proposed an automatic seed generation technique to tackle this problem. Moat operator and region growing are applied in [68] and [69], respectively. Pallawala et al. introduced a normalized cut-based local segmentation technique. Gardner et al. [70] trained an artificial neural network to recognize different retinal features, including microaneurysms. Lázár et al. [71] presented an approach relying on the strength of multiple cross-sectional profiles across the image.

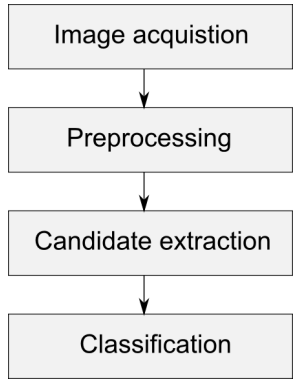


Figure 8: Common steps of microaneurysm detection.

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 [18]. However, the usual ensemble techniques aim to combine class labels or class membership likelihoods which 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 spatially close MA candidates of the individual detectors and apply a voting scheme to them.

In [72], 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 section, we propose three approaches to improve MA detection on the combination of the internal components of the detectors not only on their output as in [72]. The approaches are based on the novel concept of ⟨PP, CE⟩ ensembles.

To present the related results in details, first, a method using majority voting of candidate extraction algorithms is proposed in section 3.3. It will become clear that the results are improved using these techniques, but the sensitivity of the approach is too low and the number of false detections are still high. The corresponding results are also published in [24], [25] and [26].

The second approach (section 3.4) extends the aforementioned technique to tackle its shortcomings by combining preprocessing methods ans candidate extractors. To suppress the number of false detections, ensemble selection and voting are applied. Experimental evaluation of this approach is also performed, where it is shown to be superior over individual MA detectors. The corresponding results are also published in [19], [20], [21], [22], [23] and [27]. We investigate the performance of this approach in recognizing MAs for images where DR present, as well.

Moreover, an adaptive weighting technique to ⟨PP, CE⟩ ensembles is also presented in section 3.4. The weights assigned to ensemble members based on spatial, visual and performance information measured on a training set. More on these results are given in [28] and [29].

The rest of section 3 is organized as follows: in section 3.1 the novel concept of ⟨PP, CE⟩ ensembles are introduced. Section 3.2 contains a

brief description of the algorithms used in this section. The creation of ensembles based containing only candidate extractor algorithms is described in section 3.3. Section 3.4 is devoted to the search-based selection of ⟨PP, CE⟩ pairs to build up the ensemble, while section 3.6 introduces a context-aware weighting approach for such ensembles.

3.1 ⟨Preprocessing method, candidate extractor⟩ (⟨PP, CE⟩) ensembles

In this section, we provide some theoretical foundation for the creation of ensembles containing preprocessing methods and MA candidate extractors for the easier and precise reference in the latter parts of the thesis. We emphasize the similarities between the introduced terms and those used in the literature of classifier ensembles [18]. The concepts introduced here were used in the publications [19], [25], [26], [27] of the author.

Without the loss of generality we will work with 8-bit intensity images of size $k \times l$, that is, a digital image I is defined as:

$$I : \{1, \ldots, k\} \times \{1, \ldots, l\} \to [0, \ldots, 255]\,. \tag{1}$$

We will form ensembles by merging the output of different candidate extractors. While the number of such algorithms are limited for certain cases, we permit replacing the input images if those are modifications of the same image applying intensity-based transforms. That is, an image transform f has the general form $f : I \to I'$, where I and I' are digital images.

However, in object detection problems only such transforms are allowed that do not modify the spatial locations of the desired objects (e.g. lesions). Thus, we allow transforms that aim is to enhance the appearance of the desired objects with trying to avoid the enhancement of the others. These allowed transforms are referred as preprocessing

methods, and we will also follow this terminology. In general, it is meaningless to give a rigorous definition for preprocessing algorithms or to describe precisely whether we can find an allowed transform between two images.

Despite these difficulties, from a practical point of view we give the following intuitive definition for the preprocessed variants of an image I:

Definition 1 *Let I and I' be two digital images. I and I' are replacable, if there exists an allowed image transform between them. The set of all replaceable variants of I is denoted by $\mathcal{I}$. Note that, we also let $I \in \mathcal{I}$.*

Definition 2 *Let I be a digital image. A candidate extractor algorithm C is defined as:*

$$C : I \to P\left(\{1, \ldots, k\} \times \{1, \ldots, l\}\right), \tag{2}$$

where $P(X)$ denotes the power set of a set X. The set of all candidate extractors is denoted by $\mathcal{C}$.

In the next step, we let preprocessing methods and candidate extractor algorithms to put together to set up ensembles.

Definition 3 *Let I be a digital image. A finite set $E \subseteq \mathcal{I} \times \mathcal{C}$ is called a $\langle$preprocessing method, candidate extractor$\rangle$ ensemble ($\langle$PP, CE$\rangle$ for short) . The candidate set of the ensemble E is defined as $EC = \bigcup_{(I',C')\in E} C'(I')$.*

That is, we have an ensemble pool $\mathcal{E}$ of size $|\mathcal{I}||\mathcal{C}|$. Then, we can put together specific preprocessing methods and candidate extractors, where candidates of different candidate extractors being close to each

other will be collected and the will vote for a lesion in the corresponding position. In general, we could consider any distance measure between two candidates. However, for our current purposes we define this relation with the help of the 2D Euclidean metric.

Definition 4 *Let $c_1, c_2 \in \{1, \ldots, k\} \times \{1, \ldots, l\}$. We say that $c_1 \cong c_2$, if $||c_1 - c_2||_2 < d$ with some distance threshold $d \in \mathbb{R}_{\geq 0}$, where $||.||_2$ stands for the Euclidean norm.*

Now we can assign a confidence value to each candidate of an ensemble E with collecting the close candidates of other ensemble members.

Definition 5 *Let I be a digital image, E be a $\langle PP, CE \rangle$ ensemble, and $c \in EC$ is its candidate. Then, the confidence level of c is defined as:*

$$conf_E(c) = |\{(C', I') \in E \ : \ \exists c' \in C'(I') \text{ such that } c \cong c'\}| \ / \ |E|. \quad (3)$$

Moreover, we define the α-level candidates of E as:

$$EC_\alpha = \{c \in EC \ : \ conf_E(c) \geq \alpha\}, \quad (4)$$

where $0 \leq \alpha \leq 1$.

The level of confidence for a candidate is analogous to the number of votes in the majority voting terminology. The confidence level serves as a membership function for EC. Also note that the candidate level of each candidate of a $\langle$PP, CE$\rangle$ ensemble in the interval $[1/|E|, 1]$. Note that, majority voting [18] can be accomplished in this context by setting $\alpha = 0.5$.

Now we set up a framework to measure the accuracy of the possible ensembles to be able to select the appropriate ones for practical problems. We start with making the decision whether the candidates found by a specific extractor are true or false ones regarding some ground truth.

Definition 6 *Let $REF \subseteq \{1, \ldots, k\} \times \{1, \ldots, l\}$ be a reference set of candidates for an image I. For the classification of each candidate c of an ensemble $E \subseteq \mathcal{I} \times \mathcal{C}$ on the same image regarding REF we apply the following:*

- *c is a true positive (TP), if $\exists c_{ref} \in REF$ such that $c \cong c_{ref}$.*
- *c is a false positive (FP), if $\neg\exists c_{ref} \in REF$ such that $c \cong c_{ref}$.*

Regarding a candidate c_{ref} of REF, we apply:

- *c_{ref} is a false negative (FN), if $\neg\exists c \in EC$ such that $c \cong c_{ref}$.*

All pixels in $\{1, \ldots, k\} \times \{1, \ldots, l\} \backslash REF \cup EC$ are true negatives (TN).

Note that, REF is usually a manually annotated set created by experts of the application field. Now, we can introduce some classic measures to describe the performance of an ensemble.

Definition 7 *To measure the performance of an ensemble E regarding a reference set REF we use the the following descriptors [73]:*

- *sensitivity (SEN):*

$$SEN = \frac{TP}{TP + FP}, \tag{5}$$

- *specificity (SPE):*

$$SPE = \frac{TN}{FN + TN}, \tag{6}$$

- *false positive rate (FPR):*

$$FPR = 1 - SPE, \tag{7}$$

- *positive likelihood ratio (PLR):*

$$PLR = \frac{TP}{FP}, \tag{8}$$

- *positive predictive value (PPV):*

$$PPV = \frac{TP}{TP + FN}, \tag{9}$$

- *F-Score:*

$$F\text{-}Score = 2 \cdot \frac{SEN \cdot PPV}{SEN + PPV}, \tag{10}$$

- *average false positives per image (FPI):*

$$FPI = \frac{FP}{number\ of\ images\ in\ the\ database}, \tag{11}$$

- *accuracy (ACC):*

$$ACC = \frac{TP + TN}{FP + FN + TN + FP}. \tag{12}$$

Definition 8 *Let I be an image, E be an ensemble and REF a reference candidate set. For a confidence level $0 \leq \alpha \leq 1$ we apply the following (see definition 6):*

- *c is α-true positive (TP_α), if c is TP and $c \in EC_\alpha$,*
- *c is a α-false positive (FP_α), if c is FP and $c \in EC_\alpha$,*

Regarding a candidate c_{ref} of REF, we apply:

- *c_{ref} is α-false negative (FN_α), if $\neg\exists c \in EC_\alpha$ with $c \cong c_{ref}$.*

All pixels in $\{1, \ldots, k\} \times \{1, \ldots, l\} \backslash REF \cup EC_\alpha$ *are* α*-true negative* (TN_α).

Definition 9 *To measure the performance of an ensemble* E *regarding an image* I*, a reference set* REF *and a confidence level* α *we use the the following descriptors:*

- α*-sensitivity* (SEN_α)*:*

$$SEN_\alpha = \frac{TP_\alpha}{TP_\alpha + FP_\alpha}, \tag{13}$$

- *average false positives per image for the* α *confidence level* (FPI_α)*:*

$$FPI_\alpha = \frac{FP_\alpha}{number\ of\ images}, \tag{14}$$

- *competition performance metric (CPM) [37]:*

$$CPM = \frac{1}{|G|} \cdot \sum_{g \in G} \{SEN_\alpha | FPI_\alpha = g\}, \tag{15}$$

where $G = \left\{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\right\}$.

Definition 10 *To measure the diversity of two candidate sets* EC *and* EC' *regarding a reference set* REF*, we define the following diversity measures based on [74]:*

- *disagreement measure (DIS):*

$$DIS = n_{01} + n_{10}, \tag{16}$$

where n_{10} *and* n_{01} *are the number of TP candidates* c *where* $c \in EC \wedge \nexists c' \in EC'$ *with* $c \cong c'$ *and the number of TP candidates* c' *with* $c' \in EC' \wedge \nexists c \in EC$*, where* $c \cong c'$*, respectively.*

- *double fault measure (DF):*

$$DF = n_{00}, \tag{17}$$

where n_{00} is the number of TP candidates c, where $c \notin EC \wedge \nexists c' \in EC'$ with $c \cong c'$.

3.2 Algorithms used for microaneurysm detector ensembles

In this section, a brief description of the algorithm used for the creation of MA detector ensembles is described. Before presenting our main results, we review the preprocessing methods and candidate extractors that will give the basis for ensemble creation.

3.2.1 Preprocessing methods

This section describes 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 microaneurysm detection has not been published yet. Since preprocessing methods need to be highly interchangeable, we must select algorithms which can be used before any candidate extractor and do not change the characteristics of the original images (unlike e.g. shade correction [75]). Thus, all these preprocessing methods can be considered as allowed transformations in the sense of Definition 1. We also found some techniques to generate too much noise for MA detection (histogram equalization [75], adaptive histogram equalization [75] or color normalization [75]). Thus, we have selected methods which are well-known in medical image processing and preserve image characteristics. More specifically, we have chosen a

preprocessing method proposed for an MA detector by Walter et al. [76], a widely used medical image preprocessing method, Contrast Limited Adaptive Histogram Equalization (CLAHE) [77], a vessel removal and inpainting technique, which has proved to be successful in other retina-based detection algorithms [54], and illumination equalization to suppress the vignetting effect of fundus images [78]. Naturally, the proposed system can be improved in the future with adding new preprocessing methods.

3.2.1.1 No preprocessing (NO) We consider the results of the candidate extractors obtained for the original images without any preprocessing in the sense of Definition 1. That is, we formally consider a "No preprocessing" operation, as well. A sample image without preprocessing can be seen in Figure 9.

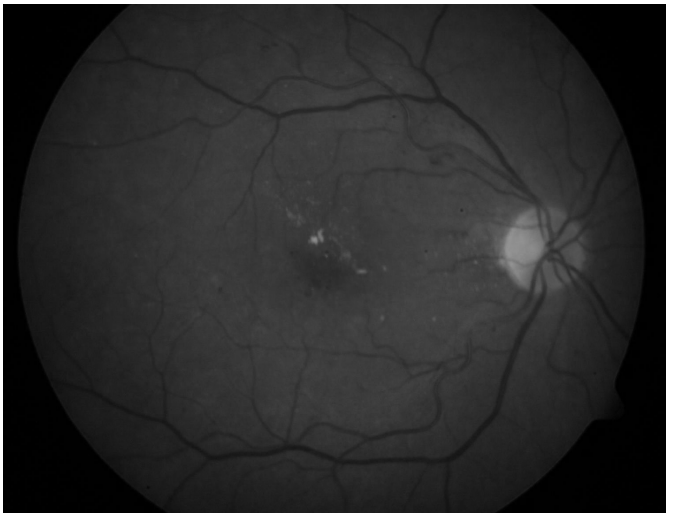


Figure 9: A sample fundus image from the DiaretDB1 database with no preprocessing applied.

3.2.1.2 Walter-Klein contrast enhancement (WK) [76] This preprocessing method aims to enhance the contrast of fundus images by applying a gray level transformation using the following operator:

$$f'(x,y) = \begin{cases} \dfrac{\frac{1}{2}\left(f'_{max} - f'_{min}\right)}{\left(\mu - f_{min}\right)^r} \cdot \left(f(x,y) - f_{min}\right)^r + f'_{min}, & f(x,y) \leq \mu, \\ \dfrac{\frac{1}{2}\left(f'_{min} - f'_{max}\right)}{\left(\mu - f_{max}\right)^r} \cdot \left(f(x,y) - f_{max}\right)^r + f'_{max}, & f(x,y) \geq \mu, \end{cases} \tag{18}$$

where $\{f_{min}, \ldots, f_{max}\}$, $\{f'_{min}, \ldots, f'_{max}\}$ are the minimal and maximal intensity levels of the original and the enhanced image, respectively, $f(x,y)$ is the intensity level for the pixels (x,y) with $0 \leq x \leq k$, $0 \leq y \leq l$, μ is the mean value of the original grayscale image and $r \in \mathbb{R}$ is a transition parameter. A sample image preprocessed with Walter-Klein contrast enhancement can be seen in Figure 10.

3.2.1.3 Contrast limited adaptive histogram equalization (CL) [77] Contrast limited adaptive histogram equalization (CLAHE) is a popular technique in biomedical image processing, since it is very effective in making the 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. A sample image preprocessed with CLAHE can be seen in Figure 11.

3.2.1.4 Gray-world normalization (GN) [75] Each pixel on the green channel of the image is transformed in the following way:

$$f'(x,y) = \frac{f(x,y)}{\mu}, \tag{19}$$

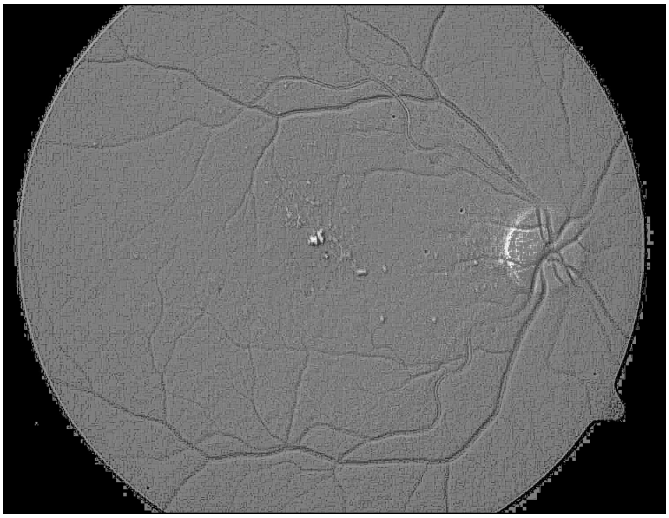


Figure 10: A sample fundus image from the DiaretDB1 database preprocessed with the Walter-Klein contrast enhancement.

where f' and f are the respective new and original pixel intensities for the pixels (x, y) with $0 \leq x \leq k$, $0 \leq y \leq l$, and μ is the average intensity of the green channel. A sample image preprocessed with gray-world normalization can be seen in Figure 12.

3.2.1.5 Intensity adjustment (IA) [79] This preprocessing method enhances the contrast of a grayscale image by saturating the lowest and highest 1% of the intensity values. A sample image preprocessed with intensity adjustment can be seen in Figure 13.

3.2.1.6 Illumination equalization (IE) [75] This preprocessing method aims to reduce the vignetting effect caused by uneven illumination of retinal images. Each pixel intensity is set according to the

Figure 11: A sample fundus image from the DiaretDB1 database preprocessed with CLAHE.

following formula:

$$f'(x,y) = f(x,y) + \mu_d - \mu_l, \tag{20}$$

where $f(x,y), f'(x,y)$ are the original and the new pixel intensity values for the pixels (x,y) with $0 \leq x \leq k$, $0 \leq y \leq l$, 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. A sample image preprocessed with illumination equalization can be seen in Figure 14.

3.2.1.7 Vessel removal and extrapolation (VR) [54] We investigate the effect of processing images with the complete vessel system being removed based on the idea proposed in [54]. We extrapolate the missing parts to fill in the holes caused by the removal using the

Figure 12: A sample fundus image from the DiaretDB1 database preprocessed with gray-world normalization.

inpainting algorithm presented in [80]. MAs appearing near vessels become more easily detectable in this way. A sample image preprocessed with vessel removal and extrapolation can be seen in Figure 15.

3.2.1.8 Background subtraction of retinal blood vessels (BS)

[81] This method is recommended for vascular system detection. Blood vessels on retinal images show similar local appearance to MAs. This approach considers the vessel system as the foreground of the image. The background is extracted by applying an averaging filter, which is followed by threshold averaging for smoothing. The background image is then subtracted from the original image. A sample image preprocessed with background subtraction of retinal blood vessels can be seen in Figure 16.

Figure 13: A sample fundus image from the DiaretDB1 database preprocessed with intensity adjustment.

3.2.1.9 Histogram equalization (HE) [75] This preprocessing method aims to enhance the global contrast of the image by redistributing the intensity values of the image. First, the accumulated normalized histogram of the image is created. Then, the histogram is transformed to reflect uniform distribution. A sample image preprocessed with histogram equalization can be seen in Figure 17.

3.2.2 Candidate extractors

Candidate extraction is a process which aims to detect 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.

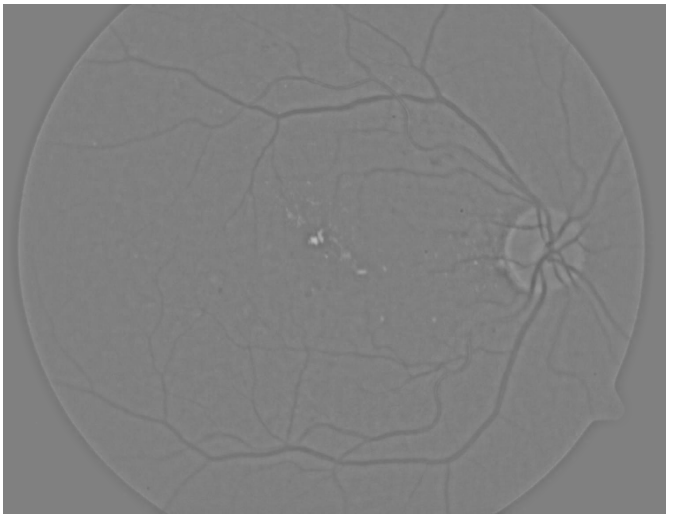


Figure 14: A sample fundus image from the DiaretDB1 database preprocessed with illumination equalization.

3.2.2.1 Walter et al. [53] 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.

3.2.2.2 Spencer et al. [44] From the input fundus image, the vascular map is extracted by applying twelve morphological top-hat transformations with twelve 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 [44] aims to detect MAs on fluorescein

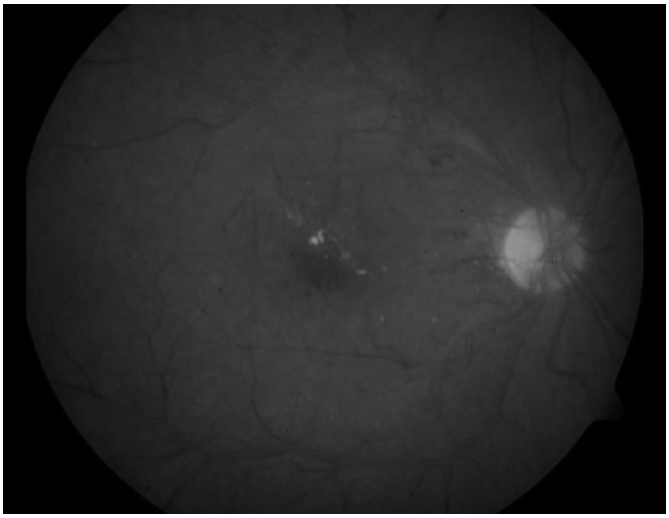


Figure 15: A sample fundus image from the DiaretDB1 database preprocessed with vessel removal and extrapolation.

angiographic images, our implementation was based on the modified version published by Fleming et al. [51].

3.2.2.3 Circular Hough-transformation [64] Following the idea presented in [64], 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.

3.2.2.4 Zhang et al. [60] 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

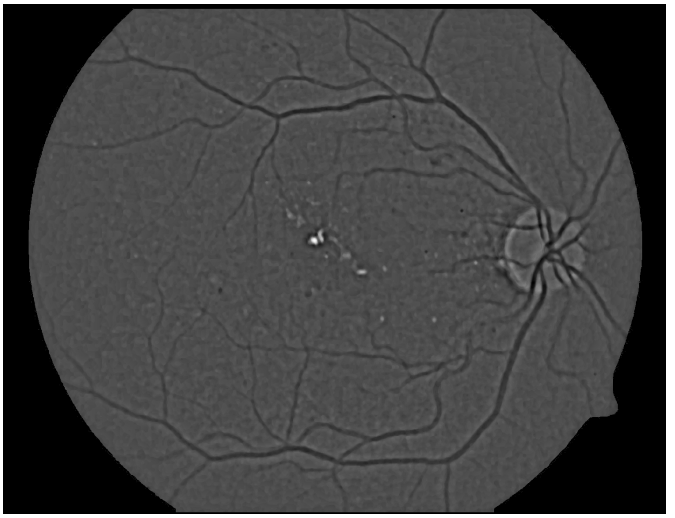


Figure 16: A sample fundus image from the DiaretDB1 database preprocessed with background subtraction of retinal blood vessels.

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.

3.2.2.5 Lázár et al. [71] Pixel-wise cross-section profiles with multiple orientations are used to construct a multi-directional 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.

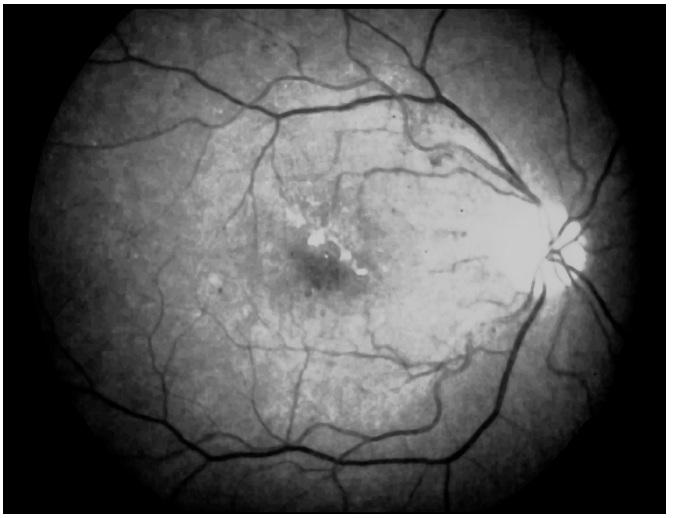


Figure 17: A sample fundus image from the DiaretDB1 database preprocessed with histogram equalization.

3.3 Majority voting of microaneurysm candidate extractors

In this section, we propose an ensemble-based approach to MA detection to suppress the errors of individual algorithms. The proposed process consists of three main stages: first, we extract MA candidates from fundus images. For this task, we select four state-of-the-art approaches, namely the candidate extractor of Walter et al. (section 3.2.2.1), Spencer et al. (section 3.2.2.2), the circular Hough transformation based approach (section 3.2.2.3) and the Lázár et al. (section 3.2.2.5) algorithm. In the second phase, we combine the results provided by the four candidate extractors. This combination framework can be regarded as a $\langle$PP, CE$\rangle$ ensemble E, where $\mathcal{C} = \{\text{Walter}, \text{Spencer}, \text{Hough}, \text{Lázár}\}$ (see (2) and $\mathcal{I} = \{I\}$ (see definition 1). That is, the set of preprocessing

methods include the identical transform ("No preprocessing") only. Finally, the number of candidates are reduced with a voting scheme. We also investigate the effect of using different number of votes for selection. Finally, we introduce a novel machine-learning based algorithm to classify the candidates. As experiments show, with this ensemble method state-of-the-art microaneurysm candidate extractors can be outperformed.

The approach proposed in this section is the work of the author and published in [24].

3.3.1 Combining the candidate extractors

As the most straightforward measure, we aim to raise the positive likelihood ratio (PLR, see (8)) using an ensemble. It is important to use diverse candidate extractors, that is, to reduce the number of false positives efficiently and keep only those candidates on which multiple methods agree. In Section 3.3.2 we show the diversity of the methods measured. For the proper comparison of the candidates extracted by the individual approaches, we will have to merge them if they are sufficiently close to each other, as defined in Definitions 4 and 5, respectively.

3.3.2 Diversity of the candidate extractors

The pairwise diversity of the classifiers can be measured by the disagreement (DIS, see (16)) and double fault (DF, see (17)) measures. DIS sums the cases where the extractors disagree, but one of them is correct, while the DF measure count the candidates, where both extractors agree and both are incorrect. For our aims, high DIS values and low DF values are ideal. As it can be seen in Table 1, the selected candidate extractors are quite diverse.

Walter	Spencer	Hough	Lázár	DIS	DF
x	x			0,73	0,09
x		x		0,77	0,04
x			x	0,49	0,10
	x	x		0,79	0,06
	x		x	0,69	0,14
		x	x	0,74	0,12

Table 1: Pairwise diversity of the microaneurysm candidate extractors.

3.3.3 Voting on the candidates

Each individual algorithm produces an initial set of microaneurysm candidates. Then, we establish a set of final candidates EC_α through majority voting. The voting procedure has the following steps formalized by Definition 4:

1. For each candidate c provided by one of the algorithms, check whether there is another candidate detected by another algorithm within a distance $r \in \mathbb{R}_{\geq 0}$ from c.

2. Let sum be the number of candidates satisfying the above proximity criterion and remove all these candidates from their respective initial sets.

3. If $sum \geq \alpha$, then add the centroid of the candidates found by step 2 to the final set of candidates EC.

4. Repeat the procedure until all the initial sets become empty.

3.3.4 Candidate classification

To improve the PLR value, we use a consequent classification step, which is based on certain unique features of MAs. Candidates are

classified as actual MAs or non-MAs in two steps. First, we train our approach with several fixed size (e.g. 21×21 pixels) subimages for both microaneurysm and random non-microaneurysm examples. Then, for each pixel of the examples, we establish a kernel density estimator for both classes. After the training step, we can classify new instances. We establish a new instance by producing a subimage of the candidate pixel and its neighbourhood with the same size as the training step. The classification procedure is then the following: for each pixel of the instance, we compare the probability provided by the kernel density estimators for both classes. Then, the candidate is considered as a microaneurysm if more comparisons confirm that this is a positive example.

3.3.5 Results and limitations

We have tested our approach on 50 images selected from the Retinopathy Online Challenge database (see section 2.1.1). To increase the PLR, we consider those candidates only, which have at least $\alpha = 0.5$ confidence as given in definition 5. Furthermore, we also considered higher α values to select candidates having more votes from the candidate extractors. As it can be seen in Table 2, PLR increases with the level of confidence. It is also clearly visible that the ensemble system (regardless of the number of votes) outperforms the individual candidate extractors. The largest PLR value has been found in the case $\alpha = 1$, that is, when all the 4 votes were assigned to the candidates. The number of candidates decreases with the raise of the votes, while the positive likelihood ratio shows increases.

It is also interesting to compare our approach to a state-of-the-art microaneurysm detector. Mizutani et al. [63] use a similar candidate extraction algorithm as the Spencer-Frame method, but it relies on a different approach for the final candidate classification. The ensemble system outperforms this algorithm (see Table 3) without even classifying

Algorithm	**Candidates**	TP	FP	PLR
Walter	2831	110	2721	0,040
Spencer	5821	115	5706	0,020
Hough	15664	94	15570	0,006
Lázár	11040	197	10843	0,018
$EC_{0.5}$	868	69	799	0,086
$EC_{0.75}$	847	72	775	0,093
EC_1	441	44	397	0,117

Table 2: Comparison of the majority voting ensemble with microanuerysm candidate extractors.

the candidates. However, if we apply our classification procedure explained in section 3.3.4, we may gain further improvement as it is also shown in Table 3. However, despite the improvement over the individual approaches, the results achieved with this approach are still not sufficient for clinical use because of the low number of true detections. In the next section, a novel extension to this approach is presented to microanuerysm detection to tackle this problem.

Algorithm	**Candidates**	TP	FP	PLR
Mizutani [63]	225	20	205	0.097
EC_1	441	44	397	0.117
EC_1 (classified)	270	36	234	0.153

Table 3: Comparison of the majority voting ensemble with a microaneurysm detector.

3.4 Microaneurysm detection with ⟨PP, CE⟩ ensembles

One of the main difficulties of MA detection is that candidate extractors detect only a limited amount of MAs, while it would be essential to achieve a high sensitivity at this stage. In this section, we propose an approach to overcome this difficulty by taking advantage of the diversity of candidate extractors by using different preprocessing methods within one framework. The merge of diverse candidate output can lead to a higher number of true detections, while the increasing number of false detections can be narrowed at a later stage (e.g. classification, voting, etc). The optimal combination of ⟨PP, CE⟩ pairs can be found via a search algorithm (see section 3.4.1), for which a proper energy function is needed. In the case of MA detection, this function should be defined to provide small values for a high number of true MA detections, while it also keeps the number of false detections as low as possible. A schematic description of ensemble creation is shown in Figure 18.

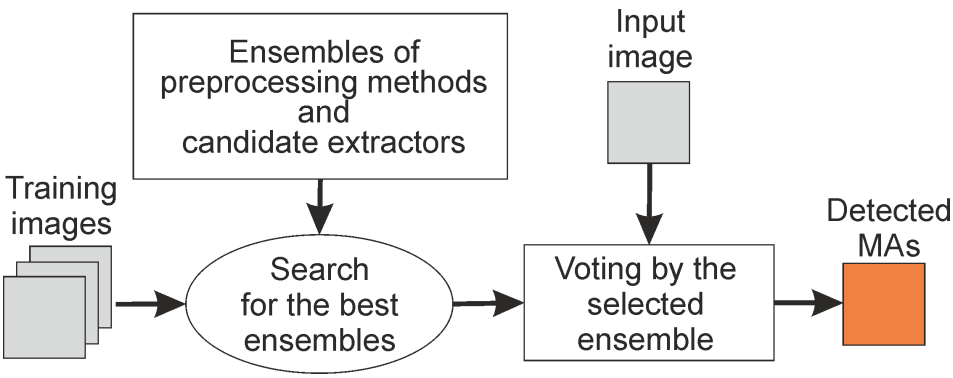


Figure 18: Flow chart of the proposed ensemble-based framework for microaneurysm detection.

The approach proposed in this section is the work of the author

and published in [19], [20], [21], [22], [23], [25], [26], [27].

3.4.1 Improving microaneurysm detection sensitivity using ⟨PP, CE⟩ ensembles

A significant number of recent works focus on the use of multiple algorithms as an ensemble, mostly for classification purposes [18]. Our task is to select the optimal combination of ⟨PP, CE⟩ pairs on the basis of the MA candidate outputs generated for each of them. A combination is considered as optimal if it detects as many microaneurysms as possible, while keeping the number of false detections low. Our search framework is based on simulated annealing [82], which is a stochastic search algorithm that does not apply any restrictions to the number of preprocessing methods and candidate extractors.

Since the number of ⟨PP, CE⟩ pairs can be very high, exhaustive search would be very resource demanding. Instead, we used simulated annealing, which is a widely used global optimization method. It is effective for large ensemble pool problems by using random sampling to avoid getting stuck in a local minimum. A crucial point in an optimization problem is the choice of the energy function. The selection of the appropriate energy function is specific to the problem, and we will provide a detailed explanation for our choice in section 3.4.3.

In order to minimize an energy function, each element of the ensemble pool $\mathcal{E}$ (see definition 3) is evaluated. A set of MA candidates belongs to each such pair, extracted by the given candidate extraction algorithm in the images with the corresponding preprocessing method applied before. The corresponding energy function value is computed on the joint candidate sets belonging to the pairs in the collection (see Definition 4). The candidates of this collection are compared to a set of microaneurysm centroids (ground truth) selected manually by clinical experts (see Definition 6). In our experiments, we set the distance threshold to 5 pixels (see Definition 4) at an image resolution

of 768 x 576 pixels, which value is calculated from the maximal MA diameter ($100\mu m$). See Algorithm 1 for the formal description of the combination procedure.

Algorithm 1 Combination of preprocessing methods and candidate extractors.

Input: An initial temperature $T \in \mathbb{R}$.

Input: A minimal temperature $T_{min} \in \mathbb{R}$.

Input: A temperature change $q \in \mathbb{R}$ with $0 \leq q \leq 1$.

Input: A set $E = \{\langle PP_i, CE_j \rangle \,|\, i = 1, \ldots, N,\, j = 1, \ldots, M\}$ containing all $\langle$PP, CE$\rangle$ pairs.

Input: A ensemble pool $\mathcal{E} = P(E)$, where P is the power set containing all possible collections of the $\langle$PP, CE$\rangle$ pairs.

Input: A function $rand(X)$, which returns a random element x from the set X.

Input: A function $neighbour(E)$, which returns the set of neighbours for the state $E \in \mathcal{E}$.

Input: A function $accept : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times [0,1] \to \{true, false\}$, which is defined in the following way:

$$accept(e,\, e_i,\, T,\, r) = \begin{cases} true, & \exp\left(\dfrac{e - e_i}{T}\right) > r, \\ false, & \text{otherwise.} \end{cases}$$

Input: An energy function $En : \mathcal{E} \to \mathbb{R}$.

Output: $E_{best} \in \mathcal{E}$, where $En(E_{best}) = \min\limits_{E \in \mathcal{E}} En(E)$.

Algorithm 1 Combination of preprocessing methods and candidate extractors (continued).

$E \leftarrow rand\,(\mathcal{E})$ // Initial state.
$e \leftarrow En\,(E)$ // Energy for the initial state.
$\mathcal{E} \leftarrow \mathcal{E} - \{E\}$ // Remove the investigated state from the ensemble pool.
while $\mathcal{E} \neq \emptyset$ and $T > T_{min}$ **do**
    $E_i \leftarrow rand\,(neighbour\,(E_i))$
    $e_i \leftarrow En\,(E_i)$
    $\mathcal{E} \leftarrow \mathcal{E} - \{E_i\}$
    if $e_i < e$ **then**
        $E \leftarrow E_i$ // If the energy of the investigated state is lower then the current minimum, then it is selected as the current optimum.
        $e \leftarrow e_i$
        $T \leftarrow T \cdot q$ // The temperature is decreased with the defined rate.
    else
        $r \leftarrow rand\,(\mathbb{R})$ // A random real number between 0 and 1 is generated for the acceptance function.
        if $accept\,(e,\, e_i,\, T,\, r) = true$ **then**
            $E \leftarrow E_i$ // The state can also be accepted via the acceptance function despite its energy value is not lower than the current minimum. The algorithm does not get stuck locally in this way.
            $e \leftarrow e_i$
            $T \leftarrow T \cdot q$
        end if
    end if
end while
return E // The state with the lowest energy is returned.

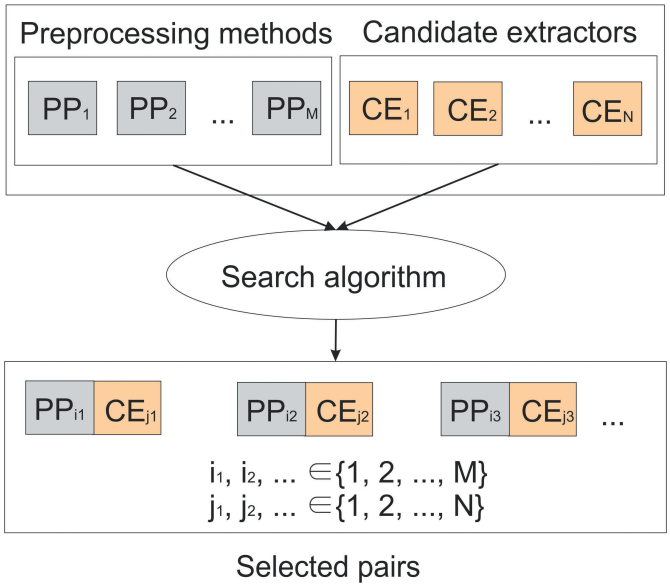


Figure 19: The proposed ensemble selection framework for improving microaneurysm candidate extraction sensitivity.

Note that, we consider $N = 4$ preprocessing methods (namely: Walter-Klein, CLAHE, Vessel removal and interpolation, No preprocessing) and $M = 5$ candidate extractors (namely: Walter, Spencer, Hough, Lázár, Zhang), respectively, but with the use of simulated annealing we can easily include more methods in the future. Currently, the largest number of $\langle$PP, CE$\rangle$ pairs in an ensemble is $5 \times 4 = 20$ in our case, and the ensemble pool contains $2^{20} - 1$ possible ensembles. For a more illustrative representation of the approach, see Figure 19.

3.4.2 Methodology

We have tested our approach on 199 images selected from three databases: the training set of the Retinopathy Online Challenge (ROC)

database (section 2.1.1), the DIARET2.1 database (section 2.1.3) and our own database, which was provided by the Moorefields Eye Hospital, London, UK (section 2.1.4). The tests were executed in two ways:

- CASE 1: We selected the preprocessing methods for each candidate extractors and measured the change in its sensitivity by using the combination. More formally, we restricted the ensemble pool to $\{\langle PP_i, CE\rangle \,|\, i = 1, \ldots, 4\}$ in Algorithm 1 with fixing a specific candidate extractor.

- CASE 2: We measured the effects using multiple candidate extractors in the selection process as it is described in Algorithm 1. That is, the set $\{\langle PP_i, CE_j\rangle \,|\, i = 1, \ldots, 4,\ j = 1, \ldots, 5\}$ was considered to be the the ensemble pool.

For both cases, we disclose the results based on 100 runs, splitting the database into two disjoint parts randomly in every run.

3.4.3 Energy functions

Since it is a crucial point to choose a suitable energy function in an optimization process, we decided to select the most appropriate one by evaluating for multiple possibilities. We have selected such energy functions, which take into account both TPs and FPs (see definition 6). Namely, we checked the following energy functions in our search framework:

$$En_{SEN} = -\frac{SEN}{FPI}, \tag{21}$$

$$En_{PPV} = -PPV, \tag{22}$$

$$En_{FSCORE} = -\text{F-score}, \tag{23}$$

$$En_{LN} = -\frac{TP}{\ln \frac{FP}{TP}}. \tag{24}$$

The components of these energy functions are defined in section 3.1. While the first three functions are based on literature recommendations, we also propose an additional one, which is competent with the others. We cannot consider any measure which relies on true negatives (TN) like specificity, due to the high number of TNs. Namely, every pixels in an image which is not a candidate centroid and not marked as MA by the experts can be regarded as a TN. For this reason, such measures are generally excluded in such image processing algorithms [83].

The evaluation of the energy functions is accomplished in the following way: we executed the selection process using all the energy functions listed above simultaneously for CASE 1 and CASE 2 described in section 3.4.2. The respective results of this evaluation for CASE 1 are disclosed in Tables 4, 5, 6 and 7 for the ROC, Diaret2.1, Moorefields database and for all images in the three databases and in Table 8 for CASE 2, respectively. Each cell in the tables contains the average sensitivity values and their standard deviations for each case using the given energy function for selection.

From these results, we can conclude that the energy function En_{PPV} defined in (21) fits our goals the most to have a large number of TPs with a rather low FP count. It is also clear, that (22) and (23) are not suitable for this optimization process, while (24) could also be taken into account. However, the variability of (24) is higher than that of (21), so we have chosen (21) in the selection process.

3.4.4 Results and discussion

The performance of the ensembles found by Algorithm 1 using the energy function (21) are shown in Tables 9, 10, 11 and 12 regarding

CE	En_{SEN}	En_{PPV}	En_{FSCORE}	En_{LN}
Walter	$\mathbf{0.5523 \pm 0.0479}$	0.1648 ± 0.0247	0.1836 ± 0.0652	0.5386 ± 0.0597
Spencer	$\mathbf{0.2400 \pm 0.0487}$	0.0811 ± 0.0291	0.0116 ± 0.0539	0.2319 ± 0.0422
Hough	$\mathbf{0.2392 \pm 0.0545}$	0.0934 ± 0.0847	0.1179 ± 0.1011	0.0265 ± 0.0350
Lázár	$\mathbf{0.5980 \pm 0.0326}$	0.2525 ± 53.0274	0.2563 ± 0.0267	0.3686 ± 0.0460
Zhang	$\mathbf{0.5583 \pm 0.0387}$	0.4671 ± 0.0296	0.4640 ± 0.0282	0.5546 ± 0.0354

Table 4: Average sensitivity of the candidate extractors on the ROC database using a set of preprocessing methods, which are selected using the energy functions shown in the respective columns.

CE	En_{SEN}	En_{PPV}	En_{FSCORE}	En_{LN}
Walter	$\mathbf{0.6651 \pm 0.0519}$	0.5174 ± 0.0452	0.5158 ± 0.0461	0.5673 ± 0.0710
Spencer	$\mathbf{0.0349 \pm 0.0121}$	0.0064 ± 0.0060	0.0090 ± 0.0086	0.0262 ± 0.0161
Hough	$\mathbf{0.1933 \pm 0.0274}$	0.0800 ± 0.0765	0.1164 ± 0.0641	0.1405 ± 0.0833
Lázár	$\mathbf{0.8066 \pm 0.0379}$	0.4963 ± 0.0385	0.4964 ± 0.0426	0.5283 ± 0.1018
Zhang	$\mathbf{0.4709 \pm 0.0289}$	0.04029 ± 0.0269	0.4015 ± 0.0289	0.3580 ± 0.1727

Table 5: Average sensitivity of the candidate extractors on the Diaret2.1 database using a set of preprocessing methods, which are selected using the energy functions shown in the respective columns.

CASE 1, and in Table 13 for CASE 2, respectively. Besides sensitivity, we also disclose other measures to describe the performance of the ensemble selection framework: the number of FPs per image (FPI (11)), the positive predictive value (PPV (9)) and the difference between the sensitivity of the ensemble and the sensitivity of the best performing individual candidate extractor on the same test database (DS).

From the DS values it can be seen, that the proposed framework increases sensitivity remarkably compared to the individual approaches. On the other hand, the number of false positives also increased in a relatively smaller extent. In Figure 20, a subimage containing both TPs and FPs is shown. FPs originate from artefacts, image compression errors or thin vessel parts. The proportion of true and false detection can be improved by using classification, voting or other post-processing techniques, as it will be shown in section 3.5.

CE	En_{SEN}	En_{PPV}	En_{FSCORE}	En_{LN}
Walter	$\mathbf{0.7543 \pm 0.0470}$	0.6474 ± 0.0855	0.6497 ± 0.0740	0.7478 ± 0.0410
Spencer	$\mathbf{0.2317 \pm 0.0225}$	0.1169 ± 0.0315	0.1181 ± 0.0326	0.2296 ± 0.0222
Hough	$\mathbf{0.2307 \pm 0.0183}$	0.2240 ± 0.0204	0.2214 ± 0.0184	0.2297 ± 0.0222
Lázár	$\mathbf{0.8077 \pm 0.0215}$	0.2723 ± 0.0251	0.2730 ± 0.0242	0.7106 ± 0.1990
Zhang	$\mathbf{0.5081 \pm 0.0208}$	0.3235 ± 0.0234	0.3278 ± 0.0230	0.4921 ± 0.0269

Table 6: Average sensitivity of the candidate extractors on the Moorefields database using a set of preprocessing methods, which are selected using the energy functions shown in the respective columns.

CE	En_{SEN}	En_{PPV}	En_{FSCORE}	En_{LN}
Walter	$\mathbf{0.5850 \pm 0.0325}$	0.5312 ± 0.0332	0.5312 ± 0.0312	0.5824 ± 0.0324
Spencer	$\mathbf{0.1755 \pm 0.0229}$	0.0623 ± 0.0170	0.1401 ± 0.0700	0.1749 ± 0.0228
Hough	$\mathbf{0.0445 \pm 0.0147}$	0.0235 ± 0.0148	0.0250 ± 0.0139	0.0428 ± 0.0167
Lázár	$\mathbf{0.6290 \pm 0.0262}$	0.3370 ± 0.0232	0.3381 ± 0.0271	0.4763 ± 0.1587
Zhang	$\mathbf{0.3824 \pm 0.0404}$	0.2698 ± 0.0210	$0.2270 \pm 0.0.0337$	0.3797 ± 0.0385

Table 7: Average sensitivity of the candidate extractors on the image of all three databases using a set of preprocessing methods, which are selected using the energy functions shown in the respective columns.

Database	En_{SEN}	En_{PPV}	En_{FSCORE}	En_{LN}
ROC	$\mathbf{0.7447 \pm 0.0453}$	0.3319 ± 0.0924	0.3061 ± 0.0796	0.7013 ± 0.1446
Diaret2.1	$\mathbf{0.9820 \pm 0.0041}$	0.5346 ± 0.0777	0.5289 ± 0.0757	0.5886 ± 0.2248
Moorefields	$\mathbf{0.9565 \pm 0.0287}$	0.3317 ± 0.0677	0.3314 ± 0.0666	0.8102 ± 0.1886
All	$\mathbf{0.8711 \pm 0.0303}$	0.3785 ± 0.0564	0.3833 ± 0.0616	0.8555 ± 0.0322

Table 8: Average sensitivity of the candidate extractor and preprocessing method combinations for the energy functions shown in the respective columns.

CE	Sensitivity	FPI	PPV	DS
Walter	0.5523 ± 0.0479	572.1612 ± 86.7554	0.0066 ± 0.0015	0.1353 ± 0.0332
Spencer	0.2400 ± 0.0487	151.8692 ± 9.5329	0.0103 ± 0.0030	0.1142 ± 0.0232
Hough	0.2392 ± 0.0545	512.4052 ± 46.6152	0.0031 ± 0.0009	0.0566 ± 0.0334
Lázár	0.5980 ± 0.0326	569.3936 ± 53.0947	0.0071 ± 0.0016	0.1934 ± 0.0275
Zhang	0.5583 ± 0.0387	540.2896 ± 27.8082	0.0069 ± 0.0014	0.0862 ± 0.0188

Table 9: Average sensitivity, false positive per image, positive predictive value and sensitivity difference of the candidate extractors on the ROC database using a combination of preprocessing methods.

CE	Sensitivity	FPI	PPV	DS
Walter	0.6651 ± 0.0519	443.9127 ± 29.7674	0.0065 ± 0.0019	0.0712 ± 0.0192
Spencer	0.0349 ± 0.0121	165.7153 ± 13.3925	0.0009 ± 0.0003	0.0162 ± 0.0081
Hough	0.1933 ± 0.0274	313.5056 ± 18.8335	0.0029 ± 0.0011	0.0428 ± 0.0133
Lázár	0.8066 ± 0.0379	663.5811 ± 16.7470	0.0055 ± 0.0013	0.1778 ± 0.0161
Zhang	0.4709 ± 0.0289	129.0984 ± 14.4546	0.0171 ± 0.0061	0.0703 ± 0.0114

Table 10: Average sensitivity, false positive per image, positive predictive value and sensitivity difference of the candidate extractors on the Diaret2.1 database using a combination of preprocessing methods.

CE	Sensitivity	FPI	PPV	DS
Walter	0.7543 ± 0.0470	792.4597 ± 32.4551	0.0080 ± 0.0017	0.0888 ± 0.0141
Spencer	0.2317 ± 0.0225	164.6963 ± 6.2115	0.0123 ± 0.0020	0.1092 ± 0.0142
Hough	0.2307 ± 0.0183	213.3273 ± 11.7408	0.0098 ± 0.0019	0.0025 ± 0.0002
Lázár	0.8077 ± 0.0215	606.1647 ± 15.2953	0.0115 ± 0.0023	0.3089 ± 0.0186
Zhang	0.5081 ± 0.0208	162.2053 ± 8.7651	0.0265 ± 0.0051	0.0983 ± 0.0137

Table 11: Average sensitivity, false positive per image, positive predictive value and sensitivity difference of the candidate extractors on the Moorefields database using a combination of preprocessing methods.

CE	Sensitivity	FPI	PPV	DS
Walter	0.5850 ± 0.0325	327.4260 ± 24.5482	0.0113 ± 0.0017	0.0556 ± 0.0066
Spencer	0.1755 ± 0.0229	165.4692 ± 2.7000	0.0067 ± 0.0011	0.1077 ± 0.0137
Hough	0.0445 ± 0.0147	133.8911 ± 11.9872	0.0021 ± 0.0007	0.0114 ± 0.0034
Lázár	0.6290 ± 0.0262	401.1449 ± 17.0726	0.0099 ± 0.0014	0.1279 ± 0.0130
Zhang	0.3824 ± 0.0404	250.7720 ± 14.3805	0.0097 ± 0.0013	0.0723 ± 0.0089

Table 12: Average sensitivity, false positive per image, positive predictive value and sensitivity difference of the candidate extractors on all three databases database using a combination of preprocessing methods.

Database	Sensitivity	FPI	PPV	DS
ROC	0.7447 ± 0.0453	808.8792 ± 47.5289	0.0060 ± 0.0012	0.2763 ± 0.0336
Diaret2.1	0.9820 ± 0.0041	1121.8218 ± 70.7609	0.0038 ± 0.0009	0.3516 ± 0.0363
Moorefields	0.9565 ± 0.0287	1135.8917 ± 36.8809	0.0074 ± 0.0014	0.3471 ± 0.0398
All	0.8711 ± 0.0303	985.8611 ± 43.9246	0.0055 ± 0.0008	0.3406 ± 0.0287

Table 13: Average sensitivity, false positive per image, positive predictive value and sensitivity difference of the candidate extractor and preprocessing method combinations on the respective databases.

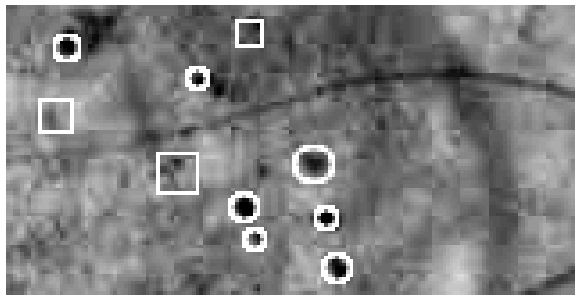


Figure 20: Results of an microaneurysm candidate extraction by an ensemble. The circles and squares depict the true and false detections, respectively.

	CASE 1			CASE 2		
Number of pairs	1	2	3	9	10	11
Occurrences	336	962	702	7	389	4

Table 14: The number of ⟨PP, CE⟩ pairs selected for the ensembles in different test runs for improving microaneurysm candidate extraction sensitivity.

We have also investigated the cardinality of the best ensembles for CASE 1 and CASE 2. Table 14 shows the occurrence of the number of the selected pairs in each case. As it can be seen, there is a dominant value for the number of pairs for both cases. In CASE 1, most selections contained 2 pairs, while in CASE 2, in almost every case 10 pairs were selected.

3.5 Suppressing the number of false microaneurysm detections

In this section, an effective microaneurysm detector based on the combination of preprocessing methods and candidate extractors is proposed. An exhaustive quantitative analysis is also given to prove the superiority of our approach over individual algorithms.

3.5.1 Ensemble selection for MA detection

We start with describing our ensemble creation approach based on the results presented in the previous section. As an energy function, we used the competition performance metric CPM (15), which is defined as the average sensitivity level at seven predefined false positive per image rate $(1/8, 1/4, 1/2, 1, 2, 4, 8)$ [37].

The ensemble creation part (see Algorithm 1) results in a set of ⟨PP, CE⟩ 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 have a confidence value (see definiton 5). 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 3.5.2.5 in detail.

3.5.2 Results

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.

3.5.2.1 MA detection We have evaluated the MA detection capabilities of the proposed method in the ROC competition for MA detectors [37], as well, as on the publicly available DiaretDB1 (section 2.1.3) and a private database (section 2.1.4). 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.

3.5.2.1.1 Testing For each database, we provide the Free-response Receiver Operating Characteristic (FROC) curves [83], 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 database, we also provide the current ranking of the competition along with the CPM values, that serves as the basis for the ranking. In addition, we also calculated a partial area under the curve (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 value 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.

3.5.2.2 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 database containing 1200 images (section 2.1.5). 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. The flow chart of this approach can be seen in Figure 21.

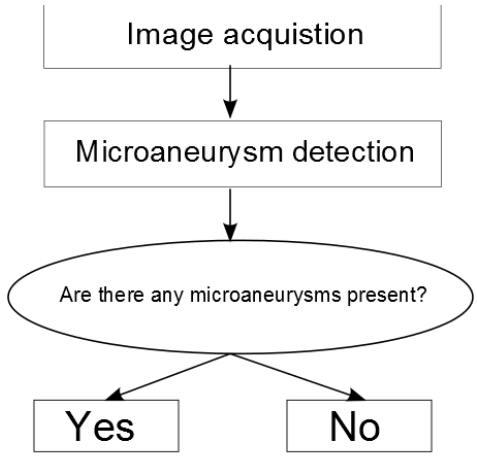


Figure 21: Flow chart of the early decision support framework.

3.5.2.2.1 Ensemble creation As there is no training set provided for the Messidor database, we used an independent database (ROC) 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.

3.5.2.2.2 Testing In our evaluation, we classified the retinal images whether they contain signs of DR (R1, R2, R3) or not (R0) (see section 2.1.5 for details). The MA detector classifies an image as diseased if at least one MA is detected, and healthy otherwise. We measured the sensitivity (5), specificity (6) and accuracy (12) of the detector at different levels by thresholding the confidence values assigned to the MA candidates. We also measured 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 [84].

3.5.2.3 Microaneurysm detection results In Table 15, we exhibit the ⟨PP, CE⟩ pairs included in the selected ensembles for the three databases, respectively. The rows of the table show the preprocessing methods , while the columns label the candidate extractor algorithms.

Table 16 contains the ranked quantitative results of the participants of the ROC competition, with the proposed ensemble (DRSCREEN) highlighted as the current leader. The performance of the ensemble is also shown in Figure 24 in terms of a FROC curve. As we can see from Table 16, the proposed ensemble earned both a higher CPM score and a higher partial AUC than the individual algorithms.

	Walter	Spencer	Hough	Lázár	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 15: ⟨PP, CE⟩ pairs selected as members of the microaneurysm detection ensemble for the three database. R, D, M denote whether the pair is selected for the ROC, Diaret2.1, or the Moorfields database, respectively.

3.5.2.4 Diabetic retinopathy grading results In Table 17, we provide the sensitivity (SEN (5)), specificity (SPE (6)) and accuracy (ACC (12)) measures of our detector corresponding to different threshold values, respectively. The fitted ROC curve of the detector can be seen in Figure 22. The empirical area under curve (AUC) is 0.875, while the AUC for the fitted curve is 0.90 ± 0.01. Table 17 also contains the percentage of the correctly recognized cases for each class.

3.5.2.5 Discussion A strong point of the proposed method is that it performs well under difficult circumstances. Figure 23 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 [18]. This diversity ensures the suppression of false detections, since diverse detectors tend to make different mistakes.

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
Lázár 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

Table 16: Quantitative results of the Retinopathy Online Challenge for microaneurysm detection. For each participating team, the competition performance metric and the partial area under the FROC curve are presented.

Thus, the false detections are likely to receive lower confidence values in the voting procedure.

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 databases. As it can be seen in Table 15, 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 [37], the FPI rate for a human expert is measured in the ROC database against the consensus of three human experts. This level is approximately 1 FP per image [37] 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

Threshold	0.4	0.5	0.6	0.7	0.8	0.9	1.0
SEN	1	1	1	0.99	0.96	0.76	0.31
SPE	0	0.01	0.03	0.14	0.51	0.88	0.98
ACC	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

Table 17: Results of diabetic retinopathy grading on the Messidor database based on the microanuerysm detection results. For each threshold, sensitivity, specificity, accuracy and the percentage of correctly recognized cases for each grade are presented.

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 it is 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 cases 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 [85], it has been suggested that sensitivity is more important for a screening system than specificity. In opposition, the British Diabetic Association (BDA) recommends 80% sensitivity and 95% specificity for DR screening [86]. In Table 17, 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

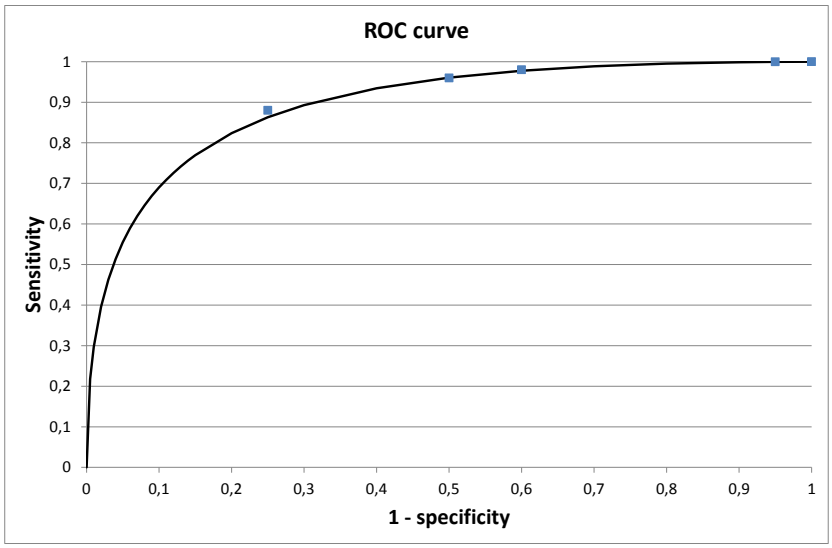


Figure 22: Receiver operating characteristic curve of the microaneurysm detector ensemble on the Messidor database.

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, since those are tested on private databases. Unfortunately, the proportion of non-DR/DR cases are varying in these experiments. Abramoff et al. [85] 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

(a) Original (b) CLAHE (c) Vessel removal

Figure 23: The effect of different preprocessing methods where microaneurysms are hard to detect.

et al. [87] tested, 74.43% and 76.26% of the 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. [88] 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. [89] in their automatic screening system.

Despite the promising results, our system still misclassifies some images, 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, as we will show it in section 4.

The FROC curves of the ensemble for the DiaretDB1 2.1 and for the Moorfields database are shown in Figures 25 and 26, 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.

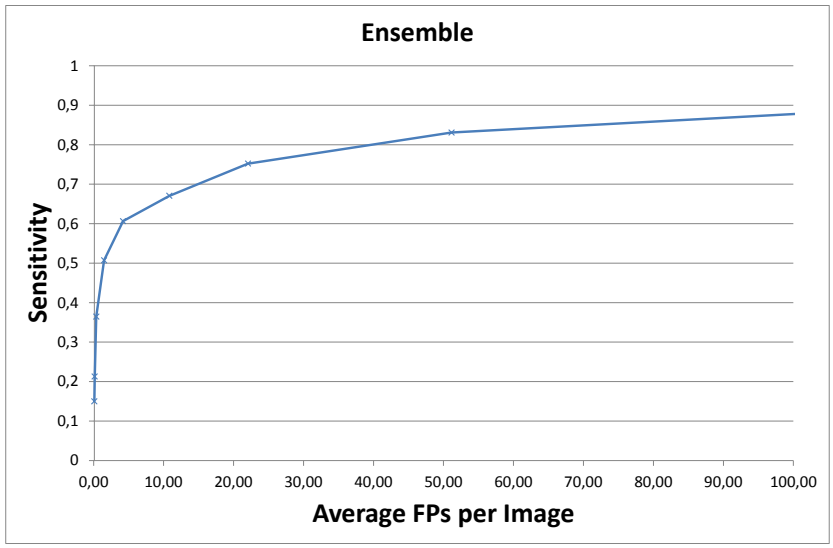


Figure 24: Free-response receiver operating characteristic curve of the microaneurysm detection ensemble on the Retinopathy Online Challenge.

3.6 Context-aware weighting of ⟨PP, CE⟩ ensembles for microaneurysm detection

The detection of MAs highly depends on the characteristics of the imaging device and other image properties (e.g. type of compression). As a result, some MAs can be easily spotted on the background of the retina, while the recognition of others are more difficult. Besides image characteristics, the spatial location also has an influence on the detection of MAs (e.g. proximity of vessel parts, etc.)

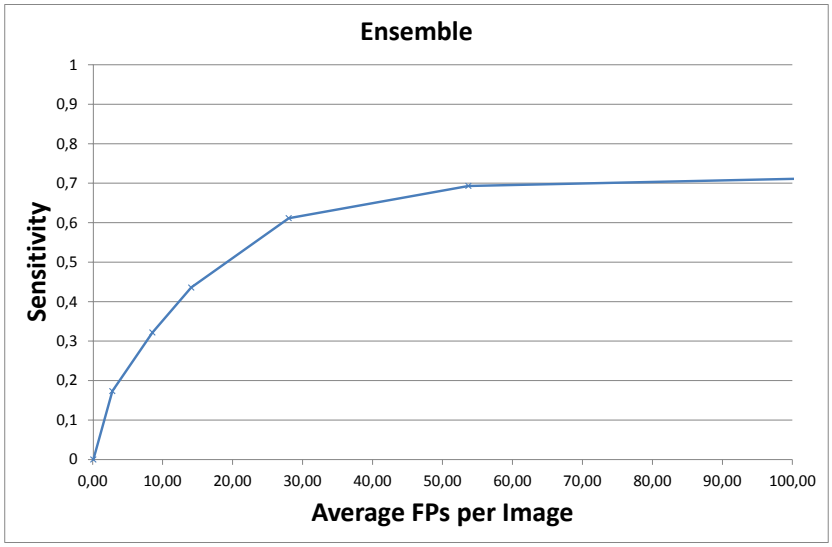


Figure 25: Free-response receiver operating characteristic curve of the microaneurysm detection ensemble on the DiaretDB1 2.1 database.

In [37], Niemeijer et al. distinguishes three categories based on visibility: subtle, regular and obvious. An example for this categorization can be seen in Figure 27. In the same study, they also investigate the detection of MAs near vessel. We extend these categorization with two additional categories with also taking into account the MAs which are detected on the macula and which are on the periphery of the image. Figure 28 shows examples for the spatial categories. We also provide a computational approach to determine the characteristics of the MAs.

To recognize MAs in the different categories, we measure the effect of using different preprocessing methods. As we can see later

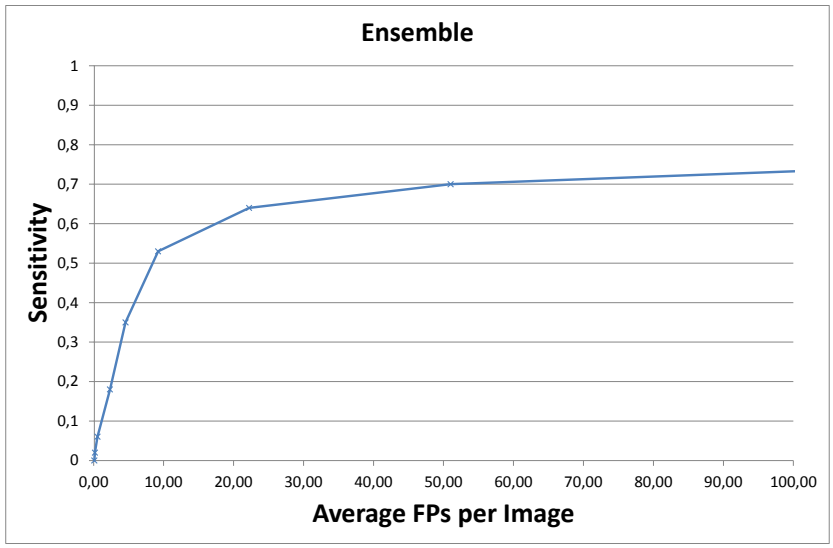


Figure 26: Free-response receiver operating characteristic curve of the microaneurysm detection ensemble on the Moorfields database.

on, a preprocessing method can enhance the detection rate in a few categories, but there is no single preprocessing method for all. To overcome this difficulty, we propose a context-aware selection approach of preprocessing methods for MA candidate extraction.

3.6.1 Context-aware selection of ⟨PP, CE⟩ pairs

In this section, we describe our context-aware preprocessing method selection approach, which is based on learning. Thus, a training database with manually labelled MAs is needed. The creation of such

(a) Subtle (b) Regular (c) Obvious

Figure 27: Microaneurysm categories based on visibility.

(a) Near vessel (b) In macula (c) Periphery

Figure 28: Microaneurysm categories based on spatial location.

database is explained in details in section 3.6.1.1.

The approach proposed in this section is published in [28], and [29].

3.6.1.1 Preparation of the training data First, the manually labelled MAs of the training database must be categorized. Opposed to the manual way described in [37], we set up an automatic method described next.

3.6.1.1.1 Categorization based on visibility To measure the visibility of an MA, we select an $A \times B$ (e.g. $A = B = 20$ in our case) window centered on the MA centroid and measure the contrast in this region in the following way:

$$\sqrt{\frac{1}{AB}\sum_{x=1}^{A}\sum_{y=1}^{B}\left(f\left(x,y\right)-\mu\right)^{2}}, \tag{25}$$

where $f(x, y)$ is the corresponding intensity of the pixel having coordinates (x, y) and μ is the average intensity of the window.

Since we do not have any prior knowledge about the distribution of MAs based on the contrast information, we aimed to divide the MAs into three sets with equal cardinalities. Thus, we categorize an MA as *subtle*, if its contrast is lower than the 33th percentile of the observed contrast values in the training set, *obvious*, if its contrast is higher than the 66th percentile and *regular,* otherwise.

3.6.1.1.2 Categorization based on spatial location We also categorized MAs into three more categories based on their spatial locations: *near to vessel*, *in the macula* and *on the periphery*. For the first category, we must detect the vessel system of the retina. For this task, we used the method presented in [90]. We consider an MA as *near to vessel*, if it is closer to a vessel part than the maximal MA diameter. For the second category, we detected macula with the detector proposed in [31]. Then, we collect the MAs falling into the area of the macula to the *in the macula* category. Finally, MAs *on the periphery* are determined in the following way: first, the radius of the retinal ROI is calculated. Then, each MA having a distance at least 90% of the radius from the center of the retina is considered as peripheral MA.

3.6.1.2 Training In the training step, each preprocessing method is applied on the training set individually and their effect on the candidate extractor is evaluated. That is, the number of true and false positives, and the number of correctly identified MAs for each category are measured, respectively. The best performing preprocessing method for each category is selected.

3.6.1.3 Testing On unknown images, the results of the candidate extractor with the selected preprocessing methods are collected and the results are merged as the union of candidate extractor output. The merged MA candidates are the output of the proposed candidate extraction approach.

3.6.1.3.1 Methodology Our experimental test is performed on the DiaretDB1 2.1 database (section 2.1.3). The selection of the preprocessing method is solely based on the results achieved in the training set. Detailed information about the number of images and MAs can be found in Table 18. For each preprocessing method listed in section 3.2.1, we used the same parameter setting for the Lázár et al. (section 3.2.2.5) MA candidate extractor.

Category	**Train**	**Test**
MA	126	278
Images	28	61
Subtle	32	101
Regular	48	86
Obvious	46	91
In macula	4	4
Near vessel	25	43
Periphery	39	69

Table 18: The number of all microaneurysms, images, and the microaneurysms belonging to each category for the train and the test databases, respectively.

3.6.1.4 Results and discussion Table 19 contains the results of the training phase with the highest number of correctly recognized MAs are highlighted with the corresponding preprocessing method. As

it can be seen, the CL, the IA and the VR preprocessing methods (see section 3.2.1) are selected, while each performed better than the rest of the algorithms in two categories.

	NO	**BS**	**WK**	**CL**	**GN**	**HE**	**IA**	**VR**
TP	57	49	29	61	16	52	54	65
FP	1572	919	1157	1799	231	859	659	1657
Subtle	14	5	8	16	0	11	6	**18**
Regular	21	15	8	**24**	2	17	18	22
Obvious	22	29	13	21	14	25	**30**	25
In macula	2	1	2	**3**	0	1	2	2
Near vessel	4	7	4	4	3	7	**8**	4
Periphery	12	7	7	16	2	16	11	**19**

Table 19: The number of true and false detections and the number of correctly recognized cases for each microaneurysm category in the training database. The CL, the IA and the VR preprocessing methods are selected since they achieved the highest number of correctly identified microaneurysms in at least one category.

In Table 20, the results of the proposed approach and the individual methods can be seen. The proposed method provided the highest number of correctly recognized MAs in each category. However, the number of false detections also increased, but it can be lowered further by applying a voting scheme.

While these results are reassuring, it should also be noted that the MA detection performance for some categories (e.g. near vessel, periphery) are rather low. In the future, the inclusion of other preprocessing methods and MA candidate extractors can lead to a better performance in these cases.

3.6.2 Adaptive weighting

In this section, we show a way to combine the output of the ⟨PP, CE⟩ pairs by weighting. We assign weights to each candidate with a

	NO	BS	WK	CL	GN	HE	IA	VR	Proposed
TP	147	67	54	154	28	60	72	116	198
FP	3023	1625	2303	3105	240	1500	765	3383	5521
Subtle	41	9	12	46	0	11	7	27	64
Regular	53	22	15	55	22	24	25	43	67
Obvious	53	36	27	53	22	25	44	46	68
In macula	2	0	0	2	0	1	1	2	3
Near vessel	7	8	2	6	6	3	9	8	18
Periphery	23	6	8	27	1	8	7	18	38

Table 20: The number of true and false detections and the number of correctly recognized cases for each microaneurysm category in the test database. The proposed combination outperformed the individual preprocessing methods in all categories.

respect to three different pieces of information: which pair extracted the output, what is the contrast in the neighbourhood of the MA and where it is located in the image.

3.6.2.1 The weighted voting approach The performance of each ⟨PP, CE⟩ pair is measured for each category in the following way: each extracted candidate is categorized both by visibility and spatial location, then compared to the ground truth whether it is actually an MA or not. Based on this evaluation, for each pair p, visual category v and spatial category s, we calculate the F-score measure (10):

$$\text{F-score}_{vs}^{p} = \frac{2 \cdot SEN_{vs}^{p} \cdot PPV_{vs}^{p}}{SEN_{vs}^{p} + PPV_{vs}^{p}}, \tag{26}$$

where for label v:

$$v \in \{\text{subtle}, \text{obvious}, \text{regular}\}, \tag{27}$$

and for label s:

$$s \in \{\text{near to vessel}, \text{in the macula}, \text{on the periphery}, \text{other}\}. \tag{28}$$

Then, we approximate the optimal weights for ⟨PP, CE⟩ pair p through the following formula [18]:

$$w_{vs}^{p} = \log \frac{\text{F-score}_{vs}^{p}}{1 - \text{F-score}_{vs}^{p}}. \tag{29}$$

The weights are normalized for each combination of visual and spatial categories to have a sum of 1.

Finally, for each candidate in an unknown image, the visual and spatial location categories are determined and the corresponding weight value is summed as the confidence value of the candidate. The final confidence value assigned to an MA candidate is the sum of the weights of the ⟨PP, CE⟩ pairs, which detected this candidate. The selected MA candidates can be filtered by thresholding their confidence values.

3.6.2.2 Results and discussion The results of the weighting approach for the DiaretDB0 (section 2.1.2) and the ROC (section 2.1.1) database can be seen in Tables 21 and 22, respectively. The search-based algorithm in the table for comparison is a former method of ours, which based on the selection of ⟨PP, CE⟩ pairs (section 3.4). As it can be seen from the results, the proposed weighting approach provides better results on the DiaretDB0 database, but not on the ROC database. The reason for the alternating performance of the weighted and the search-based method may lie in the fact that the fundus image databases are rather different. However, both ensemble-based approach outperformed the individual detectors which shows the strength of the ensembles in this field.

	1/8	1/4	1/2	1	2	4	8	avg.
DRSCREEN (search)	0.003	0.005	0.011	0.021	0.043	0.087	0.174	0.049
DRSCREEN (weighted)	0.012	0.025	0.037	0.060	0.090	0.129	0.189	0.077

Table 21: Comparison of the search and the weighting based microaneurysm detection-ensembles.

	1/8	1/4	1/2	1	2	4	8	avg.
DRSCREEN (search)	0.173	0.275	0.380	0.444	0.526	0.599	0.643	0.434
DRSCREEN (weighted)	0.172	0.201	0.323	0.426	0.478	0.560	0.638	0.399
Niemeijer et al.	0.243	0.297	0.336	0.397	0.454	0.498	0.542	0.395
LaTIM	0.166	0.230	0.318	0.385	0.434	0.534	0.598	0.381
ISMV	0.217	0.270	0.366	0.407	0.440	0.459	0.468	0.375
OKmedical II	0.175	0.242	0.297	0.370	0.437	0.493	0.569	0.369
OKmedical	0.198	0.265	0.315	0.356	0.394	0.466	0.501	0.357
Lázár et al.	0.169	0.248	0.274	0.367	0.385	0.499	0.542	0.355
GIB	0.190	0.216	0.254	0.300	0.364	0.411	0.519	0.322
Fujita	0.181	0.224	0.259	0.289	0.347	0.402	0.466	0.310
IRIA	0.041	0.160	0.192	0.242	0.321	0.397	0.493	0.264
Waikato	0.055	0.111	0.184	0.213	0.251	0.300	0.329	0.206

Table 22: Detailed quantitative results of the Retinopathy Online Challenge (including the weighted ensemble).

4 An ensemble-based grading method for diabetic retinopathy based on detailed retinal image analysis

In section 3, we have seen how efficient MA detection can be achieved by using an ensemble-based approach. Though in section 3.4, we have already discussed grading performance based on the detection of MAs, it is a natural improvement to consider more information for such a decision. To raise accuracy, we also use an ensemble-based approach for the final grading decision.

In this section, a decision-making framework for the final grading of color fundus images regarding diabetic retinopathy is proposed. The

approach classifies images based on characteristic features extracted by lesion detection and anatomical part recognition algorithms. The features are then classified using an ensemble of classifiers. As the results show, the proposed approach is highly accurate for this task. The flow chart of the final decision approach can be seen in Figure 29.

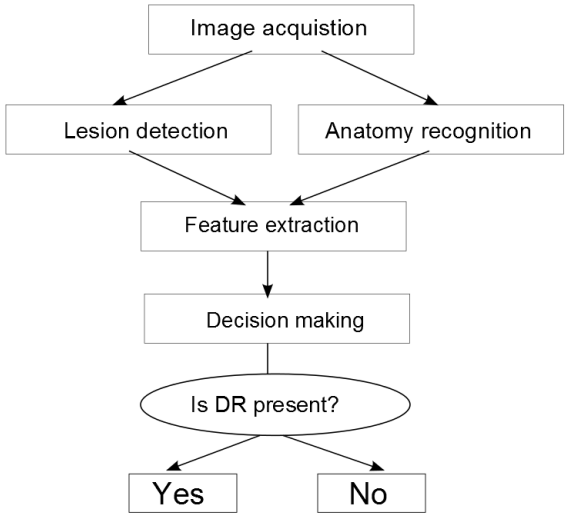


Figure 29: Flow chart of the final decision support framework.

The approach proposed in this section is also published in [30].

4.1 Ensemble learning

In this section, the basic concepts of ensemble learning is presented based on the definitions used in [18]. These concepts help to understanding of our ensemble-based system for DR grading introduced in the forthcoming sections.

Definition 11 *Let* $\Omega = \{\omega_1, \omega_2, \ldots, \omega_o\}$ *be a set of class labels. Then,*

a function $D : \mathbb{R}^n \to \Omega$ *is a called classifier, while the vector* $\bar{\chi} = [\chi_1, \chi_2, \ldots, \chi_n]^T \in \mathbb{R}^n$ *is called a feature vector.*

Definition 12 *Let* $H = \{h_1, h_2, \ldots, h_o\}$ *be discriminator functions for the class labels* $\Omega = \{\omega_1, \omega_2, \ldots, \omega_o\}$ *of the form* $h_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, o$. *Then*

$$D(\bar{\chi}) = \omega_{j*} \iff h_{j*}(\bar{\chi}) = \max_{j=1,\ldots,o} (h_j(\bar{\chi})). \tag{30}$$

Definition 13 *Let* $D_1, D_2, \ldots, D_L$ *be classifiers. Then, the majority voting ensemble classifier* $\mathcal{D}_{maj} : \mathbb{R}^n \to \Omega$ *formed from these classifiers is defined as follows:*

$$\mathcal{D}_{maj}(\bar{\chi}) = \omega_{i*} \iff \omega_{i*} = \max_{i=1,\ldots,o} \left(\sum_{j=1}^{L} \{1 | D_j(\bar{\chi}) = \omega_i\} \right), \tag{31}$$

$i = 1, \ldots o$.

Definition 14 *Let* $D_1, D_2, \ldots, D_L$ *be classifiers and* $\bar{\beta} = \{\beta_1, \beta_2, \ldots, \beta_L\}$, $\beta \in \mathbb{R}$ *a weight vector assigned to the classifiers. Then, the weighted majority voting ensemble classifier* $\mathcal{D}_{wmaj} : \mathbb{R}^n \to \Omega$ *is defined as follows:*

$$\mathcal{D}_{wmaj}(\bar{\chi}) = \omega_{i*} \iff \omega_{i*} = \max_{i=1,\ldots,o} \left(\sum_{j=1}^{L} \left\{ \beta_j \in \bar{\beta} | D_j(\bar{\chi}) = \omega_i \right\} \right). \tag{32}$$

Definition 15 *Let* $D_1, D_2, \ldots, D_L$ *be classifiers and* $h_{j,i}$ *be the discriminator function of the classifier* D_j *for the class* i, $i = 1, \ldots, o$, $j = 1, \ldots, L$. *Then, the following algebraic ensemble classifiers can be*

defined:

$$\mathcal{D}_{min}\left(\bar{\chi}\right) = \omega_{i*} \iff \omega_{i*} = \max_{i=1,\ldots,o}\left(\min_{j=1}^{L}\left(h_{j,i}\left(\bar{\chi}\right)\right)\right), \tag{33}$$

$$\mathcal{D}_{max}\left(\bar{\chi}\right) = \omega_{i*} \iff \omega_{i*} = \max_{i=1,\ldots,o}\left(\max_{j=1}^{L}\left(h_{j,i}\left(\bar{\chi}\right)\right)\right), \tag{34}$$

$$\mathcal{D}_{pro}\left(\bar{\chi}\right) = \omega_{i*} \iff \omega_{i*} = \max_{i=1,\ldots,o}\left(\prod_{j=1}^{L}\left(h_{j,i}\left(\bar{\chi}\right)\right)\right), \tag{35}$$

$$\mathcal{D}_{avg}\left(\bar{\chi}\right) = \omega_{i*} \iff \omega_{i*} = \max_{i=1,\ldots,o}\left(\frac{1}{L}\cdot\sum_{j=1}^{L}\left(h_{j,i}\left(\bar{\chi}\right)\right)\right). \tag{36}$$

4.2 Components of an automatic system for diabetic retinopathy screening

In this section, the components we used for feature extraction are described.

4.2.1 Quality assessment

We classify the images whether they have sufficient quality for a reliable decision with a supervised classifier with the box count values of the detected vessel system as features.

4.2.2 Vessel extraction

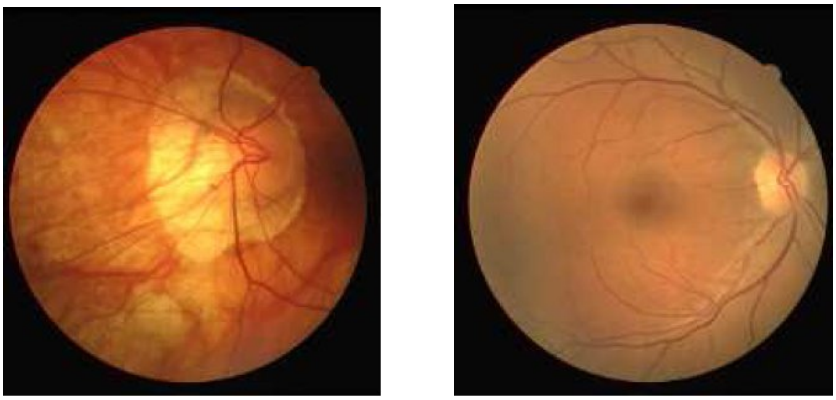
The detection of the vessels helps the localization of other anatomical parts and lesions besides the vascular disorders. In [90], an approach based on Hidden Markov Random Fields (HMRF) for the segmentation of the vascular system in retina images is proposed. The authors extend the optimization problem of HMRF models considering the tangent vector field of the image to enhance the connectivity of the vascular system consisting of elongated structures.

4.2.3 Pre-screening

As a novel contribution, we classify the images as severely diseased (abnormal) or to be forwarded for further processing. The aim of this step is twofold. On the one hand, we minimize the risk that an abnormal image passes the screening without a warning, since it is immediately spotted by the automatic system before detailed analysis. On the other hand, we save computational time, since only the not obviously abnormal fundus images are analysed in detail. Figure 30 gives an impression about these two classes. At the analysis of fundus images, machine learning algorithms are often applied to classification based on feature vectors consisting of intensity values of the image in other fields, see e.g. [91] for HIV or [92] for glaucoma detection. Thus, we consider implementing these approaches for DR screening, as well. We also improved the techniques with feature extraction based on the inhomogeneity characteristics of the diseased retina supported by clinical observations. Our algorithms are trained and tested on images from publicly available, as well as, on our own databases. The approach proposed in this section is also published in [32], [33], [34].

As first step of our approach, we check whether the fundus represented on the image has so severe abnormalities (e.g. large haemorrhages, retinal detachment) that the patient should be sent directly to a clinical expert. In the case of high-loaded automatic systems, skipping these images will enhance the performance, since detailed analyses should not take place. Pre-screening is realized with the application of machine learning algorithms. Next, we summarize the components of pre-screening by consequent steps.

4.2.3.1 Preprocessing As a preprocessing step, we convert the input RGB images to grayscale ones to get a suitable representation for possible disorders. Then, we apply adaptive histogram equalization (AHE) as an intensity normalization step proposed in [75] with an



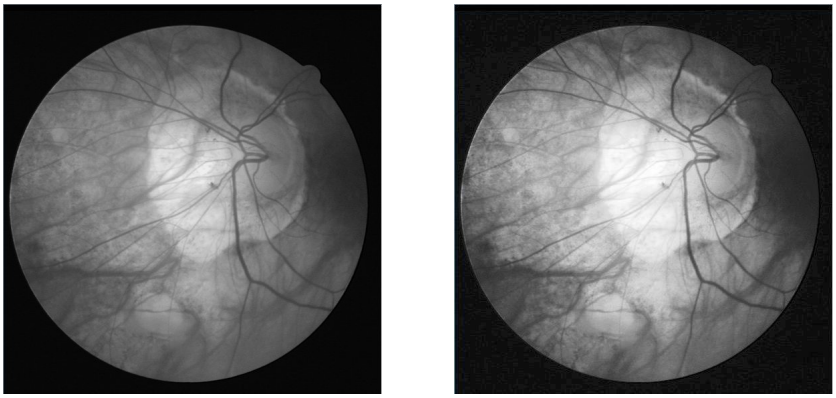
(a) (b)

Figure 30: Samples from the image set (both taken from the DRIVE database [39]); (a) abnormal fundus, (b) not abnormal fundus, the proper grading needs further analysis.

output is shown in Figure 31. Finally, we rescale the images to the size of 90×90 pixels.

4.2.3.2 Feature vectors and classifiers We also take advantage of the clinical observation that fundi with severe diabetic retinopathy often have inhomogeneity caused by retinal pigment epithelium (RPE) atrophy, which is the damage of the pigmented cell layer of the retina [93]. To extract these features, we used the following approaches:

- **Inhomogeneity:** Let the image be split into disjoint subimages of size $Z \times Z$, e.g. with $Z = 5$. Then, for each pixel within a subimage, we compute the sum of intensity differences larger than a given threshold t for every subsequent subimage pixels. If this number is larger than zero, the feature is set to 1, otherwise



(a) (b)

Figure 31: Contrast enhancement of fundus images by adaptive histogram equalization: (a) grayscale image, (b) image preprocessed by AHE. Original image is taken from the DRIVE database [39].

to 0. See Algorithm 2 for the precise formulation. The values of Z and t are determined experimentally and these values are constant across the image.

- **Standard deviation:** For each subimage we calculate the standard deviation of intensity values.
- **Combined:** We calculate both the inhomogeneity and the standard deviation features and combine them.

For the detailed evaluation of this approach, see appendix A.

4.2.4 Microaneurysm detection

A key feature to recognize DR is to detect microaneurysms in the fundus of the eye. The importance of handling MAs are two-fold.

Algorithm 2 Extraction of the inhomogeneity feature.

Input: A digital retinal image I of size $k \times l$.
Input: A subimage size $0 \leq Z \leq min\,(k,\, l)$.
Input: A threshold parameter $t \geq 0$.
Input: An empty array fv of size $(k/Z) \cdot (l/Z)$.
Output: Inhomogeneity values for each subimage stored in fv.

```
for x := 0 to k − Z do
    for y := 0 to l − Z do
        for all x′ ∈ [x, x + Z] and y′ ∈ [y, y + s] do
            sub := |I (x, y) − I (x′, y′)|
            if sub ≥ t then
                diff := diff + sub
            end if
        end for
        if diff > 0 then
            fv [count] := 1
        else
            fv [count] := 0
        end if
        count := count + 1
        x := x + Z
        y := y + Z
    end for
end for
```

First, they are 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 [85]. A more detailed description of our approach to MA detection can be found in section 3.

4.2.5 Exudate detection

Exudates are primary signs of diabetic retinopathy and occur when lipid or fat leaks from blood vessels or aneurysms. Exudates are light, small spots, which can have irregular shape. Since exudate detection is also a challenging task, we follow the same complex methodology as for microaneurysm detection [19]. Thus, we combine preprocessing methods and exudate candidate extractors in the case of exudate detection, too. More details about this approach can be found in [94].

4.2.6 Macula detection

The macula is the central region of sharp vision in the human eye, with its center referred to as the fovea (see Figure 3). Any lesions (e.g. microaneurysms) which appear within the macula can lead to severe loss of vision. Therefore, the efficient detection of the macula is essential in an automatic screening system for diabetic retinopathy. The macula is located roughly in the center of the retina, temporal to the optic nerve. It is a small and highly sensitive part of the retina, which is responsible for detailed central vision. The macula allows us to recognize details and perform tasks that require central vision such as reading.

The approach proposed in this section is also published in [31].

4.2.6.1 A novel algorithm for macula detection In this section, we present a novel approach to detect the macula in a retinal

image. The proposed macula detection algorithm is formalized in algorithm 3.

Algorithm 3 A novel macula detector.

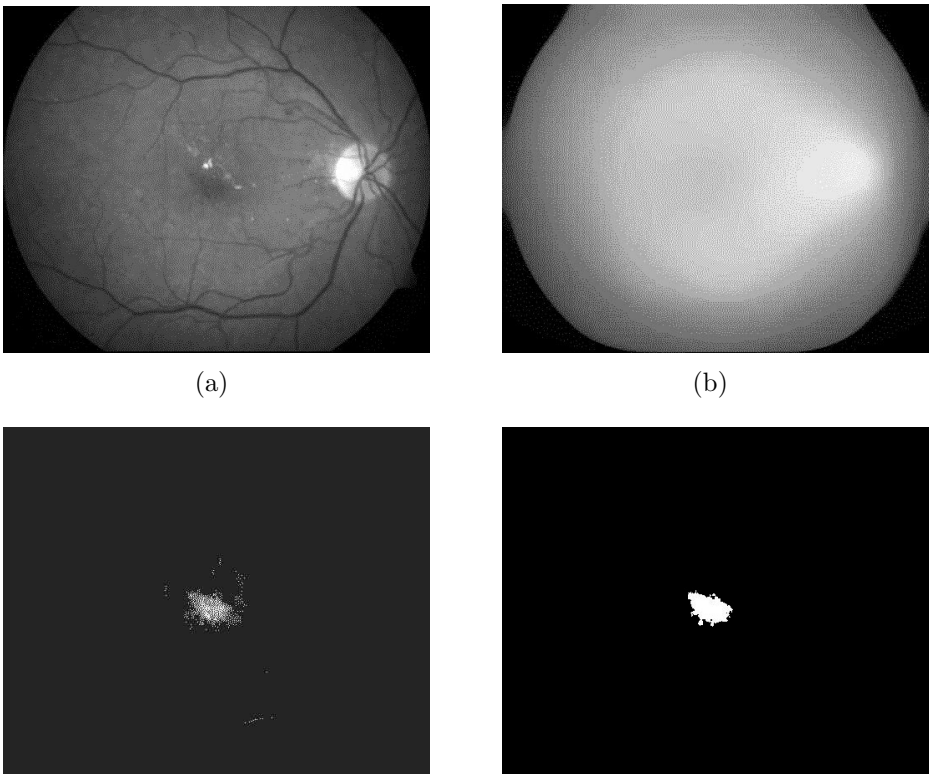
Input: A digital retinal image I of size $k \times l$.
Input: A parameter $0 \leq q \leq 1$ to adjust the mask size in the median filtering step.
Input: A threshold parameter $t \in [-255, \ldots, 255]$.
Output: A binary image containing the macula region of the eye.

1. Let $V = \lceil min\,(k,\, l) \cdot q \rceil$.
2. Produce an image I' with the same size as I by applying median filtering on I with a mask size $V \times V$.
3. Create the difference image $I_{diff} = I - I'$.
4. Produce a binary image I_{bin} by assigning all pixels with larger intensity than t in the I_{diff} to the foreground, while the rest to its background.
5. Select the largest binary component as the macula.

The results after each step of the algorithm can also be observed in Figure 32.For the detailed evaluation of this approach, see appendix B.

4.2.7 Optic disc detection

The optic disc is a circular shaped anatomical structure with a bright appearance. It is the area where the optic nerve enters the eye. If the center and the radius of the optic disc are detected correctly, then they can be used as reference data for approximating other anatomical parts e.g. the macula. Locating these anatomical parts is important from two aspects: the appearance of certain lesions can indicate a more advanced stage of DR and the presence of rare, but serious defects



(a) (b)

(c) (d)

Figure 32: The steps of the proposed macula detection algorithm. (a) The green channel of the input image. (b) The result of the median filtering. (c) The difference image. (d) The binary image after thresholding and largest component selection.

(like retinal detachment) can ruin the detection of the optic disc and macula. More details on this approach can be found in [95].

4.2.8 Multiscale AM/FM based feature extraction

The Amplitude-Modulation Frequency-Modulation (AM/FM) method extracts information from an image, decomposing the green channels of the images into different representations which reflect the intensity, geometry, and texture of the structures with signal processing techniques. The extracted information are then filtered to establish 39 different representations of the image. The images are classified using these features with a supervised learning method. More on this approach can be found in [96].

4.3 Ensemble-based decision making based on detailed retinal image analysis

The most important expectation for a computer-aided medical system is its high reliability. To ensure that, we use ensemble-based decision making [18]. Thus, we have trained several classifiers to separate DR and non-DR cases and fused their results. In this section, we present how we selected the ensemble for DR classification based on the features extracted from the output of the detector presented in section 4.2.

4.3.1 Features

To classify the images we extracted the features summarized inTtable 23.

Features $\chi_8 - \chi_{17}$ are divided with the diameter of the ROI for normalization to compensate different image sizes.

4.3.2 Ensemble selection

To select the optimal ensemble for DR classification, we trained several well-known classification algorithms (Alternating Decision Tree, kNN,

Feature	Description of feature
χ_1	The result of pre-screening (non-severe DR / severe DR).
$\chi_2 - \chi_7$	The number of microaneurysms detected at α confidence (see Definition 5), where $\alpha = 0.5, \ldots, 1.0$.
$\chi_8 - \chi_{16}$	The number of exudate pixels at α confidence, where $\alpha = 0.1, \ldots, 1.0$.
χ_{17}	The euclidean distance of the center of the optic disc and the fovea.
χ_{18}	The result of the AM/FM-based classification (non-DR / DR).

Table 23: Features for final diabetic retinopathy grading.

AdaBoost, Multilayer Perceptron, naive Bayes, Random Forest). Each ensemble is a subset of the classifiers.

4.3.2.1 Classifier output combination The first question is the combination strategy for the output of the classifiers. For this task, we tested the following approaches: majority voting (see Definition 13), weighted majority voting (see Definition 14) and the following algebraic combinations of class membership probabilities: average, product, minimum and maximum (see Definition 15).

4.3.2.2 Ensemble search strategies Several approaches have been tested for selecting the best ensemble for DR grading. The following search strategies were investigated [97]:

- **Forward search:** First, the best individual classifier is selected, which is the starting member of the ensemble. Then, further classifiers are added if the performance of the ensemble increases. The process ends when no further performance increase can be achieved by adding more classifiers.

- **Backward search:** First, all classifiers are members of the ensemble. Then, classifiers are removed from the ensemble until the performance of the ensemble increases.

- **All:** All classifiers are members of the ensemble.

To evaluate the performance of an ensemble in the search process, we compared accuracy (12), the F-score measure (10) and sensitivity (5) as energy functions.

4.3.2.3 Training 10-fold cross-validation were used for both the training phase and for the evaluation of the ensembles. The results given in section 4.4 are average values of the 10-fold cross-validation for the evaluation functions in each case on the Messidor database (section 2.1.5).

4.4 Experiments

We have evaluated the ensemble creation strategies in two scenarios. First, we have investigated whether the image contains early signs of retinopathy (R1) or not (R0) (R0 vs R1 scenario). Second, we measured the classification performance of the ensembles between all diseased categories (R1, R2, R3) and the normal one (R0) (R0 vs all scenario). Tables 24, 25 and 26 contain the sensitivity, specificity and accuracy values for the different strategies and energy functions for the first scenario, while Tables 27, 28 and 29 for the second scenario, respectively. For both scenarios, the cells containing the values for the most accurate ensembles are set in bold.

As it can be seen, the backward search strategy and the average combination strategy provided the best results for each case. However, the choice of the energy function needs to be further investigated, since different energy functions are found to be the best in the two scenarios.

Accuracy energy function			
R0 vs R1	ALL	FORWARD	BACKWARD
$\mathcal{D}_{maj}$	96%/84%/85%	77%/90%/88%	92%/88%/89%
$\mathcal{D}_{wmaj}$	85%/87%/87%	83%/88%/87%	85%/88%/88%
$\mathcal{D}_{avg}$	80%/88%/87%	77%/88%/86%	**94%/90%/90**%
$\mathcal{D}_{pro}$	100%/78%/78%	80%/88%/87%	100%/79%/80%
$\mathcal{D}_{max}$	48%/95%/69%	74%/91%/87%	83%/89%/88%
$\mathcal{D}_{min}$	95%/79%/80%	71%/91%/86%	98%/81%/83%

Table 24: R0 vs R1 final diabetic retinopathy grading results on the Messidor database with accuracy as energy function using different combination and search strategies. Each cell contains the sensitivity, the specificity and the accurracy of the ensembles, respectively.

F-score energy function			
R0 vs R1	ALL	FORWARD	BACKWARD
$\mathcal{D}_{maj}$	96%/84%/85%	75%/89%/86%	84%/89%/88%
$\mathcal{D}_{wmaj}$	85%/87%/87%	87%/87%/87%	69%/88%/83%
$\mathcal{D}_{avg}$	80%/88%/87%	82%/89%/88%	93%/90%/90%
$\mathcal{D}_{pro}$	100%/78%/78%	79%/89%/87%	100%/78%/80%
$\mathcal{D}_{max}$	48%/95%/69%	85%/88%/88%	64%/96%/85%
$\mathcal{D}_{min}$	95%/79%/80%	81%/88%/87%	76%/89%/86%

Table 25: R0 vs R1 final diabetic retinopathy grading results on the Messidor database with as F-score energy function using different combination and search strategies. Each cell contains the sensitivity, the specificity and the accurracy of the ensembles, respectively.

Sensitivity energy function			
R0 vs R1	ALL	FORWARD	BACKWARD
$\mathcal{D}_{maj}$	96%/84%/85%	98%/82%/83%	88%/87%/87%
$\mathcal{D}_{wmaj}$	85%/87%/87%	76%/90%/88%	98%/82%/84%
$\mathcal{D}_{avg}$	80%/88%/87%	86%/88%/88%	85%/89%/88%
$\mathcal{D}_{pro}$	100%/78%/78%	74%/90%/86%	100%/78%/78%
$\mathcal{D}_{max}$	48%/95%/69%	74%/90%/87%	81%/90%/88%
$\mathcal{D}_{min}$	95%/79%/80%	77%/90%/87%	98%/81%/82%

Table 26: R0 vs R1 final diabetic retinopathy grading results on the Messidor database with sensitivity as energy function using different combination and search strategies. Each cell contains the sensitivity, the specificity and the accurracy of the ensembles, respectively.

Accuracy energy function			
R0 vs all	ALL	FORWARD	BACKWARD
$\mathcal{D}_{maj}$	88%/79%/86%	91%/76%/88%	90%/80%/89%
$\mathcal{D}_{wmaj}$	88%/84%/87%	88%/88%/87%	86%/83%/85%
$\mathcal{D}_{avg}$	86%/83%/85%	88%/85%/88%	87%/80%/86%
$\mathcal{D}_{pro}$	95%/38%/60%	85%/83%/85%	88%/85%/88%
$\mathcal{D}_{max}$	80%/95%/80%	88%/82%/87%	81%/97%/82%
$\mathcal{D}_{min}$	92%/50%/72%	90%/76%/87%	93%/77%/89%

Table 27: R0 vs all final diabetic retinopathy grading results on the Messidor database with accuracy as energy function using different combination and search strategies. Each cell contains the sensitivity, the specificity and the accurracy of the ensembles, respectively.

F-score energy function			
R0 vs all	ALL	FORWARD	BACKWARD
$\mathcal{D}_{maj}$	88%/79%/86%	88%/84%/88%	90%/88%/90%
$\mathcal{D}_{wmaj}$	88%/84%/87%	91%/68%/85%	89%/90%/88%
$\mathcal{D}_{avg}$	86%/83%/85%	89%/81%/87%	89%/92%/90%
$\mathcal{D}_{pro}$	95%/38%/60%	89%/72%/85%	90%/73%/86%
$\mathcal{D}_{max}$	80%/95%/80%	87%/78%/86%	81%/98%/83%
$\mathcal{D}_{min}$	92%/50%/72%	88%/76%/86%	89%/83%/88%

Table 28: R0 vs all final diabetic retinopathy grading results on the Messidor database with F-score as energy function using different combination and search strategies. Each cell contains the sensitivity, the specificity and the accurracy of the ensembles, respectively.

Sensitivity energy function			
R0 vs all	ALL	FORWARD	BACKWARD
$\mathcal{D}_{maj}$	88%/79%/86%	88%/77%/86%	89%/78%/86%
$\mathcal{D}_{wmaj}$	88%/84%/87%	88%/78%/85%	88%/93%/89%
$\mathcal{D}_{avg}$	86%/83%/85%	92%/75%/88%	**90%/91%/90**%
$\mathcal{D}_{pro}$	95%/38%/60%	95%/72%/86%	97%/56%/80%
$\mathcal{D}_{max}$	80%/95%/80%	90%/73%/86%	81%/97%/82%
$\mathcal{D}_{min}$	92%/50%/72%	90%/78%/87%	93%/68%/86%

Table 29: R0 vs all final diabetic retinopathy grading results on the Messidor database with sensitivity as energy function using different combination and search strategies. Each cell contains the sensitivity, the specificity and the accurracy of the ensembles, respectively.

Compared to the classification performance of the MA-only early detection framework discussed in section 3, the improvement using the final decision framework is significant: the most accurate result achieved by the MA detector achieved 76% sensitivity, 88% specificity and 82% accuracy (see table 17), opposed to the 90% sensitivity, 91% specificity and 90% accuracy (see table 29) of the latter approach. Thus, the more resource-demanding final decision approach is more reliable in DR grading, as well.

5 Conclusion

In this thesis, ensemble-based approaches to medical decision making are proposed. The use case for these approaches is the detection of diabetic retinopathy (DR) in color fundus images. First, approaches to the early detection of DR are given. Normally, the earliest sign of DR is the presence of microaneurysms (MAs). As one of the major contributions of the thesis, a novel approach for image processing ensembles is presented based on the combination of preprocessing methods and MA candidate extractors. The experimental results show that this approach outperforms the individual MA detectors and is currently the ranked as the first MA detector in a world-wide online competition.

In this second part of the thesis, an ensemble-based approach for the final decision on DR grading is described. The approach is based on an ensemble of machine learning classifiers, which rely on features extracted by several retinal image processing methods. The results show that the use of this approach lead to significant improvement in DR grading compared to the approach based on MA detection only. However, the results achieved for MA detection presented in the first part of the thesis is a critical component of the developed DR grading system. The overall performance of the final DR grading is highly competitive with other state-of-the-art systems proposed ans used for DR grading.

Acknowledgement

Some elements of this thesis supported by the János Bolyai 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 (contract no.: OM-00194/2008, OM-00195/2008, OM-00196/2008). This work is also supported by the TAMOP-4.2.2/B-10/1-2010-0024 project, which is co-financed by the EU and the European Social Fund. We also acknowledge the Moorefields Eye Hospital, London for their clinical support.

I am grateful to my thesis advisor András Hajdu for the encouragement and guidance. I am also thankful for the Doctoral School of Informatics of the University of Debrecen for providing opportunity to conduct research and write the thesis.

I am honoured to work with of members of the DRSCREEN project, especially the co-authors of my papers. This thesis could not be written without the technical support of my friend, István Lázár.

Finally, I would like to thank my family and especially my fiancée, Bea for their support and patience throughout the years.

References

[1] E. S. Berner, ed., *Clinical Decision Support Systems: Theory and practice*. Springer, 2nd ed., 2006.

[2] S. Dreiseitl and M. Binder, "Do physicians value decision support? a look at the effect of decision support systems on physician opinion," *Artifficial Intelligence in Medicine*, vol. 33, pp. 25–30, Jan. 2005.

[3] R. Klein, B. E. K. Klein, and S. E. Moss, "Visual impairment in diabetes," *Ophthalmology*, vol. 91, pp. 1–9, 1984.

[4] "Global Diabetes Plan." `http://www.idf.org/sites/default/files/Global_Diabetes_Plan_Final.pdf` Downloaded on 07/16/2012.

[5] "Causes of vision impairment." `http://www.lighthouse.org/research/statistics-on-vision-impairment/causes/` Downloaded on 07/11/2012.

[6] M. Abramoff, M. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 169 –208, 2010.

[7] A. Gullstrand, "Neue methoden der reflexlosen ophthalmoskopie," *Berichte Deutsche Ophthalmologische Gesellschaft*, vol. 36, pp. 75–80, 1910.

[8] L. Allen, "Ocular fundus photography," *American Journal of Ophthalmolgy*, vol. 57, pp. 13–28, 1964.

[9] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A.

Puliafito, and J. G. Fujimoto, "Optical Coherence Tomography," *Science*, vol. 254, pp. 1178–1181, Nov. 1991.

[10] P. Mackenzie, M. Russell, P. Ma, C. Isbister, and D. Maberley, "Sensitivity and specificity of the Optos Optomap for detecting peripheral retinal lesions," *Retina*, vol. 27, no. 8, pp. 1119–1124, 2007.

[11] "Vanderbilt Eye Institute." `http://www.vanderbilthealth.com/eyeinstitute/23499` Downloaded on 07/11/2012.

[12] "EyePACS." `https://www.eyepacs.org/about.do` Downloaded on 07/11/2012.

[13] "English national screening programme for diabetic retinopathy." `http://www.retinalscreening.nhs.uk/pages/` Downloaded on 07/11/2012.

[14] "EyeCheck." `http://www.eyecheck.nl` Downloaded on 07/11/2012.

[15] S. Philip, A. D. Fleming, K. A. Goatman, S. Fonseca, P. Mcnamee, G. S. Scotland, G. J. Prescott, P. F. Sharp, and J. A. Olson, "The efficacy of automated disease/no disease grading for diabetic retinopathy in a systematic screening programme," *British Journal of Ophthalmology*, vol. 91, no. 11, pp. 1512–1517, 2007.

[16] R. Winder, P. Morrow, I. McRitchie, J. Bailie, and P. Hart, "Algorithms for digital image processing in diabetic retinopathy," *Computerized Medical Imaging and Graphics*, vol. 33, no. 8, pp. 608 – 622, 2009.

[17] 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 and Eye Research*, vol. 25, no. 1, pp. 99 – 127, 2006.

[18] L. I. Kuncheva, *Combining Pattern Classifiers. Methods and Algorithms*. Wiley, 2004.

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

[20] B. Antal and A. Hajdu, "An ensemble-based system for microaneurysm detection and diabetic retinopathy grading," *IEEE Transactions on Biomedical Engineering*, vol. 59, pp. 1720 – 1726, 2012.

[21] B. Antal, I. Lázár, A. Hajdu, Z. Török, A. Csutak, and T. Pető, "Evaluation of the grading performance of an ensemble-based microaneurysm detector," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5943–5946, 2011.

[22] B. Antal and A. Hajdu, "An ensemble-based microaneurysm detector for retinal images," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 1621–1624, 2011.

[23] B. Antal, I. Lázár, A. Hajdu, Z. Török, A. Csutak, and T. Pető, "A multi-level ensemble-based system for detecting microaneurysms in fundus images," in *Proceedings of the IEEE International Workshop on Soft Computing Applications*, pp. 137–142., 2010.

[24] B. Antal, I. Lázár, and A. Hajdu, "Novel approaches to improve microaneurysm detection in retinal images," in *Proceedings of the*

International Conference on Applied Informatics, pp. 149–156, 2010.

[25] B. Antal, I. Lázár, and A. Hajdu, "An optimal voting scheme for microaneurysm candidate extractors using simulated annealing," in *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pp. 80–87, 2010.

[26] B. Antal, I. Lázár, and A. Hajdu, *An Ensemble Approach to Improve Microaneurysm Candidate Extraction*, vol. 222 of *Communications in Computer and Information Science*, ch. Signal Processing and Multimedia Applications, pp. 378–394. Springer Verlag, 2012.

[27] B. Antal and A. Hajdu, "Improving microaneurysm detection in color fundus images by using an optimal combination of preprocessing methods and candidate extractors," in *Proceedings of the European Signal Processing Conference*, pp. 1224–1228, 2010.

[28] B. Antal, "Context-dependent selection of preprocessing methods for microaneurysm candidate extractors." under prepration.

[29] B. Antal, I. Lázár, and A. Hajdu, "An adaptive weighting approach for ensemble-based detection of microaneurysms in color fundus images," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012. To appear.

[30] B. Antal et al., "An ensemble-based grading method for diabetic retinopathy based on detailed retinal image analysis." under prepration. A Hungarian patent (ID 011593) has been requested.

[31] B. Antal and A. Hajdu, "A stochastic approach to improve macula detection in retinal images," *Acta Cybernetica*, vol. 20, pp. 5–15, 2011.

[32] B. Antal and A. Hajdu, "A prefiltering approach for an automatic screening system," in *Proceedings of the IEEE International Symposium on Intelligent Signal Processing*, pp. 265–268, 2009.

[33] B. Antal, A. Csutak, T. Pető, and A. Hajdu, "A two-phase prefiltering approach to the automatic screening of digital fundus images," in *Proceedings of the 5th International Conference on Signal Processing and Multimedia Applications*, pp. 155–158, 2010.

[34] B. Antal, A. Hajdu, Z. Szabó-Maros, Z. Török, A. Csutak, and T. Pető, "A two-phase decision support framework for the automatic screening of digital fundus images," *Journal of Computational Science*. To appear.

[35] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kalviainen, and J. Pietila, "Diaretdb1 diabetic retinopathy database and evaluation protocol," in *Proceedings of the IEEE Conference on Medical Image Understanding and Analysis*, pp. 61–65, 2007.

[36] L. Gagnon, M. Lalonde, M. Beaulieu, and M.-C. Boucher, "Procedure to detect anatomical structures in optical fundus images," in *Proceedings of the SPIE Medical Imaging*, vol. 4322, pp. 1218–1225, 2001.

[37] 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 Transactions on Medical Imaging*, vol. 29, no. 1, pp. 185–195, 2010.

[38] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, H. Uusitalo, H. Kalviainen, and J. Pietila, "Diaretdb0: Evaluation database and methodology for diabetic retinopathy algorithms," tech. rep., University of Kuopio, Finland, 2006.

[39] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, and B. van Ginneken, "Ridge based vessel segmentation in color images of the retina," *IEEE Transactions on Medical Imaging*, vol. 23, pp. 501–509, 2004.

[40] "National screening programme for diabetic retinopathy." `http://www.retinalscreening.nhs.uk/` Downloaded on 07/11/2012.

[41] B. Lay, *Analyse automatique des images angiofluorographiques au cours de la retinopathie diabetique*. PhD thesis, Centre of Mathematical Morphology, Paris School of Mines, 1983.

[42] C. E. Baudoin, B. J. Lay, and J. C. Klein, "Automatic detection of microaneurysms in diabetic fluorescein angiographies," *Revue D'Epidemiologie et de Sante Publique*, vol. 32, pp. 254–261, 1984.

[43] G. E. Oien and P. Osnes, "Diabetic retinopathy: Automatic detection of early symptoms from retinal images," in *Proceedings of the Norwegian Signal Processing Symposium*, 1995.

[44] 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," *Computers and Biomedical Research*, vol. 29, pp. 284–302, 1996.

[45] A. J. Frame, P. E. Undrill, M. J. Cree, J. A. Olson, K. C. McHardy, P. F. Sharp, and J. Forrester, "A comparison of computer based classification methods applied to the detection of

microaneurysms in ophthalmic fluorescein angiograms," *Computers in Biology and Medicine*, vol. 28, pp. 225–238, 1998.

[46] A. Mendonca, A. Campilho, and J. Nunes, "Automatic segmentation of microaneurysms in retinal angiograms of diabetic patients," in *Proceedings of the International Conference on Image Analysis and Processing*, pp. 728 –733, 1999.

[47] J. H. Hipwell, F. Strachan, J. A. Olson, K. C. McHardy, P. F. Sharp, and J. V. Forrester, "Automated detection of microaneurysms in digital red-free photographs: a diabetic retinopathy screening tool," *Diabetic Medicine*, vol. 17, pp. 588–594, 2000.

[48] G. Yang, L. Gagnon, S. Wang, and M.-C. Boucher, "Algorithm for detecting micro-aneurysms in low resolution color retinal images," in *Proceedings of the Vision Interface*, pp. 265–271, 2001.

[49] M. J. Cree, J. A. Olson, K. C. McHardy, P. F. Sharp, and J. V. Forrester, "A fully automated comparative microaneurysm digital detection system," *Eye*, vol. 11, pp. 622–628, 1997.

[50] L. Streeter and M. J. Cree, "Microaneurysm detection in colour fundus images," in *Proceedings of the Image and Vision Computing New Zealand*, 2003.

[51] A. D. Fleming, S. Philip, and K. A. Goatman, "Automated microaneurysm detection using local contrast normalization and local vessel detection," *IEEE Transactions on Medical Imaging*, vol. 25(9), pp. 1223–1232, 2006.

[52] M. Niemeijer, J. Staal, M. D. Abramoff, M. A. Suttorp-Schulten, and B. van Ginneken, "Automatic detection of red lesions in

digital color fundus photographs," *IEEE Transactions on Medical Imaging*, vol. 24, pp. 584–592, 2005.

[53] T. Walter, P. Massin, A. Arginay, R. Ordonez, C. Jeulin, and J. C. Klein, "Automatic detection of microaneurysms in color fundus images," *Medical Image Analysis*, vol. 11, pp. 555–566, 2007.

[54] S. Ravishankar, A. Jain, and A. Mittal, "Automated feature extraction for early detection of diabetic retinopathy in fundus images," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 210–217, 2009.

[55] J. Nayak, P. S. Bhat, R. Acharya U, C. M. Lim, and M. Kagathi, "Automated identification of diabetic retinopathy stages using digital fundus images," *Journal of Medical Systems*, vol. 32, pp. 107–115, 2008.

[56] M. Langroudi and H. Sadjedi, "A new method for automatic detection and diagnosis of retinopathy diseases in colour fundus images based on morphology," in *Proceedings of the International Conference on Bioinformatics and Biomedical Technology*, pp. 134 –138, 2010.

[57] G. Kande, T. Savithri, and P. Subbaiah, "Automatic detection of microaneurysms and hemorrhages in digital fundus images," *Journal of Digital Imaging*, vol. 23, no. 4, pp. 430–437, 2010.

[58] C. Marino, E. Ares, M. Penedo, M. Ortega, N. Barreira, and F. Gomez-Ulla, "Automated three stage red lesions detection in digital color fundus images," *WSEAS Transactions on Computers*, vol. 7, no. 4, pp. 207–215, 2008.

[59] A. Bhalerao, A. Patanaik, S. Anand, and P. Saravanan, "Robust detection of microaneurysms for sight threatening retinopathy screening," in *Proceedings of the Indian Conference on Computer Vision, Graphics, and Image Processing*, pp. 520 –527, 2008.

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

[61] G. Quellec, M. Lamard, P. Josselin, G. Cazuguel, B. Cochener, and C. Roux, "Optimal wavelet transform for the detection of microaneurysms in retina photographs," *IEEE Transactions on Medical Imaging*, vol. 27, no. 9, pp. 1230 –1241, 2008.

[62] C. E. Hann, J. A. Revie, D. Hewett, J. G. Chase, and G. M. Shaw, "Screening for diabetic retinopathy using computer vision and physiological markers," *Journal of Diabetes Science and Technology*, vol. 3, pp. 819–834, 2009.

[63] A. Mizutani, C. Muramatsua, Y. Hatanakab, S. Suemoria, T. Haraa, and H. Fujita, "Automated microaneurysm detection method based on double-ring filter in retinal fundus images," in *Proceedings of the SPIE Medical Imaging*, vol. 7260, pp. 1N1 – 1N8, 2009.

[64] S. Abdelazeem, "Microaneurysm detection using vessels removal and circular hough transform," in *Proceedings of the National Radio Science Conference*, pp. 421 – 426, 2002.

[65] L. Giancardo, F. Meriaudeau, T. P. Karnowski, K. W. Tobin, Y. Li, and E. Chaum, "Microaneurysms detection with the radon cliff operator in retinal fundus images," in *Proceedings of the SPIE Medical Imaging*, 2010.

[66] E. Grisan and A. Ruggeri, "Segmentation of candidate dark lesions in fundus images based on local thresholding and pixel density," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6735 –6738, 2007.

[67] S. Balasubramanian, S. Pradhan, and V. Chandrasekaran, "Red lesions detection in digital fundus images," in *Proceedings of the International Conference on Image Processing*, pp. 2932 –2935, 2008.

[68] C. Sinthanayothin, J. F. Boyce, T. H. Williamson, H. L. Cook, E. Mensah, S. Lal, and D. Usher, "Automated detection of diabetic retinopathy on digital fundus images," *Diabetic Medicine*, vol. 19, pp. 105–112, 2002.

[69] D. Usher, M. Dumskyj, M. Himaga, T. H. Williamson, S. Nussey, and J. Boyce, "Automated detection of diabetic retinopathy in digital retinal images: a tool for diabetic retinopathy screening," *Diabetic Medicine*, vol. 21, pp. 84–90, 2004.

[70] G. G. Gardner, D. Keating, T. H. Williamson, and A. T. Elliott, "Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool," *British Journal of Ophthalmology*, vol. 80, no. 11, pp. 940–944, 1996.

[71] I. Lázár and A. Hajdu, "Microaneurysm detection in retinal images using a rotating cross-section based model," in *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pp. 1405 –1409, 2011.

[72] M. Niemeijer, M. Loog, M. D. Abramoff, M. A. Viergever, M. Prokop, and B. van Ginneken, "On combining computer-

aided detection systems," *IEEE Transactions on Medical Imaging*, vol. 30, pp. 215 – 223, 2011.

[73] C. V. Rijsbergen, *Information Retrieval*. London: Butterworths, 2nd ed., 1979.

[74] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems magazine*, vol. Third Quarter, pp. 21–45, 2006.

[75] 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 *Proceedings of the Cairo International Biomedical Engineering Conference*, 2006.

[76] T. Walter and J. Klein, "Automatic detection of microaneyrysms in color fundus images of the human retina by means of the bounding box closing," *Lecture Notes in Computer Science*, vol. 2526, pp. 210–220, 2002.

[77] K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Graphics gems*, vol. IV, pp. 474–485, 1994.

[78] A. Hoover and M. Goldbaum, "Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 951 –958, 2003.

[79] R. C. Gonzalez, R. E. Woods, and E. S. L., *Digital Image Processing Using MATLAB*. Gatesmark Publishing, 2009.

[80] A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting," in *Proceedings of the Computer Vision and Pattern Recognition*, vol. 2, pp. II–721 – II–728, 2003.

[81] T. Lin and Y. Zheng, "Adaptive image enhancement for retinal blood vessel segmentation," *Electronics Letters*, vol. 38, pp. 1090–1091, 2002.

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

[83] 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.

[84] J. Eng, "ROC analysis: web-based calculator for ROC curves.." `http://www.jrocfit.org` Downloaded on 07/11/2012.

[85] 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.

[86] "Retinal photography screening for diabetic eye disease," tech. rep., British Diabetic Association, 1997.

[87] 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," *Investigative Ophthalmology & Visual Science*, vol. 52, no. 8, pp. 5862–5871, 2011.

[88] 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.," *Clinical and Experimental Optometry*, vol. 89, no. 5, pp. 299–305, 2006.

[89] 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," *British Journal of Ophthalmology*, vol. 94, no. 12, pp. 1606–1610, 2010.

[90] G. Kovács and A. Hajdu, "Extraction of vascular system in retina images using averaged one-dependence estimators and orientation estimation in hidden markov random fields," in *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pp. 693 –696, 2011.

[91] I. Kozak, P. Sample, J. Hao, W. R. Freeman, R. N. Weinreb, T. Lee, and M. H. Goldbaum, "Machine learning classifiers detect subtle field defects in eyes of HIV inviduals," *Transactions of the American Ophthalmological Society*, vol. 105, pp. 111–120, 2007.

[92] J. Meier, R. Bock, G. Michelson, L. Nyúl, and J. Hornegger, "Effects of preprocessing eye fundus images on appearance based glaucoma classification," *Lecture Notes in Computer Science*, vol. 4673/2007, pp. 165–172, 2007.

[93] C. Bellman, M. M. Neveu, H. P. N. Scholl, C. R. Hogg, P. R. Rath, S. Jenkins, A. C. Bird, and G. E. Holder, "Localized retinal electrophysiological and fundus autofluorescence imaging abnormalities in maternal inherited diabetes and deafness," *Investigative Ophthalmology & Visual Science*, vol. 45, pp. 2355–2360, 2004.

[94] B. Nagy, B. Harangi, B. Antal, and A. Hajdu, "Ensemble-based exudate detection in color fundus images," in *Proceedings of the International Symposium on Image and Signal Processing and Analysis*, pp. 700–703., 2011.

[95] R. J. Qureshi, L. Kovács, B. Harangi, B. Nagy, T. Pető, and A. Hajdu, "Combining algorithms for automatic detection of optic disc and macula in fundus images," *Computer Vision and Image Understanding*, vol. 116, pp. 138–145, 2012.

[96] C. Agurto, V. Murray, E. Barriga, S. Murillo, M. Pattichis, H. Davis, S. Russell, M. Abramoff, and P. Soliz, "Multiscale AM-FM methods for diabetic retinopathy lesion detection," *IEEE Transactions on Medical Imaging*, vol. 29, pp. 502 –512, feb. 2010.

[97] D. Ruta and B. Gabrys, "Classifier selection for majority voting," *Information Fusion*, vol. 6, no. 1, pp. 63 – 81, 2005.

[98] S. Harding, R. Greenwood, S. Aldington, J. Gibson, D. Owens, R. Taylor, E. Kohner, P. Scanlon, and G. Leese, "Grading and disease management in national screening for diabetic retinopathy in England and Wales," *Diabetic Medicine*, vol. 20, pp. 965–971, 2003.

[99] T. S. Petsatodis, A. Diamantis, and G. P. Syrcos, "A complete algorithm for automatic human recognition based on retina vascular network characteristics," *Proceedings of the Era1 International Scientific Conference*, 2006.

[100] S. Sekhar, W. Al-Nuaimy, and A. K. Nandi, "Automated localization of optic disc and fovea in retinal fundus images," *Proceedings of the European Signal Processing Conference*, 2008.

[101] A. D. Fleming, S. Philip, K. A. Goatman, J. A. Olson, and P. Sharp, "Automated assessment of diabetic retinal image quality based on clarity and field definition," *Investigative Ophthalmology and Visual Science*, vol. 47, pp. 1120–1125, 2006.

[102] F. Zana, I. Meunier, and J. C. Klein, "A region merging algorithm using mathematical morphology: application to macula detection," in *Proceedings of the International Symposium on Mathematical morphology and its applications to image and signal processing*, pp. 423–430, 1998.

Appendix

A Pre-screening results

Our first experimental dataset consisted of 34 training and 28 test images. Both the training and the test sets contain 50-50% normal and abnormal cases. Ophthalmologists selected and classified these images whether they contain serious disorder or not from three databases: the publicly available DRIVE (section 2.1.6), DIARETDB1 (section 2.1.3) and the database provided by the Moorfields Eye Hospital, London, UK (section 2.1.4). Our goal is to find images, where the fundus is abnormal to avoid obviously diseased cases to pass. These images contain sight-threatening disorders and have a grade of R3 in a usual retinopathy grading protocol [98]. Therefore, we label the elements of the test database as images with serious disorder (first class) and images to be processed further (second class). Thus, the second class expected to contain normal or not seriously diseased cases.

For pre-filtering, we used a naive Bayes classifier and trained for the combined features extracted from all regions of the images as disclosed in section 4.2.3.2. Thus, a $2 \cdot (k/Z) \cdot (l/Z)$ feature vector is extracted for each image. With this approach, we have successfully classified all elements of the test dataset. That is, the accuracy in this case is 100%.

To make the approach faster, we used backward elimination [91] for feature subset selection. That is, we have selected the best 11 regions on each image to extract the features from them for classification. In this case, our approach still provided no false predictions with an elapsed time below milliseconds.

We have also tested our approach on the Messidor database (section 2.1.5). This database is dedicated to measure the performance of screening systems by providing grading scores for each image. The grades are from R0 to R3, where R0 represents no retinopathy and

R3 is the most serious case of this disease based on the type and the number of the lesions appearing in the images. We have selected the two classes as the follows: abnormal (R3) and images that need further analysis (R0, R1, R2). Our approach achieved an accuracy of 82% with 81% sensitivity and 82% specificity on this dataset. Most of the error originates from the false classification of R2 cases (52% of all misclassified images are from this class), while a smaller portion of R1 and R3 images also classified wrongly (25% and 23% of all misclassification occurred in these images, respectively).

B Macula detection results

We have tested our approach on 199 images from three publicly available data sources: DiaretDB0 (section 2.1.2), DiaretDB1 (section 2.1.3) and DRIVE (section 2.1.6). To compare our detector to state-of-the-art macula detectors, we also evaluated the following algorithms on the same data: Petsatodis et al. [99], Sekhar et al. [100], Fleming et al. [101], and Zana et al. [102].

We have evaluated our approach from two aspects [95]: whether the detected macula center falls into the 0.5DD (Optic Disc Diameter) distance of the manually selected macula center and we also measured the Euclidean distance of them (calculated on normalized images). Table 30 and 31 contain the quantitative results using these measures, respectively. We disclose the results for each macula detector evaluated in all dataset. For the more straightforward comparison, we also calculated the simple average of these performance values. In the terms of the first measure, the use of the proposed algorithm on the novel macula detector resulted in a 85% average accuracy, while the second best method only earned 77%. However, regarding the Euclidean error it is only third in the comparison, mainly because of its difficulties on the DRIVE database.

Dataset	Petsatodis	Sekhar	Fleming	Zana	Proposed
DiaretDB0	68%	72%	85%	63%	86%
DiaretDB1	62%	76%	79%	71%	92%
DRIVE	66%	76%	53%	82%	68%
Average	66%	74%	77%	69%	85%

Table 30: Percentage of detected macula centers falling in the correct region.

Dataset	Petsatodis	Sekhar	Fleming	Zana	Proposed
DiaretDB0	26.59	26.85	37.82	24.11	24.02
DiaretDB1	26.32	27.45	35.67	24.77	25.72
DRIVE	18.15	26.20	37.29	20.85	30.25
Average	23.69	26.83	36.92	23.24	26.75

Table 31: Average euclidean distance of the detected macula centers from the manually selected ones.

C Summary

In this PhD thesis, two approaches are shown to support medical decision-making for diabetic retinopathy (DR). This disease is one of the most common causes of blindness in the developed countries. Thus, timely and precise detection is essential for a large population. Furthermore, high reliability of the diagnosis is also desired. The approaches presented in this thesis are based on the analysis of retinal images. To ensure reliability, we proposed ensemble-based approaches.

First, an approach for the early detection of DR is described. The key to early detection is the timely recognition of a lesion called microaneurysm (MA). Since MA detectors provides the spatial locations of MA candidates as output, the application of standard ensemble-based strategies does not provide appropriate solution. Thus, a spatial combination method is developed, whose details are presented in this thesis. This approach is based on the novel concept of ⟨preprocessing method, candidate extractor⟩ ensembles, which is shown to be effective in improving the sensitivity of MA candidate extraction and with the use of spatial voting, the false detections are also suppressed. For the creation of ⟨preprocessing method, candidate extractor⟩ ensembles, two approaches are introduced. First, a search-based selection approach is presented based on the performance of the ensembles on a training set.

Furthermore, we investigate the effect of using context-dependent information and use this information to assign weights to the participating ⟨preprocessing method, candidate extractor⟩ pairs in the ensembles. Experimental results show superiority over individual approaches for both cases. Especially, our approach is currently ranked as first in a world-wide online competition dedicated to the comparison of the performance of MA detectors.

In addition, an ensemble-based approach to the grading of DR is proposed in this thesis. This approach is based on the outputs of several retinal image processing algorithms, such as lesion detection (microaneurysms, exudates), anatomical parts (macula, optic disc, vascular system), image features (diameter of the region-of-interest) and global DR descriptors (AM/FM filtering, quality assessment). From the output of these methods, certain features are extracted and an ensemble of classifiers is trained. Reasurring results are obtained using this technique for DR grading with highly competitive other state-of-the-art systems. Compared to the classification performance of the MA-only early detection framework, the improvement using the final decision framework is significant. As a closing remark, we conclude that the more resource-demanding final decision approach is also more reliable in DR grading.

D Összefoglaló

Ezen PhD disszertációban két módszert mutattunk be a diabéteszes retinopátiával (DR) kapcsolatos automatikus klinikai döntéshozatal támogatásához. Ez a betegség a vakság kialakulásának egyik legjelentősebb oka a fejlett világban, így a DR gyors és precíz felismerése nagy populációt érintő feladat. A diagnózissal szemben további elvárás a nagy megbízhatóság. A disszertációban található módszerek retinaképek feldolgozásán alapulnak. A megbízhatósági igény teljesítésére a bemutatott döntéstámogató algoritmusok együttes döntéshozást végeznek.

A dolgozatban elsőként egy, a DR korai detektálását végző módszert mutatunk be. A korai detektálásban kulcsfontosságú a mikroaneurizmák kialakulásának időben történő felismerése. Mivel a mikroaneurizma-detektorok az általuk mikroaneurizmaként beazonosított objektumok képen való elhelyezkedésének koordinátáit adják meg, a hagyományos együttes döntéshozáson alapuló algoritmusok alkalmazása nem vezet megfelelő eredményre. A probléma megoldására egy új, képtérbeli kombináló módszert hozunk létre. A módszer az ⟨előfeldolgozó, jelöltállító⟩ párokból álló összetett rendszerek fogalmát vezeti be. Amint azt a dologzatban megmutatjuk, ez a mikroanurizma-jelöltállítók szenzitivitásának növeléséhez biztosít alapot, és egy képtérbeli szavazást végrehajtva a hamis detektálások számát is lecsökkenthetjük. Az ⟨előfeldolgozó, jelöltállító⟩ összetett rendszerek létrehozására két alapvető módszert mutatunk be: először egy tanulási halmazon végzett kiértékelésen alapuló keresési módszert, majd egy környzeti információt figyelembe vevő súlyozási technikát ismertetünk. Az eredmények mindkét esetben azt mutatják, hogy az egyéni módszereket felülmúló rendszereket hozhatunk létre a fenti technikákkal. Ezt bizonyítja például, hogy az itt bemutatott mikroaneurizma-detektor jelenleg az első helyen áll egy, a mikroaneurizma-detektorok teljesítmé-

nyét összehasonlító nemzetközi online versenyen.

A dolgozat másik fontos új eredménye egy összetett rendszer javaslata a retinopátia szűrésére. A módszer összetett tanuláson alapul, azaz a döntéshozás több gépi tanulási osztályozó algoritmus kimenetének felhasználásával történik. Az osztályozókhoz a retinakép feldolgozásával nyerhetünk ki jellemzőket. A jellemzőket a különböző retinakép-feldolgozó algoritmusok kimenetéből nyerjük ki, többek között elváltozások (mikroaneurizmák, exudátumok), anatómiai részek (sárgafolt, vakfolt, érhálózat), képjellemzők (a hasznos terület átmérője) és globális leírók (AM/FM szűrők kimenete, minőségellenőrzés). A módszer segítségével a DR-osztályozásában sikerült biztató eredményeket elérni: A javasolt rendszer versenyképes az aktuális nemzetközi trendet is figyelembe véve. A pusztán MA-detektáláson alapuló DR-osztályozással összehasonlítva arra a megállapításra juthatunk, hogy a részletes elemzést végrehajtó osztályozó rendszer nagyobb erőforrásigénye kifizetődő a pontosság növekedése miatt.

E List of publications

Book chapters

B. Antal, I. Lázár, and A. Hajdu, *An Ensemble Approach to Improve Microaneurysm Candidate Extraction*, vol. 222 of *Communications in Computer and Information Science*, ch. Signal Processing and Multimedia Applications, pp. 378–394. Springer Verlag, 2012. Indexed in Scopus, Mathematical Reviews, Zentralblatt MATH, DBLP.

Journal papers

B. Antal and A. Hajdu, "An ensemble-based system for microaneurysm detection and diabetic retinopathy grading," *IEEE Transactions on Biomedical Engineering*, vol. 59, pp. 1720 – 1726, 2012. IF = 2.278. Indexed in SCI, PubMed, IEEEXplore, MedLine, DBLP.

B. Antal, A. Hajdu, Z. Szabó-Maros, Z. Török, A. Csutak, and T. Pető, "A two-phase decision support framework for the automatic screening of digital fundus images," *Journal of Computational Science*. To appear. Indexed in Scopus.

B. Antal and A. Hajdu, "Improving microaneurysm detection using an optimally selected subset of candidate extractors and preprocessing methods," *Pattern Recognition*, vol. 45, no. 1, pp. 264 – 270, 2012. IF = 2.607. Indexed in SCI, Scopus, ACM Guide to Computing Literature, INSPEC, Mathematical Reviews, Zentralblatt MATH, DBLP.

B. Antal and A. Hajdu, "A stochastic approach to improve macula detection in retinal images," *Acta Cybernetica*, vol. 20, pp. 5–15, 2011. Indexed in Scopus, Mathematical Reviews, Zentralblatt MATH, DBLP.

B. Antal, "Context-dependent selection of preprocessing methods for microaneurysm candidate extractors." under prepration.

B. Antal et al., "An ensemble-based grading method for diabetic retinopathy based on detailed retinal image analysis." under prepration. A Hungarian patent (ID 011593) has been requested.

Conference proceedings

B. Antal, I. Lázár, and A. Hajdu, "An adaptive weighting approach for ensemble-based detection of microaneurysms in color fundus images," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012. To appear. Indexed in PubMed, IEEEXplore, MedLine.

B. Antal, I. Lázár, A. Hajdu, Z. Török, A. Csutak, and T. Pető, "Evaluation of the grading performance of an ensemble-based microaneurysm detector," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5943–5946, 2011. Indexed in PubMed, IEEEXplore, MedLine.

B. Antal and A. Hajdu, "An ensemble-based microaneurysm detector for retinal images," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 1621–1624, 2011. Indexed in IEEEXplore, DBLP.

B. Antal, I. Lázár, A. Hajdu, Z. Török, A. Csutak, and T. Pető, "A multi-level ensemble-based system for detecting microaneurysms in fundus images," in *Proceedings of the IEEE International Workshop on Soft Computing Applications*, pp. 137–142., 2010. Indexed in IEEEXplore.

B. Antal and A. Hajdu, "Improving microaneurysm detection in color fundus images by using an optimal combination of preprocessing

methods and candidate extractors," in *Proceedings of the European Signal Processing Conference*, pp. 1224–1228, 2010.

B. Antal, I. Lázár, and A. Hajdu, "An optimal voting scheme for microaneurysm candidate extractors using simulated annealing," in *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pp. 80–87, 2010. Indexed in Scopus, IEEEXplore, Zentralblatt MATH, DBLP.

B. Antal, A. Csutak, T. Pető, and A. Hajdu, "A two-phase pre-filtering approach to the automatic screening of digital fundus images," in *Proceedings of the 5th International Conference on Signal Processing and Multimedia Applications*, pp. 155–158, 2010. Indexed in Scopus, IEEEXplore, Zentralblatt MATH, DBLP.

B. Antal, I. Lázár, and A. Hajdu, "Novel approaches to improve microaneurysm detection in retinal images," in *Proceedings of the International Conference on Applied Informatics*, pp. 149–156, 2010.

B. Antal and A. Hajdu, "A prefiltering approach for an automatic screening system," in *Proceedings of the IEEE International Symposium on Intelligent Signal Processing*, pp. 265–268, 2009. Indexed in IEEEXplore.

B. Antal, I. Lázár, and A. Hajdu, "Mikroaneurizma-detektálás összetett rendszerrel," in *Proceedings of the Képfeldolgozók és Alakfelismerők Országos Konferenciája*, 2011. In Hungarian.

Other conference proceedings (not part of the thesis)

B. Nagy, B. Antal, and A. Hajdu, "Image database clustering to improve microaneurysm detection in color fundus images," in *Proceedings of the IEEE International Symposium on Computer-Based Medical Systems*, 2012. To appear.

B. Harangi, B. Antal, and A. Hajdu, "Automatic exudate detection with improved naive-bayes classifier," in *Proceedings of the IEEE International Symposium on Computer-Based Medical Systems*, 2012. To appear.

B. Nagy, B. Harangi, B. Antal, and A. Hajdu, "Ensemble-based exudate detection in color fundus images," in *Proceedings of the International Symposium on Image and Signal Processing and Analysis*, pp. 700–703., 2011.

Technical reports

I. Lázár, B. Antal, and A. Hajdu, "Microaneurysm detection in digital fundus images," Tech. Rep. 2010/14(387), University of Debrecen, Hungary, 2010.

B. Antal, I. Lázár, and A. Hajdu, "An ensemble-based system for microaneurysm detection," tech. rep., Retinopathy Online Challenge, 2010. `http://roc.healthcare.uiowa.edu/results/documentation/drscreen.pdf` Downloaded on 07/11/2012.

Other

B. Antal and A. Hajdu, "Component reuse in ensemble-based medical image processing applications," in *Proceedings of the Symposium on Programming Languages and Software Tools*, 2011. Invited paper of A. Hajdu.