Hindawi Publishing Corporation
Journal of Ophthalmology
Volume 2016, Article ID 5893601, 11 pages
http://dx.doi.org/10.1155/2016/5893601

Research Article

Automatic Screening and Grading of Age-Related Macular Degeneration from Texture Analysis of Fundus Images

Thanh Vân Phan,[1,2] Lama Seoud,[3] Hadi Chakor,[3] and Farida Cheriet[4]

[1]*Biomedical Engineering Institute of École Polytechnique de Montréal, Montréal, QC, Canada H3C 3A7*
[2]*Université Libre de Bruxelles, 1050 Brussels, Belgium*
[3]*Diagnos Inc., Brossard, QC, Canada J4Z 1A7*
[4]*Department of Computer and Software Engineering of École Polytechnique de Montréal, Montréal, QC, Canada H3C 3A7*

Correspondence should be addressed to Lama Seoud; lseoud@diagnos.ca

Received 27 November 2015; Revised 29 February 2016; Accepted 21 March 2016

Academic Editor: Xinjian Chen

Age-related macular degeneration (AMD) is a disease which causes visual deficien y and irreversible blindness to the elderly. In this paper, an automatic classific tion method for AMD is proposed to perform robust and reproducible assessments in a telemedicine context. First, a study was carried out to highlight the most relevant features for AMD characterization based on texture, color, and visual context in fundus images. A support vector machine and a random forest were used to classify images according to the different AMD stages following the AREDS protocol and to evaluate the features' relevance. Experiments were conducted on a database of 279 fundus images coming from a telemedicine platform. Th results demonstrate that local binary patterns in multiresolution are the most relevant for AMD classific tion, regardless of the classifie used. Depending on the classific tion task, our method achieves promising performances with areas under the ROC curve between 0.739 and 0.874 for screening and between 0.469 and 0.685 for grading. Moreover, the proposed automatic AMD classific tion system is robust with respect to image quality.

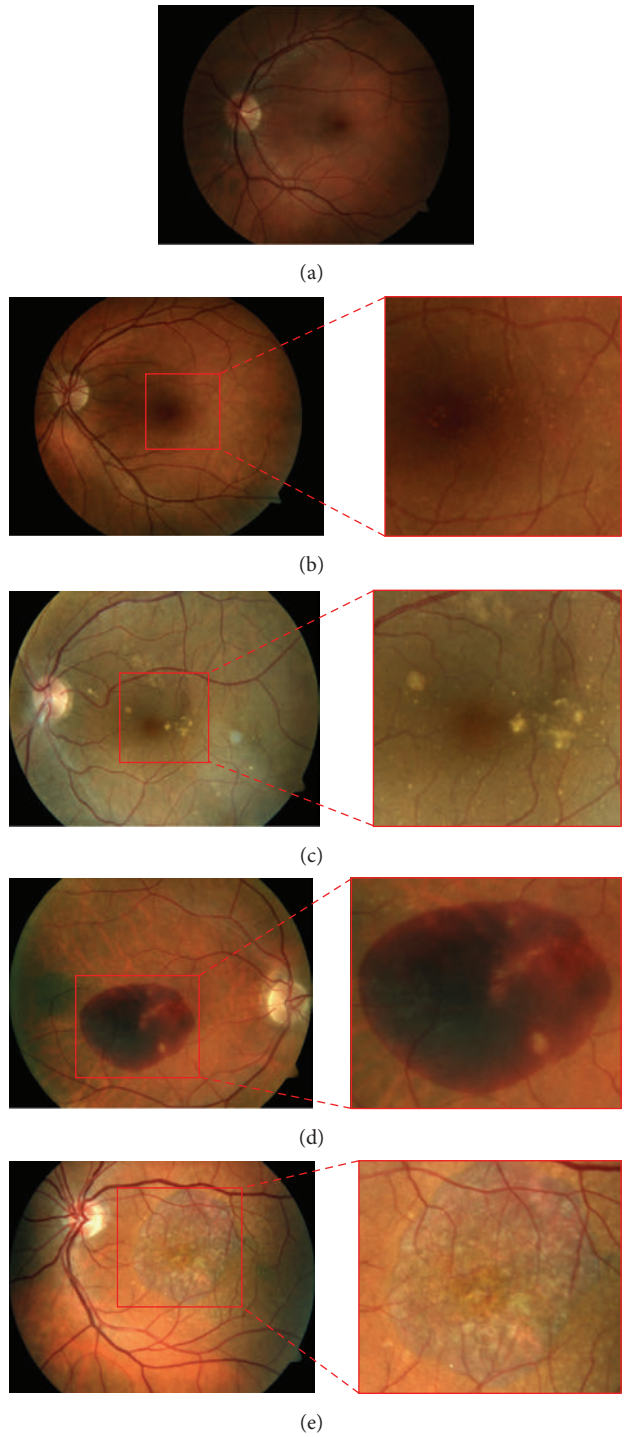
1. Introduction

Age-related macular degeneration (AMD) is the main cause of visual deficien y and irreversible blindness in the elderly in Western countries [1]. It combines a variety of disorders aff cting the macula. The early stage of AMD is asymptomatic, but small lesions, called drusen, can be revealed through examination of the retina. An increase in the size or number of drusen is a sign of the progression of the disease, leading eventually to the presence of hemorrhages (wet AMD) or to the development of geographic atrophy (late dry AMD). Th Age-Related Eye Disease Study (AREDS) [2] proposed a simplifie AMD clinical classific tion based on its stages. It comprises four categories which are illustrated in Figure 1:non-AMD {1}, mild {2}, moderate {3}, and advanced {4} AMD.

Currently, there is no approved treatment to recover from AMD. However, treatments to slow its progression exist and are different depending on the stage of the disease. These include prevention of oxidative damage, a treatment strategy based on supplements containing lutein, zeaxanthin, omega-3, vitamins C and E, and zinc, recommended for early stages [2, 3], while anti-VEGF therapy or surgical operations are used for more advanced stages [4].

With an aging population, there is urgent need for routine retinal examinations for early detection of AMD and for long-term follow-up strategies. Telescreening using fundus imaging has been extensively used for conditions like diabetic retinopathy [5, 6]. However, for AMD, it is still in its infancy. Combined with a telemedicine platform, automatic screening and grading from fundus images offer many inherent advantages. Th y allow clinicians to monitor susceptible individuals from an early age and to carry out preventive treatment.

Previous works focus mostly on dry AMD screening, based on the detection and quantific tion of drusen in fundus images [7]. Th drusen segmentation techniques are



Figur e 1:Images of macula area for different AMD categories: (a) healthy case in category {1}, (b) category {2} with hard drusen, (c) category {3} with soft drusen, and (d) category {4} with hemorrhages and (e) with geographic atrophy.

categorized into methods based on either texture analysis, thresholding, clustering, edge detection, or template matching. A number of texture-based methods use Gabor filters [8], wavelet transform [9, 10], amplitude and frequency modulation [11, 12], statistical structure information [13], or gray-level cooccurrence matrix [14]. The segmentation is based on the response of drusen to the applied texture method, which is assumed to be different from the response of the background. Th esholding-based methods aim to fin the appropriate threshold for separating the drusen from the background. Thi threshold can be set empirically [15] or automatically with Otsu's method [16]. Some image preprocessing is performed before thresholding using median or Gaussian filters [17], homomorphic filters [18], or

morphological operations [19]. Methods based on clustering are used for AMD phenotype-genotype correlation [20] or for identifying AMD [21]. Drusen segmentation can also be achieved through edge detection by identifying abrupt intensity variations using gradient or Laplacian filters [22]. Finally, template matching methods use circular or Gaussian templates [23] to model and detect drusen using similarity measurements.

Other methods first detect drusen regions and a classification based on drusen features, using, for example, linear discriminant analysis, *k*-nearest neighbors, gentle boost, random forest, or support vector machine classifiers, is then performed for AMD screening or assessing the risk of progression to the advanced stage [24–26]. The results show good performance, comparable to trained human observers. However, drusen segmentation does not provide sufficient information for a complete AMD grading. In fact, in its advanced stages, drusen are often not observed, especially when there are large hemorrhages or atrophies. Moreover, even if these methods show high accuracy for hard drusen detection (up to 100%, with the best methods [12, 18]), the segmentation of soft drusen, which characterize the moderate cases, is highly challenging because of their diffuse shape [24, 25].

Other works focus on structures characterizing advanced stages, such as what is proposed in [27] which used machine learning for GA detection and segmentation. All these works on drusen and geographic atrophy detection and classification are useful for a deep analysis of specific stage of the disease. However, a combination of segmentation methods corresponding to each AMD structure may be computationally complex for screening and grading in a telemedicine context, where a large number of images must be analyzed.

To address these limitations, automatic AMD classification methods were performed based on directly computed image features, without prior segmentation. Kankanahalli et al. proposed a method based on visual context using SURF key points as features and a random forest (RF) for classification [28]. Different binary classifications such as {1&2} versus {3&4} or {1} versus {3} and a trinary classification ({1&2} versus {3} versus {4}) were considered to discriminate the moderate cases. Indeed, close attention to moderate cases is important because even though the patient still has adequate visual acuity, there is a considerable risk of progression to a more severe stage. The proposed method achieves a good accuracy (above 90%) for AMD severity classification. However, the evaluation was conducted on 2772 images out of 11344 available in the AREDS database (24.4% of the database), selected for their good quality. Since it was trained solely on good quality images, the classifier might not be as effective on images of lower quality. In a telemedicine context, in which the acquisition conditions are not always optimal, poor quality images are often encountered.

Prior preliminary studies [29, 30] conducted by our group for the evaluation of new features demonstrated promising results with local binary patterns (LBP) in multiresolution for AMD detection. However, the validation was conducted on small datasets and the different feature subsets were evaluated individually without considering any combination thereof. Moreover, these preliminary studies were limited to a binary classification aimed only at distinguishing images with and without AMD.

The aim of this paper is to propose and to evaluate an automatic approach for clinical AMD screening and grading in a telemedicine framework. Thus, it is important to develop a system which is robust to variable image quality. To do so, various features based on texture, color, and visual context were computed, evaluated for their relevance, and used to classify the images according to the different AREDS categories. The validation was performed on a highly heterogeneous set of 279 fundus images, acquired through an existing telemedicine platform (CARA for Computer-Aided Retina Analysis, Diagnos Inc., Canada). Additionally, the robustness of the classification system to poor quality images was evaluated.

The organization of the paper is as follows. In Section 2, the main steps of the proposed AMD classification method are described in detail. The experimental setup is explained in Section 3. The results are presented in Section 4, followed by a discussion in Section 5 and a conclusion in Section 6.

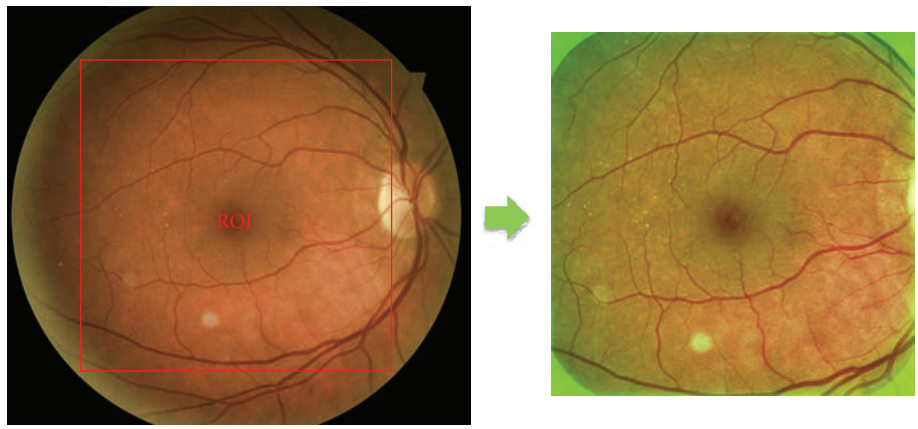
2. Materials and Methods

Fundus images acquired in a real screening context often show uneven illumination and poor contrast. To address these issues, a preprocessing step was required. Then, different features based on texture, color, and visual context were extracted to characterize the fundus image. Next, a classifier modeling step allowed us to measure the relevance of the features. Finally, two classifiers, SVM and RF, were tested on a database of 279 fundus images for performance assessment.

2.1. Image Preprocessing. Image normalization is required to correct the illumination drift introduced by the geometry of the eye and the bright flash of light used by the fundus camera. Contrast enhancement is also necessary to improve the information on details in the fundus images.

To perform these preprocessing steps, we used the same methodology as proposed in [28] for a fair comparison with their results. First, the region of interest (ROI) was defined as the square inscribed in the circle formed by the retina. Then, the green channel was extracted for preprocessing. A median filter with a kernel size of one-fourth the size of the ROI was applied in order to estimate the background illumination. The filtered image was then subtracted from the green channel of the original image. Finally, the green values were multiplied by 2 for contrast enhancement and shifted by the mean of their intensity range for visualization purposes (Figure 2).

2.2. Feature Extraction. Several features based on color, texture, and visual context were chosen because they proved to be effective in fundus image analysis. Color information is an intuitive feature, since AMD-related structures are characterized by specific colors. The texture and local gradient information also reflect the state of the retina. The image features considered in this study and their parameter settings are presented in the following subsections.



Figur e 2: Preprocessing method: ROI corresponding to the square inscribed in the circle formed by the retina and the result of preprocessing with illumination normalization and contrast enhancement in green channel.

2.2.1. Color Histograms. Blood vessels and lesions offer the highest contrast in the green channel. That is why most of the methods proposed in the literature for fundus image analysis focus solely on this channel. Still, even though the red channel is considered as saturated and with low contrast and the blue channel as very noisy in fundus images [31], all three color channels should be considered, especially to discriminate drusen from exudates, which are highly similar lesions but do not characterize the same disease [32]. In this study, the RGB and $L^*a^*b^*$ spaces were used. In RGB, the red and blue channels provide additional information to the green one. The $L^*a^*b^*$ space was also chosen because the luminance (lightness) and chrominance (colors) components are independent and color differences can be measured by a Euclidean distance.

We computed the 8-bin histograms for each channel from both color spaces as image features. The number of bins was set to 8 because there were no improvements in the results with a larger number of bins; thus we considered this suffici t for AMD classific tion.

2.2.2. Local Binary Patterns (LBP) in Multiresolution. To obtain the multiresolution information, a Lemarié wavelet transform was used with four levels of decomposition. For each level, an approximation coeffici t and three detail coeffici ts were computed, containing, respectively, the low resolution image (original size divided by two) and the high resolution details in the horizontal, vertical, and diagonal directions. From the original image and the 16 coeffici t images, textural information was extracted using LBP. This consisted in measuring the occurrence of local texture primitives, such as corners or edges. To do so, the LBP [33] was computed for each pixel of gray value g_c in a neighborhood of radius R and P neighbors with gray values g_p:

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} s\left(g_p - g_c\right) 2^P,$$

$$\text{With } s(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{Otherwise.} \end{cases} \tag{1}$$

In this study, the parameters were empirically set to $R = 1$ and $P = 8$. The magnitude component of the LBP [34] was also computed from the absolute differences of gray intensity between the central pixel and its neighbors $m_p = |g_p - g_c|$:

$$\mathrm{LBPM}_{P,R} = \sum_{p=0}^{P-1} t\left(m_p, c\right) 2^P,$$

$$\text{With } t(x,c) = \begin{cases} 1, & \text{if } x \geq c, \\ 0, & \text{Otherwise.} \end{cases} \tag{2}$$

Th threshold c was set to the image mean value.

From the sign and magnitude components of LBP, two histograms were computed by measuring the occurrence of the different patterns in the image. For each RGB color channel, LBP were computed and generated a vector of 2006 features.

2.2.3. Histograms of Oriented Gradients (HOG). Th histogram of oriented gradients is a feature generally used for edge detection [35], but it also contains local directional information which can be used for classific tion.

Th horizontal and vertical gradients were computed by applying a 1D point-centered derivative kernel [−1 0 1] to the color image. Then local histograms of the four main directions were constructed by dividing the RGB color image into 16×16 cells, with 2×2 cells for block normalization. The constructed vector contained 3600 features.

2.2.4. SURF Key Points. Starting from the hypothesis that all AMD manifestations (drusen and other lesions) were

represented in the subset of images presenting AMD, SURF key points were computed on that subset of images, previously converted into $L^*a^*b^*$. The key points were detected using ten octaves, three layers per octave, and a Hessian threshold of 600. Using the SURF features (sign of Laplacian, orientation, scale of detection, and strength of detected feature), a K-means clustering selected centroids on which the vocabulary was based to construct the features vector. For binary classific tions, K was set to 100, while for multiclass classific tions, K was set to 300. All parameters used to compute the SURF key points and to construct the vocabulary were set empirically. Once the vocabulary was established, a histogram was constructed by measuring the occurrence of key points depending on the nearest centroid. These features are implemented as proposed in [28] with unchanged parameters values.

2.3. Dimensionality Reduction and Features Importance. On one hand, a dimensionality reduction is necessary to avoid overfitting. Indeed, we have 6018 LBP features (2006 on each channel), 96 color histograms features, 3600 HOG features, and 100 or 300 SURF features. Considering the size of our dataset, a dimensionality reduction step is required before training a classifie . On the other hand, we believe that some of the features used might be more relevant than others in the discrimination between AMD stages. Thus, in order to evaluate features relevance and to select optimal subsets of features for classification, we used two approaches.

2.3.1. Fisher's Criterion. We determined the feature's relevance using the approach based on the Fisher criterion, which must be maximized [36]. This criterion measures the divergence between two classes i and j based on the estimate of their means μ and standard deviations σ when they are projected on the feature axis F:

$$D(F) = \frac{\left(\mu_i - \mu_j\right)^2}{\sigma_i^2 + \sigma_j^2}. \quad (3)$$

The maximum number of features for classifier modeling was set to one-tenth the number of training samples [37]. The fin l number of features retained was determined based on the best SVM performance obtained by varying the number of features and testing on validation samples.

2.3.2. Features Importance Using Gini Index. We also used the features' relevance assessment performed in random forest training [38]. We considered the mean decrease in Gini index to measure the features' relevance. This parameter measures the loss in Gini index on the out-of-bag samples when the feature is removed or permuted. The larger the decrease is, the more relevant the feature is. In this experiment, we used 3000 trees and we set the number of features to be selected at each node to 25 to ensure that all features are considered in the model to evaluate its importance.

2.4. Classifie Modeling. Two different classifie s were used in this study to verify if the choice of classifie has a signific nt impact on the results: a support vector machine (SVM) and a random forest (RF).

2.4.1. Support Vector Machine (SVM). The training of an SVM consists in finding the decision boundary that maximizes the margin that is the space between the elements nearest to the boundary [39].

In this study, a Gaussian kernel was chosen for the SVM because it is more effici t for systems with complex separations than a linear classifie . In addition, SVMs are useful for systems with a small number of samples because only the elements near the boundary, that is, the support vectors, contribute to the SVM modeling. For classifie modeling, the parameters to be set are γ, the parameter of the Gaussian kernel, and C, the number of elements accepted in the margin. The e parameters were set according to a performance assessment using a grid search strategy with 10-fold cross-validation to find the best pair of values in gamma = [0.001, 0.01, 0.1, 1, 10] and C = [1, 10, 50, 100].

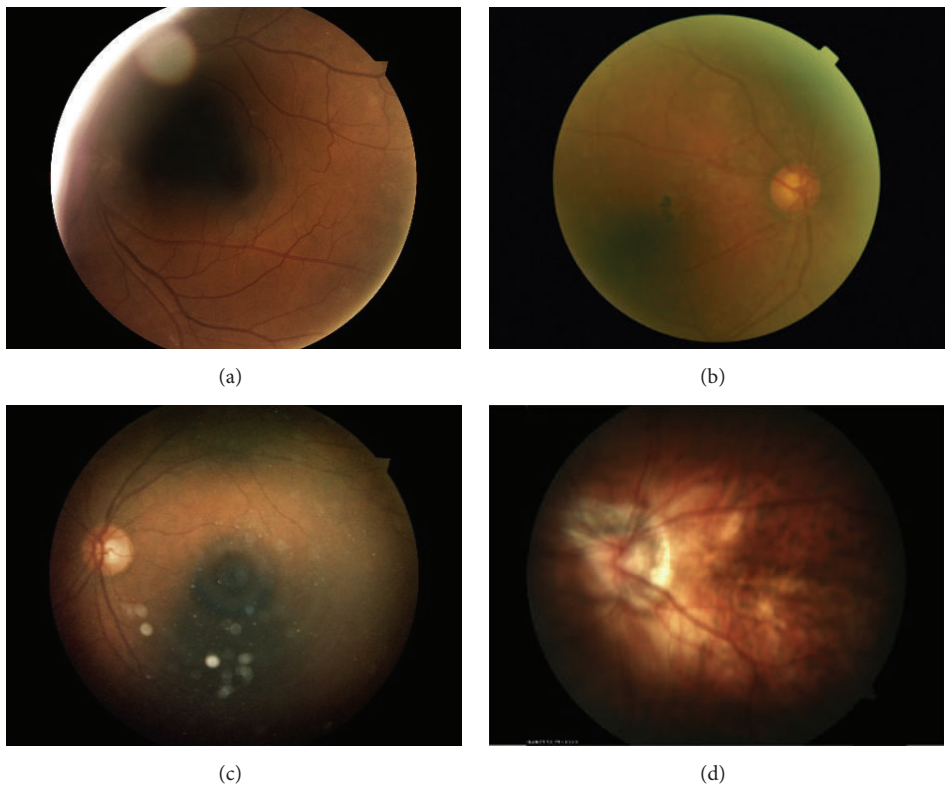
To consider more than two classes, we used the one-against-all approach. In the training phase, one SVM per class is constructed to discriminate the samples of that class from all the others. The label of a new observation is then determined by the SVM classifie that returns the highest value.

2.4.2. Random Forest (RF). The training of an RF consists in constructing decision trees, using randomly selected training samples and features. Then, the classific tion of new samples is determined by aggregating the votes of the different trees [40]. This method is quite simple to use since only two parameters need to be set: the number of features in the random subset at each tree node and the number of trees in the forest [41]. Th fi st parameter was set to the square root of the total number of features. The second parameter was set to 1,000 decision trees for binary classific tion and 2,500 decision trees for multiclass classific tion, such as what is proposed in [28].

3. Experimental Setup

3.1. Materials. The validation was conducted on a database of 279 images, all provided by Diagnos Inc. (Brossard, QC, Canada) using their telemedicine platform. The images were collected from clients in the United Arab Emirates, Mexico, Poland, India, and Canada. Th devices used for the acquisitions are different models of Zeiss, DRS, Topcon, and Canon retinal cameras. All the images are in JPEG compressed 10: 1format and acquired with a 45° field- f-view. Depending on the camera used, the size of the images varies between 1,400, 2,200, and 3,240 pixels along the diameter of the retinal images (circular imaged region excluding black background).

Depending on the acquisition conditions, the images vary in terms of resolution and illumination both of which affect the image quality [42]. Different artefacts, illustrated in Figure 3, can be encountered in fundus photography: shadows, intense refl ctions, specular refl ctions, blur, haze, or arcs. In this study, we used an automatic image quality assessment



Figur e 3: Examples of poor quality images: (a) shadows and intense refle tions, (b) haze, (c) arc and specular refle tions, and (d) blur.

Tabl e 1: Number of images in each AREDS category and for each image quality level.

Category	{1} Non-AMD	{2} Early	{3} Moderate	{4} Advanced
Good quality	50	43	24	22
Poor quality	29	36	41	34

method described in [43]. The algorithm determined if an image is of good or poor quality based on its measured color distribution and sharpness.

Two human graders were instructed to label the images into one of the four AREDS categories. Th fi st grader (Grader A) is an ophthalmologist with 10 years of experience working on fundus images. He has expertise in AREDS classific tion. Th second grader (Grader B) has 2 years of experience working on fundus images and was trained to classify fundus images following the simplifi d AMD classific tion proposed by the AREDS.

Th number of images in each AREDS category (as labeled by Grader B) and their distribution according to quality level are shown in Table 1.

3.2. Experiments

3.2.1. Dimensionality Reduction and Features Relevance. To reduce the feature space dimension, we used, on one hand, the feature selection based on Fisher's criterion and, on the other hand, the features' relevance assessment based on mean decrease of Gini index for each classific tion task. Then, we counted the number of selected features in each feature category to highlight the most relevant features for AMD classific tion.

3.2.2. Performance Assessment for Screening. To assess the performance of our method for AMD screening, we evaluated several binary classific tion tasks, using a 10-fold cross-validation approach. This consisted in taking one-tenth of the dataset as a testing set, and the rest was used to train the classifie . Th prediction result from this classific tion was kept and the process was repeated for all the elements. Receiver Operating Characteristic (ROC) curves were obtained by varying the threshold on the probabilities given by both classifie s (SVM and RF) and by reporting the sensitivity and specific ty corresponding to this threshold. The corresponding areas under the ROC curves (AUC) were then computed. We also tested statistically how the results are signific ntly different from a random classifie [44].

3.2.3. Performance Assessment for Grading. In the same way as for screening, the performance for AMD grading was assessed using a 10-fold cross-validation approach for multiclass classific tions using SVM and RF. The results were then compared to the intergrader variability. The e results are reported using the confusion matrix, the classific tion accuracy (number of elements that are well classified) and the weighted Kappa coeffici t [45].

Table 2: Number of selected features per category.

Classifications	Features selection	Features categories						
		LBP red	LBP green	LBP blue	RGB hist.	Lab hist.	HoG	SURF
All	None	2006	2006	2006	48	48	3600	100
1_234	Fisher	4	4	0	0	0	0	0
	RF Gini	92	114	27	1	1	31	0
12_34	Fisher	2	6	0	0	0	0	0
	RF Gini	63	79	18	0	0	17	0
12_3_4	Fisher	0	8	0	0	0	0	0
	RF Gini	74	94	23	1	1	23	0
1_23_4	Fisher	0	5	0	0	0	0	0
	RF Gini	82	106	25	1	1	29	0
1_2_3_4	Fisher	0	7	0	0	0	0	0
	RF Gini	92	114	29	1	1	31	0

3.2.4. Robustness to Image Quality. Selecting good quality images to train a classification system does not guarantee its efficiency for processing images of variable quality, for example, in a telemedicine context. To evaluate and to compare the robustness to variations in image quality, an assessment using only good quality images for training and poor quality images for testing was performed. In this experiment, we also performed SVM and RF training and testing using only the SURF features as proposed in [28] for ends of comparison.

Our overall approach for performance assessment aimed at determining the best solution for robust AMD classification.

4. Results

4.1. Features Relevance. The features relevance was evaluated for screening and grading to highlight the most relevant features for an automatic classification following the AREDS criteria. Table 2 shows the number of features selected according to Fisher's criterion and Gini index. For both features selection methods, LBP features are the most selected for any classification tasks, especially LBP features computed in green channel. These features are the most relevant for AMD classification.

It is also to be noted that SURF features are never selected by neither the Fisher based method nor the RF Gini method. It appears that these features are not the most relevant to discriminate between the different AMD stages.

4.2. Performance Assessment for Screening. The AMD classification for screening {1} versus {2&3&4} was assessed for both classifiers, with and without a features selection step (see Table 3). The best results were obtained with the features selected based on Gini index, with an AUC of 87.7% for SVM and an AUC of 86.9% for RF. In Figure 4, the specificity and sensitivity corresponding to mild {3}, moderate {3}, and severe {4} are presented along with the ROC curve. It shows that cases in categories {3} and {4} are better detected as AMD than category {2}.

In light of these results, we decided to assess the classification {1&2} versus {3&4}, since a large proportion of cases in category {2} were considered as non-AMD. This classification task corresponds to distinguishing cases that require treatment (moderate and advanced cases) from cases that are not at risk (healthy and mild cases). The performance is better than previously mentioned with an AUC of 89.9% for SVM and an AUC of 89.8% for RF.

Table 3: Performance assessment (AUC) for screening.

Classifier		SVM			RF		
Features selection		None	Fisher	Gini	None	Fisher	Gini
1_234	AUC	0.494	0.743*	0.877*	0.791*	0.812*	0.869*
12_34	AUC	0.491	0.879*	0.899*	0.867*	0.843*	0.898*

*: statistically different from random classifier (0.5 not included in 95% CI of AUC).

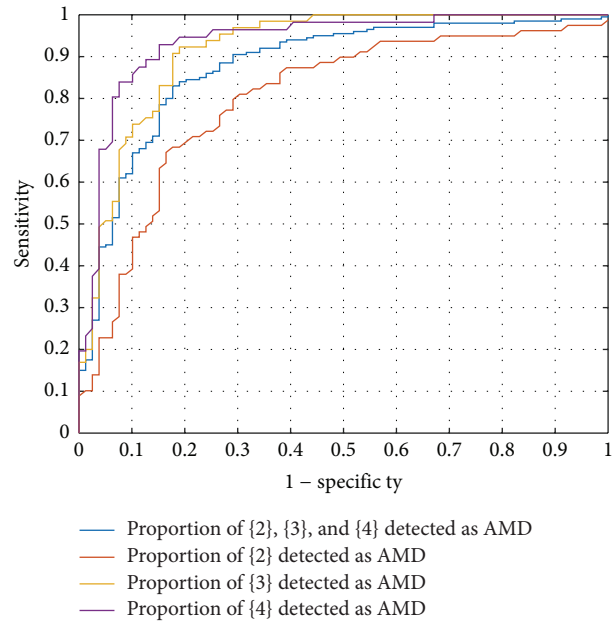


Figure 4: Screening performance for {1} versus {2&3&4} using SVM classifier and features selected using RF Gini.

Table 4: Performance assessment (accuracy) for grading.

Classifie		SVM			RF		
Features selection		None	Fisher	Gini	None	Fisher	Gini
12_3_4	Acc.	0.563	0.667	0.756	0.688	0.695	0.742
1_23_4	Acc.	0.516	0.581	0.724	0.642	0.613	0.699
1_2_3_4	Acc.	0.280	0.477	62.7	0.513	0.484	0.617

4.3. Performance Assessment for Grading. Th results of performance assessment for grading are shown in Table 4. For each classification task, the best results were obtained with the features selected based on Gini index and the SVM classifie . For the automatic classification according to AREDS protocol ({1} versus {2} versus {3} versus {4}), the method achieved an accuracy of 62.7%. Accuracies of 75.6% and 72.4% were obtained, respectively, for {1&2} versus {3} versus {4} and for {1} versus {2&3} versus {4}. The results demonstrate that the classific tion gives better performance when the number of categories to classify is lower.

Table 5 presents the confusion for {1} versus {2} versus {3} versus {4} using features selected by Gini index. Most of the misclassific tions happen between categories {1} and {2}. That explains why the performance was better when we considered {1&2} as one category. We also compared the results with intergrader variability. Th latter was assessed on a subset of 176 images annotated by both Graders A and B and measured with weighted Kappa coeffici t. The results (see Table 5) showed a weighted Kappa coeffici t of 73,6%, which corresponds to a substantial agreement between graders [45]. Th automatic method does not reach a performance on the same order as the intergrader variability.

However, we can notice that, even for graders, most disagreements happen between classes {1} and {2} and between {2} and {3}.

From these results, we also tested a classific tion in two steps. First, we classifie all images into three categories {1&2}, {3}, and {4}, since trinary classific tion gives better results. Then, the cases in {1&2} are classifie into {1} and {2}. Th results are shown in Table 6 and, indeed, improved with a weighted Kappa of 66.2% for SVM and of 61.0% for RF, which corresponds to a substantial agreement. For the SVM classifie , the weighted Kappa is in the 95% confidenc interval of the intergrader Kappa which means that there is no signific nt difference between the performance of the automatic SVM classifie and Grader B, when compared to Grader A.

4.4. Robustness to Variable Image Quality. Th robustness was assessed by measuring the performance of the system when trained with only good quality images and tested on poor quality images. We compared our results with the method proposed in [28] which is based solely on the SURF features as described in Section 2.2.4. Table 7 shows the robustness assessment for AMD screening. Th resulting AUCs are in the same range as in the 10-fold cross-validation on the whole dataset (Table 4). Table 8 shows the robustness assessment for AMD grading. Here, the classific tion accuracy decreases compared to the assessment by 10-fold cross-validation on the whole dataset (Table 5), yielding accuracies of 0.207–0.557 with SVM and 0.393–0.693 with RF.

5. Discussion

The main purpose of this paper was to propose an automatic system for clinical AMD screening and grading in a telemedicine framework and to evaluate its performance. This was achieved through a comparative study of different image features mainly based on texture, color, and visual context.

Th experiments revealed that the best results for AMD screening and grading were obtained with LBP in multiresolution applied to the green channel. The e features were considered as the most relevant for AMD classification and were favored by the Fisher criterion and Gini index. Th present work confi ms that these features are robust with respect to image quality, as suggested in our prior studies [29, 30], and extends those results from AMD detection to AMD severity classific tion. Even with small learning samples, the systems using SVM classifie and features selected by Gini index achieved AUCs between 0.877 and 0.899 for AMD screening, which is especially good considering the large proportion of poor quality images (50.2% of the database). Our best result for AMD grading was an accuracy of 75.6% for the trinary classific tion task {1&2} versus {3} versus {4}. Th automatic grading following AREDS protocol was in the same order as intergrader variability while using SVM and features selected based on Gini index.

LBP is a powerful tool for characterizing the texture and that is why these features are the most suitable for this application. First, a local characterization of the texture is more effective than a global feature such as color histograms. Then, LBP measures the proportions of the diff rent uniform patterns contained in the image (such as edges, borders, or points), which seem to be more informative than the local gradients computed in HOG or the SURF key point features. In fact, these latter features seem to be less robust to poor quality images, since they are based on detecting local maxima which can be sensitive to noise. Thus, LBP are the most reliable features taking into account the types of structures characterizing AMD images at different severity degrees. Finally, the multiresolution approach helps us to characterize the stage of the disease by identifying lesions at different scales. Indeed, a lesion detected at high resolution could correspond to large drusen or an atrophy, both being related to more advanced AMD stages.

We have proposed a method that is adapted to a real telemedicine context. This means that we processed images from variable quality levels, coming from diff rent locations and different cameras, whereas major studies on AMD in the literature have used homogeneous datasets. Furthermore, our results compare well against those of other methods. For AMD screening, a study carried out in [24] aimed to evaluate if cases were at low or high risk to progress to an advanced stage, based on drusen segmentation. Thei system achieved a Kappa coeffici t of 0.760–0.765. This is similar to our classific tion performance for {1&2} versus {3&4}, which obtained AUCs of 0.899. Nevertheless, it is difficult to compare these different methods one on one since there is no publicly available database for AMD grading containing fundus images labeled according to AREDS protocol.

Table 5: Confusion matrix in percentage for grading ({1} versus {2} versus {3} versus {4}).

%	SVM (Gini)				RF (Gini)				Grader B			
Nb img	279				279				176			
Grader A	1	2	3	4	1	2	3	4	1	2	3	4
1	20.1	6.8	1.1	0.4	19.7	6.5	1.4	0.7	31.2	9.5	0.6	0.0
2	6.5	15.8	4.7	1.4	7.2	16.5	2.9	1.8	4.5	19.3	6.2	0.6
3	1.4	4.7	13.3	3.9	2.2	5.7	13.3	2.1	0.0	3.4	7.4	2.8
4	0.7	0.7	5.0	13.6	0.7	2.2	5.0	12.2	0.0	1.1	1.1	13.1
Accuracy	62.7				61.6				71.5			
Weighted K (95% CI)	63.7 (57.3–70.2) Substantial				59.4 (52.3–66.5) Moderate				73.6 (66.1–80.2) Substantial			

Table 6: Confusion matrix in percentage for grading in two steps ({1&2} versus {3} versus {4} and then {1} versus {2}).

%	SVM (Gini)				RF (Gini)				Grader B			
Nb img	279				279				176			
Grader A	1	2	3	4	1	2	3	4	1	2	3	4
1	22.6	4.3	1.1	0.3	21.9	5.0	0.7	0.7	31.2	9.5	0.6	0.0
2	4.3	18.3	4.3	1.4	4.7	19.7	2.5	1.4	4.5	19.3	6.2	0.6
3	1.8	4.7	12.2	4.6	3.6	7.1	10.0	2.5	0.0	3.4	7.4	2.8
4	0.7	1.1	5.0	13.3	1.1	1.8	4.7	12.5	0.0	1.1	1.1	13.1
Accuracy	66.3				64.2				71.5			
Weighted K (95% CI)	66.2 (59.7–72.6) Substantial				61.0 (53.8–68.1) Substantial				73.6 (66.1–80.2) Substantial			

Table 7: Quality robustness assessment (AUC) for screening.

Classifie		SVM				RF			
Features selection		None	SURF [28]	Fisher	RF Gini	None	SURF [28]	Fisher	RF Gini
1_234	AUC	0.500	0.500	0.588	0.874*	0.797*	0.436	0.807*	0.889*
12_34	AUC	0.500	0.530	0.882*	0.812*	0.819*	0.472	0.875*	0.816*

*: statistically different from random classifie (0.5 not included in 95% CI of AUC).

Table 8: Quality robustness assessment (accuracy) for grading.

Classifie		SVM				RF			
Features selection		None	SURF [28]	Fisher	RF Gini	None	SURF [28]	Fisher	RF Gini
12_3_4	Acc.	0.466	0.464	0.529	0.557	0.607	0.493	0.571	0.586
1_23_4	Acc.	0.550	0.550	0.550	0.550	0.643	0.329	0.557	0.693
1_2_3_4	Acc.	0.207	0.300	0.450	0.507	0.486	0.350	0.393	0.521

For AMD grading, the method proposed in [28] reports an accuracy of 91.5% for classifying {1&2} versus {3} versus {4} on selected images of good quality. Our method achieved an accuracy of 75.6%, which is signific ntly lower; however all images were processed including images of poor quality. To support that furthermore, the experiment on robustness to image quality clearly demonstrates that AMD screening and grading using SURF features as proposed in [28] is not applicable in a telemedicine setting where image quality is not always guaranteed.

Our method demonstrates considerable robustness with respect to image quality. In a telemedicine context, where acquisition conditions are not strictly controlled, to only select good quality images is not adequate for AMD evaluation because we want a maximum of cases to be handled. To demonstrate the robustness to image quality, we assessed the classific tion systems performance by training them on good quality images and testing them on poor quality ones. Our system still performed well, presenting results of the same order as the ones obtained in the leave-one-out cross-validation.

In regard to the classification tasks, it is recommended to use the classific tion {1&2} versus {3&4} for AMD screening, which presented a better result using our method. Th clinical

rationale for this classification is to distinguish cases that need to be treated from those that are not at risk. We can notice that our method tends to consider a certain proportion of category {2} cases as non-AMD. For grading, a better classification performance is obtained for a two-step classification, starting with {1&2} versus {3} versus {4} classific tion and then performing a {1} versus {2} classific tion.

Our database contained a relatively small number of samples in each category. This may be the reason why a good performance for grading could not be demonstrated in this study. Moreover, even the human graders had some difficulty agreeing on the database's labeling, with an intergrader weighted Kappa of 0.736. A validation on a larger database could improve the grading results.

Future work will focus on the preprocessing step. In fact, in this study, we used a preprocessing procedure introduced in [28] for ends of comparison. Nevertheless, several improvements could be made to it. Th background illumination was estimated with a median filter, but the convolution with a high resolution image has a large computational cost. Thi aspect could be improved by using spectral filtering instead. Also, our previous work demonstrated that a local analysis focused on the macular area can improve the system performance. Indeed, features of AMD are mainly located in this area. This idea could be further explored by using an automatic detection of the macular region based on the detection of the fovea and the radius of the optic disc.

6. Conclusion

We have developed and validated an automatic clinical classific tion system for AMD screening and grading in a telemedicine context. The validation of our method reveals promising results in terms of robustness to image quality and accuracy for different AMD severity classific tion tasks. Our experimental results highlight the discriminating strength of LBP features over other tested features, whether the classifie is an RF or an SVM. Further validation must be conducted on a database containing more samples in each category in order to confi m these results. Nevertheless, the proposed approach represents an important step toward providing a reliable AMD diagnostic tool for patient monitoring and for clinical trials. Early AMD detection can facilitate timely access to treatment and consequently improve the therapeutic outcome.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

Th authors would like to acknowledge the contribution of Philippe Debanné for revising this paper. This work was funded by Diagnos Inc. and Natural Sciences and Engineering Research Council of Canada.

References

[1] D. T. Kasuga, Y. Chen, and K. Zhang, "Genetics of age-related degeneration," in *Age-Related Macular Degeneration Diagnosis and Treatment*, pp. 1–14, Springer, Philadelphia, Pa, USA, 2011.

[2] Age-Related Eye Disease Study Research Group (AREDS), "The age-related eye disease study severity scale for age-related macular degeneration," *Archives of Ophthalmology*, vol. 123, no. 11, pp. 1484–1498, 2005.

[3] A. D. Meleth, V. R. Raiji, N. Krishnadev, and E. Y. Chew, "Therapy of nonexudative age-related macular degeneration," in *Age-Related Macular Degeneration Diagnosis and Treatment*, A. C. Ho and C. D. Regillo, Eds., pp. 65–78, Springer, New York, NY, USA, 2011.

[4] F. M. Penha and P. J. Rosenfeld, "Management of neovascular AMD," in *Age-Related Macular Degeneration Diagnosis and Treatment*, A. C. Ho and C. D. Regillo, Eds., pp. 79–98, Springer, New York, NY, USA, 2011.

[5] C. M. Oliveira, L. M. Cristóvão, M. L. Ribeiro, and J. R. Faria Abreu, "Improved automated screening of diabetic retinopathy," *Ophthalmologica*, vol. 226, no. 4, pp. 191–197, 2011.

[6] L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet, and J. M. Langlois, "Red lesion detection using dynamic shape features for diabetic retinopathy screening," *IEEE Transactions on Medical Imaging*, vol. 35, no. 4, pp. 1116–1126, 2016.

[7] Y. Kanagasingam, A. Bhuiyan, M. D. Abràmoff, R. T. Smith, L. Goldschmidt, and T. Y. Wong, "Progress on retinal image analysis for age related macular degeneration," *Progress in Retinal and Eye Research*, vol. 38, pp. 20–42, 2014.

[8] S. S. Parvathi and N. Devi, "Automatic drusen detection from colour retinal images," in *Proceedings of the International Conference on Computational Intelligence and Multimedia Applications*, pp. 377–381, IEEE, Sivakasi, India, December 2007.

[9] L. Brandon and A. Hoover, "Drusen detection in a retinal image using multi-level analysis," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, pp. 618–625, Springer, Berlin, Germany, 2003.

[10] R. Priya and P. Aruna, "Automated diagnosis of Age-related macular degeneration from color retinal fundus images," in *Proceedings of the 3rd International Conference on Electronics Computer Technology (ICECT '11)*, pp. 227–230, Kanyakumari, India, April 2011.

[11] E. S. Barriga, V. Murray, C. Agurto et al., "Multi-scale AM-FM for lesion phenotyping on age-related macular degeneration," in *Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems (CBMS '09)*, Albuquerque, NM, USA, August 2009.

[12] C. Agurto, E. Barriga, V. Murray et al., "Automatic detection of diabetic retinopathy and age-related macular degeneration in digital fundus images," *Retina*, vol. 52, no. 8, pp. 5862–5871, 2011.

[13] C. Köse, U. Şevik, O. Gençalioğlu, C. İkibaş, and T. Kayıkıçıoğlu, "A statistical segmentation method for measuring age-related macular degeneration in retinal fundus images," *Journal of Medical Systems*, vol. 34, no. 1, pp. 1–13, 2010.

[14] A. R. Prasath and M. M. Ramya, "Detection of macular drusen based on texture descriptors," *Research Journal of Information Technology*, vol. 7, no. 1, pp. 70–79, 2015.

[15] W. H. Morgan, R. L. Cooper, I. J. Constable, and R. H. Eikelboom, "Automated extraction and quantific tion of macular drusen from fundal photographs," *Australian and New Zealand Journal of Ophthalmology*, vol. 22, no. 1, pp. 7–12, 1994.

[16] R. T. Smith, J. K. Chan, T. Nagasaki et al., "Automated detection of macular drusen using geometric background leveling and threshold selection," *Archives of Ophthalmology*, vol. 123, no. 2, pp. 200–206, 2005.

[17] P. Soliz, M. P. Wilson, S. C. Nemeth, and P. Nguyen, "Computer-aided methods for quantitative assessment of longitudinal changes in retinal images presenting with maculopathy," in *Proceedings of the SPIE 4681 Medical Imaging, International Society for Optics and Photonics*, pp. 159–170, 2002.

[18] K. Rapantzikos, M. Zervakis, and K. Balas, "Detection and segmentation of drusen deposits on human retina: potential in the diagnosis of age-related macular degeneration," *Medical Image Analysis*, vol. 7, no. 1, pp. 95–108, 2003.

[19] Z. Liang, D. W. K. Wong, J. Liu, K. L. Chan, and T. Y. Wong, "Towards automatic detection of age-related macular degeneration in retinal fundus images," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '10)*, pp. 4100–4103, IEEE, Buenos Aires, Argentina, September 2010.

[20] G. Quellec, S. R. Russell, and M. D. Abramoff, "Optimal filter framework for automated, instantaneous detection of lesions in retinal images," *IEEE Transactions on Medical Imaging*, vol. 30, no. 2, pp. 523–533, 2011.

[21] M. Hanafi, A. Hijazi, F. Coenen, and Y. Zheng, "Retinal image classific tion for the screening of age-related macular degeneration," in *Proceedings of the SGAI International Conference on Artifici l Intelligence*, Cambridge, UK, December 2010.

[22] A. D. Mora, P. M. Vieira, A. Manivannan, and J. M. Fonseca, "Automated drusen detection in retinal images using analytical modelling algorithms," *BioMedical Engineering Online*, vol. 10, article 59, pp. 1–15,2011.

[23] B. Remeseiro, N. Barreira, D. Calvo, M. Ortega, and M. G. Penedo, "Automatic drusen detection from digital retinal images: AMD prevention," in *Computed Aided Systems Theory-EUROCAST*, pp. 187–194, Springer, Berlin, Germany, 2009.

[24] M. J. J. P. Van Grinsven, Y. T. E. Lechanteur, J. P. H. Van De Ven, B. Van Ginneken, T. Theelen, and C. I. Sánchez, "Automatic age-related macular degeneration detection and staging," in *Proceedings of the SPIE Medical Imaging 2013: Computer-Aided Diagnosis*, Orlando, Fla, USA, February 2013.

[25] M. U. Akram, S. Mujtaba, and A. Tariq, "Automated drusen segmentation in fundus images for diagnosing age related macular degeneration," in *Proceedings of the 10th International Conference on Electronics, Computer and Computation (ICECCO '13)*, pp. 17–20, Ankara, Turkey, November 2013.

[26] V. Sundaresan, K. Ram, K. Selvaraj, N. Joshi, and M. Sivaprakasam, "Adaptative super-candidate based approach for detection and classific tion of drusen retinal fundus images," in *Proceedings of the Ophthalmic Medical Image Analysis Second Internation Workshop (OMIA '15)*, pp. 81–88, Munich, Germany, 2015.

[27] A. K. Feeny, M. Tadarati, D. E. Freund, N. M. Bressler, and P. Burlina, "Automated segmentation of geographic atrophy of the retinal epithelium via random forests in AREDS color fundus images," *Computers in Biology and Medicine*, vol. 65, pp. 124–136, 2015.

[28] S. Kankanahalli, P. M. Burlina, Y. Wolfson, D. E. Freund, and N. M. Bressler, "Automated classific tion of severity of age-related macular degeneration from fundus photographs," *Investigative Ophthalmology & Visual Science*, vol. 54, no. 3, pp. 1789–1796, 2013.

[29] M. Garnier, T. Hurtut, H. B. Tahar, and F. Cheriet, "Automatic multiresolution age-related macular degeneration detection from fundus images," in *Medical Imaging: Computer-Aided Diagnosis*, vol. 9035 of *Proceedings of SPIE*, 2014.

[30] T. V. Phan, L. Seoud, and F. Cheriet, "Towards an automatic classific tion of age-related macular degeneration," in *Proceedings of the International Conference on Image Analysis and Recognition*, Niagara Falls, NY, USA, July 2015.

[31] T. Walter, P. Massin, A. Erginay, R. Ordonez, C. Jeulin, and J.-C. Klein, "Automatic detection of microaneurysms in color fundus images," *Medical Image Analysis*, vol. 11, no. 6, pp. 555–566, 2007.

[32] M. J. J. P. van Grinsven, A. Chakravarty, J. Sivaswamy, T. Thee len, B. van Ginneken, and C. I. Sanchez, "A Bag of Words approach for discriminating between retinal images containing exudates or drusen," in *Proceedings of the IEEE 10th International Symposium on Biomedical Imaging (ISBI '13)*, pp. 1444–1447, San Francisco, Calif, USA, April 2013.

[33] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classific tion with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[34] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classific tion," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1657–1663, 2010.

[35] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 886–893, IEEE, San Diego, Calif, USA, June 2005.

[36] R. O. Duda, P. E. Hart, and D. G. Stork, "Maximum-likelihood and Bayesian parameters estimation," in *Pattern Classific tion*, pp. 84–159, Wiley-Interscience, New York, NY, USA, 2nd edition, 2009.

[37] A. R. Webb, *Statistical Pattern Recognition*, John Wiley & Sons, 2003.

[38] L. Breiman, *Manual on Setting Up, Using and Understanding Random Forests V3.1*, University of California, Berkeley, Calif, USA, 2002.

[39] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, 2008.

[40] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41] A. Liaw and M. Wiener, "Classific tion and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[42] H. Bartling, P. Wanger, and L. Martin, "Automated quality evaluation of digital fundus photographs," *Acta Ophthalmologica*, vol. 87, no. 6, pp. 643–647, 2009.

[43] M. Fasih, J. M. P. Langlois, H. B. Tahar, and F. Cheriet, "Retinal image quality assessment using generic features," in *Proceedings of the SPIE 9035 Medical Imaging 2014: Computer-Aided Diagnosis*, San Diego, Calif, USA, February 2014.

[44] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[45] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: the kappa statistic," *Family Medicine*, vol. 37, no. 5, pp. 360–363, 2005.