This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics

DREAM: Diabetic Retinopathy Analysis using Machine Learning

Sohini Roychowdhury, *Student Member, IEEE,* Dara D. Koozekanani, *Member, IEEE* and Keshab K. Parhi *Fellow, IEEE*

***Abstract*—This paper presents a computer-aided screening system (DREAM) that analyzes fundus images with varying illumination and fields of view, and generates a severity grade for diabetic retinopathy (DR) using machine learning. Classifiers such as Gaussian Mixture Model (GMM), k-nearest neighbor (kNN), support vector machine (SVM), and AdaBoost are analyzed for classifying retinopathy lesions from non-lesions. GMM and kNN classifiers are found to be the best classifiers for bright and red lesion classification, respectively. A main contribution of this paper is the *reduction in the number of features* used for lesion classification by feature ranking using Adaboost where 30 top features are selected out of 78. A novel two-step *hierarchical classification* approach is proposed where the non-lesions or false positives are rejected in the first step. In the second step, the bright lesions are classified as hard exudates and cotton wool spots, and the red lesions are classified as hemorrhages and micro-aneurysms. This lesion classification problem deals with unbalanced data sets and SVM or combination classifiers derived from SVM using the Dempster-Shafer theory are found to incur more classification error than the GMM and kNN classifiers due to the data imbalance. The DR severity grading system is tested on 1200 images from the publicly available MESSIDOR data set. The DREAM system achieves 100% sensitivity, 53.16% specificity and 0.904 AUC, compared to the best reported 96% sensitivity, 51% specificity and 0.875 AUC, for classifying images as with or without DR. The feature reduction further reduces the average computation time for DR severity per image from 59.54 seconds to 3.46 seconds.**

***Index Terms*—bright lesions, classification, diabetic retinopathy, fundus image processing, red lesions, segmentation, severity grade.**

I. Introduction

According to a study by the American Diabetes Association, diabetic retinopathy (DR) had affected more than 4.4 million Americans of age 40 and older during 2005-2008, with almost 0.7 million (4.4% of those with diabetes) having advanced DR that could lead to severe vision loss [1]. Early detection and treatment of DR can provably decrease the risk of severe vision loss by over 90% [2]. Thus, there is a high consensus for the need of efficient and cost-effective DR screening systems [3]. Unfortunately almost 50% of diabetic patients in the United States currently do not undergo any form of documented screening exams in spite of the guidelines established by the American Diabetes Association (ADA) and the American Academy of Ophthalmology (AAO) [3]. Statistics show that 60% of the patients requiring laser surgery to prevent blindness do not receive treatment [2]. The major reasons for this screening and treatment gap include insufficient referrals, economic hindrances and insufficient access to proper eye care. Telemedicine, with distributed remote retinal fundus imaging and grading at either local primary care offices or centralized grading remotely by eye care specialists, has increased access to screening and follow-up necessary treatment [4].

Computer-aided screening systems have recently gained importance for increasing the feasibility of DR screening, and several algorithms have been developed for automated detection of lesions such as exudates, hemorrhages and micro-aneurysms. However, micro-aneurysms are considered to be the primary indicators of DR, and among the noteworthy prior works that aimed at detecting DR, in [5], color normalization and contrast enhancement was applied in the pre-processing steps on fundus images followed by fuzzy C-means clustering for image segmentation, and classification of candidate regions into bright lesions and non-lesions. In [6], micro-aneurysms were detected using template matching in wavelet sub-bands followed by thresholding. Another approach to detect micro-aneurysms based on multi-scale correlation filtering and dynamic thresholding was also developed in [7]. Another method based on extraction of AM-FM features of fundus images, followed by dimensionality reduction and hierarchical clustering, and the use of partial least square method to assign DR severity grade has been presented in [8].

The novel contribution of this paper is a fully automated, fast and accurate DR detection and grading system that can be used for automated screening for DR and treatment prioritization. So far an automated DR screening system, Medalytix [9], has been used for screening normal patients without DR from abnormal patients with DR on a local data set, with sensitivity in the range 97.4-99.3% on diabetic patients in Scotland. The screening outcome combined with manual analysis of the images that are classified as abnormal by the automated system has shown to reduce the clinical workload by more than 25% in Scotland [9]. Another automated DR screening system grades images from a local data set for unacceptable quality, referable, non-referable DR with sensitivity 84% and specificity 64% [3] [10]. Both these automated systems motivate the need for a fast and more accurate DR screening and prioritization system. The proposed DR detection system is capable of detecting lesions and generating a severity grade for non-proliferative DR (NPDR) in under 6 seconds with 100% sensitivity and 53% specificity. Such a system will ensure that

This research was supported in part by a grant from the Office of Vice President for Research at the University of Minnesota.

Sohini Roychowdhury and Keshab K. Parhi are with the Department of Electrical and Computer Engineering. Dara D. Koozekanani is with the Department of Ophthalmology and Visual Neuroscience. All the authors are affiliated to University of Minnesota.(email: roych005@umn.edu; dkoozeka@umn.edu; parhi@umn.edu)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics

no patient with DR is missed for follow-up treatment, and will be critical in prioritizing eye-care delivery measures for patients with highest DR severity.

The design of the proposed DREAM system is novel since it aims at outlining 3 separate DR detection stages, and minimizing the run-time complexity of each stage to ensure a fast detection system. We propose a feature size reduction followed by selection of lesion classifiers that have low computational complexity, such that the total time to detect retinopathy lesions and to generate a severity grade is less than 6 seconds for a fundus image. Additionally, the proposed system contains an initial image enhancement module in Stage 1 that enhances the contrast and edge sharpness in fundus images so that images are not rejected based on poor image quality. Existing automated DR screening systems such as the VA screening system [10] rejects about 18% of the images, while Medalytix [9] rejects about 5% of the images based on their bad image quality. The proposed system has not rejected any image so far on account of poor image quality and hence, it is more suitable for automated DR screening and prioritization.

This paper makes two major contributions. The first major contribution is identification of the top 30 features from a set of 78 features for classifying bright and red retinopathy lesions. We implement feature ranking based on the feature weights generated by AdaBoost [11]. We demonstrate how feature reduction reduces computational complexity of lesion classification by 94-97% in terms of mean processing time per image. The second contribution is a novel *two-step hierarchical binary classification* method that rejects false positives in the first step and in the second step, bright lesions are classified as cotton wool spots (CWS) or hard exudates (HE), and red lesions are classified as hemorrhages (HA) and micro-aneurysms (MA), respectively. This hierarchical classification method reduces the time complexity by 18-24% over a parallel classification method that trains separate classifiers for identifying CWS, HE, HA and MA from false positives. While identifying the optimal bright and red lesion classifiers, we analyze a variety of feature based classifiers such as GMM, kNN, SVM, AdaBoost, and combinational classifiers, and select an optimal classifier set that ensures high specificity for bright lesion classification and high sensitivity for red lesion classification, and low computational time complexity. In our case, we found that GMM is a preferred choice of classifier for detecting bright lesions, and kNN is a preferred choice for detecting red lesions.

The algorithms used in Stage 1 for image segmentation, detection of optic disc, blood vasculature and one-step classification of red and bright lesions in Stage 2 for one public data set have been presented in our earlier work [12]. In this work, we extend our analysis to construct a complete 3-stage automated system that not only detects retinopathy lesions, but also generates a DR severity grade for every fundus image. This paper introduces a novel two-step hierarchical classification method in Stage 2, it analyzes the importance of feature reduction and feature ranking, and it combines the number of lesions in Stage 3 to generate a DR severity measure and tests the performance of the proposed automated DR detection system on another public data set.

The organization of this paper is as follows. In Section II, the functionalities of each of the three-stages of the proposed DR screening system is mathematically defined. Section III describes each of the three stages in separate sub sections. Results regarding the performance of lesion classification and DR severity grading, time complexity analysis and comparison of the proposed system with prior works are presented in Section IV followed by concluding remarks and discussion in Section V.

II. Method and Materials

The three-stage algorithm to automatically detect and grade the severity of diabetic retinopathy using ophthalmic fundus images is shown in Fig. 1. For each fundus image in JPEG format, the green plane (I_{green}) is used for information extraction.

In Stage 1, a minimum-intensity maximum-solidity (MinIMaS) algorithm [12] is invoked to detect the regions corresponding to the optic disc (R_{OD}) and vasculature (R_{vasc}) as the image background from I_{green}. Next, candidate regions corresponding to red lesions (R_{RL}) and bright lesions (R_{BL}) are detected as foreground (R_{fore}). This is mathematically represented below.

$$R_{OD}, R_{vasc}, R_{fore} \subseteq I_{green} \tag{1}$$

$$R_{fore} = R_{BL} \cup R_{RL}. \tag{2}$$

In Stage 2, classifiers are used in two hierarchical steps. In step 1, candidate lesion regions (R_{BL}, R_{RL}) are classified as true lesions (R_{TBL}, R_{TRL}) and non-lesions (R_{NBL}, R_{NRL}), respectively. In step 2, the true bright lesions are further classified as hard exudates ($\hat{R_{HE}}$) and cotton wool spots ($\hat{R_{CWS}}$), while the true red lesions are classified as micro-aneurysms ($\hat{R_{MA}}$) and hemorrhages ($\hat{R_{HA}}$). This is mathematically represented in (3)-(5). The classified regions are compared to manually annotated ground-truth regions ($R_{HE}, R_{CWS}, R_{MA}, R_{HA}$).

$$R_{BL} = R_{TBL} \cup R_{NBL}, R_{RL} = R_{TRL} \cup R_{NRL} \tag{3}$$

$$R_{TBL} = \hat{R_{HE}} \cup \hat{R_{CWS}} \tag{4}$$

$$R_{TRL} = \hat{R_{MA}} \cup \hat{R_{HA}}. \tag{5}$$

In Stage 3, the number of red lesions and bright lesions are counted and combined using a combination function (Ψ) defined in (6)-(8) to generate a DR severity grade per image (G). An image with $G = 0$ implies no DR, $G = 1$ implies mild DR, $G = 2$ implies moderate DR and $G = 3$ implies severe DR, respectively.

$$HE = |\cup_j \hat{R_{HE}}(j)|, CWS = |\cup_{j'} \hat{R_{CWS}}(j')| \tag{6}$$

$$MA = |\cup_m \hat{R_{MA}}(m)|, HA = |\cup_{m'} \hat{R_{HA}}(m')| \tag{7}$$

$$\forall\ I, G = \Psi(HE, CWS, MA, HA) \tag{8}$$

$$where,\ G = \{0, 1, 2, 3\}. \tag{9}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics

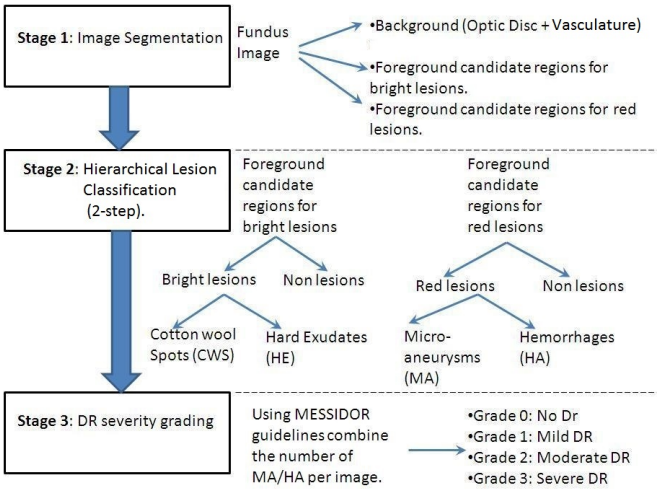


Fig. 1. The three-stage algorithm for grading DR severity using fundus images. The system flow depicts the functionalities of the three individual stages and their interconnections.

A. Data

We train and test our DR screening system for normal patients and patients with DR using two public datasets that are described as follows.

- DIARETDB1 data set [13] consists of 89 images with 50^o FOV. These 89 images have been separated to two groups: 28 training images and 61 test images, by the authors such that the training and test data sets were independent of one another. All the 89 images are manually annotated for lesions as ground truth $(R_{HE}, R_{CWS}, R_{MA}, R_{HA})$.
- MESSIDOR data set [14] contains 1200 images with 45^o FOV. Each image is manually assigned a DR severity grade.

B. Lesion Classification

Once each fundus image is segmented in stage 1, foreground regions corresponding to bright (R_{BL}) and red (R_{RL}) retinopathy lesions can be detected as candidate objects. However, classification of these candidate lesion objects is imperative to reduce instances of false positives, and to categorize bright lesions as hard exudates or cotton wool spots, and red lesions as hemorrhages or microaneurysms, respectively. Thus, to achieve lesion classification, we extract features corresponding to each candidate object, as explained in Section III-B. Every candidate object can be represented by a feature vector $(\mathbf{x})$ corresponding to the location of the object in the feature space. While training classifiers, the feature vectors of the objects from the training set of images are scaled in [0,1]. If the object is manually annotated as a lesion, then a class label $y = 1$ is assigned to the object. Based on the groundtruth, if the object is not a lesion, then $y = 0$. Object vectors from the test set of images $(\mathbf{x}^*)$ are similarly scaled in [0,1], and the class labels determined using groundtruth (y^*) are compared with the class labels, assigned by the classifiers $(\hat{y^*})$, which correspond to class 1 or class 0.

The individual classifiers analyzed here can be categorized into two kinds based on the computational complexity. The first kind of classifiers incur low computational complexity such as kNN [15], and GMM [16] with two multivariate Gaussians, one Gaussian corresponding to class 0 and another to class 1. The second kind of classifiers with high computational complexity include kernel space SVM [17] [18] and AdaBoost [19], [11]. The Bayesian combination of these two categories of classifiers using the Dempster-Shafer theory [20] is also analyzed. Such a combinational classifier, SVM+GMM combines the independent prior probabilities from the SVM and GMM classifiers, and SVM+kNN combines the priors from SVM and kNN, respectively. These classifiers are implemented in MATLAB and the priors for the SVM classifier for the combinational classifiers are computed using the libSVM toolbox [21].

III. Proposed System

In the proposed automated system, all images are preprocessed first to eliminate false photographic artifacts and illumination inconsistencies. The pre-processing module proceeds by histogram equalization and contrast enhancement on I_{green}, followed by scaling all pixel intensities in the range [0,1] resulting in image I_m. Next, since the quality of images vary among data bases, it is necessary to enhance the sharpness and illumination of certain images especially when the images are scanned film prototypes. I_m is filtered using the Laplacian of Gaussian filter to extract the gradient variations (I_h). Next, $(I_m - I_h)$ is median filtered to obtain the enhanced images as shown in Fig. 2. Following the pre-processing module, the three detection stages are invoked for DR severity detection per image.

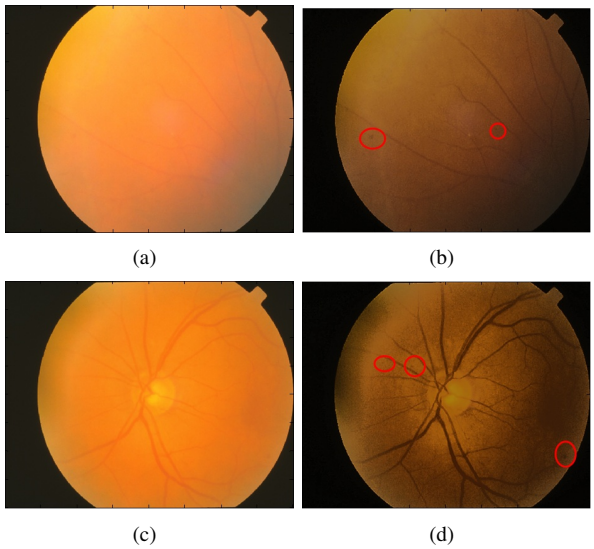


Fig. 2. Original blurry images are enhanced by spatial filtering. (a), (c) Fundus image (I), (b), (d) Filtered image with enhanced red lesions marked in red circles.

A. Stage 1: Image Segmentation

In the first stage, it is imperative to mask out the regions corresponding to the optic disc (OD) and major portions of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics

blood vasculature. This is important since a bright OD may otherwise be mistaken for a bright lesion, and the vasculature can be falsely detected as a red lesion in subsequent automated stages if not masked out in the early stage. For this purpose, a region based *MinIMaS* algorithm [12] is invoked, which detects the regions that lie on the intersection of the largest red region and the bright regions in the image. Once all such intersecting bright regions are identified (R), then the region containing the OD ($R_{OD} \subset R$) is the bright region with minimum pixel intensity sum ($Intensity$, due to the dark pixels corresponding to thick blood vessels occurring at the OD region), and maximum compactness (Sol) as represented in (10). If the region with minimum pixel intensity sum is not the same as the region with maximum solidity ($i_1 \neq i_2$), then the region with minimum pixel intensity sum is discarded ($R = R \backslash R_{i_1}$), and (10) is re-evaluated for the remaining regions till convergence is achieved.

$$\text{If, } R = \{R_1, R_2, ...R_r\}, i \in \{1, 2, ...r\}, \quad (10)$$
$$\Lambda(R) = [i_1, i_2 : i_1 = \arg\min_i Intensity(R_i), \quad i_2 = \arg\max_i Sol(R_i)].$$

$$R_{OD}(\Lambda(R)) = \begin{cases} R_{i_1} = R_{i_2} & : i_1 = i_2 \\ R_{OD}(\Lambda(R = R \backslash R_{i_1})) & : i_1 \neq i_2. \end{cases} \quad (11)$$

After R_{OD} detection, to detect major portions of the blood vasculature (R_{vasc}), each image is gradient smoothened by median filtering (I_{bg}), and subtracted from image I_m to obtain the shade corrected image ($I_{sc} = I_m - I_{bg}$), followed by thresholding and region growing. The largest red region extracted after region growing corresponds to R_{vasc}.

Once the OD and blood vessels are detected and masked out as background regions, foreground candidate regions (R_{fore}) that may represent retinopathy lesions are detected, thereby resulting in a segmented image [12]. This process is shown in Fig. 3. Once the fundus images are segmented and the foreground candidate regions are detected, the second stage of the algorithm is invoked that classifies the bright candidate regions (R_{BL}) and red candidate regions (R_{RL}) as retinopathy lesions.

B. Stage 2: Lesion Classification

Following foreground detection, for all the 6 classifiers under analysis, a training data set was used to obtain the optimal parameters for classification, and a separate test data set was used to analyze their performance.

Lesion classification is performed in two hierarchical binary (1/0) classification steps. In the first step, the bright candidate regions (R_{BL}) are classified as true bright lesions (R_{TBL}, class 1) and non-bright lesions (R_{NBL}, class 0), while the red candidate regions (R_{RL}) are classified as true red lesions (R_{TRL}, class 1) and non-red lesions (R_{NRL}, class 0), respectively. In the second level, the true bright lesions are reclassified as cotton wool spots ($\hat{R_{CWS}}$, class 1) and hard exudates ($\hat{R_{HE}}$, class 0), while the true red lesions are classified as micro-aneurysms ($\hat{R_{MA}}$, class 1) and hemorrhages ($\hat{R_{HA}}$, class 0).

Selection of features is an important aspect for lesion classification. Prior works have introduced feature sets comprising of 13 features [22] using the Spencer-Frame system, while extensions of these features up to 68 features for lesion classification have been presented in [23]. Motivated from these prior works, in this paper, we analyze a set of 78 features that consists of most of the features from [22] and [23] along with additional structural features that are obtained using the 'regionprops' command in MATLAB. Here, 14 features are structure based features such as area, convex area, solidity, orientation etc., and 16 features correspond to the mean, minimum, maximum and standard deviation of pixels within each object region in the green plane (I_{green}), red plane (I_{red}), saturation plane (I_{sat}) and intensity plane (I_{inten}) of the RGB to HSI converted image. The 14 structure based features are useful primarily for separating the non-lesion regions from the lesion candidates. For example bright lesions (HE and CWS) in general have a higher distance from the OD region compared to the non-lesions. For the red lesion classification, the non-lesion regions generally represent fine blood vessels and hence the non-lesions on an average have a smaller distance from the blood vessels than the red lesions. Also, the blood vessels are elongated structures, and hence for the non-lesion regions, the ratio of major axis to minor axis lengths is higher than the hemorrhages or micro-aneurysms. For the second step of hierarchical bright lesion classification, hard exudates (HE) are compact bright regions with well defined boundaries as compared to the cotton-wool spots (CWS) that are larger in area, and less compact (low solidity). The remaining 48 features correspond to the mean and standard deviation of 6 second-order derivative images corresponding to coefficients of the constant, first and second-order horizontal derivative and vertical derivative images ($[1, I_x, I_y, I_{xy}, I_{xx}, I_{yy}]$) scaled in [0,1], using a Gaussian filter with zero mean and $\sigma^2 = [1, 2, 4, 8]$. These 48 features are useful for the second step of hierarchical classification.

In this work, we selected the top 30 features shown in Table I out of 78 features that were ranked in decreasing order of feature weights generated by AdaBoost 12 [19]. For this feature selection operation we used the 89 images from the DIARETDB1 [13] dataset. This process of feature selection for improving the classification performance and computational speed of the algorithm can be viewed as analogous to reverse feature elimination for horizontal pruning in decision tree-based learning as mentioned in [24].

Corresponding to each foreground candidate object for bright and red lesions, 78 features are extracted ($\mathbf{x}$), and each region is manually annotated as a lesion or non-lesion (y). Next the Adaboost algorithm in (12) is invoked to generate feature weights based on the linear discriminating performance of each feature, followed by feature ranking. The top 30 features were the ones with non-zero positive feature weights for classifying bright and red lesions.

While implementing Adaboost for feature selection [11], the weight of each object ('i') is first initialized ($W_{1,i}$). Next, we analyze the $F = 78$ features over $T = 100$ iteration rounds. For each feature 'j', at iteration 't', a linear classifier classifies sample x_i as $h_{t,j}(x_i)$, such that the classification outcome

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics

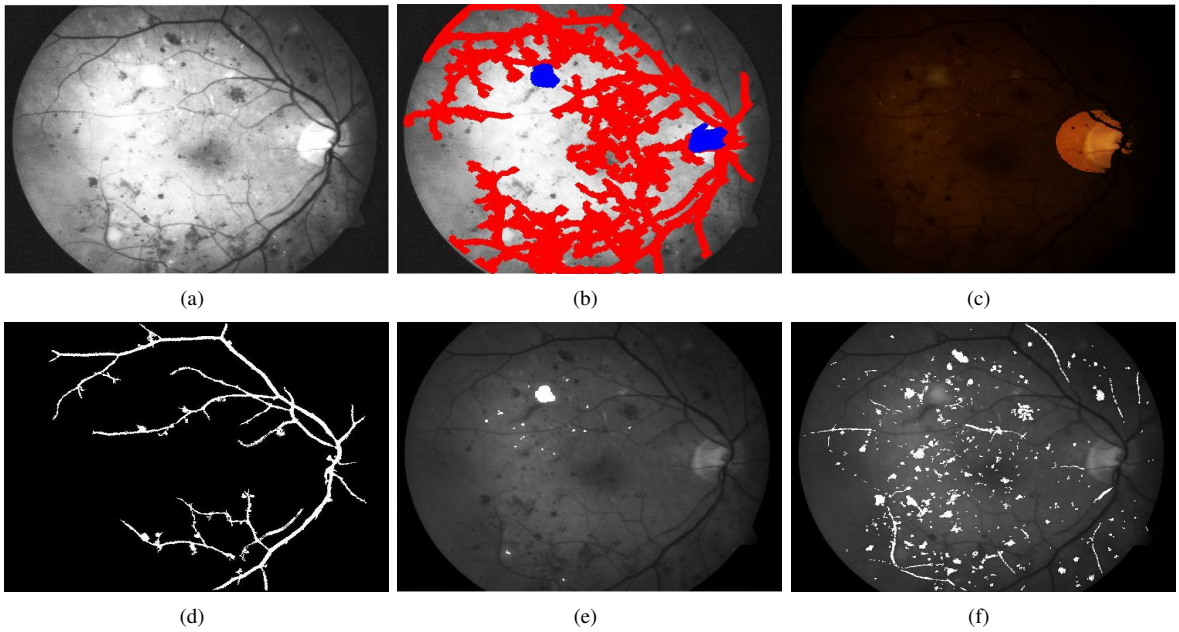


Fig. 3. Detection of OD, vasculature and foreground candidate regions. (a) illumination corrected pre-processed image. (b) Intersecting red regions and bright regions (blue), out of which OD is selected as the bright region with highest solidity. (c) OD (R_{OD}) detected. (d) vasculature (R_{vasc}) detected. (e) Candidate bright lesion regions (R_{BL}) detected. (f) Candidate red lesion regions (R_{RL}) detected.

$h_{t,j}(x_i) = 0$ or 1. Classification error due to this 1-feature based classifier on all samples is $\epsilon_{t,j}$. At any iteration 't', the classifier with minimum $\epsilon_{t,j}$ is chosen as ϵ_t. This error (ϵ_t) is then used to update the sample weights, such that the misclassified samples are weighted more than the correctly classified samples. The final resulting classifier is a weighted linear combination of the linear classifiers in each iteration step as shown in (13). Also, after 'T' iterations, the weights of selected discriminating features in each iteration are combined to find the weight of each feature in the final classifier in (14). Features are then ranked in descending order of their contributing weights.

$$\begin{aligned}
&\textbf{Initialize}: W_{1,i} = 1/N, \quad t = 1:T \\
&\textbf{Define}: \forall j = 1:F,\ h_{t,j}(\mathbf{x}_i) \in [0,1] \\
&\epsilon_{t,j} = \sum_{i=1}^{N} W_{t,i}.\,|h_{t,j}(\mathbf{x}_i) - y_i|, \\
&h_t = \arg\min_j \epsilon_{t,j},\ \epsilon_t = \min_j \epsilon_{t,j} \\
&\textbf{Stop iterating if } \epsilon_t \geq 0. \\
&\textbf{Update}: \quad W_{t+1,i} = \frac{W_{t,i}\exp(\alpha_t(er_i))}{Z_t} \\
&\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t}),\ Z_t = \texttt{Normalizer} \\
&er_i = \begin{cases} -1 & : h_t(\mathbf{x}_i) = y_i \\ 1 & : h_t(\mathbf{x}_i) \neq y_i. \end{cases} \quad \alpha_{t,j} = \begin{cases} \alpha_t & : \epsilon_t = \epsilon_{t,j} \\ 0 & : \epsilon_t \neq \epsilon_{t,j}. \end{cases} \\
&\textbf{Finally}: \quad \hat{y^*} = \begin{cases} 1 & : \sum_{t=1}^{T} \alpha_t.h_t(\mathbf{x}^*) > 0 \\ 0 & : otherwise. \end{cases} \qquad (13) \\
&\textbf{Feature Weight}: \ S(j) = \frac{\sum_{t=1}^{T} \alpha_{t,j}}{\sum_{j}^{F}\sum_{t=1}^{T} \alpha_{t,j}}. \qquad (14)
\end{aligned}$$

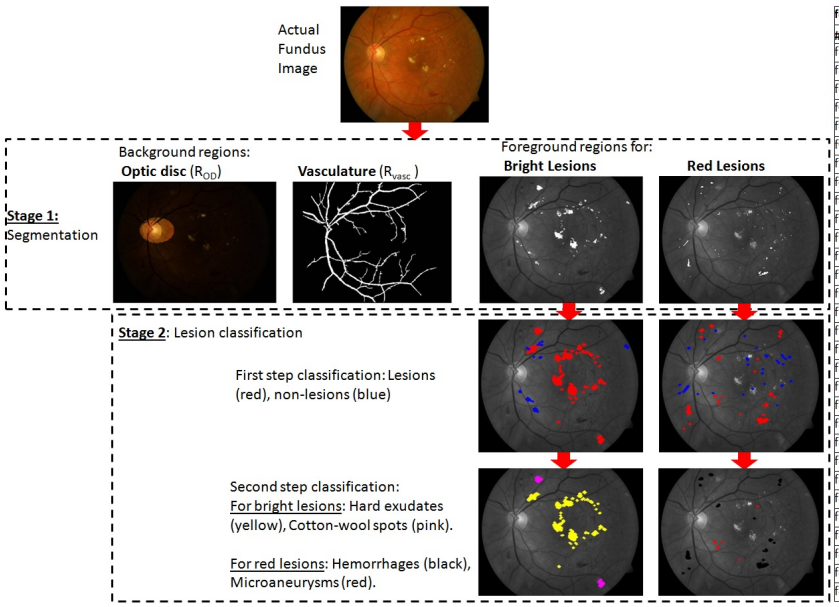
Following feature selection, each foreground candidate object is then classified using the top 30 features described in Table I. The feature vector ($\mathbf{x}$) corresponding to each object in the feature space is computed. In the training phase of lesion classification, the 30-dimensional feature space, with each dimension representing a particular feature, is populated with the objects from the training set of images. The position of each object in the feature space is determined by the feature vector ($\mathbf{x}$), and the class labels of the objects belonging to the training data set (y) are used to determine the class labels of objects belonging to the test data ($\hat{y^*}$) in the testing phase. The two-step hierarchical classification approach is illustrated for an image from the DIARETDB1 data base in Fig. 4.

After the lesions are classified for each image, the next task is to fuse the information regarding the lesions detected to generate a DR severity level for each image [3].

C. Stage 3: DR Severity Grading

Once the regions corresponding to the retinopathy lesions are detected, and the number of hemorrhages (HA), microaneurysms (MA), hard exudates (HE) and cotton-wool spots (CWS) are computed per image using (6) (7), lesion combination (ψ) operation defined in Table II for the MESSIDOR [14] data set can be used to generate a DR severity grade per image (G). The impact of lesion combination on DR severity have been studied in [25], [26] to derive the DR severity grading gold standards. These studies demonstrate that clinically relevant lesion combination operations require more accurate estimation regarding the number of red lesions than bright lesions. In case of bright lesions, over detection, or instances of false positives may imply macular degeneration or retinal abnormalities other than NPDR. Thus, detection of bright lesions must incur less false positives. However, for red

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics



feature(Rank)	NBL	HE	CWS	NRL	HA	MA
# regions	14	51	2	29	22	6
f_1	18.5714	40.6275	209	7.6207	26.4091	5.3333
f_2	117.695	175.1823	253.5167	179.6377	203.0532	133.4392
f_3	6.7548	7.8582	23.9723	8.6377	5.3569	3.5988
f_4	3.62	4.2408	15.1823	2.0476	3.9161	2.0642
f_5	4.02E-05	0.002	4.48E-04	1.79E-04	1.44E-04	1.30E-04
f_6	0.8079	0.7841	0.7684	0.354	0.3441	0.3217
f_7	0.0293	0.0671	0.0272	0.0276	0.0399	0.034
f_8	0.0322	0.0818	0.0288	0.0252	0.0311	0.0392
f_9	4.1651	5.02	16.2733	2.8412	4.8865	2.5581
f_{10}	8.00E-05	9.10E-04	2.97E-04	2.44E-04	1.49E-05	6.97E-05
f_{11}	4.52E-06	3.44E-05	2.17E-05	1.42E-05	8.64E-06	7.09E-06
f_{12}	5	5.5882	17	8.3103	5.7727	2.3333
f_{13}	0.7021	0.6371	0.5303	0.5273	0.5015	0.4833
f_{14}	0.8328	0.7975	0.7083	0.8595	0.7512	0.8024
f_{15}	8.00E-05	9.10E-04	2.97E-04	2.44E-04	1.49E-04	6.97E-05
f_{16}	18.6429	41	209	7.6207	26.5909	5.3333
f_{17}	5.4286	6.8824	25	2.3793	7	3
f_{18}	5.3199	14.2135	21.041	7.62	10.024	11.1042
f_{19}	0.0737	0.0754	0.0799	0.0545	0.0218	0.0112
f_{20}	0.7021	0.6371	0.5303	0.5873	0.5015	0.4833
f_{21}	-19.7164	9.7532	7.86055	-12.9996	29.8739	21.9229
f_{22}	18.8822	23.3977	91.7904	9.753	20.7352	5.8856
f_{23}	0.6902	0.6037	0.4902	0.5094	0.4836	0.4739
f_{24}	0.7154	0.6744	0.5686	0.5413	0.5223	0.4948
f_{25}	0.6902	0.6037	0.4902	0.5094	0.4836	0.4739
f_{26}	0.8141	0.8965	0.7194	0.82	0.898	1
f_{27}	0.9286	0.9804	1	0.91	1	1
f_{28}	0.7154	0.6744	0.5686	0.5413	0.5223	0.4948
f_{29}	0.0802	0.0897	0.1046	0.0573	0.0611	0.0619
f_{30}	0.0769	0.0825	0.0934	0.0538	0.0567	0.0589

Fig. 4. Two-step hierarchical classification of lesions for an image from DIARETDB1. In Stage 1 of the automated system, background regions corresponding to vasculature and optic disc are first detected. Candidate regions for bright (R_{BL}) and red lesions (R_{RL}) are then detected as the foreground. In Stage 2, hierarchical two-step classification is performed for identification of the type of lesions. In the first hierarchical classification step, foreground regions for bright lesions are classified as true lesions (red, R_{TBL}) and non-lesions (blue, R_{NBL}), and candidate regions for red lesions are classified as true red lesions (red, R_{TRL}) and non-lesions (blue, R_{NRL}). In hierarchical second-step classification, true bright lesions are classified into hard exudates (yellow, $\hat{R_{HE}}$), and cotton wool spots (pink, $\hat{R_{CWS}}$), while true red lesions are classified as microaneurysms (red, $\hat{R_{MA}}$), and hemorrhages (black, $\hat{R_{HA}}$). Corresponding to the 30 features mentioned in Table I, the average feature values for all the candidate lesion regions in the sample image is presented in the adjoining table (i.e., f_1 corresponds to feature with rank 1, which is area of the region). The features measuring distance are in terms of pixels, while the mean and variance of intensity values are scaled in [0, 1] range.

lesion detection, failure to detect lesions will result in false interpretation of the DR severity. Thus, it is imperative for any automated system to incur low false negatives for red lesion classification. Hence, the performance criteria for selecting lesion classifiers are as follows:

- *Bright lesion classifier*: It must incur low false positives, or high specificity.
- *Red lesion classifier*: It must incur low false negatives, or high sensitivity.

IV. Results

We analyze the performance of the three stages of the DR screening systems individually. In the first image segmentation stage, the detection of the OD region plays a vital role to ensure fewer instances of false positive while detecting bright lesions. This segmentation algorithm has shown to have an accuracy of 99.7% for OD segmentation on public data sets [12].

The metrics used for analyzing the performance of the second and third stages of the detection system are defined in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as follows.

- sensitivity (SEN)= $\frac{TP}{TP+FN}$
- specificity (SPEC)= $\frac{TN}{TN+FP}$
- area under ROC curve (AUC).

The final goal of our DR detection system is to classify the fundus images that are free from retinopathy lesions as normal, and to classify the images with bright and red lesions as abnormal. The receiver operating characteristic (ROC) curves are obtained by varying the threshold probability in increments of 0.02 in the range [0,1], and observing the variation in SEN versus (1-SPEC) each time. The following subsections present the selection of the best set of classifiers for bright and red lesion classification that have comparatively high SEN, SPEC, and AUC. The DREAM system applies the classifiers that have a consistent performance across both data sets.

A. Feature Set and Classifier Evaluation for Data Set 1: DIARETDB1

Classification of bright lesion and red lesions vs. non-lesions for the DIARETDB1 data set has been analyzed using kNN, GMM and SVM classifiers in [12]. However, in this work, we extend the analysis to two hierarchical step classification, to assess the impact of combination classifiers, and to analyze the incremental importance of the classifying feature set. In lesion classification analysis, a TP region is one which is classified as a lesion and that has more than 50% of its pixels manually marked as a lesion with greater than 50% confidence by the authors of [13].

In Section III-B, Table I, top 30 features computed using AdaBoost, for bright and red lesion classification, are ranked in decreasing order of their contribution to lesion classification. To analyze the importance of these 30 features for lesion classification, we selected the top 'F' features, where F=

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics

TABLE I
FEATURES FOR CLASSIFICATION

Rank	Feature
1	Area: Total pixels in the object.
2	Minimum distance from center of R_{OD}
3	Major Axis length: Major axis of the ellipse (E).
4	Minor Axis length: Minor axis of the ellipse (E).
5	Variance of pixels for the object region in I_{green}.
6	Mean of pixels for the object region in I_{green}.
7	Variance of pixels for the object region in $I_{yy}, [\sigma^2 = 8]$.
8	Variance of pixels for the object region in $I_{xy}, [\sigma^2 = 8]$.
9	Diameter of a circle that with same area as the object.
10	Variance of pixel intensities for the object region in I_{red}.
11	Variance of pixel intensities for the object region in I_{sat}.
12	Length of smallest rectangle (bounding box) around the object.
13	Mean pixel intensity for the object region in I_{red}.
14	Eccentricity of an ellipse (E) estimated around the object with same normalized second central moments as the object.
15	Variance of pixel intensities for the object region in I_{inten}.
16	Filled Area: Area of the object with all holes filled.
17	Width of smallest rectangle (bounding box) around the object.
18	Minimum distance from R_{vasc}.
19	Minimum pixel intensity for the object region in I_{sat}.
20	Mean pixel intensity for the object region in I_{inten}.
21	Orientation: Angle in degrees between the x-axis and the major axis of the ellipse (E).
22	Perimeter: Distance around the object region.
23	Minimum pixel intensity for the object region in I_{red}.
24	Maximum pixel intensity for the object region in I_{inten}.
25	Minimum pixel intensity for the object region in I_{inten}.
26	Solidity = $\frac{area}{convex\_area}$.
27	Euler number of the object.
28	Maximum pixel intensity for the object region in I_{red}.
29	Maximum pixel intensity for the object region in I_{sat}.
30	Mean pixel intensity for the object region in I_{sat}.

TABLE II
THE LESION COMBINATION (ψ) OPERATION PROPOSED FOR MESSIDOR [14].

Grade	Description
0	$(MA = 0)$ and $(HA = 0)$
1	$(0 < MA \leq 5)$ and $(HA = 0)$
2	$(5 < MA \leq 15)$ or $(0 < HA < 5)$
3	$(MA \geq 15)$ or $(HA \geq 5)$

$\{5, 10, 15, 20, 25, 30\}$, and observed the lesion classification performance on the test set of images from the DIARETDB1 data set. In the first hierarchical classification step, candidate regions for bright and red lesions are separated from false positives. The SEN/SPEC of each classifier with varying 'F' is shown in Tables III and IV for bright lesion and red lesion classification, respectively. The threshold probabilities obtained from the ROC curves for evaluating a candidate region as a bright or red lesion in hierarchical Step 1 classification for $\{GMM, kNN, SVM + GMM, SVM + kNN\}$ are approximately $\{0.4, 0.3, 0.45, 0.35\}$, respectively. For Step 2 classification, threshold probability is about 0.45 for all probabilistic classifiers. For the SVM and AdaBoost classifiers, a single operating point is observed since the final classifier is based on the sign of the decision principle. Also, we observe that AdaBoost generates more FNs compared to SVM, which is undesired for lesion classification. It is noteworthy that prior to this lesion classification step, the top 30 features were selected using AdaBoost on all the 89 images from the DIARETDB1 data set to obtain a better fitted feature selection strategy than using the training set of 28 images.

To compare the performance of SVM and AdaBoost with other classifiers for Step 1 bright lesion classification, using the 30 feature set, we observe that at $SEN = 0.98$, GMM, kNN, SVM+GMM, SVM+kNN have $SPEC = \{0.68, 0.45, 0.1, 0.49\}$, respectively. At $SEN = 0.92$, GMM, kNN, SVM+GMM, SVM+kNN have $SPEC = \{0.82, 0.67, 0.41, 0.71\}$, respectively. Thus, from Table III, we observe that for bright lesion classification, the GMM classifier has the highest achievable specificity at a given sensitivity for the most important Step 1 of hierarchical classification. In Step 2 of hierarchical classification, all classifiers demonstrate comparable performance. Thus, hard exudates can be very well separated from cotton-wool spots using the 30 feature set. Also, due to comparable classification performances in Step 2, the low complexity GMM classifier is preferable.

To compare the performance of SVM and AdaBoost with the other classifiers for Step 1 red lesion classification, using the 30 feature set, we observe that at $SPEC = 0.99$, SEN of GMM, kNN, SVM+GMM, SVM+kNN is $\{0.67, 0.68, 0.63, 0.63\}$, respectively. And at $SPEC = 0.97$, SEN of GMM, kNN, SVM+GMM, SVM+kNN is $\{0.69, 0.72, 0.64, 0.65\}$, respectively. Also, for Step 1 classification using AdaBoost classifier with $F = \{5, 10\}$ feature set, the training process stops after $T = \{68, 43\}$ iterations, respectively, instead of completing the $T = 100$ iterations for model selection. These correspond to the entries marked by "$[SEN/SPEC]$" in Table IV. The reason for these incomplete training processes is that the weighted error in classification exceeds random error of 0.5. Once the feature set exceeds 15 and more features, AdaBoost classification performance improves due to appropriate model selection after the complete training process.

From Table IV, we conclude that for red lesion classification, kNN achieves highest SEN for a given SPEC. Also, for Step 2 of hierarchical red lesion classification, all classifiers have comparable performances implying that micro-aneurysms can be very well separated from hemorrhages using the 30 feature set. Hence, the 30 feature set is adequate for lesion classification.

B. Classifier Evaluation for Data Set 2: MESSIDOR

For the MESSIDOR data set, classifiers are trained on 89 images from the DIARETDB1 data set with 30 features and tested on 1200 images for detection of hemorrhages and microaneurysms in each image. In this work, we do not analyze instances of neovascularization or proliferative DR, and we utilize the lesion combination (ψ) operation defined in Table II. According to the reference standard [14], out of the 1200 images, a total of 546 images are normal ($G = 0$), and 654 are abnormal images ($G > 0$) presenting signs of DR. The DR images comprise of 153 images with retinopathy grade 1, 247 images with retinopathy grade 2, and 254 images with retinopathy grade 3. Using the 89 images from the DIARETDB1 for training, the best values of k,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics

TABLE III
SEN/SPEC FOR TWO HIERARCHICAL STEP BRIGHT LESION CLASSIFICATION ON DIARETDB1 DATA SET.

Classifier	No. of features					
	5	10	15	20	25	30
Step 1:	R_{TBL}	vs	R_{NBL}			
GMM	0.62/0.85	0.66/0.85	0.73/0.85	0.79/0.85	0.87/0.85	**0.89/0.85**
kNN	0.67/0.85	0.76/0.85	0.81/0.85	0.81/0.85	0.84/0.85	0.86/0.85
SVM	0.67/0.63	0.89/0.62	0.84/0.69	0.95/0.76	0.95/0.72	0.98/0.59
SVM+GMM	0.60/0.84	0.66/0.85	0.70/0.85	0.76/0.85	0.78/0.85	0.81/0.85
SVM+kNN	0.68/0.85	0.73/0.85	0.77/0.85	0.79/0.85	0.83/0.85	0.86/0.85
Adaboost	0.61/0.90	0.62/0.94	0.73/0.95	0.85/0.85	0.85/0.87	0.92/0.79
Step 2:	$\hat{R_{HE}}$	vs	$\hat{R_{CWS}}$			
GMM	0.94/0.97	0.97/0.97	0.98/0.97	0.98/0.97	0.98/0.97	0.99/0.97
kNN	0.82/0.97	0.87/0.97	0.88/0.97	0.95/0.97	0.96/0.97	0.98/0.97
SVM	0.95/0.98	0.97/0.96	0.95/0.99	0.97/0.99	0.98/0.99	0.99/1
SVM+GMM	0.93/0.97	0.94/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.98/0.97
SVM+kNN	0.95/0.97	0.96/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.98/0.97
AdaBoost	0.93/0.92	0.95/0.93	0.96/0.91	0.96/0.98	0.97/0.95	0.98/0.98

TABLE IV
SEN/SPEC FOR FOR TWO HIERARCHICAL STEP RED LESION CLASSIFICATION ON DIARETDB1 DATA SET.

Classifier	No. of features					
	5	10	15	20	25	30
Step 1:	R_{TRL}	vs	R_{NRL}			
GMM	0.61/0.85	0.64/0.85	0.69/0.84	0.70/0.85	0.74/0.85	0.79/0.85
kNN	0.60/0.84	0.69/0.83	0.75/0.85	0.77/0.85	0.79/0.84	**0.80/0.85**
SVM	0.60/0.84	0.53/0.89	0.47/0.91	0.48/0.91	0.50/0.91	0.61/0.99
SVM+GMM	0.51/0.85	0.59/0.85	0.58/0.85	0.64/0.85	0.69/0.85	0.72/0.85
SVM+kNN	0.52/0.85	0.57/0.85	0.68/0.85	0.74/0.85	0.73/0.85	0.75/0.85
AdaBoost	[0.65/0.79]	[0.97/0.23]	0.51/0.97	0.56/0.97	0.59/0.97	0.60/0.97
Step 2:	$\hat{R_{MA}}$	vs	$\hat{R_{HA}}$			
GMM	0.74/0.90	0.75/0.90	0.79/0.90	0.80/0.90	0.81/0.90	0.82/0.90
kNN	0.75/0.90	0.76/0.90	0.84/0.90	0.93/0.90	0.94/0.90	0.95/0.90
SVM	0.83/0.94	0.86/0.95	0.87/0.96	0.85/0.97	0.94/0.97	0.97/ 1
SVM+GMM	0.82/0.90	0.85/0.90	0.88/0.90	0.91/0.90	0.92/0.90	0.97/0.90
SVM+kNN	0.80/0.90	0.87/0.90	0.90/0.90	0.92/0.90	0.94/0.9	0.95/0.90
AdaBoost	0.80/1	0.87/1	0.90/1	0.91/0.99	0.91/0.98	0.92/0.98

for kNN classifier, were found as: $k = 11$ for Step 1 red lesion classification (R_{TRL} vs R_{NRL}), and $k = 3$ for Step 2 classification of $\hat{R_{MA}}$ from $\hat{R_{HA}}$.

The performance of classifying normal images ($G = 0$) from all other images with non-zero diabetic grades ($G = 1, 2, 3$) is depicted in Table V. The number of images with non-zero DR grade that are falsely graded as $G = 0$ are represented as $\#FN$ in Table V. The operating point in this analysis is selected where SEN is greater than 0.9 and SPEC is approximately 0.5 [27], such that the threshold probabilities for Step 1 red lesion classification for GMM, kNN, SVM+GMM, SVM+kNN are $\{0.4, 0.3, 0.45, 0.35\}$, respectively. For Step 2 classification, the threshold probability is 0.45.

From Table V, we observe that SVM+kNN and kNN have good SEN/SPEC classification combination for separating normal images from abnormal ones. In such situations, the kNN classifier is preferable since it has lower computational complexity. Also, we observe the AdaBoost introduces the highest number of FNs (abnormal images incorrectly classified as normal), which is undesired. Besides, from Table V, it is evident that the lesion combination operation reduces the impact of lesion classification. However, when the number of images being screened for DR becomes significantly large, such as in the case of the MESSIDOR data set, slight variations in classification performances can be significant. For example, in Table V, for the normal to abnormal image classification, kNN

TABLE V
RESULTS OF CLASSIFICATION AND LESION COMBINATION ON MESSIDOR DATA SET.

Statistics	kNN (DREAM)	GMM	SVM	SVM+ GMM	SVM+ kNN	Ada Boost
Grade:	(0 vs 1,2,3)					
SEN	1	1	0.8474	0.9916	1	0.8062
SPEC	0.5316	0.5316	0.6975	0.5492	0.5324	0.7648
# FN	0	0	182	10	0	231
Grade:	(0 vs 1)					
SEN	0.8940	0.9070	0.8974	0.9004	0.9053	0.8689
SPEC	0.5894	0.5158	0.5189	0.5330	0.5781	0.6312
# FN	61	53	50	58	52	75
Grade:	(0 vs 2)					
SEN	0.9507	0.9295	0.8787	0.9263	0.9568	0.8484
SPEC	0.5046	0.5035	0.5811	0.5033	0.5065	0.6773
# FN	36	49	89	51	32	111
Grade:	(0 vs 3)					
SEN	0.9888	0.9814	0.9479	0.9710	0.9876	0.9350
SPEC	0.4872	0.4842	0.5662	0.495	0.5097	0.6661
# FN	8	14	34	20	9	45

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics

and SVM+GMM have SEN/SPEC 1/0.5316, 0.9916/0.5492, respectively. Although, these performances appear similar, the significant difference is that the SVM+GMM classifier falsely classifies 10 images with DR as normal images. As the number of patients being screened for DR increases, the number of missed screenings will grow, which is undesired from an automated screening system.

Additionally, from Table VI, we observe that combination classifiers do not show significant improvement over single classifiers in both data sets. Also, combination classifiers introduce computational complexity, thus discouraging their use for lesion classification.

TABLE VI
AUC ASSESSMENT ON 3 DATA SETS

Data set	kNN	SVM+kNN	GMM	SVM+GMM
DIARETDB1: R_{TBL}	0.9434	0.9533	0.9593	0.9222
DIARETDB1: R_{TRL}	0.8414	0.8356	0.8663	0.8300
MESSIDOR: (0/1,2,3)	0.9039	0.9097	0.8911	0.8826
MESSIDOR: (0/1)	0.9238	0.9151	0.923	0.9201
MESSIDOR: (0/2)	0.8915	0.8975	0.8831	0.8755
MESSIDOR: (0/3)	0.9314	0.9343	0.9087	0.9399

The important conclusion from the classification performance analysis of the two data sets is that, GMM may be preferred for classification of bright lesions, while kNN outperforms GMM and SVM for red lesion classification. Thus, for our three-stage screening system (DREAM), we use GMM for classification of bright lesions and on kNN for classification of red lesions.

C. Time complexity Analysis of the DREAM system

The time complexity of the proposed DREAM system for red lesion and bright lesion detection is analyzed in terms of the time required to classify the lesions and grade images using MATLAB programs. On a Laptop (2.53 GHz Intel Core i3 and 2 GB RAM), the times required per image to classify candidate objects corresponding to lesions on the image from DIARETDB1, and to grade images from the MESSIDOR data set are shown in Table VII.

In the DIARETDB1 data set, for each test image, the time required to classify bright candidate objects into HE/CWS, and the red candidate objects into MA/HA after rejecting false positives are analyzed using 3 classification methods. The first method (M1) uses 78 features for classifying bright objects using GMM classifier and red lesions using kNN classifier in two hierarchical steps. The second method (M2) uses the top 30 features for classification, but it uses 2 separate GMM classifiers in parallel for discriminating CWS from non-bright lesions, HE from non-bright lesions, and 2 separate kNN classifiers in parallel for separating MA from non-red lesions and HA from non-red lesions, respectively. The third method (M3) follows the 2-step hierarchical classification using 30 features as defined in (3-5). In the MESSIDOR data set, all the 3 classification methods (M1, M2, M3) are used for red lesion classification only followed by generating a DR severity grade per image.

Table VII shows a 97% time complexity reduction (from 71.63s to 2.06s) for images from DIARETDB1, and 94.18% reduction (from 59.54s to 3.46s) for images from MESSIDOR data set in the average bright and red lesion detection time required per image. This reduction in time complecity is achieved by reducing the number of features from 78 to 30 and applying the proposed 2-step hierarchical classification method.

TABLE VII
TIMING ANALYSIS PER IMAGE IN SECONDS

Classification Method	Mean	Max	Min	Std. dev.
DIARETDB1				
M1 (F=78, hierarchical)	71.63	93.69	48.85	11.27
M2 (F=30, parallel)	2.53	7.67	1.16	2.17
M3 (F=30, hierarchical)	2.06	3.33	0.98	1.50
MESSIDOR				
M1 (F=78, hierarchical)	59.54	$2.58x10^3$	3.25	212.84
M2 (F=30, parallel)	4.58	7.55	1.64	1.60
M3 (F=30, hierarchical)	3.46	5.39	1.55	0.67

Next, we observe that the proposed 2-step hierarchical classification method (M3) incurs 18.5% less time than the parallel classification method M2 (2.06s vs. 2.53s) for DIARETDB1 images, and 24.45% less time (3.46s vs. 4.58s) for MESSIDOR images. The best SEN/SPEC obtained by M2 and M3 on the DIARETDB1 and MESSIDOR data sets are 0.82/0.94, and 1/0.53, respectively.

D. Comparison with Prior Work

Comparative performance analysis of the proposed system (DREAM) in classifying bright and red lesions to existing works is shown in Table VIII. The SEN/SPEC of the DREAM system is chosen from the ROC curves corresponding to the GMM classifier with 30 feature set and threshold probability 0.7 for bright lesion classification, and kNN classifier with 30 feature set and threshold probability 0.3 for red lesion classification, respectively. We observe that the DREAM system has a favorable Step 1 classification performance for bright and red lesions compared to existing methods. Additionally, the proposed system has 95% sensitivity and 90% specificity for separating micro-aneurysms from hemorrhages, and 99% sensitivity, 97% specificity for separating hard exudates from cotton-wool spots. From Table VIII, we observe that prior works on lesion classification using the DIARETDB1 data set also demonstrated a preference for high SPEC for bright lesion classification and high SEN for red lesion classification, which is in accordance with our classifier evaluation criteria.

TABLE VIII
COMPARISON OF LESION DETECTION PERFORMANCE (%) ON DIARETDB1

Lesion Type	Method	SEN (%)	SPEC (%)
Red	Bhalero et.al.[28]	82.6	80.2
Red	Kauppi et.al.[13]	77.78	76.47
Red	Proposed (DREAM)	80	85
Bright	Sophorak et.al.[29] (47 images)	43.48	99.31
Bright	Walter et.al.[30] (47 images)	66	98.64
Bright	Welfer et.al.[31] (47 images)	70.48	98.84
Bright	DREAM (61 images)	74.2	98

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics

Comparing the performance of classification and image grading of our system with existing works on the public data set MESSIDOR [14], the proposed system achieves higher sensitivity, specificity and AUC, as shown in Table IX. Another ARVO 2008 version of the method in [30] tested on 400 images from this data set achieved an AUC of 0.89 in classifying images with grade 0 and grade 1 from the images with a higher retinopathy severity grade. Our system outperforms all existing methods that classify all the 1200 images from the MESSIDOR data set. In Table IX, the method [32] classifies normal from abnormal images by subsampling from the MESSIDOR image set to train, validate and test neural network classifiers. Thus, this method classifies about 300 images or less out of the 1200 images. Also, the performance of other DR severity grading methods on various local data sets is shown in Table IX. We observe that only the proposed system achieves 100% sensitivity, which is favorable over existing grading systems since no patient with DR is missed while screening using the DREAM system.

TABLE IX
COMPARISON OF DR SEVERITY GRADING PERFORMANCE FOR SEPARATING NORMAL FROM ABNORMAL IMAGES

Method	SEN (%)	SPEC (%)	AUC
MESSIDOR Data			
Sanchez et.al.[27]	92.2	50	0.876
Agurto et.al.[8]	92	54	0.84
Antal et.al.[33]	96	51	0.875
Esnaashari et.al.[32] ($\leq$ 300 images)	95	89.29	-
Barriga et.al.[34](400 images)	98	67	0.86
DREAM	100	53.16	0.904
Local Data			
Agurto et.al.[35](2247 images)	92	51	0.89
Acharya et.al.[36](300 images)	82	88	-
Acharya et.al.[37](331 images)	82	86	-
Usher et.al.[38](1273 images)	94.8	52.8	-

V. CONCLUSIONS AND DISCUSSION

In this paper we have proposed a three-stage computer-aided screening system for DR (DREAM) that detects and grades fundus images for the severity of NPDR. In this work, we have identified a set of best 30 features out of 78 features using feature scoring by AdaBoost [11], such that the 30 feature set classifies bright and red lesions with AUC greater than 0.83 using the classifiers such as kNN, GMM and Bayesian combination of probabilistic classifiers (SVM+GMM, SVM+kNN). Additionally, time complexity analysis of the proposed 2-step hierarchical classification method using all the 78 features and the reduced 30 feature set demonstrates computational speed enhancement by more than 94% by this feature reduction operation. This time complexity analysis is a significant contribution, since to our best knowledge no prior works have reported the time complexity of the DR detection systems. The final 30 features are capable of classifying lesions under 6 seconds per image, and these 30 features are scalable across image data sets with varying fields of view.

It is noteworthy that the first stage of the DREAM system is an important step for correctly detecting the presence of diabetic retinopathy (DR). Among all the images analyzed in this paper from the public data set of DIARETDB1 and MESSIDOR (which includes 1289 fundus images), the optic disc (OD) and blood vessels have always been correctly segmented. However, it is possible that for some fundus images with additional retinal pathologies such as myelination or peripapillary atrophy, the entire OD region is not completely masked out. This might result in an additional bright lesion being detected by the DREAM system. Also, since the minimum distance from the OD region is a feature for lesion classification in the second stage of the system, erroneous segmentation of the OD might lead to minor inaccuracies in lesion classification. However, this classification error is significantly small since the error appears due to 1 out of 30 un-weighted features, and since the minimum distance from OD is an important classification feature for mainly bright lesions regions. Since bright lesions are not considered for evaluating the severity of DR by the proposed system, this false detection of the OD region will not impact the final DR severity grade produced by the DREAM system.

In instances when the retina severely degenerates, such as the examples shown in Fig. 5, where retinal vein occlusions occur, the DREAM system classifies such images as abnormal, even if the patients don't suffer from DR. Such fundus images would result in false positive outcomes by the automated screening system. However, all false positive outcomes will be further manually assessed by retinal specialists to determine follow-up treatment procedures. Hence, a patient with retinal vein occlusion will be detected as abnormal by the DR screening system, and the patient will be further manually assessed.

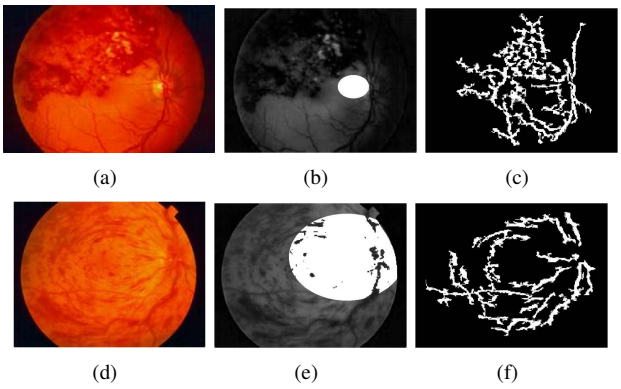


Fig. 5. DREAM system on public domain images with severe retinal degeneration. (a), (d) Original image. (b), (e) OD region detected with error. (c), (f) Vasculature detected in first stage. Both images (a), (d) are classified as images with DR by the DREAM system.

The lesion classification problem in the second stage of the DREAM system deals with unbalanced data sets. Hence, it is important to discuss the performance of the lesion classifiers on the unbalanced lesion candidate regions. A trained SVM classifier generally tends to classify each test sample as belonging to the majority class [17] [18], thereby causing a biased classification outcome. The AdaBoost classifier is accuracy-oriented, which means that its weighting strategy may be biased towards the majority class if it contributes more to the overall classification accuracy as mentioned in [39].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics

To ensure low computational complexity for the proposed DR screening system, cost-sensitive SVM and AdaBoost classifiers [39] have not been assessed in this work. Future works will be directed towards analyzing the impact of cost-sensitive SVM and AdaBoost for lesion classification. The GMM classifier is weighted by the number of samples belonging to class 1 or class 0 at the end of the Expectation Maximization operation, and hence it is quite robust to handling unbalanced data sets as mentioned in [40]. The kNN classifier has a localized impact while classification since each data sample looks at its immediate neighborhood to decide its class label. Existing works [41] have shown that kNN is robust to imbalance in data set caused due to a high number of negative samples. In this work we observe that for the red lesion classification problem, where the numbers of negative samples are 4 times more than the number of positive samples, kNN is the best classifier, an observation which is in accordance with the existing works [41]. Combination of SVM with kNN or GMM classifiers using the Dempster-Shafer theory [20] does not significantly enhance the lesion classification performance since SVM is sensitive to the imbalance in the data, and hence, the classification performance of the combined classifiers is dampened. However, other classifier combination strategies may enhance the lesion classification performance. Future work will be focused on the assessment of other classifier combination strategies for the lesion classification problem.

Although the proposed DREAM system outperforms all existing DR screening systems, further improvement in DR classification specificity may be possible by retraining the lesion classifiers for every new test set of images, by using cost-sensitive SVM or AdaBoost classifiers for lesion classification, or by using other cost-sensitive classifier combination algorithms. Future work will be directed towards further enhancement of the DR classification performance. However, classifier retraining or cost-sensitive classifiers may incur additional computational time complexity. Additional future work can be directed towards the detection of neovascularization and vascular beading caused by proliferative DR, and drusen caused by macular degeneration.

ACKNOWLEDGMENT

The authors are grateful to the anonymous reviewers and the Associate Editor for their numerous constructive comments.

REFERENCES

[1] American Diabetes Association, "Data from the 2011 national diabetes fact sheet." Jan. 26, 2011. [Online]. Available: http://www.diabetes.org/diabetes-basics/diabetes-statistics/

[2] S. Garg and R. M. Davis, "Diabetic retinopathy screening update," *CliniCal Diabetes*, vol. 27, no. 4, pp. 140–145, 2009.

[3] M. D. Abramoff, M. Niemeijer, and S. R. Russell, "Automated detection of diabetic retinopathy: barriers to translation into clinical practice," in *Expert Review of Medical Devices*, vol. 7, no. 2, 2010, pp. 287–296.

[4] J. Cuadros and G. Bresnick, "Eyepacs: An adaptable telemedicine system for diabetic retinopathy screening," *J Diabetes Sci Technol.*, vol. 3, no. 3, p. 509516, May 2009.

[5] A. Osareh, M. Mirmehdi, B. Thomas, and R. Markham, "Classification and localisation of diabetic-related eye disease," in *Computer Vision ECCV 2002*, ser. Lecture Notes in Computer Science, vol. 2353. Springer Berlin / Heidelberg, 2006, pp. 325–329.

[6] G. Quellec, M. Lamard, P. Josselin, and G. Cazuguel, "Optimal wavelet transform for the detection of microaneurysm in retina photographs," *IEEE Transactions on Medical Imaging*, vol. 27, pp. 1230 –1241, 2008.

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

[8] C. Agurto, V. Murray, E. Barriga, S. Murillo, M. Pattichis, H. Davis, S. Russell, M. Abramoff, and P. Soliz, "Multiscale am-fm methods for diabetic retinopathy lesion detection," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 502–512, February 2010.

[9] G. S. Scotland, P. McNamee, A. D. Fleming, K. A. Goatman, S. Philip, G. J. Prescott, P. F. Sharp, G. J. Williams, W. Wykes, G. P. Leese, and J. A. Olson, "Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy," *British Journal of Ophthalmology*, vol. 94, no. 6, pp. 712–719, 2010.

[10] M. D. Abramoff, M. Niemeijer, M. S. Suttorp-Schulten, M. A. Viergever, S. R. Russell, and B. van Ginneken, "Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes," *Diabetes Care*, vol. 31, no. 2, pp. 193–198, February 2008.

[11] L. Shen and L. Bai, "Abstract adaboost gabor feature selection for classification," *Proc. of Image and Vision Computing, New Zealand*, pp. 77–83, 2004.

[12] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "Screening fundus images for diabetic retinopathy," in *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2012, pp. 1641–1645.

[13] T. Kauppi, V. Kalesnykiene, J.-K. Kmrinen, L. Lensu, I. Sorr, A. Raninen, R. Voutilainen, H. Uusitalo, H. Klviinen, and J. Pietil, "Diaretdb1 diabetic retinopathy database and evaluation protocol," *Proc. of the 11th Conf. on Medical Image Understanding and Analysis (MIUA2007)*, pp. 61–65, 2007.

[14] "Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology." Accessed Sept 23, 2011. [Online]. Available: http://messidor.crihan.fr/download-en.php

[15] M. Niemeijer, B. van Ginneken, J. Staal, M. Suttorp-Schulten, and M. Abramoff, "Automatic detection of red lesions in digital color fundus photographs," *IEEE Transactions on Medical Imaging*, vol. 24, no. 5, pp. 584 –592, May 2005.

[16] A. Osareh, B. Shadgar, and R. Markham, "Comparative pixel-level exudate recognition in colour retinal images," *Image Analysis and Recognition*, vol. 3656, pp. 894–902, 2005.

[17] L. Xu and S. Luo, "Support vector machine based method for identifying hard exudates in retinal images," in *IEEE Youth Conference on Information, Computing and Telecommunication, 2009. YC-ICT '09*, Sept. 2009, pp. 138 –141.

[18] U. R. Acharya, C. K. Chua, E. Y. Ng, W. Yu, and C. Chee, "Application of higher order spectra for the identification of diabetes retinopathy stages," *Journal of Medical Systems*, vol. 32, pp. 481–488, 2008.

[19] V. Cherkassky and F. Mullier, "Learning from data," *John Wiley and sons, New York*, 1998.

[20] G. Shafer, "A mathematical theory of evidence," *Princeton, N.J.: Princeton University Press*, 1976.

[21] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[22] A. Frame, P. Undrill, M. Cree, J. Olson, K. McHardy, P. Sharp, and J. Forrester, "A comparison of computer based classi?cation methods applied to the detection of microaneurysms in ophthalmic ?uorescein angiograms," *Comput. Biol. Med.*, vol. 28, pp. 225–238, 1998.

[23] M. Niemeijer, B. van Ginneken, J. Staal, M. Suttorp-Schulten, and M. Abramoff, "Automatic detection of red lesions in digital color fundus photographs," *IEEE Transactions on Medical Imaging*, vol. 24, no. 5, pp. 584 –592, May 2005.

[24] L. Talavera, "Feature selection as retrospective pruning in hierarchical clustering," *Advances in Intelligent Data Analysis*, vol. 1642, pp. 75–86, 1999.

[25] M. G. Lawrence, "The accuracy of digital-video retinal imaging to screen for diabetic retinopathy: an analysis of two digital-video retinal imaging systems using standard stereoscopic seven-field photography and dilated clinical examination as reference standards," *Trans Am Ophthalmol Soc*, vol. 102, p. 321340, 2004.

[26] A. D. S. for the Department of Health and Ageing, "Guidelines for the management of diabetic retinopathy," 2008. [Online]. Available: http://www.icoph.org/downloads/Diabetic−Retinopathy−Detail.pdf

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2013.2294635, IEEE Journal of Biomedical and Health Informatics

[27] C. I. Sanchez, M. Niemeijer, A. V. Dumitrescu, M. S. A. Suttorp-Schulten, M. D. Abramoff, and B. van Ginneken, "Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data," *Investigative Ophthalmology and Visual Science*, vol. 52, no. 7, pp. 4866–4871, 2011.

[28] A. Bhalerao, A. Patanaik, S. Anand, and P. Saravanan, "Robust detection of microaneurysms for sight threatening retinopathy screening," in *Sixth Indian Conference on Computer Vision, Graphics Image Processing, 2008. ICVGIP '08*, Dec. 2008, pp. 520 –527.

[29] A. Sopharak, B. Uyyanonvara, S. Barman, and T. H. Williamson, "Automatic detection of diabetic retinopathy exudates from non-dilated retinal images using mathematical morphology methods," *Computerized Medical Imaging and Graphics*, vol. 32, no. 8, pp. 720 – 727, 2008.

[30] T. Walter, J.-C. Klein, P. Massin, and A. Erginay, "A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina," *IEEE Transactions on Medical Imaging*, vol. 21, no. 10, pp. 1236 –1243, Oct. 2002.

[31] D. Welfer, J. Scharcanski, and D. R. Marinho, "A coarse-to-fine strategy for automatically detecting exudates in color eye fundus images," *Computerized Medical Imaging and Graphics*, vol. 34, no. 3, pp. 228 – 235, 2010.

[32] M. Esnaashari, S. A. Monadjemi, and G. Naderian, "A content-based retinal image retrieval method for diabetes-related eye diseases diagnosis," in *International Journal of Research and reviews in Computer Science*, vol. 2, no. 6, 2011, p. 1222.

[33] B. Antal and A. Hajdu, "An ensemble-based system for microaneurysm detection and diabetic retinopathy grading," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1720 –1726, June 2012.

[34] E. Barriga, V. Murray, C. Agurto, M. Pattichis, W. Bauman, G. Zamora, and P. Soliz, "Automatic system for diabetic retinopathy screening based on am-fm, partial least squares, and support vector machines," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, April 2010, pp. 1349 –1352.

[35] C. Agurto, E. Barriga, V. Murray, S. Nemeth, R. Crammer, W. Bauman, G. Zamora, M. Pattichis, and P. Soliz, "Automatic detection of diabetic retinopathy and age-related macular degeneration in digital fundus images," *Invest Ophthalmol Vis Sci.*, vol. 52, no. 8, pp. 5862 –5871, July 29 2011.

[36] R. Acharya U, C. K. Chua, E. Y. Ng, W. Yu, and C. Chee, "Application of higher order spectra for the identification of diabetes retinopathy stages," *J. Med. Syst.*, vol. 32, no. 6, pp. 481–488, Dec 2008.

[37] U. R. Acharya, C. M. Lim, E. Y. K. Ng, C. Chee, and T. Tamura, "Computer-based detection of diabetes retinopathy stages using digital fundus images," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 223, no. 5, pp. 545–553, 2009.

[38] D. Usher, M. Dumskyj, M. Himaga, T. H. Williamson, S. Nussey, and J. Boyce, "Automated detection of diabetic retinopathy in digital retinal images: a tool for diabetic retinopathy screening," *Diabetic Medicine*, vol. 21, no. 1, pp. 84–90, 2004.

[39] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358 – 3378, 2007.

[40] D. Williams, V. Myers, and M. Silvious, "Mine classification with imbalanced data," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 3, pp. 528–532, 2009.

[41] Y. Yang, "An evaluation of statistical approaches to text categorization," *Journal of Information Retrieval*, vol. 1, pp. 67–88, 1999.

S ohini Roychowdhury received the B.Tech degree in Electronics and Communication Engineering from Birla Institute of Technology, India, in 2007, and M.S. in Electrical Engineering from Kansas State University in 2010. She is a doctoral candidate in the Electrical and Computer Engineering Department at the University of Minnesota Twin Cities, Minneapolis, MN. Her research interests include image processing, signal processing, pattern recognition, machine learning and artificial intelligence.

D ara Koozekanani received a PhD from the Ohio State University in Biomedical Engineering in 2001. His research dissertation involved application of computer vision techniques to the analysis of optical coherence tomography images. He subsequently received an MD degree from The Ohio State University in 2003, completed ophthalmology residency at the University of Wisconsin in 2007, and completed a surgical retinal fellowship at the Medical College of Wisconsin in 2009.

He is currently an assistant professor of ophthalmology on the clinical faculty at the University of Minnesota. He sees patients with a variety of surgical and medical retinal diseases. His research interests are the application of ophthalmic imaging technologies and automated analysis of those images.

K eshab K. Parhi (S'85-M'88-SM'91-F'96) received the B.Tech. degree from the Indian Institute of Technology, Kharagpur, in 1982, the M.S.E.E. degree from the University of Pennsylvania, Philadelphia, in 1984, and the Ph.D. degree from the University of California, Berkeley, in 1988. He has been with the University of Minnesota, Minneapolis, since 1988, where he is currently Distinguished McKnight University Professor and Edgar F. Johnson Professor in the Department of Electrical and Computer Engineering. He has published over 500 papers, has authored the textbook VLSI Digital Signal Processing Systems (Wiley, 1999) and coedited the reference book Digital Signal Processing for Multimedia Systems (Marcel Dekker, 1999). His research addresses VLSI architecture design and implementation of signal processing, communications and biomedical systems, error control coders and cryptography architectures, high-speed transceivers, and ultrawideband systems. He is also currently working on intelligent classification of biomedical signals and images, for applications such as seizure prediction and detection, schizophrenia classification, biomarkers for mental disorder, brain connectivity, and diabetic retinopathy screening.

Dr. Parhi is the recipient of numerous awards including the 2013 Distinguished Alumnus Award from IIT, Kharagpur, India, 2013 Graduate/Professional Teaching Award from the University of Minnesota, 2012 Charles A. Desoer Technical Achievement award from the IEEE Circuits and Systems Society, the 2004 F. E. Terman award from the American Society of Engineering Education, the 2003 IEEE Kiyo Tomiyasu Technical Field Award, the 2001 IEEE W. R. G. Baker prize paper award, and a Golden Jubilee medal from the IEEE Circuits and Systems Society in 2000. He has served on the editorial boards of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS PART I and PART II, VLSI Systems, Signal Processing, Signal Processing Letters,and Signal Processing Magazine, and served as the Editor-in-Chief of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMSPART I (2004-2005 term), and currently serves on the Editorial Board of the Journal of VLSI Signal Processing. He has served as technical program cochair of the 1995 IEEE VLSI Signal Processing workshop and the 1996 ASAP conference, and as the general chair of the 2002 IEEE Workshop on Signal Processing Systems. He was a distinguished lecturer for the IEEE Circuits and Systems society during 1996-1998. He served as an elected member of the Board of Governors of the IEEE Circuits and Systems society from 2005 to 2007.