Pattern Recognition and Image Processing Group
Institute of Visual Computing and Human-Centered Technology
TU Wien
Favoritenstrasse 9-11/193-03
A-1040 Vienna AUSTRIA
Phone: +43 (1) 58801 - 18661
Fax: +43 (1) 58801 - 18697
E-mail:
URL: http://www.prip.tuwien.ac.at/

PRIP-TR-145 Vienna February 12, 2019

Quality Assessment of Color Fundus and Fluorescein Angiography Images[1]

Michael König and Omar Ismail

Supervised by Walter G. Kropatsch

Abstract

Color Fundus and Fluorescein Angiography are medical procedures used for eye diagnosis. In Fluorescein Angiography a series of photographs is taken to analyze the blood flow in the back of the eyes. Color Fundus records color images of the interior surface of the eye to monitor and document disorders. As with every photograph, certain image distortions occur, making analysis difficult. The aim of these theses is to find a way to measure these distortions and ideally, flag bad photographs so that the analysis process can be accelerated.

[1]Supported by the Medical University of Vienna and the Vienna General Hospital

Contents

1 Introduction

Two common eye diagnostic methods are Color Fundus (CF) and Fluorescein Angiography (FA), which are used to monitor and document eye disorders. Since the first successful fundus camera in 1926 which was commercially available, CF photography is an important part of recognizing different diseases of the eye [8].
In FA Fluorescent dye is injected into the bloodstream, which highlights blood vessels in the back of the eyes. The blood flow is then analyzed through a series of photographs, which can show blood leakages, circulation problems and other abnormalities. Because of the special filter they are shot through, these images are gray-scale.

After analyzing the comments of 2,621 labeled tickets with a total of 86,997 images, it was possible to extract four quality factors: contrast, illumination, acutance (sharpness) and noise. Often, at least one of these factors is negatively impacted through human error or equipment failure. This leads to hardly readable or even completely useless images, which significantly slow the analysis process down. If there was a way to assess the quality of CF and FA images while evaluating each factor separately, the reading process could be accelerated by avoiding time consuming bad quality images.

Image quality assessment (IQA) can be split into three types: Full-Reference-IQA (FR-IQA), Reduced-Reference-IQA (RR-IQA) and No-Reference-IQA (NR-IQA) [13]. Because every eye is unique, it is not possible to define any type of reference to compare against; A "blind" (no-reference) image quality assessment is needed.

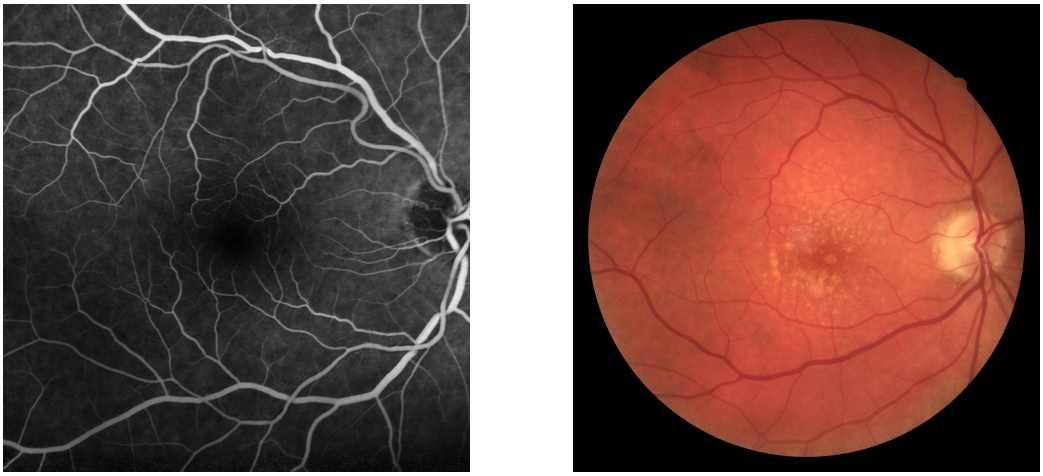
In this report, a NR-IQA framework for CF and FA images is proposed. The objective of this framework is to formalize the four factors and evaluate their impact on the image in real-time while maintaining a high generalizability. The chosen algorithms have to be efficient to ensure real-time application on weaker hardware. That way, images of insufficient quality can be detected while the patient is still present to make new, better images.

The framework extracts quality vectors from CF and FA images, which can either have a circular mask or no mask at all. Any other type of mask or image information like time stamps on top of the retinal scan will negatively impact the accuracy of the calculations.

The vectors from each of the quality features are passed to a trained classification module, which then returns a classification and its corresponding scoring. The score indicates the certainty of the classification. If the certainty (and thus the score) is low, readers can determine themselves if an image is of sufficient quality and thus mitigate the risk of misclassification.

This paper is structured as follows: Section 2 describes the various methods used to extract the quality features for each modality (written in parentheses), Section 3 focuses the training and the evaluation of the used classification modules, Section 4 lists the runtime for the two modalities, and Section 5 concludes the work done.

2 Feature Extraction



(a) (b)

Figure 1: Examples of a good FA-image (a) and a good CF-image (b).

In FA, three features (Acutance, Contrast/Illumination, and Noise) are ex-

tracted and a quality vector for each feature is calculated, while CF calculates four features(Acutance, Contrast, Illumination, and Noise). The resulting values are then evaluated using trained classification modules, which carry out the final evaluation of the feature. An overview of the framework is given in Figure 2.

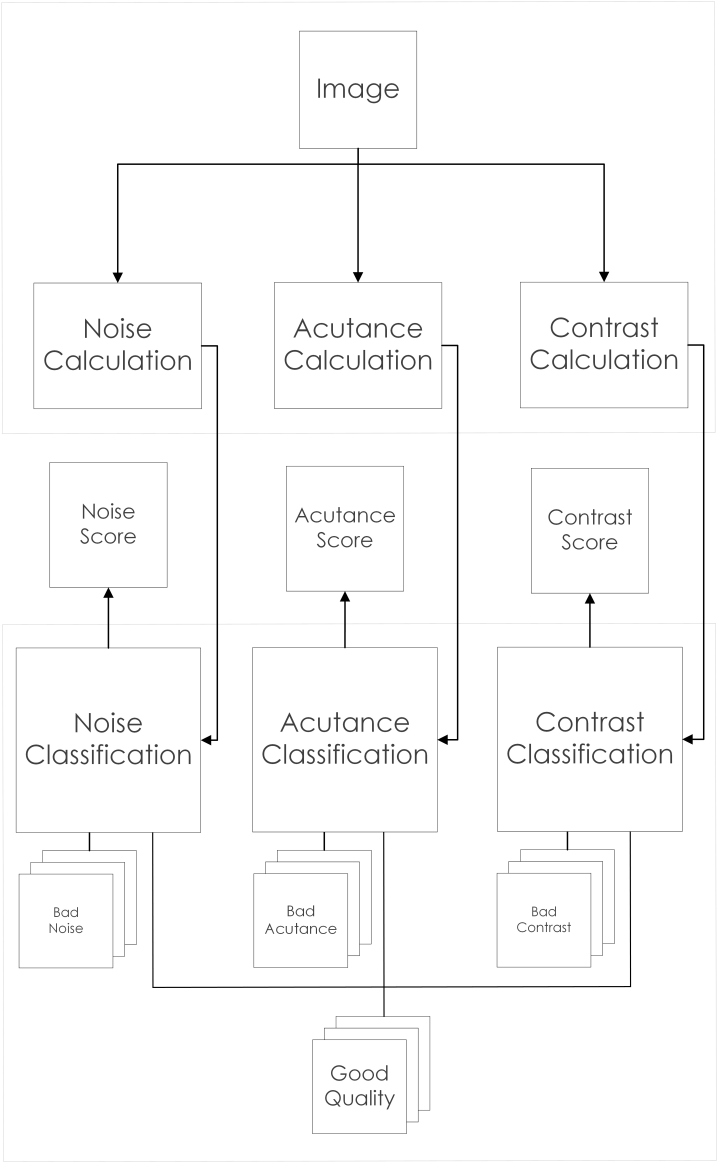


Figure 2: Overview of the proposed framework.

Because it is possible for images to be readable despite suboptimal acutance, color, contrast, illumination or noise, each feature is evaluated separately. This important for certain pathologies, where a good contrast value is not possible due to impeded blood flow.

Sometimes, CF and FA images do not fill the entire image (as seen in Figure 1b). While the background pixels can give information about the noise ratio of an image, they falsify the results of acutance and contrast calculations by adding a sharp edge and contrast to the image. To avoid this, such images need to be masked before evaluation. Another benefit of masking is the removal of timestamps in FA images, which also influence contrast and acutance calculations, as seen in Section 2.1, Figure 4.

2.1 Acutance (CF & FA)

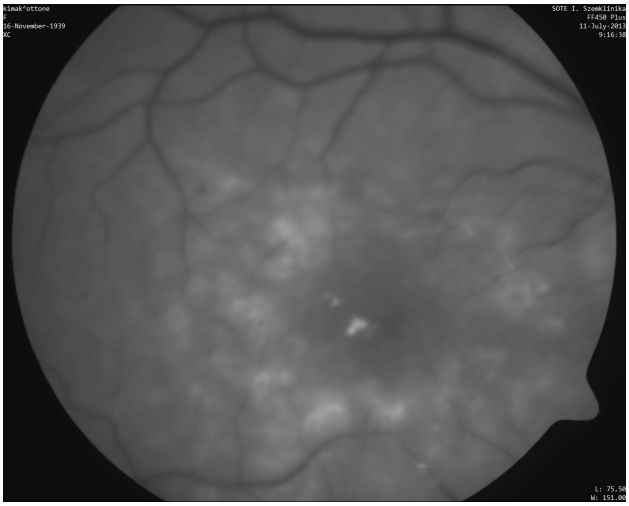


Figure 3: Example for an FA-Image with low acutance.

The acutance (or sharpness) of an image is determined by the edge contrast of an image.
In CF and FA, low acutance makes it difficult or impossible to discern thinner blood vessels, which can be critical.

There are numerous ways to measure the acutance of an image, like fre-

quency analysis via Fourier Transformation or examination of the maximum value after applying a Laplace of Gaussian filter.

The framework implements a slightly modified version of the method proposed by Dias et al. [3] for our framework. It is a combination of two explicit focus measures and is already used in NR-IQA of color fundus images, which is another eye diagnostic method.
Because it promises reliable and quick calculations, it is used in this framework, too.

To get the first gradient map O, a normalized Sobel operator is applied to the image, which consists of two convolution kernels:

$$M_v = \frac{1}{4} \cdot \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \quad \text{and} \quad M_h = \frac{1}{4} \cdot \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{1}$$

$$O = \sqrt{(M_v * I)^2 + (M_h * I)^2} \tag{2}$$

With it, FM_1 is calculated as the average value of the gradient map:

$$FM_1 = \frac{1}{n} \cdot \sum_{i=1}^{n} O_i \tag{3}$$

where O_i is the ith pixel of the gradient map. Then, a 3x3 median filter is applied to the image and again a gradient map is calculated, $L1$. FM_2 is computed as the average value of the new gradient map:

$$FM_2 = \frac{1}{n} \cdot \sum_{i=1}^{n} L1_i \tag{4}$$

Additionally, D_1 is calculated as the difference between the two averages FM_1 and FM_2:

$$D_1 = FM_1 - FM_2 \tag{5}$$

Finally, another median filter is applied to the original image, this time with a kernel size of 5x5, and the same procedure is applied for $L2$:

$$FM_3 = \frac{1}{n} \cdot \sum_{i=1}^{n} L2_i \tag{6}$$

This time, D_2 is the difference between FM_2 and FM_3:

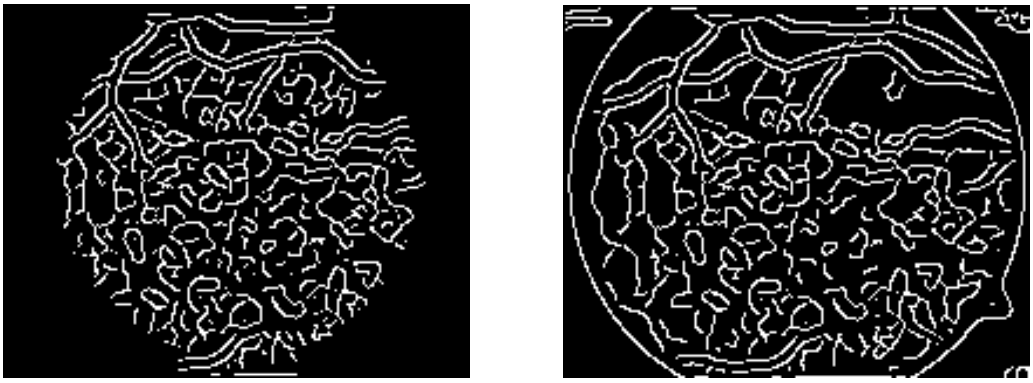
$$D_2 = FM_2 - FM_3 \tag{7}$$

While this may look similar to a Laplacian Pyramid [2], it differs in two ways:

1. The image is never resized, only a low pass filter is applied.
2. Instead of subtracting two images from each other (resulting in an edge image), two (scalar) values are subtracted, resulting in another scalar.

The idea behind these values is that sharp/focused images are more affected by low pass filtering, resulting in greater differences between the three gradient maps.

Masking is critical when measuring focus, as the high frequencies at the border of the image influence the focus rating, especially if there are timestamps or any other type of information in the corners. An example for this is shown in Figure 4 and Table 1.



(a) (b)

Figure 4: Edge image of Figure 3 with (a) and without masking (b). Resolution: 200×160 px

Table 1: Calculation results of images shown in Figures 4a and 4b.

	(5a)	(5b)
FM_1	0.0778	0.0961
FM_2	0.0518	0.0677
D_1	0.0260	0.0284
FM_3	0.0267	0.0435
D_2	0.0250	0.0242

The resulting quality vector contains FM_1, D_1 and D_2. Calculation of this feature is rather fast and returns acceptable results (see Sections 3 and 4).

2.2 Color (CF)

Many retinal scans of the dataset provided by the VRC show images with maximum intensity values of 0.08 or less, on a scale from 0 to 1. These images appear to be absolutely black, which makes any diagnosis for medical doctors impossible. In contrast, some images are too bright - as the left image of Figure 5 with an maximum intensity of 0.97 and a mean intensity of 0.66 - to recognize any features of the eye.

There are only few approaches including the color scheme of the retinal scan in the quality assessment. [10] measure the darkness and brightness together with the evenness of the illumination. Therefore, the luminance channel of the input image is processed with simple thresholding and the result is analyzed. With this method however, no use is made of the color information so that images with diverse colors, but similar lighting cannot be differentiated.

The approach of this work to assess the color scheme is mostly similar to the method proposed by [3]. The goal is to classify the input image into one of the three categories bright, dark or normal, as shown in Figure 5. Therefore, in preparation of the assessment, a predefined color map is created, using 10 representative images for each category. For each of these color maps, the algorithm then creates a so-called indexed image. This is achieved by finding the correspondent color value for every pixel of the original input image in the color map and taking this as the pixel value in the new image. If a value does not exactly appear in the used color map, the most similar

color value is chosen. Afterwards, the percentage of different occupied bins for each of these three color maps is calculated.

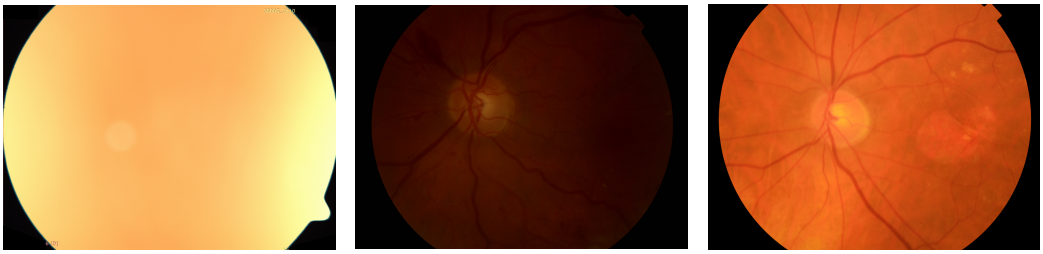


Figure 5: Examples for bright (left), dark (mid) and normal (right) color scheme for CF images.

As an additional parameter, the approach proposed in this work creates an indexed image for a fourth color map, which is displayed in Figure 6. As explained in [3], this color map naturally arranges colors, which are occurring more often in good quality images, at the right end of the map. By creating an indexed image using this color map and calculating the image mean, this method utilizes the fact that better colors are scoring significantly higher results than unusual colors.

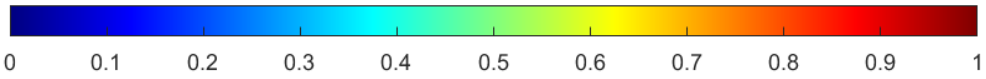
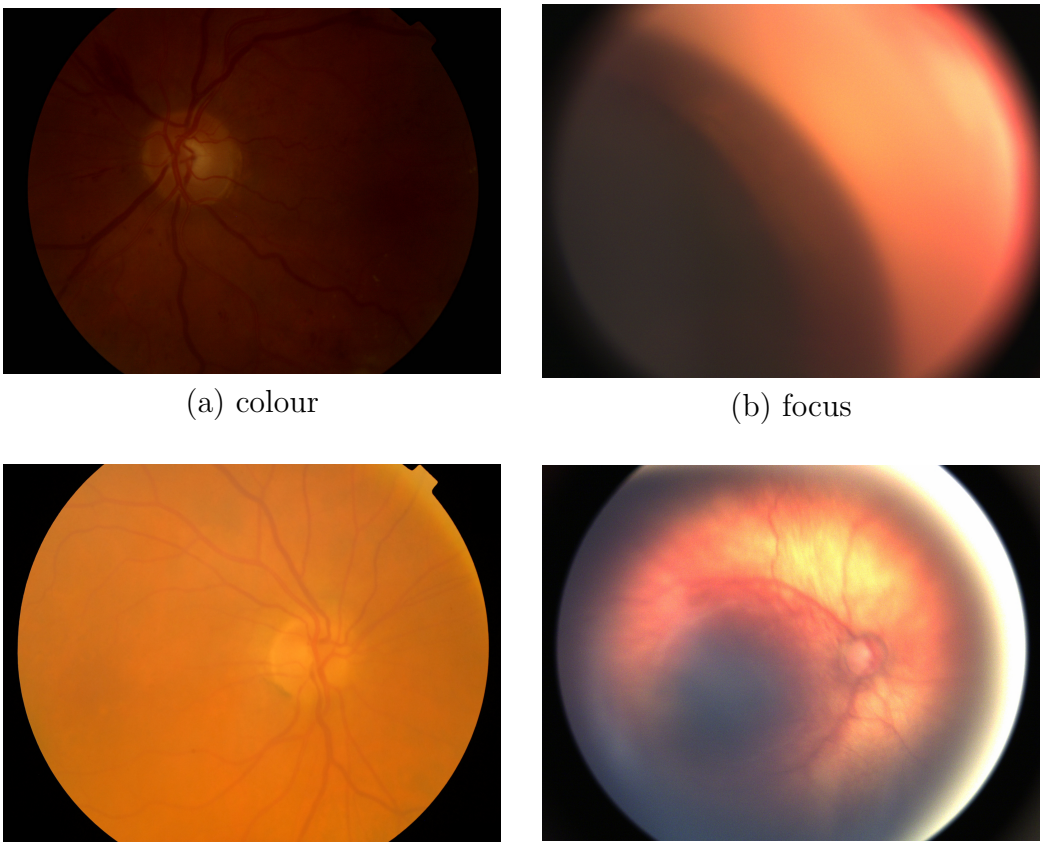


Figure 6: Color map that clusters natural colours of retinal scans.

2.3 Contrast (CF)

Contrast has as much impact on the quality of a retinal scan as the other categories. If the contrast is very low, the vessels cannot be differentiated from the background as in Figure 7(c) and so one of the most important features is lost.



(a) colour (b) focus

(c) contrast (d) illumination

Figure 7: Bad quality examples for all four categories.

Dias et Al. [3] again use a method including histogram back projection for the assessment of the contrast. Therefore, an indexed image is created for a color map, which was obtained through the analysis of 170 high contrast images. Afterwards, two different measurements are performed: As shown in the work of [3], the first measurement $CtM1$ (8) calculates the sum of the absolute difference between the pixel percentage within each bin (p_i^0). As a second measurement, the number of empty bins is counted. The same calculations are then made for a version of the image which was low-pass filtered. This approach however is not very flexible, if an image consists of not many, but only a few colors. If, for example, an image with black background and white vessels is evaluated, the two measurements would (assuming a distribution of 80% black and 20% white pixels) lead to the results of (9) and 14 empty

bins.

$$CtM1_{(p_i^0)} = \sum_{i=1}^{16}(|p_i^0 - \frac{1}{16}|) \tag{8}$$

$$CtM1_{(0.8,0.2,0,...)} = (0.8 - \frac{1}{16}) + (0.2 - \frac{1}{16}) + \frac{1}{16} * 14 = 1.75 \tag{9}$$

These results are the same for every image where only two bins are occupied, whether the colours are easy to differentiate or almost impossible to distinguish. Therefore, this method can lead to imprecise contrast assessment.

The approach followed in this work was proposed by Matkovic et al. [6], which is shown in Section 2.5.

2.4 Illumination (CF)

As the last category for image quality measurement of CF images, [3] propose the evaluation of a homogeneous illumination. Many images of the provided data show ungradable areas (mainly on the edge of the ROI as in Figure 8) due to uneven luminance.

In the last years there have been several different approaches to solve this problem, as mentioned by [3]. Some of them are based on the intensity level of gray-scale images, like [4] propose. For this task however, this approach would not make use of all the information a Color Fundus photograph has to offer and therefore was dismissed. [14] and [12] apply a different model based on smoothing the image.

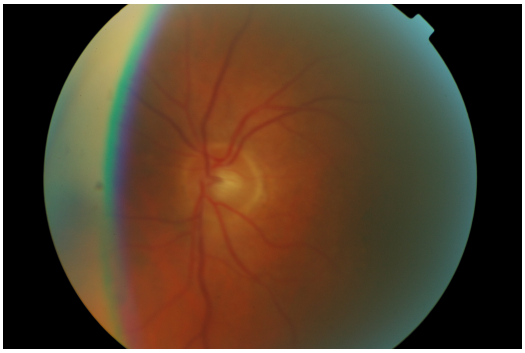


Figure 8: Example for inhomogeneous illumination.

Dias et al. [3] perform the illumination assessment by again using the modified color map in Figure 6. For the assessment, an indexed image I using this color map is created for the original input image. Afterwards, the mean M for all n pixels of this indexed image (10) is calculated. As stated by [3], this achieves a much higher value for colors which are usually occurring in better quality retinal scans.

$$M = \frac{1}{n} \sum_{i=1}^{n} I_i \tag{10}$$

In addition to calculating the mean, three other values are calculated for the indexed image: First, the variance of the occurring colors as in (11) is measured. Furthermore, for the assessment of a homogeneous illumination, the variance above and below the mean are calculated, as shown in (12) and (13).

$$V = var\{I\} \tag{11}$$

$$V_A = var\{I_i | I_i > M\} \tag{12}$$

$$V_B = var\{I_i | I_i < M\} \tag{13}$$

In this approach, the three calculated variances (V, V_A, V_B) for bad quality images are expected to show high values. The best and therefore lowest results are given by a perfectly homogeneous illuminated image, thus a single colored image. However, to prevent a completely black retinal scan from getting significantly higher scores than good quality images, the mean of the indexed image from (10) is included as a parameter in the illumination evaluation.

2.5 Contrast and Illumination (FA)

The most important factor in FA-images is the contrast. Without contrast, blood vessels cannot be distinguished from the background, making analysis impossible.

While contrast and illumination are separate factors in color images, they are interrelated in gray-scale images; It is not possible to have an FA-image

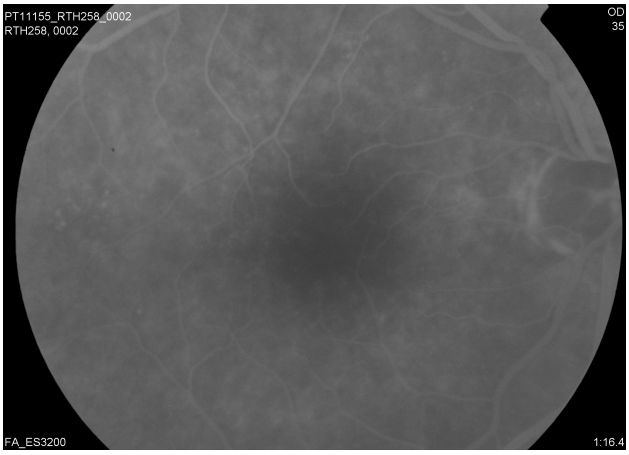


Figure 9: Example for an FA-Image with low contrast.

with good contrast but bad illumination and vice-versa, which is why this framework combines both factors into one feature.

The common definition of contrast in images is the difference in luminance and/or color. Contrast is what makes it possible for the human visual system to differentiate between objects and background or foreground.

To calculate the overall contrast of an image, the most common methods are the Michelson contrast [7], the Weber contrast and the Root-Mean-Square (RMS) contrast [9]. However, because FA-Images neither have equivalent dark and light features nor a uniform background, neither the Michelson nor the Weber methods can be used. While RMS offered usable results, it is not accurate enough, which is why another method is needed to calculate image contrast.

The proposed framework uses the Global Contrast Factor (GCF) by Matkovic et al. [6], as it promises a metric that not only measures the overall contrast of an image, but also the richness of detail as perceived by a human observer. For medical images that are analyzed by human readers, GCF appears to be ideal.

Because color information is not needed for this algorithm, and to accurately calculate the contrast quality, the image is converted from RGB (with

a pixel range of [0-255]) to double precision (with a pixel range of [0-1]). Using the original RGB intensity pixel value k (which ranges between 0 and 255) and a gamma[1] of $\gamma = 2.2$, the linear luminance

$$l = \left(\frac{k}{255}\right)^{\gamma} \tag{14}$$

and the perceptual luminance

$$L = 100 \cdot \sqrt{l} \tag{15}$$

are calculated. Once the perceptual luminance of every pixel is calculated, the local contrast lc_i of each pixel i is computed by calculating the average difference of perceptual luminance (L) between the pixel itself (L_i) and its left, right, top and bottom neighbors:

$$lc_i = \frac{|L_i - L_{left}| + |L_i - L_{right}| + |L_i - L_{top}| + |L_i - L_{bottom}|}{4} \tag{16}$$

Only existing pixels are taken into account, which means that masked pixels and pixels outside the image are ignored. For example, in calculating the local contrast of the upper left corner pixel, both the left and top pixels are ignored, leading to a calculation of $lc_i = \frac{|L_i - L_{right}| + |L_i - L_{bottom}|}{2}$.

With the local contrasts calculated, the average local contrast C_i for the resolution level i is computed via

$$C_i = \frac{1}{w \cdot h} \cdot \sum_{i=1}^{w \cdot h} lc_i \tag{17}$$

To achieve more accurate results, the average contrast of multiple resolutions has to be calculated. The paper [6] suggests nine resolutions, as more resolutions do not yield enough information to justify the computational costs. The proposed framework downsamples the image eight times by halving the width and height via bilinear interpolation and calculates the average local contrast for each resolution level. Finally, the GCF is calculated by summing up the weighted average local contrasts:

$$GCF = \sum_{i=1}^{9} w_i \cdot C_i \tag{18}$$

[1] as specified in IEC 61966-2-1

Matkovic et al. [6] conducted experiments, where users were asked to compare contrast between images to find the optimal (approximated) weights w_i that correlate with the user ratings. Since FA-images are still analyzed by human readers, the proposed framework uses the same weights for its calculation:

$$w_i = (-0.406385 \cdot \frac{i}{9} + 0.334573) \cdot \frac{i}{9} + 0.0877526 \tag{19}$$

The resulting quality vector contains every weighted average local contrast and thus nine values.

Because FA-images are recorded over time to document the flow of the fluorescent dye, there will be images (mainly at the beginning and the end of a patients medical examination) where the dye is not circulating through the eye. These images will have a negative classification, even if they were taken under perfect conditions (i.e. no human error and no hardware defects). This is not a false classification, as there cannot be any contrast without the fluorescent dye. Using the time-stamp provided with every image, readers will be able to ascertain whether the contrast classification is relevant or not.

The biggest disadvantage of this method is the amount of time-consuming calculations, making it the slowest feature of the proposed framework. To accelerate the calculations and make it possible to use this algorithm in real-time application, a 3x3/4 Pyramid[2] is employed, as it is more efficient than equivalent weighting functions [5][2]. That way, the filter kernel does not change in size, limiting the amount of calculations per pixel. Additionally, the images are resized to a width of 1,600 pixels, if they are larger. However, downscaling leads to the loss of higher frequencies and thus a less accurate result. If the hardware of the computing device is powerful enough, this algorithm can be used in real-time without resizing.

[2] An image pyramid with a 3x3 reduction window, of which the image size is quartered every level.

2.6 Noise (FA)

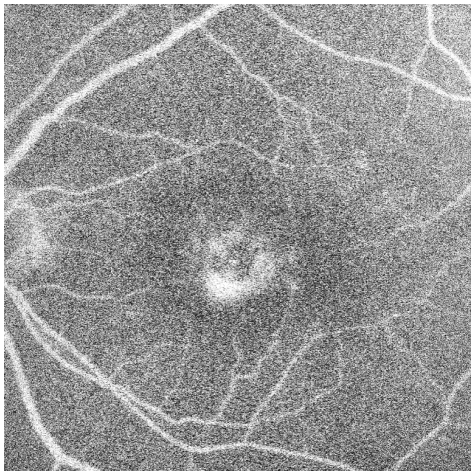


Figure 10: Example for a noisy FA-Image.

Noise is an image artifact that occurs due to hardware limitations of a camera. Because of it, random pixels change in color and brightness, making it harder to discern important image features, especially edges. This is vital in FA-images, as readers need to be able to recognize blood vessels to analyze an eye.

The most common type of noise distortion in the provided dataset is salt-and-pepper noise, as the histogram of every noisy image exhibited a fat-tailed distribution[3] (see Figure 11 for an example), which is an indicator for salt-and-pepper noise [1].

Not many reference-less methods exist to calculate the noise ratio of an image. Due to their simplicity, the Signal-To-Noise-Ratio (SNR) and its variation Peak-Signal-To-Noise-Ratio (PSNR) are tried and tested methods. However, PSNR cannot be used in this case, as it requires a reference; Every retina has a unique pattern of blood vessels [11], making it impossible to reliably define a reference image.
While each eye is photographed repeatedly in FA, it cannot be guaranteed that at least one of the pictures is good enough to use as reference for other

[3] a distribution with large skewness or kurtosis relative to a normal distirbution

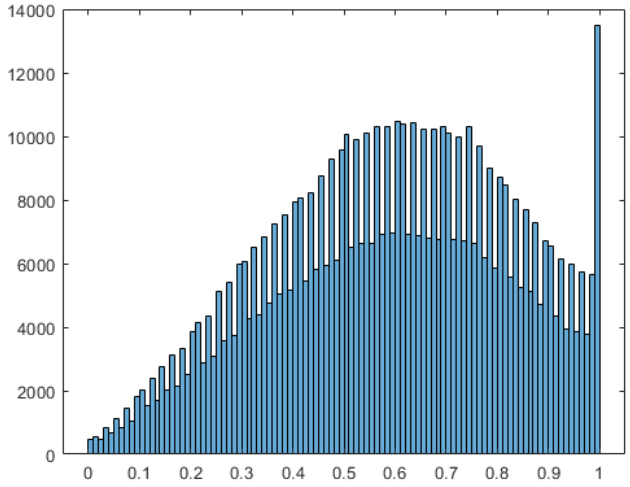


Figure 11: Histogram of example image shown in Figure 10.

pictures of the same eye; The "best" picture of a series might still not be good enough.

One of the latest algorithms is the no-reference noise metric from Yang et al. [15], which promises great accuracy and stability of the assessments. However, while this method is precise, it proved to be too computationally taxing for real-time usage together with the other quality features. The average runtime was over one minute per image on a device with the hardware specifications listed in Table 16. While this might seem like an acceptable runtime, it has to be considered that devices used in medical institutions could be significantly weaker.

Via process of elimination it was decided to use the simple Signal-To-Noise-Ratio (SNR), which is calculated as the ratio of the average signal value to its standard deviation by:

$$SNR = \frac{\mu_{sig}}{\sigma_{sig}} \tag{20}$$

Due to its simplicity, noise is the fastest feature to calculate while returning acceptable results (see Sections 3 and 4). If the computing hardware is powerful enough, it is possible to use a more accurate method like the previously

mentioned no-reference metric by Yang et al. [15] in real-time.

3 Classification Module Training and Feature Evaluation

After calculating each quality vector, another problem emerges: since each feature is calculated differently, the resulting values have different scales and ranges.

One idea was to manually find reasonable min/max values and use them to convert the range of every feature to a 0-10 or 0-100 scoring system. The problem with this approach however, is the probable limitation of the values to the provided dataset, which would significantly impact the generalizability of the framework.

To solve this problem, classification modules are used. This has two benefits, the first one being the efficiency of machine learning in finding a good threshold. The second advantage is the possibility to train the framework with another, larger dataset, further enhancing classification.

There are various classification methods, such as Support Vector Machines, Boosting, Decision Trees, or a simple Naive Bayes classification.

3.1 Classification Evaluation

One way to evaluate the performance of classification models is the Receiver Operating Characteristic (ROC) curve and its corresponding Area Under the Curve (AUC). To calculate them, two statistics are needed: the sensitivity (or true positive rate)

$$\frac{GQ_G}{GQ_G + GQ_B} \tag{21}$$

where GQ_G is the amount of good quality images that are classified "good quality" and GQ_B is the amount of good quality images that are classified "bad quality",
and the false negative rate (or $1 - specificity$)

$$1 - \frac{BQ_B}{BQ_B + BQ_G} \tag{22}$$

where BQ_B is the amount of bad quality images that are classified "bad quality" and BQ_G is the amount of bad quality images that are classifed "good quality".
A ROC curve plots sensitivity against the false negative rate. The AUC is the area under the ROC Curve and describes how good a model can distinguish between classes. An AUC of 0.5 means that it cannot distinguish at all, while an AUC of 1.0 describes perfect separability; Analogically, a value of 0.0 describes perfect separability with inverted classification. However, an AUC of 1.0 (or 0.0) does not necessarily mean that the perfect classifier has been found, but merely that the tested dataset is perfectly separable. The bigger the dataset is, the more likely it is to contain outliers that lower separability.

The proposed framework compares the AUC of each of those methods and saves the best method as the classification module.
For FA, training was done using a training set of 300 "bad quality" images for each feature and 300 "good quality" images, resulting in a total of 1,200 images. The calculated ROC curves and AUCs for each quality feature are shown in Figures 12-14 and Tables 2-6; The confusion matrices of the best classification module for each feature are shown in Tables 3-7.
For CF, about 600 images per category have been selected as the dataset: 150 images with good quality in general (which were used for all categories),

150 images with good quality regarding the evaluated features of that category and 300 images with bad quality for this specific category. These 2,087 images have been manually labeled and were used to train the classification modules. The calculated ROC curves and AUCs for each quality feature are shown in Figures 15-18 and Tables 8-14; The confusion matrices of the best classification module for each feature are shown in Tables 9-15. The best AUCs and classification losses have been marked bold.

After training, feature evaluation is done by calculating the quality vectors for every feature of the input image(s) and then passing them to their respective classification module.

To offer a more precise scoring than the binary "good" and "bad" of the classification modules, the proposed framework also returns the certainties of the classifications. The higher the value, the more likely it is for the image to be of sufficient quality.

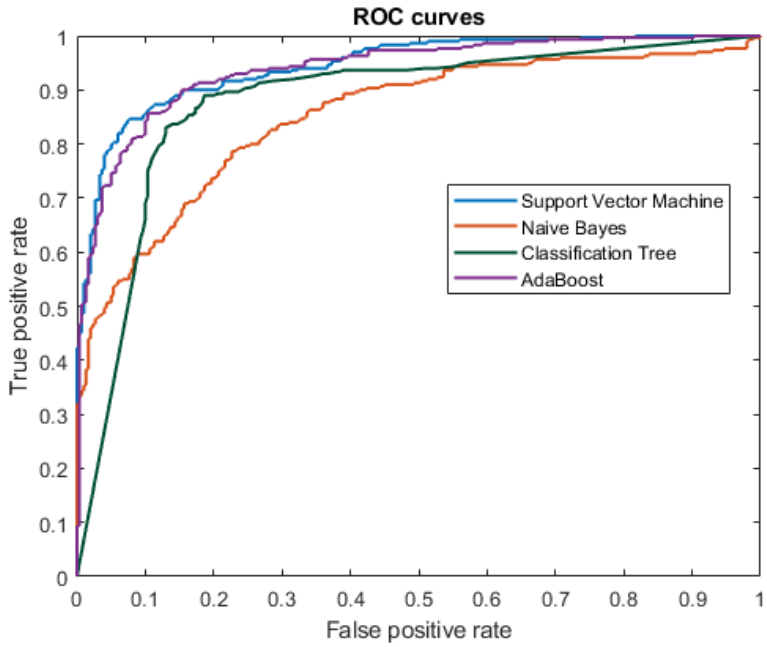


Figure 12: ROC curves of the FA contrast classification modules.

Table 2: AUCs and Classification Losses of the FA contrast classification.

Classification Method	AUC	Class Loss
Adaptive Boosting	0.9385	0.1350
Classification Trees	0.8789	0.1583
Naive Bayes	0.8513	0.2500
Support Vector Machines	**0.9444**	**0.1167**

Table 3: Confusion Matrix of Support Vector Machines for FA contrast classification

	Predicted Class	
Actual Class	Bad Quality	Good Quality
Bad Quality	279	21
Good Quality	49	251

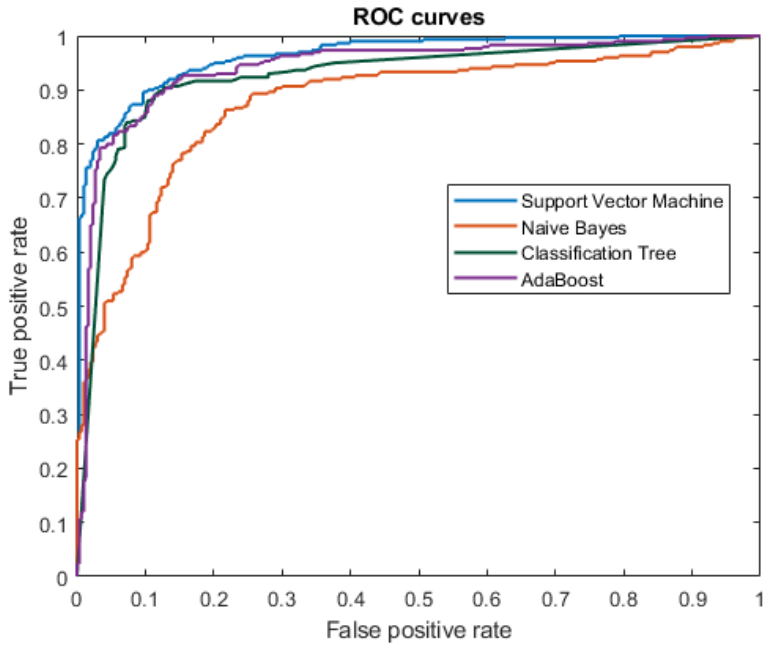


Figure 13: ROC curves of the FA acutance classification modules.

Table 4: AUCs and Classification Losses of the FA acutance classification.

Classification Method	AUC	Class Loss
Adaptive Boosting	0.9423	**0.1117**
Classification Trees	0.9256	0.1150
Naive Bayes	0.8753	0.2367
Support Vector Machines	**0.9631**	0.1150

Table 5: Confusion Matrix of Support Vector Machines for FA acutance classification

	Predicted Class	
Actual Class	Bad Quality	Good Quality
Bad Quality	282	18
Good Quality	51	249

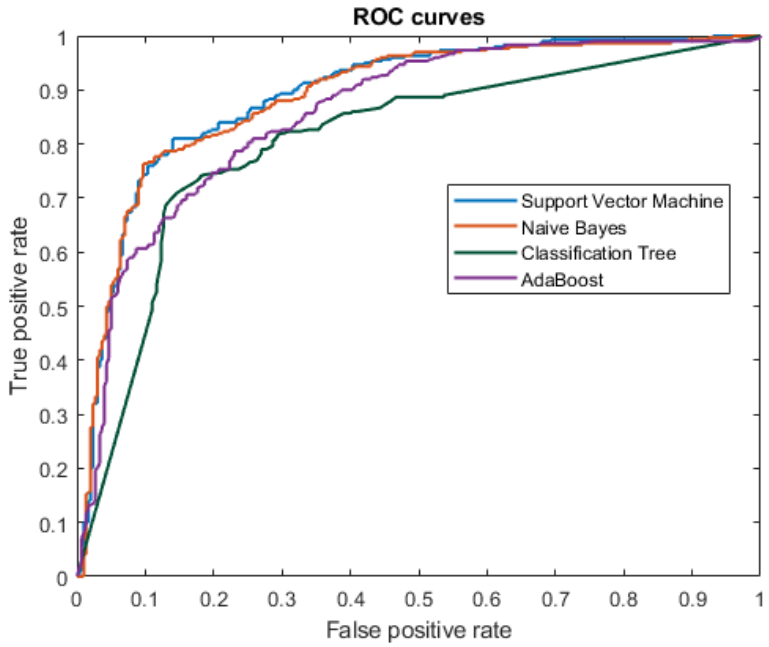


Figure 14: ROC curves of the FA noise classification modules.

Table 6: AUCs and Classification Losses of the FA noise classification.

Classification Method	AUC	Class Loss
Adaptive Boosting	0.8570	0.2267
Classification Trees	0.8076	0.2433
Naive Bayes	0.8883	0.2033
Support Vector Machines	**0.8920**	**0.1767**

Table 7: Confusion Matrix of Support Vector Machines for FA noise classification

	Predicted Class	
Actual Class	Bad Quality	Good Quality
Bad Quality	267	33
Good Quality	71	229

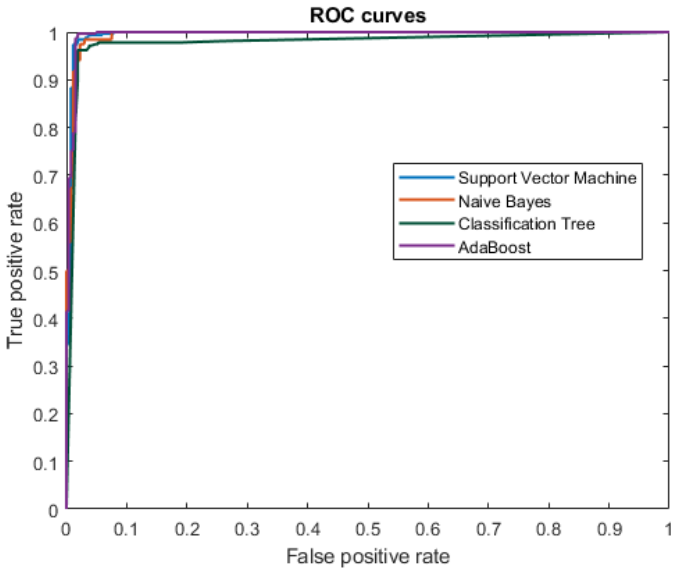


Figure 15: ROC curves for all four CF color classification modules.

Table 8: AUCs and Classification Losses for CF color classification.

	AUC	Class Loss
Adaptive Boosting	**0.9948**	**0.0137**
Classification Tree	0.9780	0.0360
Naive Bayes	0.9936	0.0240
Support Vector Machine	0.9942	0.0223

Table 9: Confusion Matrix of Adaptive Boosting for CF color classification

	Predicted Class	
Actual Class	Bad Quality	Good Quality
Bad Quality	262	4
Good Quality	4	314

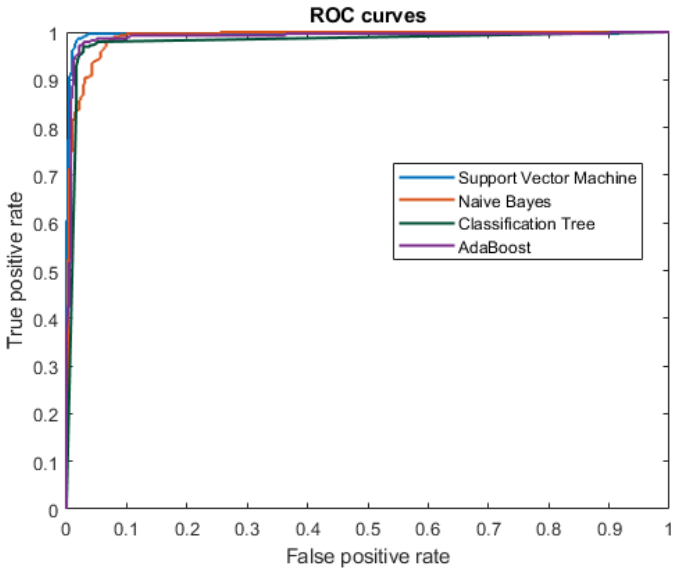


Figure 16: ROC curves for all four CF acutance classification modules.

Table 10: AUCs and Classification Losses for CF acutance classification.

	AUC	Class Loss
Adaptive Boosting	0.9901	0.0251
Classification Tree	0.9797	0.0292
Naive Bayes	0.9892	0.0501
Support Vector Machine	**0.9947**	**0.0209**

Table 11: Confusion Matrix of Support Vector Machines for CF acutance classification

	Predicted Class	
Actual Class	Bad Quality	Good Quality
Bad Quality	418	6
Good Quality	9	285

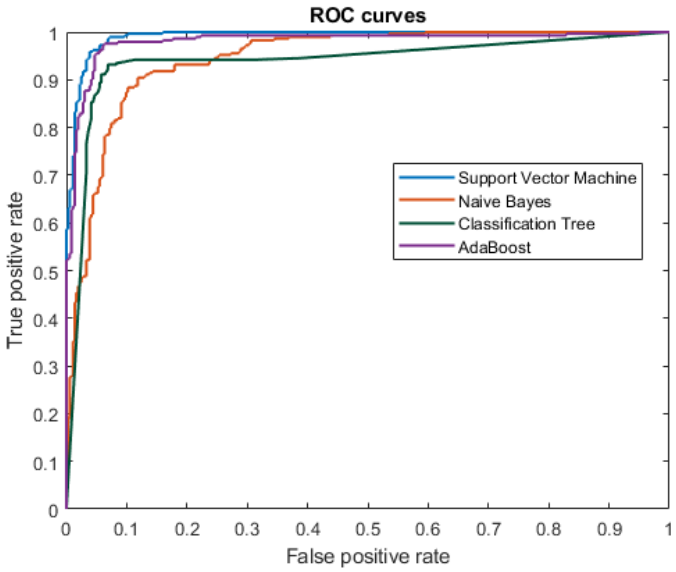


Figure 17: ROC curves for all four CF contrast classification modules.

Table 12: AUCs and Classification Losses for CF contrast classification.

	AUC	Class Loss
Adaptive Boosting	0.9810	0.0503
Classification Tree	0.9385	0.0732
Naive Bayes	0.9448	0.1128
Support Vector Machine	**0.9912**	**0.0412**

Table 13: Confusion Matrix of Support Vector Machines for CF contrast classification

	Predicted Class	
Actual Class	Bad Quality	Good Quality
Bad Quality	347	16
Good Quality	11	282

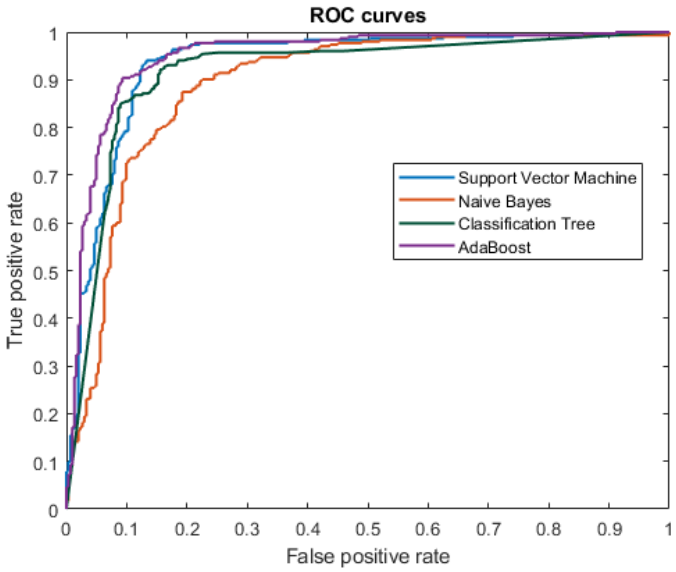


Figure 18: ROC curves for all four CF illumination classification modules.

Table 14: Classloss and AUC for CF illumination classification.

	AUC	Class Loss
Adaptive Boosting	**0.9486**	0.1025
Classification Tree	0.9160	0.1289
Naive Bayes	0.8920	0.2083
Support Vector Machine	0.9340	**0.0992**

Table 15: Confusion Matrix of Adaptive Boosting for CF illumination classification

	Predicted Class	
Actual Class	Bad Quality	Good Quality
Bad Quality	265	37
Good Quality	25	278

4 Runtime Evaluation

For FA, the average runtime of the framework has been calculated by timing the computation of 600 images and is shown in Table 17. The hardware specifications of the computing device that was used is listed in Table **??**.

Table 16: Hardware specifications of the device used to time FA method runtimes

Component Type	Component Installed
Processor	Intel Core i7-4790K
Memory	16GB DDR3 RAM
Graphics Card	NVIDIA GTX 970
Operating System	Windows 10 64bit

Table 17: Average Runtime per image (600 FA images)

Feature	Average Runtime (in seconds)
Contrast/Illumination	4.375
Acutance	0.381
Noise	0.360
Total:	**5.116**

Because Contrast/Illumination is of highest priority, the extent of accuracy reduction is limited, making it the slowest feature calculation by far. However, thanks to the efficiency of Acutance and Noise calculations, the proposed framework still manages to deliver an adequate runtime for real-time application.

For CF, the framework has been tested on a device with the specifications listed in Table 18.

Table 18: Hardware specifications of the device used to time CF method runtimes

Component Type	Component Installed
Processor	Intel Core i7-7700HQ
Memory	8GB DDR3 RAM
Graphics Card	NVIDIA GTX 1050
Operating System	Windows 10 64bit

The needed amount of time to calculate the features of every category are shown in Table 19.

Table 19: Average Runtime per CF image

Feature	Average Runtime (in seconds)
Color	1.53
Focus	0.61
Contrast	8.68
Illumination	9.11
Total	**19.93**

5 Conclusion

This report proposes an efficient way to classify and grade CF/FA-images, while discussing some of the difficulties and problems present in image quality assessment of retinal scans. With the proposed framework, medical institutions can assess the quality of CF/FA-images in real-time, provided that the classification modules are properly trained beforehand.
Furthermore, the modular structure of the framework makes it future-proof by making it possible to update the components as soon as there are breakthroughs in IQA for any of the previously mentioned quality factors.
With proper training, this framework could be used as part of retinal image segmentation, further accelerating the process of eye diagnosis.

5.1 Future research

By using multiple threads, a parallel calculation of the different categories would be possible, which would reduce the required amount of time.

Furthermore, adapting the training sets for the used classification modules would bring further improvement in accuracy. By increasing the number of used images for each category, the machine can learn to differentiate more accurately and predict border cases more precisely. Aside from the size of the training set, the suitability of each image is of high importance. Therefore, the images used to train the classification modules should be reviewed by professionals.

References

[1] M. Aggarwal, R. Kaur, and B. Kaur. Design of efficient adaptive image filters to suppress salt and pepper noise. *Journal of Emerging Trends in Computing and Information Sciences*, 5(9), 2014.

[2] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on communications*, 31(4):532–540, 1983.

[3] J. M. P. Dias, C. M. Oliveira, and L. A. da Silva Cruz. Retinal image quality assessment using generic image quality indicators. *Information Fusion*, 19:73–90, 2014.

[4] A. Hoover and M. Goldbaum. Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels. *IEEE Tansactions on Medical Imaging*, 22:951–958, 2003.

[5] W. G. Kropatsch. From equivalent weighting functions to equivalent contraction kernels. In *Sixth International Workshop on Digital Image Processing and Computer Graphics: Applications in Humanities and Natural Sciences*, volume 3346, pages 310–321. International Society for Optics and Photonics, 1998.

[6] K. Matkovic, L. Neumann, A. Neumann, T. Psik, and W. Purgathofer. Global contrast factor-a new approach to image contrast. *Computational Aesthetics*, 2005:159–168, 2005.

[7] A. A. Michelson. *Studies in optics*. Courier Corporation, 1995.

[8] N. Panwar, P. Huang, J. Lee, P. A. Keane, T. S. Chuan, A. Richhariya, S. Teoh, T. H. Lim, and R. Agrawal. Fundus photography in the 21st century - a review of recent technological advances and their implications for worldwide healthcare. *Telemedicine and e-Health*, 22:198–208, 2016.

[9] E. Peli. Contrast in complex images. *JOSA A*, 7(10):2032–2040, 1990.

[10] F. Shao, Y. Yang, Q. Jiang, G. Jiang, and Y.-S. Ho. Automated quality assessment of fundus images via analysis of illumination, naturalness and structure. *IEEE Access*, 6:806–817, 2017.

[11] C. Simon. A new scientific method of identification. *New York state journal of medicine*, 35(18):901–906, 1935.

[12] Y. Tu, F. Yi, G. Chen, S. Jiang, and Z. Huang. Illumination removal in color images using retinex method based on advanced adaptive smoothing. In *2010 International Conference on E-Health Networking, Digital Ecosystems and Technologies*, pages 219–223, 2010.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[14] G. Yang, L. Gagnon, S. Wang, and M.-C. Boucher. Algorithm for detecting micro-aneurysms in low-resolution color retinal images. In *Proceedings of Vision Interface*, pages 265–271, 2001.

[15] G. Yang, Y. Liao, Q. Zhang, D. Li, and W. Yang. No-reference quality assessment of noise-distorted images based on frequency mapping. *IEEE Access*, 2017.