RetCAD version 1.3

September 2018

Prof.dr. Bram van Ginneken, Radboud University Medical Center

Computer aided detection for Age-related Macular Degeneration and Diabetic Retinopathy

About this white paper

This white paper applies to RetCAD (version 1.3). It describes the general principles of the RetCAD software and presents validations of the software compared to human experts, on various datasets. A scientific publication is forthcoming.

Contents

Introduction

RetCAD was developed by Thirona. The software is based in part on code licensed from Radboud University Medical Center in Nijmegen, the Netherlands. RetCAD is a class IIa CE-certified medical device software product. RetCAD uses artificial intelligence to analyze color fundus images for the presence of Age-related Macular Degeneration (AMD) and Diabetic Retinopathy (DR).

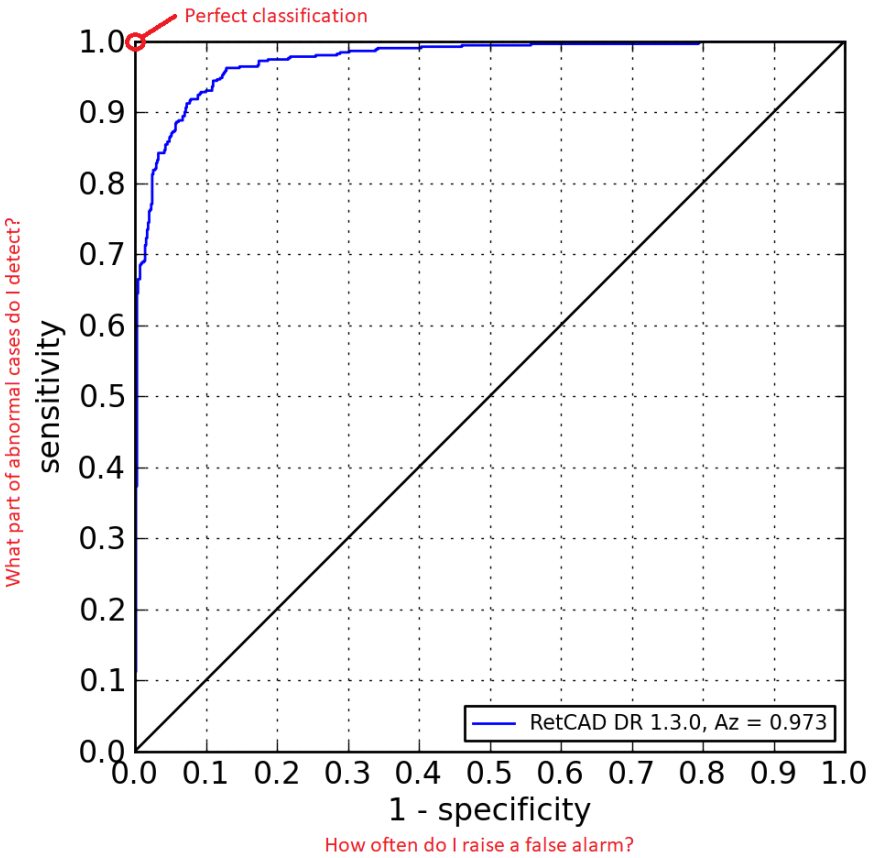
RetCAD takes a color fundus (CF) image as input and produces several outputs. These outputs include a quality assessment of the input image, heatmaps indicating possibly abnormal areas related to AMD and DR, and a score for each of these retinal diseases. The scores are monotonically related to the likelihood of presence of these diseases.

Users can take the output into account in their clinical work: they can decide if a new image should be acquired, in case the quality assessment indicates suboptimal image quality; they can decide to refer a patient for further testing for the presence of AMD, DR or other retinal abnormalities, in case the heatmaps display suspicious regions that are verified by a human operator or when the scores are above certain thresholds.

ROC analysis

Definitions:

- **Sensitivity**: proportion of positive images (i.e. having an abnormality) that have been correctly labelled as positive.
- **Specificity**: proportion of negative images (i.e. not having an abnormality) that have been correctly labelled as negative.
- **ROC curve**: This curve is created by plotting the sensitivity (also called the True Positive Rate) against the False Positive Rate (1 - specificity) at various threshold settings.
- **Az**: area under the ROC curve. This number is equivalent to the probability that a randomly picked positive cases receives a higher score than a randomly picked negative case. It is bound between 0 and 1: at 0.5 the system is equivalent to guessing, at 1 the system shows perfect classification.
- **T**: Threshold value. Different threshold values correspond to different points on the ROC curve. The point on the ROC curve closest to perfect classification (the upper left corner) is often considered as the optimal threshold, but it is not necessarily the optimal threshold for the most cost-effective screening. **T** is the value set by the user to determine which images are labelled abnormal/normal.



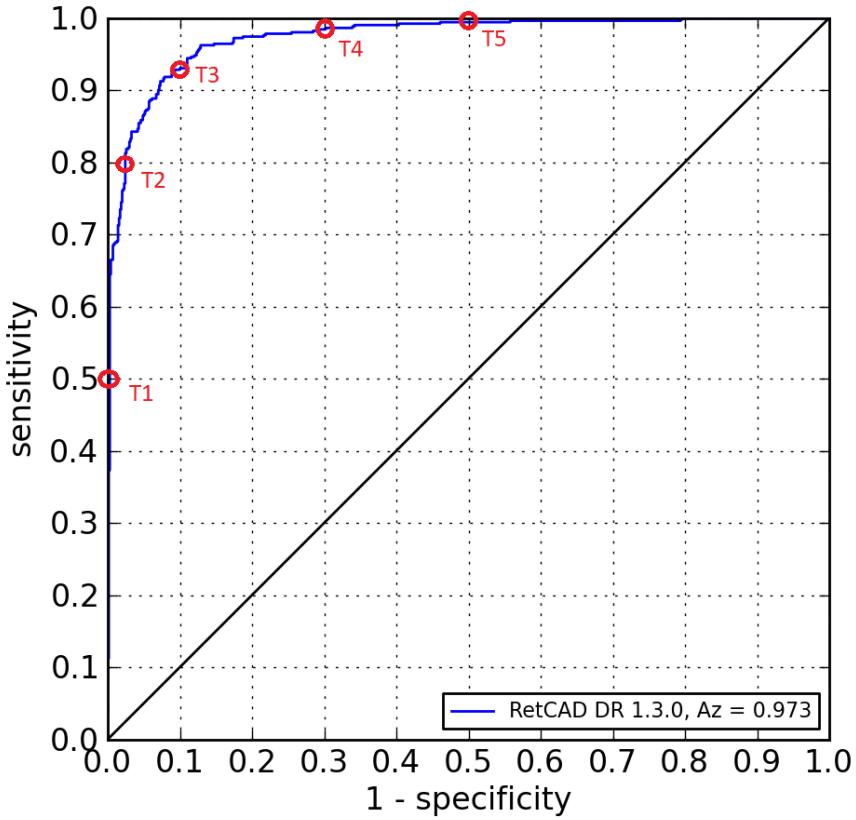
An indication for the accuracy of a diagnostic test is the traditional academic point system:

- 0.90-1 = excellent (A)
- 0.80-0.90 = good (B)
- 0.70-0.80 = fair (C)
- 0.60-0.70 = poor (D)
- 0.50-0.60 = fail (F)

In the curve shown on the left, the Az value is 0.973 which according to the above classification would be considered excellent (A).

Table 1: Different threshold values with corresponding sensitivity and specificity levels.		
Threshold	True Positive Rate (sensitivity)	True Negative Rate (specificity)
T1	50 %	100 %
T2	80 %	97 %
T3	93 %	90 %
T4	98 %	70 %
T5	100 %	50 %

This can be presented graphically into an ROC curve:



A relatively low threshold value of the software corresponds with a higher sensitivity, but at the cost of a lower specificity. A relatively high threshold value of the software corresponds with a higher specificity, but at the cost of a lower sensitivity. Hence, the threshold value is trade-off between sensitivity and specificity. Shaded regions around the ROC curve (shown later in this report) indicate the 95% confidence intervals as computed using a statistical procedure called bootstrapping.

To summarize: An ROC curve demonstrates several things/characteristics of the test:

- ✓ It shows the trade-off between sensitivity and specificity (any increase in sensitivity will often be accompanied by a decrease in specificity).
- ✓ The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- ✓ The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- ✓ The area under the curve is a measure of the test's performance.

RetCAD: How does it work?

RetCAD is software based on convolutional neural networks, a state-of-the-art technique in machine learning. In the process of analyzing the input CF image, it compares regions in the image with regions extracted from normal and abnormal CF images. These images form the training data set of the software. The software is always tested on independent data, the test or validation set.

CF cameras from different manufacturers produce images of different quality because of hardware differences. In addition, image acquisition protocols can vary across acquisition sites, for example: the illumination, angular resolution (field of view) and the resolution of the image can vary. Furthermore, the patients may originate from different populations in which the appearance of the retina, such as color and pigmentation, may vary. In some patients the fluid in their eye balls is not clear and this can make it difficult to make a good quality image. If a patient blinks during the acquisition an image, the image may be substandard.

Specific algorithms to improve and normalize the input CF image prior to analysis are included in the RetCAD software. However, these algorithms are not perfect and cannot produce a high quality image if the quality of the input image is too low. Therefore, a quality measure for each image is also computed and the user of the software could decide to obtain a new image in case the quality is considered too low by RetCAD.

RetCAD: Performance evaluation

The RetCAD software has been evaluated on several datasets. The images in these datasets were acquired using different types of CF cameras at different resolutions. The performance of the RetCAD software is directly compared with that of human experts. The following sections describe evaluations on various datasets.

Messidor

The Messidor database is a publicly available set of 1200 CF images which were acquired by three ophthalmologic departments using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinography with a 45 degree field of view. The images were captured using 8 bits per color plane at 1440x960, 2240x1488, or 2304x1536 pixels. 800 images were acquired with pupil dilation (one drop of Tropicamide at 0.5%) and 400 without dilation. More information about the database can be found following the website link[1].

For each image in the database a reference DR severity grade, set by medical experts, was provided. Four severity grades were used: No DR, mild DR, moderate DR and severe DR.

The RetCAD software was applied to each of the 1200 images in the dataset and the RetCAD software was evaluated by comparing the RetCAD DR score with the DR severity grade as set by the medical experts.

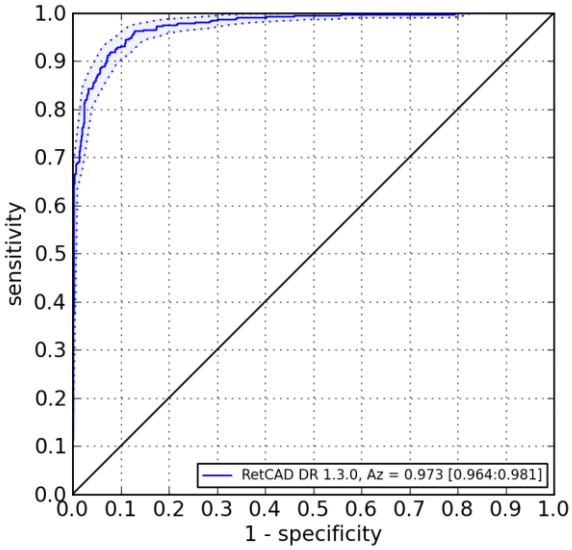
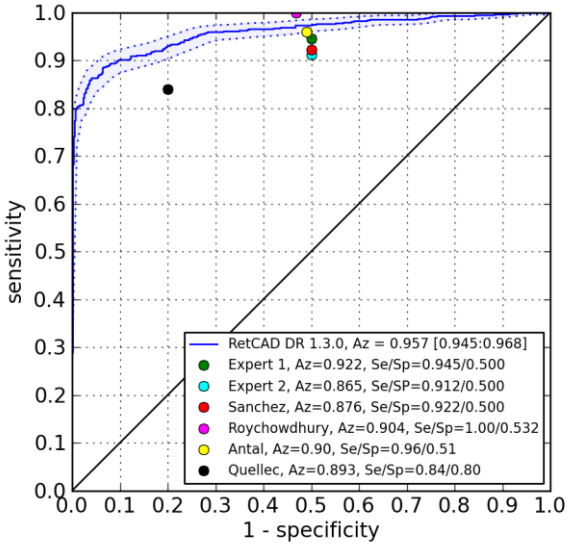
Two types of evaluations were performed:

1. The image was deemed positive if the reference grade was **referable DR**, i.e. severity level of moderate or severe DR.
2. The image was deemed positive if the reference grade was **any DR**, i.e. mild, moderate or severe DR.

The results of the evaluation are summarized in an ROC graph. In this ROC graph, the operating point of human experts is added for the second evaluation.

Note that the RetCAD software was not trained with any of the images that are part of the Messidor data set.

[1] *Kindly provided by the Messidor program partners (see http://www.adcis.net/en/DownloadThirdParty/Messidor.html).*

Reference: Referable DR	Reference: Any DR
Number of negative images: 699 Number of positive images: 501	Number of negative images: 546 Number of positive images: 654
	
Conclusion: The RetCAD software achieves an Az value of 0.973 for the identification of images with referable DR.	Conclusion: The RetCAD software achieves an Az value of 0.957 for the identification of images with any DR.

Operating points for RetCAD DR

In Table 2, sensitivity and specificity values of RetCAD for DR detection are given at several threshold values for this specific dataset.

Table 2: Operating points of RetCAD for any DR detection

Threshold	True Positive Rate (sensitivity)	True Negative Rate (specificity)
58	51 %	100 %
47	80 %	98 %
37	93 %	91 %
21	98 %	70 %
17	100 %	51 %

Comparison with other systems and human experts

Several scientific publications have presented DR detection systems that were evaluated in the Messidor data set. One publication also reported the sensitivity/specificity for two human experts. . All studies use the criteria of “any DR” for positive cases, i.e. mild or more severe are considered as the positive class. The table below reports the performances of the computer systems, including RetCAD DR, and the two human experts.

Table 3: Performance of software packages and human experts on the Messidor dataset.

Author	Az value	Se/Sp	Year	Link
RetCAD DR	0.957	0.93/0.91	2018	-
Antal et al.	0.900	0.96/0.51	2012	http://arxiv.org/abs/1410.8577
Quellec et al.	0.893	0.84/0.80	2016	http://www.ncbi.nlm.nih.gov/pubmed/26774796
Roychowdhury et al.	0.904	1.00/0.53	2014	http://www.ncbi.nlm.nih.gov/pubmed/25192577
Sánchez et al.	0.876	0.92/0.50	2011	http://www.ncbi.nlm.nih.gov/pubmed/21527381

Expert 1	0.922	0.95/0.50	2011	http://www.ncbi.nlm.nih.gov/pubmed/21527381
Expert 2	0.865	0.91/0.50	2011	http://www.ncbi.nlm.nih.gov/pubmed/21527381

Messidor-2

The Messidor-2 dataset is a collection of Diabetic Retinopathy (DR) examinations, each consisting of two macula-centered eye fundus images (one per eye). Part of the dataset (*Messidor-Original*) was kindly provided by the Messidor program partners (see http://messidor.crihan.fr). The remainder (*Messidor-Extension*) consists of examinations obtained from the Brest University Hospital.

In the original Messidor dataset, some fundus images came in pairs (one image of both the left and right eye), some others were single (one image per patient). *Messidor-Original* consists of all image pairs from the original Messidor dataset, that is 529 examinations (1058 images).

In order to populate *Messidor-Extension*, diabetic patients were recruited in the Ophthalmology department of Brest University Hospital (France) between October 16, 2009 and September 6, 2010. Eye fundi were imaged, without pharmacological dilation, using a Topcon TRC NW6 non-mydriatic fundus camera with a 45 degree field of view. Only macula-centered images were included in the dataset. *Messidor-Extension* contains 345 examinations (690 images).

Overall, Messidor-2 contains 874 examinations (1748 images). All patients in the database were graded for the presence of referable DR, i.e. moderate or more DR, by three medical experts and a consensus reference was made based on these gradings. In total, 190 patients had referable DR, and 684 patients did not have referable DR.

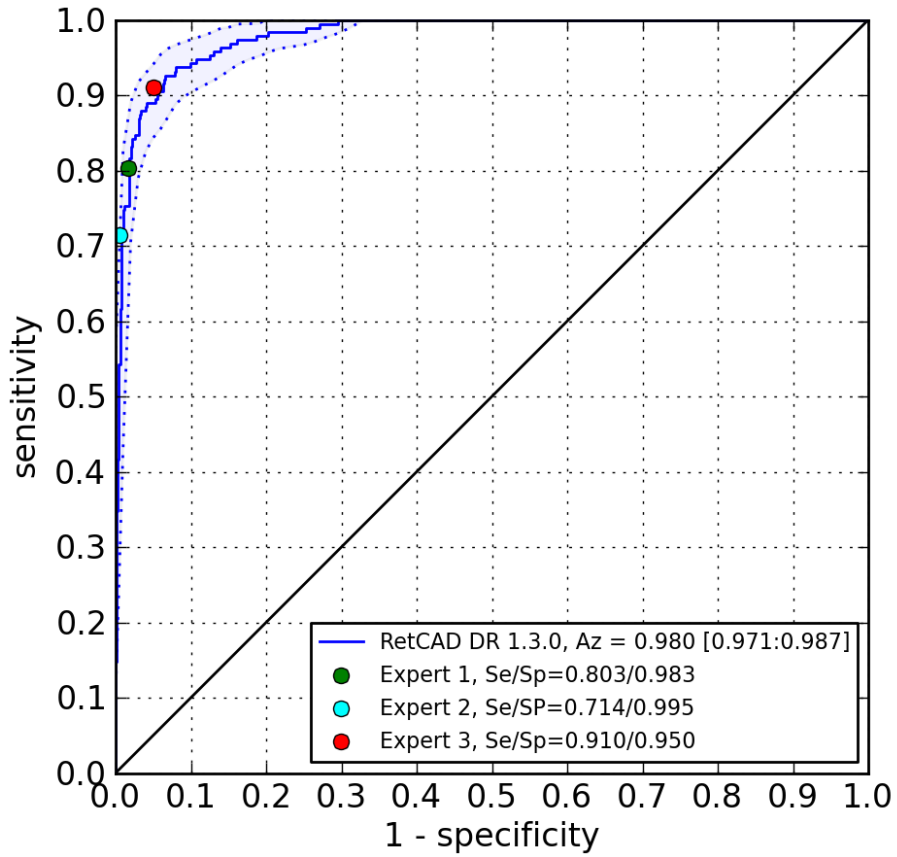
More information about the database can be found following the website link[2].

The RetCAD software was applied to each of the 1748 images in the dataset and the DR component of the RetCAD software was evaluated. In the evaluation, the highest score of the two images of a patient was set to be the patient-based score for DR. This score is compared with the provided reference score as constructed by a consensus of three medical experts (https://www.ncbi.nlm.nih.gov/pubmed/27701631).

The results of the evaluation are summarized in an ROC graph. Sensitivity and specificity of the three medical experts who scored the 874 examinations were measured by comparing the score to the consensus scoring of the other two human experts. The operating points of the human experts are added in the plot, but it thus has to be noted these were measured against a slightly difference reference standard.

Note that the RetCAD software was not trained with any of the images that are part of the Messidor2 data set.

[2] Kindly provided by the LaTIM laboratory (see http://latim.univ-brest.fr/) and the Messidor program partners (see http://messidor.crihan.fr/)

Reference: referable DR
Number of negative patients: 684 Number of positive patients: 190

Conclusion: the RetCAD software achieves an Az value of 0.980 for the identification of patients with referable DR.

Operating points for RetCAD DR

In Table 4, sensitivity and specificity values of RetCAD for DR detection are given at several threshold values for this specific dataset.

Table 4: Operating points of RetCAD for referable DR detection

Threshold	True Positive Rate (sensitivity)	True Negative Rate (specificity)
64	51 %	100 %
57	80 %	98 %
52	93 %	93 %
33	100 %	70 %
23	100 %	50 %

Comparison with other systems and human experts

Performance of other state-of-the-art DR detection systems on the Messidor2 database have been reported. Additionally, the performance of human graders were reported in one of these publications (Abramoff et al, 2013) and were added.

Table 5: Performance of other software packages on the Messidor-2 dataset

Author	Az value	Se/Sp	Year	Link
RetCAD DR	0.980	0.93/0.93	2018	-

Abramoff et al.	0.937	0.97/0.59	2013	https://www.ncbi.nlm.nih.gov/pubmed/23494039
Abramoff et al.	0.980	0.97/0.87	2016	https://www.ncbi.nlm.nih.gov/pubmed/27701631
Expert 1	-	0.80/0.98	2013	https://www.ncbi.nlm.nih.gov/pubmed/23494039
Expert 2	-	0.71/1.00	2013	https://www.ncbi.nlm.nih.gov/pubmed/23494039
Expert 3	-	0.91/0.95	2013	https://www.ncbi.nlm.nih.gov/pubmed/23494039

AMD dataset

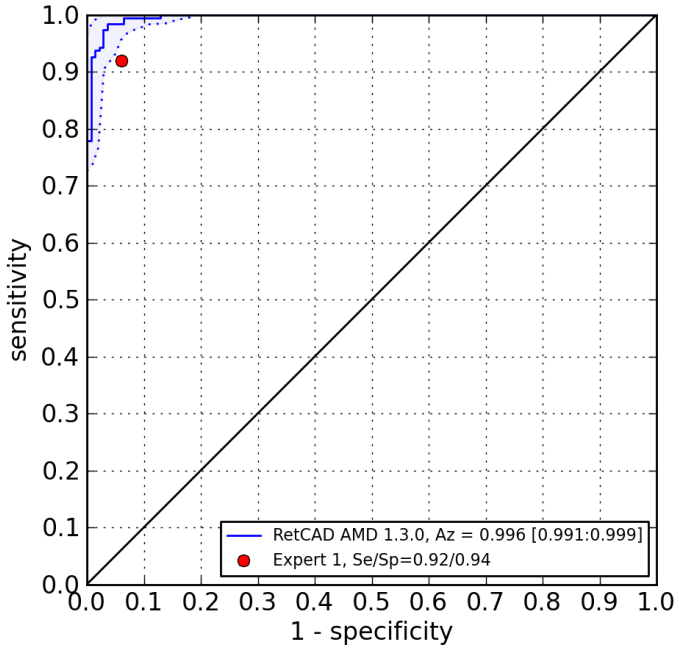
The AMD dataset is a dataset consisting of 329 macula centered images that were acquired at an ophthalmologic department in a hospital using either a Topcon TRC 501X model digital fundus camera at 50 degree field of view or a Canon CR-DGi model non-mydriatic retinal camera at 45 degree field of view. Pupil dilation was achieved with 1.0% tropicamide and 2.5% phenylephrine. All images were macula centered and image resolution varied between 1360x1024 to 3504x2336 pixels.

The images in the database were graded for presence of referable AMD by an expert with over 5 years of experience in grading fundus photographs. Referable AMD is defined as having at least 15 small drusen (>63µm) or more than one intermediate sized drusen (>126µm) or any sign of advanced AMD.

The RetCAD software was applied to each of the 329 images in the dataset and the AMD component of the RetCAD software was evaluated by comparing the RetCAD AMD score with the reference as set by the expert.

The results of the evaluation are summarized in an ROC graph. In this ROC graph, the operating point of a second human expert (over 5 years of experience in grading fundus images) is added for comparison.

Note that the RetCAD software was not trained with any of the images that are part of this data set.

Reference: referable AMD
Number of negative images: 139 Number of positive images: 190

Conclusion: the RetCAD software achieves an Az value of 0.996 for the identification of images with referable AMD.

Operating points for RetCAD AMD

In Table 6, sensitivity and specificity values of RetCAD for AMD detection are given at several threshold values for this specific dataset.

Table 6: Operating points of RetCAD for AMD detection

Threshold	True Positive Rate (sensitivity)	True Negative Rate (specificity)
67	50 %	100 %
43	80 %	99 %
23	98 %	96 %
20	100 %	84 %
17	100 %	47 %

Comparison with other systems and human experts

Table 7: Performance of other software packages on the AMD dataset

Author	Az value	Se/Sp	Year	Link
RetCAD AMD	0.996	0.97/0.97	2018	-
Expert 1	-	0.92/0.94	2018	-

Mixed AMD-DR dataset

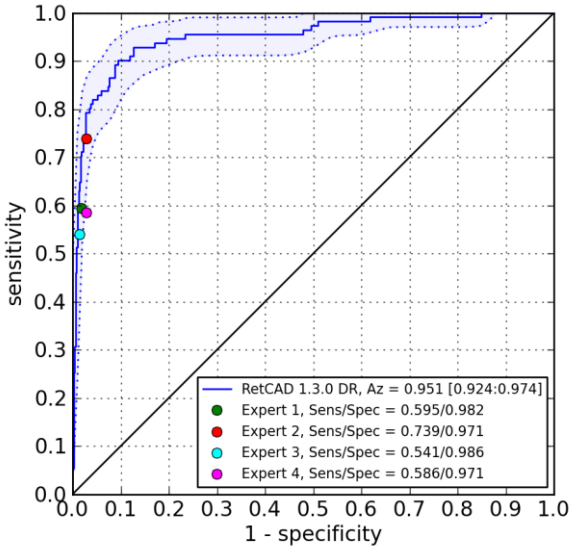
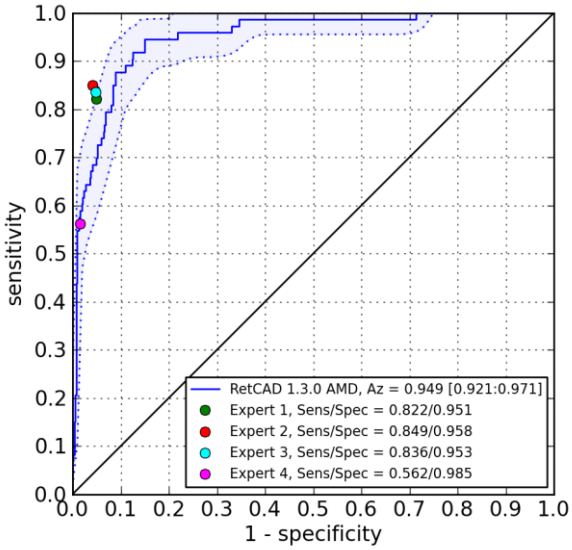
The Mixed AMD-DR dataset is a dataset consisting of 600 images that were acquired at an ophthalmologic department in a hospital using a Canon CR-2PlusAF digital fundus camera at 45 degree field of view. No pupil dilating eye-drops were administered. Image resolution varied between 2376x1584 to 3456x5184 pixels. The patients that were imaged had either signs of AMD, or DR, or both, or they were not affected by either disease.

The reference for the images in this Mixed AMD-DR dataset was set by an experienced ophthalmologist. Grading criteria were based on the ICDR and AREDS classifications. In total 78 images were graded as having referable AMD (defined as intermediate or worse AMD), 111 were graded as referable DR (defined as moderate or worse DR), 3 were graded as both referable AMD and DR, and 408 were graded as neither referable AMD nor referable DR.

The RetCAD software was applied to each of the 600 images in the dataset and both the AMD and DR component of the RetCAD software were evaluated by comparing the RetCAD scores with the reference as set by the medical expert.

The results of the evaluation are summarized in two ROC graphs, one for AMD and one for DR.

Note that the RetCAD software was not trained with any of the images that are part of this data set.

Reference: Referable DR	Reference: Referable AMD
Number of DR negative images: 489 Number of DR positive images: 111	Number of AMD negative images: 522 Number of DR positive images: 78
	
Conclusion: The RetCAD software achieves an Az value of 0.951 for the identification of images with referable DR.	Conclusion: The RetCAD software achieves an Az value of 0.949 for the identification of images with referable AMD.

Operating points for RetCAD AMD

In Table 8, sensitivity and specificity values of RetCAD for AMD and DR detection are given at several threshold values for this specific dataset.

Table 8: Operating points of RetCAD for AMD and DR detection

Threshold	True Positive Rate DR (sensitivity)	True Negative Rate DR (specificity)	True Positive Rate AMD (sensitivity)	True Negative Rate AMD (specificity)
70	25 %	100 %	62 %	98 %
60	45 %	99 %	79 %	92 %
50	70 %	98 %	92 %	87 %
40	81 %	96 %	95 %	82 %
20	93 %	87 %	99 %	59 %

Comparison with other systems and human experts

Operating points of the 4 human graders are included in the above plots. The performance obtained by the human graders is similar to that of the RetCAD software. The operating points for all observers fall within the confidence interval around the ROC curves.