Deep learning for corneal and retinal image analysis

AI FOR YOUR EYE

DEEP LEARNING FOR CORNEAL AND RETINAL IMAGE ANALYSIS

FRISO G. HESLINGA

Deep learning for corneal and retinal image analysis

"*AI for your eye*"

Friso Gerben Heslinga

This work is part of the research programme DLMedIA: Deep Learning for Medical Image Analysis - Weakly Labeled Learning with project number P15-26 Project 4, which is party financed by the Dutch Research Council (NWO).

Financial support by Intermedion B.V. for printing of this thesis is gratefully acknowledged.

Printed by Hollandridderkerk.nl
Cover design by Sadia Shadid
This thesis was typeset by the author using LaTeX.

A catalogue record is available from the Eindhoven University of Technology Library

ISBN 978-90-386-5521-5

Deep learning for corneal and retinal image analysis

AI for your eye

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens,
voor een commissie aangewezen door het College voor Promoties,
in het openbaar te verdedigen op
dinsdag 21 juni 2022 om 16:00 uur

door

Friso Gerben Heslinga

geboren te Wageningen

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr. M. Merkx
promotoren:	prof.dr. J.P.W. Pluim
	prof.dr. M. Petkovic
co-promotor:	dr. M. Veta
leden:	prof.dr. R.M.M.A. Nuijts (Maastricht UMC+)
	prof.dr. C.I. Sánchez Gutiérrez (Universiteit v. Amsterdam)
	prof.dr.ir. M. Steinbuch
	dr. T.T.J.M. Berendschot

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenkomst met de TU/e Gedragscode Wetenschapsbeoefening.

Contents

Chapter 1

Introduction

1.1 Imaging in ophthalmology

Ophthalmology is the branch of medicine that deals with diagnosis and treatment of disorders of the eye. The earliest documented reference to eye surgery dates all the way back to the Code of Hammurabi (1755 – 1750 BC), a legal text purportedly written by the king of Babylon [64]. Ironically, the same Code introduced the "an eye for an eye" principle that found its way into many justice systems since, and which perhaps fostered the demand for eye care.

Arguably, eye care has improved since Babylonian times. Nowadays, clinicians are able to diagnose a wide variety of ophthalmic disorders such as cataract, diabetic retinopathy, and corneal dystrophy. Various treatments have been developed and improved, sometimes resulting in technically demanding and complex procedures. An example is Descemet membrane endothelial keratoplasty (DMEK), a transplantation technique where the inner ~30 micrometer of the cornea is replaced with tissue from a donor using incisions of just a few millimeters wide [104].

Paramount for current day ophthalmic diagnosis and treatment is the use of imaging technology. The eye's transparency for visible light is a property that is gratefully employed by several imaging modalities, to visualize eye structures at different depths (Figure 1.1). Most notable is the anterior segment, which includes the cornea and the anterior chamber; and the fundus, which includes the retina, optic disc, and macula.

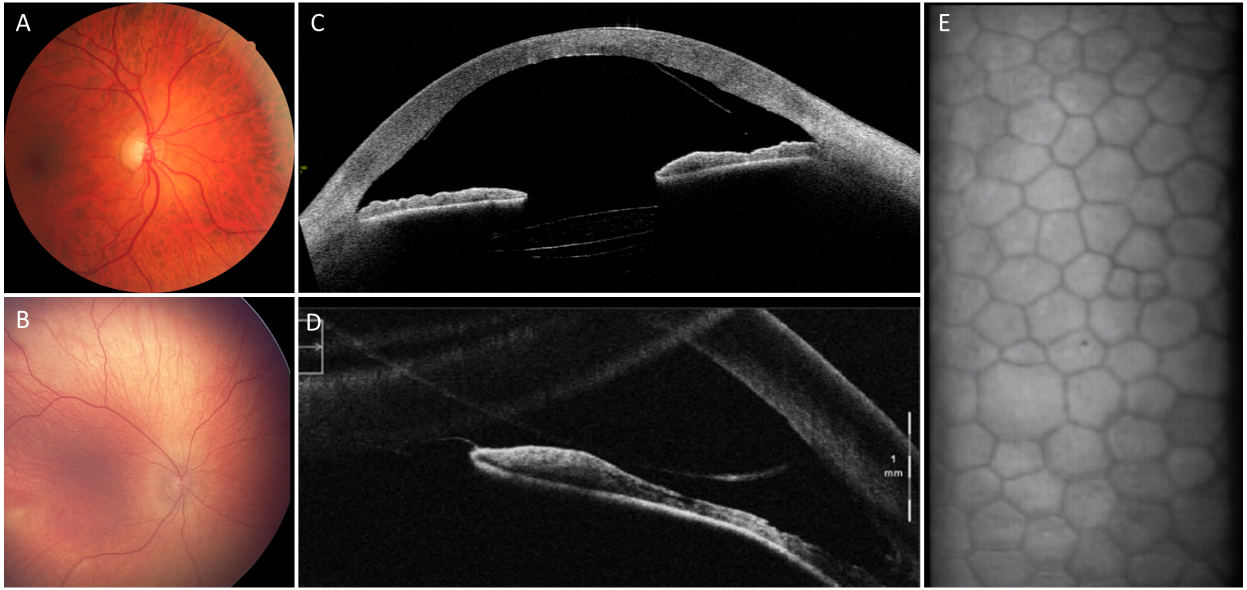

Figure 1.1: Examples of ophthalmic images. (A) Fundus photography - FF450; Carl Zeiss AG, Jena, Germany. (B) Neonatal fundus photography - RetCam III; Clarity Medical Systems, Pleasanton, California. (C) Anterior segment optical coherence tomography (OCT) - CASIA2; Tomey Corp. Nagoya, Japan. (D) Intra-operative OCT - Zeiss Lumera 700 RESCAN; Carl Zeiss Meditec, Jena, Germany. (E) Specular microscopy - Topcon SP-3000P; Topcon Corporation, Tokyo, Japan.

1.2 Automatic image analysis

Interpretation of all these images is typically done by an eye doctor, called an ophthalmologist. The overall number of image acquisitions keeps increasing, which requires a lot of time from specialists for interpretation and administration. An example is the large-scale screening for diabetic retinopathy, which results in a high demand for assessment of fundus photographs. If (part of the) assessment can be done automatically, it could save ophthalmologists a lot of time that could be spent on other parts of ophthalmic patient care.

A second motivation for automatic image analysis is the need for objective measures. An ophthalmologist might recognize that the cornea is somewhat thickened, but for a better diagnosis she would like to have a quantitative and precise measurement of the thickness at different locations of the cornea. Moreover, one would like to track changes over time, for example local increase in corneal thickness because of disease, or local decrease in corneal thickness due to a successful corneal transplantation.

Most imaging techniques create the images digitally, which allows easy handling and transformation. Automatic image analysis can be achieved using multiple steps of preprocessing, extraction of image features, and finally quantification of important clinical parameters. Until recently, image analysis was based on traditional computer vision techniques, which apply feature extraction. Feature extraction was mainly based on converting human expert knowledge about well-defined tasks into software. A wide variety of techniques has been published for ophthalmic image analysis, yet many image analysis challenges remain to be solved or even to be discovered.

1.3 Deep learning

In recent years, deep learning has shown to be very promising for automatic medical image analysis in general [95], including measurement of clinical properties [158], tissue detection [16], biomedical segmentation [128], and disease classification [47]. Deep learning is a subset of machine learning techniques with models that contain many (typically millions of) trainable weights. These weights are iteratively updated with respect to some loss function which compares model predictions to ground truth labels. In contrast to classical machine learning techniques, no handcrafted features have to be selected as the relevant features are automatically learned from the image data [84].

Deep learning has already seen many applications in the field of ophthalmology, where it often outperforms classical image analysis techniques [150]. The growing interest is visualized in Figure 1.2, showing the fast-increasing number of scientific publications about corneal and retinal image analysis in combination with deep learning. Nevertheless, at this point many challenges are yet to be solved. These challenges cover multiple image types and a wide variety of eye disorders or screening opportunities.

Choosing projects that offer most value to society is not straightforward, which is why for this thesis we focused on questions posed by clinicians in our network or by clinician that approached us. In the thesis, deep learning is used to address multiple unsolved challenges in corneal and retinal image analysis. We do so for a variety of imaging modalities and diseases by integrating state-of-the-art methods and development of novel image analysis pipelines.

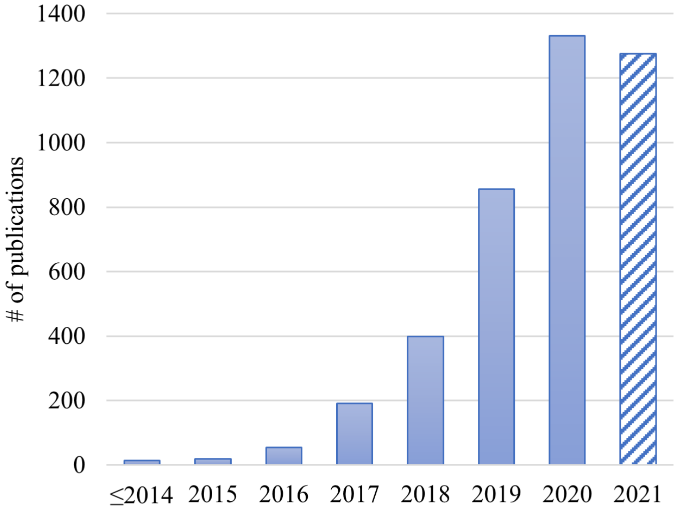

Figure 1.2: Publications about deep learning for ophthalmic image analysis in Scopus in recent years. Search performed on November 4, 2021, using the following query: TITLE-ABS-KEY (("deep learning" OR "convolutional neural network" OR "deep neural network") AND ("ophthalm*" OR "retina*" OR "cornea*" OR "fundus" OR "diabetic retinopathy" OR "retinopathy of prematurity" OR "macula*" OR "glaucoma" OR "cataract" OR "keratoconus" OR "keratitis" OR "keratoplasty")). Data for 2021 is incomplete.

1.4 Outline of this thesis

Throughout this thesis, deep learning-based methods are proposed and evaluated for challenges in corneal and retinal image analysis. Chapters 2, 3, 4, and 5 describe studies related to corneal image analysis and chapters 6 and 7 describe studies related to retinal image analysis.

In **chapter 2**, a method for automatic DMEK graft detachment quantification in anterior segment optical coherence tomography (AS-OCT) is proposed. During transplantation of the cornea, the DMEK donor graft is surgically introduced into the eye onto the posterior side of the cornea. If all goes well, the DMEK graft attaches to the cornea and supports corneal regeneration. When the graft partly detaches, it is important to quantify the extent of detachment. We developed a deep learning-based image analysis pipeline that automatically crops the relevant part of the AS-OCT image, segments the detached DMEK graft, quantifies the extent of detachment, and constructs a map of the detached regions.

In **chapter 3**, we employ the automatic cropping method proposed in chapter 2 and compare three deep learning techniques for automatic corneal thickness measurement. We develop and evaluate on the same set of AS-OCT data set used in chapter 2, acquired after DMEK surgery. This type of images poses a challenge for existing thickness measurement techniques. Furthermore, we construct detailed thickness maps that allow easy inspection and progress tracking.

In **chapter 4**, instead of using post-operative optical coherence tomography (OCT) images, we analyze intra-operative OCT images. The challenge here is to assess whether the DMEK graft, that is inside the anterior chamber but has not been positioned onto the cornea yet, is positioned correctly or upside-down. We propose an image analysis strategy that consists of deep learning-based segmentation and post-processing to obtain the graft's curvature at each point. Subsequently, we relate the graft's curvature to its orientation.

In **chapter 5**, we zoom in to the microscopic level where we can distinguish individual corneal endothelial cells using specular microscopy. Endothelial cell density (ECD) is an important biomarker of corneal health and requires accurate segmentation of corneal endothelial cells. Current methods for extracting ECD can be insufficient when image quality is suboptimal or if guttae are present. In this chapter, we present a novel deep learning method for accurate cell segmentation in specular microscopy images of varying quality and in the presence of guttae.

In **chapter 6**, we switch our focus to the retina by analyzing fundus photography images, which originate from The Maastricht Study. We develop and compare deep learning methods for classification of images that belong to individuals with normal glucose metabolism and those with type 2 diabetes. We also investigate the effect of simultaneous prediction of classical features, and we compare strategies for aggregating image-level predictions to individual-level predictions.

In **chapter 7**, we use the best practices of chapter 6 to develop a deep learning framework for detection of type 2 diabetes in a set of over 46,000 fundus images. This time we also evaluate how well the model can be used to distinguish prediabetes individuals. Moreover, we investigate how the discriminative power of the fundus images compares to that of typical diabetes risk factors such as age, sex, and waist circumference.

Chapter 8 concludes this thesis with a summary of the main findings, a discussion of the contributions made in this thesis, and an outlook for future research and development.

Chapter 2

DMEK graft detachment quantification in AS-OCT

This chapter is based on:
F.G. Heslinga, M. Alberti, J.P.W. Pluim, J. Cabrerizo, M. Veta. Quantifying graft detachment after Descemet's membrane endothelial keratoplasty with deep convolutional neural networks. *Translational Vision Science & Technology*, 9(2):48, 2020.

Abstract

In this chapter, we developed a method to automatically locate and quantify graft detachment after Descemet's membrane endothelial keratoplasty (DMEK) in anterior segment optical coherence tomography (AS-OCT) scans. A total of 1280 AS-OCT B-scans were annotated by a DMEK expert. Using the annotations, a deep learning pipeline was developed to localize scleral spur, center the AS-OCT B-scans and segment the detached graft sections.

Detachment segmentation model performance was evaluated per B-scan by comparing (1) length of detachment and (2) horizontal projection of the detached sections with the expert annotations. Horizontal projections were used to construct graft detachment maps. All final evaluations were done on a test set that was set apart during training of the models. A second DMEK expert annotated the test set to determine inter-rater performance. Mean scleral spur localization error was 0.155 mm, whereas the inter-rater difference was 0.090 mm. The estimated graft detachment lengths were in 69% of the cases within a 10-pixel (~150 μm) difference from the ground truth (77% for the second DMEK expert). Dice scores for the horizontal projections of all B-scans with detachments were 0.896 and 0.880 for our model and the second DMEK expert, respectively. In conclusion, our deep learning model can be used to automatically and instantly localize graft detachment in AS-OCT B-scans. Horizontal detachment projections can be determined with the same accuracy as a human DMEK expert, allowing for the construction of accurate graft detachment maps. Automated localization and quantification of graft detachment can support DMEK research and standardize clinical decision making.

2.1 Introduction

Descemet's Membrane Endothelial Keratoplasty (DMEK) currently offers the greatest opportunity of visual gain to patients suffering from endothelial dysfunction [104, 144]. However, partial graft detachment after DMEK remains a burden for patients and a challenge for surgeons with detachments requiring air injection in 3% to 76% of cases [36, 125].

Anterior segment optical coherence tomography (AS-OCT) allows for visualization of early graft detachment and is therefore clinically useful in guiding postoperative care [174]. However, the quantification of the detached area remains difficult because AS-OCT typically consists of multiple radial B-scans and a physician must integrate several images to have an overview of detached areas. Moreover, we found that the degree of graft detachment can be ambiguous in some regions when the graft is appositioned to the inner cornea yet not attached. No fast and objective tool to visualize all detached areas currently exists. We believe such a tool could aid in the postoperative management of DMEK patients, including the decision to rebubble or perform re-DMEK.

In this chapter, we propose an automated image analysis method (Figure 2.1) that has the potential to improve clinical decision making by objectively detecting the areas of DMEK detachment and providing an overview of all detached areas at once.

2.2 Methods

2.2.1 Ethical approval

This study was approved by the Capital Region Committee on Health Research Ethics, Denmark and adhered to the tenets of the Declaration of Helsinki. All participants provided written informed consent before participation.

2.2.2 Data

Swept-source AS-OCT scans (CASIA2; Tomey Corp. Nagoya, Japan) were collected as part of a randomized controlled trial conducted at the Department of Ophthalmology, Rigshospitalet – Glostrup, Denmark. Briefly, the randomized study included patients with Fuchs' endothelial dystrophy or pseudophakic bullous keratopathy eligible for DMEK surgery and excluded re-DMEK procedures or prior keratoplasty. The study was double-blinded and was designed to

Table 2.1: Overview of image data and annotations. Data in the training and validation column was used to design, train and optimize the deep learning models. Test data was used for the final model evaluation. *out of 276 scleral spur points annotated by the first DMEK expert, 232 points were also annotated by a second DMEK expert to evaluate inter-rater agreement.

	Training & validation data	Test data	Total
Participants	50	18	68
Hospital visits	60	20	80
AS-OCT B-scans	960	320	1280
AS-OCT B-scans without graft detachment	232 (24.2%)	104 (32.5%)	336 (26.3%)
Scleral spur points annotated	847 (44.1%)	276* (43.1%)	1123 (43.9%)
Both scleral spur points annotated in B-scan	288 (30.0%)	81* (25.3%)	369 (28.8%)

compare patients randomized to either air or sulfur hexafluoride (SF6) DMEK surgery [6]. A DMEK expert (M.A.) annotated 80 scans from 68 participants, acquired either immediately after surgery and/or postoperative day 7. Typically, due to the presence of a large gas bubble supporting most of the graft, little to no detachment is present immediately after surgery. However, these scans were included to help our model distinguish between graft detachment and intraocular gas. Each scan consists of 16 radial B-scans, corresponding to a total of 1280 images of 2133 × 1466 pixels. For each B-scan, locations where the graft had detached were manually annotated with point markings (image coordinates). Additionally, the scleral spur was annotated when clearly discernible in the inferior and superior part of the B-scan, resulting in a maximum of two points per scan. The data were randomly split on a participant level in a set for training and evaluation of our models ($N = 960$ images) and a set for final testing ($N = 320$ images). Details about the AS-OCT B-scans are shown in Table 2.1. Participant characteristics are shown in Table 2.2.

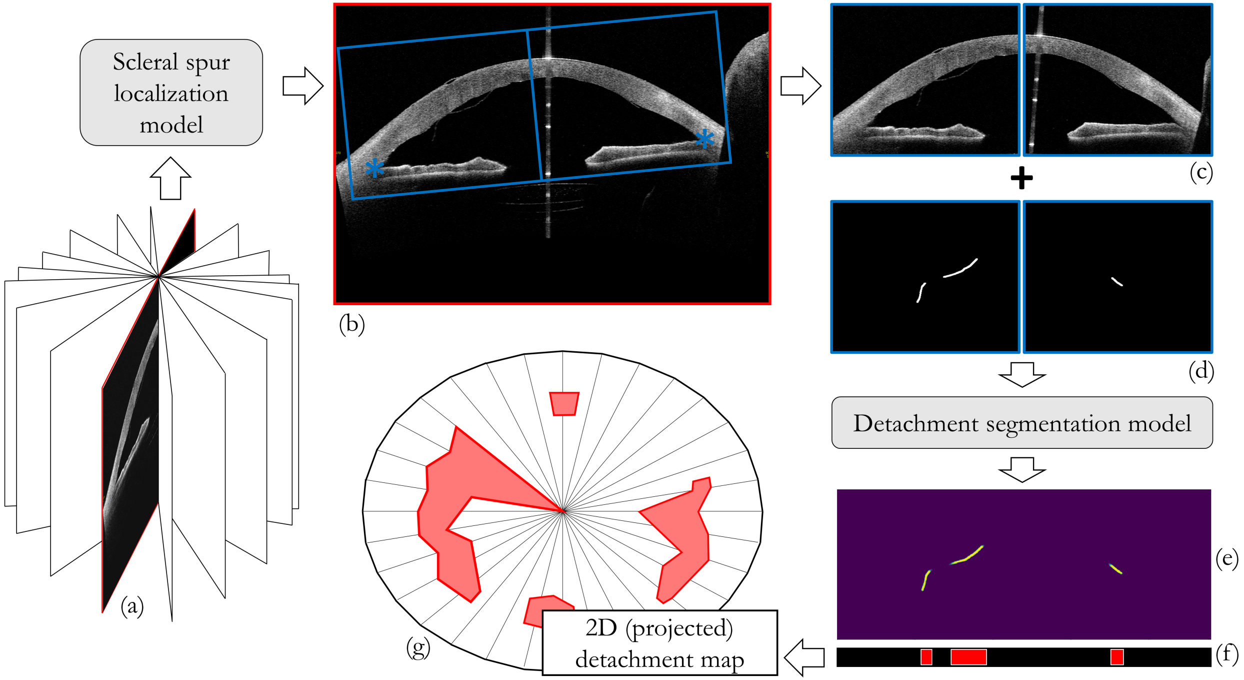
Scleral spur localization model
(a)
(b)
(c)
(d)
Detachment segmentation model
(e)
(f)
2D (projected) detachment map
(g)

Figure 2.1: Deep learning pipeline for quantification of corneal graft detachment. A scleral spur localization model is applied to a radial B-scan of an AS-OCT (a). The scleral spur estimates are used to center the B-scan (b) and obtain crops. The crops (c) are processed through a segmentation model, which was trained to output a map with detachment predictions (e) similar to expert annotations (d). Combining the horizontal projections (f) of 16 B-scans, a graft detachment map can be constructed.

Table 2.2: Participant characteristics. Values are presented as mean (SD) or ratio. P-values were determined with T-test for continuous variables and Fisher's test for categorical variables. FED = Fuchs' endothelial dystrophy. PBK = pseudophakic bullous keratopathy. POD 7 = postoperative day 7

Characteristics	Participants in training and validation set	Participants in test set	p-value
Age, mean years (SD)	70.2 (7.47)	73.3 (6.10)	0.09
Sex, male/female	27 / 23	6 / 12	0.17
Diagnosis, FED/PBK	50 / 0	17/ 1	0.26
Laterality, R/L	26 / 24	10 / 8	1.00
Tamponade, air/SF6	28 / 22	9 / 9	0.78
Visit, immediately after surgery / POD 7	10 / 50	2 / 18	0.72

2.2.3 Deep learning pipeline

For the analysis of the AS-OCT data, we used deep learning methodology [84], which has successfully been used for many medical image analysis tasks [95] including ophthalmology [41, 93, 129, 150]. In this chapter we present a framework with a four-step approach: (1) localization of the scleral spur using a deep learning-based regression model to center each AS-OCT B-scan; (2) fit of an ellipse to the scleral spur points of all radial B-scans to refine localization and centering; (3) segmentation of the detached areas with a deep learning segmentation model and (4) extraction of DMEK biomarkers from the segmentation maps. Each step is described in more detail in the following sections. An overview of the deep learning pipeline is shown in Figure 2.1. Models were implemented in Keras [29] using a TensorFlow backend [1].

2.2.4 Scleral spur localization

For a clinical evaluation of graft detachment, and to study detachment progression, it is important to find the detached areas with respect to the center of the cornea. The center of the cornea is difficult to locate – especially when corneal edema is present, which is why we used the center of the fitted scleral spur ellipse instead. The scleral spur's morphology and position have been proven to stay unaltered after surgery and therefore it has been chosen as a landmark for quantitative measurements in the anterior chamber [9, 132]. It is visible in only

70 - 78.9% of all of the radial B-scans [127], due to image artifacts from eyelids or anatomical variations [170]. The localization model is therefore only trained on B-scans for which the scleral spur could be annotated in both the superior and inferior segment ($N = 288$). Training was done using batches of ten image crops and reducing the mean squared error. Basic data augmentation (rotation and translation) was used to increase the variability of the training data [80].

AS-OCT B-scan images (2133×1466 pixels) were reduced to 512×352 pixels and converted to grayscale. A well-known deep learning architecture, ResNet-50 [67], was modified to match the input dimensions and outputs four values that represent the coordinates of two scleral spur locations per B-scan. The localization model was trained with batches of 10 images by reducing the least-square error between the model outputs and the targeted coordinates. A grid search was used to find the optimal set of model hyper-parameters and select the best-performing model. This model was then used to process all B-scans in the test set. Note that scleral spurs locations were estimated even in B-scans that were not annotated.

The anatomical structure of the scleral spur can be approximated by an ellipse in the 3D AS-OCT volume. Since the scleral spur is not clearly discernible in each B-scan, we included an extra step that exploits this ellipsoid structure and makes the localization model robust for all B-scans. First, we fitted an ellipse through the 32 scleral spur point estimates (2 scleral spur points for each of the 16 AS-OCT B-scans). Then, for each scleral spur point estimate, we updated the estimate with the location of the ellipse through that slice.

2.2.5 Detachment segmentation

We created binary masks from the point markings that represent locations along the detached graft. Examples of the masks can be seen in Figure 2.4 and Figure 2.5. The width of the detached lines was set to 15 pixels. Based on the scleral spur point estimates of the scleral spur localization model, B-scans were cropped such that the cornea was centered (1920×768 pixels). Taking advantage of the anatomical symmetry of the anterior chamber, the crops were split into an inferior half and a vertically reflected superior half. This step halved the detachment model input size, while doubling the number of training examples. The crops were downsampled by a factor of two to obtain the final size of 480×384 pixels.

As a data augmentation technique, we added random uniform noise to the locations of the scleral spurs (-60 to +60 pixels in the horizontal and vertical coordinate) prior to cropping, resulting in translated and slightly rotated crops.

The same cropping procedure and data augmentation was applied to the masks, ensuring that the OCT crops and masks remained aligned.

To localize the image pixels that illustrate graft detachment, we employed a semantic segmentation approach. A deep learning model with a U-Net architecture [128], was implemented to output a mask similar to the input. The model was trained using batches of eight image crops, using a weighted cross-entropy loss. We experimented with the weight factor of foreground pixels on the loss, and found that a factor of 2 provided the best results on the validation set. The best performing model was applied to the test set to obtain mask predictions.

2.2.6 Biomarkers extraction and evaluation

For each mask prediction in the test set, *length of detachment* was determined using a skeletonization method [176]. This skeletonization method is a morphological procedure that involves shrinking the regions in the binary image until they are one pixel wide. The remaining pixels were counted as a proxy measure for length of detachment. After processing the test set crops with the detachment segmentation model, the skeletonization method was applied to the outputs of the segmentation model as well as the annotated masks. Evaluation of length of detachment was done by comparing the model predicted detachment length with the annotated length. Although length of detachment is our primary outcome measure, it does not provide information about the relative location of detachment. To enable the construction of a 2D-map of detachment, we projected the detached locations on the horizontal axis of each cropped radial B-scan. The 16 projections can then be combined to create a 2D-map giving an overview of all detached areas in a single image (Figure 2.1). Performance of the projection of the detached sections on the horizontal axis was evaluated using Dice score [39]. The Dice score was determined for the overlap between the projections and perfect overlap would result in a Dice score of 1.

Additionally, an inter-rater analysis was performed where a second DMEK expert (J.C.) annotated the B-scans in the test set. These annotations were processed similarly as the annotations of M.A. (expert 1) and assessed for scleral spur localization error, length of detachment and overlap in horizontal axis projection of detached graft sections.

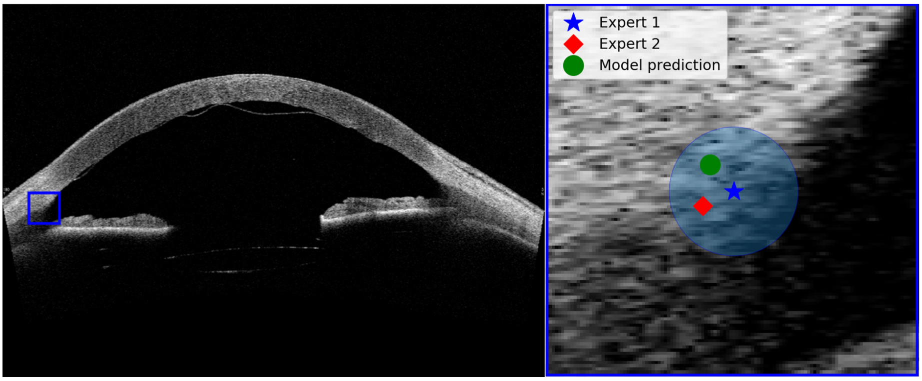

Figure 2.2: Example result of scleral spur localization model for a test case. Left: B-slice with original resolution (2133 × 1433 pixels). Right: Enlarged version of the blue box in the left image. The circle boundary represents the mean error between Expert 1 and the model prediction (0.155 mm).

2.3 Results

2.3.1 Scleral spur localization

The scleral spur localization model with the ellipse fit was applied to the downscaled B-scans of the test set. The mean Euclidean distance between the annotations of Expert 1 and the model predictions was 4.97 pixels (0.155 mm). Moreover, 95% of these errors were within 8.79 pixels (0.275 mm). In comparison, the Euclidean distance between the two experts was 2.87 pixels (0.090 mm). Only scleral spur points that were annotated by both experts were used for the final evaluation. Figure 2.2 provides a visual interpretation where we show an example case and the mean error of the scleral spur localization model. We also tested for the effect of the ellipse fit. Without ellipse fit the mean Euclidean distance between the localization model and Expert 1's annotations was found to be slightly smaller: 4.48 pixels (0.140 mm). Additional discussion of our motivation to use the ellipse fit can be found in the discussion section.

2.3.2 Length of graft detachment

The main results of the segmentation model are shown in Figure 2.3, where length of detachment is displayed in a Bland-Altman plot [18]. The original field of view of the B-scan was 16 by 11 mm, for 2133 × 1466 pixels, so after downscaling with a factor two, one pixel corresponds to 15.0 μm. The bias (6.04 pixels) is relatively small compared to the mean length of detachment, and 69% of cases are within a difference of 10 pixels (~150 μm). Some outliers are found for cases with a larger length of detachment and these mostly represent underestimations of the length of detachment. In comparison, the bias in detachment length between Expert 1 and Expert 2 was -0.9, with 1.96 SD (standard deviations) between -33.56 and 31.44. For 77% of cases, the difference in annotated length of detachment is within 10 pixels.

The numbered green dots in Figure 2.3 refer to specific B-scans that are shown in Figure 2.4 and Figure 2.5. Green dots 1-3 are examples of successfully segmented scans, while numbers 4-6 correspond to B-scans for which the segmentation model outputs show substantial deviations from the expert annotations.

The test set included two OCT scans that were acquired immediately after surgery. In all 32 B-scans, the edge of the intraocular gas bubble was visible to the human observer, but no graft detachment was present. The detachment model provided false positive regions in 2 out of the 32 B-scans.

2.3.3 Projection results

When all B-slices were included, the Dice score was found to be 0.906 (±0.190), compared to 0.916 (±0.160) for the inter-rater performance. When empty masks were excluded from this analysis, the Dice scores for the segmentation model and the inter-rater performance were found to be 0.896 (±0.149) and 0.880 (±0.172), respectively.

The detachment projections of 16 B-scan can then be plotted on a grid similar to the radial grid of the AS-OCT scan. Since all B-scans were previously centered with respect to the middle of the cornea (using the scleral spur estimates), the detached sections can directly be mapped on the radial grid. Three examples of such detachment maps are shown in Figure 2.6, in which the red structures represent the detachment segmentation model estimates and the green dotted line the expert annotations on AS-OCT images.

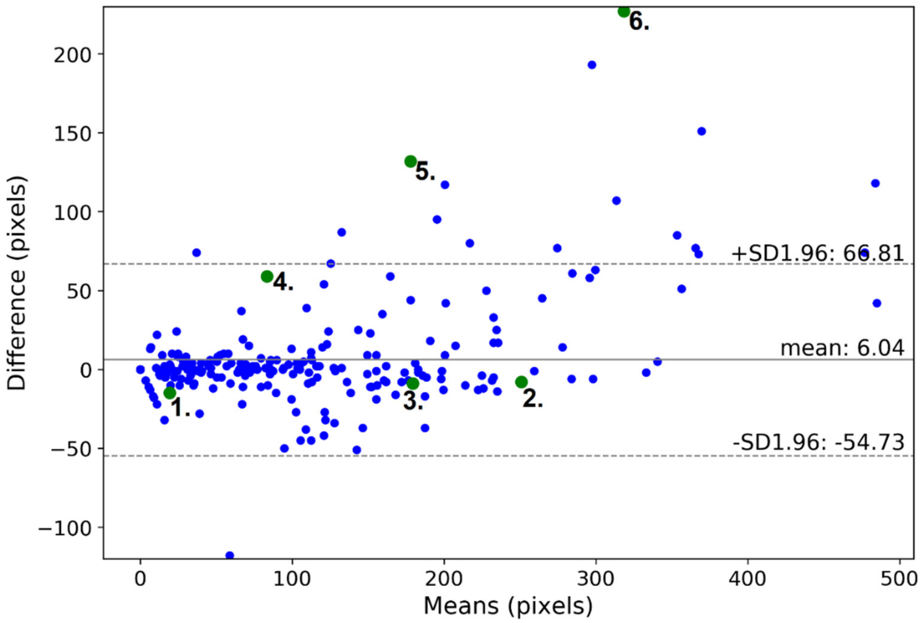

Figure 2.3: Bland-Altman plot of length of detachment determined by the segmentation model versus the annotations of Expert 1. Length is measured as the number of pixels after applying a skeletonization method to the mask. The horizontal axis describes the mean of the length of detachment as determined by Expert 1 and the segmentation model. The vertical axis is the difference between Expert 1 and the segmentation model. ±1.96 SD (standard deviation) describes the 95% confidence interval. A positive difference means that the segmentation model underestimates detachment length compared with the expert annotations. One pixel corresponds to 15.0 μm.

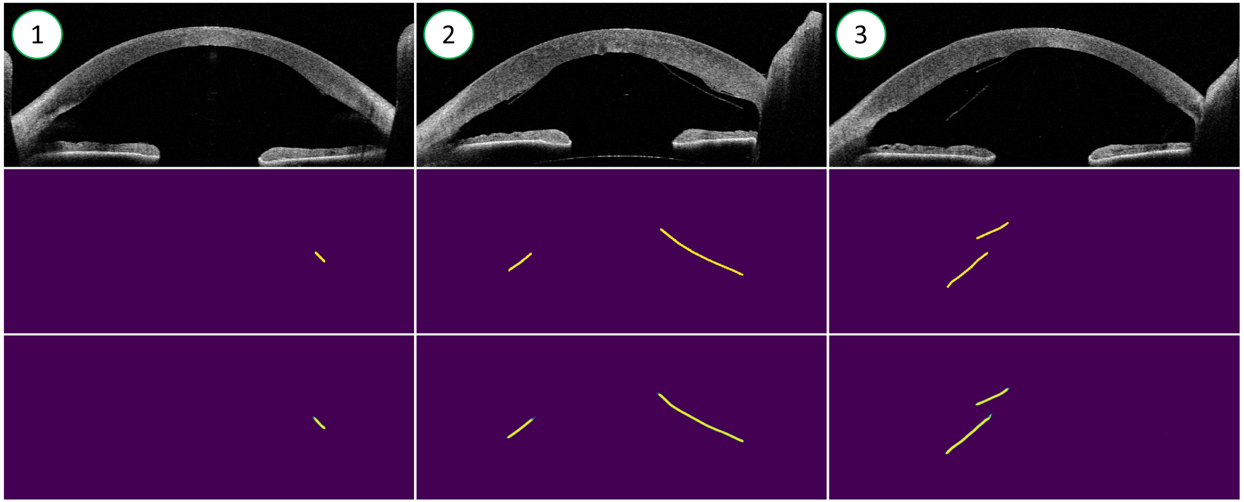

Figure 2.4: Examples of successful segmentations. Top row: OCT B-slices from the test set. Middle row: mask annotations by a DMEK expert. Bottom row: output of the segmentation model. For the predictions, yellow indicates high confidence that a section is detached, while green indicates lower confidence. The numbers in the top left corner correspond to the green dots in Figure 2.3.

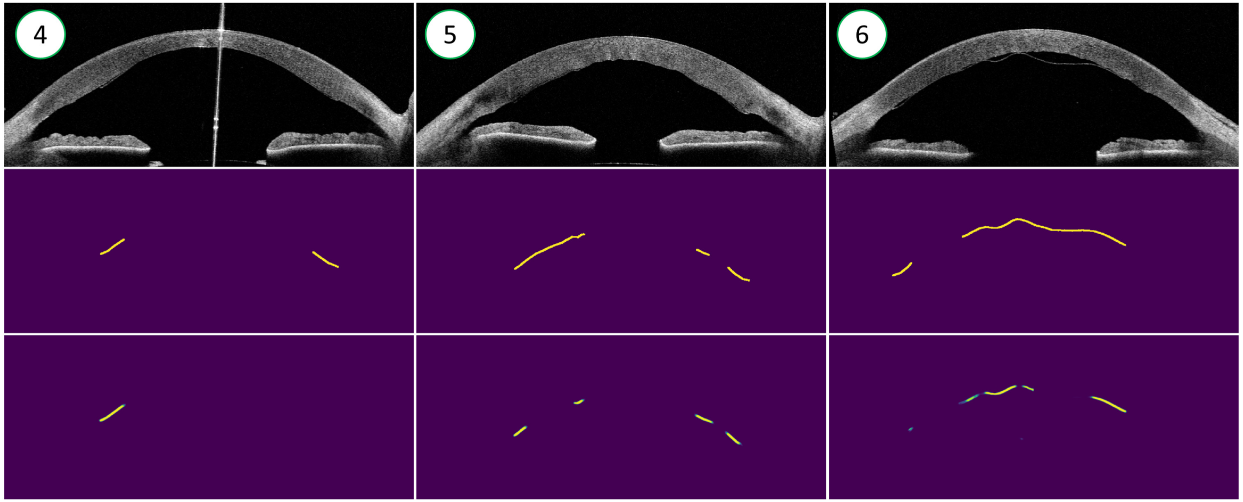

Figure 2.5: Examples of segmentations that deviate from the expert annotations. For more details, see the description in Figure 2.4

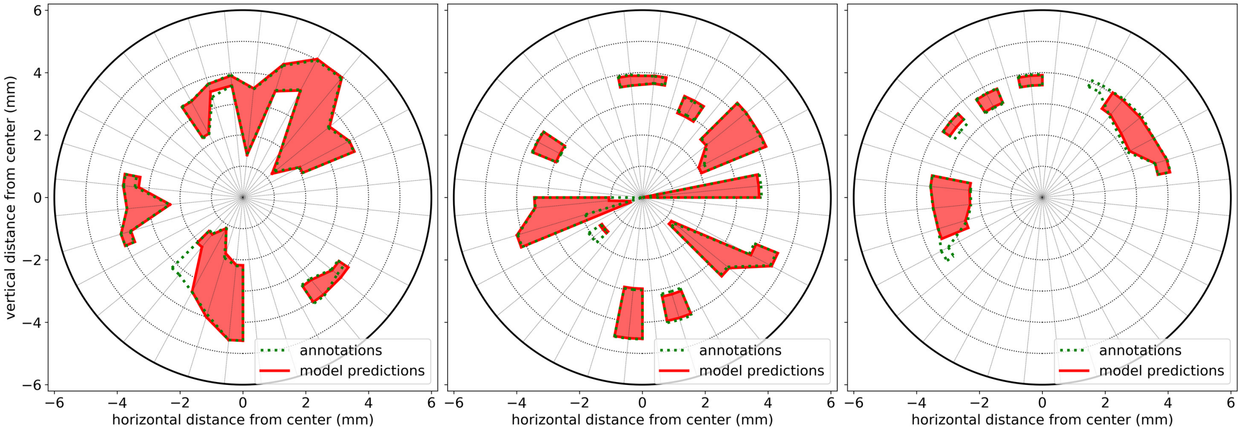

Figure 2.6: Graft detachment maps of three AS-OCT scans, connecting the detached sections of 16 radial B-scans. The red line and surface indicates the model predictions, while the green dotted line represents the expert annotations.

2.4 Discussion

Our results demonstrate that a deep convolutional neural network can accurately and automatically identify DMEK graft detachment. We believe that our deep learning pipeline has the potential to improve and standardize clinical decision making and can similarly be used as an objective and operator-independent outcome to improve DMEK research and reporting.

The number of DMEK procedures performed is rising rapidly driven by the superior visual results [100, 119]. In 2018, 41.2% more DMEKs were performed in the United States, while the total number of endothelial keratoplasty procedures increased only 4.6% in comparison with the year before [113]. This invariably increases the need for DMEK detachment management, such as the decision to await spontaneous clearance; rebubble; or perform re-DMEK [10, 40, 52]. Studies agree that management depends on the degree of detachment yet report diverging opinions for when to rebubble and varying definitions of partial detachment, including visually significant graft detachment [122], 20% detached area [124], more or less than one-third detached [40]. In current practice, the amount of detachment after DMEK is estimated by a clinician/surgeon over a succession of scans on a screen, rather than measured objectively. Thus the surgeon has to make a decision with regards to treatment without seeing all detached areas in a single image or accurately being able to quantify the total amount of detachment.

The high Dice scores for the projection results are similar to a human DMEK expert and indicate that accurate detachment maps can be constructed. Visual evaluation of some examples of these maps (Figure 2.6) indeed shows a strong similarity with the expert annotations. Since the center of the detachment map corresponds with the center of the cornea, the severity of the detached sections can be evaluated with their respective distance to the center. Moreover, follow-up OCT scans can be overlaid to assess detachment progression.

The Bland-Altman plot in Figure 2.3 also indicates that the segmentation model works well for most individual B-scans. Examples 1-3 in Figure 2.4 represent the results for the majority of the segmentations and show high segmentation accuracy, even when a graft is torn (example 2). For some cases, the predicted detachment length differed substantially from the expert annotations (Figure 2.5). Part of the disagreement could originate in the inherent uncertainty of some graft sections that are difficult to annotate. Indeed, the DMEK experts do not always agree, but the 95% confidence interval for the inter-rater study is roughly half the size of the model prediction confidence interval. Moreover, we also found that the model makes a few substantial mistakes that are obvious to the human observer (e.g. example 6). After visual inspection of the outliers, we noticed that most of the sizeable underestimations were B-slices of one specific OCT scan with a lot of detachment in the center. These errors are likely due to the lack of training examples with a large central detachment. The model might confuse some large center detachments for intraocular gas, which is only present in scans directly after surgery or rebubbling. Although the effect of these type of mistakes might be limited since they are obvious and will easily be spotted by the ophthalmologist, it could be addressed by adding more training data encompassing more variations, especially cases with center detachments. Furthermore, the current segmentation model did not take into account information from neighboring B-slices, as the DMEK expert did. Finally, some inaccuracy might result from the loss of information due to downsampling the B-slices by a factor two before processing the scans with the graft segmentation model. Given the horizontal line-like structure of the graft and the downsample factor, the horizontally most distant pixels could be misclassified. However, this error will be small compared to the whole length of the graft detachment.

Apart from missegmenting some cases with a lot of detachment in the center, it was sometimes challenging to distinguish remnant host Descemet's membrane from the DMEK graft. Furthermore, the presented models were trained and evaluated on a single data set collected with one type of AS-OCT device. For generalization towards multiple-sources, the models have to be either retrained with some images from other scanner types, or with the use of other domain generalization techniques [42].

Prior image analysis work within the realm of DMEK detachment has only focused on binary classification; i.e. whether detachment is present or not [152] and whether rebubbling was performed [66]. We believe our detachment model is of clinical value as it provides quantitative measures about length and location of graft detachment. The segmentation accurately locates detachment in most AS-OCT B-scans and is much faster than a human rater. In clinical practice an ophthalmologist would not have time to annotate the detachment regions in detail, while our deep learning pipeline could provide an instant evaluation aiding the decision. Although our aim was to develop a model for quantifying DMEK detachment, we also developed a scleral spur locating model as an intermediate step. Having this scleral spur localization model aided our AS-OCT B-scan preprocessing by cropping all images uniformly prior to the DMEK detachment model evaluation. This cropping step also provided practical benefits, as we did not have to reduce the standard U-Net model size or the resolution of the B-scans further to fit within GPU memory. However, locating the scleral spur is valuable in and of itself. Potential applications include determining limbal chamber depth parameters such as angle-opening distance (AOD) and trabecular-iris space area (TISA), relevant in glaucoma. Furthermore, it may also be a valuable tool for aligning AS-OCT scans between patient visits (e.g. to compare pachymetry map changes).

The refinement of the scleral spur estimates by fitting an ellipse resulted in a slightly bigger localization error. However, we could only evaluate for scleral spur points that were well discernable, since those were annotated by both experts. Our model also outputs an estimate for the scleral spur when the region itself is not visible (e.g. hidden behind the eyelid) [132]. We think that the ellipse fit step makes the localization more robust for these cases and reduces outliers. Whether our scleral spur model can be applied to other disease entities, such as acute angle-closure glaucoma is a topic of future research.

In summary, we have introduced a deep learning pipeline based on AS-OCT that allows automatic and accurate quantification of graft detachment after DMEK. Our future research efforts will focus on evaluating the value of our algorithm for improving clinical decision making and clinical outcomes after DMEK.

In the next chapter, we use the same AS-OCT data set to address a different challenge: Measurement of corneal thickness, also called 'pachymetry'.

Chapter 3

Corneal pachymetry in AS-OCT

This chapter is based on:
F.G. Heslinga, R.T. Lucassen, M.A. Berg, L. van der Hoek, J.P.W. Pluim, J. Cabrerizo, M. Alberti, M. Veta. Corneal pachymetry by AS-OCT after Descemet's membrane endothelial keratoplasty. *Scientific Reports*, 11(Article number: 13976), 2021.

Abstract

Corneal thickness (pachymetry) maps can be used to monitor restoration of corneal endothelial function, for example after Descemet's membrane endothelial keratoplasty (DMEK). Automated delineation of the corneal interfaces in anterior segment optical coherence tomography (AS-OCT) can be challenging for corneas that are irregularly shaped due to pathology, or as a consequence of surgery, leading to incorrect thickness measurements. In this chapter, deep learning is used to automatically delineate the corneal interfaces and measure corneal thickness with high accuracy in post-DMEK AS-OCT B-scans. Three different deep learning strategies were developed based on 960 B-scans from 50 patients. On an independent test set of 320 B-scans, corneal thickness could be measured with an error of 13.98 to 15.50 μm for the central 9 mm range, which is less than 3% of the average corneal thickness. The accurate thickness measurements were used to construct detailed pachymetry maps. Moreover, follow-up scans could be registered based on anatomical landmarks to obtain differential pachymetry maps. These maps may enable a more comprehensive understanding of the restoration of the endothelial function after DMEK, where thickness often varies throughout different regions of the cornea, and subsequently contribute to a standardized postoperative regime.

3.1 Introduction

Corneal thickness is a key biomarker for corneal disorders, including Fuchs' endothelial dystrophy [79, 118], keratoconus [7, 91], and keratitis [31, 167]. Measurements on the corneal thickness, called pachymetry, enable detection of thickness changes that are indicative of restoration of corneal endothelial function after surgical treatment. For visualization of the restoring cornea, anterior segment optical coherence tomography (AS-OCT) has become the preferred imaging modality due to its high resolution and reproducibility [94, 163] (Figure 3.1). While current OCT software works well for delineating the boundaries of healthy corneas, it often fails for corneas that are irregularly shaped due to pathology, or as a consequence of surgery (Figure 3.2). Manual correction of the delineation mistakes is time consuming and not practical for a clinical setting.

In recent years, automated image analysis using deep learning [84] has shown to be promising for ophthalmic applications [150], including the analysis of AS-OCT images [55, 69, 152, 172]. Deep learning is a subset of machine learning techniques with models that contain many (typically millions of) trainable weights. These weights are iteratively updated with respect to some loss function which compares model predictions to ground truth labels. In contrast with classical machine learning techniques, no handcrafted features have to be selected as the relevant features are automatically learned from the (image) data. Recent work already showed the potential of deep learning for corneal pachymetry by AS-OCT, specifically for keratoconus [41]. We hypothesized that a similar approach could be used for corneal pachymetry for cases with irregular inner corneal curvature and/or structures that look similar to the corneal boundaries, both of which can lead to delineation failures by standard AS-OCT software.

In this study we focus on OCT scans acquired after Descemet's Membrane Endothelial Keratoplasty (DMEK) [104]. During DMEK, the diseased corneal endothelium and Descemet's membrane are replaced with a donor graft. After placement of the graft, a gas bubble is injected into the anterior chamber to support graft attachment to the host cornea. Both the procedural gas bubble and donor graft can mimic the appearance of the corneal interface and result in incorrect delineation.

We evaluate three different deep learning techniques that were developed or used for ophthalmology applications and shown to be highly effective. We validate our thickness measurements for the central 9 mm diameter (Figure 3.1), whereas previous work only did so for 3.1 mm [41]. This is essential to assess corneal regeneration after DMEK surgery which uses a graft of ~8.5 mm. In addition, we present an automatic approach for reconstructing differential pachymetry maps that locates the center of the cornea in subsequent images and visualizes thickness differences over time.

3.2 Results

3.2.1 AS-OCT data & annotations

The AS-OCT scans used in this study are similar to those used in chapter 2. The scans were collected as part of a randomized controlled trial conducted at the Department of Ophthalmology, Rigshospitalet – Glostrup, Denmark. The trial was designed to compare air and sulfur hexafluoride (SF6) DMEK surgery in patients with Fuchs' endothelial dystrophy or pseudophakic bullous keratopathy [6]. Repeat DMEK procedures and patients with prior keratoplasty were excluded. A total of 80 swept-source AS-OCT scans *(CASIA2; Tomey Corp. Nagoya, Japan)* from 68 participants were acquired either immediately after surgery, one week after surgery, or both. Each scan consists of 16 images (B-scans) acquired in a radial pattern, corresponding to 1280 B-scans in total.

AS-OCT scans were preprocessed similar as reported in chapter 2. In brief, a deep learning-based localization model was applied to each B-scan to identify the scleral spur, a landmark in the anterior chamber of the eye [9]. B-scans were horizontally aligned and cropped based on the scleral spur locations, centering around the corneal apex (Figure 3.1). Final crop sizes were 960×384 pixels (width by height) with a pixel size of $15.0\,\mu$m. For a detailed description of the participant characteristics and processing methodology, we refer the reader to [69].

For each B-scan, the anterior corneal interface was annotated inside a 12 mm diameter from the radial center. A diameter of 10 mm was used for the posterior interface. Partial DMEK graft detachments were excluded from posterior interface annotations. The data set was randomly split on a participant level in a training set of 752 images, a validation set of 208 images and a test set of 320 images. B-scans of the training and validation set were annotated by one of three observers under supervision of a cornea specialist. The test set was annotated by all three observers to assess inter-observer variability.

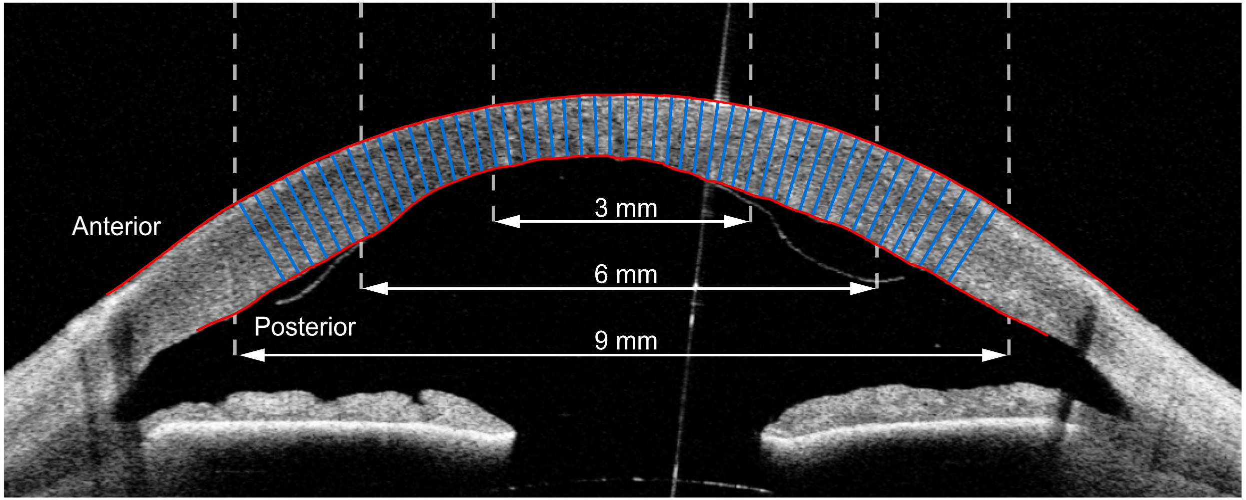

Figure 3.1: Single image (B-scan) from an AS-OCT scan, showing the cornea and the anterior chamber. This B-scan was cropped centrally and horizontally aligned as reported by Heslinga & Alberti [69]. Manual delineations of the corneal interfaces are shown in red. Corneal thickness is measured as the distance between the anterior and posterior interface, perpendicular to the anterior interface. The blue lines illustrate a subset of these thickness measurements. For evaluation of the thickness measurements, we distinguish the central 3 mm, 6 mm, and 9 mm diameter with respect to the corneal apex.

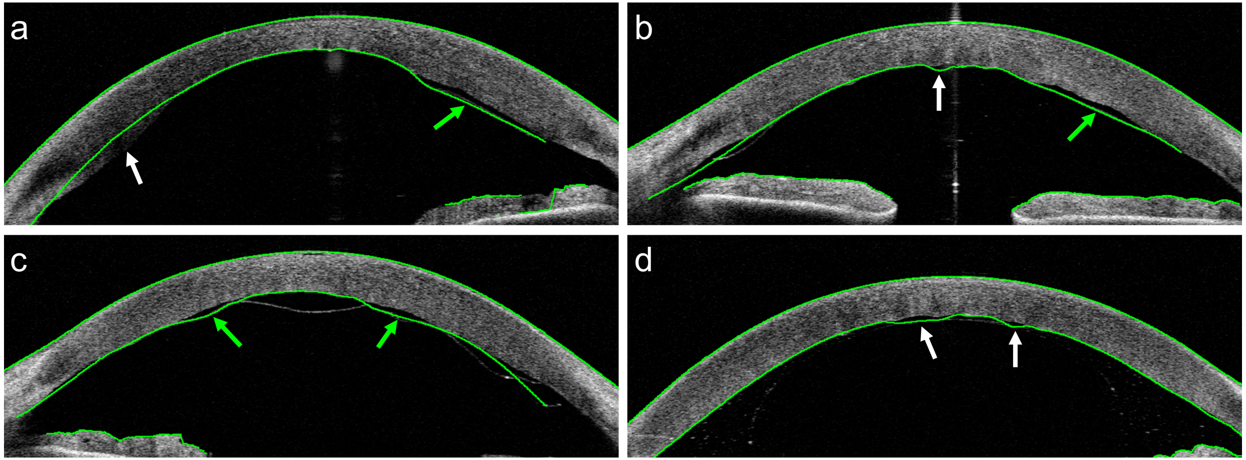

Figure 3.2: AS-OCT B-scans collected from patients after DMEK surgery. The green lines represent delineations of the (corneal) interfaces by the built-in software of the OCT system. These examples were selected to show the types of delineation errors encountered. In (a), (b), and (c) the delineation partly follows the DMEK graft (green arrows) instead of the posterior interface. Other types of mistakes are indicated by white arrows: (a) Some of the posterior part of the cornea is missed. (b) The delineation does not follow the irregularly shaped interface in the center. (d) The system confuses the boundaries of the gas bubble used in DMEK with the posterior corneal interface.

Table 3.1: Mean absolute error in μm of corneal thickness predictions on test set. Comparisons represent deep learning models versus (vs) annotations. Mean ± SD of 5 training repetitions. p-values were calculated by one-way ANOVA and represent the chance that model performances are similar. CNN w. dim. red. = CNN with dimension reduction.

	Models vs. annotations			
Diameter	Patch-based CNN	U-Net	CNN w. dim. red.	p-value
3 mm	14.40 ± 0.69	13.94 ± 0.38	13.94 ± 0.25	0.26
6 mm	14.80 ± 0.51	13.84 ± 0.22	14.17 ± 0.22	< 0.01
9 mm	15.50 ± 0.59	13.98 ± 0.15	14.40 ± 0.15	< 0.01

3.2.2 Thickness measurements

Three deep learning-based models were trained to locate the anterior and posterior corneal boundaries: (1) a patch-based convolutional neural network (CNN), (2) a U-Net [128] based model, and (3) a CNN with dimension reduction. Details about the model architectures and training process are provided in the methods section.

Corneal thickness was measured perpendicularly to the anterior interface (see Figure 3.1), similar to [92]. For each B-scan of the test set, we evaluated thickness for every pixel on the anterior interface inside a 3 mm, 6 mm, and 9 mm diameter. Corneal thickness estimates by the deep learning models were compared with all three sets of annotations (960 in total). Mean absolute errors (MAE) (shown in Table 3.1) are very similar for the three deep learning models and across different diameters. The smallest error was found for the U-Net model for the 6 mm diameter (13.84 μm), while the largest error was found for the patch-based CNN for the 9 mm diameter (15.50 μm). Apart from the latter, all mean absolute errors are smaller than one pixel (15.0 μm). The small standard deviations shown in Table 3.1 show the high repeatability over multiple training runs. In addition, we calculated the standard deviation over the mean absolute errors across the B-scans. Averaged over five training runs, these standard deviations for the central 9 mm diameter were 4.35 μm (patch-based), 4.77 μm (U-Net), and 4.90 μm (CNN with dimension reduction).

We investigated inter-observer variability by comparing the corneal thickness annotations between observers. The results of this comparison are shown in Table 3.2. Only for the combination of observer 1 vs 3, the mean absolute error is similar to that obtained with the deep learning models (13.66 - 14.69

Table 3.2: Mean absolute error in μm of corneal thickness measurements on test set. Comparisons represent deep learning annotator versus (vs) annotator.

	Inter-observer comparison		
Diameter	1 vs 2	1 vs 3	2 vs 3
3 mm	23.49	14.69	17.95
6 mm	23.71	13.91	18.80
9 mm	23.39	13.66	19.26

μm), while the differences between the other combination of observers are substantially larger (17.95 - 23.71 μm). In addition, a cornea specialist manually assessed all delineations of the test set by the built-in software (version 3H.1) of the OCT system. In 90 out of 320 (28%) B-scans a delineation mistake occurred that resulted in a clinically relevant thickness inaccuracy within the central 9 mm. Out of these 90 B-scans, 74 (82%) included an inaccuracy that overlapped with the location of a partial DMEK graft detachment. For 37 (12%) B-scans the thickness errors were considered severe. Sixteen out of 20 (80%) AS-OCT scans contained at least one B-scan with a clinically relevant thickness inaccuracy, resulting in an incorrect pachymetry map.

3.2.3 Outlier analysis

We further inspected the origin of the deviations in thickness measurements between the deep learning models and the manual annotations by investigating the cases with the largest deviations. Figure 3.3 shows two example outliers with annotations and delineations by the CNN with dimension reduction. In Figure 3.3a the remnant tissue from the host or donor prevents the graft from completely attaching at the right posterior side of the cornea. The tissue was correctly excluded from the annotation, but included in the delineation by the network.

Another example is presented in Figure 3.3b, where the enlarged region shows a shortfall in detail of the annotation compared to the network delineation. Note that the graft is not entirely attached at the right side of the posterior interface, which was correctly recognized by the network and mistakenly included in the annotation.

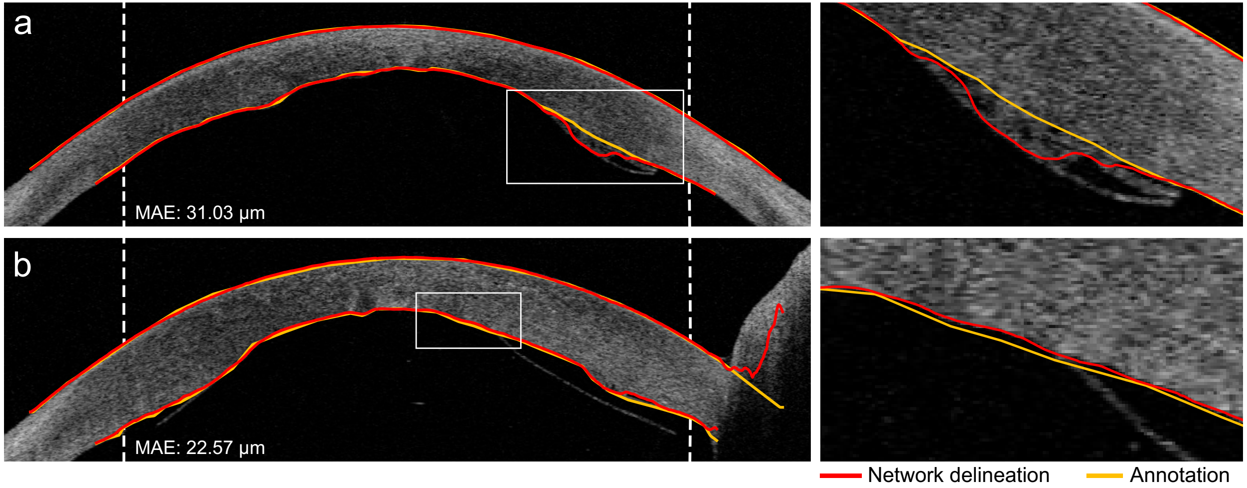

Figure 3.3: Two examples of B-scans including the annotations and delineations by the CNN with dimension reduction with substantial deviations in predicted thickness. The rectangular areas are enlarged and displayed to the right of the B-scan. Vertical dashed lines indicate the 9 mm diameter. Note that the thickness was not evaluated outside of the 9 mm diameter.

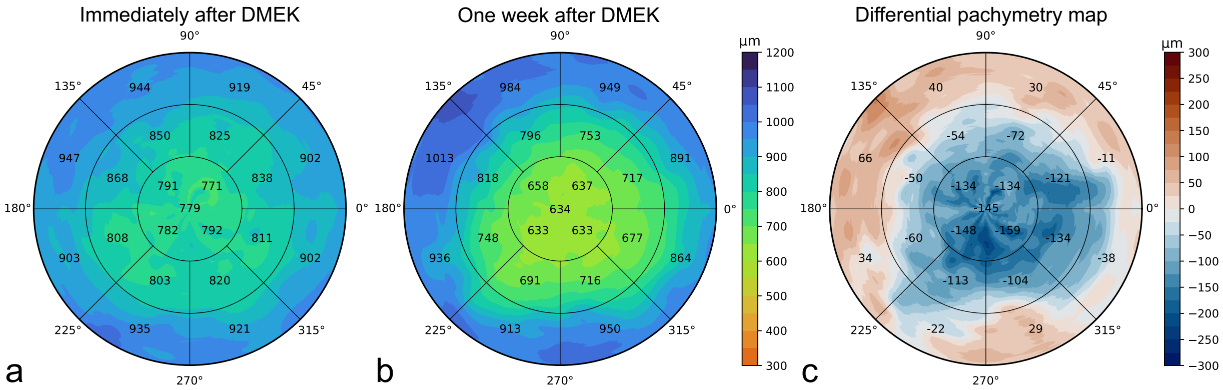

Figure 3.4: Example of pachymetry maps from one participant in the test set. (a) Pachymetry map of AS-OCT scan acquired immediately after DMEK; (b) Pachymetry map of AS-OCT scan acquired one week after DMEK; (c) Differential pachymetry map of difference in corneal thickness between (b) and (a).

3.2.4 Pachymetry mapping

Corneal thickness measurements from 16 radial B-scans were combined to construct pachymetry maps as shown in Figure 3.4. The pachymetry map was divided into three circular regions with diameters of 3 mm, 6 mm, and 9 mm. The outer two rings were divided into octants where the average thickness is displayed. The inner circle displays the average of the four quarters, as well as the average apex thickness inside the central 1 mm diameter. Thickness values were mapped to corresponding colors of a discrete colormap. Similar to conventional pachymetry colormaps the corneal thickness at 600 μm is displayed in green, thinner regions in red, and thicker regions in blue [21, 118].

3.2.5 Effect of training set size

The precise manual annotation of the corneal boundaries is a time-consuming process. For future projects it is useful to know whether a smaller set of annotated data can be used to obtain similar results. We therefore investigated the effect of the size of the annotated training set on the quality of the thickness measurements. Table 3.3 shows thickness measurement errors on the test set B-scans when the deep learning models are trained with 100%, 50%, 25%, and 10% of the original training set. For the patch-based CNN and the CNN with dimension reduction we found only a marginal increase in the mean absolute error (of about 2 μm) for the whole central 9 mm range, when trained with only 10% of the data. In contrast, the U-Net model performance did decrease more substantially for the 9 mm range when trained with 10-50% of the data. Further inspection learned that this was caused by some substantial error in the 6-9 mm range for a small portion of the B-scans.

Table 3.3: Mean absolute error in μm of corneal thickness predictions on test set for varying partitions of the training set. Mean ± standard deviation of 5 training repetitions. p-values were calculated by one-way ANOVA and represent the chance that model performances are similar.

Diameter	Training data	Patch-based CNN	U-Net	CNN with dim. red.	p-value
3 mm	100%	14.40 ± 0.69	13.94 ± 0.38	13.94 ± 0.25	0.26
	50%	14.28 ± 0.33	14.36 ± 0.51	15.04 ± 0.69	0.08
	25%	15.04 ± 1.13	16.22 ± 1.45	15.03 ± 0.49	0.19
	10%	15.93 ± 1.30	16.19 ± 1.38	16.76 ± 1.18	0.60
6 mm	100%	14.80 ± 0.51	13.84 ± 0.22	14.17 ± 0.22	< 0.01
	50%	14.65 ± 0.39	14.30 ± 0.46	15.10 ± 0.47	0.04
	25%	15.16 ± 0.95	15.25 ± 0.86	14.92 ± 0.25	0.77
	10%	16.08 ± 1.39	15.35 ± 0.74	15.88 ± 0.87	0.54
9 mm	100%	15.50 ± 0.59	13.98 ± 0.15	14.40 ± 0.15	< 0.01
	50%	15.23 ± 0.43	17.25 ± 5.97	15.24 ± 0.36	0.58
	25%	15.71 ± 0.89	20.53 ± 7.94	15.12 ± 0.20	0.17
	10%	16.76 ± 1.73	21.12 ± 6.24	16.08 ± 0.76	0.11

3.2.6 Effect of data split

We studied the effect of heterogeneity within our data set using different splits of the data into training, validation and test sets. Next to the original split of the data, two more unique splits were created. Results of the experiment are shown in Table 3.4. For all three deep learning models, the mean absolute errors over the three splits were similar or slightly smaller than for the original test set.

Table 3.4: Results for multiple splits of the image data. Presented is the mean absolute error ± standard deviation in μm over 3 unique test sets. p-values were calculated by one-way ANOVA and represent the chance that model performances are similar. CNN w. dim. red. = CNN with dimension reduction.

Diameter	Patch-based CNN	U-Net	CNN w. dim. red.	p-value
3 mm	13.83 ± 0.84	12.53 ± 1.42	13.36 ± 0.50	0.34
6 mm	14.39 ± 0.71	12.58 ± 1.31	13.34 ± 0.48	0.12
9 mm	15.26 ± 0.88	12.77 ± 1.26	13.68 ± 0.50	0.04

3.3 Discussion

This research shows the feasibility of automated corneal thickness measurements in post-DMEK AS-OCT scans using deep learning. While our data set contains many examples of irregularly shaped corneas and partly detached DMEK grafts, all models are able to measure corneal thickness with an average error of 13.98 to 15.50 μm for the central 9 mm range. In comparison with a typical central corneal thickness of 540 μm [65, 98], this corresponds to an error of less than 3%. The quality of our thickness measurements is at least on par with manual annotations, as indicated by the inter-observer distance of 13.66 to 23.39 μm for the central 9 mm range. Based on the accurate thickness measurements, detailed pachymetry maps can be constructed. The preprocessing based on the scleral spur locations largely eliminates spatial translation in the coronal plane and sagittal rotation between follow-up scans [69]. Our method does not correct for coronal rotation, but we expect the effect of this type of head tilt or eye rotation to be small. The mean localization error for a single scleral spur point was previously reported to be 0.155 mm. Given that the corneal apex location estimate follows from an average of 32 scleral spur point estimates, the registration error with respect to the corneal apex should be smaller than 0.155 mm.

Since follow-up scans can be registered, differential pachymetry maps can be constructed to monitor thickness changes. This may enable a more comprehensive understanding of the restoration of the endothelial function after DMEK, where thickness often varies throughout different regions of the cornea and the restoration of corneal thickness is associated with success of the procedure [156]. Typically, only the central corneal thickness (CCT) is reported, while this single parameter does not necessarily reflect restoration of the full cornea after DMEK. The DMEK graft is about 8.5 mm [124] and partial graft detachment happens most frequently in the peripheral region [25,37]. Detachment can sometimes be ambiguous, even in high-quality OCT imaging, as the graft can be close to the inner cornea yet not attached [69]. Local changes in corneal thickness could then be indicative of corneal restoration and thus graft attachment. The differential pachymetry maps presented in this research enable both qualitative and quantitative progression tracking options within the central 9 mm range, covering the whole region of the DMEK graft.

The data for this study only included post-DMEK AS-OCT scans of patients with previous Fuchs' endothelial dystrophy or pseudophakic bullous keratopathy. The application of the here presented deep learning models for scans from other pathologies or taken in different centra requires further research.

Three deep learning methods were compared in this research. Despite substantial differences in the approaches, the results for the different models were all at least on par with manual annotations. This could indicate that delineating the corneal boundaries is well-defined and relatively easily solvable using deep learning. This finding is in line with other research aiming at delineating layers in ophthalmic OCT imaging [41,51,63,82,83,93,129,164]. Based on the thickness results for the models trained with 100% of the training data, none of the models seems to outperform the other. However, construction of a pachymetry map with the patch-based CNN takes considerably longer because of the large number of patches involved and the extensive post-processing steps. Kugelman et al. [82] compared several patch-based and fully-convolutional deep learning methods for segmentation of retinal layers and the chorio-scleral interface. They found that all models performed comparable for retinal layers, which are considered to have well-defined boundaries. However, for the more ambiguous chorio-sclerar interface, the fully-convolutional methods outperformed the patch-based methods. The authors attributed this to the additional context available to the network when the whole image is processed at once.

Results of the deep learning models could not directly be compared with the delineations by the built-in software of the OCT system. Nevertheless, 28% of B-scans were found to contain delineations mistakes by the built-in software leading to a clinically relevant thickness inaccuracy. Moreover, 80% of the AS-OCT scans of the test set contained at least one B-scan with a thickness inaccuracy of clinical relevance. We observed that these errors often occurred at locations of high clinical relevance, such as an irregularly shaped corneal center, or where the DMEK graft detached.

Training the models with smaller partitions of the data provides insight in the value of adding extra annotated data. Since no improvements were obtained by using 100% of the available training data compared to only 50%, it can be concluded that the performance has saturated, and additional data would not contribute to much further improvement. An exception could be the addition of data from rare cases or examples that led to errors in the current test set (e.g. Figure 3.3a). For the U-Net model, training with 10% or 25% of the training data did result in an increase in the thickness error in the 6 to 9 mm region. Further inspection revealed that this was due to a small number of B-scans for which the segmentation failed. Interestingly, for the CNN with dimension reduction and the patch-based CNN, even training with 80 B-scans from 5 patients does not seem to reduce the performance of the thickness measurements substantially (MAE of 16.08 μm and 16.76 μm, respectively). These results indicate that future projects on delineation of corneal boundaries could already be developed with less annotated data, yet still obtain reasonable results. In addition, the experiment with different data splits showed that similar performance could be achieved when different splits of the data were used.

The high resolution of the AS-OCT combined with deep learning for automated image processing supports fast and accurate analysis of the corneal thickness after DMEK. The here presented (differential) pachymetry maps enable tracking of local corneal thickness changes indicative of corneal restoration. As such, these tools can contribute to the ongoing research efforts towards further improvement of the DMEK procedure and management of the postoperative regime.

3.4 Methods

3.4.1 Models & Training

We implemented three different deep learning models based on recent successful applications related to segmentation or interface delineation in ophthalmic

OCT imaging. The first model is based on the patch-based approach by Wang *et al.* [164] which was adopted from earlier work on CNNs and graph search methods by Fang *et al.* [51]. The patch-based model was previously used for the identification of five different retinal interfaces in spectral domain OCT images [164]. Using patches of 33 × 33 pixels and a relatively shallow network of 5 trainable layers, Wang et al. achieved localization accuracies of 89-98%. The second model is based on the U-Net architecture which has become the de facto standard method for medical image segmentation [128]. Multiple adaptations of U-Net were evaluated by dos Santos et al. to segment three corneal layers in images captured with a custom-built ultrahigh-resolution OCT system [41]. Using cross-validation, a mean segmentation-accuracy of 99.56% was achieved. The third model was also inspired by U-Net, but modified by Liefers et al. to reduce the dimensionality and output one-dimensional arrays with y-locations of three retinal layers in OCT images [93]. For the localization of these retinal layers, the authors obtained a mean absolute difference between the predictions and annotations of 1.31 pixels.

For our application of corneal interface localization, both the patch-based approach and the CNN with dimension reduction allow for direct delineation of the corneal interfaces. In contrast, a U-Net approach is used to segment the cornea and requires a post-processing step to obtain the interface delineations from the segmented mask. Details about the model adaptations, implementations, and training procedures are described below for each model. Depending on the model requirements we also adapted the preprocessing and post-processing steps. All models were implemented in Keras [29] with TensorFlow [1] backend and optimized using Adam [78].

Patch-based CNN

The architecture of the patch-based model was similar to that of Wang et al. [164] with 2 modifications: (1) 5 × 5 convolutional layers were replaced by two 3 × 3 convolutional layers as factorization is considered more efficient [146]; (2) 3 × 3 average pooling operations in the final two layers were replaced with 2 × 2 max pooling operations. From the full images of training and validation sets, patches of 33 × 33 pixels were extracted for each x-coordinate where the interfaces had been annotated (center 12 mm for the anterior interface and center 10 mm for the posterior interface). Anterior and posterior patches were sampled using the respective annotations as center pixel locations. Similarly, for each x-coordinate within the central 12 mm, a non-interface patch was constructed for one random pixel not part of the interface annotations. All patches

(1.70 million for the training set and 0.47 million for the validation set) were extracted prior to training to speed up the training process.

The model was optimized by minimizing the categorical cross-entropy between the pixel ground truth and model predictions. Online data augmentation was added by rotating the patches with a maximum of 30 degrees. Based on preliminary experiments, the model was trained for 20 epochs with a variable learning rate: 0.001 for epoch 1 to 12, 0.0001 for epoch 13 to 16, and 0.00001 for epoch 17 to 20.

For evaluations on the test set, patches were extracted for all pixels within the center 12 mm (width) and processed by the trained model, predicting either anterior interface, posterior interface, or background for the center pixel of the patch. Based on preliminary results on the validation set, the following post-processing steps were performed for the pixels identified as interface: (1) small connected regions (0 - 250 pixels) were removed; (2) the largest connected region was considered to be true; (3) other regions were considered to be part of the true prediction only when those would be at the same height as the largest connected region; (4) per interface, y-values of positive pixels were averaged to obtain a single value per x-coordinate; (5) any remaining gaps were filled using linear interpolation of adjacent interface locations.

U-Net

As an alternative to directly delineating the interfaces, a U-Net [128] was implemented to segment the whole cornea. The U-Net consisted of the standard 4 downsampling (and upsampling) segments and we included batch normalization and residual layers to accelerate training. Binary masks of the cornea were created using the interface annotations. As a preprocessing step, the original images and masks were cropped to 800×256 pixels (width by height) and split into a superior and inferior half. This step was included to reduce the size of the input to the U-Net while doubling the number of training examples.

For optimization of the U-Net we experimented with different loss functions (Dice, binary cross-entropy, and weighted binary cross-entropy) and learning rate schedules. We found only minor differences in the results on the validation set and used binary cross-entropy for the final model. We trained for 30 epochs with an initial learning rate of 0.001 that was divided by two at every 3 epochs. We also experimented with data augmentation (brightness adaptation and addition of Gaussian noise) but did not identify any improvements on the validation results.

For evaluations on the test set, the inferior an posterior crops were processed by the trained U-Net and combined. From the predicted mask, the maximum and minimum y-values were used to reconstruct the anterior and posterior interface, respectively. In contrast to the patch-based model, the predicted y-values of the interfaces only consisted of integers.

CNN with dimension reduction

The architecture of the third model was designed by Liefers et al. [93] to return a one-dimensional array of y-coordinates for a two-dimensional image as network input. The model consists of a downsampling and upsampling path to incorporate a large contextual region. While U-Net uses direct shortcut connections to provide local context, this architecture resorts to so-called funneling subnetworks between the downsampling and upsampling path to resolve the mismatch in activation map height. The original network architecture was designed for images of 512×512 pixels. To avoid unnecessary computations, we adapted the architecture to work for images of 512×256 pixels. The downsampling path and all funneling subnetworks therefore contain one less downsampling operation and the upsampling path one less upsampling operation. 1×1 residual blocks at the lowest level were replaced by 1×3 residual blocks in our architecture. Furthermore, we experimented with the addition of batch normalization layers, which did not result in improved performance.

As a preprocessing step, the original images were cropped to 960×256 pixels (width by height) and split into a superior and inferior half with some overlap along the horizontal axis, resulting in model inputs of 512×256 pixels. Since the annotations did not span the entire width of 512 pixels, the network output layer was cropped to only include the annotated positions. The mean squared error between the annotation and predicted delineation was used as loss function. Training was done for 200 epochs with a learning rate set to 0.0002 at the start and divided by two after every 50 epochs. Based on preliminary experiments, we used the following data augmentation: B-scans were translated (≤10 pixels) or rotated (≤12 degrees) before inferior and superior crops were made. We also added uniform noise (≤0.05), Gaussian blurring with $\sigma \leq 1$, and sigmoidal contrast changes with a gain between 4 and 5.

For evaluations on the test set, the superior and inferior crops were processed and combined by averaging the central overlapping section.

3.4.2 Thickness measurements & evaluation

Outputs of the deep learning models were y-values, describing the height of the anterior and posterior interface for each x-coordinate within the central 9 mm. The posterior interface of the cornea generally contains more irregular shapes after DMEK. We therefore measured corneal thickness perpendicularly to the anterior interface (see Figure 3.1), similar to [92]. First, the coefficient of proportionality was determined for the anterior interface. A 71 point moving average filter was used to reduce the effect of small deviations from the general curvature. We found that the proportionality coefficient was still affected by local irregularities after filtering with smaller filters, whereas larger filters introduced inaccuracies near the sides. The distance was then measured perpendicularly to a tangent with the corresponding coefficient of proportionality for every pixel on the anterior interface inside a 9 mm diameter. Performance of the models was measured by comparing the thicknesses predictions with the thicknesses following from the three sets of manual annotations. The mean absolute error was then calculated for a 3 mm, 6 mm, and 9 mm diameter. To obtain a measure of variation, we trained all models five times from scratch using different random seeds. Mean absolute errors and sample standard deviations shown in Table 3.1 and Table 3.3 were based on these five repetitions. Comparison of the model performances was done for each region (3 mm, 6 mm, and 9 mm) to test for statistically significant differences. p-values were calculated based one a one-way ANOVA and represent the chance that model performances were similar.

For assessment of the interface delineations errors by the built-in software of the OCT system, all B-scans of the test set were semi-quantitatively assessed by a cornea specialist. With the built-in drawing tool, missed parts of the cornea or areas mistakenly classified as cornea were selected. This was done approximately for the central 9 mm diameter although these B-scans were not centered or horizontally aligned. Sometimes the posterior delineation was not complete for the peripheral cornea. In such cases no extra misclassified area error was added; these regions were simply ignored. B-scans with a total incorrectly classified area of more than 0.1 mm^2 were considered to contain a clinically relevant inaccuracy. When this area was larger than 0.25 mm^2 the inaccuracy was considered severe.

3.4.3 Pachymetry mapping

Corneal pachymetry maps were constructed with the thickness profiles of the 16 radial B-scans (cross-sections) of a single AS-OCT scan. For the maps shown in Figure 3.4 we used the CNN with dimension reduction to obtain one-dimensional thickness profiles of length 600 pixels corresponding to cross-sections within the central 9 mm. These arrays were plotted on a two-dimensional polar coordinate axis based on the spatial configuration of the cross-sections in the original AS-OCT scan. We then used cubic interpolation along the polar coordinates to fill in space between the cross-sections.

Before applying the deep learning models, B-scans were cropped based on the scleral spur locations, similar to [69]. First, a deep convolutional neural network was trained to find the scleral spur in the right and left side of each B-scan. Per full AS-OCT scan, an ellipse was fitted through the scleral spur points of all 16 B-scans to ensure that the points lie in the same plane and to refine point locations. The refined scleral spur locations were used to horizontally align the B-scans by assuming that the left and right scleral spur points should be at the same height in the image. Then, the B-scan was centered by assuming that the corneal apex is located equidistant from both scleral spur points.

Since this preprocessing step automatically centers the (cropped) B-scans, follow-up scans and pachymetry maps are registered with respect to the corneal apex. The differential pachymetry map shown in Figure 3.4 was obtained by subtracting the pachymetry map obtained after the procedure from the map obtained postoperative day 7.

3.4.4 Reduced training set & cross-validation

Partitions of the training set were made by randomly selecting all B-scans from a subset of the study participants. For 50% of the training set this equaled 24 participants (384 B-scans). For 25% 12 participants (192 B-scans) and for 10% 5 participants (80 B-scans). Partitions were randomly sampled for each of the five training repetitions, and partitions were the same for each deep learning model. All other training parameters were similar as for the 100% training set size models.

In order to examine the influence of the data split into training, validation, and test sets, two extra splits of the data were created, next to the original split. Splitting the data was done on a participant-level, with 18 participants in each test set and without overlap in participants between the test sets of the three splits. Training was done once per split, using the same hyperparameters as described for the other experiments. For the original split, only the first of five training repetitions was used. Since only a single set of manual annotations was available for the original training and validation data, we opted to also use only a single annotation set for the original test data set for this experiment.

In the next chapter, instead of using post-operative AS-OCT scans, we focus on intra-operative OCT scans. For these intra-operative images, we aim to automatically assess the orientation of the DMEK graft.

Chapter 4

DMEK graft orientation in intraoperative OCT

This chapter is based on:
M.B. Muijzer, F.G. Heslinga, F. Couwenberg, H.J. Noordmans, A. Oahalou, J.P.W. Pluim, M. Veta, R.P.L. Wisse. Automatic evaluation of graft orientation during Descemet membrane endothelial keratoplasty using intraoperative OCT. *Biomedical Optics Express*, 13(5):2683-2694, 2022.

Abstract

Correct Descemet's membrane endothelial keratoplasty (DMEK) graft orientation is imperative for success of DMEK surgery, but intraoperative evaluation can be challenging. In this chapter, we present a method for automatic evaluation of the graft orientation in intraoperative optical coherence tomography (iOCT), exploiting the natural rolling behavior of the graft. The method encompasses a deep learning model for graft segmentation, post-processing to obtain a smooth line representation, and curvature calculations to determine graft orientation. For an independent test set of 100 iOCT-frames, the automatic method correctly identified graft orientation in 78 frames and obtained an area under the receiver operating characteristic curve (AUC) of 0.84. When we replaced the automatic segmentation with the manual masks, the AUC increased to 0.92, corresponding to an accuracy of 86%. In comparison, two corneal specialists correctly identified graft orientation in 90% and 91% of the iOCT-frames.

4.1 Introduction

Descemet's membrane endothelial keratoplasty (DMEK) is the preferred posterior lamellar keratoplasty procedure for treating cases of symptomatic irreversible corneal endothelial cell dysfunction [116,144]. Posterior lamellar surgeries constitute the majority of grafting procedures in the developed world [108]. The thin (~30 μm) and vulnerable DMEK graft – consisting of the Descemet's membrane and endothelium – is inserted as a roll and unfolded in the anterior chamber of the eye before fixation on the posterior surface of the recipient cornea [34]. A correct orientation of the graft – with the endothelium facing away from the cornea – is imperative. An inadvertently incorrectly positioned graft (i.e. upside-down) will result in severe corneal edema, damage to the graft's endothelial cell layer, and the subsequent need for repeated surgery [14].

The assessment of the graft's orientation can be challenging and several methods have been described to aid the surgeon in determining the orientation. Currently, the Moutsouris sign, ink-stamps, and circular cuts are used to determine intraocular graft orientation [34, 106, 157, 165]. However, poor visualization of the anterior chamber and graft hinders a proper assessment [32, 117, 140]. In addition, the presence of the Moutsouris sign is not always self-evident and both stamps and cuts damage the graft resulting in endothelial cell loss. More recently, intraoperative optical coherence tomography (iOCT) has been used to determine graft orientation, as the iOCT signal is not perturbed by corneal edema [32,117,140]. Residual stromal fibers in the Descemet's membrane of the DMEK graft result in a distinctive inward curve of the graft's ends indicative of a correct orientation, which can be visualized and assessed using iOCT (Figure 4.1) [107, 143]. This natural curling behavior of DMEK grafts can be well appreciated on the iOCT image, thereby preventing the need to use manipulation, cutting, or marking to determine the graft orientation, thus preventing endothelial cell loss.

Several studies have reported on the use of iOCT during DMEK surgery for determining the orientation of the graft. In all studies the graft orientation could be correctly determined based on the inward rolling of the graft edges visible on the cross-sectional iOCT image [32, 109, 117, 130, 140, 143]. Importantly, the surgeon was able to assess the graft orientation in cases where assessment of the Moutsouris sign or S-stamp was challenging or not possible [32, 117, 140]. However, manual assessment of graft orientation on iOCT images can be time consuming and prone to interpretation errors. In particular, when OCT image quality is suboptimal or the graft edges display little inward rolling. We believe an automated tool will aid the surgeon in fast and accurate evaluation of the orientation, thereby improving surgical workflow and reducing the risk of errors.

In this chapter, we present an automated image analysis method for evaluation of the DMEK graft orientation using iOCT. The method includes a deep learning-based segmentation model to extract the DMEK graft from the iOCT scan. Then the degree of inward rolling by the graft is assessed and related to graft orientation.

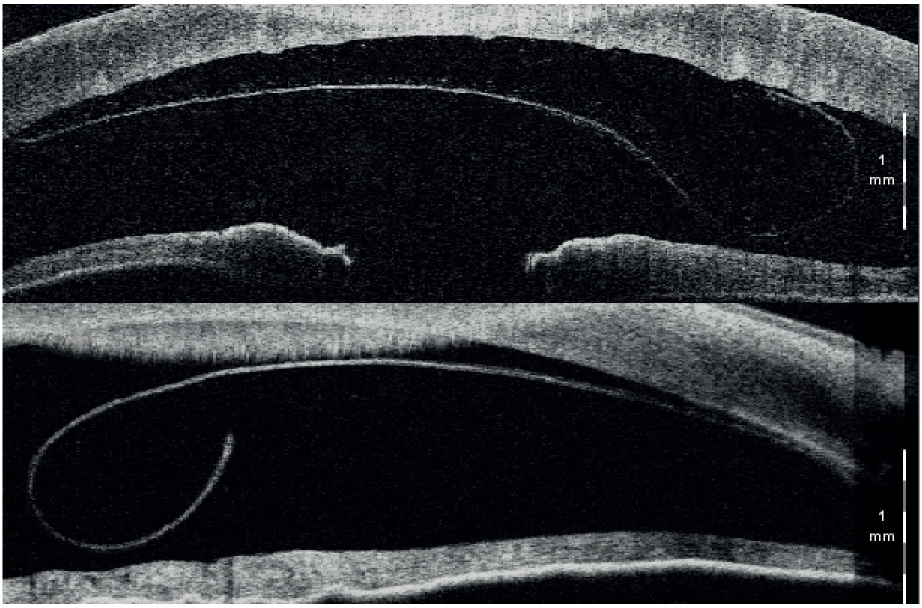

Figure 4.1: Two cross-sectional intraoperative OCT scans of the cornea. The natural rolling motion of the graft in Descemet's membrane endothelial keratoplasty (DMEK) can be used to determine the graft's orientation. The top image depicts a correctly oriented DMEK graft, indicated by the distinctive upward curve towards the recipient's cornea. The bottom image depicts an incorrectly oriented graft (i.e., upside-down), indicated by the curling motion away from the recipient's cornea.

4.2 Methods

4.2.1 Data and preprocessing

All OCT-scans in this study were acquired during DMEK surgery at the ophthalmology department of the University Medical Centre Utrecht between May 2016 and October 2020 using the "No-Touch" technique for DMEK as described by Dapena et al. [34]. DMEK grafts were cultured and provided pre-cut by the Euro Cornea Bank (Beverwijk, the Netherlands) and Amnitrans (Rotterdam, the Netherlands). During surgery, iOCT-scans of the anterior segment were made with a commercially available spectral domain microscope integrated OCT system (Zeiss Lumera 700 RESCAN, Carl Zeiss Meditec, Jena, Germany), using the two-line cross-sectional setting. The iOCT system has a wavelength of $850\,\text{nm}$ and an axial resolution of $5.5\,\mu\text{m}$. The system acquires 25 two-line cross-sectional scans per second. This study was performed in accordance with the Declaration of Helsinki and Dutch law regarding research involving human subjects. Ethical approval for this study was waived by the Ethics Review Board of University Medical Center Utrecht (METC no. 18-370).

iOCT-scans of the DMEK procedures are embedded in the surgical video feed. The video feed was qualitatively reviewed for scan quality and visibility of the graft during determination of graft orientation (i.e., before adhering the graft). Scans were excluded if the graft was not visible at all or not unfolded. Included iOCT-scans were manually extracted from the video feed using the FFmpeg tool (version 3, 2016, FFmpeg Developers). Each cropped frame contained a single cross-sectional iOCT-scan (iOCT-frame). The ground truth of the graft orientation, either correctly oriented or upside-down, in each iOCT-frame was set by an experienced grader (M.B.M.) who had access to the preceding and follow-up frames and postoperative clinical information. The orientation of each graft was subsequently graded by two corneal surgeons (R.W. and A.O.) based on a single iOCT-frame and blinded for the outcome (i.e., without access to the preceding and follow-up frames or postoperative clinical information).

A total of 335 iOCT-frames from 89 DMEK surgeries were obtained; 127 iOCT-frames measuring 550×275 (width × height) pixels acquired before 1-1-2019 and 208 iOCT-frames measuring 610×275 pixels acquired from 1-1-2019 onwards. The more recently acquired scans were of better image quality due to an improved scan protocol and we selected 100 recent iOCT-frames from 21 patients as a test set for final evaluation of our models. All other iOCT-frames ($N = 235$) were used for development and optimization of the image analysis methods and will be referred to as the development set.

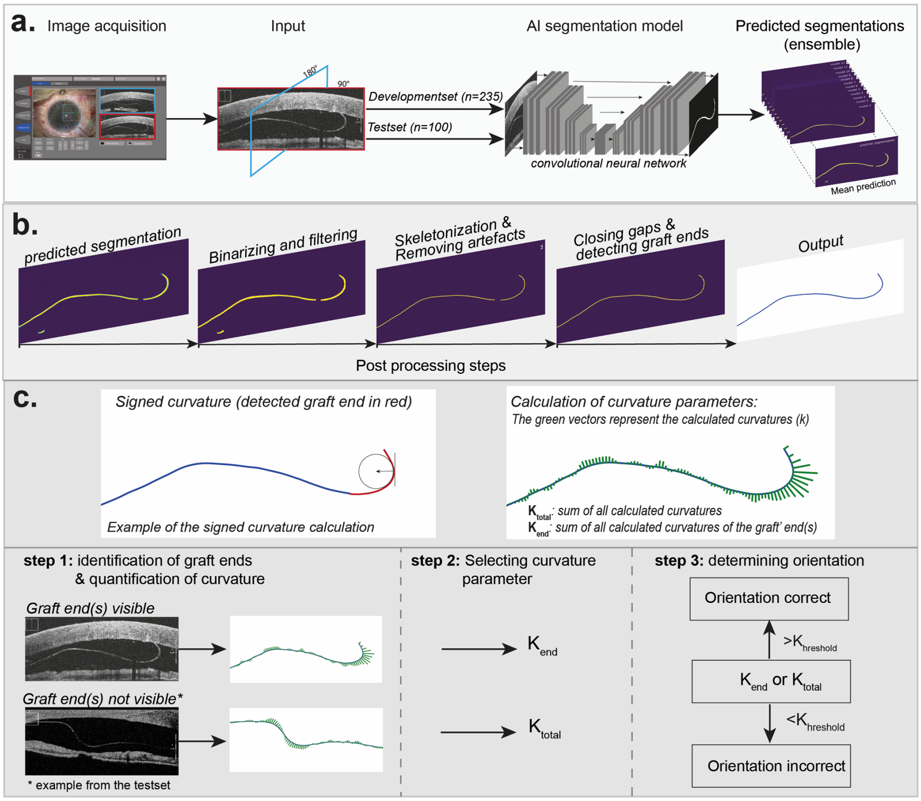
a.
Image acquisition
Input
AI segmentation model
Predicted segmentations (ensemble)
180°
90°
Developmentset (n=235)
Testset (n=100)
convolutional neural network
Mean prediction
b.
predicted segmentation
Binarizing and filtering
Skeletonization & Removing artefacts
Closing gaps & detecting graft ends
Output
Post processing steps
c.
Signed curvature (detected graft end in red)
Example of the signed curvature calculation
Calculation of curvature parameters:
The green vectors represent the calculated curvatures (k)
K_total: sum of all calculated curvatures
K_end: sum of all calculated curvatures of the graft' end(s)
step 1: identification of graft ends & quantification of curvature
Graft end(s) visible
Graft end(s) not visible*
* example from the testset
step 2: Selecting curvature parameter
K_end
K_total
step 3: determining orientation
Orientation correct
>K_hreshold
K_end or K_total
<K_hreshold
Orientation incorrect

Figure 4.2: A schematic representation of the pipeline of the intraoperative OCT DMEK graft orientation model. Shown in section (a) are the image acquisition process and the automatic segmentation model. The predicted segmentation is the mean of an ensemble of 12 deep learning models. Section (b) shows the key post-processing steps to obtain a one-pixel line representing the graft. In section C the left top image is a schematic representation of the signed curvature. The top right image shows the polygons fitted to the line representing the graft and the defined curvature parameters. The bottom images of section (c) shows the choice for selecting the curvature parameter and determining the orientation.

The development set was again divided into a training set ($N = 202$) and a validation set ($N = 33$) to determine the optimal model. The data split was done on a patient level to ensure no overlap exists between the train, validation, and test sets. The graft locations were manually annotated in the iOCT frames with marking points (image coordinates) along the graft and converting the resulting contour to a binary mask of an area containing the graft. Zero-padding was used to ensure all iOCT-frames were of width 610 for training the AI segmentation model. As a final preprocessing step, all frames were resized to 576 × 256 pixels for compatibility with the U-Net architecture. Our image analysis tool consists of three steps (Fig. 4.2). First, the area containing the DMEK graft was segmented from the iOCT-frame using a deep learning-based segmentation model. In the subsequent post-processing step, the resulting mask was converted into a one-pixel thick line representation of the graft. Artifacts and gaps in the line were removed and the graft's endings located. Finally, we build upon the work by Steven et al. to assess the curling behavior of the graft [143] and we relate curvature of the line segment to graft orientation. The predicted graft orientation was then compared to the ground truth and classification by the corneal surgeons.

4.2.2 Segmentation

For segmentation of the DMEK graft from the iOCT frame, we used a deep learning approach [84]. Our model consists of an ensemble of 2D U-Nets [128]. The U-Net architecture incorporates a large contextual region and has resulted in state-of-the-art performance for many biomedical image segmentation tasks [95]. Training was done using iOCT-frames and the corresponding manually annotated masks of the trainset ($N = 202$). Data-augmentation was used to expand the variability in appearance of the training data set. Augmentations included random affine transformations that were applied to the iOCT frames and corresponding mask annotations: translation (≤ 10 pixels), rotation (≤ 3 degrees), scaling (≤ 10%) and vertical reflection. In addition, we applied intensity shift (≤ 10/256), contrast shift (≤ 0.1) and addition of white noise (≤ 10/256) to the iOCT frames. Experiments with different learning rates and loss functions indicated that different models lead to different types of segmentation errors for the validation set ($N = 33$). We therefore constructed an ensemble of 12 U-Nets: Five models were trained using Dice loss and initial learning rates of 0.0001, 0.0002, 0.0003, 0.004, and 0.0005. Another seven models were trained based on a weighted binary cross-entropy (WBCE) loss, with a weight determining the relative penalty for misclassified foreground pixels (= DMEK graft) in comparison to background pixels. Beta values of 0.5, 1, 2, 4, 8, 12, and 16 were used, and the WBCE models were trained with an initial learning rate of 0.0003. All models were optimized with Adam for 3,500 iterations where the initial learning rate was multiplied by 0.3 every 1400 iterations [78]. Each U-Net in the ensemble provides a segmentation prediction and the final segmentation was obtained by taking the mean across the 12 segmentation maps (Figure 4.2a).

4.2.3 Post-processing

To ensure the graft is represented as a single smooth line, a post-processing algorithm was developed (Figure 4.2b) consisting of the following steps: (1) Median filtering (filter size = 2×2) to reduce noise; (2) Binarization to assign pixels to either background or graft class; (3) Skeletonization to obtain the topological skeleton (one-pixel thick) of the segmented areas [86]; (4) Removal of small islands (≤100 pixels) to get rid of small areas falsely identified as graft; (5) Morphological pruning to remove side-branches from the remaining skeletonized line segments. We implemented the pruning by finding the longest pathway for each segment and removing any pixels not belonging to these paths; (6) Closing of gaps between endings of line segments with a Euclidian distance less than 100 pixels, using a straight line. For the post-processing steps, all design choices and parameter selections were based on results for the validation set.

Next, the largest line segment was identified and the coordinates of every 15th pixel along the line were used to compute a parametric cubic smoothing spline curve. The parametrization was then used to resample 100 points along a smooth line representing the graft.

4.2.4 Graft orientation

To determine the orientation of the graft, we first assessed the rolling behavior of the graft. The rolling behavior can be measured as the signed curvature κ, similar to the previously described method by Steven et al. [143].

A Python implementation of the Matlab LineCurvature2D package [81] was used to calculate the local curvature at each of the 100 graft points obtained with the post-processing step (Figure 4.2c). Summing all local curvatures for the length of the graft (L), the total curvature (κ_{total}) can be calculated, taking into account the distance arc length steps (ds):

$$\kappa_{total} = ds \sum_{i=1}^{L} \kappa_i$$

We are however mostly interested in the graft curvature at the endings (κ_{end}), since this is typically used by our corneal specialists to determine graft orientation. The graft ending is here defined as the first and last 20% of the graft points:

$$\kappa_{end} = ds \sum_{i=1}^{20} \kappa_i + ds \sum_{i=81}^{100} \kappa_i$$

The curvature of a graft ending was only calculated if it was visible in the iOCT-frame. A graft end is classified as invisible (out of iOCT-frame bounds) when the first or last point of the calculated curve is within 10 pixels of the original iOCT-frame boundary. Prediction of the graft's orientation is primarily based on the curvature of the graft's endings κ_{end}. Alternatively, the overall curvature κ_{total} is used to determine the orientation only when: (1) both the graft's endings are not visible in the iOCT-frame or (2) the graft's endings show no curvature. To determine the orientation of the graft, the curvature of the graft was compared with a threshold value ($\kappa_{threshold}$). A graft with a curvature smaller than this threshold was considered incorrectly oriented.

4.2.5 Evaluation and statistical analysis

Performance of the automatic DMEK orientation model was evaluated for the test set iOCT frames ($N = 100$). The predicted orientation was compared to the ground truth orientation and a receiver operating characteristic (ROC) curve was determined by varying the $\kappa_{threshold}$ threshold. Sensitivity was defined as the accurate prediction of correctly oriented grafts while specificity represents true prediction of incorrectly oriented grafts (i.e., upside-down). For comparison of the automatic method with the corneal specialists, an operating point was chosen by setting a single value for $\kappa_{threshold}$, based on an optimal F1-score. The set $\kappa_{threshold}$ was used for all prediction methods. All statistical analysis were performed using R statistical software version 4.0.3 (CRAN, Vienna, Austria). The ROC plots were produced using the ROCR package (version 1.0-11).

Quality of the segmentations was evaluated using the Dice score. Additionally, we evaluated a pipeline that uses the manual annotated masks instead of deep learning-based segmentations. The post-processing of these segmentations was similar to the end-to-end pipeline, although steps (4) removal of pixel islands and (6) closing of gaps were skipped. This 'semi-automatic method' was evaluated on the test set as well as the recently acquired frames of the development set (n = 108), since these are comparable to the frames in the test set in terms of frame size and resolution.

4.3 Results

Of the 335 iOCT-frames included in this study, 255 frames contained correctly oriented grafts versus 80 incorrectly oriented grafts (i.e., upside-down). In 195 iOCT-frames the graft was free floating (i.e., no contact with other ocular structures) and in 134 iOCT-frames a mirroring artefact of the cornea was present, which (partially) overlapped with the graft in 65 iOCT-frames. Mean age of the graft donors was 74 years (range: 55 - 88). The indications were Fuchs endothelial corneal dystrophy (N = 79), Pseudophakic bullous keratopathy (N = 9), and graft failure (N = 1). Segmentation performance on iOCT-frames of the test set was similar across the 12 deep learning models, with Dice scores ranging from 0.72 to 0.74. For the ensemble, where the mean prediction of the 12 models was used, the Dice score was 0.75.

4.3.1 Performance of the DMEK orientation model

In Figure 4.3 the ROC curves are displayed for the DMEK orientation model using the deep learning-based segmentations (automatic method) and manually annotated grafts (semi-automatic method). Additionally, the performance of the corneal specialists is shown for both datasets. The automatic method achieves an AUC of 0.84, which is considered a good to excellent predictive power [99]. The semi-automatic method performs even better than the automatic method, with an AUC of 0.92 for both the development set and test set and is comparable to the performance of the corneal specialists using the same information (i.e., a single iOCT-frame). Causes for the gap in performance between the automatic and semi-automatic methods include segmentation and post-processing errors, which are described in detail in the qualitative analysis.

In line with the aim of this study – determining graft orientation using iOCT – the optimal trade-off between sensitivity and specificity was selected to determine $\kappa_{threshold}$ (Figure 4.3). The detailed results of the DMEK orientation model at $\kappa_{threshold}$ are shown in Table 4.1. The automated method was able to correctly identify the graft's orientation in the iOCT frames in 78% of the iOCT-frames in the test set and in 86% of the iOCT-frames for both the development and test set using manually segmented grafts. The automatic method achieved a high sensitivity (0.82) and moderate specificity (0.69). Thus, the model was

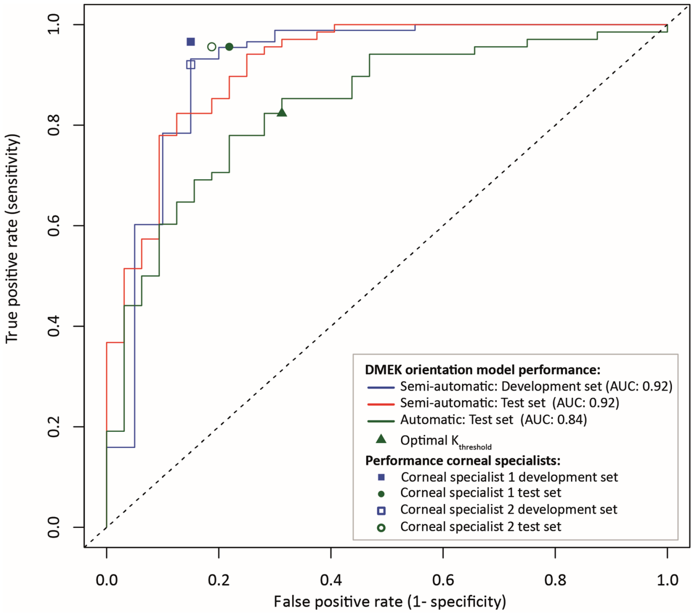

Figure 4.3: Receiver operating characteristic curves of the performance of the DMEK orientation model in the test set ($N = 100$) and the most recent frames of the development set ($N = 108$), obtained by varying the curvature threshold. The circles and squares represent the performance by the corneal specialists. The dashed 45-degree line constitutes a model with no discriminative power.

able to correctly classify the majority of the correctly oriented grafts, though it had only a moderate predictive power to correctly classify incorrect oriented grafts. Using the manually annotated grafts leads to slightly better sensitivity and markedly higher specificity compared to the automatic method. The outcomes of the semi-automatic methods were comparable to the performance of the corneal specialist (Table 4.1).

4.3.2 Qualitative analysis

All deep learning-based segmentations in the test set were qualitatively evaluated for errors in the predicted segmentation or post-processing. In 54 iOCT-frames a near perfect representation of the graft was achieved after post-processing

Table 4.1: Performance analysis of the orientation model and corneal specialists. AUC = area under the curve. *Development set consisting of only recently acquired frames measuring 610 pixels by 275 pixels were included for comparability with the test set.

	Images	Segmentation	Sensitivity	Specificity	Accuracy	AUC
DMEK orientation model	108*	semi-automatic	0.86	0.85	0.86	0.92
Cornea specialist 1	108*	-	0.97	0.85	0.94	-
Cornea specialist 2	108*	-	0.92	0.85	0.91	-
DMEK orientation model	100	semi-automatic	0.90	0.78	0.86	0.92
DMEK orientation model	100	semi-automatic	0.82	0.69	0.78	0.84
Cornea specialist 1	100	-	0.96	0.78	0.90	-
Cornea specialist 2	100	-	0.96	0.81	0.91	-

compared to the manually labeled results (Figure 4.4A) and in 46 iOCT-frames noticeable segmentation and/or post-processing errors were present after post-processing (Figure 4.4B-D and Figure 4.5E-G). A total of 22 grafts were incorrectly classified using the automatic method of the orientation model, because of segmentation errors in 8 frames, post-processing errors in 2 frames, and a limited discriminative predictive power of the model in 12 frames (i.e., in both the automated and manual method these grafts were incorrectly classified regardless of any errors; Figure 4.5H). In 29 iOCT-frames containing errors the model still correctly predicted the orientation.

The majority of the segmentation errors were considered minor, such as slightly incomplete segmentation of the graft ends or at the image boundary (Figure 4.4B and C). Notwithstanding, despite considered minor these errors may affect the algorithms performance. Partial segmentation of the graft or large gaps between segments were considered large segmentation errors (Figure 4.4D). Causes for large segmentation errors included: corneal mirror artefacts, background noise, hypo reflectance of the graft, and contact of the graft with the cornea or iris. All post-processing errors occurred during filtering of the frames and connecting the line segments resulting in partial or wrong segmentation. During filtering smaller segments (≤100 pixels) were removed, which resulted in gaps too large to bridge in the subsequent step. Similarly, in cases with large gaps (≥100 pixels) the line segments were not connected and the smaller segments were removed after identification of the largest segment (Figure 4.5E). In some frames the line segments were connected with wrong segments or an image artefact falsely identified as graft (e.g., fluid reflection, the lens capsule) resulting in an incorrect representation of the graft Figure 4.5F and G).

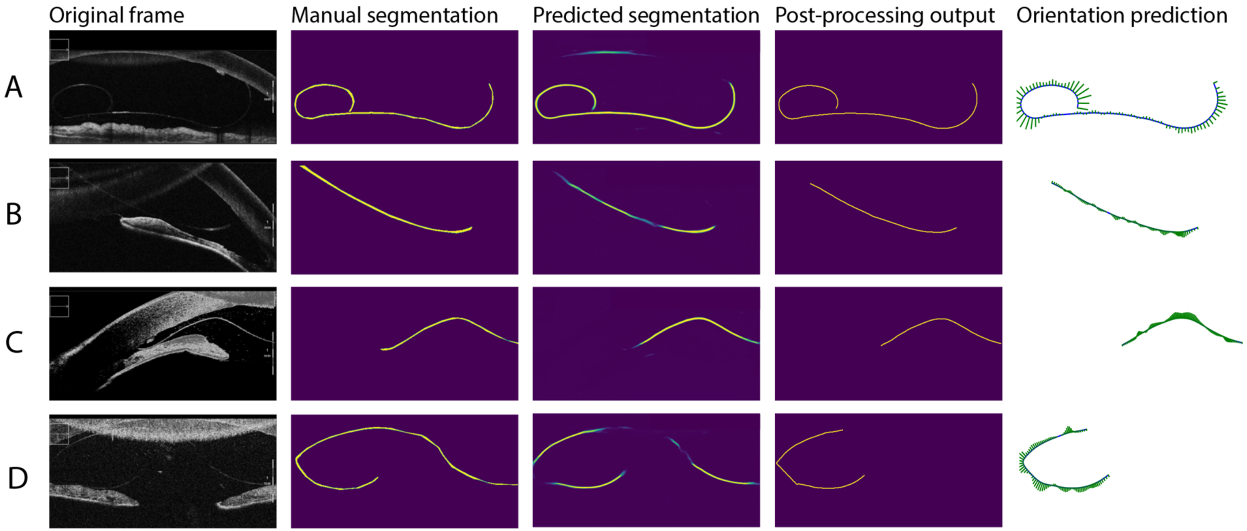

Figure 4.4: Examples of correct and incorrect segmentation and post-processing: (A) a near perfect segmentation, (B) a segmentation error at image boundaries, (C) a segmentation error at the graft end, (D) segmentation gaps resulting in partial segmentation.

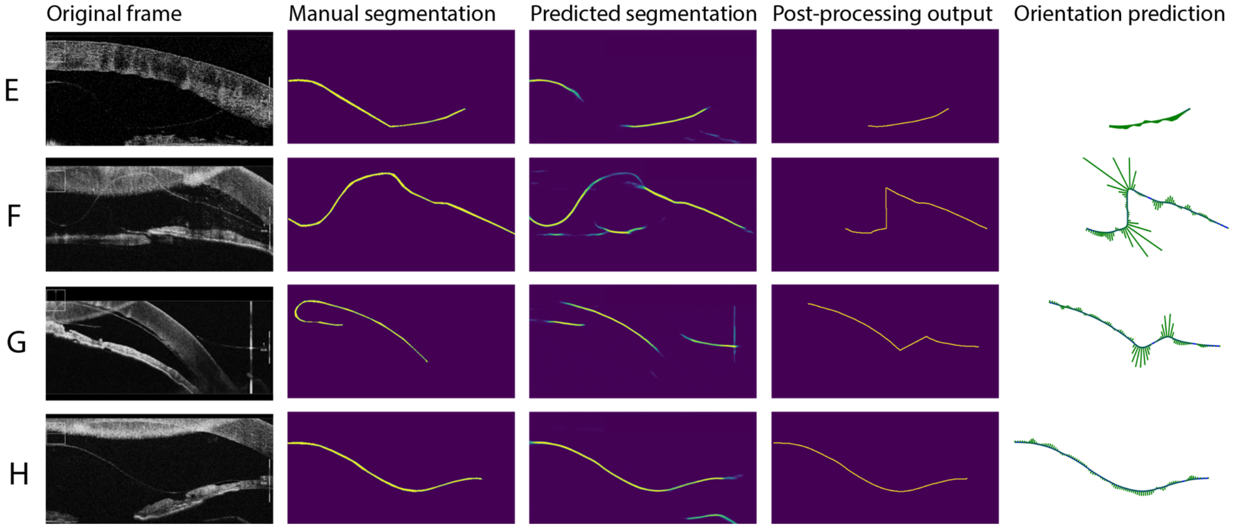

Figure 4.5: More examples. (E) Segmentation gaps too wide to connect in the post-processing, (F and G) segments wrongly connected during post-processing, (H) correct segmentation resulting in an incorrect prediction.

4.4 Discussion and conclusion

In this exploratory study we presented an image analysis tool that can automatically identify the orientation of a DMEK graft using iOCT. We believe the presented tool has the potential to improve and standardize clinical decision making. Moreover, the tool could help ease the learning curve for starting surgeons and aid experienced surgeons in the transition towards DMEK [143].

Determining graft orientation using iOCT is arguably more reliable and safer compared to other methods in use (i.e., the Moutsouris sign and various stamps / cuts) [32, 109, 117, 130, 140, 143]. However, current manual review of both live and static iOCT-scans for DMEK orientation can be time consuming and disrupt the surgical workflow hindering implementation and sustainable use of iOCT [44, 108]. Automatic image analysis may alleviate these hurdles by aiding the surgeon in determining graft orientation and may reduce interpretation errors. Several studies have pointed out the lack of (integrated) image analysis tools and clinical decision support systems (CDSS) for iOCT that can support implementation and improve the clinical value [43–45, 108]. Only a handful studies have reported on (automated) image analysis of iOCT images, although their results show promising potential for improving clinical decision making and clinical outcomes [46, 61, 173].

Computerized CDSS are increasingly developed to assist physicians with decisions in (complex) clinical situations and have the potential to improve outcomes, optimize treatments, and improve workflow efficiency [23, 111, 141]. Development of CDSS for ophthalmology has a strong focus on medical image analysis, because of the visual component during diagnosis and treatment. Automated image analysis tools using fundus photographs and OCT-scans have been developed for detection of diabetic retinopathy, glaucoma, and age-related macular degeneration. Validation studies of these tools report clinical accuracy rates comparable to expert clinicians for these specific tasks [8, 75, 126]. The tools have the ability to improve delivery of care and unlock information for treatment and research purposes. We believe our tool might be of similar value for iOCT and facilitate its use.

In recent years corneal OCT image analysis is emerging. Several studies showed the ability of automatic tools to detect and quantify graft detachment after lamellar corneal transplant surgery [38, 66, 69, 152, 173]. Our automatic method has a good to excellent predictive power [99] and when using manually annotated grafts the performance of our model improves considerably and is comparable to the performance of both corneal specialists. The gap in performance between the automatic and manual method was primarily the result of segmentation and post-processing errors, which in turn resulted in wrong predictions as shown with the qualitative analysis of the end-to-end outcome. Automatic segmentation of iOCT imaging is challenging because of the design and dynamic use of iOCT, which may result in higher signal noise, variable image quality, image decentration, and prevents standardized image acquisition [173]. We consider our dataset a realistic representation of images acquired in practice for determining the orientation and therefor consider the results generalizable to other datasets.

The difference in AUC between the automatic method and the pipeline with manual annotated grafts indicates that improvements for the automatic method can be achieved by improving the automatic segmentation. In particular, correct segmentation of the graft endings could contribute to a better estimate of the graft curvature. The deep learning-based segmentation can potentially be improved by the addition of more training data, including a wider variety of anatomies and image artifacts. If a large enough training set could be obtained, an end-to-end deep learning method could be considered, where a classification model is trained only on orientation labels. However, even if enough training data would be available, such a method would come at the cost of having a CDDS without explanation for the decision-making, which could hamper acceptation by the end users. Alternatively, future research could investigate a segmentation approach that uses shape constraints [20, 169], such as the fact that the graft is a continuous and smooth structure. Such an approach should take into account that not the whole graft necessarily lies in the field of view.

We also experimented with the addition of extra frames to the input taken shortly before or after the investigated iOCT image, in which the location and orientation of the DMEK graft slightly differed from the center frame. For example, we added the 5th and 10th frames before and after the center frame as additional channels to the input, similar to Vu et al. [162], hypothesizing that the extra information would help the learning process. However, no benefits were found from this step and it was omitted for the final ensemble.

It should be noted that assessment of graft orientation based on a single frame does not reflect clinical practice. Instead, a corneal specialist would reduce uncertainty by assessing multiple frames or manipulate the graft until orientation is evident. Future work could incorporate such a strategy in the automatic image analysis pipeline, for example by using a recurrent neural network on follow-up frames [26, 153]. For clinical implementation, the image analysis pipeline needs to be directly applied to the video-feed. In this research, iOCT frames were qualitatively reviewed for image quality and presence of characteristics on which orientation could be determined. However, not every frame contains enough information for evaluation of the orientation. Future research could include an automatic frame-based quality assessment, or an uncertainty estimate and only provide a prediction if the certainty is high. A challenge for real-time image analysis is the speed at which the segmentation and post-processing can be performed. Here an ensemble of 12 U-Nets was used for the segmentation, but this might require more computational power than standardly offered with an iOCT system. Perhaps an ensemble is not required if more annotated training data is used. Another solution could be the use of knowledge distillation techniques, which have recently been proposed to train a single segmentation model that performs similar to an ensemble [96, 148].

The threshold for the results in Table 4.1 was slightly negative for all dataset after optimizing the F1-score (i.e., optimal operating point), which corresponds to a slight curve downwards. This makes sense since the cornea itself also curves downward and the floating DMEK typically partly follows the shape of the cornea. In this study the optimal trade-off between sensitivity and specificity was chosen to optimize the predictive power of the model. However, it can be argued that depending on the use case or user expectations, either sensitivity or specificity may be more important.

In conclusion, we present an automated image analysis method for iOCT to detect a DMEK graft, quantify the curvature, and determine the graft's orientation. Our future research efforts will focus on improving automatic segmentation and predictive certainty of our algorithm.

In the next chapter, we zoom in on the corneal endothelium using specular microscopy. In these microscopy images individual corneal cells can be distinguished and the health of the cornea can be assessed by determining biomarkers such as endothelial cell density.

Chapter 5

Robust corneal endothelial cell density measurements in the presence of guttae

This chapter is based on:
F.G. Heslinga D.F.M. Timmers, G.B.A. Haijen, S.L. Dunker, T.T.J.M. Berendschot, J.P.W. Pluim, M.M. Dickman, M. Veta. Automatic and robust corneal endothelial cell density measurement in the presence of guttae. *In preparation*.

Abstract

Assessment of corneal endothelial cells is essential for diagnosis of corneal diseases and monitoring of disease progression. The aim of this study is robust corneal endothelial biomarker extraction based on automatic segmentation of corneal endothelial cells and guttae in real-world specular microscopy images. We developed a deep learning method for simultaneous mapping of cell objects, distances to cell border, cell edges, and guttae in specular microscopy images. Endothelial cell density (ECD) and coefficient of variance (CV) were calculated and compared with manual annotations. Similarly, the semi-automatic approach of the Topcon microscope was compared with manual annotations. A novel method for flagging of cases with substantial guttae presence was proposed, by considering total guttae area for ECD calculations. Evaluation was done for 300 images from 100 patients, with varying image quality and a broad range of referrals. Mean absolute percentage errors for ECD and CV were 6.1% and 15.7% which is on par and better than the semi-automatic method, respectively. Thirty images were automatically flagged for having an ECD that was affected due to guttae presence. Intra-patient variance of the deep learning method was smaller (3.0%) than that of the semi-automatic method (3.6%). In conclusion, our deep learning method can extract endothelial biomarkers in specular images of varying quality, without requiring any user input. Moreover, the flagging of cases with guttae prevents overestimation of ECD. The automatic extraction of endothelial biomarkers in real-world specular microscopy images supports robust and accurate corneal disease evaluation.

5.1 Introduction

Corneal diseases are a major cause of blindness worldwide [166]. Essential for a well-functioning cornea is the corneal endothelium, a monolayer of cells that lines the posterior corneal surface, controlling corneal hydration and nutrient supply [22]. Assessment of the corneal endothelium is typically done using specular microscopy, a non-contact technique that allows for clear visualization of the morphology of individual cells [102, 145].

The primary metric of assessment is endothelial cell density (ECD), which describes the number of cells per square millimeter. ECD is a key biomarker for diagnosis and evaluation of success and safety of treatment for e.g. Fuchs' endothelial dystrophy [103], corneal transplantation [62], and glaucoma surgery [50]. In addition to cell density, information can be obtained about cell shape (e.g. hexagonality) and cell size distribution (polymegathism). Polymegathism is typically described by the coefficient of variance (CV) which is obtained by dividing the standard deviation of cell size by the mean.

Accurate estimation of endothelial biomarkers requires detailed segmentation of individual cells (cell instances). Manual segmentation of cell instances is a time-consuming process that is unfeasible in clinical practice. Currently, most specular microscopy systems provide a built-in (semi-)automatic method for segmentation of the cells and estimation of the endothelial cell parameters. However, these techniques have been reported to lack accuracy [56], especially for images associated with high polymegathism [74], large cell size [74], and corneal grafts [97, 121].

An additional challenge is the presence of guttae in the corneal endothelium. Guttae are depositions of focal endothelial excrescences and are associated with Fuchs' endothelial dystrophy [4]. Guttae appear as dark hyporeflective round bodies in specular microscopy images of the endothelium [114]. The presence of guttae can be an extra reason for an inaccurate ECD or CV measurements [97]. Moreover, the built-in software typically calculates ECD as the inverse of the mean cell size, neglecting the guttae area that does not contain any cells, leading to an overestimation of the ECD. Alternatives, such as an 'effective ECD' [103] have been proposed but have not been developed as an automatic tool.

In recent years, deep leaning [84] has shown to be a promising tool for specular microscopy image analysis. Most of these deep learning approaches are based on pixel-to-pixel mapping using a U-Net architecture [33, 49, 76]. Typically, a map is predicted that classifies each pixel as either belonging to an endothelial cell or to the background (including cell borders). The downside of this strategy is that a small mistake like misclassifying the border between two cells leads to the incorrect merging of two neighboring cells, which can have a substantial effect on the ECD. Some studies aimed to solve this challenge with post-processing [159, 161]. However, none have yet aimed for segmenting cell instances directly nor have there been attempts to automatically assess the influence of guttae.

In this study, we propose a novel deep learning method, specifically designed for cell instance segmentation in the densely occupied endothelium. We compare the ECD and CV estimates with manual annotations and the semi-automatic method provided with the microscopy system for an extensive set of 300 images gathered in routine clinical care. This set includes many low-quality images and represent a variety of syndromes and ophthalmic procedures, including a substantial number of corneal transplantations. In addition, we present a technique to automatically flag cases with substantial guttae presence and evaluate intra-patient variability for each method to evaluate the robustness of our method.

5.2 Methods

5.2.1 Data

For the development of our methods, we used specular microscopy images (Topcon SP-3000) from the publicly available data set presented by Daniel et al. [33]. Each image represents a magnified 0.5×0.25 mm section of the endothelial surface. From this set, 46 images were selected to include cells with varying sizes, bright lighting, dark shading, and presence of guttae. All visible endothelial cells were individually segmented and saved as a separate binary mask per cell. Guttae were also segmented and saved as a binary mask containing all guttae per image. In addition, 52 images were selected with varying degrees of presence of guttae or shading with pixel intensities similar to that of guttae. For this subset, only the guttae were manually segmented to increase the size of training examples, while limiting the annotation time. The total number of images in the development set was 98, belonging to 98 unique patients. All manual segmentations were made using QuPath software [13].

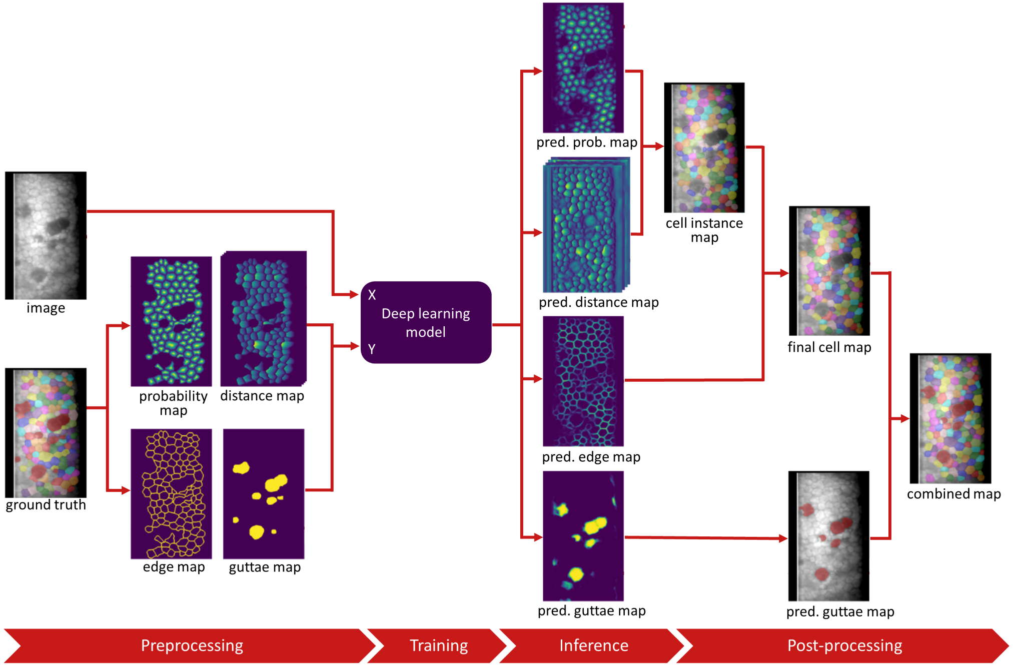

Figure 5.1: Overview of our image analysis method. During preprocessing the ground truth annotations are converted to probability, distance, edge, and guttae maps. The deep learning model is trained to predict all four maps based on the original image. The predicted probability, distance, and edge maps are post-processed to obtain the final cell map, which is in turn is combined with the guttae map. pred. = predicted.

For evaluation, a separate set of 300 specular microscopy images was collected at the University Eye Clinic Maastricht, the Netherlands, between April 2021 and July 2021. For 100 consecutive patients, three images of the same eye were acquired using the Topcon SP-3000P specular microscope. The built-in software of the microscope can semi-automatically determine biomarkers such as ECD and CV, but requires a single mouse click in each cell that should be included in the analysis. In between capturing each image, the patient was asked to lean backwards while the operator selected the cells. This process resulted in minor head position differences and variation in the region of the endothelial layer that was captured. Images were saved with and without analysis grid by the microscope's software. No specific inclusion criteria were used, meaning that patients with a broad variety of pathological backgrounds were included. As a result, the image quality varied widely, and patients were included if more than 10 cells could be identified by the operator.

To attain accurate density measures, the operator aimed to select 40-50 adjacent cells for the semi-automatic biomarker extraction. Similarly, manual segmentation of the cells was done for 40-50 adjacent cells per image, but by a different annotator to prevent selection bias. The actual number of annotated cells could be smaller depending on the image quality.

5.2.2 Deep learning model

Our image analysis method is based on deep learning [84] and includes a novel model architecture. We built upon the well-known U-Net [128], which has become the de facto standard for biomedical image segmentation and StarDist [135] which was developed to segment individual cells. An overview of our method is shown in Figure 5.1.

StarDist was presented by Schmidt and Weigert et al. in 2018 as a method for cell nuclei detection in fluorescence microscopy images [135]. The model was built upon a U-Net architecture and trained to predict two pixel maps: (1) object probabilities, defined as the normalized Euclidean distance to the nearest pixel outside the cell; and (2) star-convex polygon distances, which represent the distances to the border of the cell, calculated for n radial directions with equidistant angles. These distances are only computed for pixels belonging to a cell. Non-maximum suppression was used to obtain the final cell segmentation maps from the object probability and distance maps.

Early experiments with StarDist on our development set of specular microscopy images showed that the model was well-equipped for detecting corneal endothelial cells. However, the cell segmentation maps contained space between cell instances, mostly because the sharp corners of the endothelial cells were missed. We hypothesized this to be the result of the star-convex polygons which are perhaps better suited for round shapes, but we could not resolve this by adding more radial directions or changing the weights of the losses. We therefore chose to add a cell edge detection task to the model and use the cell segmentation maps as input for a marker-controlled watershed operation [105].

Ground truth cell edge maps were obtained from the annotated cell maps by subsequently (1) eroding each cell in the cell instance map, (2) binarizing the cell map and the eroded cell map, (3) subtracting the binarized eroded cell map from the binazired original map, (4) skeletonization of the subtracted map, and (5) dilation of the skeletonized map such that cell borders have a thickness of three pixels.

Since the presence of guttae can have a substantial influence on the effective endothelial cell density [103], we added guttae segmentation as an additional task to the model. The flagging can support clinicians and researchers in assessing the value of the ECD measure.

5.2.3 Model training

For training and optimization of the model, the development set was split into a set for training ($N = 65$) and a set for model selection ($N = 33$). The split was done such that both contained a similar ratio of images with only guttae segmented. Data augmentation was used to increase the size of the training set, by random zooming (range = 0.5 to 1.5), horizontal and vertical flipping, and changing the range of the pixel intensities (multiplication between 0.6 and 2 and addition between -0.2 and 0.2). The model was trained to predict four targets simultaneously (Figure 5.1). Cell probability maps, cell edge maps, and guttae maps were optimized using binary cross-entropy losses. Distance maps were optimized using a Mean Absolute Error (MAE) loss, which was multiplied by the cell probability maps. The aggregated training loss L was defined as

$$L = \delta(L_{CP} + \alpha L_{CE} + \beta L_D) + \gamma L_G$$

where L_{CP} is the cell probability loss, L_{CE} the cell edge loss, L_D the distance loss and L_G the guttae loss. α, β, and γ represent the loss weight. These weights were set to 1, 0.2, and 1, respectively, to have roughly equal contributions to the aggregated loss, which was supported by an optimal performance for the validation set. δ is an additional binary weight factor that was set to 0 for images where only guttae were annotated and set to 1 for images with both cells and guttae annotated.

Training was done with a batch size of four and a learning rate of 0.0003 that was halved after every 4,000 iterations if the validation loss did not decrease. Training was continued for 60,000 iterations and the model of the epoch with the lowest validation loss was selected as the final model. All other training parameters were similar to those described in [135], except for the depth of the U-Net blocks which was set to six.

5.2.4 Post-processing

Predicted guttae maps and cell edge maps were binarized. Non-maximum suppression was used to obtain cell instances from the predicted cell probability map. The centers of the cell instances were used as seeds for marker-controlled watershed – using the cell edge map as boundaries – to obtain a refined cell instance map. After this, pixels part of the edge map between two cells were assigned to one of the neighboring cells to remove empty space between the cells to obtain the final cell map. The final predicted map was obtained by combining the cell map and guttae map (Figure 5.1).

Endothelial cell density (ECD) was calculated for the deep learning model output and the manual annotations as the reciprocal of the average cell size in mm^2. The coefficient of variance (CV) was obtained by dividing the standard deviation of the cell size by the average cell size.

Additionally, we calculated a secondary cell density measure that takes into account the presence of guttae, by adding the guttae area to the total cell area before dividing by the total number of cells n:

$$ECD_{guttae} = \frac{n}{\sum_i cell\_size_i + \sum_i guttae\_size_i}$$

5.2.5 Evaluation and statistical analysis

Performance of the deep learning method and semi-automatic method were measured by comparing the ECD and CV with the manual annotations, based on the Mean Absolute Percentage Error (MAPE). Bland-Altman plots were used to visualize the results. Since the evaluation data set contains three images per patient, we could also determine the intra-patient variability of the ECD and CV as measures of robustness. Intra-patient variability was measured as the relative standard deviation of the ECD and CV for the three images per patient. Relative standard deviation was computed by dividing the standard deviation of cell size across three images by the mean cell size, which was then averaged over all patients. Welch's unequal variances one-sided t-test was used to compare whether the deep learning method performed better than the semi-automatic method. This step was repeated for both the MAPE results and the intra-patient variability results.

Cases with many guttae were automatically flagged if the ECD was found to be more than 10% larger than the ECD_{guttae}. The flagging functions as an indicator that the ECD might be overestimated and should be considered less accurate. Flagged cases were not excluded in the evaluation and statistical analysis.

5.3 Results

The evaluation set consisted of 300 specular microscopy images from 100 patients in which a total of 12,079 cells were manually segmented. The mean age of the patients was 59.8 (±19.4) and 56 of 100 patients were female. Most frequent reasons for image acquisition were related to corneal transplantation (N = 39), phakic intraocular lens (N = 21), glaucoma surgery (N = 11), Fuch's dystrophy (N = 7), and corneal cross linking (N = 6). A total of 12,718 cells were selected as input for the semi-automatic method, while the deep learning model detected 33,665 cells.

In Figure 5.2 the outputs of the deep learning method and the semi-automatic method are shown for three cases. The top row is an example where the endothelial cells are relatively large, the image quality is good and no guttae are present. Cells are easily distinguishable in all parts of the image and the delineations of the cell borders by the semi-automatic and deep learning method are generally correct. In the middle row, the endothelial cells are well visible except for the top left corner. A single small gutta, recognizable as a dark spot in the bottom right corner, is present, but the effect on the ECD is small. The example in the third row includes a large area where cell borders are indistinguishable and contains a substantial area of guttae. The guttae on the right side can easily be spotted but those on the left are more ambiguous. When the guttae are not taken into account, the ECD is 1404 cell/mm^2 according to the ground truth annotations for this case. The semi-automatic method finds 1460 cell/mm^2 (MAPE = 4.0%) and the deep learning method finds 1392 cell/mm^2

(MAPE = 0.9%). These results are quite close, but when the area of the guttae is added to the total area of the cells, the ECD_{guttae} is 1149 cell/mm^2 according to our deep learning method. The difference with the baseline ECD is (1392 - 1149) / 1149 = 21.1%, which is larger than 10%, resulting in a flagged case. Using a threshold of 10% for the effect of guttae presence on the ECD, 30 (10%) specular microscopy images were flagged.

Table 5.1 shows the results for obtaining the ECD and CV as compared with the ground truth annotations. For the ECD, the performance of the semi-automatic method and deep learning method are similar, while the deep learning model does not require any user input. For the CV, the MAPE of the deep learning method is substantially lower (15.7%) than the semi-automatic method (25.7%), indicating that the deep learning method is more accurate for assessing CV.

Table 5.1: Mean absolute percentage error (MAPE) for endothelial cell density (ECD) and Coefficient of Variance (CV) measures.

	semi-automatic method	deep learning method	p-value
ECD MAPE (± std)	6.5% (7.7)	6.1% (9.0)	0.24
CV MAPE (± std)	25.7% (40.5)	15.7% (17.3)	>0.001

Figure 5.3 shows Bland-Altman plots for comparing the semi-automatic method and deep learning method with the ground truth annotations. The flagged images are indicated in orange. Both the semi-automatic method and deep learning method seem to underestimate the ECD on average when the ECD is high (>2000 cell/mm^2). For the CV, the ninety-five percent confidence interval is wider for the semi-automatic method than for the deep learning method. Similarly, the bias is more than twice as large and indicates an overestimation of the CV on average.

The intra-patient variability results are shown in Table 5.2 and represent the relative standard deviations between three images of each patient. For both the ECD and CV the deep learning method results in a lower variability than the semi-automatic method. Noteworthy, the deep learning method even has a lower variability than the manual annotations.

Table 5.2: Intra-patient variability results. Relative std = relative standard deviation. *p-values were obtained by comparing the semi-automatic method with the deep learning method.

	manual	semi-automatic	deep learning	p-value
ECD, relative std (± std)	3.6% (3.0)	3.8% (2.7)	3.0% (3.3)	0.028
CV, relative std (± std)	9.8% (7.4)	12.1% (10.3)	8.7% (7.7)	0.004

5.4 Discussion

The use of specular microscopy for the assessment of corneal endothelium dates back to 1968 [101]. Techniques for (semi-)automatic calculation of endothelial cell biomarkers have since been developed and are still an active topic of research. In this study we developed a deep learning-based method for fully automatic segmentation of corneal endothelial cells to obtain accurate ECD and CV measurements in real-world images. Our deep learning model can correctly segment endothelial cells of different sizes, even when contrast is low and cell borders are somewhat unclear. Moreover, guttae are automatically identified, providing a tool for automatic flagging of cases where the guttae substantially affect the ECD measure.

Cell segmentation can be relatively easy in images with sufficient contrast but challenging for cases with low contrast, bright lighting, or dark shading. Even when only good quality images are included, the presence of guttae can hamper accurate ECD measurements [74]. In 2014, McLaren et al. [103] proposed the concept of effective ECD, where the number of cells within a fixed frame is divided by the size of that frame. The advantage of this approach is that it directly measures the number of cells per unit of area. In contrast, by defining the ECD as the inverse of the average cell size, any non-cell area (e.g. guttae) is not taken into account. Although the effective ECD is arguably the most objective measure, we found it is not practical in cases where image quality is limited. In such cases, not all cell borders within a fixed frame can be distinguished, and sometimes guttae can be hard to differentiate from dark shadings in low-contrast regions. Our approach for flagging is based on finding all cells and guttae that are visible within the full image. This strategy comes at the risk of segmenting cells more easily than guttae or vice versa, introducing a potential bias to our ECD_{guttae} measure. Nevertheless, this approach provides a tool for flagging, which is arguably the best alternative when the effective ECD cannot be determined accurately. The flagging can support clinicians and researchers in assessing the value of the ECD measure.

Our deep learning model provides a MAPE for ECD which is similar to that obtained with the semi-automatic method of the Topcon system, while not requiring any user input. The Bland-Altman plots indicate that the ECD is sometimes underestimated by both methods when the ECD is large. For the CV calculation, our deep learning outperforms the semi-automatic method with a MAPE that is 39% smaller. This difference can also be visually appreciated in the Bland-Altman plot by the wider spread of the point cloud for the semi-automatic method. Polymegathism is less frequently used in the clinic than ECD even though it is a key indicator of the corneal wound repair mechanism [102]. This might be because the MAPE for CV is reported to be large in comparison to the MAPE for ECD [48, 49, 160], which is also supported by our results. The reduction of the MAPE for CV with a deep learning method could improve the clinical usefulness of the polymegathism metric.

Analysis of intra-patient variability showed that a lower relative standard deviation is achieved for both ECD and CV with the deep learning approach as compared to the semi-automatic method. The lower variability implies that the deep learning method is more robust, which is useful when multiple images are collected over time to keep track of disease progression. Interestingly, the deep learning method leads to a smaller variability than the ground truth annotations. A possible explanation is that the total number of detected cells by the deep learning model is almost three times as large the number of annotated cells, reducing the effect of variance. However, it could also be that the variations in manual annotations (sometimes over-segmenting, sometimes under-segmenting) in the training set have been averaged out during training of deep learning model, resulting in a model that is more reproducible than the human annotator.

Before the introduction of deep learning, already a variety of techniques had been suggested for corneal endothelial cell segmentation. These methods were developed for in vivo specular microscopy or confocal microscopy image analysis and include classical morphological methods [12], marker-controlled watershed [57], Bayesian shape models [54], Fourier analysis [137], and active contour models [139]. Most methods can be described as a pipeline of multiple image processing steps. Others have focused on additional post-processing to improve existing segmentation methods [159].

Deep learning has proven to be a valuable tool for ophthalmic image analysis. Examples include grading of diabetic retinopathy [59], retinopathy of prematurity classification in fundus images of premature babies [171], corneal pachymetry in anterior segment optical coherence tomography (AS-OCT) scans

[70], donor graft detachment quantification in AS-OCT scans [69, 154], and diagnosis of diabetic neuropathy using corneal confocal microscopy [168]. In recent years, deep learning has become the technique of choice for corneal endothelial cell segmentation and biomarker calculations [33, 48, 49, 76, 77, 112, 160, 161]. Our approach is novel in a sense that we use cell instance segmentation, instead of binary background foreground segmentation.

It is difficult to compare results between studies about automatic corneal endothelium assessment, since evaluation is mostly done for different data sets of varying image quality. Our evaluation set represents real-world data with images typically acquired in the clinic. Most studies only included high-quality images, except for Daniel et al. [33] who also focused on real-world images. However, in the study by Daniel et al. manual segmentations were not available, and evaluation was done for cell detection, not for calculating biomarkers such as ECD and CV.

In conclusion, we have developed a novel deep learning-based method for endothelial cell segmentation and subsequent biomarker calculation in specular microscopy images. Evaluation on an extensive set of real-world images showed that the results are on par with the semi-automatic method for measuring ECD and better for measuring CV, while not requiring any user input. In addition, the lower intra-patient variances for both biomarkers indicated that the deep learning method is more robust than the semi-automatic method. Lastly, the addition of a novel method for automatic flagging of cases with substantial guttae presence provides support to clinicians and researchers in assessing the value of the ECD measure.

After this chapter, we switch our focus from the cornea to the retina. The next two chapters are about (early) detection of type 2 diabetes using fundus photography images.

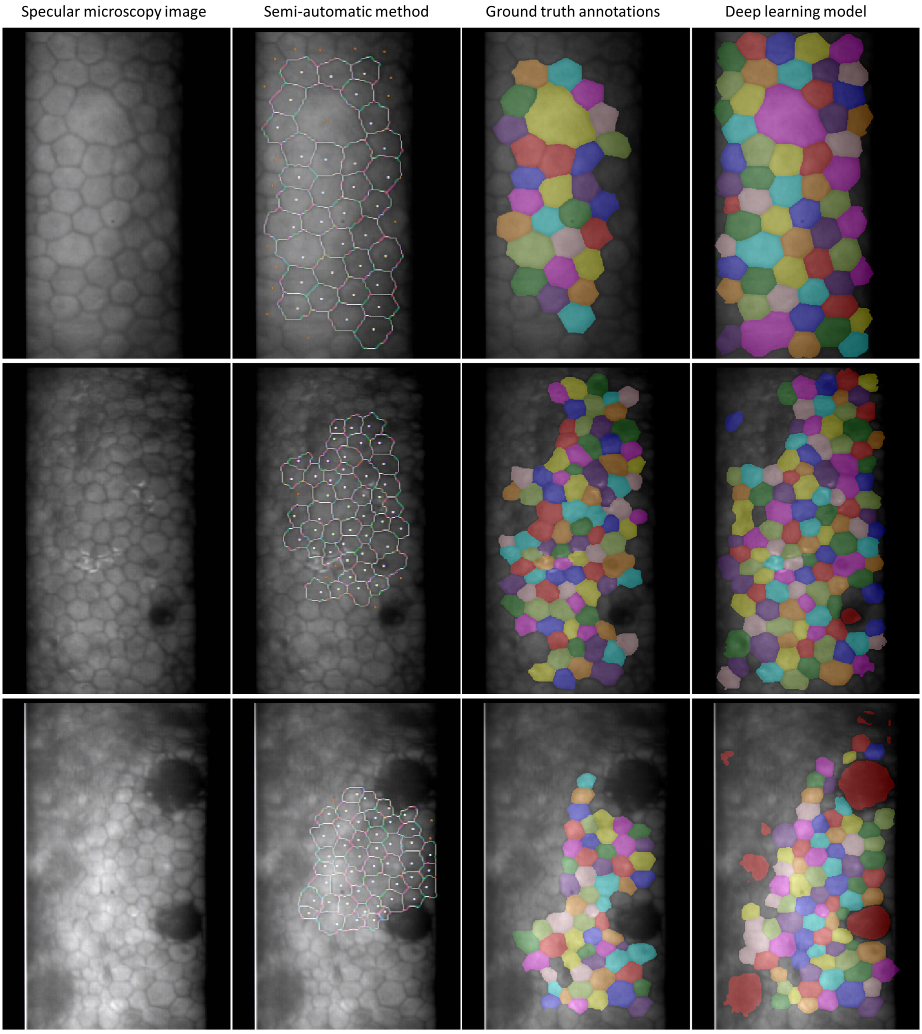

Figure 5.2: Three examples of the cell delineations by the semi-automatic method (second column) and the deep learning method (fourth column) and the comparison with the ground truth annotations (third column). The guttae are marked as dark red in the fourth column.

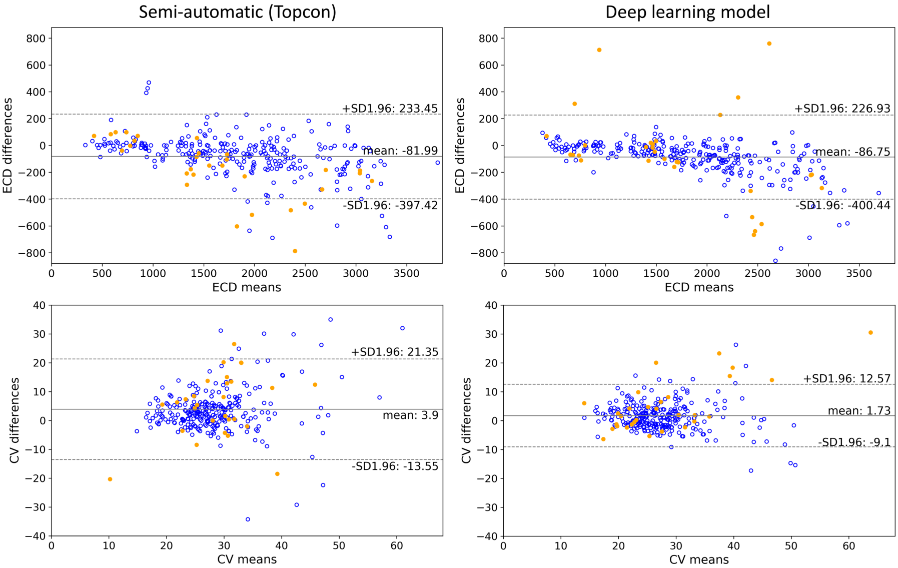

Figure 5.3: Bland-Altman plots. Top row: endothelial cell density (ECD) measurements. Bottom row: coefficient of variance (CV). The filled orange circles represent images that are flagged for an ECD affected by more than 10% by the presence of guttae. Blue open circles represent unflagged images.

Chapter 6

Methodological considerations for T2D classification from fundus images

This chapter is based on:
F.G. Heslinga, J.P.W. Pluim, A.J.H.M. Houben, M.T. Schram, R.M.A. Henry, C.D.A. Stehouwer, M.J. van Greevenbroek, T.T.J.M. Berendschot, M. Veta. Direct Classification of Type 2 Diabetes From Retinal Fundus Images in a Population-based Sample From The Maastricht Study. *SPIE Medical Imaging: Computer-Aided Diagnosis 11314*, 2020.

Abstract

Type 2 diabetes (T2D) is a chronic metabolic disorder that can lead to blindness and cardiovascular disease. Information about early stage T2D might be present in retinal fundus images, but to what extent these images can be used for a screening setting is still unknown. In this chapter, deep neural networks were employed to differentiate between fundus images from individuals with and without T2D. We investigated three methods to achieve high classification performance, measured by the area under the receiver operating curve (ROC-AUC). A multi-target learning approach to simultaneously output retinal biomarkers as well as T2D works best (AUC = 0.746 [±0.001]). Furthermore, the classification performance can be improved when images with high prediction uncertainty are referred to a specialist. We also show that the combination of images of the left and right eye per individual can further improve the classification performance (AUC = 0.758 [±0.003]), using a simple averaging approach. The results are promising, suggesting the feasibility of screening for T2D from retinal fundus images.

6.1 Introduction

Type 2 diabetes mellitus (T2D) is a chronic metabolic disorder characterized by hyperglycemia, insulin resistance, and relative insulin deficiency. Late detection of T2D can lead to long-term damage, including blindness [30] and cardiovascular disease [58]. Although T2D diagnosis based on blood-glucose measurements works well, half of the people living with diabetes worldwide were undiagnosed in 2017 [28]. This is unfortunate, especially since major health benefits are expected from early detection and treatment [68]. Non-invasive, easy-accessible screening methods could improve early detection.

Retinal fundus imaging is widely used for the detection of diabetic retinopathy (DR), one of the complications of T2D. Over the last few years, deep learning has been proposed for automated analysis of retinal fundus images [150]. DR is relatively unambiguous and deep learning models have shown excellent detection performance. For example, Gulshan et al. [59] obtained an area under the receiver operating curve (ROC-AUC) of 0.99 for detection of referable DR.

Despite the promising results for DR detection, retinal fundus images are not used for early T2D detection, even though the vascular geometrical structures of the retina have been related to early T2D [175]. In this chapter we investigate to what extent a deep learning model is able to distinguish T2D and non-T2D cases in retinal fundus images and we evaluate what techniques can be used to improve the classification.

6.1.1 Related work

To the best of our knowledge, only one study has explored the value of deep learning for direct classification of T2D [2]. In addition, Poplin et al. [120] used deep learning to extract cardiovascular risk factors from retinal fundus images, including a key diagnostic measure for T2D, haemoglobin A1c (HbAIc). While high predictive performance was reported for age and sex, and some predictive information was found for smoking history and systolic blood pressure, model predictions for HbAIc levels correlated poorly with the HbAIc labels ($R^2 = 0.09$).

Others have focused on the extraction of handcrafted features from fundus images [35, 149]. Features such as vessel tortuosity, mean arteriolar width and venular width are considered biomarkers for T2D [175]. In another study, not part of this thesis, we showed that these biomarkers can be approximated with a deep learning approach [71]. In this study we investigated the added value of these biomarkers for the training process of a deep learning model that directly classifies fundus images.

Table 6.1: Training iterations, batch size and learning rate

	Training	Validation	Test	Total
Total number of individuals	1,376	464	496	2,336
age (years) [±std]	60.0 [±8.5]	59.6 [±8.1]	60.4 [±8.2]	59.9 [±8.2]
sex (% men)	47.2	50.2	54.1	51.2
T2D individuals [%]	466 [33.9%]	159 [34.3%]	182 [36.7%]	807 [34.5%]
Number of images	5,222	1,802	1,900	8,924

6.2 Methods

The color fundus images used for this research originate from The Maastricht Study, an observational prospective population-based cohort study. The rationale and methodology have been described elsewhere [136]. Eligible for participation in The Maastricht Study were all individuals aged between 40 and 75 years and living in the southern part of the Netherlands. The study population was enriched with T2D participants for reasons of statistical power. For our study only images from individuals with T2D and normal glucose metabolism were included (8,924 images from 2,336 individuals in total). Other diabetes types and prediabetes individuals were excluded.

The data was divided into sets for training, validation and testing according to a 60%/20%/20% split. All images of a single individual were assigned to the same set. An overview of the sets is shown in Tabel 6.1. The sets comprise images of left and right eyes that are centered either on the fovea or on the optic disc. The images were resized to 1024 × 1024 pixels and channel-wise global contrast normalization was applied before further processing [53].

All experiments were performed using deep learning models based on a VGG-19 [142] architecture for which the output layer was replaced. Data-augmentation was used to expand the number of training images, encompassing translation (0 - 20 pixels), rotation (0 - 360 degrees), horizontal and vertical reflection, intensity shift(0 - 20/256), color shift (0 - 30/256) and contrast shift (0 - 0.1). Inputs for the deep learning models are 800 x 800 pixels centered crops of the 1024 x 1024 augmented images. Models were implemented in Keras [29] using a TensorFlow [1] backend and training was done with balanced batches of 18 on 3 GPU's. Optimization of the model weights was done using Adam. Target labels are either 0 = normal glucose metabolism or 1 = T2D. The best performing model was selected based on the validation set. Final performance of the models was evaluated by the ROC-AUC on the test set.

6.2.1 Model setup and initialization

First, we evaluated the effect of initialization of the model's weights for the classification of T2D images versus non-T2D images. We compared five different strategies: (1) random initialization; (2) ImageNet weights; (3) model pretrained on global retinal microvascular measurements (T2D biomarkers), including vessel caliber and vessel tortuosity [71]; (4) A multi-target learning (MTL) approach with random initialization and (5) Multi-target learning with ImageNet weights. For the T2D biomarker approach (3) we first trained a model to predict four microvascular measures as described elsewhere [71] and then replaced the output layer for the classification task. For the multi-target approaches (4 and 5), we simultaneously predicted four T2D biomarkers and T2D status. The learning rate schedule and L_2 regularization were optimized on the validation set after which all experiments were repeated three times using different random seeds to obtain a measure for standard deviation.

6.2.2 Aleatoric uncertainty estimation

In a clinical setting one can decide to refer an image for further inspection if the assessor is too uncertain about the decision. Ahyan et al. [5] showed that test-time augmentation (TTA) can be used to define a measure for the aleatoric uncertainty. We applied 30-fold TTA to the model that performed best on the validation set using the same augmentation settings as applied during training to find the posterior distribution of the T2D predictions. We used variance of the prediction distribution, *var(Pred)*, as a measure for aleatoric uncertainty. Additionally, proximity of the mean of the prediction distribution to 0.5, *abs(mean(Pred)-0.5)*, was evaluated as a measure of uncertainty, since this is exactly half-way the labels of healthy and T2D. We show the effect of the referral of images that the model is uncertain about, by excluding these from the results and recalculating the ROC-AUC for different referral fractions.

Table 6.2: Model setup and initialization results.

Initialization	ROC-AUC [±std]	30-fold ROC-AUC [±std]
random initialization	0.726 [±0.006]	0.729 [±0.009]
ImageNet weights	0.733 [±0.003]	0.737 [±0.008]
T2D biomarker weights	0.734 [±0.004]	0.738 [±0.006]
MTL w. random initialization	0.733 [±0.010]	**0.746 [±0.001]**
MTL w. ImageNet weights	0.739 [±0.002]	0.741 [±0.001]

6.2.3 Individual-level estimation

Multiple fundus images (1 to 12) were available per individual, providing a similar number of T2D predictions. Different strategies for the aggregation of image-level predictions to individual-level predictions were evaluated for the model that performed best on the validation set: (1) mean of the soft predictions for the left and right eye; (2) maximum of the predictions for the left and right eye; (3) logistic regression and (4) Gaussian Naive Bayes. For the machine learning techniques (3 and 4) the following features were selected: Mean, variance and number of images for each of the combinations left/right eye and optic disc and fovea centered images, resulting in 12 features per individual. Average padding was used for missing values: if for one eye no optic disc centered image or fovea-centered images was available, the prediction for the opposite-centered image was used. If no image was available for one eye, the prediction for the other eye was used.

6.3 Results

An overview of the results for different model setups and weight initialization is shown in Tabel 6.2. If a single (non-augmented) image was used for evaluation, the ROC-AUC was found to be in the range of 0.726-0.739. When 30-fold TTA was applied, the ROC-AUC slightly increased for all strategies, with the best performance found for the MTL approach with randomly initialized weights (AUC = 0.746 [±0.001]).

The model that performed best on the validation set was one of the MTL models with ImageNet weights. Its performance on the test set was found to be

Table 6.3: Individual-level evaluation.

	image-level	mean of left and right eye	max of left and right eye	logistic regression	Gaussian Naive Bayes
ROC-AUC [±std]	0.733 [±0.010]	0.758 [±0.003]	0.755 [±0.005]	0.761 [±0.004]	0.757 [±0.002]

0.740 with 30-fold TTA. When a fraction of the images was left out for referral, based on high uncertainty of the prediction for those images, the ROC-AUC substantially increased (Figure 6.1). For example, when 20% of the images was excluded, the ROC-AUC increased to 0.765. Interestingly, the effect on the ROC-AUC seemed similar for both uncertainty measures.

The combination of multiple images to obtain an individual-level prediction resulted in a higher ROC-AUC (e.g. 0.758 [±0.003] for mean of both eyes) than for single images (0.733 [±0.010]). The use of more complex classifiers did not lead to significantly better classification performance than a simple mean over the images of the left and right eye, as is shown in Table 6.3

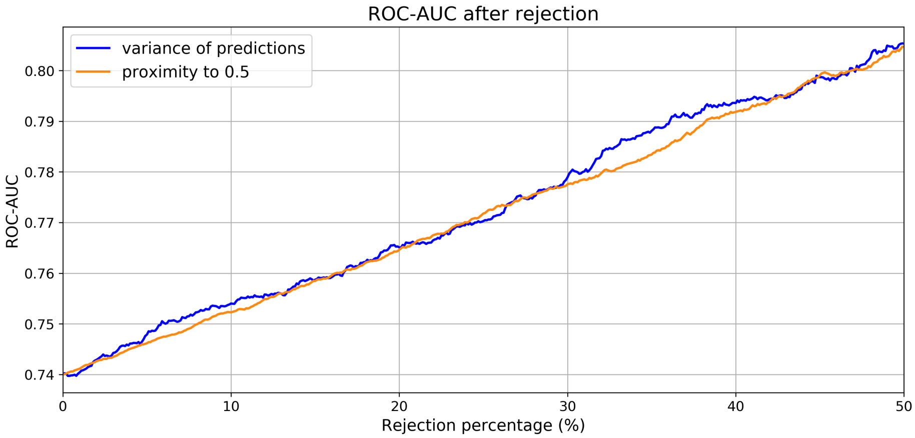

Figure 6.1: ROC-AUC after rejection of images with high prediction uncertainty.

6.4 Conclusion & discussion

Individuals with type 2 diabetes can be distinguished quite well from individuals with normal glucose metabolism in The Maastricht Study population using retinal fundus images and deep learning techniques. Minor benefits can be expected from optimization of the model setup and weight initialization. We found that an MTL approach with randomly initialized weights works marginally better than the other models. Classification performance improvement can be achieved with referral of the most uncertain cases and the use of multiple images per individual. This result is in line with the finding of Leibig et al. [88] who leveraged prediction uncertainty to successfully refer fundus images with signs of diabetic retinopathy that were difficult to grade. This step will however lead to the referral of more false positives, which could hemper the cost-effectiveness in a screening setting.

One possibility to use retinal fundus imaging as a screening technique is the use of smartphone fundus photography [60]. More research is needed to evaluate the value of the addition of basic risk factors, such as sex, age, and body mass index. Also, the inclusion criteria should be extended to comprise early T2D cases (prediabetes), which were excluded for this chapter. These topics are addressed on the next chapter, were we perform a thorough clinical evaluation of the value of fundus photographs for early detection of T2D.

Chapter 7

(Pre)diabetes detection using fundus images

This chapter is based on:
F.G. Heslinga, T.T.J.M. Berendschot, M.T. Schram, C.D.A. Stehouwer, M.J. van Greevenbroek, J.P.W. Pluim, A.J.H.M Houben, M. Veta. The use of fundus images to discriminate prediabetes and type 2 diabetes. The Maastricht Study. *In preparation*.

Abstract

We studied to what extent fundus images can be used to discriminate between individuals with normal glucose metabolism, prediabetes, and type 2 diabetes. In addition, the discriminative value of fundus images was compared with typical risk factors for type 2 diabetes. To achieve this, a deep learning model was developed to determine a glucose metabolism score (GMscore) from fundus images using data from The Maastricht Study. The discriminative power of this GMscore for classifying individuals with type 2 diabetes versus normal glucose metabolism was assessed for a hold-out dataset by the area under receiver operating characteristic (AUROC) curve. For comparison, AUROCs for risk factors such as age, waist circumference, and family history were also calculated and combined with the GMscore to evaluate the additional value of the fundus images. The GMscore based on the fundus images obtained an AUROC of 0.757 (95% CI 0.731 – 0.783), which is higher than 5 out of 6 risk factors for diabetes. Only waist circumference resulted in a higher AUROC. When information from the fundus images was combined with other risk factors, the AUROC increased to 0.895 (95% 0.878 - 0.912). Moreover, prediabetes individuals were found to have a distinct GMscore distribution, approximately half-way between normal and type 2 diabetes individuals. In conclusion, fundus images are informative for discriminating individuals with prediabetes and type 2 diabetes from those with normal glucose metabolism. When combined with other typical risk factors, fundus imaging can contribute modestly to screening for early-stage type 2 diabetes.

7.1 Introduction

Still half of all people living with diabetes worldwide are undiagnosed [131] while it is well known that late detection can lead to long-term complications, including blindness [30] and cardiovascular disease [58]. Research efforts into non-invasive screening techniques are ongoing, using risk factors such as waist circumference [24], family history [155] and a combination of factors [27, 87].

Type 2 diabetes affects the cardiovascular system and early changes in the retinal vascular tree are associated with type 2 diabetes, such as vessel caliber [89, 110, 110, 123] and vascular tortuosity [133, 134]. Fundus photography allows for non-invasive visualization of the retina and automated retinal image analysis has become increasingly popular [71, 149]. Specifically, the use of deep learning [84] has been promising, and several studies have shown excellent results for detection of diabetic retinopathy [3, 59, 151].

So far, limited research has been done on the value of deep learning on fundus images for early type 2 diabetes detection. This could be due to the challenging nature, as early signs of type 2 diabetes are much more subtle than retinopathy, but also due to the lack of a large good-quality data set.

Since 2010, a large set of fundus images from individuals with (pre)diabetes has been collected as part of The Maastricht Study [136]. Here, we aim to utilize these fundus images to develop an automated method to discriminate between individuals with normal glucose metabolism and type 2 diabetes. We also compare the discriminative power of the fundus images with that of some typical risk factors such as age, sex, and family history. In addition, we investigate to what extent fundus images can be used to discriminate between individuals with normal glucose metabolism and individuals with prediabetes.

7.2 Research design and methods

7.2.1 Data

We used data from The Maastricht Study, an observational prospective population-based cohort study. The rationale and methodology have been described previously [136]. In brief, the study focuses on the etiology, pathophysiology, complications, and comorbidities of type 2 diabetes mellitus and is characterized by an extensive phenotyping approach. Eligible for participation were all individuals aged between 40 and 75 years and living in the southern part of the Netherlands. Participants were recruited through mass media campaigns and from the municipal registries and the regional Diabetes Patient Registry via mailings. Recruitment was stratified according to known type 2 diabetes status, with an oversampling of individuals with type 2 diabetes, for reasons of efficiency. The present report includes cross-sectional data from the first 7,689 participants, from whom fundus photography was available for 6,539 participants, and who completed the baseline survey between November 2010 and December 2017. The examinations of each participant were performed within a time window of three months. The study has been approved by the institutional medical ethical committee (NL31329.068.10) and the Ministry of Health, Welfare and Sports of the Netherlands (Permit 131088-105234-PG). All participants gave written informed consent.

A total of 58,722 color fundus images (FF450; Carl Zeiss AG, Jena, Germany) from 6,539 individuals were initially included. The data set comprised images from both eyes, fixated on the optic disc, macula, or periphery. Fundus image quality was assessed automatically, and low-contrast images (N = 12,000) were excluded (Figure 7.1). In addition, we excluded individuals with other types of diabetes than type 2 (N = 41), resulting in a final set of 46,371 images from 6,453 individuals.

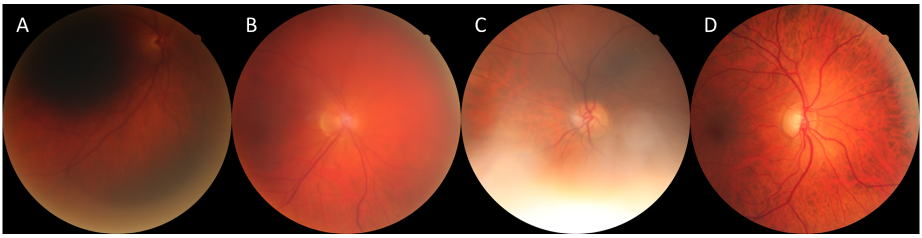

Figure 7.1: Exclusion of low-contrast fundus images. Image contrast was calculated by subtracting a blurred version from the original image and summing over all non-background pixels of the difference map. Blurring was done using a 2-D Gaussian filter (sigma = 3×3). A threshold was manually selected to exclude 12,000 (20.4%) images. (A) Low-contract image. (B) Example just below the threshold for inclusion. (C) Example just above the threshold. (D) High-quality image.

(Pre)diabetes classification

Glucose metabolism status was based on the World Health Organization definitions for fasting glucose, 2-hour oral glucose tolerance test [115], and use of glucose-lowering medication. Consequently, we distinguished individuals with normal glucose metabolism, type 2 diabetes, and prediabetes. Prediabetes was defined as impaired fasting glucose and/or impaired glucose tolerance. Individual-level glucose metabolism statuses were extended to the fundus images, meaning that every fundus image from an individual with e.g. type 2 diabetes was labeled as type 2 diabetes, independent of the actual information in the image.

7.2.2 Image preprocessing and experimental setup

All fundus images were cropped, resized, and preprocessed [53] to increase contrast for more efficient training of the image analysis algorithms (Figure 7.2). We split the data randomly on an individual level into a set for model development (N = 28,153 images) and a hold-out set for final validation (N = 14,476 images). Additionally, a separate set (N = 3,742 images) was created for a matching cohort experiment. This set included 275 individuals that were newly identified as having type 2 diabetes during The Maastricht Study, matched with non-diabetic individuals based on typical type 2 diabetes risk factors.

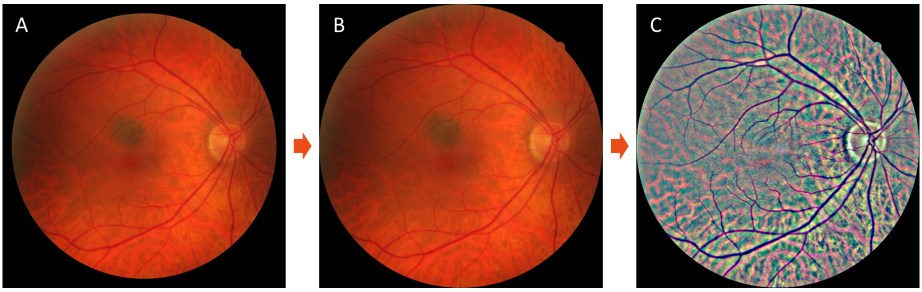

Figure 7.2: Fundus image preprocessing. (A) The original image (3744 × 3744 pixels) is cropped and resized (B) to 512 × 512 pixels. (C) Image luminosity and contrast is normalized channel-wise, similar to Foracchia et al. [53]

7.2.3 Deep learning model

The development set was split on an individual-level into a set for model training and a set for model selection. Data augmentation was used to increase the variety in the training set: translation (≤ 20 pixels), rotation (< 360 degrees), scaling (0.95-1.05), horizontal and vertical reflection, intensity shift (≤ 10/255), color (≤ 10/255) and contrast shift (0.95-1.05) and addition of white noise (≤ 3/255).

Our image analysis algorithm is based on deep learning [84]. A convolutional neural network with an EfficientNet B4 [147] architecture was trained to classify the glucose metabolism status of individual fundus images. During training, only normal glucose metabolism cases and known type 2 diabetes cases were used. Prediabetes cases were not used at this point. The output of the model is a glucose metabolism score (GMscore) in the range 0 to 1 where an output value closer to 0 represent normal glucose metabolism, and close to 1 represents type 2 diabetes.

EfficientNet was implemented in Keras with a Tensorflow backend to process images of size 512×512 pixels on a single GPU (Nvidia Titan Xp). The final model layer was replaced by a single node with a sigmoid activation to output a probability value that is interpreted as glucose metabolism score (GMscore). The model was initialized with random weights, which were iteratively updated by minimizing the binary cross-entropy loss between the labels and the model prediction using the Adam optimizer. The model was trained for 245,760 iterations using batches of eight images and an initial learning rate of 0.001. The learning rate was multiplied with a factor of 0.3 every 61,440 iterations. For evaluation, fundus images of the validation set were processed resulting in a glucose metabolism score between 0 and 1.

For the heat maps, grad-CAM was used. Since the horizontal and vertical dimension of the final convolutional layer were only 16×16, the main model provided a heat map with limited detail. Therefore, a separate model was trained for the grad-CAM, using the same EfficientNet B4 architecture but with the removal of the final downscaling operation (originally implemented as a stride = 2 in a convolutional layer). To train this model, the batch size was decreased to six to fit in GPU memory. The resulting heat maps were sized 31×31 which were upscaled and overlayed on a gray-level image of the original fundus image. Heat maps were normalized per image, such that regions with high activations are always shown in red and low activations are shown in blue. A mask was used to remove any heat map signal in the background.

7.2.4 Evaluation

For evaluation, the fundus images of the validation set were processed by the trained network to obtain glucose metabolism scores. For an individual-level model GMscore, image-level scores were combined by averaging over all fundus images of the left and right eye. Our primary evaluation was the discriminative power of the GMscore prediction for classifying individuals with normal glucose metabolism versus type 2 diabetes individuals, measured by the area under receiver operating characteristic (AUROC) curve. For comparison, risk factors for type 2 diabetes (sex, age, waist circumference, smoking, hypertension, and family history) were used to train logistic regression classifiers to discriminate normal glucose metabolism from type 2 diabetes individuals. Smoking was divided into 3 categories (non-smoker, former, and current), based on self-reported data. Hypertension was defined as an office blood pressure greater 140/90 mm Hg or use of blood pressure lowering medication. Family history represents self-reported data about a first or second degree relative with diabetes. Ninety-five percent confidence intervals (CI) for the AUROC were obtained using 1,000 bootstrap samples. Prediabetes individuals in the validation set were also processed and glucose metabolism scores for this group were compared with normal glucose metabolism individuals. All p-values reported are based on the Welch's unequal variances two-sided t-test and describe the probability that compared groups are similar.

7.2.5 Matching cohort experiment

Risk factors for diabetes such as age and sex can, to some extent, be extracted from fundus images using deep learning [120]. Since these risk factors are easily obtained in a screening setting and can be strong confounders for prediction of type 2 diabetes from fundus images, a matching cohort experiment was designed. For this cohort, all individuals were selected that were newly identified as having type 2 diabetes in The Maastricht Study ($N = 275$), meaning that these individuals were previously unaware of having type 2 diabetes. Each newly identified case was matched with an individual with normal glucose metabolism, identical sex, and similar age and waist circumference.

Table 7.1: Study population demographics and fundus image details. Normal = Normal glucose metabolism. * Missing data for 7 individuals. † Missing data for 45 individuals.

	Development	Validation	Matching cohort
Individuals	9,903	2,000	550
Normal (%)	2,460 (63.0)	1,249 (62.5)	275 (50.0)
Prediabetes (%)	632 (16.2)	335 (16.8)	–
Type 2 diabetes (%)	811(20.8	416 (20.8)	275 (50.0)
Newly identified type 2 (%) diabetes individuals	–	–	275 (50.0)
Females (%)	1977 (50.7)	1044 (52.2)	216 (39.3)
Age [years] (std)	59.4 (8.8)	59.3 (8.6)	63.4 (7.9)
Waist [cm] (std)	94.5 (13.5)	94.0 (13.3)	102.7 (12.4)
Hypertensie* (%)	2,031 (52.1)	1,039 (52.1)	368 (66.9)
Smoking†, Current(%)	497 (12.8)	249 (12.5)	60 (11.0)
Former (%)	1,912 (49.4)	967 (48.6)	284 (52.0)
Non-smoking (%)	1.464 (37.8)	773 (38.9)	202 (37.0)
Fundus images	28,153	14,476	3,742
Left eye (%)	14,314 (50.8)	7,339 (50.7)	1,872 (50.0)
Optic disc centered (%)	8,763 (31.1)	4.409 (30.5)	1,242 (33.2)
Macula centered (%)	9,468 (31.1)	4,862 (33.6)	1,316 (35.2)
Periphery centered (%)	5,242 (18.6)	2,620 (18.1)	663 (17.7)
Other (%)	4,680 (16.6)	2,585 (17.9)	521 (13.9)

7.3 Results

The study population demographics and fundus image details are displayed in Table 7.1. Selection for the match-based cohort set was done before randomly assigning the remainder of individuals to the development set or validation set. Of the 6,453 individuals included in this study 1,502 (23.3%) had type 2 diabetes, while 967 (15.0%) had prediabetes. Most of the fundus images were centered on the optic disc or macula (64.8%). The remainder of the images was either fixated on the temporal periphery or 'other' (e.g. superior of the optic disc).

The results for discriminating individuals with known type 2 diabetes from individuals with normal glucose metabolism are presented in Table 7.2. The GMscore obtained with the deep learning model achieves an AUROC of 0.757 (95% CI 0.731 – 0.783) (see also Figure 7.3) which is higher than the individual risk factors age, sex, smoking, hypertension, or family history with AUROCs in the range 0.607 – 0.727. In contrast, waist circumference has a stronger predictive value as compared to the GMscore (AUROC of 0.832 (95% CI 0.810 - 0.853)). The GMscore also provides additional predictive value when combined with risk factors. For example, the AUROC for age, sex and waist circumference increases from 0.853 (95% CI 0.832 - 0.873) to 0.867 (95% CI 0.846 - 0.888) when combined with GMscore. Even when all six individual factors are combined, the addition of the GMscore still provides extra predictive power with the AUROC increasing from 0.888 (95% CI 0.870 - 0.906) to 0.895 (95% CI 0.878 - 0.912) (p-value < 0.001).

Prediabetes individuals were excluded for the calculations of AUROCs in Table 7.2. This means that AUROCs obtained here are somewhat higher than when the prediabetes group would have been included. For example, the AUROC of 0.757 obtained with the GMscore decreases to 0.736 (95% CI 0.710 - 0.762), when the prediabetes group is added to those with normal glucose metabolism. Similarly, the AUROC of 0.832 for waist circumference would decrease to 0.792 (95% CI 0.770 - 0.815).

In this study, waist circumference was included as a risk factor since multiple studies have shown waist circumference to be a stronger discriminator for type 2 diabetes than Body Mass Index (BMI) (4,25). A post-hoc analysis shows that this is also true for our dataset, where the AUROC for BMI was 0.773 (95% CI 0.748 - 0.798).

On average, the data set contains 7.2 (± 3.6) fundus images per individual and the final glucose metabolism score is obtained by averaging across images of the left and right eye. The use of multiple fundus images per individual potentially improves the accuracy of the GMscore for two reasons: (1) different images focus on different parts of the retina and (2) averaging across multiple examples makes the prediction more robust. However, additional images also require more time to collect. We therefore studied the effect of the number of fundus images per individual by selecting all individuals for whom at least five images were available (83% of individuals). For this subset we recalculated the AUROC using 1, 2, 3, 4 or 5 images which were sampled randomly without replacement. Results are shown in Figure 7.3. When only one image is used per individual, the AUROC decreases to 0.715 (95% CI 0.685 – 0.745). Starting at three images per individual, the AUROC is similar to the one obtained when all

Table 7.2: Performance of the algorithm in comparison with other risk factors. AUROC = Area under receiver operating characteristic. GMscore = glucose metabolism score, obtained using the deep learning algorithm. *Some individuals (< 1.3%) were excluded for AUROC calculation because of missing data points.

Factors used for classification	**AUROC (95% CI)**
GMscore	0.757 (0.731 - 0.783)
Age	0.691 (0.663 - 0.719)
Sex	0.607 (0.580 - 0.634)
Age, Sex	0.711 (0.682 - 0.740)
Waist	0.832 (0.810 - 0.853)
Age, Sex, Waist	0.853 (0.832 - 0.873)
Smoking*	0.581 (0.551 - 0.610)
Hypertension*	0.727 (0.704 - 0.749)
Family history*	0.657 (0.625 - 0.689)
Age, Sex, Waist, Smoking, Hypertension, Family history*	0.888 (0.870 - 0.906)
GMscore, Age, Sex	0.773 (0.747 - 0.799)
GMscore, Age, Sex, Waist	0.867 (0.846 - 0.888)
GMscore, Age, Sex, Waist, Smoking, Hypertension, Family history*	0.895 (0.878 - 0.912)

images are used.

The performance stratified per different fixation of the images varied. The image-level AUROC for images centered on the optic disc was found to be 0.731 (95% CI 0.713 - 0.750); macula: 0.737 (95% CI 0.720 - 0.754); periphery: 0.715 (95% CI 0.690 - 0.740); and other: 0.713 (95% CI 0.688 - 0.738).

Figure 7.4 shows boxplots of the glucose metabolism scores as computed by the deep learning algorithm for the individuals of the validation set. Even though the algorithm was trained using only examples from individuals with normal glucose metabolism or type 2 diabetes, the prediabetes group has a distinct distribution, in between the two other groups. The means of the normal and prediabetes groups are significantly different (p-value < 0.001). The AUROC for discriminating prediabetes individuals from individuals with normal glucose metabolism was 0.611 (95% CI 0.575 - 0.646).

For the matching cohort experiment, the mean matching distance was 0.98 years (age) and 1.5 cm (waist circumference). The AUROC for discriminating between individuals with normal glucose metabolism and newly discovered type 2 diabetes in this set was found to be 0.549 (0.500-0.597).

Heat maps 7.5 were constructed to visualize which regions of the fundus image contribute to high glucose metabolism scores [138]. The examples shown in Figure 7.5 are from individuals with type 2 diabetes, correctly identified as such (true positives). For these individuals, the algorithm focusses on selective parts of the vascular tree. Although some of the heat maps are more diffuse than the presented examples, in general the focus seems to be on the venules and arterioles and not on the optic disc. We also looked at heat maps of false positives and observed that they looked similar to those of true positives, focusing on parts of the vascular tree.

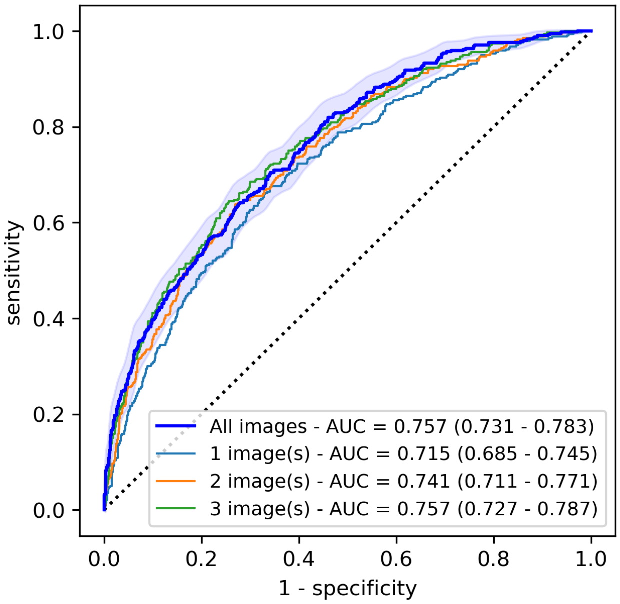

Figure 7.3: Receiver operating characteristic (ROC) curve. The light-blue area represents the 95% CI for the ROC calculated with all images.

7.4 Conclusion

In this study, a deep learning model was developed to obtain a glucose metabolism score (GMscore) from fundus images that can be used to discriminate between individuals with normal glucose metabolism and those with known type 2 diabetes. The AUROC obtained with the GMscore is higher than those obtained with age, sex, smoking, hypertension, and family history. Only waist circumference provided a higher AUROC. These results indicate that fundus images could be more informative for discriminating type 2 diabetes than some of the other well-known risk factors. A possible explanation is that the retina contains information about multiple of these factors. For example, Poplin et al. [120] showed that deep learning can be used to determine age, sex, and smoking status from fundus images relatively well. However, fundus images seem to contain additional information about the glucose tolerance status of individuals. This is supported by our results, which show that even when multiple risk factors are

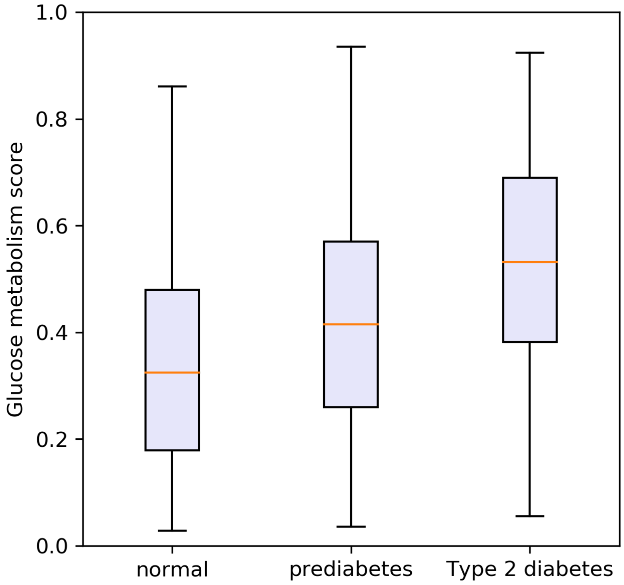

Figure 7.4: Boxplots of glucose metabolism scores for individuals with normal glucose metabolism, prediabetes, and known type 2 diabetes.

combined, the AUROC increases when the GMscore is added. Similarly, this finding is supported by the matching cohort experiment results that showed that the AUROC is modestly higher than 0.5, even though matching was done for age, sex, and waist circumference.

Averaging across multiple fundus images per individual seems to have a beneficial effect on the quality of the GMscore score, as it leads to a higher AUROC. There is a flattening effect from 3 images onwards, and future research could consider including 3 images instead of just one per individual. It should be noted that for this analysis only individuals were included that had at least 5 fundus images available which could introduce a selection bias. However, since the subset for this analysis contains 83% of individuals, the effect on the AUROC should be small. We also found that the fixation of fundus images influences the

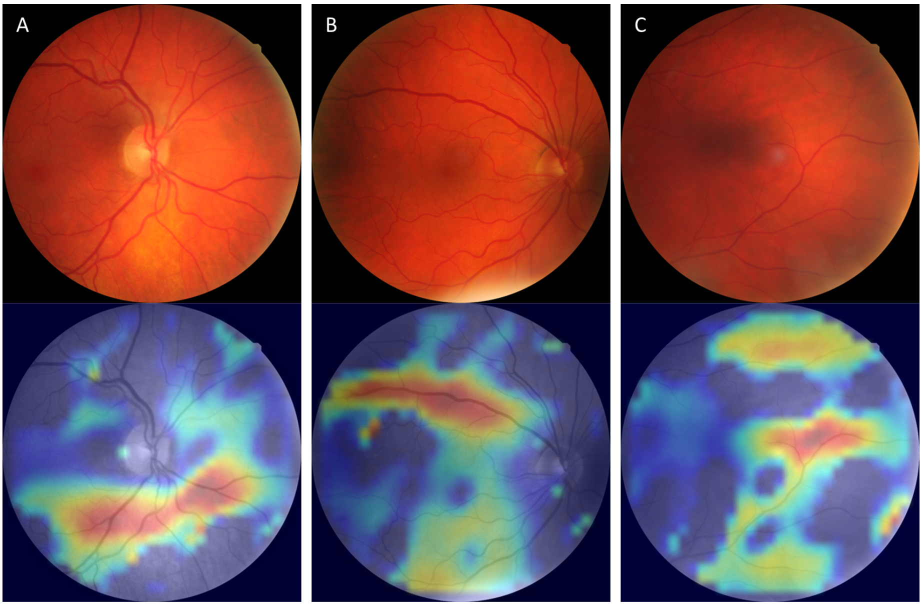

Figure 7.5: Heat maps (bottom) with corresponding fundus image (top) with high glucose metabolism scores from individuals with type 2 diabetes. Red regions in the heat map represent areas that contribute strongly to the glucose metabolism score. (A) Optic disc centered. (B) Macula centered. (C) Periphery centered.

image-level AUROC. The GMscore obtained from fundus images centered at the optic disc or macula results in a higher AUROC than those obtained from the periphery. A possible explanation for this is that more of the vascular tree is visible in optic disc and macula centered images, compared to other fixations.

Even though the algorithm was trained to separate normal glucose metabolism from type 2 diabetes fundus images, we found that the GMscore can actually be used to discriminate prediabetes from normal glucose metabolism individuals to some extent, with an AUROC of 0.611. The distribution of GMscores for prediabetes individuals falls in between that for type 2 diabetes and normal glucose metabolism, indicating that the deep learning model finds modest signs of type 2 diabetes for the prediabetes group.

Our deep learning approach does not require any handcrafted features, such as arteriolar width, but learns the relevant features directly from the data. An advantage of this strategy is that no priors are needed and that no unknown features are missed. A disadvantage that is often attributed to deep learning is the limited explainability of the decision making due to the large number of model weights. The use of heat maps [138] in this study allows for insight into the focus of the algorithm. For individuals with a high GMscore, the prediction often seems to originate from selective parts of the vascular tree. This makes sense, since studies have shown the effect of (pre)diabetes on the microvessels, such as the impaired microvascular function and different calibers as described by Sörensen et al. [19] and Li et al. [90].

In contrast to the extensive research on diabetic retinopathy detection [11, 17], only limited work has been done on direct classification of type 2 diabetes from fundus images via deep learning. Exploratory studies [2, 72] have shown the feasibility of deep learning for type 2 diabetes detection, while one other study tried to determine haemoglobin A1c (HbA1c) levels with limited success [120]. We used the results from previous research [72] for model design choice, including weight initialization, image preprocessing, data augmentation, training strategies, and patient-level aggregation of GMscores. To the best of our knowledge, the study presented here is the first clinical evaluation of the value of deep learning on fundus images for type 2 diabetes detection.

We showed that the AUROC obtained with waist circumference decreased from 0.832 to 0.792 if the prediabetes group was added to those with normal glucose metabolism. Interestingly this is still higher than reported by others. For example, in a meta-analysis by Lee et al. the AUROC for waist circumference was reported to be 0.70 for men and 0.74 for women, although with great disparity between ethnic groups [85].

One limitation of our study is that we developed our methods and validated our results on data from the same distribution (e.g. demographics, fundus photography settings). Future research should focus on applying these models to fundus and glucose metabolism status data collected at different centers.

Another limitation is that we did not take into account the length of an individual's diabetes history, nor the delay between the glucose metabolism status measurements and the fundus photo acquisition. Both time components can potentially impact the manifestation in e.g. the retinal microvessels. Similarly, we did not consider the effect of diabetes treatment on the microvascular system, even though medication reduces the harmful effects of high blood glucose, potentially partly reversing the impaired microvascular function and the changes in the fundus images. The effect of the reversibility and time components could be topic of future research, for example by specifying groups for which medication is (in)effective, or by grouping participants by length of their diabetes history.

In conclusion, it was shown that fundus images are informative for discriminating individuals with normal glucose metabolism, prediabetes, and type 2 diabetes. Using deep learning, a glucose metabolism score was obtained that proved more predictive than other risk factors, except for waist circumference.

7.5 Acknowledgements

The data used in chapters 6 and 7 originates from The Maastricht Study which was supported by the European Regional Development Fund via OP-Zuid, the Province of Limburg, the Dutch Ministry of Economic Affairs (grant 31O.041), Stichting De Weijerhorst (Maastricht, The Netherlands), the Pearl String Initiative Diabetes (Amsterdam, The Netherlands), CARIM School for Cardiovascular Diseases (Maastricht, The Netherlands), Stichting Annadal (Maastricht, The Netherlands), Health Foundation Limburg (Maastricht,The Netherlands) and by unrestricted grants from Janssen-Cilag B.V. (Tilburg, The Netherlands), Novo Nordisk Farma B.V. (Alphen aan den Rijn, The Netherlands), and Sano-Aventis Netherlands B.V. (Gouda, The Netherlands).

Chapter 8

Discussion

The research described in this thesis shows that deep learning can be used to tackle a wide variety of unsolved challenges in corneal and retinal image analysis. In post-operative optical coherence tomography (OCT) images of the cornea, we have shown that it is possible to automatically quantify graft detachment after transplantation surgery (Chapter 2) and measure corneal thickness (Chapter 3) with high accuracy. Using intra-operative OCT images, the orientation of the transplanted corneal graft can be determined automatically (Chapter 4). Zooming in on the corneal endothelium with specular microscopy, we have shown that endothelial cell density can be determined without the need for user input (Chapter 5). Finally, we have shown that fundus photography images can be used to detect type 2 diabetes (Chapter 6) and prediabetes (Chapter 7).

The exponentially growing interest in deep learning in recent years has led to a large toolbox for research and applications. This toolbox includes theoretical frameworks and methods such as model architectures, data preprocessing methods, and training strategies. It also includes dedicated software packages and implementations of various methods, which can function as building blocks to address new research topics. The contributions in this thesis consists of development of novel image analysis pipelines, aggregating and building upon state-of-the-art deep learning methods. We did so for a variety of applications, including previously untouched clinical topics. Each proposed pipeline consists of several steps, selected or designed to meet the requirements of the respective image analysis challenge. For example, in Chapter 2, we used the well-known U-Net [128] model for segmentation of graft detachments in anterior segment

OCT (AS-OCT) images. To ensure that the images were centered, horizontally aligned, and cropped such that only the relevant parts were visible, we used the location of the scleral spur, an anatomical landmark that is relatively unaltered during surgery. For localization of the scleral spur we modified a ResNet-50 [67] architecture to output the horizontal and vertical coordinates of the scleral spur in left and right side of the image. For visualization of the graft detachments, we developed an algorithm to project the graft detachments of multiple cross-sections onto a grid, resulting in a map of the detachments.

For most image analysis methods it makes sense to compare their performance on various aspects to that of clinical specialists. A major benefit of deep learning methods is the speed at which data can be processed. Although we did not always report inference times in this thesis, it took typically less than a second to process an image. Manual delineation of the cornea in AS-OCT to obtain thickness profiles (Chapter 3) is laborious and takes up to 10 minutes for a scan consisting of 16 cross-sectional images, making it unfeasible for standard clinical care. Similarly, manual segmentation of corneal endothelial cells in specular microscopy images (Chapter 5) for accurate cell density measurements takes too long and is not a viable option. As a result, manual quantification of biomarkers can be challenging. In Chapter 2 we pointed out that graft detachments are often described as either present or not, or detached more than one-third or not. These simplifications are undesirable but needed since more precise assessment takes too long. In contrast, our deep learning methods provide the ability to quantify biomarkers precisely and fast.

Other researchers have shown that, for specific medical image analysis tasks, deep learning methods can be more accurate than humans. Examples include classification of skin cancer [47] and grading of diabetic retinopathy [59]. We found something similar in Chapter 3, where the error for measuring corneal thickness with three deep learning strategies was smaller than the inter-observer error. In our case, the inter-observer error sometimes resulted from the lack of detail of the human delineations. Variability is inherent to human performance [73] and repetitive tasks can affect human reliability, which is not an issue for automated image analysis tools.

In some cases, deep learning can even be used to obtain information from images that is not considered by clinicians. Although fundus photographs are used by clinicians to detect diabetic retinopathy, these images are not assessed by clinicians to find signs of type 2 diabetes. Nevertheless, in Chapter 6 and Chapter 7 we showed that deep learning methods can be used to discriminate between individuals with normal glucose metabolism and those with type 2 diabetes. Moreover, in Chapter 7 we described that fundus images are more informative for glucose metabolism classification than some well-known factors that are currently used for risk assessment.

Saving time or outperforming clinical specialists are not always the primary goals of an automatic image analysis method. In Chapter 4, we showed that the orientation of a corneal graft can be automatically determined during transplantation surgery. Although the performance of our method was slightly worse than that of two corneal specialists, the automated tool could still benefit the surgery by functioning as an additional safety measure. If the method would be implemented in real-time, it could notify the surgeon when the graft is considered upside-down, and the surgeon could double check their assessment.

Another contribution of this thesis is the ability to relate the image analysis results of multiple images per patient. In Chapter 3, we showed that AS-OCT scans from subsequent visits by a patient can be used to obtain differential corneal thickness maps, which are key for assessing corneal degeneration or regeneration. In Chapter 5, multiple specular microscopy images of a single patient visit were processed. Since processing happens almost instantly, the analysis does not take additional time while it has the potential to be more robust in comparison with analysing only a single image.

So far, we mentioned multiple benefits of using deep learning systems for analysis of ophthalmic images. These systems can potentially be improved further by incorporating assessment strategies used by specialists. For example, for the assessment of corneal graft detachment (Chapter 2), we processed each cross-sectional image individually. A corneal specialist, however, would consider neighboring slices for more contextual information. Incorporating neighboring slices into our deep learning pipeline can be achieved using multiple approaches [154]. Another example is that clinicians will often take into account shape constraints and anatomical priors. In Chapter 4, our method for graft orientation detection includes a segmentation step. In some instances, the automatic segmentation would clearly be wrong, for example when multiple gaps are present in the segmented body. When we say ’clearly’ we refer to the clinical perspective, where the surgeon knows that the graft is in fact a continuous structure and gaps should not be present. Building such shape constrains into the deep learning methods should be topic of future research.

The methods proposed in this thesis have all been developed using specific data sets. Limitations of these data sets (e.g. single scanner, demographics, procedures) should be considered when applying the methods to other situations. Performance of deep learning methods can strongly depend on the data characteristics and domain adaptations might be needed. This could include retraining with data augmentations that overcome the differences in data distributions, or retraining with additional data from a target domain. In Chapter 3, we investigated the effect of the training set size and found that for this example good results can be obtained using only a fraction of the original data set, which should be considered if retraining is required.

At the time of writing this thesis, analysis of ophthalmic images using deep learning is still limited to a research setting, with the noteworthy exception of detection of diabetic retinopathy from fundus photographs [15]. Notwithstanding, the fast pace of research and advances in corneal and retinal image analysis is exciting and likely to lead to the development of products that can improve ophthalmic care. Development of these products should be done with great care and extensive validations will need to be performed. Validation studies should include a diverse population that represents the general patient population as well as patient with non-typical characteristics. Acceptance of artificial intelligence in the clinical workflow will strongly depend on the trust by physicians and patients. Introduction of deep learning should be done cautiously and with clear communication about its abilities and limitations. It will not always be possible to explain the full decision process that follows from millions of weights that make up a deep learning model. Nevertheless, development efforts should be focused towards explainability of the decision making, including visualization of intermediate analysis steps and areas of interest.

In the next decades, we can expect deep learning to find its way into ophthalmic care. Arguments will include improved patient care, additional safety, objectivity, and resource savings. But whichever the motivation, at some point, physicians will be likely to suggest AI for your eye.

Bibliography

[1] TensorFlow: large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. (Cited on pages 12, 40, and 85.)

[2] S. Abbasi-Sureshjani, B. Dashtbozorg, B.M. ter Haar Romeny, and F. Fleuret. Exploratory study on direct prediction of diabetes using deep residual networks. *Lecture Notes in Computational Vision and Biomechanics*, 27:797–802, 2017. (Cited on pages 83 and 102.)

[3] M.D. Abramoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J.C. Folk, and M. Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology and Visual Science*, 57(13:5200–5206, 2016. (Cited on page 91.)

[4] A. P. Adamis, V. Filatov, B. J. Tripathi, and R. C. Tripathi. Comparison of 4 specular microscopes in healthy eyes and eyes with cornea guttata or corneal grafts. *Survey of Ophthalmology*, 38(2):149–168, 1993. (Cited on page 67.)

[5] M.S. Ahyan and P. Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. *1st Conference on Medical Imaging with Deep Learning*, 2018. (Cited on page 85.)

[6] M. Alberti. Air versus SF6 for Descemet's membrane endothelial keratoplasty (DMEK). *https://clinicaltrials.gov/ct2/show/NCT03407755*. (Cited on pages 10 and 28.)

[7] R. Ambrósio Jr, R. S. Alonso, A. Luz, and L. G. Coca Velarde. Corneal-thickness spatial profile and corneal-volume distribution: Tomographic indices to detect keratoconus. *Journal of Cataract and Refractive Surgery*, 32(11):1851–1859, 2006. (Cited on page 27.)

[8] A. Tufail amd C. Rudisill, C. Egan, V.V. Kapetanakis, S. Salas-Vega, C.G. Owen, A. Lee, V. Louw, J. Anderson, G. Liew, L. Bolter, S. Srinivas, M. Nittala, S.V. Sadda, P. Taylor, and A.R. Rudnicka. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. *Ophthalmology*, 124(3):343–351, 2017. (Cited on page 61.)

[9] M. Ang, M. Baskaran, R. M. Werkmeister, J. Chua, D. Schmidl, and V. A. dos Santos. Anterior segment optical coherence tomography. *Progess in Retinal and Eye Research*, 66:132–156, 2018. (Cited on pages 12 and 28.)

[10] M. Ang, M. R. Wilkins, J. S. Mehta, and D. Tan. Descemet membrane endothelial keratoplasty. *British Journal of Ophthalmology*, 100(1):15–21, 2016. (Cited on page 21.)

[11] N. Asiri, M. Hussain, F. Al Adel, and N. Alzaidi. Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. *Artificial Intelligence in Medicine*, 99, 2019. (Cited on page 102.)

[12] G. Ayala, M. E. Díaz, and L. Marínez-Cost. Fully convolutional architecture vs sliding-window CNN for corneal endothelium cell segmentation. *Pattern Recognition*, 34(6):1219–1227, 2001. (Cited on page 76.)

[13] P. Bankhead, M. B. Loughrey, J. A. Fernández, and et al. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1):Article number: 16878, 2017. (Cited on page 68.)

[14] A.S. Bardan, M.B. Goweida, H.F. El Goweini, and C.S.C. Liu. Management of upside-down descemet membrane endothelial keratoplasty: A case series. *Journal of Current Ophthalmology*, 32(4):142–148, 2020. (Cited on page 49.)

[15] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L.M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic Retinopathy. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–12, 2020. (Cited on page 108.)

[16] B.E. Bejnordi, M. Veta, and P.J. van Diest et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 318(22):2199–2210, 2017. (Cited on page 3.)

[17] V. Bellemo, G. Lim, T.H. Rim, G.S.W. Tan, C.Y. Cheung, S.V. Sadda, M.G. He, A. Tufail, M.L. Lee, W. Hsu, and D.S.W. Ting. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. *Current Diabetes Reports*, 19(9):Article number 72, 2019. (Cited on page 102.)

[18] J. M. Bland and D. G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1(8476):307–310, 1986. (Cited on page 15.)

[19] A.J.H.M. Houben B.M. Sörensen, T.T.J.M. Berendschot, J.S.A.G. Schouten, A.A. Kroon, C.J.H. van der Kallen, R.M.A Henry, A. Koster, S.J.S. Sep, P.C. Dagnelie, N.C. Schaper, M.T. Schram, and C.D.A. Stehouwer. Prediabetes and type 2 diabetes are associated with generalized microvascular dysfunction: The Maastricht Study. *Circulation*, 134(18):1339–52, 2016. (Cited on page 102.)

[20] S. Bohlender, I. Oksuz, and A. Mukhopadhyay. A survey on shape-constraint deep learning for medical image segmentation, 2021. (Cited on page 62.)

[21] J. L. Bourges, N. Alfonsi, J. F. Laliberté, M Chagnon, G. Renard, J. M. Legeais, and Brunette I. Average 3-dimensional models for the comparison of Orbscan II and Pentacam pachymetry maps in normal corneas. *Ophthalmology*, 116(11):2064–2071, 2009. (Cited on page 35.)

[22] W.M. Bourne. Biology of the corneal endothelium in health and disease. *Eye*, 17(8):912–918, 2003. (Cited on page 67.)

[23] T.J. Bright, A. Wong, R. Dhurjati, E. Bristow, L. Bastian, R.R. Coeytaux, G. Samsa, V. Hasselblad, J.W. Williams, M.D. Musty, L. Wing, A.S. Kendrick, G.D. Sanders, and D. Lobach. Effect of clinical decision-support systems: a systematic review. *Annals of Internal Medicine*, 157(1):29–43, 2012. (Cited on page 61.)

[24] L.M. Browning, S.D. Hsieh, and M. Ashwell. A systematic review of waist-to-height ratio as a screening tool for the prediction of cardiovascular disease and diabetes: 0.5 could be a suitable global boundary value. *Nutrition Research Reviews*, 23(2):247–269, 2010. (Cited on page 91.)

[25] F. Bucher, D. Hos, S. Müller-Schwefe, P. Steven, C. Cursiefen, and Heindl L. M. Spontaneous long-term course of persistent peripheral graft detachments after Descemet's membrane endothelial keratoplasty. *British Journal of Ophthalmology*, 99:768–772, 2015. (Cited on page 38.)

[26] J. Chen, L. Yang, Y. Zhang, M. Alber, and D.Z. Chen. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. *Advances in Neural Information Processing Systems*, page 3044–3052, 2016. (Cited on page 62.)

[27] E. Cho, D. Min, and H.S. Lee. Development and validation of an undiagnosed diabetes screening tool: Based on the Korean national health and nutrition examination survey (2010 - 2016). *Healthcare*, 9:1138, 2021. (Cited on page 91.)

[28] N.H. Cho, J.E. Shaw, S. Karuranga, Y. Huang, J.D. da Rocha Fernandes, A.W. Ohlrogge, and B. Malanda. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*, 138:271–281, 2018. (Cited on page 83.)

[29] François Chollet et al. Keras. `https://keras.io`, 2015. (Cited on pages 12, 40, and 84.)

[30] N. G. Congdon, D. S. Friedman, and T. Lietman. Important causes of visual impairment in the world today. *JAMA*, 290(15):2057–2060, 2003. (Cited on pages 83 and 91.)

[31] C. Cook and M. Langham. Corneal thickness in interstitial keratitis. *British Journal of Ophthalmology*, 37:301–304, 1953. (Cited on page 27.)

[32] B. Cost, J.M. Goshe, S. Srivastava, and J.P. Ehlers. Intraoperative optical coherence tomography-assisted descemet membrane endothelial keratoplasty in the discover study. *American Journal of Ophthalmology*, 160(3):430–437, 2015. (Cited on pages 49 and 61.)

[33] M. C. Daniel, L. Atzrodt, F. Bucher, and et al. Automated segmentation of the corneal endothelium in a large set of 'real-world' specular microscopy

images using the U-Net architecture. *Scientific Reports*, 9(1):Article number: 4752, 2019. (Cited on pages 67, 68, and 77.)

[34] I. Dapena, K. Moutsouris, K. Droutsas, L. Ham, K. van Dijk, and G.R.J. Melles. Standardized "no-touch" technique for descemet membrane endothelial keratoplasty. *Archives of Ophthalmology*, 129(1):88–94, 2011. (Cited on pages 49 and 51.)

[35] B. Dashtbozorg, S. Abbasi-Sureshjani, J. Zhang, F. Huang, E. Bekkers, and B.M. ter Haar Romeny. Infrastructure for retinal image analysis. *Medical Image Analysis Third International Workshop, OMIA 2016*, page 105–112, 2016. (Cited on page 83.)

[36] S. X. Deng, W. B. Lee, K. M. Hammersmith, A. N. Kuo, Li J. Y., J. F. Shen, M. P. Weikert, and R. M. Shtein. Descemet membrane endothelial keratoplasty: Safety and outcomes: A report by the American Academy of Ophthalmology. *Ophthalmology*, 125(2):295–310, 2018. (Cited on page 9.)

[37] S. X. Deng, P. J. Sanchez, and L. Chen. Clinical outcomes of Descemet membrane endothelial keratoplasty using eye bank–prepared tissues. *American Journal of Ophthalmology*, 159:590–596, 2015. (Cited on page 38.)

[38] V.G. Dhommati, K.K. Vupparaboina, K. Challa, S. Jana, A. Richhariya, and J.C. Reddy. Automated 2d-3d quantitative analysis of corneal graft detachment post dsaek based on as-oct images. *Computer Methods and Programs in Biomedicine*, 167:1–12, 2018. (Cited on page 61.)

[39] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. (Cited on page 14.)

[40] M. Dirisamer, K. van Dijk, I. Dapena, L. Ham, O. Oganes, L. E. Frank, and G. R. J. Melles. Prevention and management of graft detachment in Descemet membrane endothelial keratoplasty. *Archives of Ophthalmology*, 130(3):280–291, 2012. (Cited on page 21.)

[41] V. A. Dos Santos, L. Schmetterer, H. Stegmann, M. Pfister, A. Messner, G. Schmidinger, G. Garhofer, and R. M. Werkmeister. CorneaNet: fast segmentation of cornea OCT scans of healthy and keratoconic eyes using deep learning. *Biomedical Optics Express*, 10(2):622–641, 2019. (Cited on pages 12, 27, 38, and 40.)

[42] Q. Dou, D. C. Castro, K. Kamnitsas, and B. Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32:6447–6458, 2019. (Cited on page 22.)

[43] J.P. Ehlers. Intraoperative optical coherence tomography: Past, present, and future. *Eye*, 30(2):193–201, 2016. (Cited on page 61.)

[44] J.P. Ehlers, Y.S. Modi, P.E. Pecen, J. Goshe, W.J. Dupps, A. Rachitskaya, S. Sharma, A. Yuan, R. Singh, P.K. Kaiser, J.L. Reese, C. Calabrise, A. Watts, and S.K. Srivastava. The DISCOVER study 3-year results: Feasibility and usefulness of microscope-integrated intraoperative OCT during ophthalmic surgery. *Ophthalmology*, 125(7):1014–1027, 2018. (Cited on page 61.)

[45] J.P. Ehlers, A. Uchida, and S. K. Srivastava. The integrative surgical theater: combining intraoperative OCT and 3D digital visualization for vitreoretinal surgery in the DISCOVER study. *Retina*, 38(Suppl 1):S88–S96, 2018. (Cited on page 61.)

[46] J.P. Ehlers, A. Uchida, S. K. Srivastava, and M. Hu. Predictive model for macular hole closure speed: Insights from intraoperative optical coherence tomography. *Translational Vision Science & Technology*, 8(1):18, 2019. (Cited on page 61.)

[47] A. Esteva, B. Kuprel, and R. Novoa et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017. (Cited on pages 3 and 106.)

[48] A. Fabijanska. Corneal endothelium image segmentation using feedforward neural network. *Proceedings of FedCSIS 2017*, pages 629–637, 2017. (Cited on pages 76 and 77.)

[49] A. Fabijańska. Segmentation of corneal endothelium images using a U-Net-based convolutional neural network. *Artificial Intelligence in Medicine*, 88:1–13, 2018. (Cited on pages 67, 76, and 77.)

[50] C. E. H. Fang, P. T. Khaw, R. G. Mathew, and C. Henein. Corneal endothelial cell density loss following glaucoma surgery alone or in combination with cataract surgery: A systematic review protocol. *BMJ Open*, 11(9), 2021. (Cited on page 67.)

[51] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomedical Optics Express*, 8(5):2732–2744, 2017. (Cited on pages 38 and 40.)

[52] E. Fernández López, L. Baydoun, N. Gerber-Hollbach, I. Dapena, V. S. Liarakos, L. Ham, and G. R. J. Melles. Rebubbling techniques for graft detachment after Descemet membrane endothelial keratoplasty. *Cornea*, 35(6):759–764, 2016. (Cited on page 21.)

[53] M. Foracchia, E. Grisan, and A. Ruggeri. Luminosity and contrast normalization in retinal images. *Medical Image Analysis*, 9(3):179–190, 2005. (Cited on pages 84 and 93.)

[54] M. Foracchia and A. Ruggeri. Corneal endothelium cell field analysis by means of interacting Bayesian shape models. *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6035–6038, 2007. (Cited on page 76.)

[55] H. Fu, M. Baskaran, Y. Xu, S. Lin, D. W. K. Wong, J. Lui, T. A. Tun, M. Mahesh, S. A. Perera, and T. Aung. A deep learning system for automated angle-closure detection in anterior segment optical coherence tomography images. *American Journal of Ophthalmology*, 203:37–45, 2019. (Cited on page 27.)

[56] L. Gasser, T. Reinhard, and D. Böhringer. Comparison of corneal endothelial cell measurements by two non-contact specular microscopes. *BMC Ophthalmology*, 15(1):Article number: 87, 2015. (Cited on page 67.)

[57] Y. Gavet and Pinoli J. C. Visual perception based automatic recognition of cell mosaics in human corneal endothelium microscopy images. *Image Analysis and Stereology*, 27:53–61, 2008. (Cited on page 76.)

[58] K. Gu, C. C. Cowie, and M. I. Harris. Diabetes and decline in heart disease mortality in US adults. *JAMA*, 281(14):1291–1297, 1999. (Cited on pages 83 and 91.)

[59] V. Gulshan, L. Peng, and M. Coram. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Journal of the American Medical Association*, 316(22):2402–2410, 2016. (Cited on pages 76, 83, 91, and 106.)

[60] L.J. Haddock, D.Y. Kim, and S. Mukai. Simple, inexpensive technique for high-quality smartphone fundus photography in human and animal eyes. *Journal of Ophthalmology*, Article ID 518479, 2013. (Cited on page 88.)

[61] K.M. Hallahan, B. Cost, J.M. Goshe, W.J. Dupps, S.K. Srivastava, and J.P. Ehlers. Intraoperative interface fluid dynamics and clinical outcomes for intraoperative optical coherence tomography–assisted descemet stripping automated endothelial keratoplasty from the pioneer study. *American Journal of Ophthalmology*, 173:16–22, 2017. (Cited on page 61.)

[62] L. Ham, C. van Luijk, I. Dapena, T. H. Wong, R. Birbal, J. van der Wees, and G. R. J. Melles. Endothelial Cell Density after Descemet Membrane Endothelial Keratoplasty: 1- to 2-Year Follow-up. *American Journal of Ophthalmology*, 148(4):521–527, 2009. (Cited on page 67.)

[63] J. Hamwood, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins. Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers. *Biomedical Optics Express*, 9:3049–3066, 2018. (Cited on page 38.)

[64] R. Harper. The Code of Hammurabi, King of Babylon, about 2250 B.C. *Chicago: University of Chicago Press*, 1904. (Cited on page 1.)

[65] H. Hashemi, S. Asgari, S. Mehravaran, M. .H. Emamian, M. Shariati, and A. Fotouhi. The distribution of corneal thickness in a 40- to 64-year-old population of Shahroud, Iran. *Cornea*, 30(12):1409–1413, 2011. (Cited on page 37.)

[66] T. Hayashi, H. Tabuchi, H. Masumoto, S. Morita, I. Oyakawa, S. Inoda, N. Kato, and H. Takahashi. A deep learning approach in rebubbling after Descemet's membrane endothelial keratoplasty. *Eye & Contact Lens*, 46(2):121–126, 2020. (Cited on pages 22 and 61.)

[67] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. (Cited on pages 13 and 106.)

[68] W. H. Herman, W. Ye, S. J. Griffin, R. K. Simmons, M. J. Davies, K. Khunti, G. E. H. M. Rutten, A. Sandbaek, T. Lauritzen, K. Borch-Johnsen, M. B. Brown, and N. J. Wareham. Early detection and treatment of type 2 diabetes reduce cardiovascular morbidity and mortality: A simulation of the results of the Anglo-Danish-Dutch study of intensive treatment in people

with screen-detected diabetes in primary care (ADDITION-Europe). *Diabetes Care*, 38(8):1449–1455, 2015. (Cited on page 83.)

[69] F. G. Heslinga, M. Alberti, J. P. W. Pluim, J. Cabrerizo, and M. Veta. Quantifying graft detachment after Descemet's membrane endothelial keratoplasty with deep convolutional neural networks. *Translational Vision Science & Technology*, 9(2):48, 2020. (Cited on pages 27, 28, 29, 37, 38, 44, 61, and 77.)

[70] F. G. Heslinga, R. T. Lucassen, M. A. van den Berg, L. van der Hoek, J. P. W. Pluim, J. Cabrerizo, M. Alberti, and M. Veta. Corneal pachymetry by AS-OCT after Descemet's membrane endothelial keratoplasty. *Scientific Reports*, 11(1):13976, 2021. (Cited on page 77.)

[71] F.G. Heslinga, J.P.W. Pluim, B. Dashtbozorg, T.T.J.M. Berendschot, A.J.H.M. Houben, R.M.A. Henry, and M. Veta. Approximation of a pipeline of unsupervised retina image analysis methods with a CNN. *Proceedings of SPIE 10949, Medical Imaging 2019: Image Processing*, 10949N, 2019. (Cited on pages 83, 85, and 91.)

[72] F.G. Heslinga, J.P.W. Pluim, A.J.H.M. Houben, M.T. Schram, R.M.A. Henry, D.A. Stehouwer, M.J. Van Greevenbroek, T.T.J.M. Berendschot, and M. Veta. Direct classification of type 2 diabetes from retinal fundus images in a population-based sample from The Maastricht Study. *Proceedings of SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis*, 11314, 2020. (Cited on page 102.)

[73] G. Heslinga. PhD thesis: Technique for human-error-sequence identification and signification. *Bibliotheek Technische Universiteit Delft*, 1988. (Cited on page 106.)

[74] J. Huang, J. Maram, T. C. Tepelus, and et al. Comparison of manual & automated analysis methods for corneal endothelial cell density measurements by specular microscopy. *Journal of Optometry*, 11(3):182–191, 2018. (Cited on pages 67 and 75.)

[75] D.K. Hwang, C.C. Hsu, K.J. Chang, D. Chao, C.H. Sun, Y.C. Jheng, A.A. Yarmishyn, J.C. Wu, C.Y. Tsai, M.L. Wang, C.H. Peng, K.H. Chien, C.L. Kao, T.C. Lin, L.C. Woung, S.J. Chen, and S.H. Chiou. Artificial intelligence-based decision-making for age-related macular degeneration. *Theranostics*, 9(1):232–245, 2019. (Cited on page 61.)

[76] N. Joseph, C. Kolluru, B. A. M. Benetz, H. J. Menegay, J. H. Lass, and D. L. Wilson. Quantitative and qualitative evaluation of deep learning automatic segmentations of corneal endothelial cell images of reduced image quality obtained following cornea transplant. *Journal of Medical Imaging*, 7(1):014503, 2020. (Cited on pages 67 and 77.)

[77] S. Katafuchi and M. Yoshimura. Convolution neural network for contour extraction of corneal endothelial cells. *International Conference on Quality Control by Artificial Vision*, 2017. (Cited on page 77.)

[78] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. (Cited on pages 40 and 54.)

[79] L. J. Kopplin, K. Przepyszny, B. Schmotzer, K. Rudo, Babineau D. C., S. V. Patel, D. D. Verdier, U. Jurkunas, S. K. Iyengar, and J. H. Lass. Relationship of Fuchs' endothelial corneal dystrophy severity to central corneal thickness. *Archives of Ophthalmology*, 130(4):433–439, 2012. (Cited on page 27.)

[80] A. Krizhevsky, I. Stutskever, and Hinton G. E. ImageNet classification with deep convolutional neural networks. *Archives of Ophthalmology*, 60(6):84–90, 2017. (Cited on page 13.)

[81] D.J. Kroon. 2d line curvature and normals. `https://www.mathworks.com/matlabcentral/fileexchange/32696-2d-line-curvature-and-normals`, 2011. (Cited on page 54.)

[82] J. Kugelman, D. Alonso-Caneiro, S. A. Read, J. Hamwood, S. J. Vincent, F. K. Chen, and M. J. Collins. Automatic choroidal segmentation in OCT images using supervised deep learning methods. *Scientific Reports*, 9(1):art. no. 13298, 2019. (Cited on page 38.)

[83] J. Kugelman, D. Alonso-Caneiro, S. A. Read, and M. J. Vincent, S. J. Collins. Automatic segmentation of OCT retinal boundaries using recurrent neural networks and graph search. *Biomedical Optics Express*, 9:5759–5777, 2018. (Cited on page 38.)

[84] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015. (Cited on pages 3, 12, 27, 53, 67, 70, 91, and 94.)

[85] C.M.Y. Lee, R.R. Huxley, R.P. Wildman, and M. Woodward. Indices of abdominal obesity are better discriminators of cardiovascular risk factors

than bmi: a meta-analysis. *Journal of Clinical Epidemiology*, 61(7):646–653, 2008. (Cited on page 102.)

[86] T.C. Lee, R.L. Kashyap, and C.N. Chu. Building skeleton models via 3-d medial surface axis thinning algorithms. *CVGIP: Graphical Models and Image Processing*, 56(6):462–478, 1994. (Cited on page 54.)

[87] Y.H. Lee, H. Bang, H.C. Kim, H.M. Kim, S.W. Park, and D.J. Kim. A simple screening score for diabetes for the Korean population: Development, validation, and comparison with other scores. *Diabetes Care*, 35(8):1723–1730, 2012. (Cited on page 91.)

[88] C. Leibig, V. Allken, M.S. Ayhan, P. Berens, and S. Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Nature Scientific Reports*, 7:Article ID 17816, 2017. (Cited on page 88.)

[89] W. Li, M.T. Schram, T.T.J.M. Berendschot, C.A.B. Webers, A.A. Kroon, C.J.H. van der Kallen, R.M.A. Henry, N.C. Schaper, F. Huang, B. Dashtbozorg, T. Tan, J. Zhang, S. Abbasi-Sureshjani, B.M. ter Haar Romeny, C.D.A. Stehouwer, and A.J.H.M. Houben. Type 2 diabetes and HbA1c are independently associated with wider retinal arterioles: The Maastricht Study. *Diabetologia*, 63(7):1408–1417, 2020. (Cited on page 91.)

[90] W. Li, M.T. Schram, Ben M. Sörensen, M.J.M. van Agtmaal, T.T.J.M. Berendschot, C.A.B. Webers, J.F.A.Jansen, W.H. Backes, E.H.B.M. Gronenschild, C.G. Schalkwijk, C.D.A. Stehouwer, and A.J.H.M. Houben. Microvascular phenotyping in The Maastricht Study: Design and main findings, 2010-2018. *American Journal of Epidemiology*, 189(9):873–884, 2020. (Cited on page 102.)

[91] Y. Li, D. M. Meisler, M. Tang, A. T. H. Lu, V. Thakrar, B. J. Reiser, and D. Huang. Keratoconus diagnosis with optical coherence tomography pachymetry mapping. *Ophthalmology*, 115(12):2159–2166, 2008. (Cited on page 27.)

[92] Y. Li, R. Shekhar, and D. Huang. Corneal pachymetry mapping with high-speed optical coherence tomography. *Ophthalmology*, 113(5):792–799, 2006. (Cited on pages 31 and 43.)

[93] B. Liefers, C. González-Gonzalo, B. van Ginneken, and C. I. Sánchez. Dense segmentation in selected dimensions: application to retinal optical coherence tomography. *International Conference on Medical Imaging*

with Deep Learning, pages 337–346, 2019. (Cited on pages 12, 38, 40, and 42.)

[94] S. H. Lim. Clinical applications of anterior segment optical coherence tomography. *Journal of Ophthalmology*, pages 1–12, 2015. (Cited on page 27.)

[95] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, and C.I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. (Cited on pages 3, 12, and 53.)

[96] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang. Structured knowledge distillation for semantic segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 2599–2608, 2019. (Cited on page 63.)

[97] N. Luft, N. Hirnschall, S. Schuschitz, P. Draschl, and O. Findl. Comparison of 4 specular microscopes in healthy eyes and eyes with cornea guttata or corneal grafts. *Cornea*, 34:381–386, 2015. (Cited on page 67.)

[98] R. Ma, Y. Liu, L. Zhang, Y. Lei, J. Hou, Z. Shen, X. Yi, and Y. Wang. Distribution and trends in corneal thickness parameters in a large population-based multicenter study of young Chinese adults. *Investigative Ophthalmology Visual Science*, 59:3366–3374, 2018. (Cited on page 37.)

[99] J.N. Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010. (Cited on pages 56 and 61.)

[100] R. E. Marques, P. S. Guerra, D. C. Sousa, A. I. Gonçalves, A. M. Quintas, and W. Rodrigues. DMEK versus DSAEK for Fuchs' endothelial dystrophy: A meta-analysis. *European Journal of Ophthalmology*, 29(1):15–22, 2019. (Cited on page 21.)

[101] D. M. Maurice. Cellular membrane activity in the corneal endothelium of the intact eye. *Experientia*, 24:1094–1095, 1968. (Cited on page 75.)

[102] B. E. McCarey, H. F. Edelhauser, and M. J. Lynn. Review of corneal endothelial specular microscopy for FDA clinical trials of refractive procedures, surgical devices, and new intraocular drugs and solutions. *Cornea*, 27(1):1–16, 2008. (Cited on pages 67 and 76.)

[103] J. W. McLaren, L. A. Bachman, K. M. Kane, and S. V. Patel. Objective assessment of the corneal endothelium in Fuchs' endothelial dystrophy. *Investigative Ophthalmology and Visual Science*, 55(2):1184–1190, 2014. (Cited on pages 67, 71, and 75.)

[104] G. R. J. Melles, T. S. Ong, B. Ververs, and J. van der Wees. Descemet membrane endothelial keratoplasty (DMEK). *Cornea*, 25(8):987–990, 2006. (Cited on pages 1, 9, and 27.)

[105] F. Meyer and Beucher. S. Morphological segmentation. *Journal of visual communication and image representation*, 1(1):21–46, 1990. (Cited on page 71.)

[106] M. Modabber, J.C. Talajic, M. Mabon, M. Mercier, S. Jabbour, and J. Choremis. The role of novel dmek graft shapes in facilitating intraoperative unscrolling. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 256(12):2385–2390, 2018. (Cited on page 49.)

[107] I. Mohammed, A.R. Ross, J.O. Britton, D.G. Said, and H.S. Duan. Elastin content and distribution in endothelial keratoplasty tissue determines direction of scrolling. *American Journal of Ophthalmology*, 194:16–25, 2018. (Cited on page 49.)

[108] M.B. Muijzer, P.A.W.J. Schellekens, H.J.M. Beckers, J.H. de Boer, S.M. Imhof, and R.P.L. Wisse. Clinical applications for intraoperative optical coherence tomography: a systematic review. *Eye*, page 1–13, 2021. (Cited on pages 49 and 61.)

[109] M.B. Muijzer, N. Soeters, D.A. Godefrooij, C.M. van Luijk, and R.P.L. Wisse. Intraoperative optical coherence tomography-assisted descemet membrane endothelial keratoplasty. *Cornea*, 39(6):674–679, 2020. (Cited on pages 49 and 61.)

[110] T.T. Nguyen, J.J. Wang, A. Richey Sharrett, F.M. Amirul Islam, R. Klein, B.E.K. Klein, M.F. Cotch, and T.Y. Wong. Relationship of retinal vascular caliber with diabetes and retinopathy: the Multi-Ethnic Study of Atherosclerosis (MESA). *Diabetes Care*, 31:544–549, 2008. (Cited on page 91.)

[111] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi. Clinical decision support in the era of artificial intelligence. *The American Journal of Medicine*, 132(7):795–801, 2020. (Cited on page 61.)

[112] K. Nurzynska. Deep learning as a tool for automatic segmentation of corneal endothelium images. *Diabetologia*, 10(3), 2018. (Cited on page 77.)

[113] Eye Bank Association of America. 2018 eye banking statistical report. `https://restoresight.org/what-we-do/publications/statistical-report/`, 2018. (Cited on page 21.)

[114] S. Ong Tone and U. Jurkunas. Imaging the corneal endothelium in Fuchs corneal endothelial dystrophy. *Seminars in Ophthalmology*, 34(4):340–346, 2019. (Cited on page 67.)

[115] World Health Organization and International Diabetes Federation. definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a who/idf consultation. (Cited on page 93.)

[116] J. Parker, J.S. Parker, and G.R.J. Melles. Descemet membrane endothelial keratoplasty — a review. *US Ophthalmic Review*, 6(1):29–32, 2013. (Cited on page 49.)

[117] A.S. Patel, J.M. Goshe, S. Srivastava, and J.P. Ehlers. Intraoperative optical coherence tomography-assisted descemet membrane endothelial keratoplasty in the discover study: First 100 cases. *American Journal of Ophthalmology*, 210:167–173, 2020. (Cited on pages 49 and 61.)

[118] S. V. Patel, D. O. Hodge, E. J. Treichel, M. R. Spiegel, and K. H. Baratz. Predicting the prognosis of Fuchs endothelial corneal dystrophy by using Scheimpflug tomography. *Ophthalmology*, 127(3):315–323, 2020. (Cited on pages 27 and 35.)

[119] P. M. Phillips, L. J. Phillips, V. Muthappan, C. M. Maloney, and Carver C. N. Experienced DSAEK surgeon's transition to DMEK: Outcomes comparing the last 100 DSAEK surgeries with the first 100 DMEK surgeries exclusively using previously published techniques. *Cornea*, 36(3):275–279, 2017. (Cited on page 21.)

[120] R. Poplin, A.V. Varadarajan, K. Blumer, Y. Liu, M.V. McConnell, G.S. Corrado, L. Peng, and D.R. Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2:158–164, 2018. (Cited on pages 83, 95, 100, and 102.)

[121] M. O. Price, K. M. Fairchild, and F. W. Price. Automated endothelial cell density analysis in normal eyes and DSEK eyes. *Cornea*, 32:567–573, 2013. (Cited on page 67.)

[122] R. Quilendrino, M. Rodríguez-Calvo-de-Mora, L. Baydoun, L. Ham, K. van Dijk, and I. Dapena. Prevention and management of Descemet membrane endothelial keratoplasty complications. *Cornea*, 36(9):1089–1095, 2017. (Cited on page 21.)

[123] N. Quinn, A. Jenkins, C. Ryan, A. Januszewski, T. Peto, and L. Brazionis. Imaging the eye and its relevance to diabetes care. *Journal of Diabetes Investigation*, 12(6):897–908, 2021. (Cited on page 91.)

[124] T. Röck, M. Bramkamp, K. U. Bartz-Schmidt, D. Röck, and E. Yörük. Causes that influence the detachment rate after Descemet membrane endothelial keratoplasty. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 253:2217–2222, 2015. (Cited on pages 21 and 38.)

[125] M. Rodríguez-Calvo-de-Mora, R. Quilendrino, L. Ham, V. S. Liarakos, K. van Dijk, L. Baydoun, I. Dapena, S. Oellerich, and G. R. J. Melles. Clinical outcome of 500 consecutive cases undergoing Descemet's membrane endothelial keratoplasty. *Ophthalmology*, 122(3):464–470, 2015. (Cited on page 9.)

[126] T.W. Rogers, N. Jaccard, F. Carbonaro, H.G. Lemij, K.A. Vermeer, N.J. Reus, and S. Trikha. Evaluation of an ai system for the automated detection of glaucoma from stereoscopic optic disc photographs: the european optic disc assessment study. *Eye*, 33(11):1791–1797, 2019. (Cited on page 61.)

[127] H. C. Römkens, H. J. Beckers, Berendschot T. T. J. M. Schouten, J. S., and C. A. Webers. Reference values for anterior chamber morphometrics with swept-source optical coherence tomography in a Caucasian population. *Clinical Ophthalmology*, 12:411–417, 2018. (Cited on page 13.)

[128] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. (Cited on pages 3, 14, 31, 40, 41, 53, 70, and 105.)

[129] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomedical Optics Express*, 8(8):3627–3642, 2017. (Cited on pages 12 and 38.)

[130] A. Saad, E. Guilbert, A. Grise-Dulac, P. Sabatier, and D. Gatinel. Intraoperative oct-assisted dmek: 14 consecutive cases. *Cornea*, 34(7):802–807, 2015. (Cited on pages 49 and 61.)

[131] Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Belma Malanda, Suvi Karuranga, Nigel Unwin, Stephen Colagiuri, Leonor Guariguata, Ayesha A. Motala, Katherine Ogurtsova, Jonathan E. Shaw, Dominic Bright, and Rhys Williams. global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition. *Diabetes Research and Clinical Practice*. (Cited on page 91.)

[132] L. M. Sakata, R. Lavanya, D. S. Friedman, H. T. Aung, S. K. Seah, P. J. Foster, and T. Aung. Assessment of the scleral spur in anterior segment optical coherence tomography images. *Archives of Ophthalmology*, 126(2):181–185, 2008. (Cited on pages 12 and 23.)

[133] M.B. Sasongko, T.Y. Wong, T.T. Nguyen, C.Y. Cheung, J.E. Shaw, , R. Kawasaki, E.L. Lamoureux, and J.J. Wang. Retinal vessel tortuosity and its relation to traditional and novel vascular risk markers in persons with diabetes. *Current Eye Research*, 41(4):551–557, 2016. (Cited on page 91.)

[134] M.B. Sasongko, T.Y. Wong, T.T. Nguyen, C.Y. Cheung, J.E. Shaw, and J.J. Wang. Retinal vascular tortuosity in persons with diabetes and diabetic retinopathy. *Diabetologia*, 54(9):2409–2416, 2011. (Cited on page 91.)

[135] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers. Cell detection with star-convex polygons. *Medical Image Computing and Computer Assisted Intervention*, 11071:265–273, 2018. (Cited on pages 70 and 72.)

[136] M.T. Schram, S.J.S. Sep, C.J. van der Kallen, P.C. Dagnelie, A. Koster, N. Schaper, R.M.A. Henry, and C.D.A. Stehouwer. The Maastricht Study: An extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *European Journal of Epidemiology*, 29(6):439–451, 2014. (Cited on pages 84 and 91.)

[137] B. Selig, K. A. Vermeer, B. Rieger, T. Hillenaar, and C. L. Luengo Hendriks. Fully automatic evaluation of the corneal endothelium from in vivo confocal microscopy. *BMC Medical Imaging*, 15(1):13, 2015. (Cited on page 76.)

[138] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. (Cited on pages 99 and 102.)

[139] M. S. Sharif, R. Qahwaji, E. Shahamatnia, R. Alzubaidi, S. Ipson, and A. Brahma. An efficient intelligent analysis system for confocal corneal endothelium images. *Computer Methods and Programs in Biomedicine*, 122(3):421–436, 2015. (Cited on page 76.)

[140] N. Sharma, P. Sahay, P.K. Maharana, P. Kumar, S. Ahsan, and J.S. Titiyal. Microscope integrated intraoperative optical coherence tomography-guided dmek in corneas with poor visualization. *Clinical Ophthalmology*, 14:643–651, 2020. (Cited on pages 49 and 61.)

[141] E.H. Shortliffe and M.J. Sepúlveda. Clinical decision support in the era of artificial intelligence. *Journal of the American Medical Association*, 320(21):2199–2200, 2018. (Cited on page 61.)

[142] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. (Cited on page 84.)

[143] P. Steven, C. Le Blanc, K. Velten, E. Lankenau, M. Krug, S. Oelckers, L.M. Heindl, U. Gehlsen, G. Hüttmann, and C. Cursiefen. Optimizing descemet membrane endothelial keratoplasty using intraoperative optical coherence tomography. *JAMA Ophthalmology*, 131(9):1135–1142, 2013. (Cited on pages 49, 53, 54, and 61.)

[144] A. J. Stuart, V. Romano, G. Virgili, and Shortt A. J. Descemet's membrane endothelial keratoplasty (DMEK) versus Descemet's stripping automated endothelial keratoplasty (DSAEK) for corneal endothelial failures. *Cochrane Database of Systematic Reviews*, 6, 2018. (Cited on pages 9 and 49.)

[145] G. D. Sturrock, E. S. Sherrard, and N. S. C. Rice. Specular microscopy of the corneal endothelium. *British Journal of Ophthalmology*, 62(12):809–814, 1978. (Cited on page 67.)

[146] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception architecture for computer vision. *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016. (Cited on page 40.)

[147] M. Tan and Q.V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*, 97:6105–6114, 2019. (Cited on page 94.)

[148] D. Tellez, M. Balkenhol, I. Otte-Höller, R. Van De Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer, G. Litjens, J. Van Der Laak, and F. Ciompi. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Transactions on Medical Imaging*, 37(9):2126–2136, 2018. (Cited on page 63.)

[149] B.M. ter Haar Romeny, E.J. Bekkers, J. Zhang, S. Abbasi-Sureshjani, F. Huang, R. Duits, B. Dashtbozorg, T.T.J.M. Berendschot, I. Smit-Ockeloen, K.A.J. Eppenhof, J. Feng, J. Hannink, J. Schouten, M. Tong, H. Wul, H.W. van Triest, S. Zhu, D. Chen, W. He, L. Xu, P. Hand, and Y. Kang. Brain-inspired algorithms for retinal image analysis. *Machine Vision and Applications*, 27(8):1117–1135, 2016. (Cited on pages 83 and 91.)

[150] D. S. W. Ting, L. R. Pasquale, L. Peng, J. P. Campbell, A. Y. Lee, R. Raman, G. S. W. Tan, and L. Schmetterer. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103:167–175, 2019. (Cited on pages 4, 12, 27, and 83.)

[151] D.S.W. Ting, C. Yim-Lui Cheung, G. Lim, G.S.W. Tan, N.D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I.Y.S. Yeo, S.Y. Lee, E.Y.M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N.C. Tan, and E.A. Finkelstein et. al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22):2211–2223, 2017. (Cited on page 91.)

[152] M. Treder, J. L. Lauermann, M. Alnawaiseh, and N. Eter. Using deep learning in automated detection of graft detachment in Descemet membrane endothelial keratoplasty: A pilot study. *Cornea*, 38(2):157–161, 2019. (Cited on pages 22, 27, and 61.)

[153] S. Valipour, M. Siam, M. Jagersand, and N. Ray. Recurrent fully convolutional networks for video segmentation. *Winter Conference on Applications of Computer Vision*, page 29–36, 2017. (Cited on page 62.)

[154] B. M. van der Velden, M. Veta, J. P. W. Pluim, M. Alberti, and F. G. Heslinga. Radial U-Net: Improving DMEK Graft Detachment Segmentation in Radial AS-OCT Scans. *Ophthalmic Medical Image Analysis*, pages 72–81, 2021. (Cited on pages 77 and 107.)

[155] E. van 't Riet, J.M. Dekker, Q. Sun, G. Nijpels, F.B. Hu, and R.M. van Dam. Role of adiposity and lifestyle in the relationship between family history of diabetes and 20-year incidence of type 2 diabetes in U.S. women. *Diabetes Care*, 33(4):763–767, 2010. (Cited on page 91.)

[156] I. Vasiliauskaitė, S. Oellerich, L. Ham, I. Dapena, L. Baydoun, K. van Dijk, and G. R. J. Melles. Descemet membrane endothelial keratoplasty: Ten-year graft survival and clinical outcomes. *American Journal of Ophthalmology*, 217:114–120, 2020. (Cited on page 38.)

[157] P.B. Veldman, P.K. Dye, J.D. Holiman, Z.M. Mayko, C.S. Sales, M.D. Straiko, J.D. Galloway, and M.A. Terry. The s-stamp in descemet membrane endothelial keratoplasty safely eliminates upside-down graft implantation. *Ophthalmology*, 123(1):161–164, 2016. (Cited on page 49.)

[158] M. Veta, P.J. van Driest, and J.P.W. Cutting out the middleman: measuring nuclear area in histopathology slides without segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 11314:632–639, 2016. (Cited on page 3.)

[159] J. P. Vigueras-Guillén, E.R. Andrinopoulou, A. Engel, H. G. Lemij, J. van Rooij, K. A. Vermeer, and L.J. van Vliet. Corneal endothelial cell segmentation by classifier-driven merging of oversegmented images. *IEEE Transactions on Medical Imaging*, 37(10):2278–2289, 2018. (Cited on pages 68 and 76.)

[160] J. P. Vigueras-Guillén, B. Sari, S. F. Goes, H. G. Lemij, J. van Rooij, K. A. Vermeer, and L.J. van Vliet. Fully convolutional architecture vs sliding-window CNN for corneal endothelium cell segmentation. *BMC Biomedical Engineering*, 1(1):Article number: 4, 2019. (Cited on pages 76 and 77.)

[161] J. P. Vigueras-Guillén, J. van Rooij, A. Engel, H. G. Lemij, L.J. van Vliet, and K. A. Vermeer. Deep learning for assessing the corneal endothelium from specular microscopy images up to 1 year after ultrathin-DSAEK surgery. *Translational Vision Science and Technology*, 9(2):1–12, 2020. (Cited on pages 68 and 77.)

[162] M.H. Vu, G. Grimbergen, T. Nyholm, and T. Löfstedt. Evaluation of multi-slice inputs to convolutional neural networks for medical image segmentation. *Medical Physics*, 47(12):6216–6231, 2020. (Cited on page 62.)

[163] S. B. Wang, E. E. Cornish, J. R. Grigg, and P. J. McCluskey. Anterior segment optical coherence tomography and its clinical applications. *Clinical % Experimental Optometry*, 102(3):195–207, 2019. (Cited on page 27.)

[164] Y. Z. Wang, D. Galles, M. Klein, D. G. Locke, and D. G. Birch. Application of a deep machine learning model for automatic measurement of EZ Width in SD-OCT Images of RP. *Translational Vision Science & Technology*, 9(2):15, 2020. (Cited on pages 38 and 40.)

[165] J. Wasielica-Poslednik, A.K. Schuster, L. Rauch, J. Glaner, A. Musayeva, J.C. Riedl, N. Pfeiffer, and A. Gericke. How to avoid an upside-down orientation of the graft during descemet membrane endothelial keratoplasty. *Journal of Ophthalmology*, page Article ID 7813482, 2019. (Cited on page 49.)

[166] J.P. Whitcher, M. Srinivasan, and M.P. Upadhyay. Corneal blindness: a global perspective. *Bulletin of the World Health Organization*, 79:214–221, 2001. (Cited on page 67.)

[167] K. R. Wilhelmus, J. Sugar, R. A. Hyndiuk, and R. D. Stulting. Corneal thickness changes during herpes simplex virus disciform keratitis. *Cornea*, 23(2):154–157, 2006. (Cited on page 27.)

[168] B.M. Williams, D. Borroni, R. Liu, Y. Zhao, J. Zhang, J. Lim, B. Ma, V. Romano, H. Qi, M. Ferdousi, I.N. Petropoulos, G. Ponirakis, S. Kaye, R.A. Malik, U. Alam, and Y. Zheng. An artificial intelligence-based deep learning algorithm for the diagnosis of diabetic neuropathy using corneal confocal microscopy: a development and validation study. *Diabetologia*, 63(2):419–430, 2020. (Cited on page 77.)

[169] J.M. Wolterink, T. Leiner, and I. Išgum. Graph convolutional networks for coronary artery segmentation in cardiac ct angiography. *International*

Workshop on Graph Learning in Medical Imaging, pages 62–69, 2019. (Cited on page 62.)

[170] H. T. Wong, M. C. Lim, L. M. Sakata, H. T. Aung, N. Amerasinghe, D. S. Friedman, and T. Aung. High-definition optical coherence tomography imaging of the iridocorneal angle of the eye. *Archives of Ophthalmology*, 127(3):256–260, 2009. (Cited on page 13.)

[171] D. E. Worrall, C. M. Wilson, and G. J. Brostow. Automated retinopathy of prematurity case detection with convolutional neural networks. *Deep Learning and Data Labeling for Medical Applications*, pages 68–76, 2016. (Cited on page 76.)

[172] B. Y. Xu, M. Chiang, Pardeshi A. A., S. Moghimi, and R. Varme. Deep neural network for scleral spur detection in anterior segment OCT images: The Chinese American eye study. *Translational Vision Science & Technology*, 9(2):18, 2020. (Cited on page 27.)

[173] D. Xu, W.J.J. Dupps, S.K. Srivastava, and J.P. Ehlers. Automated volumetric analysis of interface fluid in descemet stripping automated endothelial keratoplasty using intraoperative optical coherence tomography. *Investigative Ophthalmology Visual Science*, 55(9):5610–5615, 2014. (Cited on pages 61 and 62.)

[174] R. Y. Yeh, R. Quilendrino, F. U. Musa, V. S. Liarakos, I. Dapena, and G. R. J. Melles. Predictive value of optical coherence tomography in graft attachment after Descemet's membrane endothelial keratoplasty. *Ophthalmology*, 120(2):240–245, 2013. (Cited on page 9.)

[175] J. Zhang, B. Dashtbozorg, F. Huang, T.T.J.M. Berendschot, and B.M. ter Haar Romeny. Analysis of retinal vascular biomarkers for early detection of diabetes. *Lecture Notes in Computational Vision and Biomechanics*, 27:811–817, 2017. (Cited on page 83.)

[176] T. Y. Zhang and C. Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984. (Cited on page 14.)

Summary

AI for your eye: deep learning for corneal and retinal image analysis

Ophthalmologists (eye doctors) are able to diagnose a wide variety of eye disorders such as cataract, diabetic retinopathy, and corneal dystrophy. Various treatments have been developed and improved, sometimes resulting in technically demanding and complex procedures. Imaging technology to visualize the different parts of the eye is often essential for effective diagnosis and treatment. Interpretation of these images requires a lot of time from specialists and if (part of the) assessment can be done automatically, it could save ophthalmologists a lot of time. In addition, automatic image analysis can be used to obtain objective measures, such as (changes in) corneal thickness.

In recent years, deep learning has shown to be very promising for automatic medical image analysis. Deep learning is a subset of machine learning techniques in which features are automatically learned from the image data. In this thesis, deep learning is used to address multiple unsolved challenges in corneal and retinal image analysis. We do so for a variety of imaging modalities and diseases by integrating state-of-the-art methods and development of novel image analysis pipelines.

In chapter 2, a method is proposed for the assessment of donor tissue detachment after corneal transplantation surgery. Sometimes the donor tissue does not fully attach, and it is important to know the extent of the detachment. We develop a deep learning-based image analysis pipeline that automatically crops the relevant part of the optical coherence tomography (OCT) image, segments the detached tissue, quantifies the extent of detachment, and constructs a map of the detached regions.

In chapter 3, we employ the automatic cropping method proposed in chapter 2 and compare three deep learning techniques for automatic corneal thickness

measurement. We develop and evaluate on the same set of OCT data set used in chapter 2, acquired after transplantation surgery. Furthermore, we construct detailed thickness maps that allow easy inspection and progress tracking.

In chapter 4, we analyze OCT images that are obtained during transplantation surgery. The challenge here is to assess whether the donor tissue (graft) is positioned correctly or upside-down. We propose an image analysis strategy that consists of deep learning-based segmentation and post-processing to obtain the graft's curvature at each point. Subsequently, we relate the graft's curvature to its orientation.

In chapter 5, we analyse specular microscopy images that show individual corneal endothelial cells. Endothelial cell density (ECD) is an important biomarker of corneal health and requires accurate segmentation of corneal endothelial cells. Current methods for extracting ECD can be insufficient when image quality is suboptimal or if guttae are present. In this chapter, we present a novel deep learning method for accurate cell segmentation in specular microscopy images of varying quality and in the presence of guttae.

In chapter 6, we focus on the retina by analyzing fundus photography images. We develop and compare deep learning methods for classification of images that belong to individuals with normal glucose metabolism and those with type 2 diabetes. We also investigate the effect of simultaneous prediction of classical features, and we compare strategies for aggregating image-level predictions to individual-level predictions.

In chapter 7, we use the best practices of chapter 6 to develop a deep learning framework for detection of type 2 diabetes in a set of over 46,000 fundus images. This time we also evaluate how well the model can be used to distinguish prediabetes individuals. Moreover, we investigate how the discriminative power of the fundus images compares to that of typical diabetes risk factors such as age, sex, and waist circumference.

Our research shows the value of deep learning for corneal and retinal image analysis. In the next decades, we can expect deep learning to find its way into ophthalmic care. Arguments will include improved patient care, additional safety, objectivity, and resource savings. But whichever the motivation, at some point, physicians will be likely to suggest AI for your eye.

List of abbreviations

AI - Artificial Intelligence

ANOVA - Analysis of Variance

AOD - Angle-Opening Distance

AS-OCT - Anterior Segment Optical Coherence Tomography

AUC - Area Under the Curve

CCT - Central Corneal Thickness

CDSS - Clinical Decision Support System

CI - Confidence Interval

CNN - Convolutional Neural Network

CV - Coefficient of Variance

DMEK - Descemet Membrane Endothelial Keratoplasty

DL - Deep Learning

DR - Diabetic Retinopathy

ECD - Endothelial Cell Density

GMscore - Glucose Metabolism Score

GPU - Graphics Processing Unit

iOCT - Intraoperative Optical Coherence Tomography

MEA - Mean Absolute Error

MAPE - Mean Absolute Percentage Error

MTL - Multi-Target Learning

OCT - Optical Coherence Tomography

ROC - Receiver Operating Characteristic

SD - Standard Deviation

SF6 - Sulfur Hexafluoride

T2D - Type 2 Diabetes

TISA - Trabecular-Iris Space Area

TTA - Test-Time Augmentation

WBCE - Weighted Binary Cross-Entropy

About the author

Friso Gerben Heslinga was born in Wageningen, the Netherlands, on November 23, 1990. Friso did his secondary education at Pantarijn in Wageningen where he was actively involved in the student council, the representative council, and the school newspaper. In 2009 Friso moved to Enschede where he studied biomedical engineering (BME) at the University of Twente. For his Master BME, he specialized in biomedical physics and imaging, supervised by Prof. Bennie ten Haken. Friso did an internship at the University of Western Australia, in Perth, where he worked with Prof. Tim St. Pierre on the development of an MRI phantom. For his Master thesis, Friso went to the University of California in Berkeley. Under supervision of Prof. Steve Conolly he compared three imaging modalities for their sensitivity for stem cell tracking. While writing his Master's thesis, Friso started a second Master's program in health sciences, specializing in health technology assessment. Here, Friso compared new imaging technologies for the hybrid operating room, supervised by Dr. Marjan Hummel.

Friso continued as a junior researcher in the Health Technology and Services Research group. As part of this year, Friso was a visiting researcher at Harvard Medical School in Boston, where he compared the value of new technologies for the AMIGO suite in the Brigham and Women's Hospital, with Dr. Tina Kapur. In Boston Friso got intrigued by the enormous potential of deep learning for medical image analysis and he soon applied for a PhD at the Medical Image Analysis Group of Eindhoven University of Technology. Since 2017, he has been a PhD candidate under the supervision of Prof.dr. Josien Pluim and dr. Mitko Veta.

Publication list

Journals

F.G. Heslinga, R.T. Lucassen, M.A. Berg, L. van der Hoek, J.P.W. Pluim, J. Cabrerizo, M. Alberti, M. Veta. Corneal pachymetry by AS-OCT after Descemet's membrane endothelial keratoplasty. *Scientific Reports*, 11(Article number: 13976), 2021.

F.G. Heslinga, M. Alberti, J.P.W. Pluim, J. Cabrerizo, M. Veta. Quantifying graft detachment after Descemet's membrane endothelial keratoplasty with deep convolutional neural networks. *Translational Vision Science & Technology*, 9(2):48, 2020.

M.B. Muijzer, **F.G. Heslinga**, F. Couwenberg, H.J. Noordmans, A. Oahalou, J.P.W. Pluim, M. Veta, R.P.L. Wisse. Automatic evaluation of graft orientation during Descemet membrane endothelial keratoplasty using intraoperative OCT. *Biomedical Optics Express*, 13(5):2683-2694, 2022.

F.G. Heslinga, T.T.J.M. Berendschot, M.T. Schram, C.D.A. Stehouwer, M.J. van Greevenbroek, J.P.W. Pluim, A.J.H.M Houben, M. Veta. The use of fundus images to discriminate prediabetes and type 2 diabetes. The Maastricht Study. *In preparation*.

F.G. Heslinga D.F.M. Timmers, G.B.A. Haijen, S.L. Dunker, T.T.J.M. Berendschot, J.P.W. Pluim, M.M. Dickman, M. Veta. Automatic and robust corneal endothelial cell density measurement in the presence of guttae. *In preparation*.

A. Mehrtash, M. Ghafoorian, G. Pernelle, A. Ziaei, **F.G. Heslinga**, K. Tuncali, A. Fedorov, R. Kikinis, C.M. Tempany, W.M. Wells, P. Abolmaesumi, T. Kapur.

Automatic needle segmentation and localization in MRI with 3-D convolutional neural networks: Application to MRI-targeted prostate biopsy. *IEEE Transactions on Medical Imaging*, 38(4):1026-1036, 2018.

G. Captur, P. Gatehouse, K.E. Keenan, **F.G. Heslinga**, R. Bruehl, M. Prothmann, M.J. Graves, R.J. Eames, C. Torlasco, G. Benedetti, J. Donovan, B. Ittermann, R. Boubertakh, A. Bathgate, C. Royet, W. Pang, R. Nezafat, M. Salerno, P. Kellman, J.C. Moon. A medical device-grade T1 and ECV phantom for global T1 mapping quality assurance - the T1 Mapping and ECV Standardization in cardiovascular magnetic resonance (T1MES) program. *Journal of Cardiovascular Magnetic Resonance*, 18(1):58, 2016.

J.K. van Zandwijk, F.F.J. Simonis, **F.G. Heslinga**, E.I.S. Hofmeijer, R.H. Geelkerken, B. ten Haken. Comparing the signal enhancement of a gadolinium based and an iron-oxide based contrast agent in low-field MRI. *PloS One*, 16(8):e0256252, 2021.

J.M.H. Noothout, N. Lessmann, M.C. van Eede, L.D. van Harten, E. Sogancioglu, **F.G. Heslinga**, M. Veta, B. van Ginneken, I. Išgum. Medical image segmentation through CNN ensembles with knowledge distillation. *In preparation*.

R.T. Lucassen, M.H. Jafari, N. Duggan, N. Jowkar, A. Mehrtash, C. Fischetti, K. Prentice, E. Duhaime, P. Abolmaesumi, **F.G. Heslinga**, M. Veta, M.A.D. Mendicuti, S. Frisken, P. Shyn, E. Boyer, W.M. Wells, A.J. Goldsmith, T. Kapur. Comprehensive multi-level deep learning for detection of B-lines using a new open lung ultrasound dataset. *In preparation*.

Conference proceedings

B.M Velden, M. Veta, J.P.W. Pluim, M. Alberti, **F.G. Heslinga**. Radial U-Net: Improving DMEK Graft Detachment Segmentation in Radial AS-OCT Scans. *International Workshop on Ophthalmic Medical Image Analysis*, 8:72-81, 2021.

F.G. Heslinga, J.P.W. Pluim, A.J.H.M. Houben, M.T. Schram, R.M.A. Henry, C.D.A. Stehouwer, M.J. van Greevenbroek, T.T.J.M. Berendschot, M. Veta. Direct Classification of Type 2 Diabetes From Retinal Fundus Images in a Population-based Sample From The Maastricht Study. *SPIE Medical Imaging: Computer-*

Aided Diagnosis 11314, 2020.

F.G. Heslinga, H. Koffijberg, R.H. Geelkerken, R. Meerwaldt, T.G. ter Mors, C.J.M. Doggen, M. Hummel. Value based decision support to prioritize development of innovative technologies for image-guided vascular surgery in the hybrid operating theater. *SPIE Medical Imaging: Image-Guided Procedures, Robotic Interventions, and Modeling 11315*, 2020.

F.G. Heslinga, J.P.W. Pluim, B. Dashtbozorg, T.T.J.M. Berendschot, A.J.H.M. Houben, R.M.A. Henry, M. Veta. Approximation of a pipeline of unsupervised retina image analysis methods with a CNN. *SPIE Medical Imaging: Image Processing 10949*, 2019.

F.G Heslinga, S. Bruns, E. Yu, P. Keselman, X.Y. Zhou, B. Zheng, S. Waanders, P.W. Goodwill, M. Wendland, B. Haken, S.M. Conolly. Stem cell tracking potential of magnetic particle imaging compared with 19F magnetic resonance imaging. *International Workshop on Magnetic Particle Imaging (IWMPI)*, 2016.

Master Theses

Biomedical Engineering - University of Twente - The future of stem cell tracking: a comparison between magnetic particle imaging and 1H and 19F magnetic resonance imaging.

Health Sciences - University of Twente - Value based decision support to prioritize innovative technologies for vascular surgery in the hybrid operating theater.

Acknowledgments

As a PhD candidate I have had help from a considerable number of people, and this thesis would not have been the same without them. I owe a lot of thanks to everyone involved.

In the first place, to Mitko Veta. I have been very lucky to have you as a first supervisor. You are very knowledgeable, friendly, and funny, which are arguably the key skills for a good PhD supervisor. Also, you have been very patient and supportive while I was getting to know the ins and outs of deep learning, and you gave me a lot of freedom to decide on the direction of the research.

To my promoter, Josien Pluim, on whom I could always count for extensive feedback and a quick reply to any question. You have been very helpful and kind. Also to my promoter Milan Petkovic, for the good talks and your detailed feedback on the thesis.

To Prof. Rudy Nuijts, Prof. Clarisa Sánchez Gutiérrez, Prof. Maarten Steinbuch, Dr. Tos Berendschot, thank you for joining my PhD committee and for assessing the thesis. It is a privilege to have so much expertise from different fields getting together.

To the students who I supervised, and whose BEPs and Master's projects have very much contributed to this thesis. PhD research can be very individualistic, especially when working from home. The frequent meetings with you have been a lot of fun, and the fresh and creative discussions have been very motivating.

To Boy Houben, Mor Dickman, Suryan Dunker, Gilles Haijen, Dieter Timmers, Bas Muijzer, Robert Wisse, Floor Couwenberg, Javier Cabrerizo, Ruben Lucassen, Myrthe van den Berg, Luuk van der Hoek, and all other collaborators and co-authors involved. Needless to say, without you the work in this thesis would not have been possible.

To Mark Alberti, who reached out to us with the most interesting research opportunities. I am very happy that you did, and it ended up redefining the focus of my PhD research. I have really enjoyed working together and our collaboration has been fruitful and a lot of fun.

I would also like to thank a few people who have been at the basis of my scientific curiosity and career. Sebastiaan Waanders, Bennie ten Haken, Marjan Hummel, Erik Koffijberg, and Carine Doggen, from the University of Twente, thank you for all your help and showing me much fun science can be. Tim St. Pierre, Steve Conolly, and Tina Kapur, thank you for the opportunities you gave me for getting to know the international scientific environment. Being in each of your labs has been very inspirational and has led me to pursue a PhD.

Suzanne, we started this PhD journey together, always in the same office, and always open to discuss anything. Thank you for all the gezelligheid. Thanks to Koen, Maxime, Andrey, Seb, Samaneh, Veronika, and Maureen for all the good discussions and everything I learned from you. Also, thanks to all other members of the Medical Image Analysis Group, it was a pleasure having as colleagues.

Steffen, Julia, Sanne, Cristina, and Boukje, thank you for experiencing the PhD-life together, for placing things in perspective, for the coffee talks, and for all the fun experiences we had.

I have been very lucky with the unconditional support I received from family and friends during the last four years. I know I haven't expressed this enough, but without you I would not have been where I am now and I am very grateful for that.

A special shout-out to the friends who were ready to make the most out of every weekend during the past four years. Daniel, Ilse, Jade, Kirsten, Lieke, Mirjam, Niels, Nienke, Ruby, Sjoerd, Tijs, Tycho, Ward, and everyone else, the outlook of an adventurous weekend with you has made sure I started every week with a big smile.

Laura, you are amazing. I was considering some kind words for your support, but you get to be on the cover of the thesis instead.