The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/159478

Please be advised that this information was generated on 2019-04-11 and may be subject to change.
Automated analysis of retinal images for detection of age-related macular degeneration and diabetic retinopathy

Mark van Grinsven
Automated analysis of retinal images for detection of age-related macular degeneration and diabetic retinopathy

Proefschrift

Ter verkrijging van de graad van doctor aan de Radboud Universiteit Nijmegen op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken, volgens besluit van het college van decanen in het openbaar te verdedigen op donderdag 29 september om 12.30 uur precies.

Door

Mark Johannes Josephus Petrus van Grinsven

Geboren op 22 april 1987 te Sint-Oedenrode
The research described in this thesis was carried out at the Diagnostic Image Analysis Group, department of Radiology and Nuclear Medicine, Radboud University Medical Center (Nijmegen, The Netherlands).

This work was funded by grant ZonMW 11.631.0003 from the Dutch Organization for Health and Healthcare Innovation.

Financial support for publication of this thesis was kindly provided by the department of Radiology and Nuclear Medicine at the Radboud University Medical Center, Nijmegen, The Netherlands.
# TABLE OF CONTENTS

## 1 Introduction

1.1 The human eye .................................................. 5
1.2 Retinal imaging ............................................... 6
1.3 Retinal diseases: Diabetic Retinopathy ..................... 7
1.4 Retinal diseases: Age-related Macular Degeneration ...... 10
1.5 Screening for retinal diseases ............................... 13
1.6 Computer aided detection .................................... 15
1.7 A review on CAD systems for automatic retinal image analysis .. 22
1.8 Thesis outline .................................................. 27

## 2 Drusen quantification and risk assessment of Age-related Macular Degeneration

2.1 Introduction ................................................... 31
2.2 Methods ....................................................... 32
2.3 Results ......................................................... 36
2.4 Discussion ..................................................... 39
2.5 Appendix A .................................................... 42

## 3 Automatic identification of reticular pseudodrusen

3.1 Introduction ................................................... 47
3.2 Methods ....................................................... 48
3.3 Results ......................................................... 52
3.4 Discussion ..................................................... 55

## 4 Self-learning systems for reticular pseudodrusen detection

4.1 Introduction ................................................... 61
4.2 Methods ....................................................... 63
4.3 Results ......................................................... 65
4.4 Discussion ..................................................... 69
4.5 Appendix A .................................................... 70

## 5 Automatic differentiation of drusen and exudates

5.1 Introduction ................................................... 77
5.2 Materials and Methods ....................................... 78
5.3 Results ......................................................... 85
5.4 Discussion ..................................................... 87
6 Hemorrhage detection using convolutional networks with a fast training scheme 95
6.1 Introduction .................................................. 97
6.2 Materials ..................................................... 99
6.3 Methods ....................................................... 100
6.4 Experimental design ........................................ 106
6.5 Results ......................................................... 107
6.6 Discussion ..................................................... 111
6.7 Conclusion .................................................... 117

7 General discussion ............................................... 119

8 Summary .......................................................... 129

9 Samenvatting ..................................................... 135

Publications .......................................................... 141

Bibliography ......................................................... 147

Dankwoord .......................................................... 183

Curriculum Vitae .................................................... 189
Introduction
In 1999, the World Health Organization (WHO) and the International Agency for the Prevention of Blindness (IAPB) have launched a global initiative known as "VISION 2020: the right to sight". The main objective of this initiative is to eliminate the main causes of avoidable blindness by the year 2020 by facilitating the planning, development and implementation of sustainable national eye-care programs. The targets of the program have been updated and now include a 25% reduction of avoidable blindness and visual impairment by 2019 as compared to 2010. Every year on the second Thursday of October, World Sight Day (WSD) is held to focus global attention on blindness and visual impairment as major international public health issues. VISION 2020 comprises three main components including cost-effective disease control, human resource development, and infrastructure and technology.

In this thesis, an effort is made to pursue the VISION 2020 goals of eliminating avoidable blindness and visual impairment worldwide. It describes and validates new automatic methods to detect Diabetic Retinopathy (DR) and Age-related Macular Degeneration (AMD), two of the most common retinal diseases worldwide. Achieving automatic detection of these diseases will facilitate and accelerate implementation of screening programs for retinal diseases worldwide. It is estimated that 80% of blindness is preventable if timely awareness of presence of these diseases is achieved.
1.1 The human eye

The human eye is an organ which allows humans to interact with the environment. Light-sensitive cells, called photoreceptors, generate electrical signals that are sent to the brain. The brain interprets these signals and generates an interpretation of the world: sight. Figure 1.1a shows a schematic overview of the human eye.

![Schematic overview of the human eye](image)

Light enters the eye through several structures, such as the cornea and the lens, which focus the light on the retina where all photoreceptors are located. Figure 1.1b shows a color fundus image (see Section 1.2) with several anatomical structures highlighted. The retina contains two types of photoreceptors: rods and cones. The number of rods is in the order of 60 million. They are mainly located at the peripheral part of the retina and allow for peripheral vision, light and dark contrast perception and detection of motion. The 3 million cones are used for color and central vision. Cones are concentrated in the macular region, with a peak in density at the central macula, called the fovea, which allows for sharp central vision. The retina itself consists of multiple layers, each with specific properties, functions and characteristics, see Figure 1.2d. Electrical signals, produced by the photoreceptors, exit the eye through the optic nerve fiber which is connected to the visual cortex of the brain. The spot where the optic nerve fibers exit the eye, i.e. the optic disc, does not contain any photoreceptors and is therefore referred to as the blind spot. The retinal vasculature, also entering and exiting the eye through the optic disc, forms a fine network of vessels that supplies the retina.

---

*Source: image courtesy of NIH, National Eye Institute*

*Source: kindly provided by the Messidor program partners (see [http://messidor.crihan.fr](http://messidor.crihan.fr))*

Figure 1.1: (a): Schematic overview of the human eye. (b): Color fundus image.
with nutrients and oxygen. Damage to any of these structures in the eye can lead to visual distortions, vision loss, or, in the worst case, blindness.

1.2 Retinal imaging

Retinal imaging has become an indispensable diagnostic tool in the field of ophthalmology. It allows to observe the retina with enormous detail and can help to identify eye diseases in early stages, even before any visual symptoms appear. Different characteristics of the retina can be observed with a variety of imaging techniques, such as color fundus (CF) photography, fundus autofluorescence (FAF) imaging, near infrared (NIR) imaging and optical coherence tomography (OCT) amongst many others\textsuperscript{2-4}. This has resulted in a better understanding and characterization of retinal diseases and helps in monitoring disease progression. Figure 1.2 shows examples acquired with these different modalities.

![Figure 1.2: (a): Color fundus image, (b): Fundus autofluorescence image, (c): Near infra-red image of the same eye. (d): Optical coherence tomography: Single B-scan extracted from a volume scan.](www.hoaclr.org)
1.3 Retinal diseases: Diabetic Retinopathy

In CF photography, a color image is acquired as in standard photography. A CF camera encapsulates a set of special lenses focusing the imaging plane of the camera onto the retina and a digital camera, attached to the fundus camera, is used to store the obtained images digitally. During image acquisition, a bright flash is used to lighten the retina to acquire well illuminated CF images. This form of retinal imaging is the most widely used technique to image the retina.

In FAF imaging, advantage is taken of the autofluorescent properties of lipofuscin. Lipofuscin is a byproduct of photoreceptors that accumulates in the retinal pigment epithelium (RPE). When exposed to short- to medium-wavelength visible light (300nm-600nm), lipofuscin autofluoresces and FAF imaging is used to document its accumulation. The optic disc and blood vessels appear as dark structures on FAF images as they do not contain lipofuscin and blood vessels also cover up the underlying lipofuscin.

NIR imaging is very similar to CF imaging, except for the difference in wavelength used to image the retina. Furthermore, no flash is required for acquisition of a NIR image. Typical wavelengths used for NIR imaging are around 820nm. NIR imaging allows for visualizing structures which are in the deeper layers of the retina as a higher wavelength has a higher penetration depth.

OCT is an imaging technique based on interferometry to capture depth-resolved images of the retina. A beam of broadband near-infrared light is directed at a mirror and the retina. The light reflected back from the retina and mirror is rejoined in an optical fiber, resulting in coherent interference of the light reflected from the retina with the reference light returning from the mirror. The resulting light is spectrally analyzed by a spectrometer. By Fourier-transformation of this spectrum, the amount of light reflected from different depths of the retina can be determined. By repeating this procedure for laterally adjacent sites of the retina, a three-dimensional image can be obtained. Differences in the optical properties of the retinal layers typically cause a variable amount of back-reflected light, which is visualized as a multi-layered retinal structure. OCT imaging is increasingly used in clinical retinal imaging, but in this thesis we focus on 2D retinal images.

1.3 Retinal diseases: Diabetic Retinopathy

DR is an eye disorder affecting the retinal vasculature that gradually develops in patients with diabetes. Patients with either type 1 or type 2 diabetes can be affected by the disease. Eventually, the majority of patients with diabetes will develop DR to some extent. The increased blood glucose levels in patients with diabetes has several effects on the retinal vessels. One of the effects is that the permeability of the retinal vessels increases, causing weakness of the capillary walls and leading to leakages of blood
and lipid components into the retina. Another effect is that retinal vessels may be occluded and this leads to a local lack of oxygen and nutrients, called ischemia\(^7\). Visual complaints come in the form of dark spots in the vision, correlated with the location of the retina affected. Figure 1.3 shows a scene as viewed by a person with normal vision and a simulation of how a person with DR would see the scene.

![Scene with Normal Vision](image1.png) ![Scene with Diabetic Retinopathy](image2.png)

Source: images courtesy of NIH, National Eye Institute

Figure 1.3: (a): Scene as viewed by a person with normal vision. (b): Simulated scene as viewed by a person with diabetic retinopathy.

### 1.3.1 Prevalence of Diabetic Retinopathy

DR is the most common cause of preventable blindness in the working-age population\(^8\). Worldwide, around 250 million people have diabetes. Incidence rates of DR vary between 74% for patients having 10 years of diabetes to 97% of patients having 25 years of diabetes\(^9\)\(^-\)\(^{11}\). Prevalence rates of DR are projected to increase rapidly in the near future\(^12\),\(^13\). Although complete prevention of development of DR for patients with diabetes is nearly impossible, strict control of glucose levels reduces the risk of development and progression of DR in both types of diabetes\(^8\).

### 1.3.2 Diagnosis of Diabetic Retinopathy

According to clinical guidelines, DR can be categorized into several stages ranging from mild, moderate and severe non-proliferative diabetic retinopathy (NPDR) to proliferative diabetic retinopathy (PDR)\(^14\),\(^15\). Each of the stages is characterized by the presence of different types and extent of retinal lesions. Common retinal microvascular signs of NPDR include microaneurysms (small swellings that forms in the wall of tiny blood vessels), hemorrhages (bleedings), hard exudates (leakages of lipids), cotton wool spots (accumulations of axoplasmic\(^16\)) and venous beading\(^14\). Figure 1.4 shows a CF image with lesions associated with DR. Table 1.1 shows the findings for each DR severity level.
1.3 Retinal diseases: Diabetic Retinopathy

Table 1.1: International clinical diabetic retinopathy diseases severity scale\textsuperscript{15}.

<table>
<thead>
<tr>
<th>DR severity level</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>0: No DR</td>
<td>No abnormalities</td>
</tr>
<tr>
<td>1: Mild NPDR</td>
<td>Microaneurysms only</td>
</tr>
<tr>
<td>2: Moderated NPDR</td>
<td>More than just microaneurysms but less than severe NPDR</td>
</tr>
<tr>
<td>3: Severe NPDR</td>
<td>More than 20 intra-retinal hemorrhages in each of 4 quadrants or definite venous beading in 2 or more quadrants or prominent intra-retinal microvascular abnormalities in 1 or more quadrants and no signs of PDR</td>
</tr>
<tr>
<td>4: PDR</td>
<td>Neovascularization and/or vitreous/preretinal hemorrhage</td>
</tr>
</tbody>
</table>

In the mild NPDR stage, only a few microaneurysms appear. These microaneurysms may leak fluid into the retina. As the disease progresses, the retinal blood vessels may swell and distort. The number of microaneurysms starts to increase and hemorrhages start to appear. At this phase, the DR stage is considered moderate NPDR. In Chapter 6 of this thesis, an automatic system to detect hemorrhages is presented and evaluated. In the severe NPDR stage, the disease has progressed and hard exudates or cotton wools spots, which are caused by the lack of blood supply due to blocked vessels, can appear. An extensive number of microaneurysms and hemorrhages appear in this phase of the disease. In the PDR stage, new fragile vessels (neovascularization) start to form in the retina. These can form on the optic disc (NVD) or elsewhere on the retina (NVE). These newly formed vessels are weak and prone to break easily, causing serious damage to the retina.

CF imaging is the mainstream modality to assess DR as it is cheap, non-invasive and widely available\textsuperscript{17}. DR can be detected using this modality by identifying individual lesions. When DR is identified and lesions are detected, other imaging modalities, such as fluorescein angiography and OCT, are used to provide additional information on treatment response and allow for precise and reproducible measurements. DR progression is assessed by identifying newly formed lesions as compared with previous examinations, if available.

1.3.3 Treatment for Diabetic Retinopathy

Treatment options to slow down the progression and prevent severe vision loss exist for DR. However, for the treatment to be effective, the disease has to be detected while residing in the early stages. Monitoring and managing blood sugar levels is the most important factor for controlling DR and preventing progression to the late stages and can be done by the patients themselves\textsuperscript{18}. Treatment to arrest blood leakages
is applied by using laser photocoagulation\textsuperscript{17,19,20}. This laser treatment exists in two forms: panretinal photocoagulation and macular (focal or grid) photocoagulation. In panretinal photocoagulation, the goal is to place laser burns over the entire peripheral retina to arrest progression of retinal neovascularization. In macular photocoagulation, laser beams are applied to leaking microaneurysms (focal) or to areas of non-perfusion (grid) in the retina, preventing further bleedings. Other treatment comes in the form of anti-vascular endothelial growth factor (anti-VEGF) drugs\textsuperscript{21}. Anti-VEGF drugs are injected into the vitreous gel to block a protein called vascular endothelial growth factor (VEGF), which stimulate growth of abnormal, fragile blood vessels. Blocking VEGF can reduce abnormal blood vessel growth and decrease fluid in the retina. Available anti-VEGF drugs include Avastin (bevacizumab), Lucentis (ranibizumab), and Eylea (aflibercept)\textsuperscript{22}. Lucentis and Eylea are approved by the U.S. Food and Drug Administration (FDA) for treating DR. Avastin was approved by the FDA to treat cancer, but is commonly used to treat eye conditions, including DR.

1.4 Retinal diseases: Age-related Macular Degeneration

AMD is the most common eye disease in the elderly which causes irreversible vision loss and ultimately blindness in the final stages of the disease\textsuperscript{23}. AMD progresses from early and intermediate AMD, with no or only subtle visual complaints, to advanced AMD, where severe vision loss can occur. Vision loss due to AMD usually becomes
noticeable at the age of 60 to 70 and tends to worsen over time. AMD affects the central vision hampering patients to perform detailed tasks such as reading, driving a car or recognizing faces. Figure 1.5 shows a scene as viewed by a person with normal vision and a simulated scene as viewed by a person with AMD.

![Scene viewed by normal vision and AMD](image)

Source: images courtesy of NIH, National Eye Institute

Figure 1.5: (a): Scene as viewed by a person with normal vision. (b): Simulated scene as viewed by a person with age-related macular degeneration.

1.4.1 Prevalence of Age-related Macular Degeneration

The number of people suffering from AMD is estimated to be 196 million in 2020 worldwide, with an expected increase to 288 million in 2040\(^24\). Without treatment, the number of visual impaired and blind people due to AMD is estimated to triple in the coming years\(^25\). Although little is known about the disease progression, certain factors such as increased age, family history, gender, race and lifestyle habits such as smoking have shown to increase the risk for development of advanced AMD\(^{26-29}\).

1.4.2 Diagnosis of Age-related Macular Degeneration

AMD is typically categorized into several stages, including early AMD, intermediate AMD and advanced AMD\(^{30,31}\). Table 1.2 shows an overview of the AMD severity levels with associated findings. The early and intermediate stages of AMD are characterized by the presence of drusen and pigmentary changes\(^{30,31}\). Figure 1.6 shows a CF image with drusen. Drusen are extracellular deposits located between the Bruch’s membrane and the basal lamina of the RPE. The stage is early AMD when only a few small drusen are present; if there are more and/or larger drusen, the stage is intermediate AMD\(^{31-34}\). In Chapter 2, an automatic system to detect and quantify drusen is described and evaluated.
Introduction

Table 1.2: Criteria for grading AMD according to the CIRCL Grading Protocol.

<table>
<thead>
<tr>
<th>AMD severity level</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: No AMD</td>
<td>No drusen or small, hard drusen only.</td>
</tr>
<tr>
<td>2: Early AMD</td>
<td>&gt;10 small (&lt;63 μm), hard drusen + pigmentary changes or 1-15 intermediate (63-124 μm) drusen.</td>
</tr>
<tr>
<td>3: Intermediate AMD</td>
<td>&gt;15 intermediate (63-124 μm) drusen or any large (≥125 μm) drusen or GA not in the central circle of the ETDRS grid.</td>
</tr>
<tr>
<td>4: Advanced AMD (GA)</td>
<td>Presence of central GA.</td>
</tr>
<tr>
<td>5: Advanced AMD (CNV)</td>
<td>Evidence of active or previous CNV lesion.</td>
</tr>
<tr>
<td>6: CNV without signs for AMD</td>
<td>Chosen if CNV is present but no drusen of any size are present within the Field 2.</td>
</tr>
<tr>
<td>7: Cannot grade</td>
<td>Image is regarded as not gradable.</td>
</tr>
</tbody>
</table>

AMD, age-related macular degeneration; CIRCL, Cologne Image Reading Center and Laboratory; CNV, choroidal neovascularization; ETDRS, Early Treatment Diabetic Retinopathy Study; GA, geographic atrophy.

Recent research shows that more attention should be given to extracellular lesions in AMD commonly termed as "subretinal drusenoid deposits" or "reticular pseudodrusen" (RPD). These lesions have different characteristics and a different distribution compared to drusen and they are a strong risk factor for progression to advanced AMD\textsuperscript{35–37}. In Chapter 3 and Chapter 4, a study on the detection of RPD using multi-modal imaging is performed and automatic systems to detect these lesions using multi-modal imaging are presented.

Around 16% of people with AMD will progress to advanced AMD\textsuperscript{38}. Two forms of advanced AMD exists: advanced AMD with presence of geographic atrophy (GA) (often referred to as dry AMD); and advanced AMD with presence of choroidal neovascularization (CNV) (often referred to as wet AMD). The dry form of AMD is much
more common and accounts for approximately 85% of all advanced AMD cases. In the dry form, there is a gradual breakdown of photoreceptors in the macula and of the supporting tissue beneath the macula. The wet form is more aggressive. Here abnormal, fragile blood vessels start to grow underneath the macula, which are prone to leak blood and fluid causing damage to the macula.

Identification of AMD is mainly done using manual assessment of CF images by trained graders or ophthalmologists. Additional imaging modalities such as FAF imaging and OCT imaging are often used after identification of AMD to make a more detailed diagnosis of AMD. Multi-modal imaging can provide more details on disease severity and disease progression.

1.4.3 Treatment for Age-related Macular Degeneration

Treatment options for AMD are still limited. Lifestyle changes such as cessation of smoking and prophylactic regimens like vitamin supplementation are recommended for patients in the non-advanced AMD stages to slow down disease progression. Treatment for the advanced stages of AMD is only available for the CNV type in the form of anti-VEGF treatment. However, this treatment is uncomfortable and expensive with no guarantee of a positive outcome for the patient. Other treatment options for wet AMD include photodynamic therapy and laser surgery, in which the newly formed blood vessels are blocked or destroyed. However, damage done to the retina is irreversible and therefore the best option is to identify AMD in the early stages and try to prevent progression to advanced AMD.

1.5 Screening for retinal diseases

Patients with AMD or DR do often not experience any visual complaints until the late stages of the disease, where irreversible damage to the retina has already occurred. To identify patients at early stages, screening programs have been established. Screening programs for DR are widely implemented and are proven to be effective. Nearly all patients with diabetes will develop some form of DR. Visual complications due to DR can be prevented in up to 80% of all the cases when it is timely detected and adequate action is taken. Therefore, all patients diagnosed with diabetes are advised to have a yearly retinal eye examination to check for signs of DR. Due to recent developments in treatment options for AMD, screening for the disease has received interest, and AMD screening programs are starting to be implemented. International authorities such as the World Health Organization (WHO) and the International Agency for the Prevention of Blindness (IAPB) recommend screening to detect patients in early AMD.
stages\textsuperscript{61}. However, to screen for AMD in a cost-effective manner and to deal with the enormous screening population, usage of automated software solutions is needed\textsuperscript{59,62}.

### 1.5.1 Screening protocols

Screening programs make use of CF imaging as this modality is cheap and widely available and enough detail is visible to assess presence of DR and AMD. During a screening examination, one or more CF images of both patient’s eyes are acquired and examined for abnormalities\textsuperscript{63,64}. Figure 1.7 shows a standard color fundus camera used for the acquisition of CF images. Images can be acquired with or without mydriasis (administration of pupil dilating eye drops), depending on the specifications of the camera\textsuperscript{64}. Acquisition is done at special screening centers or hospitals, but can also be done at a local optic shop or general practitioner’s center, if a fundus camera is available. Acquiring CF images using a camera does not require medical expertise. However, reading the acquired CF images requires specially trained personnel and heavily relies on the experience of the individual reader\textsuperscript{65,66}. If abnormalities are found in the images, the patient is referred to an ophthalmologist for a more detailed diagnosis and treatment planning. If no abnormalities are found, the patient is sent home and is advised to come back after one year for the next examination.

Source: \url{http://www.truetex.com/cr6-45nm_2.jpg}

Figure 1.7: Fundus camera used to acquire digital color fundus images of the retina.
1.5.2 Challenges and solutions

All people with diabetes and persons above the age of 50 years are advised to participate in screening for DR and AMD\textsuperscript{56,61}. One of the challenges in retinal screening programs is to ensure sufficient availability of trained personnel to assess all CF images. Assessment of the screening examinations is time-consuming and puts an enormous load on available medical expertise. Additionally, the amount of screening examinations that needs to be analyzed will increase rapidly in the near future due to the increasing prevalence of diabetes and the aging population\textsuperscript{12,24}. Training more people to cope with the large amounts of data is a solution, but this will increase healthcare costs. There is a need for a solution to deal with the massive screening population without hampering the healthcare system.

A solution comes in the form of automated systems\textsuperscript{13,67}. These computer aided detection (CAD) systems can help to maintain screening recommendations by providing a fully automatic analysis of the screening examination images without the need for specially trained personnel. CAD systems use the same input images currently acquired during the screening examination. Image analysis techniques are applied to these images and a system output is generated\textsuperscript{68}. The output of a CAD system depends on the task at hand and can vary from detection of individual lesions or retinal structures to, in the case of retinal screening, generating a decision if a patient needs to be sent for referral or not.

1.6 Computer aided detection

CAD refers to technology which assists doctors in the interpretation of medical images. In recent years, CAD systems have been applied successfully to many medical image analysis tasks such as lung cancer screening\textsuperscript{69–71}, breast cancer screening\textsuperscript{72,73}, histopathology image analysis\textsuperscript{74}, detection of colorectal cancer\textsuperscript{75,76} and screening for tuberculosis\textsuperscript{77,78}. Although most of these systems are intended to be used as an aid for doctors, nowadays these CAD systems are being used independently for a number of tasks including detection of tuberculosis on chest X-rays\textsuperscript{79}, bone age assessment on X-rays\textsuperscript{80} and breast density estimation on mammograms\textsuperscript{81}. Since the introduction of digital imaging systems in the 1990s, CAD systems to analyze retinal images have been developed\textsuperscript{82–84}. The goal of these CAD systems is to aid in the detection of retinal abnormalities and potentially substitute the retinal specialist for dedicated interpretation tasks. One of these tasks is screening for patients with signs of retinal disease such as AMD and DR. In this thesis, CAD systems for the automatic detection of abnormalities related to AMD and DR have been developed and described. These systems are intended to assist in blindness reduction by providing automated tools to analyze
retinal screening examinations. In the following sections, more details are given on the components of a CAD system.

1.6.1 Preprocessing

Preprocessing of input images is performed by CAD systems to improve image quality and to reduce variations between data sources. These variations occur as a result of the use of different acquisition protocols at the imaging locations. Different fundus cameras are used at the imaging locations and camera settings such as focus, illumination, resolution and angle, i.e. the degree of the field of view, might vary. Another reason for variations are the intrinsic properties of the human retina: race influences the pigment and color of the retina and due to the three-dimensional concave shape of the retina, illumination across the retina is different. Humans are good in dealing with these variations as the brain is able to correct for them. However, CAD systems trained with images from one source are not good at dealing with data from other sources. Most commonly used preprocessing techniques which we have used extensively in our works include histogram based approaches, color remapping and contrast enhancement.

Preprocessing can also involve spatial registration of images. This includes registration of images of the same eye which are centered on different anatomical structures, follow-up images of the same patient and images taken with different modalities. We have used registration for these tasks in Chapter 3 and Chapter 4. The retinal vasculature is a well recognizable structure in retinal images and most automatic systems are based on correctly matching this structure. Other systems made use of mutual information for the registration of multi-modal retinal images.

1.6.2 Segmentation

Segmentation of relevant structures is important to reduce classification errors and grade the extent of the disease. To assess an image, humans rely on prior knowledge such as spatial relations between anatomical structures. Similarly, CAD systems use this information to make an assessment of the image. To limit the region of interest, algorithms based on circular Hough transform or thresholding are used to segment the field of view. Segmentation of anatomical structures such as retinal vessels, the optic disk and the fovea is accomplished by using template matching, matched filters, pixel classification and morphological operations. Segmentation of retinal abnormalities such as drusen, microaneurysms and exudates is performed to make an assessment of disease severity. In Chapter 3 and Chapter 4, we suppress the retinal vasculature by using a previously developed vessel segmentation technique.
algorithm\textsuperscript{103}. In Chapter 2 and Chapter 5, we use the optic disk and fovea location as internal reference position during analysis.

### 1.6.3 Features

Numerical values, or so-called features, are used to discriminate between two or more classes by CAD systems. Features can encode information at different levels of abstraction, ranging from pixel level to patient level. Pixel level features include filtering operations, such as Gaussian, Gabor and wavelets; or derivative operations, such as Canny, Sobel or Laplacian\textsuperscript{109}. Pixel level features are used to identify pixels which are part of retinal structures or lesions. Region or lesion features can be subdivided into texture, morphology, context, color or intensity features. Texture features include higher order moments, local histogram and local binary patterns\textsuperscript{109}. In Chapter 3 we used higher order moment features to identify regions affected by AMD. Morphology and context features encode the shape of objects and the relation of them with respect to other retinal structures or objects. Color and intensity features provide information about lesion characteristics and substance. In Chapters 2 and 5 we used these types of features to identify lesions associated with AMD and DR. At the highest levels of abstraction, i.e. image- and patient-level, features encode information of the full image or of multiple images. These types of features are used to make an assessment on the presence of a disease or the severity of a disease. In Chapter 2 we computed image level features based on individual drusen detections and in Chapter 5 we computed image level features based on contextual information.

### 1.6.4 Classification

Classification of regions or images using the extracted features can be performed using statistical models or statistical classifiers\textsuperscript{68}. To construct such a model, features extracted from training examples are provided to the model in a training phase. In case of supervised classification, these training examples are manually labeled with a reference label. In unsupervised classification, these training examples do not come with a reference label. The classifier creates a decision boundary in the feature space to separate the different classes by minimizing a loss function or optimizing a criterion. The use of a classifier highly depends on the data and it is difficult to determine the best classifier beforehand. Several classifiers have been proposed in literature\textsuperscript{110}. In this thesis, we have examined different classifiers such as $k$-Nearest Neighbor\textsuperscript{111}, Linear Discriminant Analysis\textsuperscript{110}, Random Forest\textsuperscript{112}, Support Vector Machine\textsuperscript{113} and Gentle-Boost\textsuperscript{114} and we chose the best performing classifier for each problem at hand.
After training has finished, an unknown case can be classified by the classifier. The classifier either assigns the label of the predicted class or computes a probability for each class. These probabilities are calculated based on, for example, the distance to the decision boundary in the feature space. In this thesis, we have mainly focused on two-class classification problems, i.e. classifying candidate pixels and regions to be (part of) true lesions and classifying images as being normal or abnormal. Pixel classification scores can also be used to form candidate regions, which are consecutively used in region based analysis. In Chapter 2 and Chapter 5, we used the detected regions as input in a second classification round. In Chapter 3, we used a rule-based approach in which we took a percentile of the pixel probabilities to compute an image level score for being normal or affected. Region based classification scores can be used to compute an image level score by using another classification round or rule-based approach.

1.6.5 Convolutional neural networks

Deep learning is a technique in machine learning which has received great interest in the last few years. In deep learning, multiple hierarchical layers of individual model architectures are stacked to perform feature learning and classification simultaneously. The underlying model architectures include sparse auto-encoders, restricted Boltzmann machines and neural networks. A sub-class of neural networks called convolutional neural networks (CNNs), already introduced in the 1980s, has been used extensively in the past few years as the computational power needed to train these types of models has risen exponentially. A CNN is a deep learning architecture in which convolutional operations are applied to each layer’s input. In this way, inspired by how the visual cortex of the human brain works, a representation of a pattern or object is generated which can be used for classification. CNNs are similar to neural networks but are easier to train as they have fewer parameters because of their convolutional kernels. These convolutional kernels are learned from the data during the training phase of the CNN, therefore being tailored to the problem at hand and omitting the need of manual defined features. In Chapter 4 we applied this technique to the problem of identification of RPD and compared the results to the ones obtained in Chapter 3. In Chapter 6 we improved the training scheme of CNNs and applied CNNs to the detection of hemorrhages.

1.6.6 Evaluation procedures

To validate CAD systems and see how well the CAD systems perform as compared to human experts, several evaluation procedures have been proposed in literature. In these evaluation procedures a measure is computed on an independent test data set,
which has not been used to train or design the CAD system. This test data set is provided with a reference label for each of the cases. In this thesis, we have used an independent human grader or a consensus of human graders as reference to which we compare our CAD performance. In the following subsections, we will elaborate more on the evaluation procedures we have used in this thesis.

**Receiver Operating Characteristics analysis**

A case in a two-class classification problem can be either normal (where a reference score of 0 is assigned) or abnormal (where a reference score of 1 is assigned) according to the reference. As a CAD system typically has a probabilistic output for a case, first the CAD scores are thresholded and cases with scores below the threshold are assigned a 0, and the ones with scores above the threshold are assigned a 1. A case is a true positive (TP) if both the reference score and the CAD score are 1. If both these scores are 0, the case is a true negative (TN). If there is disagreement between the CAD score and the reference score, the case is either a false positive (FP) if the reference score is 0 and the CAD score is 1; or a false negative (FN) if this is vice versa. Performance measures are often based on a combination of these four numbers.

Sensitivity is calculated as the number of TPs divided by the sum of the number of TPs and FNs, see Equation 1.1. This metric indicates the percentage of correctly identified abnormal cases by the CAD system. Specificity, calculated as the number of TNs divided by the sum of the number of TNs and FPs, is used as metric to indicate the percentage of correctly identified normal cases, see Equation 1.2.

\[
\text{Sensitivity} = \frac{\text{TPs}}{\text{TPs} + \text{FNs}} \tag{1.1}
\]

\[
\text{Specificity} = \frac{\text{TNs}}{\text{TNs} + \text{FPs}} \tag{1.2}
\]

Ideally, both these measures should be high, indicating high performance of the CAD system. However, in a real world scenario, a trade-off between these metrics needs to be made. This can be accomplished by using Receiver Operating Characteristics (ROC) analysis\textsuperscript{120,121}. ROC analysis is mostly employed for image or patient based analysis, but can also be used for pixel based analysis. In ROC analysis, a wide variety of thresholds to create binary CAD scores are evaluated and sensitivity/specificity pairs at each threshold are calculated. These pairs are then plotted in a graph, constructing a so-called ROC curve. See Figure 1.8 for an example of an ROC curve. The diagonal line in this graph is the performance obtained if the outcome for each case is randomly guessed. The closer the curve gets to the upper left corner, i.e. a sensitivity of 1 and a specificity of 1, the better the performance of the CAD system. Consequently, the area
(Az) under the ROC curve is a good indicator of the performance and is commonly used. As human observers typically read a case as being either normal or abnormal, only a single point in the ROC graph is computed for observers and used to compare observer performance with the CAD performance.

Free-response Receiver Operating Characteristic analysis

As an image can have multiple lesions or affected regions, an image based ROC analysis is not appropriate to measure CAD performance for individual lesion detection. To evaluate system performance for the identification and localization of individual lesions or regions, Free-response Receiver Operating Characteristic (FROC) analysis can be employed\textsuperscript{120,122}. In FROC analysis, the threshold for assigning binary scores to the CAD system output is varied. At each threshold, the lesion based sensitivity and the number of FPs per image (or patient) are computed. The obtained measures can be visualized in an FROC curve which reflects the CAD system’s performance. Figure 1.9 shows an example of an FROC curve. A better CAD system will have a higher sensitivity at similar number of false positive detections. Choosing an operating point from the FROC curve is problem dependent and is mostly done based on the number of acceptable FPs.
at a reasonable sensitivity. In Chapter 2 and Chapter 6 we have employed FROC analysis to measure system performance for the detection of drusen and hemorrhages.

Statistical analysis

To assess whether a CAD system is significantly better than another CAD system or diagnostic test, the performance difference needs to be evaluated for significance. Confidence intervals for evaluation metrics can be computed. In this thesis, we have used bootstrap analysis to compute 95% confidence intervals for ROC curves, and consequently for the Az values\(^{123}\). In bootstrap analysis, multiple data subsets are constructed using random sampling with replacement, where cases are randomly drawn from the total test set until the same number of cases as in the original test set is reached. During this process, it is possible that one case is drawn multiple times while another case might not be drawn at all. In each subset, or so-called bootstrap, ROC analysis is performed and a 95% confidence interval can be calculated from these individual ROC analyses. Two systems or diagnostic tests can be compared for significance by comparing the Az values of each of the individual bootstraps of both systems or tests. \(p\)-values are defined as the number of times that the Az value of one system is higher than the other divided by the total number of bootstraps. If this \(p\)-value is lower than 0.05, the CAD system is statistically considered significantly different from the other system or diagnostic test.

Other tests which we have performed in this thesis include kappa (\(\kappa\)) agreement and intraclass correlation (ICC) analysis\(^{124,125}\). \(\kappa\) agreement is defined based on the
difference between the actual agreement compared to the expected agreement which can occur by chance and is computed as follows:

\[ \kappa = \frac{p_o - p_e}{1 - p_e} \]  

(1.3)

where \( p_o \) is the relative observed agreement among raters and \( p_e \) is the hypothetical probability of agreement by chance. As this measure takes into account the agreement by chance, using this measure is generally thought to be more robust than using only sensitivity and specificity\(^{124}\). ICC is a descriptive statistic that can be used when quantitative measures are made by two or more raters\(^{125}\). It describes how strongly ratings of cases in the same population resemble each other. In Chapter 2, Chapter 3 and Chapter 4, we used these metrics to compute the agreement between our proposed CAD system and human graders.

1.7 A review on CAD systems for automatic retinal image analysis

1.7.1 Automatic image quality analysis

Performance of CAD systems is influenced by the quality of the input data. Automatic analysis of image quality to ensure good quality images is therefore important for a CAD system to be operational in practice. As mentioned before, preprocessing steps to enhance image quality can be performed. However, this might not always be sufficient; when image quality is too low it is not possible to perform accurate automatic analysis. Previous works have described methods to automatically assess the quality of the images\(^{126-129}\). These works focused on contrast and clarity of the retina based on appearance of anatomical structures, i.e. vessels, to make a differentiation between good and bad quality images. If the image is rated as being of insufficient quality, it is recommended to discard the image for further analysis or acquire a new image if possible.

1.7.2 Automatic detection of anatomical structures in the retina

Automatic localization and segmentation of anatomical structures allow removal of false positive lesion detections and can help in the assessment of disease severity by providing information on location of lesions with respect to these anatomical structures. The methods proposed in literature are based on a wide variety of image analysis techniques. Studies to segment blood vessels have used various techniques, including matched filters\(^{102,130,131}\), mathematical morphology\(^{132-134}\), and supervised classification
methods\textsuperscript{103,135,136}. Figure 1.10 shows an example of the output of such a supervised classification system which has been developed previously in our group\textsuperscript{103}. Detection of the optic disc was performed using template matching\textsuperscript{101,137}, intensity analysis\textsuperscript{138,139}, shape analysis\textsuperscript{97}, supervised classification\textsuperscript{100} and geometrical relations\textsuperscript{140,141}. The fovea has been localized using template matching\textsuperscript{141,142}, geometrical relations\textsuperscript{140,143} and deformable models\textsuperscript{97}.

Figure 1.10: (a) and (c): Color fundus image. (b) and (d): Blood vessels as segmented by a supervised classification method\textsuperscript{103}. 


1.7.3 Automatic identification and quantification of retinal lesions

Identification of individual retinal lesions is important to make an assessment of diseases present in the retina and to make an assessment of disease severity. Previous works have proposed methods for the automatic identification and quantification of lesions related to AMD and DR.

In AMD, the hallmark lesions are drusen. Proposed methods for drusen identification include histogram based methods\textsuperscript{106,144}. These methods tried to improve the contrast between drusen and the background in order to obtain a segmentation of each individual druse. Other works have focused on local texture differences between drusen and the background\textsuperscript{145–150}. Some studies model individual drusen to identify these lesions\textsuperscript{105,151} or use supervised classification methods to detect and quantify drusen\textsuperscript{152}. In Chapter 2, we have developed an automatic system to perform automatic detection and quantification of drusen based on supervised classification. Figure 1.11 shows an example of the output of this system. No studies have been previously described in

![Figure 1.11: (a) and (c): Color fundus image. (b) and (d): Automatic drusen identification by the automatic system as described in Chapter 2.](image-url)
literature on the automatic detection of RPD. Chapter 3 and Chapter 4 describe the first automatic systems performing this task. Detecting advanced signs of AMD such as geographic atrophy and neovascularization has been addressed in previous works as well and were based on supervised classification using statistical classifiers\textsuperscript{153} and image segmentation using clustering\textsuperscript{154}.

![Figure 1.12: (a): Color fundus image. Bottom: example of the output of a supervised classification system for the automatic detection of bright lesions (b) and microaneurysms (c)\textsuperscript{155}.](image)

In DR, red appearing lesions, i.e. microaneurysms and hemorrhages; and bright appearing lesions, i.e. hard exudates and cotton wool spots, are the first visible signs. Previous works to automatically detect red appearing lesions include approaches based on filtering\textsuperscript{156}, morphology\textsuperscript{157} and supervised classification\textsuperscript{107,155,158}. Works on hard
exudate and cotton wool spot detection employ neural networks\textsuperscript{150} and supervised classification\textsuperscript{107,155,160–164}. Figure 1.12 shows the output of a supervised classification system for the detection of bright- and red lesions which has previously been developed in our group\textsuperscript{155}. Automatic detection of neovascularization has been addressed in literature as well\textsuperscript{165–168} and is mainly based on analysis of retinal vessel segmentations.

1.7.4 Automatic identification and quantification of retinal diseases

The goal of automated screening systems is to identify patients with early signs of retinal diseases. Most developed CAD systems make use of outputs of CAD systems for individual lesion detection and segmentation to make an overall assessment of an image or a patient\textsuperscript{13,82–84}. These systems typically assign a score indicating the likelihood whether retinal diseases such as AMD or DR are present. Apart from only detecting presence of a disease, severity grading of the abnormality is also addressed in some of these works. The majority of previous works were based on analysis of CF images, although methods using other modalities such as OCT and FAF have been developed as well\textsuperscript{169–171}. As this thesis is mainly based on analysis of color fundus images, we will only briefly describe the works using color fundus imaging.

\textit{Diabetic retinopathy}

Software solutions to identify patients with early signs of DR have been developed since the introduction of digital retinal imaging. A system to perform automatic identification of presence of DR was proposed and evaluated in large population studies in Scotland\textsuperscript{172–174}. The study consisted of images of over 33,500 patients in which 6.6% had referable DR. The CAD system was capable of identifying 97.8% of all patients in need for a referral at a specificity rate of 41.1%. Another system, previously developed in our group, was based on the fusion of information obtained from individual lesion detectors operating on multiple images from the same patient\textsuperscript{175–177}. This system was evaluated in multiple studies including a large number of patients. The systems achieved sensitivities of 92.9% and 96.8% with specificities of 60% and 59.4% respectively. Another approach, using a combination of detected red and bright appearing lesions was developed and evaluated in the UK\textsuperscript{178–181}. The system achieved sensitivities ranging between 93.1% and 97% at specificities between 71.4% and 78.0%.

Other types of CAD systems which do not require individual lesion detections but generate a decision based analysis of the whole image have been proposed as well. Systems based on image content were developed for the identification of images with signs of DR\textsuperscript{182,183}. These systems compared query images with a large database of labeled cases in terms of pictorial content to make an assessment of presence of diseases.
However, no large scale evaluation study using these system has been performed. A study with similar methodology showed the potential of this technique in a larger study dataset of over 1,000 images and achieved a sensitivity of 90% at 85% specificity\textsuperscript{158,184}.

Besides automatic DR identification systems, CAD systems which are capable of categorizing images into one of the DR severity stages have been proposed in literature\textsuperscript{185,186}. These systems were based on similar image analysis techniques by identifying individual lesions to make a multi-class classification into one of the DR disease stages, achieving an accuracy of 93% in a set of 130 images.

\textit{Age-related macular degeneration}

Automatic systems for the identification and severity classification of AMD have been developed to a lesser extent than systems for DR. Interest for CAD systems for AMD detection has only started recently with the introduction of treatment options for AMD. Existing systems for AMD identification were mainly based on the quantification of drusen\textsuperscript{145,148,150,187}. Although many of these studies used only a small number of patients, a study using almost 400 cases reported a sensitivity of 94% at a specificity 50%\textsuperscript{150}.

Other systems have bypassed drusen detection and directly predicted presence of AMD using pictorial content and image based features\textsuperscript{188–190}. A sensitivity of 98.6% at a specificity of 96.3% was achieved when classifying no AMD and early AMD versus intermediate and advanced AMD in a large set of 2145 color fundus images\textsuperscript{188}. Another study using 540 cases developed a method based on statistical moments, wavelet energy and image entropy and achieved a sensitivity and specificity of 91.1% and 96.3% for the identification of AMD\textsuperscript{190}.

Although severity classification of AMD is important, this aspect has not been addressed in most previous works. The closest related work we identified focused on differentiating each individual AMD category from the other categories\textsuperscript{188}.

\section*{1.8 Thesis outline}

In this thesis, we describe and evaluate CAD systems for the analysis of retinal images. The goal of these systems is to be used in eye screening programs for the early detection of retinal diseases. In our research group, we have previously developed CAD systems for the automatic detection of DR. In this thesis, we continued the work on several aspects of these systems and we have developed additional CAD systems for the identification of AMD signs.

In Chapter 2, a system based on supervised classification to automatically detect and quantify drusen on color fundus images is presented and compared with human
experts. A statistical classifier is used in combination with features derived from the detected drusen to separate patients with high risk to advance to the advanced stage of AMD from patients with low risk.

In Chapter 3, we used multiple retinal imaging modalities to identify RPD, lesions associated with a high risk to advance to the late stages of AMD. Besides evaluating human performance using single-modality grading and multi-modality grading, an automatic CAD system capable of identifying and quantifying RPD using multi-modal information is described and evaluated.

Chapter 4 continued the work on automatic identification and quantification of RPD. In contrast to the system described in Chapter 3, the system described in this chapter is based on CNNs and does not require manual crafted features to detect and quantify RPD. This self-learning system is able to learn an optimal set of filters from multi-modal imaging data to perform this task.

Chapter 5 described and evaluated an automatic CAD system for the detection of images containing either AMD or DR. It combined information from the CAD system described in Chapter 2 and a previously developed CAD system for DR detection to make an assessment of presence of abnormalities and to make a differentiation of the underlying disease.

In Chapter 6, a method based on CNNs is used to automatically detect hemorrhages using a large dataset of labeled examples. In this approach, we improved the CNN training procedure by guiding the training procedure and showed improved performance over a conventionally trained CNN.

The last Chapter provides a general discussion followed by a summary.
Drusen quantification and risk assessment of Age-related Macular Degeneration

Mark J. J. P. van Grinsven, Yara T. E. Lechanteur, Johannes P. H. van de Ven, Bram van Ginneken, Carel B. Hoyng, Thomas Theelen, Clara I. Sánchez

*Original title:* Automatic Drusen Quantification and Risk Assessment of Age-Related Macular Degeneration on Color Fundus Images

*Published in:* Investigative Ophthalmology and Visual Science (2013);54(4):3019–3027
Abstract

Purpose: To evaluate a machine learning algorithm that allows for computer-aided diagnosis (CAD) of non-advanced age-related macular degeneration (AMD) by providing an accurate detection and quantification of drusen location, area, and size.

Methods: Color fundus photographs of 407 eyes without AMD or with early to moderate AMD were randomly selected from a large European multicenter database. A machine learning system was developed to automatically detect and quantify drusen on each image. Based on detected drusen, the CAD software provided a risk assessment to develop advanced AMD. Evaluation of the CAD system was performed using annotations made by two blinded human graders.

Results: Free-response receiver operating characteristics (FROC) analysis showed that the proposed system approaches the performance of human observers in detecting drusen. The estimated drusen area showed excellent agreement with both observers, with mean intraclass correlation coefficients (ICC) larger than 0.85. Maximum druse diameter agreement was lower, with a maximum ICC of 0.69, but comparable to the inter-observer agreement (ICC = 0.79). For automatic AMD risk assessment, the system achieved areas under the receiver operating characteristic (ROC) curve of 0.948 and 0.954, reaching similar performance as human observers.

Conclusions: A machine learning system capable of separating high-risk from low-risk patients with non-advanced AMD by providing accurate detection and quantification of drusen, was developed. The proposed method allows for quick and reliable diagnosis of AMD, opening the way for large dataset analysis within population studies and genotype-phenotype correlation analysis.
2.1 Introduction

Age-related macular degeneration (AMD) is the leading cause of irreversible vision loss in developed countries among individuals older than 50 years. AMD is a gradually progressive disease that evolves from early and intermediate stages, with no or subtle visual changes, to an advanced stage, where the loss of central vision can occur. Patients with intermediate AMD are at higher risk of developing advanced AMD and thus suffering from severe visual loss, and they should undergo routine- and self-monitoring for a timely diagnosis. Lifestyle changes such as cessation of smoking and prophylactic regimens like vitamin supplementation are recommended for patients at risk in order to slow down the progression of the disease.

Deposits of extracellular material localized between the inner collagenous layer of the Bruchs membrane and the basal lamina of the RPE, known as drusen, are considered the hallmark feature of AMD. Macular drusen are important in the context of AMD grading and certain drusen characteristics are associated with progressing toward end-stage AMD. On fundus photography they appear as yellowish-white spots and different drusen phenotypes can be distinguished. Hard drusen are defined as small (≤63 μm) nodular lesions with well-defined borders. Soft drusen, on the other hand, tend to be larger and are generally characterized by poorly demarcated boundaries.

Identification and classification of eyes with AMD are performed mainly using color fundus images by manually determining the size and extension of drusen. However, other imaging modalities, such as optical coherence tomography, are gaining traction as well. Human observer classification is time-consuming and prone to inter-observer variations. Aside from speed, objectivity and reproducibility, implementation of an automatic drusen detection and quantification system could prove useful in many ways. It may allow for a cost-efficient screening program for patients at risk and help to identify and classify AMD patients in large cohort studies. Additionally, accurate quantitative measurements can help in large clinical studies for the evaluation and progression of drusen area, for example, in clinical trials concerning new therapeutic strategies for dry AMD, and it could help in applying inclusion criteria for large-scale clinical studies and genotype-phenotype correlation analysis.

Previously proposed methods automatically assessed the presence of drusen on color fundus photographs. However, the presence of drusen alone is not directly correlated with the risk of progression to advanced AMD. Other works focused on the automatic quantification of drusen without identifying patients at high risk or the AMD stage. Here, we describe and evaluate a machine learning algorithm that automatically distinguishes between images from low-risk and those from...
high-risk AMD patients by providing an accurate quantification of drusen location, area and size.

2.2 Methods

Study dataset

A total of 407 images of eyes with non-advanced stages of AMD (i.e. stage 1, 2 and 3 according to the criteria shown in Table 2.1), with sufficient grading quality for human graders, was selected in consecutive fashion from the European Genetic Database (EUGENDA, http://www.eugenda.org), a large multicenter database for clinical and molecular analysis of AMD\textsuperscript{207,208}. For each subject, images of both eyes were eligible for inclusion, but we did not select multiple images of the same eye. Images with presence of reticular pseudodrusen were excluded from analyses. Number of drusen, age, or ethnicity was not taken into account for the selection of data. Written informed consent was obtained before enrolling patients in EUGENDA. The study was performed according to the tenets set forth in the Declaration of Helsinki, and Investigational Review Board approval was obtained.

Digital nonstereoscopic color fundus photographs were acquired with a Topcon TRC 501X model digital fundus camera at 50\(^\circ\) (Topcon Corp., Tokyo, Japan) or with a Canon CR-DGi model non-mydriatic retinal camera at 45\(^\circ\) (Canon, Inc., Tokyo, Japan), and pupil dilation was achieved with topical 1.0\% tropicamide and 2.5\% phenylephrine. All images were macula-centered. Image size varied from 1360x1024 to 3504x2336 pixels. Before analysis, images were resized in a preprocessing step to have a field of view with a standardized diameter of 630 pixels independently of the image resolution. The data were divided randomly into two sets: set A, consisting of 52 images, for the evaluation of automatic drusen quantification, and set B, consisting of 355 images, for the evaluation of automatic risk assessment. Images from the same patients were kept in the same set.

Observer annotations

Resampled images were displayed on an LCD monitor similar to those used in ophthalmology practice and with the ability to zoom and pan. All visible drusen were manually outlined in set A by both observers using a specifically developed annotation tool. Whether confluent drusen were annotated as separate drusen or as one large drusenoid patch was left to the judgment of the observers. Two trained graders (JPHV, designated Observer 1, and YTEL, designated Observer 2) manually performed a risk assessment to develop late-stage AMD in all images of sets A and B. No AMD and early AMD were defined as low-risk stages, and intermediate AMD was considered
Table 2.1: Criteria for grading AMD according to the CIRCL Grading Protocol.

<table>
<thead>
<tr>
<th>AMD stage</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: No AMD</td>
<td>No drusen or small, hard drusen only.</td>
</tr>
<tr>
<td>2: Early AMD</td>
<td>&gt;10 small (&lt;63µm), hard drusen + pigmentary changes or 1-15 intermediate (63-124µm) drusen.</td>
</tr>
<tr>
<td>3: Intermediate AMD</td>
<td>&gt;15 intermediate (63-124µm) drusen or any large (&gt;125µm) drusen or GA not in the central circle of the ETDRS grid.</td>
</tr>
<tr>
<td>4: Advanced AMD (GA)</td>
<td>Presence of central GA.</td>
</tr>
<tr>
<td>5: Advanced AMD (CNV)</td>
<td>Evidence of active or previous CNV lesion.</td>
</tr>
<tr>
<td>6: CNV without signs for AMD</td>
<td>Chosen if CNV is present but no drusen of any size are present within the Field 2.</td>
</tr>
<tr>
<td>7: Cannot grade</td>
<td>Image is regarded as not gradable.</td>
</tr>
</tbody>
</table>

AMD, age-related macular degeneration; CIRCL, Cologne Image Reading Center and Laboratory; CNV, choroidal neovascularization; ETDRS, Early Treatment Diabetic Retinopathy Study; GA, geographic atrophy.

Druse size is measured as the diameter of the smallest enclosing circle of the druse.

high-risk stage. AMD was defined according to the standard protocol of the Cologne Image Reading Center and Laboratory (CIRCL) (Table 2.1)\textsuperscript{208}, an AMD classification that was adopted specifically for EUGENDA based on different international staging systems\textsuperscript{30,33,198,199,209}. Both observers performed one session of AMD grading for each study eye in set B and performed one session of drusen annotation for each study eye in set A on a different day.

Machine learning algorithm

The proposed CAD system analyzed all color fundus images to automatically quantify the visible drusen and assigned each image a probability between 0 and 1, with 1 indicating a high risk of developing advanced AMD and 0 indicating low risk. To accomplish this, the system performed the following steps:

1. Druse candidate extraction: Each pixel in the image was assigned a probability that the pixel was part of a bright lesion structure, using a supervised pixel classification method. Supervised classification is a machine learning technique where manually labeled training examples are used to infer the classification rule\textsuperscript{164}. Neighboring pixels with similar probability, not located close to the automatically detected optic disc\textsuperscript{164}, were grouped into druse candidates.

2. Druse candidate segmentation: The boundary of each druse candidate was automatically delineated using intensity and contrast characteristics.

3. Druse candidate classification: Druse characteristics and a supervised lesion classification method were used to assign a probability to each segmented candidate
which indicated the likelihood of being a true druse, creating a so-called drusen probability map.

4. AMD risk assessment: Based on the drusen probability map, a supervised image classification method assigned each image a probability to be at high risk of developing advanced AMD.

Figure 2.1: Example of the outputs obtained in each step of the proposed CAD system. (a): Original color fundus image. (b): Each pixel was assigned a probability of being part of a bright structure after the druse candidate extraction step. A higher intensity indicates a higher probability. (c): The boundary of each druse candidate (shown overlaid on the original image) was delineated during the drusen candidate segmentation step. (d): Candidates were classified as true drusen in the drusen candidate classification step. The final detected drusen are shown overlaid on the original image. Brighter color represents a higher probability of being a true druse.

Figure 2.1 shows the steps of the CAD algorithm. The classification steps in the system were performed using statistical classifiers that could differentiate between different types of pixels, candidates, or images by using a training set of labeled examples and extracting numerical characteristics (features). Several supervised classifiers were tested for each step, and the classifiers that performed best were chosen, namely a k-nearest neighbor (kNN) classifier for step 1, a linear discriminant (LDA) classifier for druse candidate classification (step 3), and a random forest (RF) classifier for step 4. A more detailed description of the CAD system can be found in Appendix A.
an image, the CAD system provides two outputs: (1) detection of all visible drusen in
the image and quantification of the drusen area and maximum druse diameter; and (2)
a probability indicating the likelihood that the patient was at high risk of developing
advanced AMD based on the drusen probability map.

Data analysis

To evaluate the proposed CAD system, two types of analyses were performed: (1) eva-
uation of automatic drusen quantification; and (2) evaluation of automatic AMD risk
assessment. Due to the lack of a single gold standard, each evaluation was performed
twice, taking Observer 1 as reference standard and comparing the CAD results with
those obtained by Observer 2, and vice versa.

For drusen quantification, a five-fold cross-validation approach was performed to
train and test the CAD system by using data from set A. Cross-validation analysis
allows determination of the system performance in an unbiased manner\textsuperscript{212}. Using the
test folds, the lesion sensitivity (fraction of drusen marked in the reference standard
that were detected as drusen by the CAD system) and the number of false positives
per image were calculated after setting a threshold for the druse probabilities obtained
in step 3. Varying this threshold, different lesion sensitivity-false positives pairs were
calculated and summarized in a free receiver operating characteristic (FROC)\textsuperscript{120} to
evaluate the CAD performance on the detection of drusen. The observer performance
compared to the reference standard, which corresponds with one lesion sensitivity-false
positives pair, was also calculated and included in the obtained FROC curve.

Total drusen area and maximum druse diameter obtained by the CAD system were
calculated using the distance between the fovea and the border of the optic disc as a
reference distance of 3000 \( \mu \text{m} \textsuperscript{31,33} \). After thresholding the drusen probability map, total
drusen area and maximum druse diameter were measured and compared to the observers
opinions, using intraclass correlation coefficient (ICC) analysis\textsuperscript{125}. This threshold was
set at the same false-positives rate as Observer 1, as this observer had fewer false
positives than Observer 2. During the analysis, the performance of the CAD system
and the observers of drusen quantification were evaluated inside and outside the Early
Treatment Diabetic Retinopathy Study (ETDRS) grid\textsuperscript{17}, which was manually set before
the analysis.

For AMD risk assessment, a leave-one-out cross-validation approach\textsuperscript{212} was per-
formed to train and test the CAD system, using data from set B. The leave-one-out
cross-validation allows measurement of the predictive performance measure of a statis-
tical model by testing a single sample while training with the remaining samples. This
is repeated such that each sample is used once as test data. Using the test folds, image
sensitivity (fraction of images correctly classified by the CAD system in the high-risk
stage) and image specificity (fraction of images correctly classified by the CAD system in the low-risk stage) were calculated after setting a threshold for the estimated risk obtained in step 4. Varying this threshold, different image sensitivity-image specificity pairs were calculated and summarized in a receiver operating characteristic (ROC) to evaluate the CAD performance of distinguishing between low-risk and high-risk patients. The area (Az) under the ROC was used as a measure of performance. The observer performance compared to the reference standard, which corresponds with one image sensitivity-image specificity pair, was also calculated and included in the obtained ROC curve. Overall agreement on risk assessment between the observers was calculated using $\kappa$ statistics (version 17.0.0 software; SPSS, Chicago, IL).

2.3 Results

Of the 407 images, 145 were captured with the Topcon camera and 262 were captured with the Canon camera. Table 2.2 shows some statistics of the performed observer annotations for AMD risk assessment and drusen quantification.

Table 2.2: Summary of the manual annotations performed on Set A and B by the two observers.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Set A</th>
<th>Set B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs 1</td>
<td>Obs 2</td>
</tr>
<tr>
<td>Risk assessment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No AMD</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>Early AMD</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Intermediate AMD</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>Drusen quantification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average number of drusen</td>
<td>130.4 ± 178.1</td>
<td>198.5 ± 243.1</td>
</tr>
<tr>
<td>Average size of drusen, (\mu\text{m}^2)</td>
<td>5,873 ± 10,027</td>
<td>5,115 ± 8,257</td>
</tr>
</tbody>
</table>

Obs, observer; AMD, age-related macular degeneration.
No AMD and early AMD are defined as low-risk stages and intermediate AMD as high-risk stage.
Average number of drusen is the average number of annotated drusen per image. Average size of drusen is the average size of annotated drusen.

Drusen quantification

Figure 2.2 shows the automatically detected drusen by the CAD system from a sample image and shows the annotations of the observers. Figure 2.3a and Figure 2.3b show the FROC curves for the CAD system inside and outside the ETDRS grid using Observer 1 and Observer 2 as a reference standard, respectively.

Table 2.3 and Table 2.4 summarize the mean drusen area and maximum druse diameter obtained by the CAD system and the observers. The corresponding ICCs are shown in Figure 2.4a and Figure 2.4b. For estimated drusen area, ICCs of 0.91
2.3 Results

Figure 2.2: (a): Original color fundus image. (b): Drusen detected by the CAD system overlaid on the original image. (c): Drusen annotated by Observer 1 on the original image. (d): Drusen annotated by Observer 2 on the original image.

and 0.86 were obtained for the CAD system compared to Observer 1 and Observer 2, respectively, whereas the inter-observer agreement reached an ICC equal to 0.87. The CAD system showed similar agreement with the observers independently of the camera used for the acquisition, reaching ICC values of 0.80 and 0.88 on images acquired with the Topcon digital fundus camera at 50° and the Canon nonmydriatic retinal camera at 45°, respectively. For the estimation of maximum druse diameter, defined as the diameter of the smallest enclosing circle of a druse, the agreement with the observers was lower with a maximum ICC of 0.69, whereas observers had an agreement with ICC of 0.79.

**AMD risk assessment**

Figure 2.5a and Figure 2.5b show the ROC curves using Observer 1 and Observer 2 as reference standard, obtaining Az values of 0.948 and 0.954, respectively. Observer 2 reaches an image sensitivity of 0.85 and an image specificity of 0.96, as shown in Figure 2.5a, whereas Observer 1 obtained an image sensitivity of 0.84 and image specificity of 0.96 (Figure 2.5b). Table 2.5 shows the contingency table and kappa (κ)
Figure 2.3: FROC curves for the CAD system inside and outside the ETDRS grid, considering Observer 1 (a) and Observer 2 (b) as reference standards. The corresponding observer performance compared to the reference standard is also plotted as a point in the graph.

Table 2.3: Mean area and percentage covered by drusen inside and outside the ETDRS grid and in the total image.

<table>
<thead>
<tr>
<th>Coverage</th>
<th>CAD</th>
<th>Observer 1</th>
<th>Observer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inside grid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean area, mm²</td>
<td>0.43 ± 0.57</td>
<td>0.44 ± 0.68</td>
<td>0.56 ± 0.73</td>
</tr>
<tr>
<td>Area, %</td>
<td>1.52 ± 2.01</td>
<td>1.55 ± 2.40</td>
<td>1.98 ± 2.58</td>
</tr>
<tr>
<td>Outside grid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean area, mm²</td>
<td>0.35 ± 0.70</td>
<td>0.33 ± 0.72</td>
<td>0.46 ± 0.81</td>
</tr>
<tr>
<td>Area, %</td>
<td>0.36 ± 0.73</td>
<td>0.34 ± 0.75</td>
<td>0.49 ± 0.87</td>
</tr>
<tr>
<td>Total image</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean area, mm²</td>
<td>0.78 ± 1.00</td>
<td>0.77 ± 1.07</td>
<td>1.01 ± 1.21</td>
</tr>
<tr>
<td>Area, %</td>
<td>0.67 ± 0.86</td>
<td>0.65 ± 0.96</td>
<td>0.89 ± 1.16</td>
</tr>
</tbody>
</table>

Table 2.4: Mean maximum druse diameter (mm) and pixels inside and outside the ETDRS grid and in the total image.

<table>
<thead>
<tr>
<th>Coverage</th>
<th>CAD</th>
<th>Observer 1</th>
<th>Observer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inside grid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum diameter, pix</td>
<td>11.54 ± 6.74</td>
<td>13.00 ± 12.53</td>
<td>11.79 ± 10.06</td>
</tr>
<tr>
<td>Maximum diameter, mm</td>
<td>0.21 ± 0.12</td>
<td>0.23 ± 0.23</td>
<td>0.21 ± 0.17</td>
</tr>
<tr>
<td>Outside grid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum diameter, pix</td>
<td>10.16 ± 6.30</td>
<td>7.15 ± 7.22</td>
<td>8.91 ± 8.72</td>
</tr>
<tr>
<td>Maximum diameter, mm</td>
<td>0.18 ± 0.12</td>
<td>0.13 ± 0.14</td>
<td>0.16 ± 0.15</td>
</tr>
<tr>
<td>Total image</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum diameter, pix</td>
<td>13.88 ± 6.27</td>
<td>14.67 ± 12.33</td>
<td>13.74 ± 10.21</td>
</tr>
<tr>
<td>Maximum diameter, mm</td>
<td>0.25 ± 0.12</td>
<td>0.27 ± 0.23</td>
<td>0.25 ± 0.18</td>
</tr>
</tbody>
</table>

agreement between the observers and between the CAD system and the observers for
2.4 Discussion

In this study, a supervised machine learning algorithm for automated AMD classification based on drusen identification and quantification, was developed. Our system was able to perform equally as experienced human graders with respect to AMD risk assessment. The threshold for the CAD system was set at the cutoff point that maximizes sensitivity + specificity.

Figure 2.5: ROC curves for the CAD system considering Observer 1 (a) and Observer 2 (b) as reference standard, respectively. The corresponding observer performance compared to the reference standard is also plotted as a point in the graph.

2.4 Discussion

In this study, a supervised machine learning algorithm for automated AMD classification based on drusen identification and quantification, was developed. Our system was able to perform equally as experienced human graders with respect to AMD risk assessment. The threshold for the CAD system was set at the cutoff point that maximizes sensitivity + specificity.

Figure 2.5: ROC curves for the CAD system considering Observer 1 (a) and Observer 2 (b) as reference standard, respectively. The corresponding observer performance compared to the reference standard is also plotted as a point in the graph.
Table 2.5: Contingency table, $\kappa$ agreement and 95% CI for AMD risk assessment between Observer 1 and Observer 2.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Low risk</th>
<th>High risk</th>
<th>$\kappa$</th>
<th>95 % CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer 1 vs. Observer 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low risk</td>
<td>268</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High risk</td>
<td>12</td>
<td>64</td>
<td>0.807</td>
<td>0.731-0.833</td>
</tr>
<tr>
<td>CAD vs. Observer 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low risk</td>
<td>261</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High risk</td>
<td>19</td>
<td>65</td>
<td>0.765</td>
<td>0.684-0.846</td>
</tr>
<tr>
<td>CAD vs. Observer 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low risk</td>
<td>259</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High risk</td>
<td>20</td>
<td>66</td>
<td>0.760</td>
<td>0.679-0.841</td>
</tr>
</tbody>
</table>

CI, confidence interval.

assessment, drusen localization, and determination of mean drusen area with a dataset considerably larger than those used in previous publications\textsuperscript{105,106,146,148,149,206,213–217}.

Detecting drusen on color fundus images is a challenging task, as shown by the differences in observer annotations in Figure 2.2. These differences illustrate the need for a robust and accurate system for drusen detection. This would help in eliminating intra-and inter-observer variability and the subjective character of manual drusen detection. For this reason, automated drusen detection on color fundus photographs has been a field of interest for the last couple of decades. However, many systems still require human adjustments or close supervision by experts and are therefore still amenable to subjective input\textsuperscript{146,200,206,215–217}. Unsupervised automatic detection systems have been developed, but most have failed to achieve acceptable performances compared to human graders\textsuperscript{149,213,214} or are only able to give categorized outcome values\textsuperscript{148,152}.

In addition to detection, accurate localization and segmentation of drusen are very important to adequate quantification of drusen load in an image. In contrast to other methods\textsuperscript{105,106}, where the performance analysis was carried out by pixel-to-pixel comparison, we performed FROC analysis\textsuperscript{120,218}, stressing the importance of a correct localization and segmentation of individual lesions and providing higher statistical power than conventional ROC analysis for this task\textsuperscript{218}.

With respect to quantification of the total drusen area, there is high agreement between the observers and the proposed CAD system (Figure 2.4). The CAD system also showed similar agreement with the observers independently of the camera used
for the acquisition. However, a more exhaustive analysis should be made in order to evaluate the effect of the image quality of manual and automatic drusen quantification. In a previously proposed drusen quantification method\textsuperscript{105}, a slightly higher ICC value was reported (ICC = 0.92) than the ground truth based on the average grading of eight experts. However, in that study, images with the highest variability among observers were excluded from the study. This was the case for five images, resulting in the exclusion of more than 20% of the total dataset, which is likely to influence the outcome.

For the estimation of maximum druse diameter, agreement between the CAD system and the observers was lower, with a maximum ICC value of 0.69. However, the inter-observer agreement on this measurement also decreased. These lower values might be explained by the fact that a correct druse diameter depends on accurate druse delineation, which may be hampered by several factors. For human observers, the main problem lies in the analysis and classification of complex morphological patterns that characterize drusen\textsuperscript{106}, whereas the CAD system is impeded mostly by low image quality, poor contrast, or neighboring artifacts.

In this study, we also examined AMD risk assessment, and we showed that the CAD system performs as well as the human observers (Figure 2.4, Table 2.4). Images incorrectly classified by the system in the low-risk stage corresponded mainly to cases of disagreement between the observers (40% of misclassified images) or low quality images where the system was unable to localize low contrast drusen. Other studies have tried to identify AMD with automatic methods\textsuperscript{150,189}. However, these were aimed primarily at identifying the presence or absence of disease instead of trying to separate high-risk from low-risk AMD patients, which is clinically more relevant. Zheng et al.\textsuperscript{189} developed an algorithm for identification of AMD with a sensitivity of 99.4% and a specificity of 100%. However, the authors compared their CAD system only to a single human observer, which can lead to false high performance.

In contrast to previously published CAD systems\textsuperscript{105,105,106,146,149,200,206,214–217}, our software performs drusen quantification independently of a fixed region of interest. With full image drusen detection, more information from the image is extracted which can be beneficial if our method would be deployed in clinical studies of AMD. For example, in studies of the cuticular drusen subtype of AMD, diagnosis is based on a typical pattern of innumerable small drusen on fluorescein angiography (FA), not only in the macular region but also in the peripheral retina\textsuperscript{207,219}. It would be very valuable to evaluate this drusen pattern to see if regions identified on FA were also detected by the CAD system on color images.

In our system, misclassification of candidates as true drusen often occurred due to reflections of the internal limiting membrane or because of the presence of non AMD-related abnormalities. Adding better features to characterize these regions or including
them as samples in the learning process of the CAD system might solve this problem in the future. Depending on their number and size, false-positive drusen detection might lead to incorrect AMD risk assessment. However, in our study, this did not occur very often. It is possible that these false-positive drusen have a relatively low probability of being a true druse, which is accounted for during computation of AMD risk. In addition, we did not consider pigmented changes for automatic AMD risk assessment. This could be unfortunate if we wanted to separate patients with early AMD from healthy controls. For detection of patients at high risk of developing late-stage AMD, this distinction is not relevant because no AMD and early AMD were both considered low risk. However, if we wanted to use the system for classifying groups of patients in different AMD stages in new studies, the need for a well-defined control group is high. We will investigate automatic detection of pigmentedary changes in upcoming studies.

We are not aware of any implementation of AMD screening programs, but there have been studies evaluating cost effectiveness of such programs\(^{59,220}\). However, the proposed programs are based on self-testing, whereas screening based on evaluation of color fundus images would be preferable. Deployment of human graders in such a broad setting would be costly and time consuming, and implementation of an automatic detection system would circumvent these problems. The CAD system could, for example, be installed in opticians offices and be implemented in routine evaluation of elderly people. High-risk individuals would be selected on site and referred for further ophthalmologic evaluation.

In conclusion, we have developed and evaluated a machine learning system for identification of high-risk AMD patients. Our system allows for accurate detection and quantification of drusen location, area, and size, with a performance equal to human observers under stringent testing conditions. Implementation of our system allows for quick and reliable diagnosis of AMD in screening as well as in research programs. Additionally, there is a need for detailed phenotyping of large datasets in order to gain more insight into risk factors and disease mechanisms involved in AMD\(^{207}\). With the use of an automatic detection system, identification of homogeneous AMD subgroups and genotype-phenotype correlations should be achievable in a broader context\(^{204}\).

### 2.5 Appendix A

**Druse candidate extraction**

In this step, pixels that are potentially bright lesion pixels are extracted by convolving the green channel of the color fundus image with a group of Gaussian filters. These filters are based on Gaussian derivatives up to second order at different scales of information\(^{164}\). A kNN classifier is then trained to classify every pixel in the image on the
basis of the filter responses\textsuperscript{210}. No preprocessing of the image is needed previous to the druse candidate extraction step, such as suppression of luteal pigmentation\textsuperscript{146,200,206}. After classification, a pixel probability map is obtained that indicates the probability of each pixel to be part of a bright lesion. Neighboring pixels with similar probability were grouped into druse candidates. Algorithms that perform optic disc segmentation and vessel segmentation\textsuperscript{164} were also applied in order to remove candidates that overlapped with these anatomical landmarks and to use in further processing.

**Druse candidate segmentation**

In order to find the border of the drusen candidates, dynamic programming\textsuperscript{221} is applied around the local maxima of the calculated pixel probability map. During this process, the gradient magnitude of the Gaussian derivatives is used as cost function to guide the algorithm to the candidate borders.

**Druse candidate classification**

In order to determine whether a druse candidate is a true druse or not, a classification step using a linear discriminant classifier (LDA) is performed\textsuperscript{210}. For each druse candidate, a total of 109 features based on color, intensity, contextual information and shape are extracted (Table 2.6)\textsuperscript{164}. These features exploit the different characteristics that the drusen show in color fundus images.

Table 2.6: Features for druse candidate classification.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Number</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape</td>
<td>1-5</td>
<td>Area, perimeter, compactness, length and width of the candidate.</td>
</tr>
<tr>
<td>Context</td>
<td>6,7</td>
<td>Average and standard deviation of vessel pixel probability at the candidate border.</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Distance to the closest candidate.</td>
</tr>
<tr>
<td></td>
<td>9,10</td>
<td>Number and average pixel probability of neighboring candidates in a radius of 50 pixels.</td>
</tr>
<tr>
<td>Intensity</td>
<td>11-33</td>
<td>Features measuring the contrast of the candidate in the RGB channels.</td>
</tr>
<tr>
<td></td>
<td>33-81</td>
<td>Mean and standard deviation of Gaussian filter bank outputs.</td>
</tr>
<tr>
<td>Color</td>
<td>82-105</td>
<td>Average and standard deviation inside and outside the candidate using the planes of the Luv color space and HSI color space.</td>
</tr>
<tr>
<td>Misc.</td>
<td>106-109</td>
<td>Average, standard deviation, maximum and median pixel probability inside the candidate.</td>
</tr>
</tbody>
</table>

HSI, hue-saturation-intensity; Luv, luminescence-saturation-hue angle color space adopted by the International Commission on Illumination (CIE); RGB, red-green-blue.
AMD Risk Assessment

To separate high risk from low risk patients, a weighted histogram of the calculated drusen probabilities in the image is created to encode the drusen extension and size. The value $h_n$ of the histogram bin $n$ is defined as:

$$h_n = \sum_{i \in L_n} p_i$$  \hspace{1cm} (2.1)

where $p_i$ is the posterior probability of druse candidate $i$ and $L_n$ is the group of candidates whose size is $\tau n \leq d_i < \tau(n + 1)$, with $d_i$ the size (µm) of candidate $i$, $\tau$ and $n$ control the bin size and the histogram resolution, respectively and were chosen as $n = 0, ..., 36$ and $\tau = 10\mu m$. The last bin ($n = 36$) takes all the candidates with sizes $d_i$ larger than $360 \mu m$ into account. A random forest (RF) classifier$^{211}$ is then trained using the histogram bins as features to distinguish high risk patients.
Automatic identification of reticular pseudodrusen

Mark J. J. P. van Grinsven, Gabriëlle H. S. Buitendijk, Corina Brussee, Bram van Ginneken, Carel B. Hoyng, Thomas Theelen, Caroline C. W. Klaver, Clara I. Sánchez

Original title: Automatic Identification of Reticular Pseudodrusen Using Multimodal Retinal Image Analysis

Abstract

Purpose: To examine human performance and agreement on reticular pseudodrusen (RPD) detection and quantification by using single- and multi-modality grading protocols and to describe and evaluate a machine learning system for the automatic detection and quantification of reticular pseudodrusen by using single- and multi-modality information.

Methods: Color fundus, fundus autofluorescence, and near-infrared images of 278 eyes from 230 patients with or without presence of RPD were used in this study. All eyes were scored for presence of RPD during single- and multi-modality setups by two experienced observers and a developed machine learning system. Furthermore, automatic quantification of RPD area was performed by the proposed system and compared with human delineations.

Results: Observers obtained a higher performance and better inter-observer agreement for RPD detection with multi-modality grading, achieving areas under the receiver operating characteristic (ROC) curve of 0.940 and 0.958, and a $\kappa$ agreement of 0.911. The proposed automatic system achieved an area under the ROC of 0.941 with a multi-modality setup. Automatic RPD quantification resulted in an intraclass correlation (ICC) value of 0.704, which was comparable with ICC values obtained between single-modality manual delineations.

Conclusions: Observer performance and agreement for RPD identification improved significantly by using a multi-modality grading approach. The developed automatic system showed similar performance as observers, and automatic RPD area quantification was in concordance with manual delineations. The proposed automatic system allows for a fast and accurate identification and quantification of RPD, opening the way for efficient quantitative imaging biomarkers in large data set analysis.
3.1 Introduction

Age-related macular degeneration (AMD) is a progressive eye disease affecting mainly the elderly and causing vision loss at advanced stages. The early stages of AMD are characterized by the presence of pigmentary changes and drusen, which are deposits accumulating between the retinal pigment epithelium (RPE) and the Bruchs membrane. A newly appreciated extracellular lesion in AMD, commonly termed “subretinal drusenoid deposits” or “reticular pseudodrusen” (RPD), presents different characteristics and distribution than normal drusen and is a strong risk factor for progression to advanced AMD. Therefore, its identification and quantification is of paramount importance for a better understanding of disease progression.

Reticular pseudodrusen are visible on color fundus (CF) photography, fundus autofluorescence (FAF) imaging, and near infra-red (NIR) imaging among other retinal imaging modalities such as confocal blue reflectance, indocyanine green angiography, spectral-domain optical coherence tomography (SDOCT), and fluorescein angiography. On CF images, RPD are described as indistinct, yellowish interlacing networks with a width of 125 to 250 \( \mu \text{m} \). On FAF images, RPD are characterized as hypofluorescent lesions, while on NIR images, RPD are characterized as groups of hyporeflectant lesions against a mild hyperreflectant background. Previous studies have reported a difference in sensitivities for RPD detection among image techniques. However, RPD identification using a single-image modality is challenging, as the characteristic changes associated with RPD are often subtle and may not always be detected when using only one imaging technique. Therefore, for an accurate diagnosis, RPD detection should be performed with two or more image modalities.

Although other studies have investigated and compared the performance of individual image techniques for RPD detection, a study of the performance obtained by using multiple image modalities simultaneously has not been performed yet, to the best of our knowledge.

Despite its expected higher accuracy, grading of multi-modality images represents a considerable workload for a human grader. Machine learning algorithms have huge potential for dealing with complex information extracted from different image modalities. Furthermore, automatic systems are not influenced by fatigue and mindset and, therefore, are less prone to variability than humans. Previously developed systems for the automatic detection of drusen showed good performance on CF images. Whether they also perform well fusing information from different image modalities is currently unknown. To the best of our knowledge, there is no method for the automatic identification of RPD fusing information from different image modalities.
The aim of the present study was twofold. Firstly, we evaluated the performance and the agreement between human observers by using single- as well as multi-modality grading approaches for RPD detection. In the single-modality approach, RPD detection was performed by using only one image technique (namely, CF, FAF, or NIR). In contrast, during the multi-modality grading session, the observers evaluated the three available image modalities simultaneously. Secondly, we aimed to investigate the effectiveness of a novel machine learning algorithm for the automatic identification and quantification of RPD by using combined information from different image modalities and comparing its performance to that of human observers.

3.2 Methods

Study dataset

A set of subjects with and without RPD was selected from the Rotterdam Study, a prospective cohort study investigating risk factors for chronic diseases in the elderly. The study adhered to the tenets set forth in the Declaration of Helsinki, and Investigational Review Board approval was obtained. Only patients with CF, FAF, and NIR images available were included in this study. Color fundus images were taken by using a 35° field-of-view Topcon TRC 50EX fundus camera (Topcon Optical Company, Tokyo, Japan) with a Sony DXC-950P digital camera with a resolution of 768x576 pixels (Sony Electronics, Inc., New York, NY, USA). FAF and NIR images were taken with a Heidelberg Retina Angiograph 2 (Heidelberg Engineering, Heidelberg, Germany) with a field of view of 30° and a resolution of 768x768 pixels. In total 278 eyes of 230 patients aged 65 years and older were selected from the last examination round of the Rotterdam Study. All CF images were graded according to the Wisconsin Age-Related Maculopathy Grading and the International Classification and Grading System for Age-Related Maculopathy and Age-Related Macular Degeneration by local graders of the Rotterdam study, using visual assessment. These annotations constituted the reference standard for our study. We selected all the eyes for which RPD were identified in this round (N=72) from CF images. Status of RPD was also confirmed on FAF and NIR imaging. For positive and negative controls, we selected eyes that were graded by the local Rotterdam study graders as having soft distinct or soft indistinct drusen but without RPD (N=108) and eyes that did not contain any type of drusen (N=98), respectively. The positive and negative controls did not have any signs of RPD in the other modalities (FAF and NIR). As the database did not contain any information about the extent of RPD area, two human observers (G.H.S.B., C.B.) made RPD area delineations in consensus by using the three modalities simultaneously for the eyes con-
3.2 Methods

taining RPD. These delineations were used as reference standard for the quantification of RPD area.

Observer study: single- versus multi-modality grading

All images were evaluated independently by two human observers (G.H.S.B., C.B.) for evidence of RPD. Reticular pseudodrusen were defined as indistinct, yellowish interlacing networks with a width of 125 to 250 μm on CF images; groups of hyporeflectant lesions in regular patterns on FAF images; and groups of hyporeflectant lesions against a mildly hyperreflectant background in regular patterns on NIR images. Observer 1 had 4 years of reading experience for all three imaging modalities, whereas Observer 2 had 19 years of reading experience on CF imaging and 5 years on FAF and NIR imaging. The observers were asked to use a scoring system ranging from 0 to 1, indicating the likelihood of presence of RPD. Two different grading approaches were used: single- and multi-modality grading. During single-modality grading, the observers graded each image modality separately in a randomized order. Color fundus, FAF, and NIR images were pooled and shown randomly to the observers. Observers were also asked to indicate whether the image was of sufficient quality for grading. The quality of an image is deemed insufficient when it becomes difficult or impossible to make a confident assessment regarding the presence of RPD. During multi-modality grading, observers were asked to diagnose RPD after observing CF, FAF, and NIR images from the same eye simultaneously. The eyes were shown in randomized order in this grading session as well.

In a separate grading session, the observers manually delineated in consensus the area covered by RPD, based on one single modality, that is, single-modality RPD delineation on CF, FAF, or NIR images. Only the 72 eyes containing RPD as indicated by the reference were taken into account for the quantification of RPD area.

Automatic reticular pseudodrusen identification

The proposed machine learning algorithm simultaneously analyzed the available modalities from an eye examination to automatically identify reticular pseudodrusen areas. The algorithm assigned the complete eye examination a probability between 0 and 1, indicating the probability of presence of RPD and providing a quantification of the area covered by the lesions. To accomplish this, the algorithm performed three steps: preprocessing, feature extraction, and classification and quantification.
Preprocessing

In the preprocessing step, two different methods were applied to the images: image registration and vessel removal.

1. Registration will provide a geometric alignment across modalities to identify corresponding pixels that represent the same scene. This multi-modal image registration was performed by using a semiautomatic affine method, where the images are deformed to accurately match user-specified points or landmarks. In this study, three corresponding landmarks on prominent image locations, such as vessel bifurcations, were manually selected on each modality and used to perform the registration.

2. To reduce intensity variations due to presence of vessels, the retinal vasculature was removed from the images. The vasculature was automatically extracted by using a previously developed algorithm and used as input in an inpainting algorithm, which removes the vessels by interpolating intensities at the supplied image locations. Figure 3.1 shows an example CF image (Figure 3.1a), FAF image (Figure 3.1b), and NIR image (Figure 3.1c) of an eye and Figure 3.1d, Figure 3.1e and Figure 3.1f show their corresponding images after vessel removal, respectively.

Figure 3.1: Co-registered (a) CF, (b) FAF and (c) NIR images and their corresponding results after vessel removal (d-f).
3.2 Methods

**Feature extraction**

To perform an automatic analysis of the images, the machine learning algorithm uses information that is extracted from the images and encoded in numerical values or so-called features. To do so, each color channel of the CF image, as well as the FAF and NIR image, was separately convolved with a set of Gaussian filters. These filters are based on Gaussian derivatives up to second order at different scales and are invariant to rotation and translation. For each resulting filtered image, the mean, standard deviation, skewness, and kurtosis values in a circular neighborhood around each pixel were calculated. The corresponding features for each pixel were then obtained by concatenating these extracted values in a single feature vector.

**Classification and quantification**

To determine whether a pixel is part of an RPD area, a random forest classifier was used to obtain an automatic classification based on the calculated features. This classifier operates by constructing a multitude of decision boundaries (trees) to make a separation between multiple classes. After training, the random forest classifier provided a probability between 0 and 1 indicating the probability that the pixel belongs to an RPD area, based on labeled training examples and the input pixel feature vector. Figure 3.2 shows the images of the modalities of an example eye and the output of the classifier. Finally, an image score indicating the likelihood of the eye examination to contain RPD was assigned by taking the 99th percentile of the obtained probability map.

To quantify the area covered by RPD, a threshold was set on the probability map. This threshold was image based and experimentally determined as the 55th percentage of the maximum value of the probability map. Only the area inside the Early Treatment Diabetic Retinopathy Study (ETDRS) grading grid was taken into account for the quantification.

**Statistical analysis**

The performance of the observers and the proposed machine learning algorithm for the single- and multi-modality approaches was evaluated by measuring the area (Az) under the receiver operating characteristic (ROC) curve. Statistical comparisons were made by using bootstrap analysis with 5000 bootstraps. Bootstrap analysis is an nonparametric test that is commonly used to estimate the variance of ROC analysis. Results with a P value lower than 0.05 were seen as statistically significant. Bonferroni correction was applied to counteract the problem of multiple comparisons. For observers, statistics were also reported to assess inter-observer variability. As the
Figure 3.2: Example of the classification result obtained by the proposed machine learning algorithm. Given an eye exam consisted of (a) a CF image, (b) an FAF image and (c) an NIR image; the algorithm outputs (d) a probability map indicating the likelihood for each pixel to be part of a RPD area. Red values indicate higher probability to be RPD.

The proposed machine learning algorithm requires labeled example data for training, the evaluation was performed by using a patient-based leave-one-out strategy. Automatic quantification of RPD area was evaluated by calculating the percentage of detected RPD area inside the ETDRS grading grid and was compared with the observer delineations. The RPD area agreement with observers was measured by using intraclass correlation (ICC) statistics.

3.3 Results

Image quality assessment

Table 3.1 shows the image quality analysis of the observers for the different image modalities. Of the 278 eyes, only 172 (61.9%) were graded by both observers as having all image modalities with good quality and were established as the good quality set for
the subsequent data analysis. Bad quality of the FAF image was the main reason for a bad-quality indication for the multi-modal examination (CF+FAF+NIR).

Table 3.1: Number and percentage of good-quality images as indicated by observers for the different image modalities independently.

<table>
<thead>
<tr>
<th>Image Modality</th>
<th>Observer 1</th>
<th>Observer 2</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF</td>
<td>272 (97.8%)</td>
<td>268 (96.4%)</td>
<td>264 (95.0%)</td>
</tr>
<tr>
<td>FAF</td>
<td>211 (75.9%)</td>
<td>195 (70.1%)</td>
<td>185 (66.5%)</td>
</tr>
<tr>
<td>NIR</td>
<td>269 (96.8%)</td>
<td>265 (95.3%)</td>
<td>264 (95.0%)</td>
</tr>
</tbody>
</table>

Last column shows the number of images where both observers agree that the image is of good quality.

**Comparison of single- and multi-modality grading**

Figure 3.3 shows the ROC curves for the single modality approaches using CF (Figure 3.3a), FAF (Figure 3.3b), or NIR (Figure 3.3c) images and the multi-modality grading approach (Figure 3.3d). The point on the curve closest to the upper left corner in the ROC curve is used to compute sensitivity/specificity pairs.

Table 3.2 shows the Az values and sensitivity/specificity pairs for the single- and multi-modality grading of observer 1 and observer 2, respectively, calculated on the full data set and on the subset of good-quality images as indicated by both observers. The performance of both observers for RPD detection considerably increased when performing a multi-modality grading.

κ statistics were calculated to measure inter-observer variability during single- and multi-modality grading sessions. Table 3.3 shows the κ values between the observers for the different grading sessions. Observers achieved a higher agreement with multi-modality grading. When considering only good-quality images, observers also achieved high agreement when using FAF images.

**Performance of the automatic method**

The ROC curves for the proposed machine learning algorithm are shown in Figure 3.3. The corresponding Az values and the sensitivity/specificity pairs for the single- and multi-modality approaches are summarized in Table 5.4.

**Quantification of the area covered by RPD**

The box plots in Figure 3.4 show the RPD area percentage inside the ETDRS grading grid as delineated by the observers and as identified by the automatic system. Only eyes that were of good quality as indicated by both observers were taken into account. The
Figure 3.3: Receiver operating characteristics curves for the identification of eyes with RPD using (a) CF images, (b) FAF images, (c) NIR images and (d) a multi-modality setup.

multi-modality area delineations made during consensus grading of the two observers was used as the reference for the RPD area quantification.

The agreement between single-modality RPD area delineations made by the observers and the reference delineations set by using multi-modal information reached ICC values of 0.580 (-0.034; 0.830), 0.790 (0.409; 0.920), and 0.930 (0.763; 0.976) for the CF, FAF, and NIR delineations, respectively. For the automatic quantification of the RPD area, ICC values of 0.637 (0.395; 0.796), 0.389 (0.082; 0.631), and 0.557 (0.280; 0.747) were obtained for the single-modality analysis of CF, FAF, and NIR images with respect to the reference delineations. Comparing the automatic multi-
3.4 Discussion

Table 3.2: Performance of Observers 1 and 2 for RPD detection using single- and multi-modality grading.

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>All</th>
<th>Good quality</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Az  Se  Sp</td>
<td>Az  Se  Sp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single modality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF</td>
<td>0.879*</td>
<td>0.778 0.951</td>
<td>0.888* 0.789 0.974</td>
<td></td>
</tr>
<tr>
<td>FAF</td>
<td>0.881*</td>
<td>0.889 0.806</td>
<td>0.959 0.946 0.966</td>
<td></td>
</tr>
<tr>
<td>NIR</td>
<td>0.936</td>
<td>0.903 0.956</td>
<td>0.936 0.929 0.918</td>
<td></td>
</tr>
<tr>
<td>Multimodality</td>
<td>0.940 0.917 0.961</td>
<td>0.956 0.944 0.963</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observer 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single modality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF</td>
<td>0.944</td>
<td>0.944 0.989</td>
<td>0.944 0.930 0.917</td>
<td></td>
</tr>
<tr>
<td>FAF</td>
<td>0.793*</td>
<td>0.653 0.951</td>
<td>0.961 0.973 0.946</td>
<td></td>
</tr>
<tr>
<td>NIR</td>
<td>0.932</td>
<td>0.903 0.922</td>
<td>0.929* 0.900 0.918</td>
<td></td>
</tr>
<tr>
<td>Multimodality</td>
<td>0.958 0.972 0.942</td>
<td>0.974 1.000 0.949</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Area (Az) under the ROC values and optimal sensitivity (Se) and specificity (Sp) values are reported.

* Indicates a statistical significant difference of the Az value with respect to the multi-modality approach.

Table 3.3: Kappa agreement (κ) and 95% confidence intervals (CI) between observers for single-modality and multi-modality reading sessions.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Good quality</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>κ  95% CI</td>
<td>κ  95% CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single modality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF</td>
<td>0.654</td>
<td>(0.556-0.752) 0.724 (0.632-0.817)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAF</td>
<td>0.468</td>
<td>(0.363-0.572) 0.938 (0.879-0.998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIR</td>
<td>0.884</td>
<td>(0.822-0.945) 0.839 (0.767-0.910)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multimodality</td>
<td>0.911 (0.857-0.965)</td>
<td>0.936 (0.874-0.998)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CI, confidence interval.

modality approach with the reference standard, an ICC value of 0.704 (0.495; 0.837) was obtained.

3.4 Discussion

In this study, we assessed the performance achieved for RPD detection by using multi-modal information and compared it to the one obtained by using several single-image techniques. In our larger data set\textsuperscript{228-234}, we have demonstrated that a significantly higher performance, as well as a better inter-rater agreement, is achieved when the reticular pattern is assessed in a multi-modality grading approach. Moreover, our automatic machine learning algorithm for RPD detection and quantification using multi-modal information performed within the same range as the human graders.
Table 3.4: Performance of the automatic system for RPD detection using single- and multi-modality grading.

<table>
<thead>
<tr>
<th></th>
<th>All Good quality</th>
<th>All Good quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Az</td>
<td>Se</td>
</tr>
<tr>
<td>Single modality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF</td>
<td>0.942</td>
<td>0.833</td>
</tr>
<tr>
<td>FAF</td>
<td>0.844*</td>
<td>0.806</td>
</tr>
<tr>
<td>NIR</td>
<td>0.927</td>
<td>0.847</td>
</tr>
<tr>
<td>Multimodality</td>
<td>0.941</td>
<td>0.875</td>
</tr>
</tbody>
</table>

Area (Az) under the ROC values and Se and Sp values are reported.
* Indicates a statistical significant difference of the Az value with respect to the multi-modality approach.

Two independent human observers identified RPD areas by using two different grading protocols. During the single modality grading session, only information from a single-image technique was available, whereas during the multi-modality approach, the observers evaluated evidence of RPD by using all the modalities simultaneously. Both observers achieved higher performance with the multi-modality approach, reaching Az values of 0.940 and 0.958 (Figure 3.3; Table 3.2). Although previous studies have evaluated the accuracy for detecting RPD of single-image modalities only, our results confirm their conclusions that a more accurate diagnosis of RPD is obtained by using multiple image modalities.

In contrast to observer 1, observer 2 achieved high performance on RPD assessment with CF images. Possible reasons for this observation include the vaster experience of this observer on this modality and the lower sensitivity of this image technique. The disparity between the observers performance was substantially reduced when the assessment was performed by using multiple image modalities (Table 3.2). When the observers scored FAF images, the performance was significantly lower than when they used multi-modality reading. This may be due to the poor quality level of the FAF images. Only 66.5% of the images were considered of good quality by both observers, as shown in Table 3.1. During FAF acquisition, a mean intensity image is constructed to reduce noise in the image. However, eye movements may cause displacement errors, resulting in a lower contrast and thus lower quality of the FAF image. Another reason is the presence of cataracts in the study population. The wavelength used for FAF imaging is affected more by cataracts than the one used in NIR imaging, resulting in lower image quality. As shown in Table 3.2, the adoption of a multi-modality grading approach can overcome image quality issues, maintaining a high detection performance independently of the quality level of a particular image technique. When considering only the subset of good-quality images, the performance of both observers increased for both single- and multi-modality gradings.
3.4 Discussion

Inter-observer agreement was also investigated by using the two grading protocols. Table 3.3 shows that the agreement between observers substantially increased when multiple imaging techniques were used to evaluate the evidence of RPD. When taking only the subset of good-quality images into account, the agreement between observers improved when using CF, FAF, and the multi-modal approach. However, agreement when using CF images is still substantially lower than when using the other modalities. Other studies\textsuperscript{229} have included multiple graders but no information about inter-observer agreement has been reported.

In this study, we also developed and evaluated a machine learning algorithm for the automatic identification and quantification of RPD using multi-modal information. The results showed that the proposed system achieved similar performance as the observers (Figure 3.3; Table 5.4). Similar to the observers gradings, the incorporation of multi-modal information improved the performance of the algorithm. Using multi-modal information, the proposed algorithm achieved an $Az$ value of 0.941 and a sensitivity/specificity pair of 0.875/0.873. Compared to the observers, who reached a $\kappa$ agreement of 0.87 with the reference, the automatic system had a $\kappa$ agreement of 0.70. However, 20% of the misclassified cases correspond to cases where there was disagreement between the observers. Of the false-positive cases, 9 cases contained low-quality images, 3 cases presented geographic atrophy, 1 case showed a neovascular macular detachment, and 12 cases contained soft indistinct drusen. As described in other pu-
RPD and drusen have very similar characteristics and they might therefore more easily be misinterpreted by the automatic system. Better discriminant features, such as image context information or local intensity changes, might improve the performance of the automatic system, but this has to be further investigated.

Quantification of RPD area is a more difficult task owing to the undefined boundaries of RPD. When comparing the manual delineations performed on CF images with the reference delineations based on multi-modal information, an ICC value of 0.580 was achieved. When comparing the FAF or NIR delineations with the reference delineations, the agreement was better, reaching ICC values of 0.790 and 0.930, respectively. As presented in Figure 3.4, the RPD area was underestimated when using CF images as compared with the other image techniques. As reported in previous publications, the visibility of RPD differs over imaging modalities, causing these differences. As RPD are more pronounced on FAF and NIR, the delineations on these modalities were more similar to reference delineations. The quantified RPD area, which was automatically obtained by the proposed algorithm, was in agreement with the area delineated by the observers, reaching an ICC value of 0.704. Of note, only images of good quality were used for RPD area quantification because images with insufficient quality were not suitable, as it was nearly impossible for observers to delineate RPD area on these images. Another limitation of this study was that the multi-modal approach included only fundus images, excluding information obtained with SD-OCT. Including this modality in the multi-modal protocol might result in better understanding of the reticular pattern, thus increasing accuracy in their identification. Spectral-domain OCT can provide 3-D information about RPD formation and is essential for RPD volume measurements. This enhancement will be of great importance for clinical trials studying the development and progression of RPD. We will investigate this improvement in further studies.

In conclusion, we were able to show that a multi-modal approach significantly increased observer performance and inter-observer agreement for detection of RPD in fundus images when the information of different imaging modalities was evaluated simultaneously. Furthermore, an automatic machine learning algorithm for detection and quantification of RPD using multi-modal information was developed and evaluated, showing comparable results with those obtained by observers. The area covered by RPD was also automatically quantified by the algorithm, tallying the values manually provided by the observers. The absence of SD-OCT is regarded as a limitation of this study and will be investigated in future work. This automatic algorithm yields a quick and reliable diagnosis and quantification of RPD, for large data set analysis within population studies and for gaining insights into risk factors involved in AMD and disease progression.
Self-learning systems for reticular pseudodrusen detection

Mark J. J. P. van Grinsven, Thomas Theelen, Geert J. S. Litjens, Corina Brussee, Gabriëlle H.S. Buitendijk, Carel B. Hoyng, Caroline C. W. Klaver, Bram van Ginneken, Clara I. Sánchez

*Original title:* Self-learning computer algorithms can identify visual threatening retinal diseases from multimodal imaging

*Submitted*
Abstract

Purpose: To automatically identify and quantify areas of reticular pseudodrusen (RPD), a retinal abnormality highly correlated with increased risk of vision loss by end-stage macular degeneration, using a self-learning system based on deep learning which independently learns discriminative patterns from multimodal retinal imaging data, eliminating the need for ill-defined manual descriptors.

Subjects: Retinal imaging data including 278 eyes of 230 patients, of which 72 eyes had RPD, 108 had drusen and 98 were control eyes, were drawn from the Rotterdam Study database.

Methods: A self-learning system based on deep learning, trained with multimodal patches from color fundus and near infra-red images, was developed for automatic detection and quantification of RPD. The performance of the self-learning system was compared with a previously developed human-guided system and three human experts.

Results: The self-learning system obtained an area (Az) under the Receiving Operating Characteristics curve of 0.966 with 95% confidence interval (CI) of [0.943-0.985] for the identification of RPD and was significantly better (p=0.0018) than the human-guided system (Az=0.927, CI=[0.893-0.956]). With sensitivity (Se), specificity (Sp) and kappa (κ) values of 0.92, 0.93 and 0.82, respectively, the self-learning system performed on par with human experts (Se values ranging from 0.86 to 0.92, Sp values from 0.87 to 0.97, and κ values from 0.72 to 0.84). Automatic quantification of RPD area by the self-learning system (dice similarity coefficient (DSC) of 0.703; intra-class correlation (ICC) of 0.572) was in line with human experts (DSC values from 0.608 to 0.711 and ICC values from 0.428 to 0.626 and better than the human-guided system (DSC=0.680; ICC=0.375).

Conclusions: This study showed that self-learning systems based on deep learning reached human performance for the automatic identification and quantification of RPD. The self-learning system learned to efficiently leverage information contained in multimodal imaging, avoiding ill-defined characteristics of RPD and increasing the final classification performance. These self-learning systems allow automatic and accurate interpretation of large multimodal image data sets, opening the way for larger clinical studies on AMD disease mechanisms and AMD treatment development.
4.1 Introduction

Retinal imaging has become an indispensable diagnostic tool in ophthalmology. Technological advances in imaging in the last decades allow to observe the retina with more details and to identify eye diseases in earlier stages, even before symptoms occur, so they can be treated in a timely manner to prevent vision loss\cite{25,38}. Different characteristics of eye diseases are highlighted with a variety of image modalities, including color fundus photography, near-infrared imaging and optical coherence tomography (OCT). This multi-modal approach greatly improves the identification of subtle lesions and provides a better understanding of disease status and progression\cite{39,41}.

Age related macular degeneration (AMD) is the most common eye disease in the elderly population which leads to vision loss and ultimately blindness if it progresses to final stages\cite{23,24}. Reticular pseudodrusen (RPD), (see Figure 4.1), an extracellular lesion associated with AMD, is a strong indicator for progression to late stage AMD\cite{36,37,39,222}. Multimodal imaging allows for a better identification of RPD\cite{228,230}. Reliable and accurate identification and quantification of RPD provides an important measure of AMD progression, essential for timely and cost-effective treatment decisions.

Multi-modal image interpretation and lesion recognition is a complex and costly task if done by human readers. In the view of this situation, computer learning systems have been developed and trained to automatically interpret retinal images\cite{150,187,235,245,246}. Our group developed a learning system that analyzes several image modalities simultaneously to automatically identify and quantify RPD regions\cite{246}. These learning systems, or as referred in the paper, human-guided learning systems, are designed based on visual characteristics or descriptors that human experts consider important for abnormality discrimination, such as color, texture and size. However, due to the complexity of the recognition task and the variability of visual characteristics in the different image modalities, good discriminative descriptors are difficult to define and hard to translate to a machine learning system, limiting the system performance and, consequently, impeding their adoption in clinical settings.

Deep learning is a revolutionary learning approach that has been recently introduced in the field of machine learning and computer vision\cite{116,117}. In deep learning, the learning system independently learns the discriminative descriptors from the data, omitting the use of ill-defined visual characteristic\cite{116,117}. These self-learning systems try to mimic the neural network of the human brain where information is processed in several layers of abstraction. Self-learning systems based on deep learning have outperformed previously proposed human-guided learning systems and have reached the same performance as that of humans for specific visual tasks, such as object recognition in natural images\cite{247,248}. Therefore, deep learning is increasingly used for computer vision
Figure 4.1: Examples of color fundus and near infra-red images showing presence of reticular pseudodrusen (top), drusen (middle) and a control case (bottom).

tasks in large data sets. We hypothesize that these self-learning systems can also be very useful for more complex tasks, such as medical image interpretation in multi-modal approaches.

In this paper, we design and evaluate a self-learning system based on deep learning for the identification and quantification of RPD using different image modalities. We compare its performance to the opinion of three human experts and to a previously developed system based on human-guided learning. We demonstrate that self-learning systems can be of great value in medical image interpretation and can potentially be applied to diverse complex visual tasks.
4.2 Methods

Study dataset

The dataset used in this study was selected from a cross-sectional round of the Rotterdam Study database\textsuperscript{237}. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study). All participants provided written informed consent to participate in the study and to obtain information from their treating physicians. The study complied with the tenets sets forth in the declaration of Helsinki. All subjects which had color fundus (CF) images and near infra-red (NIR) images taken were eligible for inclusion in this study. All subjects with presence of RPD as graded by the Rotterdam Study graders and which had both CF and NIR images taken, were included in this study. Color fundus images were acquired using a 35° field-of-view Topcon TRC 50EX fundus camera (Topcon Optical Company, Tokyo, Japan) with a Sony DXC-950P digital camera with a resolution of 768 x 576 pixels (Sony Electronics, Inc., New York, NY, USA). NIR images were acquired with a Heidelberg Retina Angiograph 2 (Heidelberg Engineering, Heidelberg, Germany) with a field of view of 30° and a resolution of 768 x 768 pixels. Images were centered at the macula with at least part of the optic nerve visible.

Table 4.1: Patient demographics of the study population.

<table>
<thead>
<tr>
<th></th>
<th>Total N=230</th>
<th>Controls N=85</th>
<th>Drusen N=101</th>
<th>RPD N=44</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, yrs (sd)</td>
<td>78.1 (6.5)</td>
<td>75.3 (5.8)</td>
<td>78.7 (6.1)</td>
<td>81.8 (6.4)</td>
</tr>
<tr>
<td>Sex, % male</td>
<td>41.7</td>
<td>47.1</td>
<td>43.6</td>
<td>27.3</td>
</tr>
<tr>
<td>BMI, kg/m(^2)</td>
<td>27.5 (4.0)</td>
<td>27.3 (8.5)</td>
<td>27.8 (4.1)</td>
<td>27.2 (3.5)</td>
</tr>
<tr>
<td>Waist circumference, cm (sd)</td>
<td>93.3 (11.0)</td>
<td>92.5 (11.1)</td>
<td>94.3 (11.4)</td>
<td>92.6 (10.0)</td>
</tr>
<tr>
<td>Hip circumference, cm (sd)</td>
<td>103.3 (8.3)</td>
<td>102.9 (8.5)</td>
<td>103.6 (8.1)</td>
<td>103.2 (8.7)</td>
</tr>
<tr>
<td>Systolic blood pressure, mmHg (sd)</td>
<td>154.8 (22.4)</td>
<td>154.9 (21.6)</td>
<td>155.8 (23.4)</td>
<td>152.2 (21.8)</td>
</tr>
<tr>
<td>Diastolic blood pressure, mmHg (sd)</td>
<td>85.8 (11.4)</td>
<td>86.8 (10.9)</td>
<td>84.9 (11.5)</td>
<td>85.9 (12.1)</td>
</tr>
<tr>
<td>Hypertension, %</td>
<td>89.1</td>
<td>88.2</td>
<td>91.1</td>
<td>86.4</td>
</tr>
<tr>
<td>Never</td>
<td>25.7</td>
<td>27.1</td>
<td>23.8</td>
<td>27.3</td>
</tr>
<tr>
<td>Smoking, %</td>
<td>Past</td>
<td>62.6</td>
<td>61.2</td>
<td>65.3</td>
</tr>
<tr>
<td>Current</td>
<td>11.7</td>
<td>11.8</td>
<td>10.9</td>
<td>13.6</td>
</tr>
</tbody>
</table>

N: number of patients; BMI: body mass index; sd: standard deviation; yrs: years.

Table 4.1 shows more information about the study population demographics. According to the graders of the Rotterdam Study database, 72 eyes showed presence of RPD. Subjects with drusen and healthy subjects as confirmed by the Rotterdam Study graders were randomly added to the study dataset. In total 278 eyes of 230 patients were
Self-learning systems for reticular pseudodrusen detection

included, of which 72 eyes showed presence of RPD, 108 showed presence of drusen and 98 were graded as control\textsuperscript{31,33}. Drusen were characterized as small round deposits on CF and NIR images, having a yellowish color on CF images\textsuperscript{228}, whereas RPD are characterized as yellowish interlacing networks on CF images and as groups of hyporeflectant lesions against a mildly hyperreflectant background on NIR images\textsuperscript{231,234}. Delineation of the area affected with RPD were made in consensus by two human experts (G.H.S.B and C.B.) on the 72 eyes showing RPD using CF and NIR imaging\textsuperscript{246}. These consensus delineations were used to train the automatic systems and as reference for RPD area quantification.

\textbf{Self-learning system: System description}

In this work, we implemented a self-learning system based on a deep convolutional neural network for the automatic detection and quantification of RPD using multi-modal information\textsuperscript{116,117}. Using a group of training examples, the self-learning system learned an optimal representation of RPD to discriminate them from non-RPD regions. The self-learning system consisted of a neural network of multiple convolutional layers. In each layer the inputs were convolved with a set of filters and the filter response maps were passed down to the next layer. The inputs of the first layer are small multi-modal image patches extracted from the CF image and the NIR image. The filters of the network were learned from the data during the training phase and were therefore tailored to the problem of RPD detection. After the learning phase was completed, the self-learning system was used to automatically analyze an unseen test case in order to identify the presence of RPD and quantify the area covered by them. More details about the system can be found in Appendix A.

\textbf{Evaluation of the automatic method}

All data was randomly split into five folds, allowing for a fair evaluation on the whole data set using a 5-fold cross-validation scheme. In case data from both eyes of a subject was present, both eyes were kept in the same fold. Three folds were used for training, one fold for network monitoring and one fold for evaluation. Rotating this scheme resulted in all images being evaluated.

To evaluate the performance on the identification of RPD, Receiver Operating Characteristic (ROC) analysis\textsuperscript{120} was performed and sensitivity/specificity pairs were calculated as well as kappa agreement. Bonferroni correction was applied to correct for multiple comparisons\textsuperscript{241}. For the evaluation of RPD area quantifications, dice similarity coefficient (DSC)\textsuperscript{249} and intra-class correlation (ICC)\textsuperscript{125} values were calculated and
compared with the values obtained by the human experts and a previously developed human-guided learning system.

**Human-guided learning system: System description**

Previously, we have developed a human-guided learning system for the automatic detection of RPD which makes use of visual characteristics or descriptors, considered important by humans. A full description of this human-guided learning system can be found following the reference. We have adapted this human-guided learning system according to the parameters used in this study to be able to make a direct comparison between the two different types of systems. The human-guided learning system incorporated also the green and blue channel of the CF image as well as the NIR image. Patches to train the system were exactly the same as the ones used to train the self-learning system. The human-guided learning system was trained and evaluated following the same cross-validation scheme and data division as used for the self-learning system. The same evaluation protocol, including ROC analysis, DSC and ICC values, was used.

### 4.3 Results

**Recognizing RPD: Performance of the human experts**

Three human experts (C.B.H., C.C.W.K. and T.T.) independently indicated for each eye if RPD were present. Each expert was shown the CF and NIR images simultaneously during grading while being blinded to the reference and other observers’ annotations. Compared to the reference, human experts obtained sensitivity/specificity pairs of 0.92/0.87, 0.89/0.95 and 0.86/0.97, respectively, and kappa agreement of 0.72, 0.83 and 0.84, respectively. See Table 4.2 for an overview.

<table>
<thead>
<tr>
<th>Expert</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Kappa agreement [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1</td>
<td>0.92</td>
<td>0.87</td>
<td>0.72 [0.63,0.81]</td>
</tr>
<tr>
<td>Expert 2</td>
<td>0.89</td>
<td>0.95</td>
<td>0.83 [0.76,0.91]</td>
</tr>
<tr>
<td>Expert 3</td>
<td>0.86</td>
<td>0.97</td>
<td>0.84 [0.76,0.91]</td>
</tr>
</tbody>
</table>

**Recognizing RPD: Performance of the human-guided learning system**

The human-guided system obtained an area (Az) under the ROC curve of 0.927 with 95% confidence interval of [0.893;0.956]. A sensitivity/specificity pair of 0.81/0.90 and
kappa agreement of 0.68 [0.58;0.78] was achieved. Figure 4.2 shows the ROC curve of the human-guided learning system. Operating points of the three human experts are also added in the plot for the sake of comparison.

Figure 4.2: Receiver operating characteristics (ROC) curves for the identification of eyes with reticular pseudodrusen. The plot shows the ROC curves for the self-learning system and the human-guided learning system as well as the three human expert operating points, for the identification of eyes with presence of reticular pseudodrusen.

Recognizing RPD: Performance of self-learning system

The proposed self-learning system obtained an area under the ROC curve of 0.966 [0.943;0.985] and a sensitivity/specificity pair of 0.92/0.93 with kappa agreement of 0.82 [0.74;0.90]. Figure 4.2 shows the ROC curve of the self-learning system, together with the performance of the human experts and the human-guided system. In Table 4.3, the performance obtained using different system configurations, as explained in Appendix A, is reported and compared to the human-guided system. The self-learning
system reached a performance similar to the three human experts and significantly outperformed the performance of the human-guided system ($p$-value=0.0018).

Table 4.3: Performance of different configurations of the self-learning system and the human-guided system for the identification of eyes with reticular pseudodrusen. Area (Az) under the ROC values for the identification of eyes with reticular pseudodrusen for the different configurations and the human-guided system are reported including the 95% confidence interval (CI). Last column shows $p$-values for statistical comparison with the human-guided system after Bonferroni correction was applied. Values marked with * are statistical significant different compared to the human-guided system.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Az value [95% CI]</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-learning systems</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double</td>
<td>0.961 [0.937,0.980]</td>
<td>0.01260*</td>
</tr>
<tr>
<td>Base</td>
<td>0.953 [0.924,0.976]</td>
<td>0.02970*</td>
</tr>
<tr>
<td></td>
<td>0.964 [0.941,0.982]</td>
<td>0.00180*</td>
</tr>
<tr>
<td>Model B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double</td>
<td>0.949 [0.919,0.974]</td>
<td>0.24210</td>
</tr>
<tr>
<td>Base</td>
<td>0.958 [0.934,0.978]</td>
<td>0.04950*</td>
</tr>
<tr>
<td></td>
<td>0.966 [0.943,0.985]</td>
<td>0.00180*</td>
</tr>
<tr>
<td>Model C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double</td>
<td>0.962 [0.934,0.983]</td>
<td>0.00990*</td>
</tr>
<tr>
<td>Base</td>
<td>0.964 [0.943,0.982]</td>
<td>0.00360*</td>
</tr>
<tr>
<td></td>
<td>0.947 [0.920,0.970]</td>
<td>0.26910</td>
</tr>
<tr>
<td><strong>Human-guided system</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.927 [0.893,0.956]</td>
<td>-</td>
</tr>
</tbody>
</table>

**RPD quantification: Beyond recognition**

RPD area quantification results of the self-learning system and the human-guided learning system were compared with three human expert delineations of RPD. These human expert delineations were made by the same three human experts in a second, independent grading session on the 72 RPD cases of which they were able to delineate the area affected by RPD. For this, the same grading protocol was used by showing CF and NIR images simultaneously to the experts. Figure 4.3 shows some examples of the obtained RPD delineations of the two learning systems, the human experts and the reference. DSC values and ICC values of the area affected by RPD with respect to the reference were calculated (Table 4.4). The self-learning system achieved a DSC value of 0.703 and a mean ICC value of 0.527 with 95% confidence interval of [0.326;0.682], which is in the range of the values of the human experts. The human-guided learning system achieved a lower DSC than the self-learning system and two experts and a lower mean ICC value than the self-learning system and all three human experts.
Table 4.4: Reticular pseudodrusen area quantification. DSC: Dice similarity coefficient values for the area of reticular pseudodrusen compared with the reference. ICC: Intraclass correlation coefficient values including 95% confidence interval (CI) for the area covered by reticular pseudodrusen.

<table>
<thead>
<tr>
<th></th>
<th>Mean DSC</th>
<th>st. dev. DSC</th>
<th>ICC value [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1</td>
<td>0.708</td>
<td>0.167</td>
<td>0.607 [0.417,0.746]</td>
</tr>
<tr>
<td>Expert 2</td>
<td>0.711</td>
<td>0.158</td>
<td>0.626 [0.210,0.812]</td>
</tr>
<tr>
<td>Expert 3</td>
<td>0.608</td>
<td>0.205</td>
<td>0.428 [-0.068,0.715]</td>
</tr>
<tr>
<td>Self-learning system</td>
<td>0.703</td>
<td>0.160</td>
<td>0.527 [0.326,0.682]</td>
</tr>
<tr>
<td>Human-guided system</td>
<td>0.680</td>
<td>0.154</td>
<td>0.375 [0.131,0.576]</td>
</tr>
</tbody>
</table>

Figure 4.3: Delineations of reticular pseudodrusen area by the self-learning system, human-guided system, three human experts and the reference standard. First column shows the color fundus image, second column shows the co-registered near infra-red image and last column shows the delineations overlaid on the co-registered near infra-red image. In black: reference standard; purple: self-learning system; red: human-guided system; light-blue, yellow and green: three human experts, respectively.
4.4 Discussion

We have developed and evaluated a self-learning system for the automatic identification and quantification of RPD using multi-modal retinal imaging. In contrast to a human-guided learning system, the developed self-learning system was able to independently learn from image data and characterize the abnormal pattern of RPD without the need of externally defined descriptors.

Results show that the self-learning system performs at the same level of three human experts and significantly outperforms a previously developed human-guided learning system (see Table 4.3, Table 4.4 and Figure 4.2). The difference in performance can be attributed to the definition of abnormality descriptors: whereas the human-guided system uses externally defined descriptors based on texture and intensity, the proposed self-learning system has learned from the data to discriminate between RPD regions and non-RPD regions. By learning these descriptors from the data, these descriptors are better tailored to the problem and would allow for a better discrimination of RPD and non-RPD regions. In total, 14 eyes were incorrectly classified as having RPD by the self-learning system of which 13 eyes had drusen and 1 eye was a control case. In 8 out of those 14 cases, there was a disagreement between at least one of the three human experts and the reference, showing the difficulty and subjectivity in identifying RPD, especially in the presence of other abnormalities such as drusen.

In addition to RPD identification, the system was able to identify RPD areas accurately with a mean DSC value of 0.703 and ICC value of 0.527 (see Table 4.4). Those values were in concordance with the performance of expert 1 and 2, while expert 3 has a lower performance. The visual inspection of the results showed that human expert 3 tends to undersegment the RPD area as compared to the reference standard.

We tested several self-learning system configurations, shown in Table 4.5, which all obtained Az values higher than the human-guided system (see Table 4.3). The configuration settings of the self-learning system have only little effect on the performance of the overall system. An explanation for this observation could be that the self-learning system is complex enough to capture the variety in the RPD patterns and small changes to the self-learning system configuration may not lead to big overall performance differences. In other computer vision tasks, such as discrimination between a large number of diverse objects in natural images, larger and more complex self-learning systems were used in order to discern a wider range of patterns. Results of the difference evaluated network configurations show that such complex networks are not needed for this specific task, (Table 4.3 and Table 4.6).

Our study has some limitations. A recurring problem with medical image analysis tasks is the absence of a proper external reference. Therefore, using manual gradings
and delineations are the best available option we could use in this study. The reference for the presence of RPD was set by human graders and the RPD delineations used to train the system were made by a consensus of two human experts. Errors in this reference are inevitable and, after inspection and comparison of the reference with the three human experts, we found that in 5 cases, the reference identified RPD while neither of three experts did; and in 6 cases, RPD was not identified by the reference whereas all three experts found RPD areas. However, this problem might be hard to solve. A large study on fundus image analysis by a considerable number of experienced retinal experts showed a large inter- and intra-observer variation\textsuperscript{251}. Adding OCT data might help to set a more reliable reference.

The use of multi-modal information has the advantage of providing more complementary information, resulting in a better performance for both humans and machine learning systems\textsuperscript{225,229,246}. Manually devising a way to combine this information is a challenging task and would hardly result in an optimal combination. Self-learning systems are able to deal with such a rich amount of information due to the self-learning capabilities, obtaining an optimal combination of information from both CF and NIR modalities. However, using multi-modality information requires a proper co-registration of the modalities to be spatially aligned. In literature, methods to register multiple retinal imaging modalities have been described\textsuperscript{91,252–254}. However, to reduce registration errors to the best possible extent, we used a semi-automatic technique. For introduction to a fully operational system in clinical practice, a fully automatic registration has to be incorporated in the overall framework.

In conclusion, we have shown that self-learning systems based on deep learning achieve performance on par with human experts for the automatic identification and quantification of RPD using multi-modal retinal images. The self-learning system learned to efficiently leverage information from multi-modal imaging data, avoiding the use of ill-defined manual descriptors and resulting in an increased classification performance. These self-learning systems could be more widely implemented in retinal image analysis tasks and can potentially take over visual recognition tasks of human experts without giving in on the performance. This opens the way for larger clinical studies on AMD disease mechanism as well as for AMD treatment development.

4.5 Appendix A

In this Appendix we describe in detail the proposed self-learning algorithm based on deep learning for the identification and quantification of RPD. Detailed information about the self-learning system parameters and configurations are also provided.
4.5 Appendix A

4.5.1 Preprocessing

Before any processing, CF images and NIR images of each patient were co-registered using a semi-automatic registration technique to ensure spatial correspondence\textsuperscript{246}. CF and NIR images were spatially aligned by automatically rotating, scaling and translating the NIR image to match the CF image. The field of view in both the CF and NIR images was automatically segmented by applying a fixed threshold. Image pixels which had no corresponding locations in the other modality after co-registration or which were outside the field of view were discarded from further analysis.

Self-learning networks: General configuration details

The self-learning system was designed as an architecture of multiple, repetitive convolutional layers\textsuperscript{116,250}. Figure 4.4 shows an overview of the self-learning network configurations. In each layer, the input was convolved with a number of filters (i.e. kernels) \( K \) with size \( n \times m \times c \) (width \( \times \) height \( \times \) channels). After convolution, the resulting \( K \) feature maps were passed through a non-linear mapping; and spatial 2 by 2 pixels max-pooling with a stride of 2 pixels was applied. This reduced the complexity of the network by down-sampling the feature maps, becoming invariant to small spatial translations\textsuperscript{116}. Rectified linear units (RELUs) were used as non-linear mapping as they have shown to improve performance in other computer vision tasks\textsuperscript{255–257}. The input of the first layer was a three-dimensional patch of 40 by 40 pixels centered on the pixel to be evaluated. These patches consisted of the green and blue channel of the CF image and the NIR image, where each of these was stacked in the "channel" dimension. The red channel of the CF image was left out as this color plane is usually saturated due to the intrinsic properties of the human retina. After the last convolutional layer, a soft-classification using logistic regression of the pixel was made into either RPD or non-RPD, providing a numerical output between 0 and 1 representing the probability of the pixel of being RPD. Applying this scheme to all pixels in the image resulted in a pixel probability map.

Self-learning networks: Training the system

Three-dimensional patches of size 40 by 40 pixels, constructed from the green and blue channel of the CF image and the NIR image, were randomly extracted from a group of training examples. The patch size of 40 was chosen to ensure enough contextual information in the patch, while keeping the patch as small as possible to reduce the number of parameters and computational cost.

All network parameters (weights of the convolutional filters and bias terms) were optimized during the training phase of the network using stochastic gradient descent\textsuperscript{258}. 
This method optimized the parameters by minimizing a cost $C(y, p)$, defined as the negative log-likelihood with $p$ the predicted label of patch $i$ and $y$ the true label, see Equation 4.1.

$$C(y, p) = -\sum_{i=0}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \quad (4.1)$$

In gradient descent, all parameters are updated after a full pass through the training data (or epoch) has been performed. To speed up this process, in stochastic gradient descent, all parameters are updated after a much smaller amount of data (or mini-batch), resulting in a faster convergence during training$^{258}$.

From training images without RPD, 1000 patches were randomly extracted and labeled negative; whereas from training images with RPD, 300 patches were randomly extracted at locations where the center pixel was annotated as RPD according to the reference. To overcome over-fitting of the network, a monitoring set was used to measure performance during training of the network. Both the error on the training set and on the monitoring set will decrease in the early training phase. When the error rate on the monitoring set starts to increase, the training phase is considered finished as the network starts over-training, losing its ability to generalize on unseen data.

**Self-learning networks: Variations in network configurations**

We tested different network configurations to evaluate the influence of network configurations on performance. We used networks with three different depths, specifically with two, three or four convolutional layers with a final soft-classification layer at the end. We named these configurations model A, model B and model C, respectively. For each of the models, we used different number $f_i$ of kernels $K$ in the $i^{th}$ convolutional layer. For the “Base” configuration, $f_1=8$, $f_2=16$, $f_3=64$ and $f_4=64$. The “Double” configuration has $f_1=16$, $f_2=32$, $f_3=128$ and $f_4=128$ kernels whereas the “Quadruple”
configuration has \( f_1=32, f_2=64, f_3=256 \) and \( f_4=256 \) kernels. See Table 4.5 for a full overview of the different configurations.

Table 4.5: Configurations of the different self-learning network models. For each convolutional layer, the width x height x channels of the kernels are reported with the number of kernels \( S_k \) in the \( i^{th} \) layer equal to \( f_i \). Max-pooling is applied with stride \( s \) in the first two layers.

<table>
<thead>
<tr>
<th></th>
<th>Input size</th>
<th>Model A</th>
<th>Model B</th>
<th>Model C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolutional</td>
<td>40x40</td>
<td>7x7x3, ( S_k=f_1 )</td>
<td>7x7x3, ( S_k=f_1 )</td>
<td>7x7x3, ( S_k=f_1 )</td>
</tr>
<tr>
<td>Layer 1</td>
<td></td>
<td>2x2 max-pool, ( s=2 )</td>
<td>2x2 max-pool, ( s=2 )</td>
<td>2x2 max-pool, ( s=2 )</td>
</tr>
<tr>
<td>Convolutional</td>
<td>17x17</td>
<td>5x5x3, ( S_k=f_2 )</td>
<td>5x5x3, ( S_k=f_2 )</td>
<td>5x5x3, ( S_k=f_2 )</td>
</tr>
<tr>
<td>Layer 2</td>
<td></td>
<td>2x2 max-pool, ( s=2 )</td>
<td>2x2 max-pool, ( s=2 )</td>
<td>2x2 max-pool, ( s=2 )</td>
</tr>
<tr>
<td>Convolutional</td>
<td>7x7</td>
<td></td>
<td>3x3x3, ( S_k=f_3 )</td>
<td>3x3x3, ( S_k=f_3 )</td>
</tr>
<tr>
<td>Layer 3</td>
<td></td>
<td></td>
<td>no max-pool</td>
<td>no max-pool</td>
</tr>
<tr>
<td>Convolutional</td>
<td>5x5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Layer 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soft-classification</td>
<td>3x3</td>
<td>2 classes</td>
<td>2 classes</td>
<td>2 classes</td>
</tr>
</tbody>
</table>

Self-learning networks: Identification and quantification of eyes with RPD

For the unseen images, a patch for each pixel in the image was extracted and classified using the trained network. To speed up the computation process, a stride of 4 pixels in both horizontal and vertical direction was used, meaning a patch was extracted every four pixels in the image. After applying the trained network to all the image pixels, a pixel probability map was obtained, indicating the probability of each pixel to be part of a RPD structure. On an Intel Xeon PC with 2.4Ghz memory and a GeForce GTX 570 video card, processing of an entire image took around 85 seconds.

This pixel probability map was noisy and therefore we applied a smoothing step yielding a smoothed pixel probability map\(^{259}\). Table 4.6 shows the effect of the amount of smoothing on the RPD identification performance as measured with ROC analysis using the pixel probabilities. A final image score, indicating the probability of RPD presence is then obtained as the maximum value of the smoothed pixel probability map. To compute the area affected by RPD on images identified with RPD, an adaptive threshold, experimentally determined as 65% of the image score was applied to the smoothed pixel probability map, resulting in a binary RPD segmentation.
Table 4.6: Performance measured with ROC analysis on pixel level identification of reticular pseudodrusen with and without applying a smoothing operation. The table shows the area under the ROC (Az) values measured without (Raw Az) and with smoothing with different smoothing factors $\sigma$ applied.

<table>
<thead>
<tr>
<th>System</th>
<th>Raw Az</th>
<th>$\sigma=1$</th>
<th>$\sigma=2$</th>
<th>$\sigma=4$</th>
<th>$\sigma=8$</th>
<th>$\sigma=16$</th>
<th>$\sigma=32$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-guided</td>
<td>0.890</td>
<td>0.896</td>
<td>0.903</td>
<td>0.913</td>
<td>0.923</td>
<td>0.927</td>
<td>0.923</td>
</tr>
<tr>
<td>Model A</td>
<td>Quadruple</td>
<td>0.886</td>
<td>0.912</td>
<td>0.924</td>
<td>0.936</td>
<td>0.945</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td>Double</td>
<td>0.887</td>
<td>0.908</td>
<td>0.920</td>
<td>0.933</td>
<td>0.943</td>
<td>0.946</td>
</tr>
<tr>
<td></td>
<td>Base</td>
<td>0.897</td>
<td>0.915</td>
<td>0.926</td>
<td>0.938</td>
<td>0.948</td>
<td>0.954</td>
</tr>
<tr>
<td>Model B</td>
<td>Quadruple</td>
<td>0.881</td>
<td>0.907</td>
<td>0.921</td>
<td>0.935</td>
<td>0.944</td>
<td>0.948</td>
</tr>
<tr>
<td></td>
<td>Double</td>
<td>0.892</td>
<td>0.916</td>
<td>0.929</td>
<td>0.942</td>
<td>0.950</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>Base</td>
<td>0.883</td>
<td>0.912</td>
<td>0.926</td>
<td>0.940</td>
<td>0.949</td>
<td>0.956</td>
</tr>
<tr>
<td>Model C</td>
<td>Quadruple</td>
<td>0.885</td>
<td>0.912</td>
<td>0.926</td>
<td>0.941</td>
<td>0.950</td>
<td>0.953</td>
</tr>
<tr>
<td></td>
<td>Double</td>
<td>0.900</td>
<td>0.922</td>
<td>0.932</td>
<td>0.943</td>
<td>0.949</td>
<td>0.953</td>
</tr>
<tr>
<td></td>
<td>Base</td>
<td>0.874</td>
<td>0.900</td>
<td>0.915</td>
<td>0.930</td>
<td>0.940</td>
<td>0.944</td>
</tr>
</tbody>
</table>
Automatic differentiation of drusen and exudates

Mark J. J. P. van Grinsven, Thomas Theelen, Leonard Witkamp, Job van der Heijden, Johannes P. H. van de Ven, Carel B. Hoyng, Bram van Ginneken, Clara I. Sánchez

Original title: Automatic differentiation of color fundus images containing drusen or exudates using a contextual spatial pyramid approach

Published in: Biomedical Optics Express (2016); 7(3):709-725
Abstract

We developed an automatic system to identify and differentiate color fundus images containing no lesions, drusen or exudates. Drusen and exudates are lesions with a bright appearance, associated with age-related macular degeneration and diabetic retinopathy, respectively. The system consists of three lesion detectors operating at pixel-level, combining their outputs using spatial pooling and classification with a random forest classifier. System performance was compared with ratings of two independent human observers using human-expert annotations as reference. Kappa agreements of 0.89, 0.97 and 0.92 and accuracies of 0.93, 0.98 and 0.95 were obtained for the system and observers, respectively.
5.1 Introduction

Diabetic retinopathy (DR) is the leading cause of blindness worldwide in the working population with an estimated number of affected patients of 93 million in 2010\textsuperscript{260-262}. Age-related macular degeneration (AMD) is another sight threatening disease and the most common cause of blindness in the elderly. The estimated number of patients affected by AMD was 170 million in 2014\textsuperscript{24}. Both diseases progress without any visual complaints in early stages, while leading to visual impairment and, ultimately, vision loss in advanced stages. Major risk factors for these diseases include diabetes and an increased age. With the rising prevalence of diabetes and the aging population, the incidence of DR and AMD is expected to increase rapidly in the near future\textsuperscript{14,27}.

To improve early detection of these diseases, screening for both DR and AMD through retinal examination using color fundus (CF) images is recommended by national authorities\textsuperscript{55,59}. However, only screening for DR is currently routinely implemented for diabetic patients. The main limitation of widespread implementation of DR and AMD screening is the enormous screening population and the associated costs of acquiring and analyzing the large amount of images generated. Automatic software solutions have been proposed to allow for more cost-effective mass-screening, reducing the amount of specialized personnel required and making mass-screening feasible. Most of these automatic software solutions analyze CF images for presence of lesions which are associated with DR or AMD\textsuperscript{105,107,150,151,155,163,235,263,264}. Lesions associated with DR include microaneurysms, hemorrhages, exudates and cotton wool spots, whereas for AMD, these include drusen. After fusing the information obtained from the individual lesion detections, a final decision is made and in case of a positive finding, the patient is referred to an ophthalmologist for a precise diagnosis and follow-up procedure\textsuperscript{175}.

To make a correct decision for patient referral, it is evidently important to identify which types of lesions are present. On CF images, drusen have a similar bright appearance as exudates and cotton wool spots. Whereas cotton wool spots can be more easily identified by their generally larger size and different color appearance, exudates and drusen have very similar characteristics and differentiation is difficult. Following the criteria of international grading schemes for DR and AMD, a patient referral decision is different when a patient presents with drusen or exudates\textsuperscript{15,31,265}. A patient can have a few drusen present without the need for a direct referral to an ophthalmologist, whereas the presence of only a few exudates is an indication for more severe levels of DR and the patient should be referred immediately.

As the underlying cause and disease mechanisms for DR and AMD are different, exudates and drusen appear with different location patterns in the retina. Exudates are caused by leaking fatty deposits from blood vessels and appear in compact groups,
whereas drusen are believed to be a result of a reduced capacity of the retina to cleanse waste products from the photoreceptors and can appear over the whole retina. In our previous work we have shown that including contextual information is beneficial for individual lesion detection.\textsuperscript{164} We believe that including spatial information is an important factor to differentiate between drusen and exudates.

In this work, we therefore propose an automatic classification scheme that employs spatial information to both identify and discriminate between images containing drusen and exudates. The method makes use of information of individual lesions as detected by three separate systems focusing on the detection of red lesions, bright lesions in general, and drusen in particular. We introduce a concentric spatial pyramid approach with a twofold purpose: 1) to incorporate spatial information in the classification step and 2) to structurally transform lesion-level information into image-level information. Features encoding local lesion load are computed in an increasingly fine concentric spatial grid and used in combination with a random forest classifier to make the final three-class classification. We compare our system performance with that of two independent human graders.

5.2 Materials and Methods

The automatic method to identify and differentiate between drusen and exudates consists of four steps. First, to standardize image resolution across different datasets, the field of view is automatically extracted and resized to 650 pixels in diameter. Next, three lesion detection and classification algorithms are applied to the images for the detection and classification of: (1) bright appearing lesions, i.e. hard exudates, cotton wools spots and drusen; (2) drusen; and (3) red lesions, i.e. microaneurysms and hemorrhages. In the third step, results of these systems, consisting of the detected lesions and their associated posterior probability of being a true lesion, are combined. After removing lesions with a low posterior probability, the detected lesions are used as inputs to the spatial pyramid framework by creating histograms encoding the lesion load for each of the different type of detected lesions. Red lesion information is also included as the presence of red lesions is considered to be an indicator of DR, and therefore bright appearing lesions are more likely to be exudates. Finally, a multi-class classification is obtained by using a random forest classifier in a one-versus-all classification scheme using the spatial pyramid features. Using such a general framework allows to incorporate contextual information for differentiation between lesion types and allows to transform lesion-level information into image-level information.
5.2 Materials and Methods

5.2.1 Study Dataset

For this study, images were taken from several public datasets as well as from private datasets. In total, 130, 89, 397, 1200 and 569 images were taken from DiaretB0, DiaretB1, Stare, Messidor and DR1/DR2 datasets, respectively. After adding 488 and 1777 images, which were consecutively selected from the European Genetic Database (EUGENDA) and from the diabetic screening database of KSYOS Tele-Medical Center, we obtained a total of 4650 images. Images from the DiaretB0 and DiaretB1 dataset were taken at 50° field of view with a resolution of 1500x1152 pixels, while the type of camera is not reported. Images from the Stare dataset were taken with a TopCon TRV-50 fundus camera at 35° field of view and subsequently digitized at 605x700 pixels in resolution. Images from the Messidor dataset were captured with a Topcon TRC NW6 non-mydriatic retinograph with a 45° field of view and had a resolution of 1440x960, 2240x1488 or 2304x1536 pixels. Images from the DR1/DR2 dataset were captured using a Topcon TRC50X mydriatic camera with a 45° field of view and 640x480 pixel resolution. Images from the EUGENDA dataset were taken with either a Topcon TRC501X at 50° field of view or with a Canon CR-DGi at 45° field of view. Resolution of the EUGENDA images varied between 1360x1024 and 3504x2336 pixels. Images from the DR screening dataset had a resolution varying between 1024x768 and 4992x3328 pixels. Camera type and field of view were not reported for this set.

![Example images](https://via.placeholder.com/150)

Figure 5.1: Examples of color fundus images. (a): Control case, (b): image showing the presence of drusen and (c): image showing the presence of exudates.

Images which were not macula centered, had insufficient quality, or contained other bright appearing abnormalities such as myelinated nerve fibers or abnormalities not related to DR or AMD were discarded from the study dataset. The number of discarded images for each of these criteria was 1002, 270 and 302, respectively. The remaining set of images was graded by a human expert, who had over ten years of experience, into one of following categories: control, containing only drusen, containing only exudates, containing both.

---

1Kindly provided by the Messidor program partners (see [http://messidor.crihan.fr](http://messidor.crihan.fr))
2[http://www.eugenda.org](http://www.eugenda.org)
3[http://www.ksyos.org](http://www.ksyos.org)
or containing both drusen and exudates. All images containing either drusen (N=362) or exudates (N=186) as graded by the expert were included and complemented with randomly selected images graded as control (N=552), resulting in a final study dataset of 1100 images of 1100 eyes. Figure 5.1 shows example images of the three different classes. Gradings of the human expert were considered as the reference standard in this study. Table 5.1 shows the distribution of the images among the different data sources. Two independent human observers, referred to as Observer 1 and Observer 2, also graded this set and rated every image into one of three categories: control, containing drusen or containing exudates. Observer 1 and Observer 2 had four and six years of experience, respectively. In addition, Observer 1 also identified subtle cases within the study dataset, where his/her confidence level in discriminating the underlying disease was low. Image quality, image focus, lesion size and lesion number influenced the level of confidence of the observer.

Table 5.1: Distribution of images across different data sources. Only macular centered images with sufficient grading quality showing presence of only drusen or exudates were included and complemented with random control images.

<table>
<thead>
<tr>
<th></th>
<th>DB0</th>
<th>DB1</th>
<th>Eugenda</th>
<th>KSYOS</th>
<th>Messidor</th>
<th>Stare</th>
<th>DR1/DR2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>18</td>
<td>10</td>
<td>47</td>
<td>142</td>
<td>306</td>
<td>5</td>
<td>24</td>
<td>552</td>
</tr>
<tr>
<td>Drusen</td>
<td>2</td>
<td>1</td>
<td>163</td>
<td>92</td>
<td>68</td>
<td>21</td>
<td>15</td>
<td>362</td>
</tr>
<tr>
<td>Exudates</td>
<td>25</td>
<td>14</td>
<td>1</td>
<td>12</td>
<td>99</td>
<td>25</td>
<td>10</td>
<td>186</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>25</td>
<td>211</td>
<td>246</td>
<td>473</td>
<td>51</td>
<td>49</td>
<td>1100</td>
</tr>
</tbody>
</table>

DB0: DiaretB0; DB1: DiaretB1.

5.2.2 Preprocessing

In a preprocessing step, all images were resized to have the same field of view diameter of 650 pixels. No correction of illumination or shade correction was applied to the images. The optic disk location and fovea location were automatically detected using a previously developed algorithm\textsuperscript{100}, and were used later in the spatial pyramid framework.

5.2.3 Automatic lesion identification

Three types of lesions were automatically detected using previously developed algorithms: (1) Bright appearing lesions, i.e. hard exudates, cotton wool spots and drusen\textsuperscript{155}; (2) drusen\textsuperscript{235}; and (3) red lesions, i.e. microaneurysms and hemorrhages\textsuperscript{266}. The systems generally consisted of a candidate extraction step, followed by a candidate classification step. Features based on shape, color and intensity were extracted from each candidate and used in a supervised classification framework to classify each individual
candidate as being a true lesion. Table 5.2 shows a brief overview of the different features used for lesion classification. Candidates which overlapped with the optic disc location were automatically discarded in these previous works using an automatic optic disc detection system\textsuperscript{100,269}. More details on the individual system’s implementations can be found following the provided references\textsuperscript{155,235,266}. The outputs of these systems are the detected bright appearing lesions, drusen and red lesions, respectively, and their associated posterior probability of being a true lesion. Figure 5.2 shows examples of the individual system outputs.

Table 5.2: Features for classification of lesion candidates as used the works for bright appearing lesion detection, drusen detection and red lesion detection\textsuperscript{155,235,266}.

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape</td>
<td>Area, perimeter, compactness, length and width of the lesion candidate.</td>
</tr>
<tr>
<td>Context</td>
<td>Average and standard deviation of vessel pixel probability at the lesion candidate border. Distance to the closest lesion candidate. Number and average pixel probability of neighboring lesion candidates within a certain radius.</td>
</tr>
<tr>
<td>Intensity</td>
<td>Features measuring the difference in intensity in the RGB color space of the lesion candidate as compared to its surrounding area. Mean and standard deviation of Gaussian filter bank outputs.</td>
</tr>
<tr>
<td>Color</td>
<td>Average and standard deviation inside and outside the lesion candidate using the planes of the Luv and HSI color space.</td>
</tr>
<tr>
<td>Misc.</td>
<td>Average, standard deviation, maximum and median pixel probability inside the lesion candidate.</td>
</tr>
</tbody>
</table>


5.2.4 Feature descriptor

Using the three lesion type probabilities, an estimate of the lesion load was made. This was done by constructing a weighted histogram of the detected lesions taking the size of the lesion into account. The size of the detected lesions was included as exudates are in general smaller in size compared to drusen. Given the probability $P_i$ of the detected lesion $i$, the value $h_n$ of the histogram bin $n$ is defined as:

$$h_n = \sum_{i \in L_n} P_i$$  (5.1)

where $L_n$ is the group of lesions whose size is $n$ pixels, calculated as the smallest enclosing diameter of the lesion. The number of bins was set to 25 as most lesions have sizes smaller than 25 pixels. The last bin also takes into account lesions with size larger
Figure 5.2: Examples of images with the outputs of the individual automatic lesion detection systems. Top row: (a): Control case, (b): image showing the presence of drusen and (c): image showing the presence of exudates. Second row: output of the bright lesion detection system\textsuperscript{155}. Third row: output of the red lesion detection system\textsuperscript{266}. Bottom row: output of the drusen detection system\textsuperscript{235}. In each of the lesion detection output maps, a brighter color indicates a higher likelihood of being a true lesion.

than 25 pixels. To remove false positive detections by the individual lesion identification systems, all bright appearing lesions and drusen with $P_i$ smaller than 0.5 and all red lesions with $P_i$ smaller than 0.3 were neglected during construction of the weighted
histograms. The values for these thresholds were determined in pilot experiments with each of the individual lesion detection systems.

5.2.5 Spatial pyramids

The image was divided into increasingly coarser concentric circular regions. With increasing coarseness, i.e. level, the number of concentric regions was doubled. Each circular region was centered on the automatically detected fovea location. At the 0th level, the radius of the circular region was equal to twice the distance between the automatically detected fovea and the optic disc center. This radius was chosen in such a way that, for most images, the complete field of view of the image fell inside the 0th level region. At each increasing level, the number of regions was doubled with equally spaced radii. See Figure 5.3 for an example of the regions at different levels.

For each region, three lesion encoding weighted histograms, Equation 5.1, were computed including only lesions whose center of gravity fell inside the region. Histograms for each of the lesion types were concatenated, resulting in a feature vector of 75 features.
per region. For higher level regions, these regions can be seen as hollow circles, i.e. a lesion was included only once in a weighted histogram. Finally, one image encoding feature vector was composed for each image following either a single-scale approach or a multi-scale approach. In the single-scale approach, all feature vectors of the different regions at the level of evaluation were concatenated. In the multi-scale approach, all feature vectors of regions at the current level and the feature vectors of the regions at the previous, lower levels, were concatenated. Figure 5.4 shows a graphical representation of the image feature vector computed at spatial pyramid level 0.

![Graphical representation of the image feature vector at the 0th level spatial pyramid.](image)

**Figure 5.4:** Graphical representation of the image feature vector at the 0th level spatial pyramid. (a): Feature vector of the image shown in Figure 5.2(a), (b): feature vector of the image shown in Figure 5.2(b) and (c): feature vector of the image shown in Figure 5.2(c).

### 5.2.6 Multi-class differentiation of controls, drusen and exudates

Classification of the image into either control, containing drusen, or containing exudates was done using a one-versus-all classification scheme using the constructed image feature vector and a random forest classifier. In each of the one-versus-all classifications, 100 decision trees with a tree depth of 18 were used for the random forest classifier. All experiments were performed in a 10-fold cross-validation scheme. The total dataset was split up into 10 subsets. For each of the 10 folds, the random forest classifiers in the one-versus-all classification scheme were trained using 9 subsets and the left out subset was classified. Rotating this scheme 10 times resulted in classification of all subsets, without training and testing on the same data. All classified subsets were then pooled back together and performance was calculated by assigning the label of the class with the highest assigned probability. Receiver Operating Characteristic (ROC) analysis was performed using the assigned soft-probabilities for each of the individual classes in a one-versus-all classification scheme. Area (Az) under the ROC curve and 95% confidence intervals were calculated using bootstrap analysis\(^{271}\).
5.3 Results

5.3.1 Observer grading

Two independent observers independently graded all images into one of the three categories: control, drusen or exudates. Results of the human observers are reported in a contingency table, shown in Table 5.3.

Table 5.3: Contingency table showing the results for the human observer gradings.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Observer 1</th>
<th></th>
<th>Observer 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Drusen</td>
<td>Exudates</td>
<td>Control</td>
</tr>
<tr>
<td>Control</td>
<td>479</td>
<td>55</td>
<td>18</td>
<td>404</td>
</tr>
<tr>
<td>Drusen</td>
<td>33</td>
<td>302</td>
<td>27</td>
<td>13</td>
</tr>
<tr>
<td>Exudates</td>
<td>1</td>
<td>2</td>
<td>183</td>
<td>0</td>
</tr>
</tbody>
</table>

Observer 1 had an overall accuracy of 0.876 and kappa agreement of 0.802 with 95% confidence interval (CI) of [0.771;0.833], whereas observer 2 had an overall accuracy of 0.835 and kappa agreement of 0.740 [0.706;0.775]. Eyes that were graded as control by the reference and as having drusen by the observers contributed mostly to the errors made by the observers. Individual sensitivity/specificity pairs for differentiating controls, drusen and exudates were computed in a one-versus-all fashion and are reported in Table 5.6.

5.3.2 Automatic differentiation between controls, drusen and exudates

System performance was measured by applying two spatial pyramid schemes: one using single-scale analysis and one using multi-scale analysis. Table 5.4 shows the results for the single- and multi-scale spatial pyramid approaches. Kappa agreement and overall accuracy is reported.

Table 5.4: Classification performance of the automatic system using single- and multi-scale spatial pyramid approaches.

<table>
<thead>
<tr>
<th>Pyramid level</th>
<th>Single-scale</th>
<th></th>
<th>Multi-scale</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kappa [95% CI]</td>
<td>Accuracy</td>
<td>Kappa [95% CI]</td>
<td>Accuracy</td>
</tr>
<tr>
<td>0</td>
<td>0.666 [0.626;0.705]</td>
<td>0.801</td>
<td>0.666 [0.626;0.705]</td>
<td>0.801</td>
</tr>
<tr>
<td>1</td>
<td>0.665 [0.625;0.705]</td>
<td>0.800</td>
<td>0.664 [0.624;0.703]</td>
<td>0.799</td>
</tr>
<tr>
<td>2</td>
<td>0.635 [0.594;0.676]</td>
<td>0.785</td>
<td>0.663 [0.623;0.703]</td>
<td>0.799</td>
</tr>
<tr>
<td>3</td>
<td>0.608 [0.566;0.651]</td>
<td>0.772</td>
<td>0.663 [0.623;0.703]</td>
<td>0.799</td>
</tr>
</tbody>
</table>

CI: Confidence interval.

Figure 5.5 shows the ROC curves for each of the one-versus-all classifications. The operating points of both human observers are added to each graph. The CAD system
achieved an area Az under the ROC curve of 0.929, 0.883 and 0.956 for the detection of controls, images containing drusen and images containing exudates, respectively. Table 5.5 shows the contingency table for the automatic system for differentiation between controls, drusen and exudates at the 0th pyramid level. Note that the results for the 0th level was the same for the single- and multi-scale approaches. The multi-class approach achieved a kappa agreement of 0.666 [0.626;0.705] and accuracy of 0.801. Table 5.6 shows the sensitivity and specificity of the automatic system for identifying controls, drusen or exudates. The 0th level spatial pyramid approach achieved sensitivity/specificity pairs of 0.904/0.781, 0.699/0.896 and 0.694/0.976 for controls, drusen and exudates, respectively.

Table 5.5: Contingency table showing the results for the multi-class automatic system differentiation between controls, drusen and exudates using features of the 0th pyramid level.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Control</th>
<th>Drusen</th>
<th>Exudates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>499</td>
<td>42</td>
<td>11</td>
</tr>
<tr>
<td>Drusen</td>
<td>98</td>
<td>253</td>
<td>11</td>
</tr>
<tr>
<td>Exudates</td>
<td>22</td>
<td>35</td>
<td>129</td>
</tr>
</tbody>
</table>

5.3.3 Evaluation of influence of image subtlety

Observer 1 indicated for each image whether he/she was confident about his/her grading. Image gradings of which observer 1 was less confident were marked as subtle cases (N=582), whereas others were marked as clear cases (N=518). Table 5.7 shows the performance of the observers and the spatial pyramid method on the subset of subtle and clear cases. Total number of subtle cases was 292, 226 and 64 for normal, drusen and exudates, respectively. For clear cases, these numbers were 260, 136 and 122 for normal,
Table 5.6: Performance of the observers and the automatic spatial pyramid approach using features of the 0th pyramid level.

<table>
<thead>
<tr>
<th>Grader</th>
<th>Accuracy</th>
<th>Kappa [95% CI]</th>
<th>Control</th>
<th>Drusen</th>
<th>Exudates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs 1</td>
<td>0.876</td>
<td>0.802 [0.771;0.833]</td>
<td>0.868/0.938</td>
<td>0.834/0.923</td>
<td>0.984/0.951</td>
</tr>
<tr>
<td>Obs 2</td>
<td>0.835</td>
<td>0.740 [0.706;0.775]</td>
<td>0.732/0.976</td>
<td>0.939/0.798</td>
<td>0.941/0.979</td>
</tr>
<tr>
<td>System</td>
<td>0.801</td>
<td>0.666 [0.626;0.705]</td>
<td>0.904/0.781</td>
<td>0.699/0.896</td>
<td>0.694/0.976</td>
</tr>
</tbody>
</table>

Obs 1: Observer 1, Obs 2: Observer 2, System: automatic system.
Control, Drusen, Exudates: Sensitivity/specificity pair for identification of the corresponding class using one-vs-all approach.
CI: Confidence interval.

drusen and exudates, respectively. Figure 5.6 shows the ROC curves obtained using only the clear cases and the 0th level spatial pyramid. The CAD system achieved Az values of 0.997, 0.973 and 0.978 for the detection of controls, images containing drusen and images containing exudates, respectively.

Table 5.7: Performance of the observers and the proposed spatial pyramid approach on the subset of subtle (N=582) and clear (N=518) cases.

<table>
<thead>
<tr>
<th>Grader</th>
<th>Accuracy</th>
<th>Kappa [95% CI]</th>
<th>Control</th>
<th>Drusen</th>
<th>Exudates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subtle cases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obs 1</td>
<td>0.784</td>
<td>0.645 [0.590;0.700]</td>
<td>0.760/0.883</td>
<td>0.765/0.848</td>
<td>0.953/0.927</td>
</tr>
<tr>
<td>Obs 2</td>
<td>0.734</td>
<td>0.564 [0.505;0.623]</td>
<td>0.558/0.955</td>
<td>0.925/0.640</td>
<td>0.859/0.973</td>
</tr>
<tr>
<td>System</td>
<td>0.687</td>
<td>0.444 [0.377;0.511]</td>
<td>0.842/0.590</td>
<td>0.549/0.865</td>
<td>0.469/0.971</td>
</tr>
<tr>
<td><strong>Clear cases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obs 1</td>
<td>0.981</td>
<td>0.969 [0.950;0.988]</td>
<td>0.988/1.000</td>
<td>0.949/0.992</td>
<td>1.000/0.982</td>
</tr>
<tr>
<td>Obs 2</td>
<td>0.950</td>
<td>0.921 [0.891;0.950]</td>
<td>0.927/1.000</td>
<td>0.963/0.945</td>
<td>0.984/0.987</td>
</tr>
<tr>
<td>System</td>
<td>0.929</td>
<td>0.886 [0.850;0.921]</td>
<td>0.973/0.996</td>
<td>0.949/0.924</td>
<td>0.811/0.982</td>
</tr>
</tbody>
</table>

Obs 1: Observer 1, Obs 2: Observer 2, System: automatic system.
Control, Drusen, Exudates: Sensitivity/specificity pair for identifying the corresponding class.
CI: Confidence interval.

5.4 Discussion

Correctly identifying the type of bright lesions is important for correct disease diagnosis in a screening setup as presence of drusen or exudates result in different referral criteria and might alter automatic referral decision of a patient. An automatic method using information of individual lesion detections in a global spatial pyramid framework to identify and differentiate between drusen and exudates was presented and evaluated in a heterogeneous dataset consisting of images from multiple sources and cameras. Using spatial information, the automatic method approaches the level of performance of human observers. The automatic system achieved an overall accuracy of 0.80 and kappa agreement of 0.67 with the reference. Observers achieved slightly higher performances
Automatic differentiation of drusen and exudates

Figure 5.6: Receiver operating characteristics curves for each of the one-vs-all classifications using the set consisting of clear cases (N=518). Area (Az) under the curve and 95% confidence intervals are reported in the legend. (a): Control versus all, (b): drusen versus all and (c): exudates versus all.

with accuracies of 0.88 and 0.84 and kappa agreements of 0.80 and 0.74, respectively. Higher level spatial pyramids, in which the image was divided into more concentric sub-regions, did not lead to an increase in performance. Using multi-scale analysis is beneficial over single-scale analysis when using higher levels of spatial pyramids, but does not increase overall performance.

In our proposed framework an image level decision was generated by combining individual lesion detections in the whole image. Other methods have used a rule-based approach where the image level score was determined by the maximum lesion probability of the target class. To investigate the added value of our approach over such maximum rule based approach, we have applied the bright lesion detection system and the drusen detection system in our dataset. The bright lesion detection system achieved Az values of 0.906, 0.702 and 0.804 for the identification of controls, cases with drusen and cases with exudates, respectively. The drusen detection system achieved an Az value of 0.711 for the identification of cases with drusen. Our proposed method achieved substantially higher performance, with Az values of 0.929, 0.883 and 0.956 for controls, drusen and exudates, respectively. By not using only the maximum lesion detection response for the image score, our method is more robust to outliers and confounding lesions present in the data (i.e. drusen are confounding lesions when identifying cases with exudates and vice versa).

From Table 5.3 and Table 5.6, it can be observed that both observers achieved good performance for the identification of each of the individual classes. Control images which were graded as having drusen contributed most to the errors made by both observers. This can also be seen in the sensitivity for identifying control cases and specificity for drusen cases, which is slightly lower than the ones for the other classes for both observers, see Table 5.6. A possible explanation for this is that drusen can be very subtle and small reflections in the retina can be misinterpreted as drusen. For
exudates, this poses less of a problem as exudates in general have a higher contrast with the background. The proposed spatial pyramid approach achieved a sensitivity of 0.904 for identification of normal cases with a slightly lower specificity than observers. Lower sensitivities of the automatic system compared to observer for the classes drusen and exudates can be observed in Table 5.6.

One reason for the lower sensitivities of the automatic detection of images with drusen or exudates is the fact that the three automatic systems that detected various lesions are not perfect. The outputs of these systems are used in the spatial pyramid framework. Lesions which are not detected by these systems will not be incorporated in the spatial framework and hamper thus the ability of the automatic system to make a correct decision. This is reflected in the larger number of images assigned to the control class, while being a drusen or exudates case, see Table 5.5, and the lower sensitivities for identification of drusen and exudates cases, see Table 5.6.

Lowering the threshold for the lesions to be included in the framework increases the chance of subtle lesions to be included in the spatial pyramid framework, but also contributes to a larger number of false positive detections. The thresholds were determined using pilot experiments with each of the individual lesion detection systems. Additional analysis of the influence of these system hyper-parameters on the final performance of our proposed system showed similar performance. Table 5.8 shows the accuracy values obtained by the 0th pyramid level CAD system. A change in threshold values led to only minor changes on the overall CAD system’s accuracy. However, these thresholds might not be optimal for the identification of the individual classes, such as drusen or exudates.

Table 5.8: Accuracy values obtained using the 0th pyramid level CAD system using different threshold settings for discarding bright appearing lesion, drusen and red lesion information. Threshold 1: threshold for discarding bright appearing lesions and drusen; Threshold 2: threshold for discarding red lesions.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Threshold 1</th>
<th>Threshold 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.795</td>
<td>0.804</td>
</tr>
<tr>
<td>0.1</td>
<td>0.792</td>
<td>0.796</td>
</tr>
<tr>
<td>0.2</td>
<td>0.791</td>
<td>0.797</td>
</tr>
<tr>
<td>0.3</td>
<td>0.799</td>
<td>0.809</td>
</tr>
<tr>
<td>0.4</td>
<td>0.799</td>
<td>0.803</td>
</tr>
<tr>
<td>0.5</td>
<td>0.800</td>
<td>0.812</td>
</tr>
<tr>
<td>0.6</td>
<td>0.804</td>
<td>0.812</td>
</tr>
</tbody>
</table>

Another reason for the fact that the automatic system is performing worse than human experts may be the large amount of subtle cases in the dataset. Many images only contain a single or only a few drusen or exudates. Figure 5.7 shows examples of
Automatic differentiation of drusen and exudates

subtle cases included in the dataset. Images presenting with only a single subtle lesion are more likely to be missed by the automatic lesion detection systems. Furthermore, based on a single lesion, it is hard or even impossible for the automatic system to make a differentiation between drusen or exudates using spatial information. Table 5.5 shows the number of images correctly and incorrectly classified by the automatic system. To analyze the influence of subtlety of the data on the system performance, we evaluated performance on the subset of images marked as clear cases by Observer 1. System performance and also observer performance is much higher in this subset compared to the subtle subset, Table 5.7, reaching an accuracy of 0.93 and kappa agreement of 0.89. The ability to detect and differentiate normal, drusen and exudates cases appears to be heavily influence by the subtlety of the images under consideration.

In our study, we have excluded images if they contained myelinated nerve fibers or abnormalities which are not related to AMD or DR. However, in a screening setup, such images can be encountered. Therefore, we have investigated how our system would respond in such a case. Figure 5.8 shows an example of an image containing myelinated nerve fibers. The 0th level spatial pyramid classified this image as containing exudates. Although this classification might not be correct, the system has identified that the image does not belong to the control class. Although identification of this image as not being a control is important in a screening setting, an automated system which identifies multiple eye abnormalities would lead to a more accurate patient referral and treatment decision.

Table 5.6 and Table 5.7 show that the incorporation of spatial information through spatial pyramids did not improve the final performance. The value of using higher spatial pyramids might not be well expressed for this task, but could potentially increase performance for other related classification tasks. In retinal screening, not only the type of lesions, but also the location of lesions is important to make a decision on whether a patient should be referred. Lesions close to the macular region represent a higher risk of disease progression for the patient and, therefore, influence the decision for referral. Higher level spatial pyramids could improve the performance as compared to the 0th level as higher levels are capable of capturing this spatial information in the image. Another application where higher level spatial pyramids might prove to have added value is for disease severity grading. In the case of AMD severity grading, the level of severity is based on the location of the abnormalities with respect to the fovea. The 0th level spatial pyramid is unable to encode this information as all lesions in the image are encoded in the same weighted histogram regardless their position. Higher level spatial pyramids are, however, able to capture this spatial information: the concentric circles of the higher level spatial pyramids are centered on the fovea and each ring encodes information at a fixed distance from the fovea. The property of encoding
Figure 5.7: Examples of images showing subtle bright lesions. (a): image showing the presence of drusen with (b) showing a close up view of the region as indicated by the black box in (a). (c): image showing the presence of exudates with (d) showing a close up view of the region as indicated by the black box in (b). Both images (a) and (c) were graded by the automatic system as control.

spatial information might have a higher impact on such classification tasks. We will further investigate the impact of higher level spatial pyramids for these tasks in future work.

We have used the ratings of a single expert as reference, mimicking the clinical protocol where a single observer grades the images. However, as the availability of multiple ratings for each image allowed for construction of a potentially more robust reference, we also examined the use of a consensus reference, constructed by using majority voting of the three observers. Training and evaluating the system using this
The choice of a suitable classifier is difficult to make in advance as the classifier performance depends on the dataset and the classification task at hand. We have also compared other classifiers such as a support vector machine and a k-nearest neighbor classifier, obtaining an accuracy of 0.736 and 0.696 for the 0th level spatial pyramid, respectively. The random forest classifier obtained higher performance with a moderate computational cost of 12 seconds for training and classification, computed on an Intel Xeon CPU with 2.4 GHz memory.

In literature, other methods which have focused on the identification of either drusen or exudates have been described\(^{105,107,150,151,163,263,264}\). In these works, the focus was put on discriminating control cases from abnormal cases, i.e. cases containing either abnormalities related to AMD or abnormalities related to DR. The datasets for this task consisted of control cases and cases with either exudates or drusen, but not simultaneously. Features to differentiate between exudates and drusen need to be more robust as these abnormalities might have similar appearance on color fundus images, making differentiation more difficult. This difficulty is also reflected in Figure 5.6 which shows that the CAD system achieves higher performance for the identification of controls with an Az value of 0.997, as compared to the identification of drusen (Az=0.978) or exudates (Az=0.973).

Previous works also described methods to discriminate between images containing bright lesions. A method, making use of sparse coded features and a support vector machine achieved near perfect discrimination results\(^{161}\). However, it should be noted
that a fair comparison between methods is hard to make since results are based on different datasets. The composition of the datasets has a large influence on algorithm performance, as is evident from the difference in performance between subtle and clear cases in our experiments (see Table 5.7). In our study, images showing presence of single drusen or exudates were included. Based on a single lesion, it can be difficult or even impossible for an algorithm to make a differentiation. Reference criteria for data inclusion in the previous study\textsuperscript{161} were not mentioned by the authors. Furthermore, no human observer was used in the previous study, which would allow to compare system performance to human level performance, giving a measure of overall performance. Including human observers is therefore important to draw conclusions on algorithm performance.

Other studies have proposed frameworks based on bag-of-visual-words (BoVW) representation which was used to detect bright lesions and red lesions\textsuperscript{158,272,273}. In these approaches, a visual vocabulary is constructed using features extracted from the color fundus images. These methods achieve good performance on specific datasets. An evaluation of these methods on a diverse dataset consisting of image from multiple centers and cameras has not been performed. Furthermore, these methods have many system parameters which need to be tuned in order to obtain good results. These parameters include the number of visual words for image description and the manual crafted features used to map local image patches to one of the visual words. Furthermore, interpretation of the system is difficult as the visual vocabulary cannot be mapped directly to individual lesions or image characteristics.

In this work, we have combined features extracted from solely color fundus images. Adding information from different imaging modalities could, however, increase the CAD performance. For example, pathologies related to AMD and DR manifest at different retinal depths and affect different layers of the retina. As Optical Coherence Tomography (OCT) provides valuable depth information, features derived from OCT could potentially be beneficial for disease differentiation. Previous works employing OCT imaging data have shown good performance for the automatic differentiation of retinal diseases\textsuperscript{274,275}. A method incorporating histogram of oriented gradient features and a support vector machine correctly identified 100% of the cases with AMD, 100% of the cases with diabetic macular edema, and 86.7% of the normal subjects in a small study dataset of 15 subjects for each class\textsuperscript{274}. Another work focused on the identification of different macular pathologies: macular hole, macular edema and AMD\textsuperscript{275}. This method uses multiscale texture features and shape features in combination with support vector machines to identify control cases and each of the pathologies separately. The method obtained Az values ranging between 0.941 and 0.978 in a dataset comprising of 131 scans from 37 subjects. Extending our proposed framework by adding features
derived from OCT data could therefore potentially increase the classification performance. However, in current screening settings, only color fundus imaging is performed, which limits the inclusion of OCT derived features in our proposed framework.

To conclude, we presented an automatic system for the identification and differentiation of images containing drusen or exudates using a spatial pyramid framework. We provide a general framework to generate an image level decision based on individual lesion detections and we show that the automatic system approaches human level performance, although results do not improve when using higher level spatial pyramids. System performance is limited by the subtlety of the images in the dataset and leaves room for improvement. Improvements could include the addition of lesion level characteristics, such as color and shape in the spatial pyramid approach.
Hemorrhage detection using convolutional networks with a fast training scheme

Mark J. J. P. van Grinsven, Bram van Ginneken, Carel B. Hoyng, Thomas Theelen, Clara I. Sánchez

*Original title:* Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images

*Published in:* Transactions on Medical Imaging (2016); 35(5):1273-1284
Abstract

Convolutional neural networks (CNNs) are deep learning network architectures that have pushed forward the state-of-the-art in a range of computer vision applications and are increasingly popular in medical image analysis. However, training of CNNs is time-consuming and challenging. In medical image analysis tasks, the majority of training examples are easy to classify and therefore contribute little to the CNN learning process. In this paper, we propose a method to improve and speed-up the CNN training for medical image analysis tasks by dynamically selecting misclassified negative samples during training. Training samples are heuristically sampled based on classification by the current status of the CNN. Weights are assigned to the training samples and informative samples are more likely to be included in the next CNN training iteration. We evaluated and compared our proposed method by training a CNN with (SeS) and without (NSeS) the selective sampling method. We focus on the detection of hemorrhages in color fundus images. A decreased training time from 170 epochs to 60 epochs with an increased performance – on par with two human experts – was achieved with areas under the receiver operating characteristics curve of 0.894 and 0.972 on two data sets. The SeS CNN statistically outperformed the NSeS CNN on an independent test set.
6.1 Introduction

Convolutional neural networks (CNNs) have been widely adopted in the field of computer vision\textsuperscript{116,117}. These models are based on convolution operations applied to the input image at multiple hierarchical layers. CNNs are very powerful because they can be trained end-to-end in a supervised manner and thus obviate the need to manually devise features, and have substantially outperformed the state-of-the-art for classification of natural images on large and well established databases\textsuperscript{247,250,276}. In medical image analysis, CNNs are also increasingly used. Their capability to learn a complex, hierarchical representation of the data makes CNNs useful to discern the complex disease specific patterns, difficult to be encoded by humans and by simpler traditional classifiers. Recent works on cancer detection and brain segmentation have shown CNN achieved remarkable performance\textsuperscript{259,277,278}. However, the need of large high-quality training sets to accurately train CNNs prevent a wider adoption of these networks in medical imaging.

CNN training process is a sequential process requiring many iterations (or epochs) to optimize the network parameters and learn discriminative features\textsuperscript{117}. In every epoch, a subset of samples is randomly selected from the training data and is presented to the network to update its parameters through back-propagation, minimizing a cost function. In this work we focus on finding diseased regions in images, a common task in medical image analysis. In such a classification task, CNNs are trained with small patches centered on pixels of interest. Although this results in vast training sets of image patches, the quality of the data is suboptimal: the normal class is extremely over-represented in this classification task and, moreover, the majority of normal training samples are highly correlated due to the repetitive pattern of normal tissues in each image. Only a small fraction of these samples are informative. Treating uniformly this data during the learning process leads to many training iterations wasted on non-informative samples, making the CNN training process unnecessarily time-consuming. An approach to identify informative normal samples will help to increase the efficiency of the CNN learning process and to reduce the training time.

Boosting techniques have been previously proposed to focus the learning process on informative samples in order to increase the performance of simple classifiers\textsuperscript{279}. These techniques create an ensemble of learners, each trained consecutively, where more emphasis is put on samples misclassified by the previous learners\textsuperscript{279,280}. Classification is performed by combining the outputs of each of the individual learners. A simplified version of the boosting strategy is a two-step approach in which misclassified samples of an initial model are used as the training set of a second, independent learner\textsuperscript{259,281}. The second learner, which is trained then with only informative samples, is used for
final classification. In general, boosting strategies rely on the optimization of different classifiers in cascade in order to discover informative samples (i.e. misclassified samples) for the next learner. Considering the high computational expense of CNN optimization, a boosted cascade of CNNs is inefficient, increasing the time complexity with the number of CNNs in the ensemble. In contrast to boosting techniques, dynamically sampling strategies focus the learner on informative samples during its optimization process, in order to simultaneously increase the classification performance and reduce training time. To achieve this, the training set is dynamically updated during the learning process of a single learner, putting more emphasis on informative samples\cite{282-284}. These dynamic sampling strategies have shown to reduce the training time and outperform boosting types of strategies\cite{283}. However, the challenge of these sampling techniques is defining a sampling heuristic optimal for the learner and the characteristics of the data and task at hand. To the best of our knowledge, the incorporation of a dynamic sampling strategy in the CNN learning process for medical image tasks has not been proposed yet.

In this paper, we propose an innovative sampling heuristic to identify informative training samples in a common medical image classification task, namely abnormality detection. The proposed heuristic will dynamically increase the probability of misclassified normal samples to be selected in each training iteration. We integrate this heuristic in the CNN learning process in order to increase its efficiency and reduce its training time, while maintaining its performance. The performance of the proposed method is
then validated in two large datasets for the detection and localization of hemorrhages on color fundus images. Hemorrhages are one of the visible signs on color fundus images of diabetic retinopathy (DR), a vision threatening disease affecting patients with diabetes\textsuperscript{14}. Figure 6.1 shows an example of a color fundus image including hemorrhages and typical confounding elements in hemorrhage classification.

Hemorrhage detection is of high importance for the automated detection and staging of DR, the most important cause of blindness in the working population. Whereas a lot of methods have been presented for the automated detection of micro-aneurysms in color fundus photographs, detection and segmentation of larger hemorrhages has received less attention\textsuperscript{84,285}. Hemorrhages and micro-aneurysms are mostly detected together and associated with a single label. In previous works, approaches based on morphological operations\textsuperscript{157}, wavelet operations\textsuperscript{286} and manual designed features in combination with statistical classifiers\textsuperscript{107,266,287,288} were used for the detection of hemorrhages and micro-aneurysms. Although hemorrhages are different in size and shape and pose different clinical relevance\textsuperscript{14}, only few works have addressed the identification of hemorrhages separately on color fundus images\textsuperscript{289,290}.

Section 6.2 provides a description of the different data sets used in this work. The proposed method and experimental design are described in detail in Section 6.3 and Section 6.4. In Section 6.5, the results are shown which are discussed in Section 6.6. Section 6.7 concludes the proposed work.

6.2 Materials

Two independent data sets were used in this study for training and evaluating the proposed method. 1) a subset of images from the "Diabetic Retinopathy Detection" challenge database from Kaggle\textsuperscript{1} and 2) images from the publicly available Messidor database\textsuperscript{2}.

Dataset description

Kaggle database

The Kaggle data set consist of 35,126 training images graded into five DR stages and 53,576 test images with undisclosed DR stage. Images were acquired using multiple fundus cameras and different field of view. Details about image acquisition, such as camera type and field of view, are not revealed. More information about the data can be found in the challenge website.

\textsuperscript{1}https://www.kaggle.com/c/diabetic-retinopathy-detection
\textsuperscript{2}Kindly provided by the Messidor program partners (see http://messidor.crihan.fr)
A subset consisting of 6,679 images was selected from the Kaggle training set. This subset consists of 4,450 randomly selected images from DR stage 0 (normal), 488 randomly selected images from DR stage 1 (mild), 1,058 randomly selected images from DR stage 2 (moderate) and 593 randomly selected images from DR stage 3 (severe). Images on which the retina was not visible were not included in this study dataset.

The selected 6,679 images were further split into a fixed training, monitoring and test set according to a 60-20-20 split. Images from the same patient were kept in the same subset.

**Messidor database**

The publicly available Messidor database consists of 1200 images acquired at three different sites. Images were acquired using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinograph with a 45 degree field of view. The images have resolutions of 1440x960, 2240x1488 or 2304x1536 pixels. More details about the database can be found in the corresponding website. The Messidor set will be exclusively used as an independent set for testing.

**Reference standard and observer annotations**

In this study, annotations were performed by three different independent observers, having 5 years, 12 years and over 15 years of experience, respectively. The first observer annotated and graded training, monitoring and test data. We referred to this observer as the reference observer. The two other observers (referred to as Observer 1 and Observer 2) graded only the test sets. These two observers were used to report human performance on the test data.

The reference observer indicated presence of hemorrhages on both the Kaggle and Messidor set. In the Kaggle set, this observer also annotated the center point of each individual hemorrhage in the training, monitoring and test sets. Furthermore, the reference observer indicated good or poor quality for each of the test images in both sets. An overview of the reference set can be seen in Table 6.1. No individual hemorrhage lesion annotations were performed in the Messidor set.

**6.3 Methods**

A dynamic CNN training strategy is presented where informative normal samples are dynamically selected at each training epoch from a large pool of medical images. A dynamic weight is assigned to each pixel in the negative training pool indicating its informativeness level. After each CNN training epoch, the weight of each negative
Table 6.1: Reference annotation statistics. Plus and minus signs indicate the number of positive and negative images, respectively. The numbers between brackets indicate the number of good quality images and number of hemorrhages on good quality images.

<table>
<thead>
<tr>
<th>Lesions</th>
<th>Training stage</th>
<th>Test stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Monitoring</td>
</tr>
<tr>
<td>+</td>
<td>655</td>
<td>224</td>
</tr>
<tr>
<td>-</td>
<td>3304</td>
<td>1104</td>
</tr>
</tbody>
</table>

training pixel is updated. This process is repeated until a stopping criterion is reached. The final trained CNN is used to classify each pixel in the test images, resulting in a pixel probability map for each test image.

**Preprocessing**

In a preprocessing step, the field of view of the color fundus images is segmented to limit the analysis of the CNN to the region of interest. Circular template matching is used to extract the field of view and images are cropped to the square bounding box of this circular field of view. Images are resized to 512 x 512 pixel dimension to reduce the computational costs and preprocessing was applied to improve image contrast. A contrast enhanced image $I_{ce}(x, y; \sigma)$ is obtained as follows:

$$I_{ce}(x, y; \sigma) = \alpha I(x, y) + \beta G(x, y; \sigma) \ast I(x, y) + \gamma$$ (6.1)

where $\ast$ represents the convolution operator and $G(x, y; \sigma)$ a Gaussian filter with scale $\sigma$. Values of the parameters were empirically chosen as: $\alpha = 4$, $\beta = -4$, $\sigma = 512/30$ and $\gamma = 128$. The contrast enhanced image values are used as input for the CNN. Figure 6.2 shows an example image before and after applying the contrast enhancement step.

**Training data preparation and augmentation**

Images which do not contain any hemorrhage are defined as negative images, whereas images with hemorrhages are defined as positive images. To construct the CNN training data, pixels are extracted from these images, where negative pixels are extracted only from negative images and positive pixels are extracted only from positive images at hemorrhage locations. Corresponding training patches, centered on the extracted pixels, of size 41x41 pixels and 3 channels depth are created during the CNN training routine. The patch label is determined by the label of the central pixel. Data augmentation by spatial translation with one pixel in both horizontal and vertical direction and vertical and horizontal flipping is applied to the positive patches to artificially increase
the number of positives. Negative patches were also randomly flipped vertically and horizontally to counter for possible over-fitting.

**Network details**

The CNN architecture used in this study consists of five convolutional layers followed by rectified linear units (ReLUs) and spatial max-pooling. The final layers of the network consist of a fully connected layer and a final soft-max classification layer. Inspired by the OxfordNet which showed good performance for the classification of images of natural scenes, we use 32 small size filters of size 3x3 pixels in each convolutional layer. Max-pooling of size 2x2 and a stride of 2 is applied after the first two convolutional layers, halving the feature map sizes after the operations. Max-pooling reduces the number of free parameters and introduces small spatial invariance in the network. The fully connected layer consist of 1024 nodes followed by a soft-max logistic regression which outputs a score ranging between 0 and 1, indicating the probability of the pixel to belong to the positive class. Weight-decay of $5 \cdot 10^{-5}$ is added to each layer to penalize large weight parameters during back-propagation of the gradient in the optimization routine. Table 6.2 and Figure 6.3 show an overview of the network architecture with the omission of the ReLUs. All network parameters are randomly initialized according to a normal distribution with variance equal to 0.05. The CNN is trained using stochastic gradient descent with learning rate of $5 \cdot 10^{-5}$, minimizing a cost function $C$ defined as follows:
$$C(l, s) = - \sum_{i=0}^{B} l_i \log(s_i) + (1 - l_i) \log(1 - s_i)$$  \hspace{1cm} (6.2)

where \( s \) is the assigned pixel probability score, \( l \) the reference pixel label and \( B \) the total number of samples in one mini-batch. A mini-batch size of 256 patches is used and one epoch is defined as 4000 mini-batches. This means that around one million samples, of which half are positive and half are negative, are used in one epoch to train the CNN.

![Figure 6.3: Schematic overview of the CNN architecture containing convolutional layers, max-pooling layers, a fully connected layer and a soft-max logistic regression classification.](image)

Table 6.2: Architecture of the CNN. For each convolutional layer, the width x height x depth of the kernels is reported with the \( K \) number of kernels. In each max-pooling layer, 2x2 max-pooling is applied with stride \( a \) pixels.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Operation</th>
<th>Input size</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer 1</td>
<td>convolution</td>
<td>41x41</td>
<td>3x3x3, ( K = 32 )</td>
</tr>
<tr>
<td>Layer 2</td>
<td>max-pooling</td>
<td>39x39</td>
<td>2x2, ( a = 2 )</td>
</tr>
<tr>
<td>Layer 3</td>
<td>convolution</td>
<td>20x20</td>
<td>3x3x3, ( K = 32 )</td>
</tr>
<tr>
<td>Layer 4</td>
<td>max-pooling</td>
<td>18x18</td>
<td>2x2, ( a = 2 )</td>
</tr>
<tr>
<td>Layer 5</td>
<td>convolution</td>
<td>9x9</td>
<td>3x3x3, ( K = 32 )</td>
</tr>
<tr>
<td>Layer 6</td>
<td>convolution</td>
<td>7x7</td>
<td>3x3x3, ( K = 32 )</td>
</tr>
<tr>
<td>Layer 7</td>
<td>convolution</td>
<td>5x5</td>
<td>3x3x3, ( K = 32 )</td>
</tr>
<tr>
<td>Layer 8</td>
<td>fully connected</td>
<td>3x3</td>
<td>1024 nodes</td>
</tr>
<tr>
<td>Layer 9</td>
<td>soft-max</td>
<td>1024x1</td>
<td>2 classes</td>
</tr>
</tbody>
</table>
Selective sampling

At each CNN training epoch, a weight is assigned to each negative sample, proportional to their sampling probability: higher weight means a higher probability to be selected for the next epoch. In order to reduce the number of redundant samples in the training set, higher weights are assigned to representative samples. In this work, representative sample are considered those negative samples with a larger classification error at the current CNN state.

Given $X = \{(x_i, l_i)\}$ the set of $N$ training pixels $x_i$ and their corresponding reference label $l_i$ with $i = \{1, ..., N\}$, let $X_+$ and $X_-$ be the sets of positive and negative pixels:

$$X_+ = \{(x_i, l_i), \forall x_i \text{ with } l_i = 1\}$$
$$X_- = \{(x_i, l_i), \forall x_i \text{ with } l_i = 0\}$$

where $X = X_+ \cup X_-$.  

The proposed iterative algorithm for dynamically selecting training pixels to train a CNN $c$ follows these steps:

1. Initialize the sets of positive pixels $X_+^t \subset X_+$ and negative pixels $X_-^t \subset X_-$ by randomly selecting $M$ samples with replacement for each class from $X_+$ and $X_-$, respectively.

2. Train the network $c$ with $X^t = X_+^t \cup X_-^t$ using stochastic gradient descent.

3. Classify each pixel $x_i$ in $X_-$ with the trained network $c^t$. A pixel probability score $s^t_i$ is obtained for each $x_i$ in $X_-$ after classification.

4. Assign each $x_i$ in $X_-$ a weight $w^t_i = |s^t_i - l_i|$. A higher weight is assigned to those pixels of which the preliminary network prediction differs the most from the initial reference label.

5. Update $X_+^t$ and $X_-^t$ by selecting $M$ samples for each class. $x_i$ in $X_+^t$ is selected randomly while $x_i$ in $X_-^t$ is selected with probability $p^t_i$:

$$p^t_i = \frac{w^t_i}{\sum_{x_j \in X_-} w^t_j}$$

6. Train the network $c$ with $X^t = X_+^t \cup X_-^t$ using stochastic gradient descent.

7. Repeat steps (3)-(6) until a stopping criterion is reached.

In this proposed iterative selective sampling (SeS) algorithm, the pool of negative and positive training pixels is dynamically changed at each training epoch, preventing
the training process to focus on redundant negative samples and efficiently train the CNN. The parameter $M$ is not tunable by itself but is dependent on the mini-batch size and the number of mini-batches in one epoch. Changing the value of $M$ can be done by modifying either one of these two. In order to obtain a more efficient scheme, we consider applying step (3) and (4) once every five epochs.

**CNN training monitoring**

To determine when the CNN training process is completed and avoid over-fitting, the CNN performance during training is monitored during training on an independent monitoring set. Although the problem of over-fitting is countered by using different training pixels in each training epoch, an independent measure to determine when to stop the training procedure is still required. One way to measure performance is by measuring the pixel classification performance using the area (Az) under the Receiver Operating Characteristics (ROC) curve$^{120}$. However, Az values of pixel-based ROC curves are misleading due to the unequal distribution of positive and negative pixels. Therefore, we measure the Az value based on image classification performance. A score for each image is obtained by classifying all pixels in the image and considering the maximum pixel probability as the image score. When the Az value on the monitoring set has reached a stable maximum, determined after visual inspection, the CNN training phase is considered finished.

**Hemorrhage classification**

Given an unseen test input image, the CNN classifies every pixel in the image and returns a probability map indicating for each pixel the probability to belong to a hemorrhage. We postprocess this probability map to extract hemorrhage candidates and compute an image score describing the probability of the image to contain hemorrhages.

**Hemorrhage lesion identification**

The obtained pixel probability map is convolved with a Gaussian filter with scale $\sigma = 1$ to smooth the values. Candidate hemorrhage regions are identified by detecting local maxima in the smoothed pixel probability map. The local maxima locations are used as seed points for dynamic programming to segment the individual hemorrhage candidates$^{235}$. The dynamic programming algorithm is driven by a cost function computed as the gradient magnitude of the smoothed pixel probability map. The segmented candidate is assigned a final probability equal to the average of the pixel probabilities inside the candidate.
Identification of images with hemorrhages

To determine if an image contains hemorrhages, an image score is computed from the obtained pixel probability map. After the Gaussian smoothing step is applied to the pixel probability map, the maximum pixel probability is assigned as image score.

6.4 Experimental design

To compare the performance of the proposed SeS algorithm, a second CNN with the same network architecture was trained using the same pool of training images. However, at each training epoch random sampling of training pixels was performed. This means that each pixel has an equal chance of being used in the training procedure. This non-selective sampling (NSeS) CNN was also monitored using the same monitoring data set and the same stopping criteria. To train and monitor both CNNs, the Kaggle training and monitoring set were used, respectively. After both CNNs were trained, results were computed on the two test sets.

We evaluated the proposed SeS scheme by conducting the following experiments:

1. Evaluation of CNN performance during training: The Az value on the monitoring set was measured during CNN training for both the SeS and NSeS CNNs and the required number of training epochs was compared.

2. Evaluation of hemorrhage lesion identification: Free-response ROC (FROC) analysis was employed to compare the CNN performance for the detection of individual hemorrhages. In here, only false positives encountered on negative images were taken into account to prevent ambiguities in the reference annotation to influence the result. Additionally, there is no clinical relevance in a screening setting for false positives detected on images containing hemorrhages as these patients should be sent for referral. To determine if a hemorrhage was detected by the CNN, the distance between the manually annotated hemorrhage center location and the seed point of the segmented candidate was used, with a maximum tolerance of 8 pixels. This value was empirically determined using visual inspection of the average hemorrhage size on the 512x512 pixel resolution images. Detected regions which had no reference hemorrhage center location within this 8 pixel circular radius were considered as false positives.

3. Evaluation of identification of images with hemorrhages: ROC analysis was performed to evaluate the performance on identification of images with hemorrhages. Bootstrap analysis with 10,000 bootstraps was used to compute 95% confidence intervals for the Az values. A level of significance of \( \alpha = 0.05 \) was used for
6.5 Results

6.5.1 CNN performance during training

During training, performance was measured on the monitoring set during the CNN training process. Figure 6.4 shows the Az values measured on image level as function of the number of training epochs. The performance of the CNNs increased over time and finally converged to a stable maximum performance. For the SeS CNN, this maximum performance was achieved after 60 training epochs and for the NSeS CNN this maximum was achieved after 170 training epochs. Both these networks, i.e. the SeS CNN after 60 epochs (SeS CNN 60) and the NSeS CNN after 170 epochs (NSeS CNN 170) were used to compute hemorrhage detection results on the two independent test sets.

4. Evaluation of the influence of image quality: In order to assess the influence of image quality on the CNN performance, images graded by the reference observer as having poor image quality were removed from both test sets. CNN performance on image level was measured using ROC analysis and sensitivity, specificity and kappa agreement values were calculated.

Figure 6.4: Image-based Az values on the monitoring set for the SeS CNN and NSeS CNN over a number of training epochs. Shaded regions indicate the 95% confidence intervals of the Az values. After 60 and 170 epochs, the training phases of the SeS CNN and NSeS CNN were considered finished.

statistical comparison of the CNNs. Sensitivity, specificity and kappa agreement of the CNNs with respect to the reference were calculated at the operating point on the ROC curve closest to the upper left corner of the graph. These measures were also computed for Observer 1 and Observer 2.
Figure 6.5 shows example outputs of the SeS CNN and NSeS CNN after different numbers of training epochs as heat-map overlays on the example input image shown on top and in Figure 6.2a. After training for a small number of training epochs, both CNNs incorrectly classified all dark normal structures, such as vessels and fovea, but were able to correctly classify the normal background pixels. As CNN training continues, the CNNs learn to separate hemorrhages and the normal retinal structures are correctly classified as negative. For the SeS CNN, this learning process required less training epochs.

Figure 6.5: Pixel probability maps obtained by applying the SeS and NSeS CNNs to a sample image from the training set after training the network for a different number of epochs. Overlays are shown using a heat-map color coding, where red codes for high probabilities and blue for low probabilities. The SeS CNN required 60 epochs to reach final performance while the NSeS CNN required 170 epochs to reach final performance.
6.5 Results

Figure 6.6: Left column: example color fundus images from the Kaggle test set. Middle column: reference hemorrhage center locations. Right column: output of the SeS CNN 60.

**Hemorrhage lesion identification**

Figure 6.7 shows the FROC curves for the SeS CNN 60 and the NSeS CNN 170. The NSeS CNN after 60 epochs is included for direct comparison with the SeS CNN 60, showing a lower overall FROC curve compared to the SeS CNN 60 and NSeS CNN 170. At 1 false positive per normal image the SeS CNN 60 and NSeS CNN 170 achieve sensitivities of 0.786 and 0.753, whereas at 0.1 false positives per normal image, both CNNs achieve sensitivities of 0.511 and 0.316, respectively. In Figure 6.6, example images with annotated hemorrhage center locations and the outputs of the SeS CNN 60 networks are shown.

**Identification of images with hemorrhages**

Figure 6.8 shows the ROC curves of both CNNs on the Kaggle and Messidor test set. For the NSeS CNN, the performance after 60 epochs was also calculated and shown in the graphs for direct comparison with the SeS CNN 60. Operating points of both human observers are added in the plots. On the Kaggle test set, there was no significant difference ($p$-value=0.509) between the SeS CNN 60 and NSeS CNN 170, whereas on the Messidor test set, the SeS CNN 60 significantly outperformed the NSeS CNN 170 ($p$-value=0.0028).
Hemorrhage detection using convolutional networks with a fast training scheme

Figure 6.7: FROC curves of the SeS CNN 60 and NSeS CNN 170 on the Kaggle test set. The FROC curve of the NSeS CNN after 60 epochs of training is added for direct comparison with the SeS CNN after 60 epochs of training.

(a) Kaggle
(b) Messidor

Figure 6.8: Image-based ROC curves on Kaggle (a) and Messidor (b) test sets. Observer operating points of the human observers are added in the graphs.

Table 6.3 shows the contingency table for the observer gradings, the SeS CNN 60 and the NSeS 170 CNN as compared with the reference. Kappa agreements (κ) with 95% confidence intervals and sensitivity/specificity pairs are included.
Table 6.3: Contingency table of human observers, SeS CNN 60 and NSeS CNN 170 on the Kaggle and Messidor test sets. K = Kappa agreement with 95% confidence intervals, Se/Sp = sensitivity and specificity, Az = area under the ROC with 95% confidence intervals.

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>Observer 2</th>
<th>SeS CNN 60</th>
<th>NSeS CNN 170</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaggle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>1046</td>
<td>58</td>
<td>940</td>
</tr>
<tr>
<td>+</td>
<td>56</td>
<td>232</td>
<td>+</td>
</tr>
<tr>
<td>K = 0.759 [0.716-0.802]</td>
<td>K = 0.740 [0.696-0.785]</td>
<td>K = 0.598 [0.549-0.648]</td>
<td>K = 0.554 [0.501-0.607]</td>
</tr>
<tr>
<td>Se/Sp = 0.816/0.947</td>
<td>Se/Sp = 0.806/0.942</td>
<td>Az = 0.894 [0.867-0.919]</td>
<td>Az = 0.774/0.851</td>
</tr>
<tr>
<td>Messidor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>1040</td>
<td>64</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>53</td>
<td>235</td>
<td>+</td>
</tr>
<tr>
<td>K = 0.791 [0.753-0.829]</td>
<td>K = 0.742 [0.701-0.783]</td>
<td>K = 0.793 [0.755-0.832]</td>
<td>K = 0.783 [0.743-0.822]</td>
</tr>
<tr>
<td>Se/Sp = 0.972/0.889</td>
<td>Se/Sp = 0.944/0.869</td>
<td>Az = 0.972 [0.963-0.980]</td>
<td>Az = 0.894/0.918</td>
</tr>
</tbody>
</table>

Table 6.4: Contingency table of human observers, SeS CNN 60 and NSeS CNN 170 on the Kaggle and Messidor test sets after removal of poor quality images. K = Kappa agreement with 95% confidence intervals, Se/Sp = sensitivity and specificity, Az = area under the ROC with 95% confidence intervals.

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>Observer 2</th>
<th>SeS CNN 60</th>
<th>NSeS CNN 170</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaggle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>568</td>
<td>25</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>23</td>
<td>174</td>
<td>+</td>
</tr>
<tr>
<td>K = 0.838 [0.794-0.883]</td>
<td>K = 0.879 [0.776-0.931]</td>
<td>K = 0.917 [0.888-0.944]</td>
<td>K = 0.909 [0.884-0.932]</td>
</tr>
<tr>
<td>Se/Sp = 0.883/0.958</td>
<td>Se/Sp = 0.885/0.949</td>
<td>Az = 0.972 [0.963-0.980]</td>
<td>Az = 0.909 [0.884-0.932]</td>
</tr>
<tr>
<td>Messidor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>727</td>
<td>86</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>7</td>
<td>282</td>
<td>+</td>
</tr>
<tr>
<td>K = 0.806 [0.761-0.850]</td>
<td>K = 0.735 [0.719-0.751]</td>
<td>K = 0.979 [0.970-0.985]</td>
<td>K = 0.966 [0.954-0.976]</td>
</tr>
<tr>
<td>Se/Sp = 0.976/0.894</td>
<td>Se/Sp = 0.958/0.872</td>
<td>Az = 0.979 [0.970-0.985]</td>
<td>Az = 0.966 [0.954-0.976]</td>
</tr>
</tbody>
</table>

6.5.2 Influence of image quality

The number of poor quality images as indicated by the reference observer was 602 and 98 in the Kaggle and Messidor test sets, respectively. Table 6.4 shows the contingency table after removing the mentioned poor quality images for the observer gradings, the SeS CNN 60 and the NSeS 170 CNN as compared with the reference on the two test sets. Kappa agreements (K) with 95% confidence intervals and sensitivity/specificity pairs are included.

6.6 Discussion

During the time-consuming training process of a CNN, the majority of samples that are presented to the network are easy to classify correctly. In this work we hypothesized that we can speed up the training process by selecting difficult normal samples to present to the network. We achieved this by classifying normal images with the current state of the network after a number of epochs of training, and select more patches from those regions in normal images that the network considered abnormal. More precisely, in the SeS method, dynamic weights based on the CNN’s preliminary classification were
computed for each training sample at selected snapshots during training. Samples with higher weights were more likely to be selected for training in the next epochs. Using this scheme, the training procedure was guided to learn from the more informative samples. We applied the proposed SeS strategy to the detection of hemorrhages on color fundus images to show the potential of this technique in an important medical image analysis application. The results showed that the CNN with SeS employed in the training procedure required a considerably smaller number of training epochs to achieve a high performance when compared to a CNN without selective sampling.

Figure 6.9: Examples of errors by the SeS CNN system. (a): example of a retinal image with different type of abnormality (bright and dark regions on the right side of the image), (b): output of the SeS CNN 60 computed on (a), (c): example or a retinal image with hemorrhages (bottom left and top middle) which was graded as negative by the reference, (d): output of the SeS CNN 60 computed on (c). See text in the discussion section for more details.
6.6 Discussion

The SeS CNN required 60 epochs for the training phase to obtain similar performance as the NSeS after 170 epochs of training, as illustrated in Figure 6.4. When training is conducted in an iterative approach, which is the case with CNNs, it is likely that the importance of training samples changes during this learning process. The ability of the SeS CNN to dynamically change the focus of the learning process attributed to the speed-up of the learning process, as training time is not wasted on samples which the networks has already “learned” to classify correctly. Figure 6.5 displays the evolution of the pixel probability maps when evaluated on one unseen example image. It can be observed that the SeS CNN learns to differentiate between background tissue (i.e. blood vessels, fovea and micro-aneurysms) and hemorrhages faster than the NSeS CNN. Although these structures are specific to the retina, a similar learning behavior can be expected on other data sets as the training procedure with SeS is guided by its own capability to classify these structures.

Performance for the identification of hemorrhages of the SeS CNN is higher on both test sets as compared to the NSeS CNN, see Figure 6.8 and Figure 6.7 and comparable to human observer performance for the identification of images with hemorrhages. On the independent Messidor test set, this difference was statistically significant (p-value=0.0028). The image scores for presence of hemorrhages were calculated based on the maximum posterior probability in each probability map. Images containing challenging confounding structures are therefore more prone to misclassification. As the SeS CNN was guided to learn these challenging structures, overall classification rates compared to the NSeS CNN increase. There is a difference in performance obtained by the CNNs on the Kaggle and Messidor test set. Performance on the independent Messidor test set is higher as compared to the one obtained on the Kaggle test set for both CNNs. An explanation for this can be the presence of other abnormalities and the quality of the images in both data sets, see Figure 6.9((a) and (b)).

Assessment of image quality showed that 602 images (43.2%) were graded as having poor image quality by the reference observer in the Kaggle dataset, whereas for the Messidor test 98 poor quality images (8%) were identified. Figure 6.10 shows two examples of images which were graded as having poor image quality. This is an indication that the overall image quality in the Messidor test set is better than the overall quality in the Kaggle test set, allowing the CNNs to achieve higher performance.

In this study, we applied the SeS strategy only to the training samples that belong to the negative class. There is no fundamental reason why the same strategy could not be applied to the positive class as well. In our case, the set of positive samples was limited in numbers and each sample was already presented multiple times per epoch. If we would increase the number of positive training samples, either by increasing the amount of available training images or by applying more data augmentation, and applying the
SeS strategy to the positive class as well could potentially further increase the detection rates and speed up training. In this way, also difficult positive samples are presented more frequently during training. This would guide the CNN learning procedure to also better recognize the more difficult hemorrhage structures.

Although we obtained excellent performance for the detection of hemorrhages, an in-depth optimization of the network hyper-parameters was not performed in this study. This optimization is a challenging task: The depth of the network, i.e. the number of layers, and the number of kernels per layer, as well as the use of fully connected layers should be more thoroughly investigated. Pilot experiments using MSRA weight initialization showed an equal number of epochs required to train the SeS with similar end-performance on the test sets. Additions such as the inclusion of drop-out or batch normalization could potentially further increase performance. Furthermore, a kernel size of 3x3 pixels was chosen. The rationale of using such a small size kernels is that each larger size kernel, e.g. a 5x5 kernel, can be represented by multiple smaller sized kernels, i.e. two times a 3x3 kernel sized layer. Using multiple smaller sized layers with non-linear rectifications makes the CNN more discriminative and less parameters need to be optimized. The use of multi-scale patches could potentially be beneficial as usage of multi-scale patches has shown promising results in other applications. Optimizing CNN hyper-parameters is challenging and trying many combinations is common practice. However, it should be noted that the SeS CNN was compared to a NSeS CNN using the same network architecture. A similar improvement in training time and classification performance may be expected with a different network architecture.

Previous works have shown that adding more informative samples to the training set can improve the performance of the learner substantially. In boosting
techniques, an ensemble of learners is trained where each of the consecutive learners uses a fixed, more informative training set. Samples that are misclassified by the previous learners are typically added to the training set of the next learner. Application of this boosting approach to CNNs is highly time-consuming and inefficient as each of the learners is optimized independently and no information, such as network parameters, are shared between the learners. Taking into account the training process of a CNN is an iterative process, dynamically updating the training set in each iterations will avoid the use of multiple learners, focusing the attention of the learner on informative samples and optimizing the CNN parameters simultaneously. Other works have used a two-step approach in which representative samples are first identified by an initial learner. The first learner can be either the same learner, or a simplified, faster to train learner. Using this approach, a new dataset is created which is used for training a second, independent learner. Apart from the fact that a cascade of two independent learners are still needed and, consequently, more extra training time, using a simplified learner to select informative training samples does not guarantee that these samples are also informative for the second, more complex learner.

Similar to our approach, a previous work has used a dynamic sampling approach to train a multi-layer perceptron (MLP). In each training epoch, each training sample was first classified by the current state MLP to assign a sampling weight and it was determined using a sampling heuristic if this sample should be included for training. In this case, the sampling heuristic was designed to include all misclassified samples and a selection of correctly classified samples based on the class balance in the training set and the confidence level of the current state MLP. However, applying this heuristic to CNN training for patch classification is not feasible. First, including all misclassified samples would lead to the over-fitting of the network as millions of patches, mainly normal, would be misclassified, especially in the first iterations. Additionally, positive samples are highly under-represented in medical images. Therefore, all the positive samples should be considered as informative and no prioritizing selection is needed. For that reason, a sampling heuristic specifically designed for abnormality identification using CNN in medical images was proposed in this work, where a selection of informative negative samples was performed in each iteration and all positive samples were randomly included.

A limitation of this study is the use of a manual reference as provided by a single human expert. As hemorrhages and micro-aneurysms are similar in characteristics and are only differentiable by their size and color on color fundus images, they can be easily confused. Figure 6.9 (c) and (d) show a retinal image and the SeS CNN 60 output, respectively. The image which was graded as normal by the reference but both Observer 1 and Observer 2 indicated presence of hemorrhages. Combining human
observer annotations to create a consensus annotation might improve the reference, but prohibits a fair comparison with the performance obtained by one of these observers. Using an additional external reference such as fluorescein angiography, in which the contrast of blood (and therefore also hemorrhages) is enhanced by a contrast agent, might help to set a better reference standard\textsuperscript{84}. As the data sets used in this study are retrospectively analyzed and only contain color fundus images, expert grading on color fundus images was the best strategy available to us. Furthermore, the reference observer only indicated the hemorrhage center locations. Therefore, no detailed analysis on the individual hemorrhage segmentation could be performed. This would however be of added value for clinical studies and more research is needed for a more thoroughly evaluation of this task.

After analyzing the Kaggle test set further, we noticed that the majority of the errors made by human observers were on images from DR stages 1 and 2. Images from DR stage 1 and 2 contain numerous confounding lesions, such as microaneurysms, which are very difficult to differentiate from hemorrhages and introduce a high inter-reader variability. When taking only the images from DR stage 0 and 3 into account, the agreement of both observers with the reference was higher with $\kappa$ values of 0.776 and 0.771. This indicated a more reliable grading could be made on the images from DR stage 0 and 3. To investigate further the influence of a more reliable annotated training set, only image from DR stage 0 and DR stage 3 were used from the Kaggle dataset to train and evaluate the CNNs. The training time for the SeS CNN and NSeS was reduced to 40 and 140 epochs, respectively. The image level performance increased slightly to $Az$ values of 0.919 and 0.907; and 0.981 and 0.967 for the SeS and NSeS CNNs on the Kaggle and Messidor test sets, respectively. In this subset, we have also investigated the influence of the color normalization preprocessing step on the CNN performance. Training the SeS CNN without color normalization took five epochs longer to converge but achieved the same performance as the SeS CNN using color normalization in both test sets. This demonstrates that CNN is capable to deal with the large variability of medical data but it requires more time to learn this variability during the training phase.

All experiments were performed on an Intel Xeon PC with 2.4Ghz memory and a GeForce GTX 570 video card. The training time per epoch was around 16 minutes for both the SeS CNN and NSeS CNN and classifying all pixels in one image, i.e. computing a probability map, took around 0.82 seconds using a sliding window approach\textsuperscript{305}. In our current implementation, all the weights for the selective sampling were generated sequentially during one pass over the training set consisting of 3,959 images, i.e. this means a total time of 54 minutes for weight calculation. The SeS CNN required 60 epochs with 11 weight updates for the training process. The total time for the SeS
CNN to complete the training phase was then \(60 \cdot 16 + 11 \cdot 54 = 1554\) minutes, whereas for the NSeS, this was \(170 \cdot 16 = 2720\) minutes. However, the training time for the SeS CNN can be reduced significantly by parallelizing the generation of weights and CNN training. By doing so, the weights for the training samples can be computed during CNN training and the total time for the SeS can be reduced to \(60 \cdot 16 = 960\) minutes.

We have evaluated our proposed strategy using two different datasets in order to analyze the generalization of the method. However, more experiments with larger datasets or different training sets can be done in order to test the strategy more thoroughly. Increasing the amount of data with robust reference labels to train the CNNs may also help to further improve classification performance\(^{306}\). A data set with more training images contains a larger number of diverse training samples which can help the CNN to generalize better on unseen data. Using the proposed SeS strategy, the CNN will figure out which samples to use for training, forestalling an increase in training time. Using a much larger data set for training will be part of future work.

Despite these considerations, it is worthy to note that this is the first work on automated detection of hemorrhages in color fundus images that reports performance on par with two human experts. This result was obtained on a large, completely independent, publicly available test set, the Messidor database. Our method has a substantial higher Az value of 0.972 on this set as compared to previous work which reported an Az value of 0.87\(^{290}\). However, it has to be noted that a direct comparison cannot be performed as the training sets differ and the previous work used only a subset of 900 cases for evaluation. This excellent result confirms that convolutional neural networks have great potential to push forward the state-of-the-art in medical image analysis, similar to what has been achieved with this exceptionally powerful class of models in computer vision.

### 6.7 Conclusion

We have presented a method to substantially speed-up the time-consuming training process of convolutional neural networks with a selective sampling strategy, named SeS, embedded in the training procedure. We have demonstrated excellent results in the identification of hemorrhages on color fundus images. The SeS method addresses the common issue in medical image analysis tasks that challenging examples comprise only a small subset of the available data. By dynamically focusing the training effort on these samples that pose greater difficulty, we have shown that an increased overall performance can be achieved while a smaller number of epochs is required to train the network.
General discussion
In this thesis, we described and evaluated computer aided detection and diagnosis systems (here collectively referred to as CAD) for the analysis of retinal images. These systems can be used for a variety of tasks in research and clinical practice, but the focus of this thesis was to consider them to be used in eye screening programs for the early detection of retinal diseases. Currently in such programs, assessments are made by trained human graders. These assessments are time-consuming and the process generates a large burden on the health care system. Moreover, it is becoming increasingly difficult to comply with national screening recommendations due to the rapid growth of the screening population. To cope with this problem, the analysis of retinal images acquired during a screening examination can be performed by CAD systems. In our research group, we have previously developed CAD systems for the automatic detection of diabetic retinopathy (DR). In this thesis, we continued the work on several aspects of these systems and we have developed additional CAD systems for the identification of signs of age-related macular degeneration (AMD). In the following sections, we elaborate on eye screening, the role of CAD in screening, and we identify possible directions for future research.

**Eye screening: present and the future**

In the early stages, eye diseases often cause no visual loss and no other symptoms. The best option to identify individuals with early stage disease is then to screen subjects at increased risk. For this purpose national screening programs have been started in many countries, in particular for patients with diabetes, who have an increased risk to develop DR and are recommended to have a yearly eye examination. Studies have shown that these DR screening settings are effective and timely detection prevents these patients from losing their vision. With the development of new treatments to slow down the progression of AMD and the success of DR screening programs, screening for AMD is advocated and screening programs for AMD are starting to emerge. During a screening examination, images are acquired from the patient and examined for abnormalities by trained personnel. If abnormalities are found, the patient is referred to a specialist for a more detailed diagnosis.

CF imaging is currently the only imaging modality used in retinal screening programs. The concept of fundus photography dates back to the late 20th century. Years of research have led to digital fundus cameras with increased speed and processing power. Furthermore, costs of a fundus camera have decreased from ten thousands of dollars to prices in the range of a few thousand dollars. A CF photography exam is cheap with costs varying between $20 and $50, non-invasive and provides a high resolution view of the retina, sufficient to identify different lesions characteristics.
of early stage disease. Color fundus photography is not the only modality available. We envision that multi-modal imaging and image analysis will become more important in the near future. Multi-modality analysis can potentially increase the performance of systems that detect retinal diseases, and will be discussed in the next section. Although many studies have shown improved performance using multi-modality analysis in clinical settings \(^{225,311-313}\), studies on the cost-effectiveness and the performance in a screening setting have not yet been examined.

Screening programs put a large burden on the health care system with costs ranging between a baseline cost of $100 to $400 when a patient is referred \(^{314,315}\). In addition, the screening population is expected to increase rapidly in the near future. The number of people aged over 60 is expected to increase to 1.2 billion in 2025 \(^{316}\) whereas prevalence rates of diabetes are estimated to increase to 552 million in 2030 \(^{13}\). Solutions to this problem can be sought in automatic methods which can aid in, or potentially take over, the analysis of these retinal screening examinations by providing an automatic assessment. With the introduction of digital CF photography around 1995, research on automatic tools to analyze CF images has accelerated. Early works have focused on the detection and segmentation of anatomical landmarks, such as the fovea and optic disc, and structures such as the vasculature. The focus of research then targeted automatic detection of lesions indicative of specific diseases. This led to the development of fully automatic systems for the detection of retinal diseases \(^{82,176,317,318}\). Most of these works have focused on the detection of DR, but AMD is getting more attention in recent years. In Chapter 2, we have described a system which can identify patients in need for a referral for AMD based on the identification of drusen. These automatic systems can help in the identification of patients in need for follow-up or referral for more detailed diagnosis. The first automatic systems are currently enrolled in real-world screening settings. As of now, several systems are commercially available and have been CE certified \(^{172,174,178,318,319}\). So far, these systems have only focused on the identification of DR.

Incorporating automatic image analysis into screening programs may make them more efficient and cost-effective. This would be especially the case if it is demonstrated that CAD systems perform on par with human experts, or even better than the average human reader. In that case, automatic systems could perform the first line assessment of retinal examinations, and, just like human graders, identify patients in need for referral. In case of doubt, for example when the images are low quality or contain unusual structures, the computer could err on the side of caution and present the exam to a human grader, a strategy proposed previously in literature \(^{175}\). Advantages of using CAD systems instead of human graders include lower costs, improved throughput, and consistent output. I believe that in the foreseeable future, CAD systems can take over
initial assessment of retinal screening examinations. New image analysis techniques such as deep learning and the use of multi-modal imaging, could improve the performance and potential of these automatic systems even further.

**The role of multi-modal imaging**

Besides CF photography, other imaging techniques, fundus autofluorescence (FAF) imaging, near infra-red (NIR) imaging and optical coherence tomography (OCT), can be used to image the retina\(^2\)\(^-\)\(^4\). These imaging modalities can be complementary to each other. Fluorescein angiography allows for a clear visualization of the retinal vasculature, whereas FAF can document metabolic changes from the accumulation of fluorophores in the retinal pigment epithelium. NIR imaging allows for visualizing deeper structures in the retina as NIR imaging uses a higher wavelength for imaging. In OCT imaging, an optical beam is directed at the retina and the reflecting light, which is in coherence with a reference arm, is acquired\(^8\)\(^4\). Backscatters are typically caused by differences in refractive index in transitions from one tissue to another. By moving this reference arm, each layer of the retina can be visualized individually, allowing for a detailed analysis of the retina. OCT is currently widely used in clinical practice\(^8\)\(^4\).

In this thesis, we incorporated multi-modal information acquired using CF imaging, FAF imaging and NIR imaging in the analysis. Although CF photography is usually sufficient to identify signs of DR and AMD, multi-modality analysis has added value in certain situations. In Chapter 3, we showed that the use of FAF, NIR and CF together is beneficial for the identification and quantification of reticular pseudodrusen (RPD). Both expert human graders, as well as our proposed automatic system performed better when using multi-modality information. In Chapter 4, we continued the work on RPD detection using multi-modal information and showed even higher performance for our automatic system. By using multi-modal analysis, the number of false detections caused by drusen were reduced substantially. In these studies, we used 2D imaging techniques, which only allows assessment of a projection of the retina. OCT imaging could potentially help to differentiate between different types of lesions. RPD and drusen, two different lesions associated with AMD have very similar characteristics but form at different layer locations in the retina. Therefore, including OCT could potentially help to increase performance further and this is an important direction for future research.

While studies have shown increased performance of humans using multi-modal imaging\(^2\(^{29}\)\(^,\)\(^3\(^{20}\)\), automatic systems which make use of multiple modalities are still scarce. Many automatic systems have been developed to detect signs of retinal diseases using various single modalities. Combining image modalities in a single automatic system has
General discussion

not received much attention yet. One reason might be that CF images are still most widely used for diagnosis and screening and therefore have received the most attention. Combining multi-modal information can be achieved in various ways. We have shown that combining features derived from multiple imaging modalities improves the detection rate of RPD. We see no limitation for the inclusion of more imaging modalities or for application to other types of diseases.

Advanced image processing techniques

For most types of retinal lesions, detection is not considered as a difficult task by humans, because a clear difference with normal retinal images can be observed. However, this is misleading as lesions can also be subtle, and there are confounders, such as camera artifacts, that also for human readers are hard to differentiate from true lesions. As a result, observer studies demonstrate that human observers usually have good, but not outstanding agreement with the reference standard (often also set by human experts). In this thesis this is evident in Figure 2.3 and Figure 2.5 in Chapter 2, Figure 3.3 in Chapter 3, Figure 4.2 in Chapter 4, Table 5.6 in Chapter 5 and Figure 6.8 in Chapter 6. The same figures and table also show that the performance of computer systems is also not perfect. One explanation for this could be that most automatic systems to date rely on descriptors designed by humans to make a decision on presence of a disease. Human are well capable of detecting differences between lesions, background, and confounding structures, but explaining how they perform this task, and encoding their knowledge in computable features is difficult and cumbersome.

Image analysis techniques have been introduced which omit the use of these manual disease descriptors. In Chapter 4 and Chapter 6 we have made use of such a technique: deep learning\textsuperscript{116,117}, in particular convolutional networks. In deep learning, abnormal patterns are learned by the system from the available data. These learned descriptors are therefore tailored to the problem at hand and higher system performance can be obtained. Although the concept of deep learning is far from new and dates back about half a century\textsuperscript{119}, this technique has received great attention lately because of recent advances in computational power needed for this technique. Algorithms using deep learning have outperformed many state-of-the-art automatic systems in natural image processing tasks using large scale data sets\textsuperscript{247,250,276}. Recently, we can observe a rapid rise in applications of deep learning to medical image analysis and current state-of-the-art algorithms are being outperformed by deep learning systems\textsuperscript{259,277,278,321}. This technique has the potential to drive performance higher and can be applied to more diverse tasks. Furthermore, this technique naturally allows to embed multi-modal information in the automatic analysis. Although not covered in this thesis, I believe
deep learning can have a great impact on the detection of multiple diseases simultaneously as well. Therefore, further investigation of convolutional networks and other deep architectures for retinal image assessment is an important topic for future research.

**Data availability**

In order to further improve the performance of automated retinal screening, large, diverse, and carefully annotated databases are needed, likely in the order of millions of images. New techniques, such as deep learning have shown very promising results in computer vision tasks when extensively trained with millions of examples \(^250,276\) and are currently considered the method of choice for many image interpretation tasks, including those encountered in medical image analysis \(^259,277,278\). In Chapter 4 and in Chapter 6, we showed the potential of this technique by applying it to the detection of RPD and hemorrhages. We showed that a selective sampling technique can effectively filter informative samples from a large data set during training, to improve the training speed of such models, and in some cases obtain better results as well. Deep learning can circumvent the problem of designing optimal descriptors, but large numbers of images with expert annotations indicating presence of a disease need to be available. These annotations can be delineations of lesions or images or even exam level gradings on the severity of diseases. This is widely recognized in the retinal image analysis community and the last few years more and more of such databases have become available.

In February 2015, a challenge was organized by Kaggle to identify signs of DR in CF images\(^1\). This challenge titled "Diabetic Retinopathy Detection" invited people to design a fully automatic software solution which is able to differentiate CF photographs into the five severity stages of DR. In this challenge, the organizers accumulated CF photographs from a DR screening setting and provided this to the public: one training data set of over 35,000 images where each image was scored for DR severity, and one evaluation set of over 53,000 images where the image severity level was kept secret to the public until the challenge was completed in July 2015. The task for participants was to develop an algorithm which correctly assigned the DR severity level to the test images. After the challenge ended, a final ranking score was computed for each of the participants. In total, over 650 teams participated and a wide variety of software solutions were proposed. Interestingly, the majority of the top ranking algorithms were based on convolutional networks that used the full image as input and a single prediction was generated for each of the test images. The leading algorithms used deep networks with many layers and a range of additional tweaks including color normalization \(^88\), fusion of information from both eyes of a patient, state-of-the-art techniques such as dropout \(^300\)

\(^1\)https://www.kaggle.com/c/diabetic-retinopathy-detection
and special kinds of pooling operations\textsuperscript{322}. In this thesis, data of this challenge has been used to improve convolutional network training for the automatic detection of hemorrhages, see Chapter 6. As the challenge data originated from a real DR screening setting, the challenging conditions one has to deal with in a real-world setting were part of the problem: poor image quality, camera artifacts and a wide variety of cameras used. This challenge has given momentum to the development of automatic tools for the detection of DR by providing the necessity of large amounts of labeled data to develop such algorithms. Furthermore, it has made the machine learning community aware of the need of automatic tools for detection of retinal diseases and the need of practical solutions which can be applied in current screening programs.

In 2001, the Age Related Eye Disease Study (AREDS) was completed\textsuperscript{42}. This is the largest study to date about the natural history and risk factors of AMD performed. In 2010, 72,000 CF images with corresponding human grading information were made available. Researchers had to provide a detailed description of the study in which the data would be involved. In 2014, an additional set of 134,500 CF images were made available together with corresponding human grading information. All available CF images were acquired by digitizing film photographs by means of scanning and were captured at different time points with a follow-up up to 12 years for each patient. Although the image quality poses challenges, this database has given researchers great opportunities to develop new CAD systems for the automatic detection of AMD.

For glaucoma, no large data sets have been made available to facilitate development of automatic systems for the detection of glaucoma, but would be of great value for the automatic detection of this disease.

Population imaging is the large-scale application and analysis of medical images in controlled population cohorts, aimed to find imaging biomarkers that allow prediction and early diagnosis of diseases and preventive therapy. An increasing number of large population studies are conducted worldwide, and they often follow a cohort over time. Increasingly, retinal imaging is part of the work-up in population studies, not only to analyze eye diseases per se, but also because these images may contain biomarkers for major diseases such as stroke and dementia\textsuperscript{323}. The Rotterdam Study is a prospective cohort study which started in 1990 in The Netherlands and includes over 15,000 participants as of today\textsuperscript{237}. This study targets several diseases including ophthalmic, cardiovascular, neurological, oncological and respiratory diseases. Retinal images of this study have been included in Chapter 3 and Chapter 4. The European Genetic Database (EUGENDA) is a collaboration between the RadboudUMC and Uniklinik Köln and focuses to find answers to important questions regarding the development and therapy of AMD\textsuperscript{207,208}. It includes follow-up data of over 4,000 patients with detailed ocular information and imaging data. Part of this study data has been used
in Chapter 2. The Beaver Dam Eye Study was launched in 1987 and funded by the National Eye Institute\textsuperscript{324}. The purpose of the study was to collect information on the prevalence and incidence of cataract, AMD and DR. A total of 5,000 participants were recruited and the 20-year follow-up period was completed in 2010. Another large population database has been established in the UK: the UK Biobank. The UK Biobank recruited 500,000 people between 2006 and 2010 from all over the country. Detailed information including blood samples, urine samples, saliva samples and retinal images were acquired from each of the participants. Data from this population study has been used in several studies related to eye diseases\textsuperscript{325,326}.

In addition, a number of smaller databases have been made publicly available in the research community, including Messidor\textsuperscript{286,327}, DiareTB0\textsuperscript{267}, DiareTB1\textsuperscript{268}, Stare\textsuperscript{102}, DR1/DR2\textsuperscript{158}, the Digital Retinal Images for Vessel Extraction database (DRIVE)\textsuperscript{328}, REVIEW\textsuperscript{329} and the Hamilton Eye Institute Macular Edema Dataset (HEI-MED)\textsuperscript{330}. We expect this trend of making large annotated databases available to researchers to continue, and to be of paramount importance to further progress in the field.

**Detection of multiple diseases simultaneously**

Most screening programs focus on the identification of patients with a specific disease by only screening a population with a precisely targeted risk profile. However, if a patient presents with an abnormality associated with a different disease, he or she will also be referred to an ophthalmologist for a more detailed diagnosis. This is quite common as risk populations for different retinal diseases overlap. A large part of the diabetes population are older than fifty years, which is also the population at higher risk for AMD\textsuperscript{5,331,332}. As a result, patients having early signs of AMD are often encountered in DR screening settings. For automatic systems to be valuable in practice, these systems also need to cope with presence of signs of multiple diseases. Identification of abnormalities is top priority in screening settings and not referring a patient in need can have large consequences for the individual. CAD systems are typically designed to detect one specific disease or lesion. In order to operate in real-world eye screening, especially when first reading by computers to select subsets of exams is considered, automatic systems need to identify all relevant abnormalities in order to maintain high screening quality.

The most straightforward approach to automatically detect multiple diseases is to run multiple CAD systems in parallel, each for a different disease, and refer a patient if one or more automatic systems flag the patient as in need for a referral. An alternative is to design a system that can detect multiple diseases simultaneously. These systems can be more challenging to build, as they have to deal with a larger variety of disease
patterns. The focus of these algorithms will be to identify patients with any abnormality instead of identification of patients with a specific abnormality. In chapter 5, we have described a system to identify early signs of both AMD and DR and to make a differentiation of them using a set of features based on context. Although we obtained promising performance, the experiments indicated that human observers outperformed the computer system. One reason for this may be that it is difficult to create a good set of features which are optimal for the detection of different types of lesions. Deep learning systems are a good candidate for this task, as an optimal set of features is obtained automatically during training, rather than designing or selecting these features a priori using human intuition. No studies have been performed yet to show the potential of deep learning systems for the identification of multiple retinal diseases and this poses an avenue for future research.

Implementation of automatic screening systems

Some studies have reported performance of CAD systems on par with human observers in tasks related to retinal image interpretation\textsuperscript{13,82,155} and in this thesis we have shown the same in Chapter 2, Chapter 3, Chapter 4 and Chapter 6. However, these automatic systems are not yet widely implemented in real-world screening programs. Scientific research often shows the potential of a novel technique, but actual implementation is not realized. Implementation brings new challenges which need to be dealt with. Besides meeting high performance requirements, automated tools also need to be accepted in screening settings by the patients, the screening providers, regulatory authorities and health insurance companies. Without the support of these parties, automatic screening of retinal diseases will not become a reality. In the final stage of my PhD project, we have successfully applied for a STW TakeOff grant. This grant is made available by the Dutch organization "Stichting Techniek en Wetenschap" (STW), and various Dutch ministries and aims to stimulate young entrepreneurs to bring their research ideas into practice. This gives me the opportunity to perform a study on the feasibility to implement our developed software solutions for automatic screening for both AMD and DR in a real-world screening settings. Not only the technical feasibility, but also the commercial feasibility and regulatory issues will be investigated.

Concluding remarks

In this thesis we have developed new automatic techniques for the identification of retinal eye diseases in a screening setting. Technological advances have been made and new and improved image processing techniques for automated analysis of retinal
images, some based on deep learning, some based on the analysis of multiple types of images of the same patient, have been developed and validated. This thesis has focused in particular on automated detection of signs associated with AMD, but also considered DR and the differentiation between AMD and DR. In screening programs today, CF imaging is the mainstream modality because it is cheap, non-invasive and widely available. There are various other non-invasive techniques available to image the retina. In the future, we therefore expect multi-modality analysis to be increasingly important, and multi-modal imaging already plays a crucial role in detailed diagnosis when a patient is referred after a screening examination. Automatic systems for multi-modal image analysis are still in their infancy. The work presented in this thesis is a first step towards automated image interpretation in a multi-disease, multi-modal screening. I hope that this thesis will contribute to the goals of VISION 2020 of facilitating and accelerating implementation of eye screening programs for the early identification of retinal diseases.
The World Health Organization (WHO) and the International Agency for the Prevention of Blindness (IAPB) have launched a global initiative known as "VISION 2020: the right to sight". This objective aims to reduce and eliminate the main causes of preventable blindness. Achieving these goals can be accomplished by planning, development and implementation of sustainable national eye-care programs. The three main focus points of VISION 2020 include cost-effective disease control, human resource development, and infrastructure and technology.

In order to reach the VISION 2020 goals of eliminating preventable blindness, self-awareness of eye diseases has to be achieved. The best approach to accomplish this is to screen the population for eye diseases. National screening programs have been implemented for early detection of patients with Diabetic Retinopathy (DR) and programs for Age-related Macular Degeneration (AMD) are starting to emerge. These diseases are two of the most common retinal diseases worldwide that lead to visual impairment and blindness. They slowly progress to advanced stages where irreversible blindness occurs. Fortunately, treatment to slow down or halt the progression of the diseases is available.

In this thesis, an effort is made to help facilitating global eye screening programs. It describes and validates new automatic methods to detect AMD and DR using CF images. Automatic detection of these diseases can help achieving global screening and reduce the human suffering from preventable blindness.

In Chapter 2, we describe and evaluate a fully automatic machine learning system that allows for computer-aided diagnosis (CAD) of AMD by providing an accurate detection and quantification of drusen location, area and size. Data used in this study consisted of CF images of 407 eyes without AMD or with early to moderate AMD. The proposed CAD system consisted of a drusen detection step, followed by a step to provide a risk assessment to develop advanced AMD. First, pixels were identified which were likely to be part of a druse. Next, these candidate pixels were grouped into candidate druse regions. To complete the druse detection step, the candidate druse regions were classified using a statistical classifier and a set of features based on the shape, color, texture and intensity of the druse candidate. Finally, a risk assessment for patients to be at high risk of developing advanced AMD was performed using the detected drusen. Evaluation of the CAD system was performed using manual annotations made by two human experts. Free-response receiver operating characteristics (FROC) analysis showed that the proposed system approached the performance of human observers for the detection of individual drusen. The automatically computed drusen area showed excellent agreement with both observers, with mean intraclass correlation coefficients (ICC) larger than 0.85. Agreement on maximum druse diameter was slightly lower with a maximum ICC of 0.69, but comparable to the inter-observer
agreement (ICC=0.79). The performance of the CAD system for AMD risk assessment achieved areas under the Receiver Operating Characteristics (ROC) curve of 0.948 and 0.954, reaching similar performance as the human observers. The proposed CAD system allows for quick and reliable diagnosis of AMD, opening the way for large data set analysis within populations studies and genotype-phenotype correlation analysis.

In Chapter 3, we present a CAD system for the automatic detection and quantification of reticular pseudodrusen (RPD), a strong risk factor for advancement to late stage AMD. We evaluated the performance of the system using single- and multi-modal information and compared its performance with the ones obtained by two human observers. In this study, we used CF images, FAF images and NIR images of 278 eyes of patients with or without presence of RPD and patients with normal type drusen. The two human observers scored each image during single- and multi-modality grading setups for the presence of RPD. The developed CAD system consisted of a semi-automatic co-registration of image modalities to ensure spatial correspondence. After that, areas affected by RPD were identified using features derived from Gaussian filterbank responses computed from the different image modalities. The CAD system scored all images for presence of RPD using only single-modality information as well as when using multi-modality information. Furthermore, automatic quantification of RPD area was performed by the proposed system and compared with human delineations. The observers obtained a higher performance and better inter-observer agreement for RPD detection with multi-modality grading, achieving areas under the ROC curve of 0.940 and 0.958, and a kappa agreement of 0.911. The automatic CAD system achieved an area under the ROC of 0.941 with the multi-modality setup. Automatic RPD quantification resulted in an ICC value of 0.704, which was comparable with ICC values obtained between single-modality manual delineations. In this study, we showed that human performance and agreement for RPD identification improved significantly by using a multi-modality grading approach. The developed CAD system showed similar performance as observers, and automatic RPD area quantification was in concordance with manual delineations. The proposed automatic system allows for a fast and accurate identification and quantification of RPD, which opens the way for efficient quantitative imaging biomarkers in large data set analysis.

Chapter 4 continues the work presented in Chapter 3 and makes use of a more advanced image processing technique to improve RPD detection performance further. In Chapter 3, we presented a CAD system which required a description of disease patterns as seen on images by human experts in order to identify them in unseen cases. However, disease pattern description is difficult and mainly subjective by humans, especially in a multi-modal approach. In this Chapter, we proposed a self-learning system which independently learns discriminative patterns from multi-modal data, eliminating the need
for ill-defined manual descriptors. This technique made use of convolutional neural networks (CNNs) and is better known under the terminology of deep learning. Inspired by how the human brain works, disease patterns were captured in multiple hierarchical layers in order to make a distinction between normal and abnormal regions. Using this technique, we automatically identified and quantified RPD and compared the performance of the self-learning system to that of three human experts and the human-guided system as described in Chapter 3. The data we used in this study consisted CF images and NIR images; the same as used in Chapter 3. Results showed that the proposed self-learning system significantly outperformed the previous developed CAD system with an area under the ROC curve of 0.966 and obtained similar performance as human experts for the identification and quantification of RPD. The ability to independently learn an optimal set of patterns together with its high reliability makes these self-learning systems of great value for medical image analysis tasks and can potentially take over complex interpretation tasks.

In Chapter 5, we describe and evaluate a method to automatically differentiate between CF images containing either no lesions, drusen or exudates. Drusen and exudates, two bright appearing lesions associated with AMD and DR, respectively, can have very similar characteristics on CF images. Our method consists of three bright lesions detectors operating at pixel-level, combining their outputs using spatial pooling and classification with a random forest classifier. We used a large diverse dataset constructed from images from multiple public and private datasets. System performance was compared with ratings of two independent human observers using human-expert annotations as reference. Kappa agreements of 0.89, 0.97 and 0.92 and accuracies of 0.93, 0.98 and 0.95 were obtained for the system and observers, respectively.

In Chapter 6, we present a technique to improve and speed-up training of CNNs and applied it to the detection of hemorrhages on CF images. Although CNNs have pushed forward the state-of-the-art in a range of computer vision applications and are increasingly popular in medical image analysis, training of CNNs is time-consuming and challenging. In medical image analysis tasks, the majority of training examples are easy to classify and therefore contribute little to the CNN learning process. In this Chapter, we proposed a method to improve and speed-up the CNN training for medical image analysis tasks by dynamically selecting misclassified negative samples during training. Training samples were heuristically sampled based on classification by the current status of the CNN. Weights were assigned to the training samples and informative samples were more likely to be included in the next CNN training iteration. We evaluated and compared our proposed method by training a CNN with and without the selective sampling method. We used data from two large DR screening databases containing over 35,000 and 1,200 CF images, where the latter was solely used for eval-
uation purposes. Results show that a decreased training time from 170 epochs to 60 epochs with an increased performance, on par with two human experts, was achieved with areas under the ROC curve of 0.894 and 0.972 on the two data sets. The CNN with the proposed training strategy embedded statistically outperformed the standard CNN on the independent test set. By using the proposed training strategy, CNNs can be trained more efficiently and better performance can be obtained.

The studies performed in this thesis show that the developed CAD systems achieve performance comparable with human experts for the identification of both AMD and DR. Implementation of these systems in clinical practice can have a significant impact on eye health care and can help in the fight against preventable blindness.
Samenvatting
De Wereldgezondheidsorganisatie (WHO) en de International Agency for the Prevention of Blindness (IAPB) hebben een wereldwijd initiatief opgezet bekend onder de naam ”VISION 2020: the right to sight”. Het doel van dit initiatief is het verminderen en elimineren van alle gevallen van vermijdbare slechtziendheid en blindheid. Dit doel kan worden bereikt door planning, ontwikkeling en implementatie van duurzame nationale oog-screening programma’s. De drie belangrijkste focus punten van VISION 2020 omvatten kosteneffectieve ziektebestrijding, educatie, en infrastructuur en techniek.

Om het VISION 2020 doel te verwezenlijken moeten mensen zich ervan bewust worden wanneer ze een oogziekte hebben. De beste manier om dit te bereiken is om de populatie te screenen op oogziektes. Nationale screening programma’s voor de vroege detectie van patiënten met Diabetische Retinopathie (DR) zijn geïmplementeerd en programma’s voor Leeflijden Macula Degeneratie (AMD) worden ook steeds meer toegepast. Deze oogziektes zijn wereldwijd de twee meest voorkomende netvlies aandoeningen die leiden tot slechtziendheid en blindheid. Ze ontwikkelen zich langzaam totdat ze in een laat stadium onherstelbare schade toerichten die uiteindelijk leidt tot blindheid. Gelukkig bestaan er behandelingen die deze progressie kunnen vertragen of zelfs helemaal kunnen stoppen.

Dit proefschrift beschrijft en valideert nieuwe automatische methodes om AMD en DR te detecteren door het analyseren van kleuren netvliesfoto’s. Met behulp van automatische detectie van deze ziektes kunnen wereldwijde screening programma’s worden geïmplementeerd en kan het leed dat wordt veroorzaakt door vermijdbare blindheid worden verminderd.

In Hoofdstuk 2 wordt een volledig automatisch comput ersysteem beschreven dat AMD kan detecteren door middel van detectie en kwantificatie van drusen. De data die gebruikt wordt in deze studie bestaat uit 407 netvliesfoto’s van ogen zonder de kenmerken van AMD, of met kenmerken van vroege of gematigde AMD. Het computersysteem bestaat uit een drusen detectie stap, gevolgd door een risicoanalyse voor het ontwikkelen van een vergevorderd stadium van AMD. Allereerst worden pixels geïdentificeerd die mogelijk deel uit maken van een druse. Deze pixels worden vervolgens gegroepeerd om potentiële drusen kandidaten te vormen. Vervolgens worden deze kandidaten geclassificeerd op basis van de vorm, kleur, textuur en intensiteit van de kandidaten. Met deze gedetecteerde drusen wordt vervolgens een risicoanalyse voor het ontwikkelen van vergevorderde AMD uitgevoerd. Het computersysteem is geëvalueerd met behulp van handmatige annotaties welke zijn gemaakt door twee menselijke experts. Free-response Receiver Operating Characteristics (FROC) analyse laat zien dat het computersysteem net zo goed is als de menselijke experts in het detecteren van de individuele drusen. Het berekende drusen oppervlak kwam in grote mate overeen
Samenvatting

met die van de experts en de maximale grootte van de drusen kwam in mindere mate overeen met de experts. Hier was echter de overeenstemming tussen de twee experts ook lager. Voor de risicoadanalyse voor het ontwikkelen van vergevorderde AMD was Receiver Operating Characteristic (ROC) analyse uitgevoerd en haalde het computersysteem hetzelfde niveau als de twee experts en bereikte een oppervlak onder de ROC curve van 0.954. Dit computersysteem kan een snelle en betrouwbare diagnose van AMD uitvoeren en opent de weg voor het analyseren van grote populaties en genotype-fenotype correlatie studies.

Hoofdstuk 3 presenteert een computersysteem voor de automatische detectie en kwantificatie van reticulaire pseudodrusen (RPD), welke een sterke risico factor zijn voor de ontwikkeling van vergevorderde AMD. De prestaties van het systeem waren geëvalueerd en vergeleken met twee menselijke experts bij het gebruik van slechts één modaliteit en bij gebruik van meerdere modaliteiten tegelijkertijd. In deze studie zijn kleuren netvliesfoto's, autofluorescentie en infrarood beelden van het netvlies van 278 ogen van patiënten met of zonder RPD of met "gewone" drusen gebruikt. De twee menselijke experts scoren alle ogen op de aanwezigheid van RPD bij het gebruik van slechts één enkele modaliteit en bij het gebruik van alle drie de modaliteiten. In het ontwikkelde computersysteem werden eerst de verschillende beeld modaliteiten met elkaar geregistreerd zodat de locaties in de beelden exact overeenkomen. Vervolgens werden gebieden met RPD automatisch gedetecteerd met behulp van berekende kenmerken gebaseerd op de verschillende beeld modaliteiten. Het computersysteem analyseerde de ogen net als de experts met behulp van slechts één enkele beeld modaliteit en met behulp van alle drie de beeld modaliteiten tegelijk. De automatisch gedetecteerde gebieden werden vergeleken met de contouren die de experts hadden gemaakt. De experts bereikten een oppervlak onder de ROC curve van 0.940 en 0.958 en een kappa waarde van 0.911 wanneer ze alle drie de beeld modaliteiten tegelijk mochten gebruiken, wat hoger was dan wanneer ze slechts één beeld modaliteit mochten gebruiken. Het computersysteem bereikte een oppervlak onder de ROC curve van 0.941 wanneer het alle modaliteiten tegelijk gebruikte. De automatische kwantificatie van RPD was overeenkomstig met de handmatige kwantificatie. Deze studie toont aan dat menselijke experts beter presteren in het detecteren van ogen met RPD wanneer ze meerdere beeld modaliteiten tegelijk gebruiken en dat het ontwikkelde computersysteem dezelfde prestaties als de menselijke experts bereikt voor detectie en kwantificatie van RPD. Dit computersysteem kan op een snelle en precieze manier RPD identifieren en opent de weg voor analyse van bio-markers in grote populatie studies.

Hoofdstuk 4 bouwt voort op Hoofdstuk 3 en presenteert een methode die gebruikt maakt van een meer geavanceerde methode om RPD te detecteren. De methode in Hoofdstuk 3 is gebaseerd op het detecteren van patronen zoals gekarakteriseerd door
menselijke experts. Echter, het beschrijven van dit patroon is moeilijk in een multi-modaliteiten opzet en verschilt tussen experts. In dit hoofdstuk hebben we een methode ontwikkeld die zelf in staat is om discriminatorende patronen te leren herkennen in verschillende beeld modaliteiten. Deze zelf-lerende techniek, beter bekend onder de naam deep learning, maakt gebruik van een neurale netwerk met convolutie operaties (CNN). Deze techniek is geïnspireerd door de werking van het menselijk brein en gebruikt meerdere hiërarchische lagen van convolutie operaties om een onderscheid te maken tussen normale en abnormale gebieden. Deze techniek hebben we gebruikt om automatisch RPD gebieden te detecteren en we hebben de resultaten vergeleken met die van het systeem beschreven in Hoofdstuk 3 en met drie menselijke experts. In deze studie hebben we kleuren netvleesfoto’s en infrarood beelden van het netvlies gebruikt; dezelfde data als in Hoofdstuk 3. De resultaten laten zien dat de zelf-lerende methode statistisch gezien significant beter is dan het systeem in Hoofdstuk 3 met een oppervlak onder de ROC curve van 0.966 wanneer het gebruik maakt van beide modaliteiten. Het zelf-lerende systeem behaalde hetzelfde niveau als de drie experts voor het detecteren en kwantisering van RPD. Het vermogen om zelf patronen te leren herkennen en de hoge betrouwbaarheid van het systeem maakt deze techniek van grote waarde voor medische beeld analyse taken en kan in de toekomst misschien meerdere complexe taken overnemen.

Hoofdstuk 5 beschrijft een methode om automatisch kleuren netvleesfoto’s te onderscheiden in drie categorieën: geen laesies, drusen of exudaten. Druisen en exudaten zijn twee type laesies die voorkomen bij respectievelijk AMD en DR en kunnen een zeer sterke gelijkenis hebben op kleuren netvleesfoto’s. Ons computersysteem bestaat uit het combineren van drie laesie detectie computersystemen door middel van spatiale pooling en klassificatie met een statistische classifier. In deze studie hebben we een grote dataset gebruikt welke bestond uit netvleesfoto’s van meerder publiek toegankelijke datasets en private datasets. De output van het computersysteem en de gradering van twee getrainde mensen waren vergeleken, waarbij de mening van een menselijke expert als referentie was gebruikt. Overeenkomsten met kappa waarden van 0.89, 0.97 en 0.92 met een nauwkeurigheid van 0.93, 0.98 en 0.95 werden bereikt voor respectievelijk het computersysteem en de twee getrainde mensen.

In Hoofdstuk 6 presenteren we een methode om de training van CNNs te versnellen en hebben dit toegepast om grote bloedingen in het netvlies automatisch te detecteren met behulp van kleuren netvleesfoto’s. CNNs hebben de prestaties op meerdere gebieden in de automatische beeldanalyse naar een hoger niveau gebracht en worden nu ook steeds meer gebruikt voor medische beeldanalyse. Echter, het trainen van een CNN is tijdrovend en uitdagend. In medische beeldanalyse taken bestaat de grote meerderheid van voorbeelden om het netwerk te trainen uit makkelijk te onderscheiden
voorbeelden en deze dragen daarom minder bij aan het leerproces van het netwerk. Daarom stellen wij in dit hoofdstuk een methode voor om training voorbeelden dynamisch te selecteren tijdens het leerproces. Deze selectie is gebaseerd op hoe moeilijk deze voorbeelden zijn te classificeren door de staat van het netwerk op dat moment in het leerproces. Elk voorbeeld krijgt zo een gewicht toegewezen en moeilijke voorbeelden hebben hierdoor een grotere kans om te worden gebruikt in de volgende iteraties van het leerproces. We hebben deze methode geëvalueerd door een CNN te trainen met en zonder onze voorgestelde methode en de resultaten met elkaar vergeleken. We hebben hiervoor data uit twee grote DR screening data sets gebruikt die bestaan uit meer dan 35,000 en 1,200 kleuren netvliesfoto’s, waarbij de laatste enkel en alleen voor evaluatie doeleinden was gebruikt. De resultaten laten zien dat met onze voorgestelde methode een gereduceerde trainingstijd, van 170 iteraties naar 60 iteraties, en een hogere prestatie, die overeenkomt met die van twee menselijke experts, wordt bereikt met oppervlakken onder de ROC curve van 0.894 en 0.972 in de twee data sets. De CNN met de voorgestelde trainingsstrategie was statistisch gezien significant beter dan de CNN zonder deze strategie in een onafhankelijk evaluatie set. Door deze trainingsstrategie te gebruiken kunnen CNNs efficiënter worden getraind en kunnen er betere prestaties worden verkregen.

De studies in dit proefschrift laten zien dat de ontwikkelde computersystemen prestaties leveren die vergelijkbaar zijn met die van menselijke experts voor de detectie van zowel AMD als DR. Implementatie van deze computersystemen in de praktijk kan een significante impact hebben in de oogheelkunde en kan helpen in de strijd tegen vermijdbare blindheid.
Papers in international journals


**Papers in conference proceedings**


**Abstracts in conference proceedings**


Bibliography


[30] Age-Related Eye Disease Study Research Group, “The age-related eye disease study system for classifying age-related macular degeneration from stereoscopic


[47] E. Y. Chew, T. E. Clemons, E. Agrón, R. D. Sperduto, J. P. Sangiovanni, N. Kurinij, M. D. Davis, and Age-Related Eye Disease Study Research Group,


Dankwoord
Nu het einde van het proefschrift nabij is, lijkt het mij een mooie gelegenheid om even stil te staan en terug te blikken op een bijzondere periode uit mijn leven. Ik kan me nog goed herinneren hoe mijn rol binnen DIAG is begonnen. Aan het einde van mijn Master traject had ik besloten verder te gaan op het gebied van medische beeldanalyse en een PhD te willen gaan doen. Tijdens mijn master was ik al in aanraking gekomen met de groep van Nico in Nijmegen. Hier is de groep van Bram uit Utrecht bijgekomen en zodoende was de DIAG groep in Nijmegen ontstaan. Binnen deze groep waren op dat moment 3 PhD posities beschikbaar: Automatisch beeldanalyse op het gebied van onder andere het brein, de longen en de ogen. De laatste hiervan stond me wel aan en na een paar nachtjes slapen had ik besloten om op dit gebied mijn PhD te willen doen. Iets waar ik tot op de dag van vandaag geen spijt van heb gekregen! Het doen van een PhD zal voor veel buitenstanders worden geassocieerd met het doen van onderzoek, behalen van resultaten en het schrijven van publicaties. Het omvat echter vele malen meer dan dit alleen en is denk ik alleen goed te begrijpen als je zelf het traject hebt doorlopen. Er wordt het uiterste gevraagd op het gebied van creativiteit, flexibiliteit en incasseringsvermogen. Gelukkig sta je er niet alleen voor en heb je een team om je heen dat je helpt om alles in goede banen te leiden. Ik wil daarom iedereen die heeft bijgedragen aan het tot stand brengen van dit proefschrift enorm bedanken. Een aantal van jullie wil ik hierbij ook graag persoonlijk noemen.

Prof. dr. van Ginneken, beste Bram, ik wil je ontzettend bedanken voor jouw bijdrage aan mijn promotieonderzoek. Het enthousiasme dat jij uitstraalt en de toewijding die jij hebt voor het doen van onderzoek is een enorme boost voor mij geweest. Jouw oog voor detail zonder het grote verhaal uit het oog te verliezen zijn bezienswaardig. De visie die jij hebt op het gebied van medische beeldanalyse laat zijn sporen na binnen het vakgebied en draagt enorm bij aan het hoge aanzien van de DIAG groep. Jouw enthousiasme en soms wat directe manier van werken heeft mij soms wel wat afgeschrikt. In het begin heb ik hier erg aan moeten wennen, maar nu, na ruim 4 jaar kan ik toch wel zeggen dat ik enorm blij ben samen met jou te mogen werken. Bedankt voor jouw vertrouwen in mij en de mogelijkheid die jij me hebt gegeven om binnen de DIAG groep te promoveren.

Prof. dr. Hoyng, beste Carel, zonder de input vanuit de oogheelkundige kant zou het een lastige opgave zijn geworden om mijn onderzoek ook toegespaste waarde mee te kunnen geven. Jouw klinische blik heeft mij laten realiseren dat de impact van het onderzoek op de patient nimmer uit het oog verloren mag gaan en altijd hoog in het vaandel moet staan. Met jouw kijk op de toekomst binnen de oogheelkunde heb je een belangrijke sturing aan mijn promotie kunnen geven. Bedankt dat jij mij hebt willen begeleiden.
Dr. ir. Sánchez, dear Clarisa, I’m grateful to have had you as my direct supervisor during my PhD. Your preciseness has played a significant role and improved my thesis a lot. I realize that I have learned many things from you during my PhD. Not only my writing skills and English language have improved a lot, but also my level of confidence and capabilities to think about things in perspective. Starting as the only PhD student on the topic of the retina was sometimes difficult for me, but you always helped me whenever I was stuck or needed any help. Thank you for being my supervisor.

Dr. Theelen, beste Thomas, ik ben blij dat ik jou als begeleider heb mogen hebben. Elke keer dat we van gedachten wisselden was ik verbaasd over jouw kennis over en passie voor de oogheelkunde. Ik heb veel van je kunnen leren op het gebied van de ogen, meer dan ik mij ooit heb voor heb kunnen stellen. Jouw inbreng heeft ertoe geleid dat de klinische kant van mijn promotie niet op de achtergrond is komen te staan en dat het technische en klinische verhaal een samenhangend geheel is geworden. Bedankt voor jouw inbreng en begeleiding.

Prof. dr. van der Wilt, prof. dr. ter Haar Romeny en dr. Verbraak, ik wil jullie graag bedanken dat jullie deel uit hebben willen maken van mijn leescommissie.

Prof. dr. Klaver, beste Caroline, bedankt voor de goede samenwerking en vele projecten die we samen hebben gedaan en welke hebben geleid tot mooie publicaties. Het screening van mensen op netvlies aandoeningen is nauw verwant met bevolkings onderzoeken en ik ben blij dit ik een kijk vanuit die hoek heb mogen nemen tijdens mijn promotie.

Beste Freerk en Bart, toen ik begon aan mijn PhD was ik de enige promovendus op het gebied van de ogen binnen DIAG. Ik ben blij dat jullie het retina team hebben versterkt. Ik was eindelijk niet meer de enkeling die aan retina onderzoek werkte binnen DIAG. Ik heb het werken met jullie altijd al zeer plezierig mogen ervaren.

Ik wil graag alle (ex) onderzoekers in DIAG bedanken. De DIAG groep heb ik door jullie mogen ervaren als een erg prettige groep om in te werken en heb ook veel van jullie kunnen leren. Niet alleen op het gebied van werk, maar ook op het gebied daarnaast heb ik een aangename tijd met jullie gehad. De sociale activiteiten zoals borrels, feestjes en natuurlijk niet te vergeten de onwijs gave DIAG weekenden. Speciale vernoeming wil ik nog naar de volgende personen laten gaan omdat ieder van jullie op zijn of haar manier invloed heeft gehad op mijn manier van werken en doen: Albert, Babak, Bart, Bram P, Colin, Eva, Francesco, Freerk, Geert, Jan, Jaime, Jan-Jurre, Jean-Paul, Kaman, Leticia, Mohsen, Oscar, Paul, Pragnya, Rashindra, Rick, Rieneke, Sarah, Sjoerd, Steven, Suzan, Sven, Tao, Thijs en Wendy.

Solang, Leonie, Charlotte, bedankt dat jullie altijd voor mij klaar stonden als ik iets nodig had.
Tijdens mijn promotie heb ik ook mogen ervaren hoe het er in de klinische setting van oogheelkunde aan toe gaat en heb met vele mensen van de afdeling oogheelkunde mogen samenwerken. Michel, jouw invalshoek en kennis op het gebied van imaging heeft mij soms positief verrast. Ik heb op dat gebied veel van jou kunnen leren. John en Yara, ik ben jullie nog steeds dankbaar voor het monnikenwerk dat jullie voor mijn eerste studie hebben gedaan: het annoteren van alle individuele laesies op netvliesbeelden. De tijd en frustraties die dit bij jullie gekost moet hebben zal waarschijnlijk groter zijn geweest dan ik mij voor heb kunnen stellen. Ik wil ook graag de overige mensen bij oogheelkunde waarmee ik heb mogen werken vernomen omdat ook zij niet mogen ontbreken: Anneke, Constantin, Eiko, Eveline, Freekje, Myrte, Nathalie, Nicole, Ramon en Stanley. Bedankt voor alles wat ik van jullie heb mogen leren en de vele conferenties waar we samen naartoe hebben mogen gaan.

Gabriëlle en Corina, bedankt voor jullie hulp en samenwerking. Het annoteren van beelden om onze software te verbeteren moet voor jullie een flink karwei geweest zijn. Gelukkig heeft dit wel tot mooie resultaten geleid en uiteindelijk tot hoofdstukken in mijn proefschrift. Ik ben jullie hier nog steeds enorm dankbaar voor.

Beste Rick en Albert, bedankt dat jullie mijn paranimfen willen zijn. Niet alleen op het gebied van werk kon ik altijd bij jullie aankloppen maar ook voor allerlei andere dingen kon ik op jullie steun rekenen. Als harde kern van de mensen die zelf hun brood meebrengen heb ik heel wat slap gezever (voetbal, huizen, huisdieren, etc) van jullie aan mogen horen tijdens de lunch. Tot op de dag van vandaag ben ik er gelukkig nog niet op uitgekeken.

Ik wil ook graag mijn vriendengroep bedanken. Dirk, Edu, Frans, Ivo, Jeroen, Joost, Luuk, Martijn, Monique, Nicole en Pim, de meesten van jullie ken ik al zeer lang, sommigen zelfs van jongs af aan. Het is top dat ik altijd op jullie steun en vertrouwen kan rekenen. Bedankt voor jullie belangstelling en broodnodige afleiding. Hopelijk hebben jullie na het lezen van dit proefschrift een beter idee over wat ik de afgelopen jaren uit heb zitten spaken.

Ceciel, lieve schoonmoeder, dank voor de steun die ik van jou heb mogen ontvangen. Wanneer ik even krap in de tijd zit, bied je altijd aan om te komen eten. Ik hoop nu wat meer tijd te hebben, maar dat wil niet zeggen dat ik minder vaak langs zal komen.

Pim, lieve broer, al sinds ik me nog kan herinneren hebben we veel dingen samen gedaan. Ik kan altijd bij je terecht voor advies en ik ben ervan overtuigd dat dit altijd zo zal blijven. Ik ben er trots op om jou als mijn broer te hebben.

Lieve pap en mam, jullie hebben al mijn gehele leven achter mij gestaan en jullie onvoorwaardelijke steun is misschien wel het mooiste dat jullie mij kunnen geven. Net voordat ik aan mijn promotie was begonnen ben ik het huis uitgegaan. Bij jullie langskomen voelt echter nog steeds als thuiskomen en ik hoop dat ik nu na het afron-
den van mijn promotie ook meer tijd hiervoor zal hebben. Door alle drukte vergeet ik weleens te zeggen dat ik enorm trots ben op jullie als mijn ouders, daarom schrijf ik het hier nu eens en voor altijd op.

Als laatste natuurlijk mijn lieve Claudia. Op jou kan ik altijd rekenen als steun en toeverlaat. De laatste loodjes van dit proefschrift waren voor jou minstens zo zwaar als voor mij. Ik denk dat we samen een onoverwinnelijk team vormen, ook in de wat moeilijkere tijden. Bedankt voor jouw liefde. Ik hou van je!
Curriculum Vitae
Curriculum Vitae

Mark van Grinsven was born in Sint-Oedenrode on April 22nd 1987. He studied Biomedical Engineering at the Eindhoven University of Technology. His master thesis was entitled “Automated tracking and segmentation of vascular trees” and the research was carried out with Philips Healthcare. In July 2011, he started at the Diagnostic Image Analysis Group as a PhD student on Computer-Aided Diagnosis of Retinal Images. Results on the work he carried out in the Diagnostic Image Analysis Group are reported in this thesis.