Research Article | Vol. 10, No. 2 | 1 Feb 2019 | BIOMEDICAL OPTICS EXPRESS 914
Biomedical Optics EXPRESS

Drusen diagnosis comparison between hyper-spectral and color retinal images

YIYANG WANG,[1] BRIAN SOETIKNO,[2,3] JACOB FURST,[1] DANIELA RAICU,[1,4] AND AMANI A. FAWZI[2,5]

[1]College of Computing and Digital Media, DePaul University, Chicago, Illinois, 60604, USA
[2]Department of Ophthalmology, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA
[3]Functional Optical Imaging Laboratory, Department of Biomedical Engineering, Northwestern University, Evanston, IL 60208, USA
[4]draicu@cdm.depaul.edu
[5]amani.fawzi@northwestern.edu

Abstract: Age-related macular degeneration (AMD) is a degenerative aging disorder, which can lead to irreversible vision loss in older individuals. The emergence of clinical applications of retinal hyper-spectral imaging provides a unique opportunity to capture important spectral signatures, with the potential to enhance the molecular diagnosis of retinal diseases. In this study, we use a machine learning classification approach to explore whether hyper-spectral images offer an improved outcome compared to standard RGB images. Our results show that the classifier performs better on hyper-spectral images with improved accuracy and sensitivity for drusen classification compared to standard imaging. By examining the most important features in the classification task, our data suggest that drusen are highly heterogeneous. Our work provides further evidence that hyper-spectral retinal image data are uniquely suited for computer-aided diagnosis and detection techniques.

1. Introduction

Age-related macular degeneration (AMD) is a major health burden that can lead to irreversible vision loss in the elderly population [1]. Early detection of AMD is extremely important to identify patients who are at high risk of permanent vision loss and who can benefit from early preventative interventions. Drusen, the characteristic early AMD lesion, appears as yellowish deposits under the retina. The molecular composition of drusen has been studied extensively using biochemical and molecular techniques. Non-invasive approaches for detecting and distinguishing these lesions in healthy retinal images have become important in the medical informatics field [2,3].

Clinically, RGB fundus imaging is the gold standard modality for drusen detection and AMD risk stratification. RGB fundus images visualize drusen based on their color and overlying pigment variation in the fundus [4]. However, these images have low contrast and suffer from heterogamous illumination [5]. Fluorescein angiography (FA) can more easily detect some forms of subtle drusen, e.g. cuticular drusen [4], with the distinct disadvantage of being invasive and requiring intravenous dye injection. Fundus auto-fluorescence (FAF) images, though non-invasive, cannot visualize all forms of drusen [4]. Optical coherence tomography (OCT) is another widely used modality in AMD [6]. The advantage of OCT is the ability to visualize the retinal structure in high resolution, especially for characterizing leakage of blood, fluid and disorganization of the retinal structure, making it ideal for non-invasive diagnosis and monitoring of neovascular AMD and the response to anti-VEGF therapy [1]. A more comprehensive review of the different imaging modalities in AMD can also be found in [7].

#346007 https://doi.org/10.1364/BOE.10.000914
Journal © 2019 Received 9 Oct 2018; revised 4 Jan 2019; accepted 7 Jan 2019; published 28 Jan 2019

Biomedical Optics EXPRESS

In the current study, in addition to standard RGB fundus images, we investigated another image modality, hyper-spectral retinal imaging. Using a prototype device, we acquired fundus reflectance images at 16 different wavelengths, as detailed by Li *et al.* [8]. The benefit of hyper-spectral images are their ability to capture, non-invasively, a large spectral data set, with the potential to identify important biomarkers for diagnosis of AMD [9]. Lee *et al*. analyzed the hyper-spectral signatures of drusen in hyper-spectral fundus images using non-negative matrix factorization (NMF) [10]. Kaluzny *et al.* and Fawzi *et al.* further investigated hyper-spectral mapping of macular pigment [11,12]. Other researchers focused on detecting the characteristics of drusen and retinal pigment epithelium using hyper-spectral auto-fluorescence images [13–15]. In this study, we extracted Haralick texture features [16] for each individual wavelength of a hyper-spectral data set and adopted a classification approach to investigate different feature characteristics comparing drusen and non-drusen regions of interest.

Previous research has mostly focused on exploring different methods to detect drusen. These methods include image processing and computer vision techniques alone [17–23] or in combination with machine learning algorithms [24–29]. For example, Mittal and Kumari implemented a combination of gradient-based segmentation and edge linking-procedure and achieved 98.55% accuracy for detecting intermediate drusen [23]. García-Floriano *et al.* adopted a Support Vector Machine (SVM) algorithm to classify images with or without drusen and achieved an accuracy of 92.16% [28]. The limitations of previous studies include considering only green channel images, which might suffer from loss of important information. In addition, the validation data sets were small; for example, García-Floriano *et al.* only used 33 drusen images and 37 healthy tissue images [28]. In our study, we address these limitations by considering all image channels as well as generating larger validation data sets by cropping all the regions of interest.

While many low-level image features such as SIFT and SURF [30], wavelets [31], and extracted image spatial information based on histogram of orthogonal vectors or triangular regions [32,33] have been used for general image classification studies, only a few have been explored for drusen diagnosis. For example, although Haralick texture features have been widely used in the computer-aided diagnosis field, such as for lung nodules [34], liver diseases [35], and parotid-gland injury [36], their use in retinal imaging and specifically for drusen diagnosis has been limited. Prasath and Ramya used drusen texture features to segment the drusen in RGB retinal images, using only the green channel because of its higher contrast compared with the other two channels [37]. In our study, instead of setting a threshold value for drusen segmentation, we employed a classification approach to classify the drusen and non-drusen images using all 12 Haralick texture features and all 16 hyper-spectral wavelength channels. To our best knowledge, this is the first study that investigates the role of texture in drusen diagnosis using machine learning techniques and hyper-spectral retinal images.

Newer machine learning approaches based on deep learning have been recently proposed to learn directly from the raw image data rather than from extracted low-level image features. Lee *et al.* [38] implemented the deep learning method to distinguish normal OCT images from images of patients with AMD. Burlina *et al.* [39] used transfer learning and universal features derived from deep convolutional neural networks (DCNN) to classify different stages of AMD images. More recently, Schmidt Erfurth *et al.* used a deep learning approach to predict AMD progression [40]. However, training and testing deep learning algorithms require a large number of images, which makes these algorithms not applicable to settings with limited image data sets.

In summary, we aim to study the effects of texture as a biomarker for drusen and to compare the classification performance between hyper-spectral retinal images and RGB retinal images. Since we focus on lesion classification rather than lesion detection, we manually cropped drusen and healthy retinal tissue region of interest in hyper-spectral images

Biomedical Optics EXPRESS

and RGB images, respectively. Using a statistical model based on Haralick features to quantify texture and random forests to learn and classify drusen from non-drusen, we show that the classifier performs better on the hyper-spectral images. Furthermore, we found that inverse difference moment, a feature describing the local homogeneity, is the most important feature in distinguishing the drusen and non-drusen images. Our work suggests that hyper-spectral images are more sensitive to the texture feature characteristics of drusen, which may offer distinct advantages in future studies for automatic drusen detection.

2. Data and methods

Our methodology consists of three steps as illustrated in Fig. 1. First, we divided the entire imaging data set into training and testing sets. After splitting each image into different wavelength channels to create the data for learning the effect of texture as a biomarker for AMD related diseases in hyper-spectral retinal images and RGB retinal images, we manually cropped the regions of interest (ROIs). Second, we extracted intensity and texture features from the cropped images and used these features to train and test machine learning classification models of drusen versus non-drusen. Lastly, as a result of the classification model, we identified the most important texture features for distinguishing drusen regions.

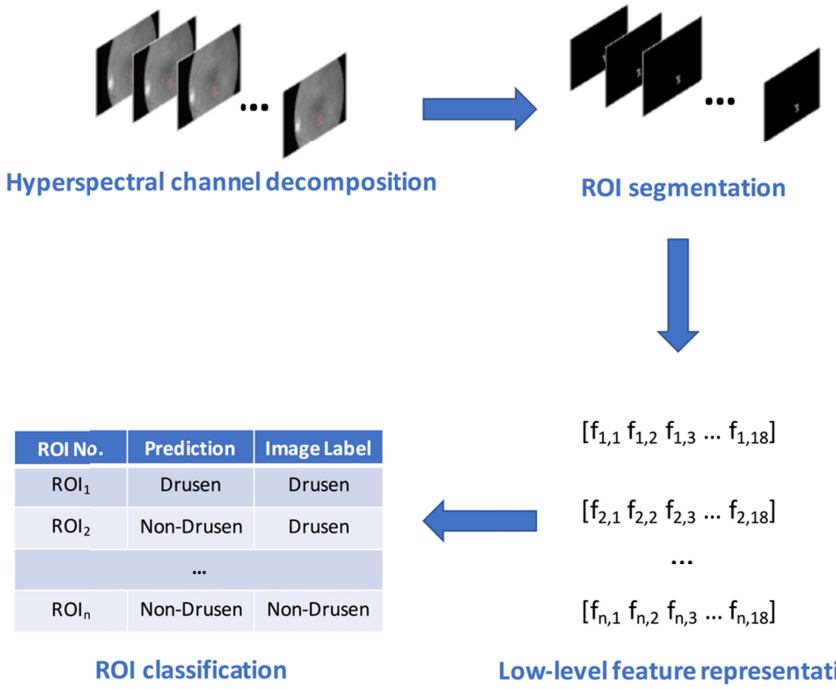

Fig. 1. The classification approach for drusen diagnosis. The process consists of cropping of ROIs, image feature extraction, and classification.

2.1 Hyper-spectral and RGB retinal images

In the study, we used hyper-spectral retinal images (1024 × 2048 pixels) generated by a compact, snapshot hyper-spectral fundus camera [8]. Each hyper-spectral image super-pixel contains 16 different wavelength channels represented by 4 × 4 complementary metal-oxide-semiconductor (CMOS) pixels. Figure 2 illustrates the structure of hyper-spectral images.

Biomedical Optics EXPRESS

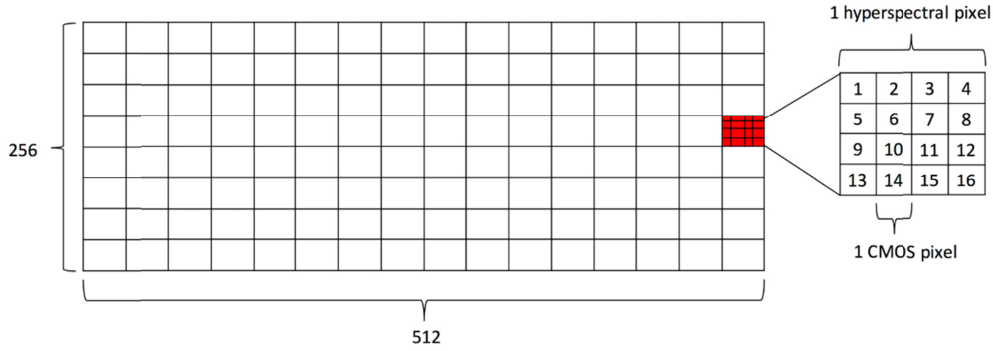

Fig. 2. An illustration of hyper-spectral imaging structure. The image resolution is 1024 × 2048. Each hyper-spectral superpixel is represented by 4 × 4 CMOS pixels.

We divided each hyper-spectral image into 16 individual wavelength channel images, ranging from 460 nm to 630 nm and each having 256 × 512 spatial resolution. The individual wavelength channel images are 8-bit grayscale images. Further details about the hyper-spectral retinal images can be found in [8,11].

We also used a separate data set of RGB retinal images and compared them with the hyper-spectral image data. The RGB retinal images were obtained using the Topcon fundus camera (TRC-50IX, Topcon, Japan) with a resolution of 2048 × 2392 × 3 channels. In our study, we used the individual channel grayscale images from the RGB retinal images with the intensity values ranging from 0 to 255.

2.2 Drusen and healthy tissue regions

For hyper-spectral images, we manually cropped drusen regions using the Matlab 'imfreehand' tool on a single wavelength image [41], and retained the drusen intensity values while converting intensity values everywhere else in the image to zero (Fig. 3). We used the same locations to automatically crop the drusen on the remaining 15 wavelength images. By doing this, for a single drusen in the hyper-spectral image, we obtain 16 cropped grayscale drusen images with 256 × 512 spatial resolution. We repeated the cropping process for all the drusen in the hyper-spectral images.

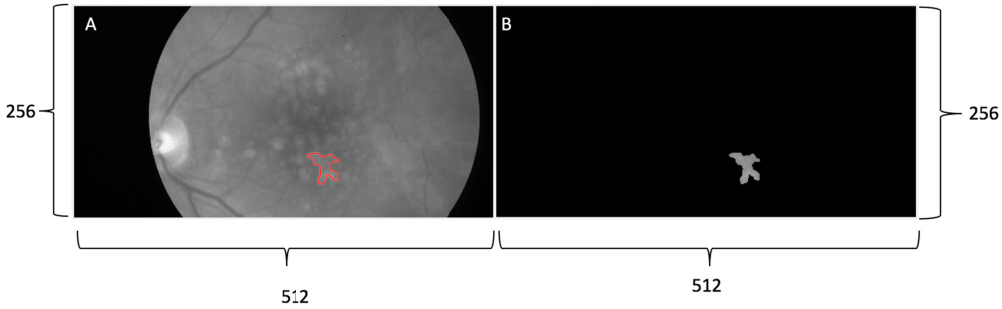

Fig. 3. The process of cropping one drusen in the image. (A) Drusen image: one single spectral image with 256 × 512 pixels of resolution. The irregular outline in red encloses the cropping area. (B) ROI drusen image: all the pixels values except for the cropping area were converted to zero. The image resolution remains the same.

For the RGB images, we employed the same cropping method. On a single channel image, we manually cropped a drusen and used the same cropped location for the other two channel images. Thus, for a single drusen in the RGB image, we obtain 3 cropped grayscale

Biomedical Optics EXPRESS

drusen images with 2048 × 2392 spatial resolution. We repeated the cropping process to extract the drusen in all the RGB images.

In order to generate healthy tissue ROIs, we applied the same method to crop healthy tissue regions in the images. During the health tissue cropping process, we avoided cropping retinal blood vessels or the optic disk and used similar sized cropped regions as used in drusen ROI.

Table 1 summarizes the hyper-spectral and RGB data used in this study. We generated 2032 cropped drusen and 2032 cropped healthy tissue images from all 16 wavelengths hyper-spectral images. In total, we cropped 1731 drusen and 1785 healthy tissue ROIs from all 3 channels of the RGB images.

Table 1. A summary of hyper-spectral data and RGB data set

	Drusen Images	Drusen ROIs	Healthy Retinal Images	Healthy Retinal ROIs
Hyper-spectral	22	2032	6	2032
RGB	30	1731	7	1785

2.3 Image feature extraction

For each ROI image, we extracted 6 intensity features and 12 Haralick texture features. The 6 intensity features are mean, median, max, min, standard deviation and the range of the intensity values within the cropped region. The 12 Haralick texture features are calculated from a gray level co-occurrence matrix (GLCM) [42] that encodes the spatial relationship of the gray-levels across a certain direction θ and pixel displacement d.

Suppose we quantized pixel values in the image into L_g levels, then the GLCM L_g has number of rows and L_g number of columns. Each element in GLCM represents the probability $p(i,j)$ that the quantized gray level value i can be found adjacent to the quantized value j at angle θ and displacement d:

$$p(i,j) = \frac{f(i,j)}{\sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} f(i,j)}, \tag{1}$$

where $f(i,j)$ is the frequency of quantized value i and value j appearing together in the search at angle θ and displacement d. Figure 4 shows an example of the GLCM.

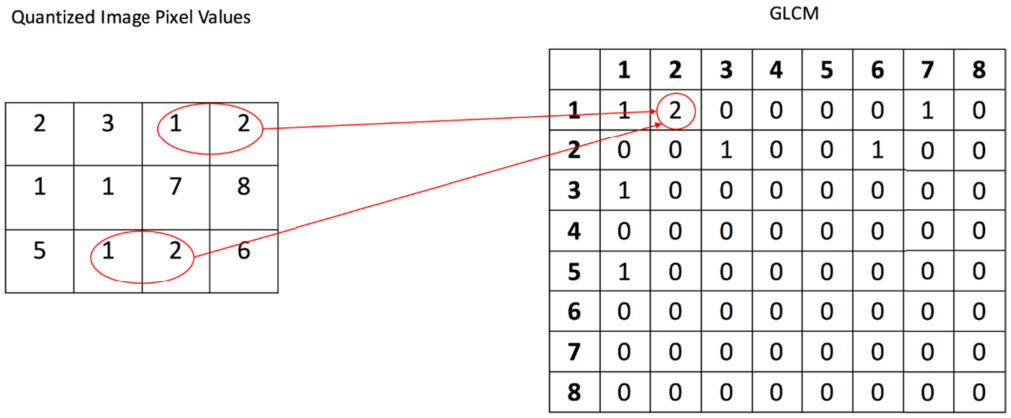

Fig. 4. An example of GLCM calculation. Left: image with quantized gray level $L_g = 8$. Right: the GLCM matrix for $d = 1$ and $\theta = 0^o$.

Biomedical Optics EXPRESS

In our study, we calculated 4 GLCM for $L_g = 8$, $d = 1$, and $\theta = 0^o, 45^o, 90^o, 180^o$; for each GLCM we extracted 12 Haralick texture features that were averaged with respect to the angle. The texture features are angular-second moment (energy), contrast, correlation, variance, inverse difference moment, sum average, sum variance, entropy, sum entropy, difference variance, difference entropy, information measure of correlation and maximal correlation coefficient. This set of texture features is chosen to quantify second-order gray level properties such as local uniformity, variance, and homogeneity. The Appendix Table 15 summarizes the definitions of these features [43]. We also provided the link to download the data sets in this study (Dataset 1 [44]).

2.4 Drusen vs. non-drusen classification

To determine whether hyper-spectral images offer an improved outcome compared to standard RGB images based on intensity and texture features, we implemented four binary classifiers to differentiate between healthy (non-drusen) and non-healthy (drusen) tissues: decision trees, naïve Bayes, AdaBoost with stump trees, and random forests. First, we examined different split ratios between the training and testing sets (80%-20%, 70%-30%, 60%-40% and 50%-50%). For each classifier with a certain training vs. testing split ratio, we repeated the process 30 times and calculated the mean accuracy, sensitivity (drusen is the positive case) and specificity (non-drusen is the negative case) under 95% confidence interval. Second, we compared the classification results for hyper-spectral and RGB data sets by testing the mean accuracies and mean sensitivities (calculated across the 30 trials) between different combinations of training vs. testing split ratio, classifier type, and image modality using Welch's *t*-test [45]. Since we had more hyper-spectral ROIs, we under-sampled the hyper-spectral ROI image data set to balance it with the number of ROIs present in the RGB images. In the rest of this section we provide the mathematical details for each one of the four classifiers.

Decision Tree is a greedy algorithm that constructs a classification tree in a top-down, recursive, divide-and-conquer manner [46]. A decision tree can be represented as a flow-chart-like tree structure, where the root node represents all the samples S, each internal node represents a test on a feature A, the outcome of the test is represented by a branch, and the leaf nodes are target class distributions for m distinct classes $C_i (i = 1, \ldots, m)$ [46]. The algorithm first starts with the root node; if the samples belong to the same class, then the node becomes a leaf node and is labeled with the class. Otherwise, the algorithm uses information gain $I(s_1, s_2, \ldots, s_m)$ to select a feature A that becomes the test feature at that node and divides the samples into different groups:

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i), \tag{2}$$

where $p_i = \frac{s_i}{s}$ and s_i is the number of samples of S in class C_i.

If the feature A has v different values, $\{a_1, a_2, \ldots, a_v\}$, then the feature A can be used to partition S into v subsets, $\{S_1, S_2, \ldots, S_v\}$. The entropy is defined as

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I(s_{1j}, \cdots, s_{mj}) \tag{3}$$

where s_{ij} represents the number of samples that have value a_j for feature A and belong to class C_i and s represents the number of samples at the partition node. The information gain by branching on feature A is

Biomedical Optics EXPRESS

$$Gain(A) = I(s_1, s_2, \cdots, s_m) - E(A) \tag{4}$$

The algorithm chooses the attribute with the highest information gain to separate the samples and uses the same process recursively for the samples at each partition node. The recursive partition stops when all the samples in a node belong to the same class or there are no more features or samples to split the node.

In this study, we implemented 10-fold cross validation method [47] on a training set to find the optimal configuration of the decision tree that leads to the minimum average cross-validation error.

Naïve Bayes is a probabilistic classifier that implements Bayes' theorem with the assumption that all the features are independent. However, Pedro *et al.* [48] found that even in the situation where features are dependent, Naïve Bayes can have a better classification performance. Suppose we have a new instance x with n features $x = (A_1, A_2, \cdots, A_n)$, the predicted class C is defined as

$$C(x) = \arg\max_{k \in \{1,\ldots,m\}} p(C_k) \prod_{l=1}^{n} p(x_i \mid C_k) \tag{5}$$

AdaBoost is an ensemble learning classifier that combines weak learners and assigns weights to training instances and weak learners h_t, $t = 1 \cdots T$, where T is the total number of learners. The algorithm assigns higher weights to most likely misclassified cases [49]. In this study, we choose stump trees as our weak learners. In the first iteration, the algorithm gives equal weight D to all the training instances $(x_1, y_1) \cdots (x_s, y_s)$ where y_i belongs to class C_i:

$$D_1(i) = \frac{1}{S}, \text{ for } i = 1, \cdots, S \tag{6}$$

The weight for a weak classifier h_t is defined as

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right), \tag{7}$$

where ε_t is the classification error at iteration t. The updated weight at iteration $t+1$ is defined as

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{z_t}, \tag{8}$$

where z_t is the normalization factor. The output of the final hypothesis is

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{9}$$

We also implemented 10-fold cross validation on training set to find the optimal number of iterations that leads to the minimal average cross validation error.

Random Forest is an ensemble of classifiers that creates a group of decision trees from the original data by bootstrapping and then randomly choosing features to build the trees [50]. Because of its randomness, random forest is robust to outliers and overfitting problems. The algorithm classifies in instance by a majority vote across all the classification outputs of the individual decision trees [50,51].

To determine the optimal number of features and trees, we examined the "Out-of-Bag" (OOB) error as a measurement of classification performance. Figure 5 illustrates the calculation process. OOB error is the average error of all the instances in the data set while each instance error is the average error of all the trees that do not select the instance.

Biomedical Optics EXPRESS

Mathematically, for each tree h_t, we denote the portion of the training set that contributes to OOB as T_t^{OOB} [52] and there are $(T - T_t^{OOB})_i$ trees that do not use observation x_i. The OOB error is defined as

$$OOB(x_i) = (1/(T - T_t^{OOB})_i)\sum_{t=1}^{T}[\hat{h}_t(x_i)I((y_i, x_i) \in T_t^{OOB}) - f_t(x_i)], \quad (10)$$

where $\hat{h}_t(x_i)$ is the prediction of $h_t(x_i)$ from the h_t tree and $I(\cdot)$ is the indicator function.

We select the optimal combination of the number of features selected at each split (n_f) and the number of trees (T) that will determine the minimum OOB error.

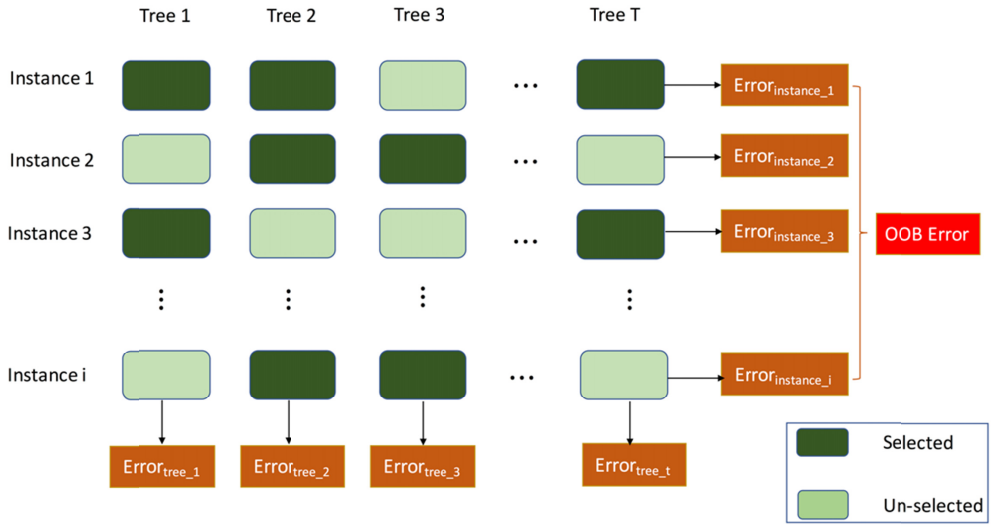

Fig. 5. An illustration of the OOB error calculation. OOB error is the average of all the instance errors. Each instance error is the average error of trees that do not select the instance.

To determine the most relevant feature when choosing a split for the random forest, we used the Gini index defined as:

$$Gini(node) = \sum_{i=1}^{m} p_i(1 - p_i), \quad (11)$$

where $p_i = \frac{s_i}{s}$ and s_i is the number of samples in class C_i at that corresponding node that has s samples in total.

3. Results

We present and compare the classification results using different split ratios in Section 3.1, compare the classification results using different classifiers within the same image modality in Section 3.2 and across modalities in Section 3.3, and analyze the importance of the image features in Section 3.4.

3.1 Classification results using different split ratios

For each combination of training-testing split ratio and type of classifiers, we repeated the process of shuffling and splitting the data 30 times. Tables 2, 3 and 4 show the mean

Biomedical Optics EXPRESS

accuracy, sensitivity and specificity under the 95% confidence interval in the hyper-spectral image testing set, respectively.

Table 2. Mean accuracy of different split ratios in the hyper-spectral image test data

	80% - 20%	70%-30%	60%-40%	50%-50%
Random Forest	**95.05% ± 0.34%**	94.73% ± 0.21%	94.58% ± 0.23%	94.26% ± 0.21%
Decision Tree	88.59% ± 0.49%	88.06% ± 0.47%	**88.72% ± 0.37%**	88.03% ± 0.55%
Naïve Bayes	79.69% ± 0.61%	**79.77% ± 0.4%**	79.70% ± 0.34%	79.77% ± 0.3%
AdaBoost	**92.79% ± 0.46%**	92.3% ± 0.41%	92.53% ± 0.32	92.43% ± 0.35%

Table 3. Mean sensitivity of different split ratios in the hyper-spectral image test Data

	80% - 20%	70%-30%	60%-40%	50%-50%
Random Forest	**94.87% ± 0.35%**	93.96 ± 0.36	93.88 ± 0.40	93.42 ± 0.41
Decision Tree	**87.19% ± 0.81%**	85.98 ± 1.08	86.78 ± 0.94	85.37 ± 1.28
Naïve Bayes	**78.53% ± 0.96%**	78.35 ± 0.62	78.36 ± 0.75	78.21 ± 0.45
AdaBoost	**91.31% ± 0.77%**	90.44 ± 0.61	90.77 ± 0.45	90.91 ± 0.53

Table 4. Mean specificity of different split ratios in the hyper-spectral image test Data

	80% - 20%	70%-30%	60%-40%	50%-50%
Random Forest	95.24% ± 0.54%	**95.48% ± 0.3%**	95.26% ± 0.33%	95.08% ± 0.22%
Decision Tree	89.95% ± 0.73%	90.09% ± 0.85%	90.60% ± 0.57%	**90.62% ± 0.86%**
Naïve Bayes	80.82% ± 0.95%	81.14% ± 0.64%	80.99% ± 0.55%	**81.30% ± 0.42%**
AdaBoost	94.23% ± 0.58%	94.10% ± 0.5%	**94.24% ± 0.42%**	93.91% ± 0.41%

We notice that all four classifiers achieved the highest mean sensitivity (Table 3) for classifying drusen regions under the split ratio 80%-20%. Random forest and AdaBoost also have the highest mean accuracy with the split ratio 80%-20% (Table 2). We further implemented Welch's *t*-test to statistically determine whether the difference of the mean sensitivity between 80%-20% with other split ratios is significant (Table 5).

Table 5. P-values of Welch's *t*-test when comparing the mean sensitivity between the 80%-20% split ratio with other split ratios (hyper-spectral image)

		70%-30%	60%-40%	50%-50%
80%-20%	Random Forest	**0.0005299**	**0.0003899**	**1.25e-06**
	Decision Tree	0.07282	0.5029	**0.01769**
	Naïve Bayes	0.7568	0.7868	0.5437
	AdaBoost	0.07719	0.2243	0.3894

From Table 5, we can conclude that when we use random forest algorithm in hyper-spectral image data, there is a significant difference of the mean sensitivity between the 80%-20% split ratio with other split ratios. Random forest algorithm can achieve the highest mean sensitivity with 80%-20% ratio. When we use decision tree algorithm, there is a different of mean sensitivity between 80%-20% and 50%-50%. However, there is no difference of mean sensitivity between different split ratios for other combinations.

We repeated the same classification process for RGB image data. Tables 6, 7 and 8 show the mean accuracy, sensitivity and specificity respectively under the 95% confidence interval when using the RGB image.

Table 6. Mean accuracy of different split ratios in the RGB image test data

	80% - 20%	70%-30%	60%-40%	50%-50%
Random Forest	**87.91% ± 0.44%**	87.50% ± 0.33%	87.04% ± 0.23%	87.14% ± 0.26%
Decision Tree	**84.05% ± 0.47%**	83.79% ± 0.37%	83.40% ± 0.29%	83.47% ± 0.42%
Naïve Bayes	67.12% ± 0.76%	**67.24% ± 0.52%**	67.16% ± 0.60%	68.23% ± 0.71%
AdaBoost	**88.25% ± 0.49%**	87.78% ± 0.36%	87.32% ± 0.23%	87.26% ± 0.26%

Biomedical Optics EXPRESS

Table 7. Mean sensitivity of different split ratios in the RGB image test data

	80% - 20%	70%-30%	60%-40%	50%-50%
Random Forest	**87.20% ± 0.54%**	86.63% ± 0.58%	86.33% ± 0.57%	86.12% ± 0.64%
Decision Tree	82.82% ± 0.96%	**82.97% ± 0.80%**	82.64% ± 1.15%	81.16% ± 1.20%
Naïve Bayes	50.57% ± 1.41%	50.21% ± 1.02%	50.49% ± 1.03%	**51.86% ± 1.10%**
AdaBoost	**87.63% ± 0.58%**	86.85% ± 0.56%	86.63% ± 0.46%	86.56% ± 0.62%

Table 8. Mean specificity of different split ratios in the RGB image test data

	80% - 20%	70%-30%	60%-40%	50%-50%
Random Forest	**88.61% ± 0.64%**	88.34% ± 0.52%	87.73% ± 0.49%	88.14% ± 0.57%
Decision Tree	85.24% ± 1.04%	84.57% ± 0.62%	84.15% ± 1.23%	**85.71% ± 1.08%**
Naïve Bayes	83.17% ± 0.78%	83.75% ± 0.65%	83.31% ± 0.61%	**84.09% ± 0.72%**
AdaBoost	**88.84% ± 0.70%**	88.69% ± 0.57%	87.99% ± 0.48%	87.94% ± 0.52%

From Table 6, we can see that random forest, decision tree and AdaBoost achieved the highest mean accuracy with the split ratio 80%-20%. When considering mean sensitivity and mean specificity, random forest and AdaBoost also performed the best with the split ratio 80%-20% (Tables 7 and 8). Similarly, we implemented Welch's *t*-test to statistically determine whether the difference of the mean accuracy between 80%-20% with other split ratios is significant (Table 9).

Table 9. P-values of Welch's *t*-test when comparing the mean accuracy between the 80%-20% split ratio with other split ratios (RGB image)

		70%-30%	60%-40%	50%-50%
80%-20%	Random Forest	0.1306	**0.0008286**	**0.003449**
	Decision Tree	0.3723	**0.02169**	0.06518
	Naïve Bayes	0.8015	0.9444	**0.03347**
	AdaBoost	0.1232	**0.001051**	**0.0007104**

From Table 9, we can see that in RGB image, there is no significant difference of mean accuracy between 80%-20% and 70%-30% for both classifiers under the 95% confidence interval. However, for random forest and AdaBoost classifiers, the difference of mean accuracy between the split ratio 80%-20% and 60%-40% and the difference between 80%-20% and 50%-50% are significant. This result indicates that we can choose either 80%-20% or 70%-30% as the spit ratio for RGB image data.

3.2 Classification results using different classifiers

Based on the split ratio results, we analyzed the classification performance across the four classifiers using the 80%-20% as the split ratio. Table 10 summarizes the results across classifiers using a 80%-20% ratio and shows that the random forest classifier achieved the highest accuracy, sensitivity and specificity for the hyper-spectral image data set based on the Welch's *t*-test at significance level of 0.05.

Table 10. Classification performance for hyper-spectral image data set; the numbers between parentheses represent P-values of Welch's *t*-test when comparing random forest with the other classifiers

	Accuracy	Sensitivity	Specificity
Random Forest	**95.05% ± 0.34%**	**94.87% ± 0.35%**	**95.24% ± 0.54%**
Decision Tree	88.59% ± 0.49% **(< 2.2e-16)**	87.19% ± 0.81% **(< 2.2e-16)**	89.95% ± 0.73% **(< 2.2e-16)**
Naïve Bayes	79.69% ± 0.61% **(< 2.2e-16)**	78.53% ± 0.96% **(< 2.2e-16)**	80.82% ± 0.95% **(< 2.2e-16)**
AdaBoost	92.79% ± 0.46% **(9.35e-11)**	91.31% ± 0.77% **(1.173e-10)**	94.23% ± 0.58% **(0.01189)**

Biomedical Optics EXPRESS

Table 11 summarizes the results across classifiers using an 80%-20% ratio and although it shows that the AdaBoost classifier achieved the highest accuracy, sensitivity and specificity for the RGB image data set, the Welch's *t*-test showed no statistically significant difference between AdaBoost and random forest at significance level of 0.05. This result indicates that for RGB image data, we can also choose random forest as the classifier. Figure 6 shows the optimal random forests parameter values for both the RGB and hyperspectral image data.

Table 11. Classification performance on the RGB image data set; the numbers between parentheses represent P-values of Welch's *t*-test when comparing AdaBoost with the other classifiers

	Accuracy	Sensitivity	Specificity
Random Forest	87.91% 7.91m % (0.304)	87.20% 7.20m % (0.2671)	88.61% 8.611% (0.6167)
Decision Tree	84.05% ± 0.47% (< 2.2e-16)	82.82% ± 0.96% (1.573e-11)	85.24% ± 1.04% (3.2e-07)
Naïve Bayes	67.12% ± 0.76% (< 2.2e-16)	50.57% ± 1.41% (< 2.2e-16)	83.17% ± 0.78% (8.039e-16)
AdaBoost	**88.25% ± 0.49%**	**87.63% ± 0.58%**	**88.84% ± 0.70%**

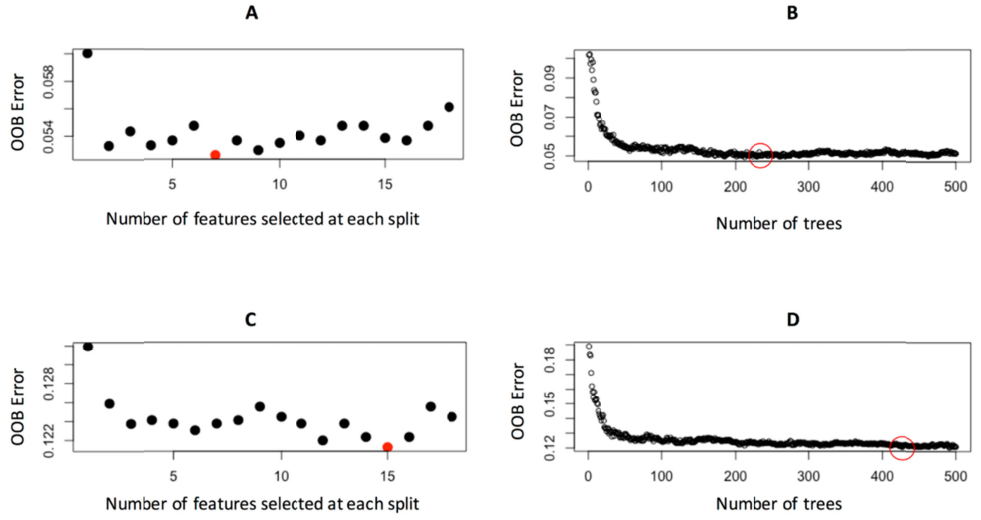

Fig. 6. The optimal parameters of random forest algorithm. (A and B) The number of features selected at each split is 7 when the number of trees is 212 on hyper-spectral image data; (C and D) The number of features selected at each split is 15 when the number of trees is 427 on RGB image data

3.3 Classification results comparing the hyper-spectral and RGB images

Furthermore, to determine whether there are significant statistical differences between the RGB and hyperspectral random forest classification results in Tables 10 and 11, we performed Welch's *t*-test (Table 12). The results indicate that the differences across all three performance metrics (accuracy, sensitivity, specificity) are significant and therefore, we can conclude that hyper-spectral images perform significantly better than RGB images for drusen ROIs identification.

Table 12. P-values of Welch's *t*-test for random forest classification performance comparison between hyper-spectral and RGB image

	RGB Data		
p-values	Accuracy	Sensitivity	Specificity
Hyper-spectral Data	**< 2.2e-16**	**< 2.2e-16**	**< 2.2e-16**

Biomedical Optics EXPRESS

3.4 Low-level image feature importance analysis

We further investigated the random forest classification performance by the type of features. We focus the analysis of the results on the hyper-spectral image data given the higher performance for drusen classification. Table 13 compares the classification results when using intensity-based features, texture-based features, and a combination of intensity and texture features. The results show that the highest performance is obtained using a combination of texture and intensity features, followed in performance by the texture features. The mean differences for all accuracy, sensitivity, and specificity values are all significant based on the Welch's *t*-test (Tables 13 and 14).

Table 13. Random forest classification result in hyper-spectral testing data using different feature sets

	Accuracy	Sensitivity	Specificity
Intensity Features	90.53% ± 0.21% (< 2.2e-16)	89.92% ± 0.40% (< 2.2e-16)	91.12% ± 0.31% (< 2.2e-16)
Texture Features	92.39% ± 0.19% (< 2.2e-16)	90.83% ± 0.47% (< 2.2e-16)	93.92% ± 0.28% 6.063e-05
Combined Features	**95.05% ± 0.34%**	**94.87% ± 0.35%**	**95.24% ± 0.54%**

Table 14. P-values of Welch's *t*-test when comparing classification performance using intensity features vs texture features

	Texture Features		
p-values	Accuracy	Sensitivity	Specificity
Intensity Features	**< 2.2e-16**	**0.003531**	**0.003531**

To understand the relevance of the individual low-level image features that distinguished drusen ROI from non-drusen ROI, we used the Gini index criterion (Eq. (11) to rank the feature importance when building the random forest on all features (both texture and intensity). Figure 7 shows the most important low-level image features with the 'inverse difference moment", a feature describing the local homogeneity in a region, being the most important (it has the largest value for the mean decrease in the Gini index).

As a result, we analyzed the differences in the inverse difference moment features for the drusen versus non-drusen ROIs. Based on the definition of the feature, a low inverse difference moment value indicates the image is heterogeneous while a higher value indicates the region is more homogeneous. Figure 8 shows that, on average, drusen images are more heterogeneous confirmed by the statistically significant Welch *t*-test at a p-value smaller than 2.2e-16.

Biomedical Optics EXPRESS

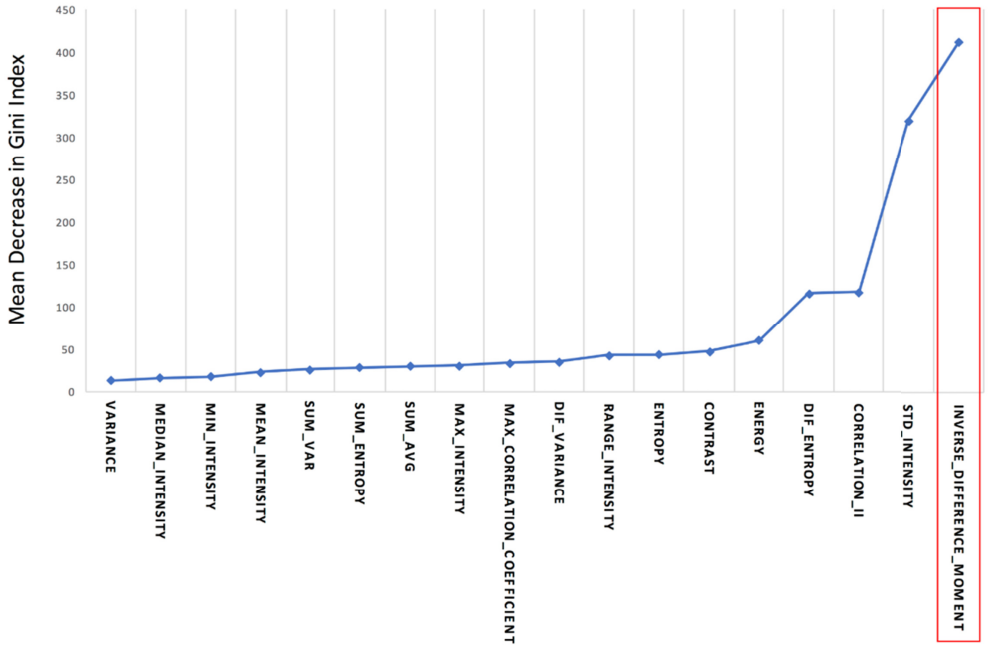

Fig. 7. Mean decrease in Gini Index of each feature. Inverse difference moment is the most important feature since it has the highest mean decrease in Gini Index.

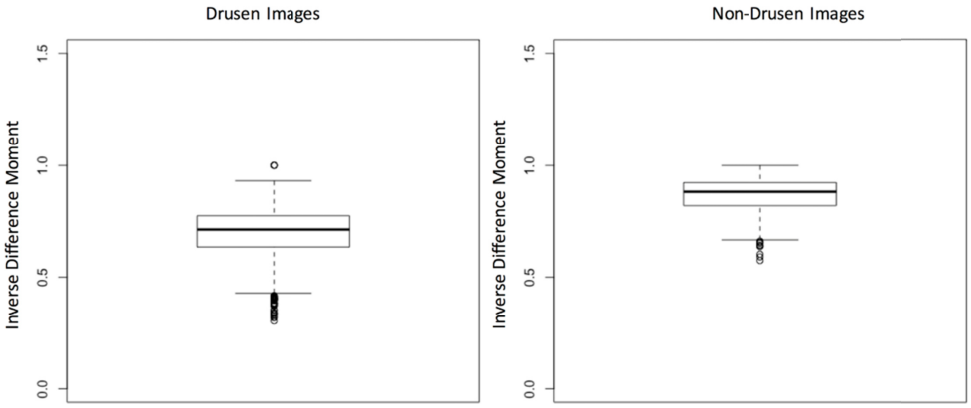

Fig. 8. The distribution of inverse difference moment. Drusen images have relatively low inverse difference moment, which indicate that drusen images are more heterogeneous.

4. Discussion

Our results demonstrate that hyper-spectral images can be superior to RGB images when used in the diagnosis of AMD-related disease process. Among four different classifiers, two base classifiers and two ensembles of classifiers, the random forest achieved the best classification performance for the hyper-spectral image data set and had a similar classification performance to the AdaBoost for the RGB data. When comparing the hyper-spectral data with the RGB data, we learned that the hyper-spectral data is characterized by certain texture properties that quantify better the differences between drusen and non-drusen. In particular,

Biomedical Optics EXPRESS

we determined mathematically that the texture heterogeneity is an important local image characteristic that has higher values for the drusen images.

Furthermore, when comparing the random forest classifiers for the hyper-spectral images and the RGB images (Fig. 6), we found that the classification model for the RGB data needed a higher number of features (15) per split and more trees (427) to achieve the optimal combination of parameters than hyper-spectral images that required only 7 features and 212 trees. These findings indicate that classification models for the hyper-spectral data are not only superior in performance but also have a lower complexity with only few image characteristics needed to distinguish between drusen and non-drusen.

In the context of previous studies, our work validates and extends the work by Prasath and Ramya [37] that showed that thresholding certain texture features can help segment drusen regions in RGB images. By using a robust Haralick set of features (averaged across different displacements and angles) and a machine learning algorithm, we determined the most important texture features and their combinations with intensity features for drusen diagnosis. Finally, instead of using only the green channel as in [37] and [53] where local binary patterns (LBP) features computed in green channel where reported to be the most important features in distinguishing drusen from non-drusen images, we showed that hyper-spectral imaging has the potential to provide the optimal combination of texture and intensity features for drusen ROIs characterization.

5. Conclusions

Using hyper-spectral retinal images containing 16 different wavelength channels generated by a compact, snapshot hyper-spectral fundus camera [8], we showed the potential advantages of hyper-spectral imaging for retinal disease diagnosis. We discovered that drusen ROIs are more heterogeneous than the surrounding retinal tissue, a property that can be quantified mathematically through one of the Haralick texture feature, the inverse difference moment.

As future work, we plan to investigate the effect of the location and size of drusen on the classification. For example, can we answer questions like ‘Is there any difference between drusen centrally located and those near the arcades using texture descriptors’? Augmenting the approach presented in this paper with a patch-based segmentation approach as proposed in [54] will allow the extension of this work to automatic segmentation of ROIs and eliminate the need for manual cropping. This would then allow us to perform automatic drusen classification as well as detection.

Furthermore, newer techniques such as deep learning have been recently explored in retinal imaging [38,39] and resulted in promising results. Since these deep learning approaches require large image training sets, we plan to acquire a larger hyper-spectral image data set and compare the performance of the feature-based random forest classifier with the deep learning classification approaches.

Biomedical Optics EXPRESS

Appendix

Table 15. Haralick's Texture Features Employed in the Study

Angular Second Moment (Energy)	$\sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} p(i,j)^2$
Contrast	$\sum_{m=0}^{L_g-1} m^2 \left\{ \sum_{i=1}^{L_g}\sum_{j=1}^{L_g} p(i,j) \,\middle|\, \lvert i-j\rvert = m \right\}$
Variance	$\sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} (i-\mu_c)^2 * p(i,j) + \sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} (j-\mu_y)^2 * p(i,j)$ where $\mu_c = \sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} i * p(i,j)$ $\mu_r = \sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} j * p(i,j)$ $\sigma_c = \sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} (i-\mu_x)^2 * p(i,j)$ $\sigma_r = \sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} (j-\mu_x)^2 * p(i,j)$
Inverse Difference Moment	$\sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} \frac{1}{1+(i-j)^2} p(i,j)$
Sum Average, μ_{c+r}	$\sum_{m=2}^{2L_g} m P_{c+r}(m)$ where $P_{c+r}(m) = \sum_{i=1}^{L_g}\sum_{j=1}^{L_g} p(i,j)$, $m = 2,3,\cdots,2L_g$ $i+j=m$
Sum Variance	$\sum_{m=2}^{2L_g} (m-\mu_{c+r})^2 P_{c+r}(m)$
Entropy	$-\sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} p(i,j)\log p(i,j)$
Sum Entropy	$-\sum_{m=2}^{2L_g} P_{c+r}(m)\log P_{c+r}(m)$
Difference Variance	$\sum_{m=0}^{Lg-1} (m-\mu_{c-r})^2 P_{c-r}(m)$ where $P_{c-r}(m) = \sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} p(i,j)$, $m = 0,1,\cdots,L_g-1$ $\lvert i-j\rvert = m$
Difference Entropy	$-\sum_{m=0}^{L_g-1} P_{c-r}(m)\log P_{c-r}(m)$

Biomedical Optics EXPRESS

Information measure of correlation 2	$\sqrt{1-\exp[-2(HXY_2 - HXY)]}$ where $HXY = -\sum_{i=0}^{L_g-1} p(i,j) * \log p(i,j)$ $HXY_2 = -\sum_{i=0}^{L_g-1}\sum_{j=0}^{L_g-1} p_c(i) * p_r(j) * \log[p_c(i) * p_r(j)]$ $p_c(i) = \sum_{j=0}^{L_g-1} p(i,j)$ $p_r(j) = \sum_{i=0}^{L_g-1} p(i,j)$
Maximal Correlation Coefficient	$\sqrt{The \text{ second largest eigenvalue of } Q}$ where $Q(i,j) = \sum_{m=0}^{L_g-1} \frac{p(i,k) * p(j,k)}{p_c(i) * p_r(m)}$

Funding

National Institutes of Health (NIH) (DP3DK108248).

Disclosures

The authors declare that there are no competing interests regarding the publication of this paper.

References

1. J. Ma, R. Desai, P. Nesper, M. Gill, A. Fawzi, and D. Skondra, "Optical coherence tomographic angiography imaging in age-related macular degeneration," Ophthalmol. Eye Dis. **9**, 1179172116686075 (2017).
2. L. Roisman and R. Goldhardt, "OCT Angiography: An Upcoming Non-invasive Tool for Diagnosis of Age-related Macular Degeneration," Curr. Ophthalmol. Rep. **5**(2), 136–140 (2017).
3. Z. Yehoshua, P. J. Rosenfeld, G. Gregori, W. J. Feuer, M. Falcão, B. J. Lujan, and C. Puliafito, "Progression of geographic atrophy in age-related macular degeneration imaged with spectral domain optical coherence tomography," Ophthalmology **118**(4), 679–686 (2011).
4. R. Zhao, A. Camino, J. Wang, A. M. Hagag, Y. Lu, S. T. Bailey, C. J. Flaxel, T. S. Hwang, D. Huang, D. Li, and Y. Jia, "Automated drusen detection in dry age-related macular degeneration by multiple-depth, *en face* optical coherence tomography," Biomed. Opt. Express **8**(11), 5049–5064 (2017).
5. E. Tsikata, I. Laíns, J. Gil, M. Marques, K. Brown, T. Mesquita, P. Melo, M. da Luz Cachulo, I. K. Kim, D. Vavvas, J. N. Murta, J. B. Miller, R. Silva, J. W. Miller, T. C. Chen, and D. Husain, "Automated brightness and contrast adjustment of color fundus photographs for the grading of age-related macular degeneration," Transl. Vis. Sci. Technol. **6**(2), 3–3 (2017).
6. M. A. Hussain, A. Bhuiyan, C. D Luu, R. Theodore Smith, R. H Guymer, H. Ishikawa, J. S Schuman, and K. Ramamohanarao, "Classification of healthy and diseased retina using SD-OCT imaging and Random Forest algorithm," PLoS One **13**(6), e0198281 (2018).
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," Prog. Retin. Eye Res. **38**, 20–42 (2014).
8. H. Li, W. Liu, B. Dong, J. V. Kaluzny, A. A. Fawzi, and H. F. Zhang, "Snapshot hyperspectral retinal imaging using compact spectral resolving detector array," J. Biophotonics **10**(6-7), 830–839 (2017).
9. M. A. Sohrab, R. T. Smith, and A. A. Fawzi, "Imaging characteristics of dry age-related macular degeneration," in *Seminars in Ophthalmology* (Taylor & Francis, 2011), pp. 156–166.
10. N. Lee, J. Wielaard, A. Fawzi, P. Sajda, A. Laine, G. Martin, M. Humayun, and R. Smith, "In vivo snapshot hyperspectral image analysis of age-related macular degeneration," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE* (IEEE, 2010), pp. 5363–5366.
11. J. Kaluzny, H. Li, W. Liu, P. Nesper, J. Park, H. F. Zhang, and A. A. Fawzi, "Bayer filter snapshot hyperspectral fundus camera for human retinal imaging," Curr. Eye Res. **42**(4), 629–635 (2017).
12. A. A. Fawzi, N. Lee, J. H. Acton, A. F. Laine, and R. T. Smith, "Recovery of macular pigment spectrum in vivo using hyperspectral image analysis," J. Biomed. Opt. **16**(10), 106008 (2011).
13. Y. Tong, T. Ben Ami, S. Hong, R. Heintzmann, G. Gerig, Z. Ablonczy, C. A. Curcio, T. Ach, and R. T. Smith, "Hyperspectral autofluorescence imaging of drusen and retinal pigment epithelium in donor eyes with age-related macular degeneration," Retina **36**(Suppl 1), S127–S136 (2016).

Biomedical Optics EXPRESS

14. T. B. Feldman, M. A. Yakovleva, A. V. Larichev, P. M. Arbukhanova, A. S. Radchenko, S. A. Borzenok, V. A. Kuzmin, and M. A. Ostrovsky, "Spectral analysis of fundus autofluorescence pattern as a tool to detect early stages of degeneration in the retina and retinal pigment epithelium," Eye (Lond.) **32**(9), 1440–1448 (2018).
15. J. J. Foubister, A. Gorman, A. Harvey, and J. van Hemert, "Spectral Autofluorescence imaging of the retina for drusen detection," in *Ophthalmic Technologies XXVIII* (International Society for Optics and Photonics, 2018), p. 104741H.
16. R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," IEEE Trans. Syst. Man Cybern. **3**(6), 610–621 (1973).
17. R. T. Smith, T. Nagasaki, J. R. Sparrow, I. Barbazetto, C. C. Klaver, and J. K. Chan, "A method of drusen measurement based on the geometry of fundus reflectance," Biomed. Eng. Online **2**(1), 10 (2003).
18. B. Remeseiro, N. Barreira, D. Calvo, M. Ortega, and M. G. Penedo, "Automatic drusen detection from digital retinal images: AMD prevention," in *International Conference on Computer Aided Systems Theory* (Springer, 2009), pp. 187–194.
19. A. D. Mora, P. M. Vieira, A. Manivannan, and J. M. Fonseca, "Automated drusen detection in retinal images using analytical modelling algorithms," Biomed. Eng. Online **10**(1), 59 (2011).
20. D. W. Wong, J. Liu, X. Cheng, J. Zhang, F. Yin, M. Bhargava, G. C. Cheung, and T. Y. Wong, "THALIA-An automatic hierarchical analysis system to detect drusen lesion images for amd assessment," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on* (IEEE, 2013), pp. 884–887.
21. M. R. K. Mookiah, U. R. Acharya, J. E. Koh, V. Chandran, C. K. Chua, J. H. Tan, C. M. Lim, E. Y. Ng, K. Noronha, L. Tong, and A. Laude, "Automated diagnosis of Age-related Macular Degeneration using greyscale features from digital fundus images," Comput. Biol. Med. **53**, 55–64 (2014).
22. K. Kumari and D. Mittal, "Automated drusen detection technique for age-related macular degeneration," Journal of Biomedical Engineering and Medical Imaging **2**, 18 (2015).
23. D. Mittal and K. Kumari, "Automated detection and segmentation of drusen in retinal fundus images," Comput. Electr. Eng. **47**, 82–95 (2015).
24. G. Raza, M. Rafique, A. Tariq, and M. U. Akram, "Hybrid classifier based drusen detection in colored fundus images," in *Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on* (IEEE, 2013), pp. 1–5.
25. Y. Zheng, B. Vanderbeek, E. Daniel, D. Stambolian, M. Maguire, D. Brainard, and J. Gee, "An automated drusen detection system for classifying age-related macular degeneration with color fundus photographs," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on* (IEEE, 2013), pp. 1448–1451.
26. A. Bhuiyan, C. Karmakar, D. Xiao, K. Ramamohanarao, and Y. Kanagasingam, "Drusen quantification for early identification of age related macular degeneration (AMD) using color fundus imaging," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (IEEE, 2013), pp. 7392–7395.
27. J. E. W. Koh, U. R. Acharya, Y. Hagiwara, U. Raghavendra, J. H. Tan, S. V. Sree, S. V. Bhandary, A. K. Rao, S. Sivaprasad, K. C. Chua, A. Laude, and L. Tong, "Diagnosis of retinal health in digital fundus images using continuous wavelet transform (CWT) and entropies," Comput. Biol. Med. **84**, 89–97 (2017).
28. A. García-Floriano, Á. Ferreira-Santiago, O. Camacho-Nieto, and C. Yáñez-Márquez, "A machine learning approach to medical image classification: Detecting age-related macular degeneration in fundus images," Comput. Electr. Eng. (2017).
29. S. Khalid, M. U. Akram, T. Hassan, A. Nasim, and A. Jameel, "Fully automated robust system to detect retinal edema, central serous chorioretinopathy, and age related macular degeneration from optical coherence tomography images," BioMed Res. Int. **2017**, 7148245 (2017).
30. N. Ali, K. B. Bajwa, R. Sablatnig, S. A. Chatzichristofis, Z. Iqbal, M. Rashid, and H. A. Habib, "A novel image retrieval based on visual words integration of SIFT and SURF," PLoS One **11**(6), e0157428 (2016).
31. A. Nazir, R. Ashraf, T. Hamdani, and N. Ali, "Content based image retrieval system by using HSV color histogram, discrete wavelet transform and edge histogram descriptor," in *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (IEEE2018), pp. 1–6.
32. N. Ali, K. B. Bajwa, R. Sablatnig, and Z. Mehmood, "Image retrieval by addition of spatial information based on histograms of triangular regions," Comput. Electr. Eng. **54**, 539–550 (2016).
33. N. Ali, B. Zafar, F. Riaz, S. Hanif Dar, N. Iqbal Ratyal, K. Bashir Bajwa, M. Kashif Iqbal, and M. Sajid, "A Hybrid Geometric Spatial Image Representation for scene classification," PLoS One **13**(9), e0203339 (2018).
34. D. Zinovev, D. Raicu, J. Furst, and S. G. Armato III, "Predicting radiological panel opinions using a panel of machine learning classifiers," Algorithms **2**(4), 1473–1502 (2009).
35. G. Layer, I. Zuna, A. Lorenz, H. Zerban, U. Haberkorn, P. Bannasch, G. van Kaick, and U. Räth, "Computerized ultrasound B-scan texture analysis of experimental fatty liver disease: influence of total lipid content and fat deposit distribution," Ultrason. Imaging **12**(3), 171–188 (1990).
36. X. Yang, S. Tridandapani, J. J. Beitler, D. S. Yu, E. J. Yoshida, W. J. Curran, and T. Liu, "Ultrasound GLCM texture analysis of radiation-induced parotid-gland injury in head-and-neck cancer radiotherapy: an in vivo study of late toxicity," Med. Phys. **39**(9), 5732–5739 (2012).
37. A. R. Prasath and M. Ramya, "Detection of macular drusen based on texture descriptors," Research Journal of Information Technology **7**(1), 70–79 (2015).
38. C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration OCT images," Ophthalmol. Retina **1**(4), 322–327 (2017).

Biomedical Optics EXPRESS

39. P. Burlina, K. D. Pacheco, N. Joshi, D. E. Freund, and N. M. Bressler, "Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis," Comput. Biol. Med. **82**, 80–86 (2017).
40. U. Schmidt-Erfurth, S. M. Waldstein, S. Klimscha, A. Sadeghipour, X. Hu, B. S. Gerendas, A. Osborne, and H. Bogunović, "Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence," Invest. Ophthalmol. Vis. Sci. **59**(8), 3199–3208 (2018).
41. "MATLAB R2018a," The mathworks Inc.: Natick, Massachusetts (2018).
42. R. M. Haralick, "Statistical and structural approaches to texture," Proc. IEEE **67**(5), 786–804 (1979).
43. P. Brynolfsson, D. Nilsson, T. Torheim, T. Asklund, C. T. Karlsson, J. Trygg, T. Nyholm, and A. Garpebring, "Haralick texture features from apparent diffusion coefficient (ADC) MRI images depend on imaging and pre-processing parameters," Sci. Rep. **7**(1), 4041 (2017).
44. W. Yiyang, *Drusen diagnosis comparison datasets* (2019). https://figshare.com/articles/paper_data_zip/7550174
45. B. L. Welch, "The generalisation of student's problems when several different population variances are involved," Biometrika **34**(1-2), 28–35 (1947).
46. J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques* (Elsevier, 2011).
47. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI* (Montreal, Canada, 1995), pp. 1137–1145.
48. P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Mach. Learn. **29**(2/3), 103–130 (1997).
49. Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci. **55**(1), 119–139 (1997).
50. L. Breiman, "Random forests," Mach. Learn. **45**(1), 5–32 (2001).
51. A. Liaw and M. Wiener, "Classification and regression by randomForest," R News **2**, 18–22 (2002).
52. G. Zhang and Y. Lu, "Bias-corrected random forests in regression," J. Appl. Stat. **39**(1), 151–160 (2012).
53. T. V. Phan, L. Seoud, H. Chakor, and F. Cheriet, "Automatic screening and grading of age-related macular degeneration from texture analysis of fundus images," J. Ophthalmol. **2016**, 5893601 (2016).
54. X. Ren, Y. Zheng, Y. Zhao, C. Luo, H. Wang, J. Lian, and Y. He, "Drusen segmentation from retinal images via supervised feature learning," IEEE Access **6**, 2952–2961 (2018).