Informatics in Medicine Unlocked 9 (2017) 44–57

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Detection of microaneurysms in retinal images using an ensemble classifier

M.M. Habib [a,*], R.A. Welikala [a], A. Hoppe [a], C.G. Owen [b], A.R. Rudnicka [b], S.A. Barman [a]

[a] School of Computer Science and Mathematics, Faculty of Science, Engineering and Computing, Kingston University, London, UK
[b] Population Health Research Institute, St. George's, University of London, United Kingdom

ARTICLE INFO

Keywords:
Image processing
Medical image analysis
Retinal imaging
Microaneurysm detection
Tree ensemble
Diabetic retinopathy

ABSTRACT

This paper introduces, and reports on the performance of, a novel combination of algorithms for automated microaneurysm (MA) detection in retinal images. The presence of MAs in retinal images is a pathognomonic sign of Diabetic Retinopathy (DR) which is one of the leading causes of blindness amongst the working age population. An extensive survey of the literature is presented and current techniques in the field are summarised. The proposed technique first detects an initial set of candidates using a Gaussian Matched Filter and then classifies this set to reduce the number of false positives. A Tree Ensemble classifier is used with a set of 70 features (the most commons features in the literature). A new set of 32 MA groundtruth images (with a total of 256 labelled MAs) based on images from the MESSIDOR dataset is introduced as a public dataset for benchmarking MA detection algorithms. We evaluate our algorithm on this dataset as well as another public dataset (DIARETDB1 v2.1) and compare it against the best available alternative. Results show that the proposed classifier is superior in terms of eliminating false positive MA detection from the initial set of candidates. The proposed method achieves an ROC score of 0.415 compared to 0.2636 achieved by the best available technique. Furthermore, results show that the classifier model maintains consistent performance across datasets, illustrating the generalisability of the classifier and that overfitting does not occur.

1. Introduction

Retinal Image Analysis (RIA) is an active area of research due to its application in screening programs for Diabetic Retinopathy (DR) – one of the leading causes of blindness in the developed world. During the screening process, fundus images of the retina are captured for the purpose of detection of diabetic retinopathy. The presence of microaneurysms (MAs) in retinal images is an early indicator of DR (Fig. 1). The automated detection of MAs from retinal images can aid in screening programs for DR diagnosis. Several algorithms have been proposed for the detection of MA, however, MA detection is still a challenging problem due to the variance in appearance of MAs in retinal images [1].

Through our review of MA detection in the literature, we have identified three main stages in MA detection algorithms: 1) preprocessing 2) MA candidate detection and 3) candidate classification. *Preprocessing* corrects non-uniform illumination in retinal images and enhances the contrast of MAs in the image. *MA candidate detection* seeks to detect an initial set of candidate regions where MAs are likely to exist. *MA candidate classification* applies machine learning techniques in order to improve the specificity of the algorithm by filtering out false positives from the candidate detection phase. Some of the proposed methods in the literature are unsupervised methods, which means they do not require the third classification stage [1–7]. A summary of MA candidate detection algorithms presented in the literature is listed in Table 1. For each algorithm the table describes image type, the initial candidates method, the classifier used, and the reported performance for each classifier. Most of the literature has differences in the method used to evaluate their algorithms or the dataset used, which makes it difficult to compare any 2 algorithms together. One of the earliest proposed techniques for MA detection was applied to fluorescein angiograms [8]. A Gaussian matched filter was used to detect the initial set of candidates. Finally, each initial candidate was classified as either a true candidate or a spurious one using some features, producing the final classification result. Cree [9] applied Spencer's technique [8] to multiple longitudinal florescence images in order to detect the 'MA turnover' – the appearance or disappearance of MA objects over time.

More recent techniques have tackled the problem of MA detection in colour fundus images. The main reason for this is that colour images are more common in screening programs and are also non-invasive to capture, unlike fluorescein images. The following methods are all based on MA detection in colour fundus images.

A number of techniques have adapted Spencer's approach in terms of

* Corresponding author.
E-mail addresses: m.habib@kingston.ac.uk (M.M. Habib), s.barman@kingston.ac.uk (S.A. Barman).

http://dx.doi.org/10.1016/j.imu.2017.05.006
Received 9 March 2017; Received in revised form 12 May 2017; Accepted 13 May 2017
Available online 17 May 2017
2352-9148/© 2017 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

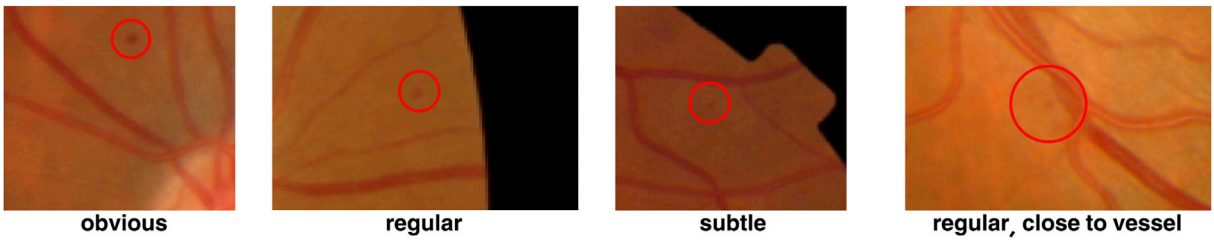

Fig. 1. Examples of various microaneurysms with varying appearances and locations.

applying morphological vessel removal followed by a Gaussian matched filter. Hipwell [10] performed a modification of Cree [9] in order to apply the algorithm to colour fundus images. Streeter [11] used a method based on Cree [9]. However, during the classification phase, 80 features are extracted and Linear Discriminant Analysis (LDA) was used to perform the classification. Feature Selection was performed to filter the features down to 16 features. Feature Selection is a process to reduce redundant features in order to reduce computational time and decrease chances of overfitting. Another Spencer-based approach was introduced in Fleming [12]. This technique introduced a novel region-growing step based on gradient values, rather than a simple threshold. In addition a paraboloid was fitted to each candidate using a parameter optimization process. The paraboloid parameters are used to compute many of the features used in the candidate classification phase. Instead of using a single Gaussian matched filter, Zhang [13] applied multiple Gaussian filters at multiple scales and computed the maximum response to produce a probability map of the likelihood of presence of MA candidates. This probability map was then thresholded to produce the initial set of MA candidates. Finally a rule-based classifier using 30 features was used to perform the final classification. Li [4] used an unsupervised method based on a Multi-orientation Sum of Matched Filter (MSMF). This filter is a modification of the classical Gaussian Matched filter. This modified filter is anisotropic in nature and is applied in multiple directions. Hence, this filter is better at suppressing responses to blood vessels than the Gaussian Matched filter. Wu [14] modified the MSMF filter to take into account the varying size of MAs.

Sánchez used a mixture model-based clustering technique to detect the initial MA candidate regions [7]. The technique fits three normal distribution histograms to the retinal image histogram. These histograms correspond to foreground, background and outliers. The foreground histogram pixels are considered as the initial set of MA candidate regions. Finally, logistic regression was used to classify each MA region as belonging to either a foreground or background region. Quellec [15] based his technique on wavelet transforms applied in different sub-bands of the colour image.

A double-ring filter was used in Mizutani [16] to detect the initial candidates. The filter used the property that MAs are dark circular regions within a brighter region to detect the MA candidates. It consists of an inner ring and an outer ring. A given pixel is considered to be a MA pixel if the average intensity of the inner ring is smaller than the average intensity of the outer ring. After the initial candidates are detected, classification is performed using 12 extracted features and an Artificial Neural Network (ANN).

Initial candidates were detected using a simple thresholding in Giancardo [5,6]. A novel Radon-based transform was used to extract the features of the initial candidates and a Support Vector Machine (SVM) classifier was used to perform the final classification. An initial set of 31 features were computed for classification. The dimensionality of the features was reduced to 10 dimensions using Principle Component Analysis (PCA), and this reduced representation was used to perform the classification. A reduced dimension for the features reduces the risk of overfitting and also makes the classification more computationally efficient.

Sinthanayothin [17] used a 'moat operator' to enhance red lesions in the image and then these regions were segmented. Vessel regions were then removed to produce the final set of candidates. Note that this method detected both MAs and haemorrhages. The moat operator was not defined in the paper and we were unable to find the exact definition in the literature.

AbdelAzeem [18] used a Hessian matrix in order to detect the initial MA candidate set. A rule based classifier is then used to detect false MA detections. The rule is simply based on the candidate 'energy'. The exact definition of the computed 'energy' was not mentioned in the paper, however, it is likely to be the same definition as in Fleming [12]. Inoue [19] relied on a Hessian matrix in order to detect the initial candidates and an Artificial Neural Network (ANN) was used to classify the features. A group of 126 features were fed into the ANN for classification. However this group of features was reduced using Principle Component Analysis (PCA) in order to reduce computational complexity and avoid overfitting. Moreover, Srivastava [20] used the eigenvalues of the hessian matrix in order to detect the initial candidates. Recently, Adal has used a hessian matrix in order to detect the initial set of MA candidates. A combination of SURF, Radon and scale-space features were extracted from the initial candidates. Multiple classifiers (Support Vector Machines, K-Nearest-Neighbours, Naive Bayes and Random Forest) were also experimented with in this technique.

An adaptation of Spencer [8] and Frame [22] is presented in Niemeijer [22]. Two main contributions were added: A pixel based classification system for the initial candidate detection phase and an extended set of features used for pixel classification.

A unique method was introduced in Lazar [1,2] since it is an unsupervised technique that does not require any training or classification steps. Moreover the reported results of this technique are comparable to other supervised methods, which make it a promising method. The essence of this technique is to discriminate between vessels and MAs by using a 1D scanline at different directions for each pixel. While a MA will have local minima in all directions of the rotated scanlines, a vessel will have only one minima corresponding to when the scanline is perpendicular to the vessel. Hence, using this property, a probability map is produced at each pixel and then simple thresholding is applied to produce the final set of candidates.

Garcia [30] compared the accuracy of four neural network variants: Multilayer Perceptron (MP), Radial Basis Function (RBF), Support Vector Machine (SVM) and Majority Voting (MV). The initial candidates were detected using a local thresholding technique based on the mean pixels of the entire image compared to mean intensity in a small window around a pixel. According to their experiments, the RBF was suggested as the preferred classifier among all 4. An interesting approach that relies on visual dictionaries was presented in Rocha [24]. The use of visual dictionaries (bag of words) makes this approach more generalizable since it does not rely on specific features during the classification. Therefore, the same approach can be used to perform detection of lesions other than MAs as well. The disadvantage of this is that it requires a larger training set. Haloi [29] recently applied deep neural networks to detect MAs in colour images. Deep neural networks have gained popularity in the field of computer vision in the recent years since they do not require manual feature engineering (selection of features). Moreover, algorithms based on deep learning have produced results that out-perform other state-of-

Table 1
Summary of MA detection algorithms in the literature. The performance superscripts are defined as follows: [a)] Lesion-based measure [b)] Image-based measure [c)] Pixel-based measure. Key: AUC – Area Under the Curve, FP/image - False positives per image, PPV – Positive Predictive Value.

Paper	Image Type	Initial candidates method	Classifier used	Reported Performance	
				Dataset	Performance
Spencer, 1995 [8]	Florescence	Gaussian Filter	Rule-based	Private dataset (4 images)	Sensitivity[a]: 0.25 FP/image[a]: 1.0
Cree, 1997 [9]	Florescence	Gaussian Filter	Rule-based	Private dataset (20 images)	Sensitivity[a]: 0.6 FP/image[a]: 1.0
Hipwell, 2000 [10]	Colour	Basic Thresholding	Rule-based	Private dataset (3783 images)	Sensitivity[a]: 0.6 FP/image[a]: 1.0
Sinthanayothin, 2002 [17]	Colour	Moat operator	N/A	Private dataset (14 images)	Sensitivity[b]: 0.885 Specificity[b]: 0.997
AbdelAzeem, 2002 [18]	Florescence	Hough transform	Rule-based	Private dataset (3 images)	Sensitivity[a]: o.6 FP/image[a]: 17.67
Streeter, 2003 [11]	Colour	Gaussian filter	Linear Discriminant Analysis	Private dataset	Sensitivity[a]: 0.3 FP/image[a]: 1.0
Niemeijer, 2005 [21]	Colour	Gaussian Filter pixel classification	K-Nearest-Neighbours	Private dataset (100 images)	Sensitivity[a]: 0.83 FP/image[a]: 1.0 Sensitivity[b]: 1.0 Specificity[b]: 0.5
Fleming, 2006 [12]	Colour	Gaussian Filter	K-Nearest-Neighbours	Private dataset (1441 images)	Sensitivity[a]: 0.51 FP/image[a]: 1.0 Sensitivity[b]: 0.91 Specificity[b]: 0.5
Quellec, 2008 [15]	Colour	N/A	N/A	ROC dataset	Sensitivity[c]: 0.90 Specificity[c]: 0.898
Mizutani, 2009 [16]	Colour	double-ring filter	Neural network	ROC dataset	Sensitivity[c]: 0.15 PPV[c]: 1.0
Sánchez, 2009 [7]	Colour	Mixture model-based clustering	N/A	ROC dataset	ROC score: 0.332 Sensitivity[a]: 0.30 FP/image[a]: 1.0
Zhang, 2010 [13]	Colour	Multiscale Gaussian	Rule-based	ROC dataset	Sensitivity[a]: 0.11 FP/image[a]: 1.0 ROC: 0.201
Giancardo, 2010 [5]	Colour	Basic Thresholding	N/A	ROC dataset	Sensitivity[a]: 0.22 FP/image[a]: 1.0
Lazar, 2011 [2]	Colour	Local Maxima scanlines	N/A	ROC dataset	Sensitivity[a]: 0.38 FP/image[a]: 1.0 ROC score: 0.355
Sopharak, 2011 [23]	Colour	extended-minima	Naïve Bayes	Private dataset (45 images)	Sensitivity[c]: 0.816 Specificity[c]: 0.99
Giancardo, 2011 [6]	Colour	Basic Thresholding	N/A	ROC dataset	Sensitivity[a]: 0.43 FP/image[a]: 1.0 ROC: 0.375
Lazar, 2013 [1]	Colour	Local Maxima scanlines	N/A	ROC dataset	Sensitivity[a]: 0.41 FP/image[a]: 1.0 ROC: 0.423
Rocha, 2012 [24]	Colour	N/A	Support Vector Machine	DIARETDB1 v1	Sensitivity[c]: 0.91 Specificity[c]: 0.5
				MESSIDOR	Sensitivity[c]: 0.93 Specificity[c]: 0.5
Sopharak, 2013 [25]	Colour	extended-minima	Bayesian	Private dataset (80 images)	Sensitivity[c]: 0.86 Specificity[c]: 0.99
Akram 2013 [26]	Colour	Gabor filter	Hybrid classifier	DIARETDB0, DIARETDB1 v1	Sensitivity[a]: 0.99 Specificity[a]: 0.997 Accuracy[a]: 0.994
Li, 2013 [4]	Colour	Multi-orientation Gaussian (MSMF)	N/A	ROC dataset	Sensitivity[a]: 0.05 FP/image[a]: 1.0
Junior, 2013 [3]	Colour	Extended Minima	N/A	DIARETDB1v1	Sensitivity[c]: 0.87 Specificity[c]: 0.92
Inoue, 2013 [19]	Colour	Hessian Matrix Eigenvalues	Neural network	ROC dataset	Sensitivity[a]: 0.18 FP/image[a]: 1.0
Adal, 2014 [21]	Colour	Hessian Matrix Eigenvalues	Support Vector Machines, K-Nearest-Neighbours, Naïve Bayes, Random Forest	ROC dataset	ROC score: 0.363 Sensitivity[a]: 0.364 FP/image[a]: 1.0
Ram, 2015 [27]	Colour	Morphological reconstruction	K-Nearest-Neighbours	ROC dataset	Sensitivity[a]: 0.31 FP/image[a]: 1.0
				DIARETDB1 v1	Sensitivity[a]: 0.73 FP/image[a]: 1.0
				Private dataset	Sensitivity[a]: 0.18 FP/image[a]: 8.0
Wu, 2015 [14]	Colour	Multiscale Multi-orientation Gaussian (MMMF)	Support Vector Machines, K-Nearest-Neighbours, Linear Discriminant Analysis	ROC dataset	Sensitivity[a]: 0.23 FP/image[a]: 1.0 Sensitivity[c]: 0.92 Specificity[c]: 0.50

(*continued on next page*)

Table 1 (*continued*)

Paper	Image Type	Initial candidates method	Classifier used	Reported Performance	
				Dataset	Performance
Srivastava, 2015 [20]	Colour	Frangi-based filters	Support Vector Machines	MESSIDOR+ DIARETDB1 v1	Sensitivity[c]: 1.00 Specificity[c]: 0.50
Romero, 2015 [28]	Colour	Hit-or-miss transform	Neural networks	DIARETDB1v2.1 ROC dataset	Sensitivity[c]: 0.93 Specificity[c]: 0.94 Sensitivity[c]: 0.88 Specificity[c]: 0.97
Haloi, 2015 [29]	Colour	N/A	Nearest-mean classifier	DIARETDB1v2.1 ROC dataset	Sensitivity[c]: 0.88 Specificity[c]: 0.97 AUC[c]: 0.98

the-art algorithms in other computer vision applications. However, deep learning requires massive datasets for training [31] and such large labelled datasets are not yet available for retinal images.

Ram [27] used a dual classifier in order classify the initial candidates. The initial candidates were detected using a simple thresholding operation after preprocessing. Two classification stages are then applied. The first classification stage was applied in order to separate MAs from vessels. The features used for this purpose are a second derivative Gaussian at multiple orientations, difference of Gaussians and inverted Gaussians. The second classification stage was applied in order to further separate MAs from other types of noise for the MA classification on a pixel level rather than at a candidate level. This means that each pixel gets classified as either an MA or not, rather than each initial candidate as a whole. After preprocessing, the extended-minima transform is used to detect the initial candidates, and a Bayesian classifier was used to perform the pixel-based MA classification. Similarly, Junior [3] presents the same technique as Sopharak, but does not apply a classification stage. Akram [26] used a Gabor filter for the detection of the initial candidates. The Gabor filter is applied at multiple scales and rotated at various angles, and the maximum response is computed. This causes a large response for vessels, microaneurysms and haemorrhages. Vessel segments are removed using a vessel segmentation technique. A hybrid classifier is used for to reduce the false positives in the initial candidates. The technique has reported a lesion-based specificity and accuracy measure, even though it is not possible to measure the number of true negatives at the lesion level [8].

The objective of the present work is as follows: 1) to present a new technique for MA detection based on an ensemble classifier for classification. 2) Introduce 70 of the most common features used in the literature and perform feature ranking in order to identify the features that are most important for discriminating MA candidates from spurious objects. 3) To introduce a new groundtruth dataset for MA detection based on the MESSIDOR dataset.

Section 2 describes the methodology of the proposed algorithm. In Section 3, a new dataset of MA groundtruth images is introduced and the experiments performed to evaluate the algorithm are discussed and the results presented. A final discussion is presented in Section 4 and concluding remarks are presented in Section 5.

2. Methodology

The proposed method is based on the method suggested by Fleming [12]. The main modifications that were made to Fleming's algorithm will be stated throughout the methodology section. This work is an extension of the algorithm published in Ref. [32] and includes a more extensive evaluation as well as detailed feature analysis. The proposed methodology consists of three phases: 1) preprocessing 2) MA Candidate Detection and 3) MA Candidate Classification. During the preprocessing stage non-uniform illumination is removed from the image using background subtraction. Noise removal is also performed during this stage. In the MA Candidates Detection phase an initial set of MA candidates are detected. Ideally all the candidates in the image should be detected with as few false positives as possible. Most of these false positives should then be removed during the Candidate Classification phase. The three stages of the proposed algorithm will be explained in the following sections.

Despite being published in 2006, Fleming's reported per-lesion performance on a large private dataset is comparable to recently published methods. This makes it reasonable to use Fleming as a baseline for comparison with the proposed technique. These methods include Wu (2015) [29], Adal (2014) [28], Inoue (2013) [20] and Li (2013) [14]. This is also illustrated in Table 1, and discussed in Section 5.1.

2.1. Preprocessing

The preprocessing steps are as follows: Given a colour retinal image (Fig. 2(a)) the green channel is extracted (Fig. 2(b)) since MA candidates appear with high contrast in this channel. Salt-and-pepper noise is removed by applying 3 × 3 median filter. Contrast-Limited Adaptive Histogram Equalisation (CLAHE) [33] is applied in order to improve the contrast in the image. Further noise removal is performed by applying a 3 × 3 Gaussian filter to the image. Let the resulting of the previous operations be I_{adapt}. Shade correction (I_{shade}) is performed by dividing the image by an estimate of the background:

$$I_{shade} = I_{adapt} / I_{bg} \tag{1}$$

where I_{bg} is the background estimate calculated by applying a 68 × 68 median filter to I_{adapt}. The filter size is chosen to be large enough in order to eliminate vessels and other features in the image. Finally, global contrast normalization is performed on the resulting image by dividing it by its standard deviation:

$$I_{con} = \frac{I_{shade}}{std(I_{shade})} \tag{2}$$

where $std(I_{shade})$ represents the standard deviation of the shade corrected image. The result of these operations is illustrated in Fig. 2(c). Following these operations we need to detect an initial set of MA candidates from the preprocessed image. This is described in the following section.

2.2. MA candidate detection

After performing noise removal and shade correction, an initial set of MA candidates can be detected. The method used is based on that proposed by Fleming [12]. A Gaussian matched filter ($\sigma = 1$) is used in order to enhance circular dark regions in the image. Since blood vessel cross-sections have intensity profiles similar to MAs, they need to be removed before applying the Gaussian matched filter. The following morphological operations are applied for vessel removal.

A closing operation is applied using a linear structuring element at multiple directions. The minimum of the application of closing operation at multiple operations was then subtracted from the shade corrected image [12].

$$I_{bothat} = I_{shade} - \min_{i=0..7}(I_{shade} \cdot strel(\pi i/8, n)) \tag{3}$$

where strel(x,n) represents a linear structuring element at an angle of x

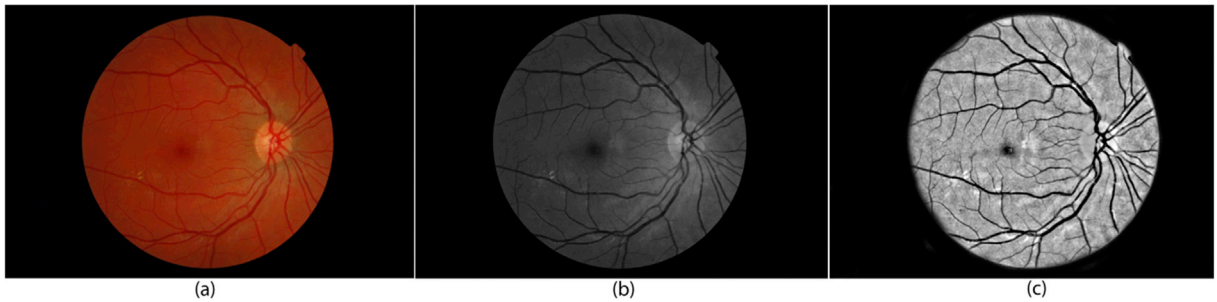

Fig. 2. An example of the preprocessing stage. a) The colour image, b) the green channel image, c) the preprocessed image. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

degrees and of length n. The size of the structuring element should be chosen to be larger than the largest vessel in the images (in our case a size of 11 pixels was selected through direct measurement in the images). This operation causes vessels to be removed from the image while retaining circular objects which resemble the shape of MAs (Fig. 2 (a) and (b)).

A Gaussian matched filter ($\sigma = 1.0$) is then convoluted with I_{bothat} in order to enhance circular dark regions $I_{gauss} = I_{bothat} {}^{*}gauss(1.0)$ (Fig. 2(c)). The resulting response probability is then thresholded as follows:

$$I_{thresh} = thresh\left(I_{gauss},\ 5\tau\right) \tag{4}$$

The value of τ is chosen to be the threshold value at which the top 5% of pixels are selected [12]. A region growing operation based on Fleming [12] is performed in order to enhance the shapes of the detected MA candidates. The set of initial candidates are used as input. The procedure involves iteratively growing along the 8-connected pixels from the minimum intensity pixel of the candidate until a stopping condition. In our case, the stopping condition is when a maxima point of the "energy function" is reached. The energy function is defined as the average value of the gradients around the boundary of the grown region. All the parameters at this stage have been kept the same except the maximum grown size.

Through our experiments it was found that the maximum grown size of 3000 pixels suggested by Fleming resulted in large blood vessel regions being falsely identified. We empirically found that a value of 100 pixels was a more suitable value for the maximum area and this parameter modification decreased the number of false positives appreciably, while achieving almost the same sensitivity. The value was chosen to be over twice the size of the average MA size in the groundtruth images (Fig. 3 (d)). The region growing operation causes the intensity profile of the boundary to be detected more accurately.

2.3. MA candidate classification

The initial candidate detection phase will inevitably produce false positives. The main reasons for this are: 1) vessel cross sections or vessels that were not removed before the Gaussian filter and 2) noise in the image that looks similar to MAs. For these reasons a classification phase was required in order to reduce the number of false positives that were detected during the candidate detection phase.

The proposed method uses a Tree Ensemble classifier for classification. A Tree Ensemble classifier is an ensemble classifier based on decision tree learning. An ensemble classifier combines the decisions of multiple weak classifiers. Our main motivation for the use of this classifier are: 1) Successful application in other fields [34,35], 2) it can rank features while performing classification, giving insights about the most important features, and 3) robustness to outliers and the ability to cope with small training sets [36].

Given a training set L consisting of data $\{(y_n, x_n), n = 1\ldots, N\}$ where y represents the classification label (1 or 0 in our case), a given **CART (Classification And Regression Trees)** classifier $T(x, L)$ will predict y given unlabeled data x.

However, in the case of an ensemble of trees we are given a sequence of training sets $\{L_k, k = 1\ldots, K\}$ and a sequence of classifiers $\{T_k(x, L_k)\}$ is produced. The j^{th} classifier $T_j(x, L_j)$ in the set will produce a label y_j. The set of labels $\{y_k\}$ produced by the K classifiers need to be aggregated to produce a final label y for unlabeled data x. In our case a majority vote of the set of class labels $\{y_k\}$ is used to produce the final classification y. It has been shown that combining the results of a set of weak classifiers $\{T_k(x,\ L_k)\}$ often outperforms using a single classifier on the whole training set $T(x, L)$ [37].

The final point that needs to be addressed is that given a training set L, how can we produce a set of training sets $\{L_k\}$ that will be used to train each tree classifier $T_j(x, L_j)$. A sampling technique known as Bootstrap aggregation (or bagging) [37] was used in order to sample the training data during the training process. In bagging, the j^{th} training set L_j is obtained by drawing M samples (with replacement) from the set of N training data, L (where $M \leq N$). In practice, in order to produce L_j a set of M random numbers $\{r_m; r_m \leq N, m = 1\ldots M\}$, and then L_j is drawn using $L_j = \{(y_{r_m},\ x_{r_m})\}$. There is no restriction that the generated random numbers are unique and therefore each sample in the set $\{(y_n, x_n)\}$ may be used more than once or not at all in L_j. After producing K training sets from L, there will be a set of samples in L that have not been drawn in any of the samples in the j^{th} classifier L_j. These unused features can be used to estimate the classification error (out-of-bag-error) for each tree and also estimate the "importance" of each feature (based on each tree and then averaged over all trees) [38]. The bagging approach is used to increase the diversity of training samples across the trees, which leads to increased prediction accuracy for unstable classifiers (including decision trees) [37,39].

We have extended Fleming's [12] feature set of 10 features to include a set of 70 features. These were based on the features that have been reported in the literature. Table 2 displays a list of the 70 features that were fed into the classifier. These features are explained below in the same order of appearance as Table 2:

- **Fleming's features:** These are the features introduced by Fleming in his technique [12]. These features rely on fitting a paraboloid to each candidate's intensity profile in order to estimate some parameters from the paraboloid. These features are based on both the shape and intensity of the object. A detailed explanation of these features can be found in the original paper [12].
- **Shape features & Moment Invariants:** These features describe various shape properties of the detected candidates. Moment Invariants (10) are 7 features that represent various shape properties of an object [40]. Other shape features include Aspect Ratio (major axis length/minor axis length), major axis length, minor axis length, Perimeter, Area, Eccentricity, Compactness. Some of these are common to Fleming's features, however these are calculated at a pixel level rather than after fitting a paraboloid to the candidate. To elaborate, Fleming estimates a paraboloid for each candidate and then

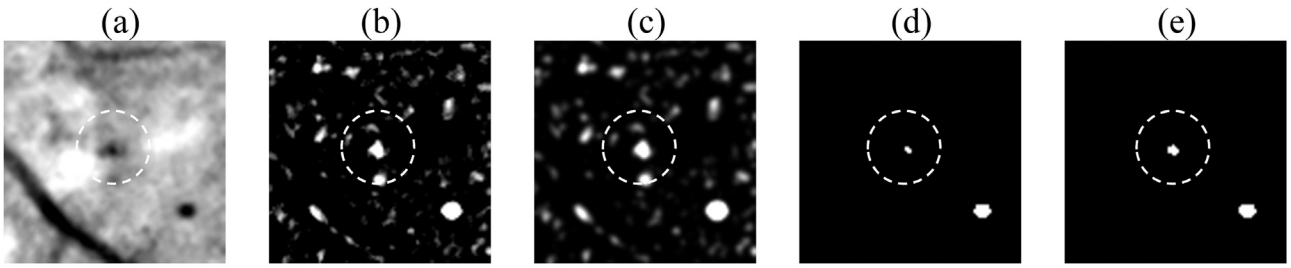

Fig. 3. An example of the candidate detection phase a) The preprocessed image, b) The result of the bottomhat operation, c) the Gaussian filter response d) The thresholded image e) the result of the region growing operation. The highlighted region is a true microaneurysm.

Table 2
Features list. The symbols below (G,t,seed,c,σ) are defined in Section 2.3. Key: std - standard deviation, max - maximum, min - minimumç.

Category	Index	Description	Parameters	Feature count
Fleming	1	Number of peaks	N/A	1
	2	Major Axis length	N/A	1
	3	Mean of minor and major axis	N/A	1
	4	Eccentricity	N/A	1
	5	Depth of candidate in the original image	N/A	1
	6	Depth of candidate in the preprocessed image	N/A	1
	7	Energy	N/A	1
	8	candidate depth/mean diameter of MA candidate	N/A	1
	9	Energy with depth correction	N/A	1
Moment Invariants	10	7 moment invariant features	N/A	7
Shape features	11	Aspect ratio	N/A	1
	12	Major axis length	N/A	1
	13	Minor axis length	N/A	1
	14	Perimeter	N/A	1
	15	Area	N/A	1
	16	Eccentricity	N/A	1
	17	Compactness	N/A	1
Gaussian features	18	Gaussian seed pixel response: $G_\sigma(\text{seed(c)})$	$\sigma = 1$	1
	19	$\underset{(x,y)\in c}{\text{mean}}(G_\sigma(x, y))$		1
	20	$\underset{(x,y)\in c}{\text{std}}(G_\sigma(x, y))$		1
Gaussian Features 1D	21	Max 1D Gaussian response at various angles: $\max_{t\in\theta}(G^{1D}_{1,t}(x, y))$	$\theta \in \{0, 10, 20, \ldots, 180\}$	1
	22	Min 1D Gaussian response at various angles: $\min_{t\in\theta}(G^{1D}_{1,t}(x, y))$		1
	23	Mean 1D Gaussian response at various angles: $\underset{t\in\theta}{\text{mean}}(G^{1D}_{1,t}(x, y))$		1
	24	Std of 1D Gaussian response at various angles: $\underset{t\in\theta}{\text{std}}(G^{1D}_{1,t}(x, y))$		1
	25	1D gaussian response at angle perpendicular to the maximum response (30)	N/A	1
	26	max (29,33)	N/A	1
Intensity features	27	Sum of candidate intensities	Applied to red, blue, green, hue, saturation, value and preprocessed channels.	7
	28	mean candidate intensity		7
	29	standard deviation of the candidate intensity		7
	30	Range (Max - min candidate value)		7
	31	candidate contrast		7
Morphological features	32	maximum candidate response of the morph close ratio	N/A	1
	33	minimum candidate response of the morph close ratio	N/A	1
	34	mean candidate response of the morph close ratio	N/A	1
Total				**70**

computes the values of eccentricity and major & minor-axis length from the paraboloid. In contrast, these features are calculated from the binary image.

- **Gaussian Features**: Features that are based on Gaussian filters have been extensively used in the literature [12–14,25,30]. In our case we have experimented with features that rely on $\sigma = 1$ since that is parameter used during the initial candidates detection phase. Some definitions related to these features will follow. The symbols mentioned below also appear in Table 2. Let I_{shade} be the shade corrected image (Section) and:

$$G_\sigma = I_{shade} * gauss(\sigma) \tag{5}$$

be the Gaussian filter response for sigma $= \sigma$ and $G_\sigma(x, y)$ be the filter response at coordinates (x, y). Let $\overline{C}$ be a set of initial candidates detected (after region growing). Each candidate (c) is a set of coordinates (x_i, y_i). Let $seed(c)$ be the coordinates (x_s, y_s) of the minimum intensity defined as follows:

$$seed(c) = (x_s, y_s) = \underset{(x,y)\in c}{\text{argmin}}(I_{shade}(x, y)) \tag{6}$$

A 1-Dimensional Gaussian is a special case of G_σ applied linearly in one direction. $G^{1D}_{s,t}(x, y)$ is the 1D Gaussian applied at angle t and a scale (standard deviation) of s. In our case we have applied the 1D Gaussian at a constant scale ($s = 1$). Let the set θ be the set of angles applied at each

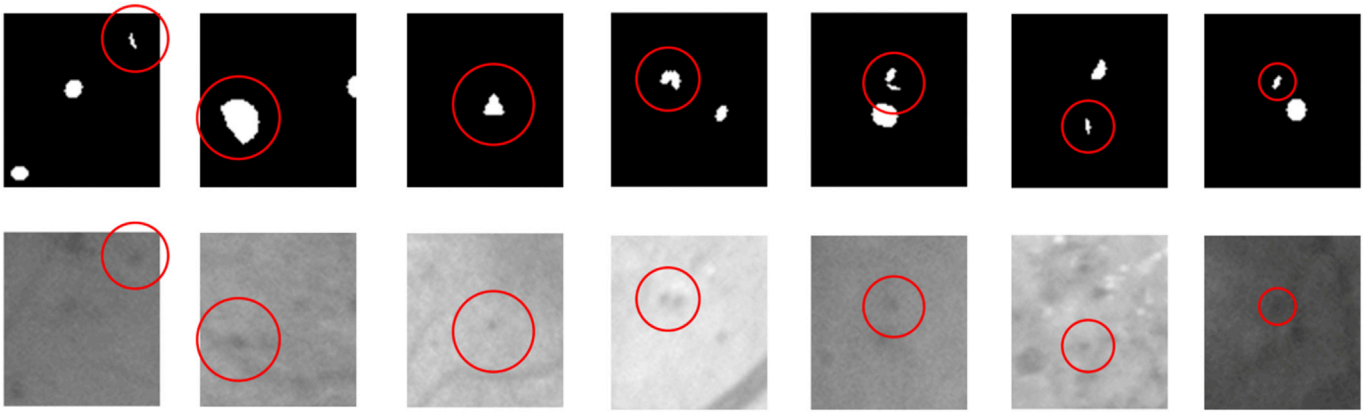

Fig. 4. Examples of DIARETDB1 groundtruth candidates that do not correspond to the microaneurysm shape in the original image. The first row shows patches MA groundtruth in the dataset. The second row shows corresponding patches from the retinal images. The retinal image patches have been enhanced to improve MA contrast.

Table 3
Distribution of DR grades (a) and resolutions (b) of images in the dataset.

Retinopathy Grade	Number of Microaneurysms	No. of images (training)	No. of images (test)
DR0	0	8	8
DR1	1–5	3	4
DR2	6–14	3	3
DR3	>15	2	1
	TOTAL	16	16

coordinate. In our experiments:

$$\theta \in \{\theta_i \ : \theta_i = 10^*i; \ \ i = [0\ldots18]\} \tag{7}$$

- **Intensity Features:** These are calculated directly from the intensity in the image at multiple bands: the red (R), blue (B), green (G) band in the RGB colour space; the Hue (H), saturation (S) and value (V) bands of the HSV space [13].
- **Morphological Features:** These three features are based on applying a linear morphological close operator (15 px size has been empirically chosen to be larger than the largest vessel in the dataset) at different angles ($\theta \in \{\theta_i \ : \theta_i = 22.5^*i; \ \ i = [0..7]\}$) and are aimed at discriminating vessels from MAs. This is because the linear structures of vessels would respond differently to different angles of the linear operator, while the circular nature of MA objects would cause the response of the morphological operator to be more uniform.

3. Experiments

This section explains the methodology that has been followed in order to evaluate the proposed algorithm. An overview of the publicly available datasets for microaneurysm detection is presented in Section 3.1. An MA groundtruth dataset based on a subset of the MESSIDOR dataset [41] is also introduced in this section. Details of the evaluation method are explained in Section 3.2.

3.1. Dataset

To the best of our knowledge, there are two public datasets for MA detection: the Retinopathy Online Challenge dataset (ROC dataset) [1] and the DIARETDB1 v2.1 dataset [42]. The ROC dataset contains 100 images split into 50 training images and 50 test images. Groundtruths are only available for the training set. The test set groundtruths are not public since the contest organizers used those to evaluate submissions. Moreover, the groundtruth of this dataset has generated discussion in the literature [6,16] due to the fact that many of the MA candidates marked in the groundtruth are invisible to the viewers or could not be seen by other expert observers. This made it difficult to rely on this dataset as a benchmarking dataset for MA algorithms. The DIARETDB1 v2.1 (henceforth DIARETDB1) dataset is a general-purpose dataset for the detection of diabetic retinopathy (DR) from retinal images. The dataset includes groundtruths for various lesions in the image including MAs, haemorrhages and exudates, as labelled by 4 experts. However, in order to reduce the bias in labelling, the experts were not instructed to mark a specific shape for each lesion. Hence, some of the experts marked large regions around a group of MAs as groundtruths and others did not. Thus, labelling of some of the MAs resulted in unusual shapes after the 4 expert labels were fused together (Fig. 4). To address the shortcomings in the current public datasets we introduced a new public dataset of MA groundtruths for the purpose of benchmarking MA detection algorithms. This is described in the next section.

Due to the reasons mentioned above, as well as to add more variety to the existing datasets, we have produced a new public MA groundtruth set based on the MESSIDOR database[1] [41]. Thirty-two images were selected from the MESSIDOR dataset to cover a wide range of retinopathy. A summary of the images in dataset in terms of retinopathy grade and number of MAs included in the set is shown in Table 3. The grade is predominantly based on the number of MAs (the presence of haemorrhages and new vessels is also factored in) [41]. We have included 16 healthy images (without MAs) in order to maintain a balanced dataset while performing per-image MA evaluation (evaluating whether or not an MA candidate exists for each image). A summary of the distribution of retinopathy grade in the 32 images is presented in Table 3. The images belonged to the same resolution of 1440 × 960 px.

The images were groundtruthed by an expert grader. The dataset has also been made publicly available 1. All the visible MAs were marked during the process. A circular marker was used rather than pixel-based marker [43]. The main reason for the use of a circular marker is that majority of the literature has relied on lesion-based metrics to measure the accuracy of detection. Using lesion-based metrics makes the results more sensible since the measure is informative of the amount of MA candidates that were detected by a given algorithm. In contrast, reliance on pixel-based metrics can be misleading due to the imbalance in proportion between very few MA pixels and a large number of background pixels.

Motivated by the Retinopathy Online Challenge [43], each MA candidate was labelled using one of the following categories: Obvious, Regular or Subtle and Close-to-Vessel (Fig. 1). The labels Obvious,

[1] The groundtruth dataset can be downloaded using the following link: http://blogs.kingston.ac.uk/retinal/?p=311.

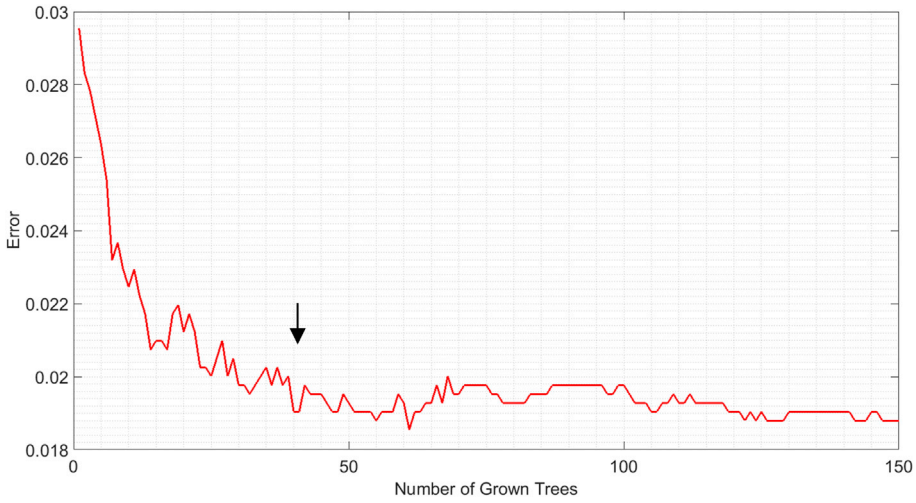

Fig. 5. Out of bag (OOB) classification error as a function of the number of trees in the Tree Ensemble classifier.

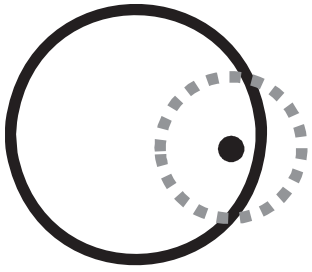

Fig. 6. An example of a candidate (dotted circle) that is considered to match a ground-truth (solid circle). There is a match since the centre of the candidate lies within the groundtruth region.3.

Regular or Subtle are based on the relative visibility and/or local contrast of the corresponding MA in the image. Close-to-Vessel is a label given to MA candidates that lie close to a blood vessel. A detailed explanation of each category is mentioned in Ref. [43].

3.2. Evaluation

We have used the public MESSIDOR dataset mentioned in the previous section to train and measure classifier model error. Hence we have built our models using the training set and measured the accuracy of the model using the 16 images in the MESSIDOR test set. In order to ensure that our model is not overfitting the MESSIDOR dataset we have also measured the performance of our model on the DIARETDB1 test set. In the case of an overfit model the error on the DIARETDB1 test set would be greater than the error on the MESSIDOR test set [35]. Hence we want to ensure that the error on both the MESSIDOR and the DIARETDB1 sets was similar for our model. The following procedure was followed in order to perform the evaluation on the dataset:

- The MESSIDOR dataset was split into 16 images for building the model (training set) and 16 images for measuring the model classification error (test set) as shown in Table 3.
- The procedure outlined in Section 2.3 was used to generate the 70 features. The training groundtruth was used to label the features. These features were used to train the Tree Ensemble classifier and generate the model. Note that the training set was undersampled in order to maintain a balance between the positive and negative samples. One parameter that needs to be set for the Tree Ensemble classifier is the number of generated trees (N). Fig. 5 shows the out-of-bag (OOB) classification error as a function of the number of trees in the Tree Ensemble classifier. We have selected a value of $N = 40$ based on Fig. 5. This is also within the range recommended by Brieman [37]. The value of N is chosen from the plot at the point where there is no more significant decrease in the OOB error.

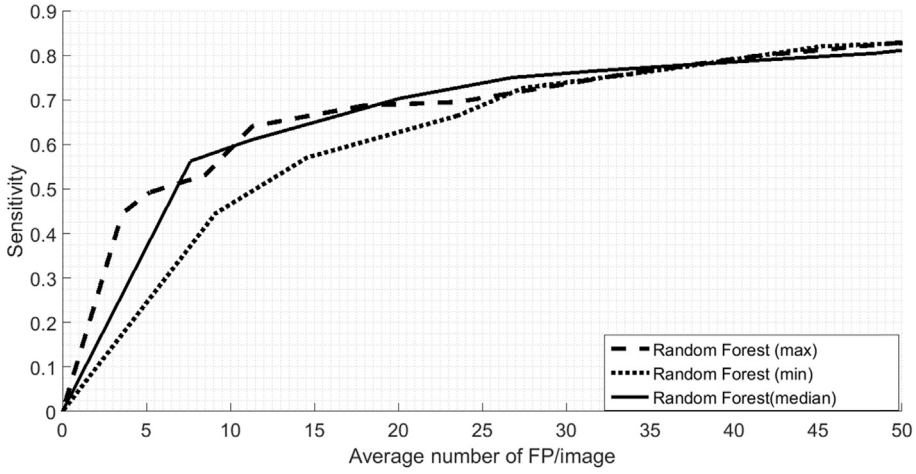

Fig. 7. An example of variability in results everytime a Tree ensemble model is built for classification. In this example the Tree ensemble classifier was run 11 times and the maximum, minimum and median results are displayed.

- For each image in the MESSIDOR test set, the procedure outlined in Section 2.2 was used to find the set of candidates and their corresponding features. Each candidate feature vector was then fed into the classifier in order to classify it as a true candidate or a false positive.
- The final classified set of candidates was then compared against the ground truths in order to assess the performance.

In addition to the proposed algorithm, we have also implemented Fleming's algorithm [12] in order to compare it against the proposed technique. Fleming uses a K-Nearest-Neighbours classifier with 9 features. We call this the "basic feature set". In addition, we have also used the K-Nearest-Neighbours classifier with all 70 features, and we call this the "extended feature set".

In order to measure the accuracy of the algorithm, we measured the sensitivity of the proposed method [8]. Given image I_i in a test set (for $i = [1..t]$, where t is the number of images in the test set), let G_i be the set of true MA objects (groundtruth) for image I_i and C_i be the set of detected candidates after classification (Section 2.3) for image I_i. The sensitivity is defined as:

$$Sensitivity = \frac{\sum_{i=1}^{t} |G_i \cap C_i|}{\sum_{i=1}^{t} |G_i|} \tag{8}$$

where $|x|$ represents set cardinality of x. Thus the sensitivity is the

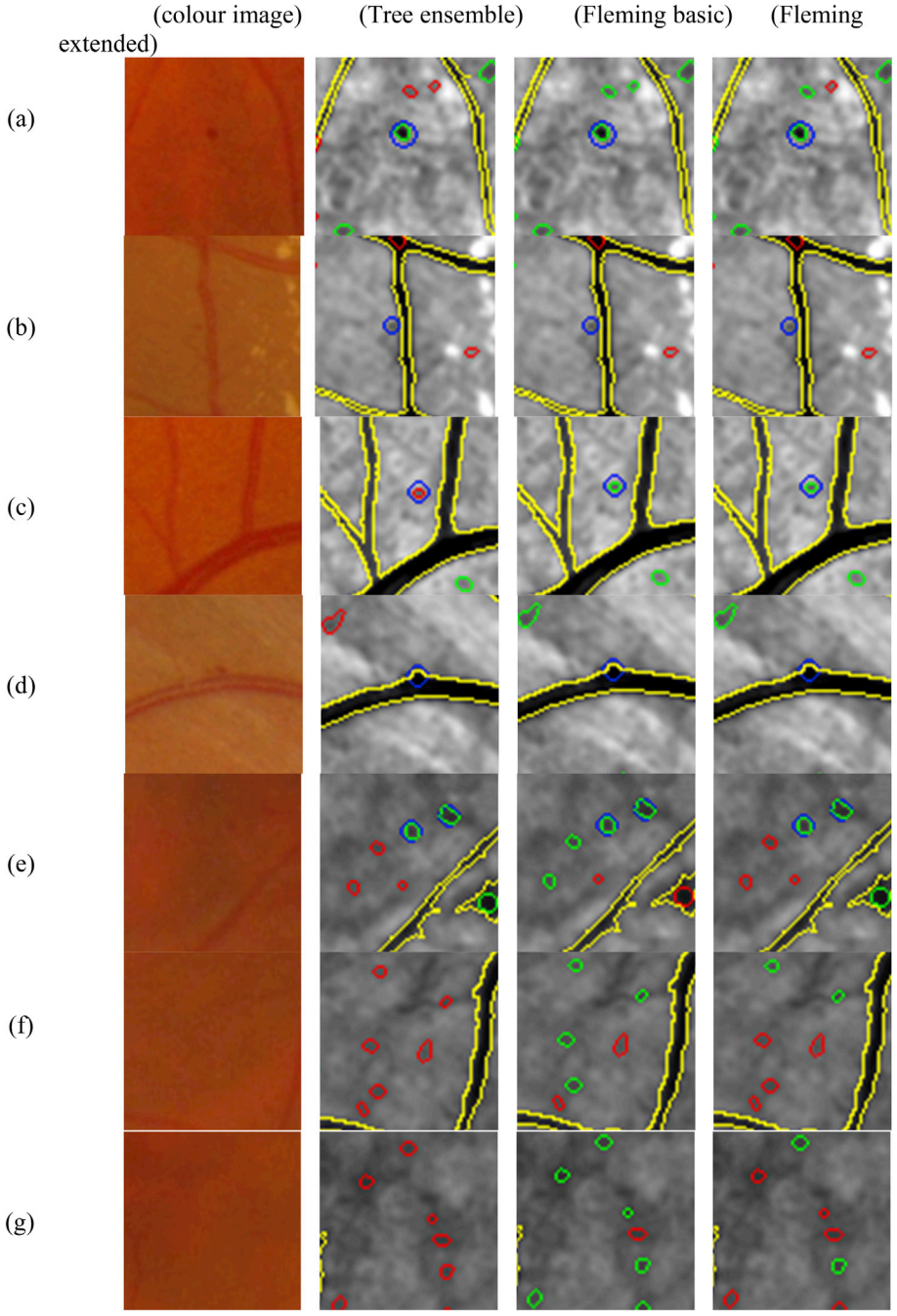

Fig. 8. Examples of microaneurysm detection algorithms applied to the MESSIDOR dataset. The first column shows the colour image patch. Columns 2–4 show the preprocessed image with the algorithm results highlighted. The blue circle represents the groundtruth labelled microaneurysm. The green circle represents an MA candidate detected by the algorithm. The red circle represents an MA candidate that was detected as a candidate MA but classified as a false positive by the classifier. The yellow boundaries shows the vessel regions detected by the QUARTZ software [44]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

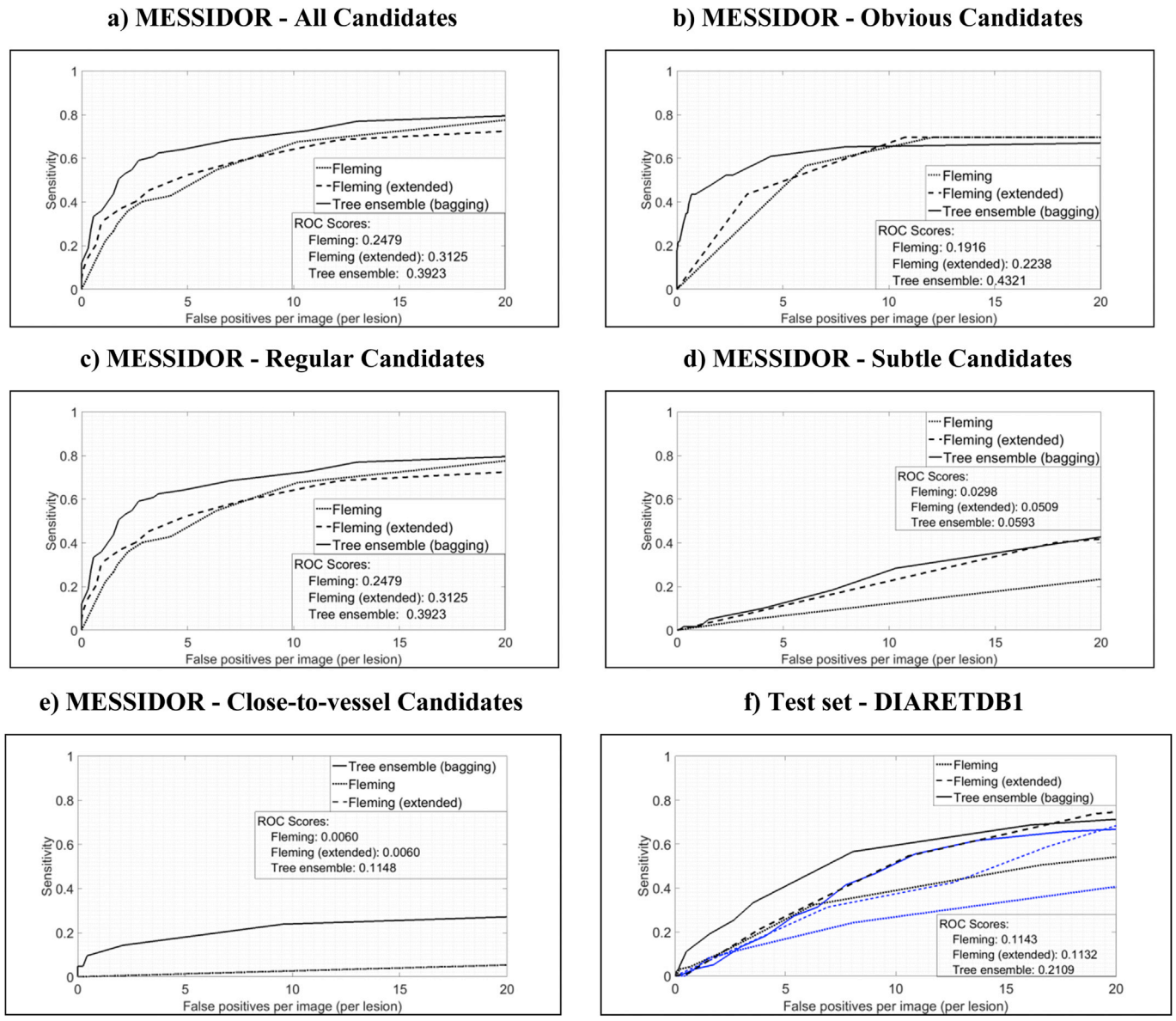

Fig. 9. Free-Receiver operating curve (FROC) for all microaneurysm candidates in the test set. In f) the black lines represent the performance of each respective model on the DIARETDB1 test set by training on the DIARETDB1 training set, while the blue lines represent the performance of each model on the DIARETDB1 test set after training on the MESSIDOR training set.

proportion of true candidates detected in proportion to the total number of true candidates. A candidate $c \in C$ is considered to be equivalent to $g \in G$ if the pixel coordinates of g and c overlap by at least 1 pixel (Fig. 6). Note that we are measuring the sensitivity at a candidate level rather than at a pixel level (lesion-based sensitivity). Since we cannot determine the number of true negatives, we used a Free Receiver Operating Curve (FROC) rather than a traditional ROC curve [9]. In an FROC curve, the x-axis is replaced with the average number of false positive candidates per image instead of the specificity. Fig. 9 shows FROC curves for both the proposed method (using the Tree Ensemble classifier) (solid) and Fleming's method (using K-Nearest Neighbours) (dotted) for multiple categories in the dataset. Each curve was generated by evaluating the trained model on the test set. The dotted curve represents the performance of Fleming's state-of-the-art algorithm on the MESSIDOR dataset, the dashed curve represents the performance of Fleming's algorithm (using the K-Nearest-Neighbours classifier) with the extended feature set. A

Table 4
ROC scores for multiple categories in the set.

Category	Method		
	Fleming (basic feature set)	Fleming (extended feature set)	Proposed method (Tree ensemble)
MESSIDOR - All candidates	0.2479	0.3125	**0.3923**
MESSIDOR (Obvious Only)	0.1916	0.2238	**0.4321**
MESSIDOR (Regular Only)	0.2479	0.3125	**0.3923**
MESSIDOR (Subtle Only)	0.0298	0.0509	**0.0593**
MESSIDOR (Close-to-vessel Only)	0.0060	0.0060	**0.1148**
DIARETDB1 (test set)	0.1143	0.1132	**0.2109**

value of $K = 15$ was used for the K-Nearest-Neighbours classifier [12]. Each classifier produces a probability value (P) between 0 and 1 representing the likelihood of a candidate belonging to 0 or 1. We use a threshold value P_t to produce the final classification (i.e. $class = 0$ if $P \leq P_t$, otherwise $class = 1$). The value of P_t was varied in order to generate the FROC curve. Tree Ensembles generate trees at random and generates the attribute splits at random as well [34,35]. Due to this feature of the classifier, every run produces results with slightly different accuracy (Fig. 7). To overcome the varying results, we have applied the Random Forest classifier multiple times and, for each run in the curve, we calculate the Area Under each Curve. Finally we display the curve with the median AUC value. This helps reduce the variability in the FROC curve. For our experiments we found that applying the classifier 11 times was sufficient to reduce the variability in the results. In an experiment run on the MESSIDOR dataset the Tree Ensemble classifier was run 100 times and the average mean squared error (MSE) for all the curves was found to be 0.0124, which shows that the variability in the classifier can be considered negligible.

4. Results

In this section experimental results of the proposed algorithms are presented. Both visual and quantitative results are presented. Section 4.1 presents patches of the algorithm which show detections of microaneurysms and the classification results. The patches are compared to both Fleming and an extended version of Fleming. Section 4.2 will present some quantitative results for both the MESSIDOR and DIARETDB1 datasets. An analysis of the features and the discriminative ability of each feature will be listed in Section 4.3.

4.1. Visual results

Fig. 8 shows example patches from the MESSIDOR images for the three methods mentioned in Section 3.2. The figure shows several patches from multiple colour images. The patches are scaled by 200% and for each patch the groundtruth, the MA candidates (after region growing) and the result of the classification have been highlighted. For the purpose of comparison the results are shown for the proposed method, Fleming's algorithm with the basic feature set, and Fleming's algorithm with the extended feature set.

The colour codes of the labels in Fig. 8 are as follows: The blue circle represents the groundtruth labelled microaneurysm. The green circle represents an MA candidate detected by the algorithm. The red circle represents an MA candidate that was detected as a candidate MA but classified as a false positive by the classifier. An analysis of these patches will be presented in Section 5.1.

4.2. Quantitative results

Fig. 9 shows the FROC curves for the three algorithms: Tree ensemble, Fleming (basic feature set) and Fleming (extended feature set). The model was built using the training set of the MESSIDOR dataset and the performance was measured using the MESSIDOR test set. The first FROC curve in Fig. 9 was generated by evaluating the classification model performance on the test set (16 images, 128 microaneurysms) after training using the entire set of MA labels in the training set (16 images, 128 microaneurysms).

In addition, evaluations for the subsets of the MESSIDOR groundtruths are also presented: obvious candidates, regular candidates, subtle candidates and close-to-vessel candidates. Each reported performance for a subset of the MESSIDOR dataset was trained on the respective subset of microaneurysm groundtruths in the dataset. This was done to highlight the variation in performance for each category of microaneurysms. Finally, the classification models were tested on the test set of the DIARETDB1 set. In order to demonstrate the overfitting process, we generated two models for each classifier: once by training on the MESSIDOR training set and the second by training on the DIARETDB1 training set. The models were then evaluated on the DIARETDB1 test set (61 images, 169 microaneurysms). In Fig. 9(f) the black curves represent the performance of each respective model on the DIARETDB1 test set using a model which was trained on the DIARETDB1 training set (28 images, 85 microaneurysms), while the blue curves represents the performance of each model on the DIARETDB1 test set using a model which was trained on the MESSIDOR training set. It is observed that the blue curve performance is comparable to the dashed black curve, which represents the performance Fleming (extended feature set) on the DIARETDB1 dataset. This shows that the Tree Ensemble classifier is generalizable across datasets since the performance is still comparable using a model trained on a different dataset.

In order to quantify the results further, we present the ROC Scores for each method in Table 4. The ROC Score [43] calculates the average sensitivity of the curve at multiple False Positive Rate intervals (1/8, 1/4, 1/2, 1, 2, 4, 8). In other words, the ROC score measures the average sensitivity of a technique at low False Positive rates. The ROC score simply captures the first section of the FROC curve (until 8 FP/image) as a simple quantifiable result. The ROC score focuses on the algorithm performance at low false positive rates. An extended discussion of the quantitative results will be presented in Section 5.1.

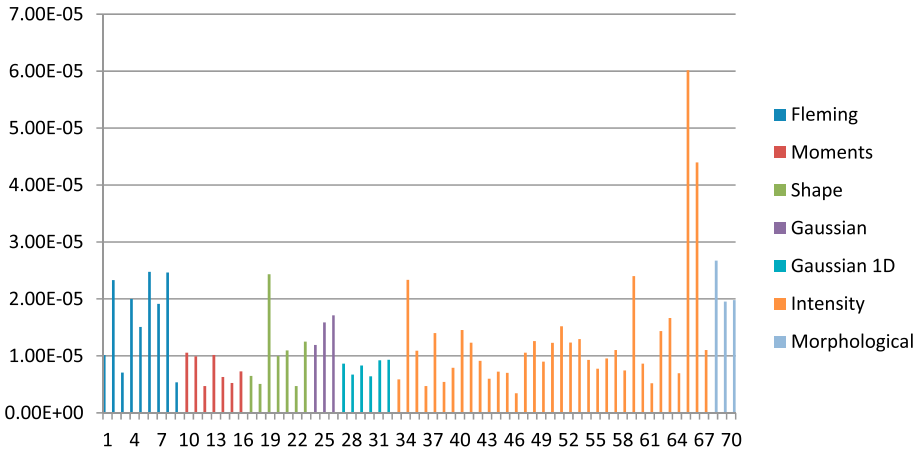

Fig. 10. Feature importance measured using the MESSIDOR dataset.

Tests were performed to measure the computational performance on the two datasets. The machine used for the tests was a core i5-4590 @ 3.30 GHz CPU with 8 GB RAM and an SSD hard drive. The average time required by the algorithm for each image in the MESSIDOR dataset was 166 s, while the average time per image in the DIARETDB1 dataset was 65 s. The algorithm's average performance on the DIARETDB1 dataset is around 40% of the time required for the MESSIDOR dataset. The main reason for this discrepancy is that the number of microaneurysms labelled per image in the DIARETDB1 is 2.85 (89 images and 254 labelled candidates) compared to 8 microaneurysms per image for MESSIDOR (32 images and 256 labelled microaneurysms). Intuitively this implies that less initial candidates will be detected per image in the DIARETDB1 dataset, which results in less computation for region growing and feature computation, since this needs to be computed for less candidates on average in DIARETDB1. Fleming has proposed several postprocessing methods and on average the reported performance in a dataset of 422 images was between 58 and 100s per image. The time performance of the hybrid classifier used by Akram [26] has been reported. The hybrid classier required $1.4 \times 10^{-3}s$ per candidate. One of the advantages of an ensemble classifier is that it is very efficient to train and test. The ensemble classifier requires less than a second for training in both DIARETDB1 and MESSIDOR datasets. Furthermore, classification is also very efficient for decision trees and it takes $4 \times 10^{-5}s$ on average.

4.3. Feature analysis

Since a large number of features are used in the proposed algorithm a presentation of the discriminative ability of each feature would be insightful to understand the most impactful feature and also for other researchers developing microaneurysm detection algorithms. Fig. 10 shows the feature importance of the features in the same order of appearance as in Table 1. The features are categorised by type. This performance was measured based on the trees generated from the tree ensemble classifier. A rough visual analysis of the chart shows that there are some important features for each category of features. A more detailed analysis of the chart is described in detail in Section 5.2.

5. Discussion

5.1. Visual and quantitative results

Fleming [12] is used as a baseline for comparison with the proposed technique. Fleming's reported per-lesion performance on a large private dataset is comparable to recently published methods. This makes it reasonable to use it as a baseline for comparison. Recent methods which are comparable to Fleming include Wu (2015) [29], Adal (2014) [28], Inoue (2013) [20] and Li (2013) [14]. While it is difficult to compare 2 FROC curves for 2 methods, the per-lesion sensitivities in Table 1 are all reported based at a value of 1 False positive/Image. This value was chosen since it is the median value of the 8 samples used while computing the ROC Score [43]. Using this value of 1 False Positives/Image makes it possible to compare between Fleming and other methods that have reported lesion-based sensitivity. The sensitivities for these methods at 1 FP/image are: Fleming (2006) 0.51, Wu (2015) 0.23, Adal (2014) 0.36, Inoue (2013) 0.23. In this section both a quantitative and visual analysis of the results of the proposed method is discussed.

In order to further analyse the results, we can understand the FROC curves in Fig. 9 by observing the patches in Fig. 8. An observation of the patches in Fig. 8 will help us understand the FROC curves in Fig. 9. It is observed that the Tree ensemble classifier is superior to the K-Nearest-Neighbours classifier in terms of eliminating the false positives in the image. This is observed more evidently in rows (e), (g) and (f), since the second column shows more red circles that do not intersect with a true candidate (blue circle). To elaborate, a red circle which does not intersect with a blue circle is a true negative. A blue circle which intersects with a green circle is a true positive. A blue circle which intersects with a red circle, or does not intersect with anything is a false negative.

The patches show that most of the time, the three methods equally detect the true positive candidates in the image. In fact, Fig. 8(c) shows that the proposed method has marked a true candidate MA as a false positive while the K-Nearest-Neighbours classifier correctly marked it as a true candidate. The conclusion drawn is that all methods are almost equivalent in terms of maintaining true positive candidates while the proposed method is superior in terms of eliminating false positives. This is important from a clinical perspective since a reduction in the number of false positives while maintaining the same sensitivity will avoid over-referral of the patients. One more interesting observation is that of Fig. 8(d) which shows an example of a close-to-vessel candidate. The figure shows that during the preprocessing phase the candidate merges with the vessel and therefore remains undetected as a candidate. This indicates why the performance of close-to-vessel candidates is very low for all methods.

The analysis of the patches in the previous paragraph will help us explain the FROC curves in Fig. 9. The FROC curves show that the proposed method performs better when all candidates are considered. In addition, it is also better at distinguishing close-to-vessel candidates. However, the curves intersect in the case of Obvious, Regular and Subtle candidates. This makes it difficult to judge which method performs better. For this purpose we needed a numerical measure in order to compare the curves in a more objective manner.

The ROC Score [43] calculates the average sensitivity of the curve at multiple False Positive Rate intervals (1/8, 1/4, 1/2, 1, 2, 4, 8). Table 4 shows the ROC scores for the corresponding ROC curves in Fig. 9. As illustrated by the FROC curves, the ROC score for the proposed method is better in terms of all candidates and close-to-vessel candidates. It also achieves a better score for regular, obvious and subtle candidates.

Fig. 9(f) shows that the proposed method can build a model that is generalizable across datasets. In this figure the black lines represent the performance of each respective model on the DIARETDB1 test set by training on the DIARETDB1 training set, while the blue curves represent the performance of each model on the DIARETDB1 test set after training on the MESSIDOR training set. It can be seen that the proposed method is much more generalizable than the Fleming technique since the performance does not deteriorate significantly even when the classifier is trained on a different dataset (training on MESSIDOR and testing DIARETDB1). This fact increases the confidence that overfitting does not occur on the model that was trained on the MESSIDOR dataset.

5.2. Feature analysis

The extended feature set of 70 features that we have used is based on features that have been applied in the literature. We have attempted to collect the most common features that were present in the literature.

However, a question which arises is whether all of the features are important features that contribute to the performance of the Tree ensemble classifier [45]. Some features may not contribute much information to the classifier and hence may be ignored. We have performed an experiment to rank the features according to the Predictor Importance. The Predictor Importance for a given attribute is calculated by computing the entropy (or Information Gain) for each tree in the Tree ensemble and then computing the average entropy for each tree. The predictor importance can be computed while building the Tree ensemble model and provides an indication of the importance of features. Fig. 10 shows the measured predictor importance for the 70 features in Fig. 10. We observe that there is varying importance for the features in the dataset.

In general by visualizing the graph it is observed that there are some important features for each category of features. It is observed that the Gaussian 1D and the Shape features are in general less important than the rest of features (though visual observation). This does not imply that they should be ignored, however, since to decide which features need to be removed a feature selection method should be utilized, and this step is left for future work 46].

Table 5
The 5 most (a) and least (b) discriminative features according to the bagging feature importance measure.

Feature number	Feature description	Category
(a)		
65	The intensity range in the value channel	Intensity
66	Intensity range in the preprocessing channel	Intensity
6	Depth of candidate in the preprocessed image	Fleming
19	minor axis length	Shape
2	Major axis length	Fleming
(b)		
46	Mean candidate intensity in red channel	Intensity
36	Range in the hue channel	Intensity
12	3^{rd} moment invariant	Moment
14	5^{th} moment invariant	Moment
15	6^{th} moment invariant	Moment

The top 5 and least 5 features in terms of discriminative ability are listed in Table 5. Some interesting observations can be made based on this table. Firstly, it is observed that intensity features appear twice in the most discriminative list and also twice in the least discriminative list. The intensity features that appear in the most discriminative list are computed from the processing channel, suggesting that computing features from this channel will help produce discriminative features. Another observation is that 2 of 9 Fleming features appear as most discriminative. The minor axis length feature also appears to be in the list of most discriminative features. Interestingly, there is another Major axis length feature that appears in the shape feature category (feature 18). The difference between the major axis length in Fleming (2) and the shape feature list (18) is that the first is measured after fitting a paraboloid to each candidate whereas the latter does not fit a paraboloid. It is observed that feature 18 is ranked low in the graph whereas feature 2 is among the most discriminative features. This raises the question about whether they are both correlated features which causes the feature to be ranked high while the other being under ranked [46]. If that is the case, then one may utilize this fact and eliminate some of the Fleming features by substituting them with shape features that are more efficient since they are calculated at the pixel level.

A final remark about the least discriminative features is that 3 out of the least 5 discriminative features are moment features, which suggests that the use of moment features does not help in the classification process. The process of experimenting with feature elimination and selecting a smaller set of the 70 features is left for future work.

6. Conclusion

In this work a new approach for MA detection is proposed. The new approach is based on Fleming's method. The proposed method relies on using a Tree ensemble classifier (ensemble classifier with bagging). The proposed method uses an extensive set of 70 features in order to perform the classification. A new public dataset of MA groundtruths is introduced based on the public MESSIDOR dataset. This set of groundtruths for 32 images is categorised according to MA appearance and closeness to blood vessels. The proposed method is evaluated using two datasets: including the new MESSIDOR dataset and DIARETDB1 (v2.1) dataset. The proposed method is compared to Fleming's method and another variant of Fleming. Results show that the proposed method is superior in terms of eliminating false positives (while maintaining the same sensitivity as the other methods) from the images and this is reflected in the plotted Free-Receiver Operating Curves (FROC). Furthermore, results show that the Tree ensemble classifier produces a model that is generalizable across datasets – this if verified by measuring error of the model trained on the MESSIDOR dataset on the DIARETDB1 dataset. The importance of the features is discussed to identify the most discriminative features among the 70 features. Feature selection for the reduction of the feature set is left for future work. The purpose of feature selection would be to increase the algorithm efficiency and reduce the chances of classifier overfitting. A summary of the performance of the algorithm on both MESSIDOR and DIARETDB1 is presented and areas which can be optimized are discussed.

Acknowledgements

This work is funded by an internal Scholarship, awarded by the Faculty of Science, Engineering and Computing, Kingston University.

References

[1] Lazar I, Hajdu A. Retinal microaneurysm detection through local rotating cross-section profile analysis. IEEE Trans Med Imaging Feb. 2013;32(2):400–7.
[2] Lazar I, Hajdu A. Microaneurysm detection in retinal images using a rotating cross-section based model. In: 2011 IEEE international symposium on biomedical imaging: from nano to macro; 2011. p. 1405–9.
[3] Júnior SB, Welfer D. Automatic detection of microaneurysms and hemorrhages in color eye fundus images. Int J Comput Sci Inf Technol Oct. 2013;5(5):21–37.
[4] Li Q, Lu R, Miao S, You J. Detection of microaneurysms in color retinal images using multi-orientation sum of matched filter. In: Proc. of the 3rd international conference on multimedia technology; 2013.
[5] Giancardo L, Mériaudeau F, Karnowski TP, Tobin KW, Li Y, Chaum E. Microaneurysms detection with the radon cliff operator in retinal fundus images. In: SPIE medical imaging; 2010. 76230U.
[6] Giancardo L, Meriaudeau F, Karnowski TP, Li Y, Tobin Jr KW, Chaum E. Microaneurysm detection with radon transform-based classification on retina images. In: Engineering in medicine and biology society, EMBC, 2011 annual international conference of the IEEE; 2011. p. 5939–42.
[7] Sánchez CI, Hornero R, Mayo A, García M. Mixture model-based clustering and logistic regression for automatic detection of microaneurysms in retinal images. In: SPIE medical imaging, vol. 7260; 2009. p. 72601M–72601M–8.
[8] Spencer T, Olson JA, McHardy KC, Sharp PF, Forrester JV. An image-processing strategy for the segmentation and quantification of microaneurysms in fluorescein angiograms of the ocular fundus. Comput Biomed Res Aug. 1996;29(4):284–302.
[9] Cree MJ, Olson JA, McHardy KC, Sharp PF, Forrester JV. A fully automated comparative microaneurysm digital detection system. Eye 1997;11:622–8.
[10] Hipwell JH, Strachan F, Olson JA, McHardy KC, Sharp PF, Forrester JV. Automated detection of microaneurysms in digital red-free photographs: a diabetic retinopathy screening tool. Diabet Med Aug. 2000;17(8):588–94.
[11] Streeter L, Cree MJ. Microaneurysm detection in colour fundus images. In: Image vision comput. New Zealand; 2003. p. 280–4.
[12] Fleming AD, Philip S, Goatman KA, Olson JA, Sharp PF. Automated microaneurysm detection using local contrast normalization and local vessel detection. IEEE Trans Med Imaging Sep. 2006;25(9):1223–32.
[13] Zhang B, Wu X, You J, Li Q, Karray F. Detection of microaneurysms using multi-scale correlation coefficients. Pattern Recognit Jun. 2010;43(6):2237–48.
[14] Wu J, Xin J, Hong L, You J, Zheng N. New hierarchical approach for microaneurysms detection with matched filter and machine learning. In: Engineering in medicine and biology society (EMBC), 2015 37th annual international conference of the IEEE; 2015. p. 4322–5.
[15] Quellec G, Lamard M, Josselin PM, Cazuguel G, Cochener B, Roux C. Optimal wavelet transform for the detection of microaneurysms in retina photographs. IEEE Trans Med Imaging Sep. 2008;27(9):1230–41.
[16] Mizutani A, Muramatsu C, Hatanaka Y, Suemori S, Hara T, Fujita H. Automated microaneurysm detection method based on double ring filter in retinal fundus images. 200972601N.
[17] Sinthanayothin C, et al. Automated detection of diabetic retinopathy on digital fundus images. Diabet Med Feb. 2002;19(2):105–12.
[18] Abdelazeem S. Micro-aneurysm detection using vessels removal and circular hough transform. In: Proceedings of the nineteenth national radio science conference, (NRSC 2002); 2002. p. 421–6.
[19] Inoue T, Hatanaka Y, Okumura S, Muramatsu C, Fujita H. Automated microaneurysm detection method based on eigenvalue analysis using hessian matrix in retinal fundus images. In: Engineering in medicine and biology society (EMBC), 2013 35th annual international conference of the IEEE; 2013. p. 5873–6.
[20] Srivastava R, Wong DW, Duan L, Liu J, Wong TY. Red lesion detection in retinal fundus images using Frangi-based filters. In: Engineering in medicine and biology society (EMBC), 2015 37th annual international conference of the IEEE; 2015. p. 5663–6.
[21] Adal KM, Sidibé D, Ali S, Chaum E, Karnowski TP, Mériaudeau F. Automated detection of microaneurysms using scale-adapted blob analysis and semi-supervised learning. Comput Methods Programs Biomed Apr. 2014;114(1):1–10.
[22] Niemeijer M, van Ginneken B, Staal J, Suttorp-Schulten MSA, Abramoff MD. Automatic detection of red lesions in digital color fundus photographs. IEEE Trans Med Imaging May 2005;24(5):584–92.
[23] Sopharak A, Uyyanonvara B, Barman S. Automatic microaneurysm detection from non-dilated diabetic retinopathy retinal images using mathematical morphology methods. IAENG Int J Comput Sci 2011;38(3):295–301.
[24] Rocha A, Carvalho T, Jelinek HF, Goldenstein S, Wainer J. Points of interest and visual dictionaries for automatic retinal lesion detection. IEEE Trans Biomed Eng Aug. 2012;59(8):2244–53.

[25] Sopharak A, Uyyanonvara B, Barman S. Simple hybrid method for fine microaneurysm detection from non-dilated diabetic retinopathy retinal images. Comput Med Imaging Graph 2013;37(5):394–402.
[26] Akram MU, Khalid S, Khan SA. Identification and classification of microaneurysms for early detection of diabetic retinopathy. Pattern Recognit Jan. 2013;46(1):107–16.
[27] Ram K, Joshi GD, Sivaswamy J. A successive clutter-rejection-based approach for early detection of diabetic retinopathy. IEEE Trans Biomed Eng Mar. 2011;58(3):664–73.
[28] Rosas-Romero R, Martínez-Carballido J, Hernández-Capistrán J, Uribe-Valencia LJ. A method to assist in the diagnosis of early diabetic retinopathy: image processing applied to detection of microaneurysms in fundus images. Comput Med Imaging Graph Sep. 2015;44:41–53.
[29] Haloi M. Improved microaneurysm detection using deep neural networks. arXiv preprint arXiv:1505.04424. 2015.
[30] García M, López MI, Álvarez D, Hornero R. Assessment of four neural network based classifiers to automatically detect red lesions in retinal images. Med Eng Phys Dec. 2010;32(10):1085–93.
[31] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in neural information processing systems, vol. 25. Curran Associates, Inc; 2012. p. 1097–105.
[32] Habib M, Welikala RA, Hoppe A, Owen CG, Rudnicka AR, Barman SA. Microaneurysm detection in retinal images using an ensemble classifier. In: Presented at the the sixth international conference on image processing theory, tools and applications, Oulu, Finland; 2016.
[33] Pizer SM, et al. Adaptive histogram equalization and its variations. Comput Vis Graph Image Process Sep. 1987;39(3):355–68.
[34] Credit O. Bootstrap-inspired techniques in computational intelligence. IEEE signal Process Mag 2007;1053(5888/07).
[35] Polikar R. Ensemble based systems in decision making. IEEE Circuits Syst Mag 2006;6(3):21–45. Third.
[36] Mitchell TM. Machine learning. first ed. New York, NY, USA: McGraw-Hill, Inc.; 1997.
[37] Breiman L. Bagging predictors. Mach Learn Aug. 1996;24(2):123–40.
[38] Fraz MM, et al. An ensemble classification-based approach applied to retinal blood vessel segmentation. IEEE Trans Biomed Eng Sep. 2012;59(9):2538–48.
[39] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning, vol. 103. New York, NY: Springer New York; 2013.
[40] Hu M-K. Visual pattern recognition by moment invariants. IRE Trans Inf Theory Feb. 1962;8(2):179–87.
[41] Decencière E, et al. Feedback on a publicly distributed image database: the messidor database. Image Anal Stereol Aug. 2014;33(3):231–4.
[42] Kauppi T, et al. Constructing benchmark databases and protocols for medical image analysis: diabetic retinopathy. Comput Math Methods Med Jun. 2013;2013:e368514.
[43] Niemeijer M, et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans Med Imaging Jan. 2010;29(1):185–95.
[44] Fraz MM, Welikala RA, Rudnicka AR, Owen CG, Strachan DP, Barman SA. QUARTZ: quantitative analysis of retinal vessel topology and size – an automated system for quantification of retinal vessels morphology. Expert Syst Appl Nov. 2015;42(20):7221–34.
[45] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3(Mar):1157–82.
[46] Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemom Intelligent Laboratory Syst Sep. 2006;83(2):83–90.